한국자료분석학회

사무국: (우)06650 서울특별시 서초구 반포대로 24길 76, 6층 602호(서초동)
전화: (02)6673-0709 / 회장 : 한상태 / 총무이사 : 이성건

문서번호 : KDAS-2401000

시행일자 : 2024. 1. 26.

(경유)

수 신 : 전 회원

참 조 :

선결			지시		
접수	일자				
	시간		결재		
	번호				
담당자					

제목: 한국자료분석학회 2023년 동계 학술논문발표대회

1. 귀 회원님의 무궁한 발전을 기원하며, 본 학회의 발전을 위한 회원 여러분의 관심과 노고에 깊이 감사를 드립니다.

2. 본 학회 2023년도 동계 학술논문발표대회를 다음과 같이 개최함을 알려 드리오니 많은 참여 바랍니다.

   1) **개최일자 : 2024년 1월 25일(목) ~ 26일(금)**

   2) **장 소 : 부경대학교 대연캠퍼스(부산광역시 남구 용소로 45)**

한 국 자 료 분 석 학 회 장 한 상 태

PROCEEDINGS OF THE KOREAN DATA ANALYSIS SOCIETY

January 25-26, 2024

at

Pukyong National University, Busan, Korea.

(https://www.pknu.ac.kr)

THE KOREAN DATA ANALYSIS SOCIETY Founded 1998

PROCEEDINGS OF THE KOREAN DATA ANALYSIS SOCIETY January 26, 2024

한국자료분석학회 2023년도 동계 학술논문발표대회 준비위원

조직위원회

위 원 장	한 상 태	호서대학교 빅데이터AI학부
부위원장	이 학 배	연세대학교 응용통계학과
	김 기 환	고려대학교 빅데이터사이언스학부
	김 용 대	서울대학교 통계학과
	홍 재 범	부경대학교 경영학부
	이 영 섭	동국대학교 통계학과
	윤 성 민	부산대학교 경제학부
	박 찬 근	한국해양대학교 데이터사이언스전공
	박 승 열	㈜ 케이스탯리서치 회장
	노 맹 석	부경대학교 데이터정보과학부
위 원	이 성 건	성신여자대학교 수리통계데이터사이언스학부
	조 형 준	고려대학교 통계학과
	최 호 식	서울시립대학교 도시빅데이터융합학과
	진 서 훈	고려대학교 빅데이터사이언스학부
	허 태 영	충북대학교 정보통계학과
	이 동 희	경기대학교 경영학부
편집위원장	강 현 철	호서대학교 빅데이터AI학부
자문위원	최 종 후	고려대학교 응용통계학과
	조 완 현	전남대학교 통계학과
	박 희 창	창원대학교 통계학과
	강 창 완	동의대학교 산업경영·빅데이터공학과

운영위원회

공동위원장	조 형 준	고려대학교 통계학과
	최 호 식	서울시립대학교 도시빅데이터융합학과
부위원장	이 성 건	성신여자대학교 수리통계데이터사이언스학부
	신 승 준	고려대학교 통계학과
	김 지 수	가천대학교 간호학과
	제 상 영	고려대학교 경제통계학과
	조 장 식	경성대학교 정보통계학과
	노 맹 석	부경대학교 데이터정보과학부
위 원	김 지 희	강원대학교 응급구조학과
	김 태 훈	경성대학교 경제금융물류학부
	김 성 환	건국대학교 응용통계학과
	김 양 진	숙명여자대학교 통계학과
	곽 일 엽	중앙대학교 통계학과
	박 만 식	성신여자대학교 수리통계데이터사이언스학부
	김 은 석	㈜ GDS 컨설팅그룹 대표이사
	김 지 연	㈜ 케이스탯리서치 대표이사

한국자료분석학회
2023년도 동계 학술논문발표대회

- 일시 : 2024년 1월 25일(목) ~ 1월 26일(금)
- 장소 : 부경대학교(대연캠퍼스)
- 공동주최 : 한국자료분석학회, 부경대학교 경영학부, 데이터정보과학부
- 주관 : 부경대학교 경영학부, 데이터정보과학부
- 후원: 한국연구재단, (주)케이스탯리서치, 트랜드리서치, BNK 투자증권, (주)GDS컨설팅, 데이터솔루션, 나이스평가정보, 부경대학교

● 1월 25일 목요일

17:00 - 19:00	편집위원회
17:00 - 19:00	이사회

● 1월 26일 금요일

09:00 - 18:00	포스터발표(경영관 1층)
09:30 - 11:00	논문발표 A1 - A8(경영관 1013, 1012, 808, 810, 611, 612, 613, 104)
11:10 - 12:10	튜토리얼(경영관 1013호) 좌장: 최호식 교수(학술이사, 서울시립대학교)
	제목 : 구조방정식모형을 이용한 연구에서 검토해야 할 체크리스트 강사 : 강현철 교수(호서대학교 빅데이터AI학부)
12:10 - 12:30	총회(경영관 1013호) 진행: 이성건 교수(총무이사, 성신여자대학교)
12:30 - 13:40	중식 시간(라일락 구내식당, 12시 30분 시작)
13:40 - 14:40	학회장 초청 강연(경영관 1013호) 좌장: 한상태 교수(학회장, 호서대학교)
	제목 : 데이터 과학과 디지털 인문학 강사 : 장원철 교수(서울대학교 통계학과)
14:50 - 16:20	논문발표 B1 - B8(경영관 1013, 1012, 808, 810, 611, 612, 613, 104)
16:30 - 18:00	논문발표 C1 - C8(경영관 1013, 1012, 808, 810, 611, 612, 613, 104)

논문발표 A: 09:30 - 11:00

논문발표 A1(경영관 1012호) **Data Science PLUS 기획세션** 좌장 : 최호식(서울시립대학교)

09:30 | 11:00

논문발표 A2(경영관 1013호) **의료/간호/보건** 좌장 : 박민희(동서대학교)

09:30 | 11:00

논문발표 A3(경영관 808호) **공공데이터의 활용** 좌장 : 전새봄(목원대학교)

09:30 | 11:00

논문발표 A4(경영관 810호) **통계** 좌장 : 이은지(충남대학교)

09:30 | 11:00

논문발표 A5(경영관 611호) 재무-금융 BNK 투자증권 특별세션 좌장 : 이용웅(한국외국어대학교)

09:30 | 11:00

논문발표 A6(경영관 612호) 교육/심리/사회 좌장 : 정혜원(충남대학교)

09:30 | 11:00

논문발표 A7(경영관 613호) **통계** 좌장 : 연규필(호서대학교)

09:30 | 11:00

논문발표 A8(경영관 104호) **통계** 좌장 : 정호현(성신여자대학교)

09:30 | 11:00

논문발표 B: 14:50 - 16:20

논문발표 B1(경영관 1012호) **데이터 사이언스와 통계적 기법의 융합** 좌장: 이영섭(동국대학교)

14:50 | 16:20

논문발표 B2(경영관 1013호) **생존분석** 좌장: 김양진(숙명여자대학교)

14:50 | 16:20

논문발표 B3(경영관 808호) **패널자료분석** 좌장 : 정병철(서울시립대학교)

14:50 | 16:20

논문발표 B4(경영관 810호) **신뢰성자료분석** 좌장 : 장인홍(조선대학교)

14:50 | 16:20

논문발표 B5(경영관 611호) **경제/경영** 좌장 : 송철종(선문대학교)

14:50 | 16:20

논문발표 B6(경영관 612호) **경제/경영** 좌장: 김명준(공주대학교)

14:50 | 16:20

논문발표 B7(경영관 613호) **통계** 좌장: 김혁주(원광대학교)

14:50 | 16:20

논문발표 B8(경영관 104호) **Data Science PLUS 기획세션** 좌장 : 김성용(호서대학교)

14:50 | 16:20

논문발표 C: 16:30 - 18:00

논문발표 C1(경영관 1012호) **신호처리를 위한 딥러닝 및 딥러닝 응용** 좌장 : 곽일엽(중앙대학교)

16:30 | 18:00

논문발표 C2(경영관 1013호) **보건의료** 좌장 : 김광환(건양대학교)

16:30 | 18:00

논문발표 C3(경영관 808호) 시뮬레이션 기반 시그널 탐지 통계 모델 성능 연구 좌장 : 김성환(건국대학교)

16:30 | 18:00

논문발표 C4(경영관 810호) 조사연구에서의 머신러닝 활용 좌장 : 강창완(동의대학교)

16:30 | 18:00

논문발표 C5(경영관 611호) **경제/경영** 좌장 : 김태훈(경성대학교)

16:30 | 18:00

논문발표 C6(경영관 612호) **의료/간호/보건** 좌장 : 이재우(고려대학교)

16:30 | 18:00

논문발표 C7(경영관 613호) **경제/경영** 좌장 : 이동희(경기대학교)

16:30 | 18:00

논문발표 C8(경영관 104호) **고려대학교 BK21 통계학교육연구팀 기획세션** 좌장 : 박관영(성신여자대학교)

16:30 | 18:00

포스터발표(경영관 1층) 09:00-18:00

포스터발표(경영관 1층) 09:00-18:00

포스터발표(경영관 1층) 09:00-18:00

튜토리얼

(1) **발표자 : 강현철 (호서대학교 빅데이터AI학부)**

(2) 제목 : 구조방정식모형을 이용한 연구에서 검토해야 할 체크리스트

구조방정식모형은 최근 매우 다양한 분야에서 가장 광범위하게 사용되는 자료분석 기법 중 하나이다. 구조방정식모형이 널리 쓰이는 이유로 다음과 같은 장점을 들 수 있다. 첫째, 구조방정식모형에서는 측정오류(measurement error)가 통제될 수 있다. 둘째, 매개변수(mediator variable)의 사용이 용이하다. 셋째, 이론모형(theoretical model)에 대한 통계적 평가가 가능하다. 즉, 연구자가 구축한 이론모형이 실제 자료에 얼마나 부합되는지를 평가하여 이를 바탕으로 연구자는 그 모형을 타당한 모형으로 받아들이거나 수정할 수 있다.

구조방정식모형의 자료분석 과정은 상당히 복잡하기 때문에 연구 방법, 연구 결과 및 논의를 보고하는 데 있어서 적절한 형식의 일치성을 보이기 어렵다. 이는 모형의 적합도, 가설적 관계의 유의성, 구조모형에서 설명되는 변이, 그리고 이론적 모형이 실제 자료와 얼마나 부합되는지 등에 대한 충분한 자료를 제시해야 하기 때문이다. 이러한 어려움으로 인해 구조방정식 모형의 장점을 효과적으로 활용하지 못하는 연구가 속출할 수 있고, 구조방정식 모형의 오용이 보고되고 있으며 구조방정식 모형의 오용과 그로 인해 나타나는 문제점을 지적하는 연구들이 수행되었다.

이번 강연에서는 구조방정식모형을 이용한 자료분석에 있어서 주의해야 할 사항들을 모형 설정 단계, 관찰변수 검토 단계, 모형 평가 및 수정 단계, 결과 보고 및 해석 단계로 구분하여 제시하였다. 실제 연구에서 연구자는 이러한 사항들을 꼼꼼히 검토함으로써 연구의 질과 서술의 완결성을 향상시킬 수 있을 것으로 기대한다.

○ 모형 설정 단계

1. 측정모형의 설정
   - ○ 각 잠재변수를 대표하여 나타낼 수 있는 관찰변수들이 설정되었는가?
   - ○ 각 잠재변수를 규정하는 관찰변수들은 동일한 개념을 측정한다고 말할 수 있는가?
   - ○ 각 잠재변수가 단일한 개념(의미)을 나타내도록 규정되어 있는가?
   - ○ 각 잠재변수들의 크기(크다/작다)를 말할 수 있는가?
2. 구조모형의 설정
   - ○ 잠재변수들 사이의 인과관계 및 연관관계가 이론적/경험적 결과에 근거하고 있는가?
   - ○ 잠재변수들 사이의 인과관계 및 연관관계가 누락된 것이 없는가?

○ 관찰변수 검토 단계

3. 타당하고 신뢰로운 측정도구에 의해 관찰변수들이 측정되었는가?
4. 관찰변수들의 기술통계량과 상관계수들이 이론적/경험적 결과와 부합하는가?

○ **모형 평가 및 수정 단계**

5. 대부분의 적합도 지수들이 우수하게 나타나고 있는가?
6. 상대적으로 큰 수정지수들이 모두 검토되었는가?
7. 모형의 간명성이 추구되었는가?
8. 동치모형 및 대안모형이 검토되었는가?
9. 측정오차변수 및 내생오차변수들의 연관관계가 검토되었는가?

○ **결과 보고 및 해석 단계**

10. 관찰변수들의 기술통계량과 상관계수들이 제시/해석되었는가?
11. 주요 적합도 지수들이 제시/해석되었는가?
12. 측정모형에 대한 주요 결과들이 적절하게 제시/해석되었는가?
13. 구조모형에 대한 주요 결과들이 적절하게 제시/해석되었는가?
14. 매개효과에 대한 주요 결과들이 적절하게 제시/해석되었는가?

※ Kang, H., Ahn J. W. (2021). Model setting and interpretation of results in research using structural equation modeling: a checklist with guide questions for reporting, *Asian Nursing Research*, 15(3), 157-162. https://doi.org/10.1016/j.anr.2021.06.001

초청강연

(1) 발표자 : 장원철 (서울대학교 통계학과)

(2) 제목 : 삶의 변화 - 돌아보다, 지켜보다, 내다보다

2003년까지 인류가 만들어 낸 자료의 크기가 5 엑사바이트 정도인데 요즘은 이틀에 한번꼴로 이 정도 규모의 자료가 생성된다고 합니다. 하지만 이러한 정보의 홍수속에서 실제로 유용한 정보를 찾는 것은 건초더미에서 바늘찾기 만큼 어러운 일입니다. 정보가 21세기의 기름이라면 분석은 연소엔진이라는 비유가 있듯이 쏟아지는 정보의 바다에서 등대를 찾아 헤메는 우리에게 데이터과학은 나침반과 같은 존재입니다.

빅데이터의 시대를 맞이하여 인류가 궁금해 왔던 많은 문제에 대해서 답변이 이루어지고 있습니다. 꿈의 재생과 유전체 정보에 대한 분석은 인류가 오랫동안 갈망해오던 시대의 서막을 알리고 있지만 한편에서는 빅데이터의 그늘을 두려워하는 시선도 존재하는 것이 사실입니다..

이러한 데이터의 시대에서 한발짝 옆에 비껴 서있는 것 같은 인문학에서도 변환의 물결은 감지되고 있습니다. 디지털 인문학은 역사, 문학등 다양한 인문학 분야의 오래된 질문에 대해 디지털화된 대규모 자료를 이용하여 새로운 답변을 얻어내고자 하는 분야입니다.

예를 들면 셰익스피어는 약 18,000개의 단어를 작품에 사용했는데 그 중 “road”, “hurry”와 같은 단어를 포함한 1700개의 단어를 그가 새로 만들어 냈다고 합니다. 그렇다면 그가 실제로 알고 있었던 총 단어의 숫자는 얼마나 될까요? 또한 그의 작품중에 진위여부에 대한 논란이 끊이지 않는 작품들에 대해서 정량적 분석을 통한 진위여부의 판단은 가능할까요?

이와 같이 문학작품을 정량적으로 분석하는 분야를 양식측정학이라고 합니다. 19세기 말에 시작된 양식측정학은 구글 도서관 프로젝트와 구텐베르크 프로젝트를 통한 대규모 문학 작품의 디지털화와 텍스트 마이닝의 등장으로 비약적인 발전을 이루게 됩니다.

이 강연에서는 디지털 인문학을 통하여 문화 예술 전반에 대한 새로운 시각을 제공하고자 합니다. 조선왕조실록의 분석을 통한 예송논쟁의 재조명에서부터 비틀즈 In my life의 실제 작곡자 발견까지, 전혀 어울리지 않는 커플 “데이터과학과 디지털 인문학”과 함께 정보의 바다로 같이 떠나보지 않으시겠습니까?

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 1-1

워드 임베딩 방법에 따른 딥러닝을 활용한 가짜뉴스 판별 모델에 관한 연구*

이진혁[1], 한상태[2]

요 약

정보화 기술과 미디어화의 가속화 속에 가짜뉴스가 우리 사회의 심각한 문제로 대두되고 있다. 이에 가짜뉴스를 판별하는 모델을 개발하여 가짜뉴스 노출에 취약한 현시점에서 유해한 영향을 줄이고 신뢰성 있는 정보 전달에 기여하고자 한다. 자연어를 처리하는 방법 중 단어 임베딩 방법을 기반으로 한 딥러닝 모델을 통해 가짜뉴스를 판별해 내고 그 모델의 성능을 높이고자 다양한 워드 임베딩 방법을 소개하고자 한다. 워드 임베딩 방법은 뉴스 텍스트 데이터에서 의미 있는 특징을 추출하고, 단어 간 의미적 유사성을 파악하는 방법인데, 이 방법을 활용하여 뉴스 기사의 실제 내용과 일치하지 않는 정보를 식별하여 가짜뉴스를 판별하고자 한다. 각 워드 임베딩 방법인 TF-IDF, Word2Vec, FastTextt의 embedding matrix를 생성하고, 가짜뉴스 판별 모델인 딥러닝 기반 LSTM 모델에 임베딩층의 가중치를 적용한 후, 모델의 정확도(accuracy)를 비교하여 더 우수한 워드 임베딩 방법을 제시하였다. 본 연구의 사례 분석을 통해 모델의 정확도를 비교 한 결과 Word2Vec 방법이 TF-IDF와 FastText보다 더 우수한 결과를 얻었다.

주요용어 : 가짜뉴스, 워드 임베딩, 딥러닝, 판별 모델, LSTM

1. 서론

오늘날 빠르게 진행되고 있는 디지털 정보화 시대에 가짜뉴스의 빠른 확산은 우리가 보는 뉴스가 진짜인지 가짜인지 구별할 시간조차 주지 않는다. 가짜뉴스 문제는 사회적으로 심각한 이슈로 부상하고 있는데, 최근만 하더라도 코로나19가 빠르게 퍼져나갈수록 코로나19와 관련해서 '코로나 바이러스는 생화학무기 개발 과정에서 유출됐다.', '코로나19는 5G 전파를 타고 확산한다.' 등 잘못된 정보가 사회에 큰 공포와 혼란을 부추겼다. 이러한 가짜뉴스가 빠르게 확산하는 형태를 정보(information)와 전염병(epidemic)의 합성어인 '인포데믹' 현상이라고 한다.

가짜뉴스는 합법적인 뉴스로 제시되어 의도적으로 조작되거나 오해의 소지가 있는 정보로 개인, 지역 사회 및 사회 전반에 여러 가지 위험을 초래한다. 따라서, 뉴스 구독자로 하여금 비판적인 태도로 뉴스를 수용해야 하며 해당 정보의 전문가를 통해 뉴스의 팩트를 구별하고, 교차 검증하는 것이 권장되지만, 이것만으로는 가짜뉴스의 빠른 생산과 확산을 막기엔 역부족이다. 이런 측면에서, 가짜뉴스를 탐지하는 모델을 개발하여 실시간으로 뿜어져 나오는 정보를 일차적으로 여과 할 수 있는 모델을 만들고자 하였다.

---

*이 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (2022M3J6A1063595).

[1]31499 충청남도 아산시 배방읍 호서로79번길 20, 호서대학교 데이터사이언스학과 석사과정. E-mail: jinhyeok_22@naver.com

[2]31499 충청남도 아산시 배방읍 호서로79번길 20, 호서대학교 빅데이터AI학부 교수. E-mail: sthan@hoseo.edu

2. 자연어 처리 방법

텍스트로 이루어진 데이터는 인간이 사용하는 언어인 자연어(Natural Language, NLP)이므로, 컴퓨터가 이해하기 위해서는 단어를 수치적으로 변환해 주는 과정이 필요하다. 단어를 표현하는 방법 중 TF-IDF(Term Frequency-Inverse Document Frequency)는 단어의 빈도와 문서 간 중요성을 고려하여 가중치를 부여하여 수치로 표현하는 방식이다.

원 핫 인코딩 방법은 단어 집합의 크기를 벡터의 차원으로 설정하고, 나타내고자 하는 단어의 인덱스에 1의 값으로 표현하고 나머지 인덱스는 0의 값으로 희소 표현 방식이다. 하지만, 이러한 단어 표현 방법은 그 단어가 무엇인지만 알려줄 뿐, 단어의 의미를 담고 있지 않다. 이러한, 한계점을 극복하고자 워드 임베딩이 제안되었는데, 본 연구에서는 TF-IDF 방법과 워드 임베딩 방법 중 Word2Vec, FastText를 제시하였다.

2.1. TF-IDF 방법

TF-IDF(Term Frequency-Inverse Document Frequency)는 단어의 빈도와 역 문서 빈도를 사용하여 단어의 상대적 중요성을 가중치를 주어 나타내는 방법이다. TF-IDF 가중치는 TF(Term Frequency)값과 IDF(Inverse Document Frequency)값을 곱한 것이다. TF값은 한 문서 내에서 특정 단어가 얼마나 자주 나타나는지를 나타낸다. 이 값은 주어진 단어가 문서 내에서 많이 등장할수록 해당 문서 내에서의 상대적인 단어의 중요도를 의미한다. 하지만, TF값만을 가지고 단어의 중요도를 표현하기에는 특정 단어가 문서 내에서 얼마나 자주 나타나는지에만 주목하게 되므로 한계가 있다(Lee and Kim, 2009).

IDF 값은 전체 문서의 수를 특정 단어가 등장한 문서의 수 (Document Frequency, DF)로 나눈 것이다. 이 값은 전체 문서에서 특정 단어가 얼마나 일반적인지를 의미한다. 상대적으로 많은 문서에 출현한 단어는 IDF값이 작아져 보편적인 단어일 가능성이 크고, 반대로 하나의 문서에 편중하여 자주 나타난 단어는 IDF값이 커져 문서 내에서 주요한 의미를 가지는 단어일 가능성이 크다. TF-IDF값은 TF값과 IDF값을 곱한 것이고 관련 수식은 다음과 같다.

$$TF-IDF(w,d) = TF(w,d) * \log\left(\frac{n}{1+df(w)}\right)$$

2.2. Word2Vec 방법

Word2Vec 방법은 기존의 피드 포워드(Feed Forward) 인공신경망 언어모델(Neural Network Language Model, NNLM)을 통해 단어 벡터 간 유사도를 구할 수 있도록 워드 임베딩의 개념에서 Goolgle의 Tomas Mikolov 외 팀원들에 의해 연구되었는데, 워드 임베딩 자체에 집중하여 NNLM의 느린 학습 속도와 정확도를 개선한 방법이다. Word2Vec은 언어 분산 표현(Distributed Representation) 방법으로 유사한 의미를 가진 단어들은 유사한 벡터값을 가진다는 분포 가설(Distributional Hypothesis)을 기반으로 단어를 학습하고, 단어의 의미를 벡터의 여러 차원에 분산하여 표현한다(Tomas et al., 2013). Word2Vec의 학습모델 방법은 CBOW(Continuous Bag of Words)와 Skip-gram 두 가지 방식이 있다. CBOW 방법은 주변에 있는 단어들(context word)을 입력으로 중간에 있는 단어들(center w

ord)을 예측하는 방법이다. 또한, Skip-gram 방법은 중심단어에서 주변 단어를 예측하는 방법이다. 본 연구에서는 가짜뉴스의 특성상 특정 대상을 중심 단어 중심으로 의도적인 목적으로 주변 단어를 생성하므로 Skip-gram 방식을 사용하였다.

2.3. FastText 방법

FastText는 2016년도에 facebook AI 연구팀에서 개발한 character n-gram과 skip-gram을 결합한 임베딩 모델이다. FastText는 Word2Vec의 한계점이었던 코퍼스(corpus)내의 모르는 단어(Out Of Vocabulary, OOV)에 대해 대처할 수 없음, 빈도수가 적은 단어(rare word)에 대해 임베딩 정확도가 높지 않음, 오타가 섞인 단어에 대한 임베딩 처리의 한계, 동음이의어의 특성을 반영하지 못함 등을 보완하고자 제안되었다.

FastText 방법은 Word2Vec 방법과 달리 하나의 단어 안에도 여러 내부 단어(subword)들이 있다 가정하고 학습한다. FastText에서 각 단어는 글자 단위 n-gram의 구성으로 취급하는데, 예를 들어 n-gram(n=3)이라면 ‘apple’에 대해서 <ap, app, ppl, ple, le>로 6개의 토큰으로 벡터화한다. 이렇게 FastText의 인공 신경망을 학습한 후에는 데이터 셋의 모든 단어의 각 n-gram에 대해서 워드 임베딩이 된다(Yoo and Ahn, 2022). 이러한 FastText의 특징으로 노이즈가 많은 코퍼스에도 강점을 가지는 것으로 알려져 있다.

3. 사례 분석

3.1. 분석 자료

본 연구의 분석 자료는 Kaggle, opendatascience, George Mclntire, Reuters가 제공하는 뉴스데이터를 병합하여 실제 뉴스 35,028개 가짜뉴스 37,106개로 구성된 72,134개의 뉴스 데이터이다. 일련번호(0부터 시작), Title(텍스트 뉴스 제목), Text(뉴스 콘텐츠 내용), Label(0 : 가짜뉴스, 1 : 진짜 뉴스) 열로 구성되었고, 분석에서는 Title과 Text를 합쳐 새로운 열을 만들어 변수로 사용하였다.

3.2. 분석 모델

LSTM(Long Short-Term Memory)은 RNN의 한 방법이다. LSTM은 RNN의 시퀀스 데이터에서 장기 의존성 문제(long-term dependencies)를 해결하기 위해 고안되었는데, 텍스트 데이터의 일련의 단어 또는 문자로 이루어진 시퀀스 데이터를 학습하여 문장의 의미를 파악하여 가짜뉴스인지 진짜뉴스인지 판별할 수 있다. 본 연구에서 이용한 LSTM 모델의 아키텍처는 Table 1과 같다.

Table 1. LSTM model Architecture

층(유형)	출력 형태	파라미터 수
embedding (임베딩 층)	(None, 700, 100)	24,182,200
lstm (LSTM 층1)	(None, 700, 128)	117,248
lstm_1 (LSTM 층2)	(None, 64)	49,408
dense (밀집 층1)	(None, 32)	2,080
dense_1 (밀집 층2)	(None, 1)	33

Table 2. Accuracy by LSTM model in valid data

Word embedding	epoch 1	epoch 2	epoch 3	epoch 4	epoch 5	epoch 6	정확도 (accuracy) 평균
TF-IDF	0.706	0.7469	0.7544	0.7542	0.7583	0.7807	.750
Word2Vec	0.893	0.948	0.928	0.954	0.961	0.967	.942
FastText	0.867	0.936	0.938	0.934	0.946	0.959	.927

3.3. 분석 결과

LSTM 모델을 이용한 분석에서 모델의 batch 크기는 128로 고정하였고, epoch는 6으로 설정하여 반복 학습하였다. 또한, 검증데이터에 대한 손실(loss)이 2회 이상 개선되지 않으면, 훈련을 조기 종료(Early Stopping) 시켰다. LSTM 분석 절차는 다음과 같다.

1) 파이썬 nltk 패키지를 사용하여 불용어를 제거하고, 어간을 추출하여 통일화(Stemming) 작업을 실시, 형태소 분석을 위해 토크나이저를 실시하는 등 데이터 전처리 작업을 시행한다.

2) 문자의 길이를 모두 700자로 맞춰주는 패딩 작업을 시행한다.

3) TF-IDF, Word2Vec, FastText 단어 임베딩을 실시하고 임베딩 행렬을 만든다.

4) LSTM의 임베딩 층에 가중치로 각 임베딩 행렬값을 투입하여 모델을 수행하고 정확도(accuracy)를 계산한다.

본 연구에서 사용한 LSTM 모델 적용은 학습용데이터(Train)와 검증용데이터(Test)를 7:3 비율로 분할한 후, 학습용 데이터를 이용하여 모델을 학습시킨 후 검증용 데이터에 적용하여 정확도를 계산하였다. 분석 결과는 Table 2.에 나타나 있다.

Table 2.결과에서 검증용데이터의 정확도(accuracy)를 계산하여 이에 대한 평균값을 구하여 제시하였다. 결과를 살펴보면, Word2Vec 방법의 정확도가 0.942로 가장 높았고, FastText가 0.927, TF-IDF가 0.750로 나타났다.

References

Lee, S. J., K. H. J. (2009). A variant of TF-IDF extracts news from electrons. *Journal of Korea Electronic Commerce Society*, 14(4), 59-73 (in Korean).

Tomas, M., Kai. C., Greg C., Jeffrey, D., “Efficient Estimation of Word Representations in Vector Space”. *In Proceedings of Workshop at ICLR*, 2013.

Tomas, M., Ilya, S., Kai, C., Greg, C., Jeffrey, D., “Distributed Representations of Words and Phrases and their Compositionality”. *In Proceedings of NIPS*, 2013.

Yoo, W. J., Ahn, S. J. (2022). *Introduction to natural language processing using deep learning*, (https://wikidocs.net/book/2155)

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 5-5

트랜스포머 모델을 활용한 건강검진 결과지 정보 추출 연구

이영[1,2], 김영화[3], 원성호[4]

요 약

문서 이해(Document understanding)는 인공지능(Artificial Intelligence, AI) 모델을 통해 문서 이미지에서 텍스트와 그림 정보를 해석하고, 사람이 필요로 하는 정보로 가공하는 작업을 의미한다. 주로 문서 분류(Document Classification), 질문 응답(Question Answering), 문서 레이아웃 분석(Document Layout Analysis), 정보 추출(Information Extraction)로 나눌 수 있다. 본 연구는 건강검진 결과지에서 사전에 정의한 검사항목과 결과치를 추출하기 위해 정보 추출 작업에 중점을 두고 진행되었다. 건강검진 결과지 양식을 선정하여 인조 데이터를 생성하고 트랜스포머 기반 Donut(Document Understanding Transformer) 모델을 미세 조정(fine tuning)하였다. 본 연구에서 사용한 모형은 간단하면서도 광학 문자 인식(Optical Character Recognition, OCR)에 의존하지 않는 특징이 있다. 또한 우수한 성능을 보여 새로운 도메인에 대해 정보 추출을 수행할 때 유용한 선택지로서 적절함을 입증하고 있다.

주요용어 : 정보추출, 딥러닝, 트랜스포머.

[1]06974 서울특별시 동작구 흑석로 84, 중앙대학교 통계학과 박사과정. E-mail: lyou7688@gmail.com
[2]05368 서울특별시 강동구 진황도로 61길 53, 중앙보훈병원 보훈의학연구소 선임연구원.
E-mail: lyou7688@gmail.com
[3](교신저자) 06974 서울특별시 동작구 흑석로 84, 중앙대학교 통계학과 교수. E-mail: gogators@cau.ac.kr
[4](교신저자) 08826 서울특별시 관악구 관악로 599, 서울대학교 보건대학원 보건학과 교수.
E-mail: won1@snu.ac.kr

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 7-7

대규모 언어 모델(LLM)의 테이블 형 데이터 분류 능력 탐구

노강준[1], 성백륜[2], 송성진[3], 송경우[4]

요 약

Large Language Model(LLM)은 강력한 사전 학습된 지식 기반을 활용하여 다양한 영역에서의 활용 가능성을 보여주었다. 테이블형 데이터 분류 작업의 경우, LLM의 성능은 XGboost나 KNN과 같은 전통적인 Machine Learning 기반 모델(ML Model)과 비교하였을 때 좋은 모습을 보여주지는 못하였다. 본 연구에서는 분류 작업에서 전통적인 ML Model들을 대체할 수 있는 수준까지 LLM의 성능을 향상하고, 기계 학습 기반 모델에는 없는 LLM의 강점을 파악하기 위해 다양한 시도를 하였다. 본 연구에서는 LLM 자체의 성능을 전통적인 ML Model들을 능가하는 수준으로 올리기 위해 출력 샘플링 방법 수정, 수치형 특징에서 범주형 특징으로의 변환, 전통적인 ML Model과의 앙상블 등의 시도를 통해 LLM의 성능을 향상했다. 또한, LLM의 Out Of Variable(OOV) 활용 가능성에 주목하였다. OOV는 사전 학습 모델을 Fine Tuning 하는 과정에서 사용된 학습 데이터에 포함되지 않은 새로운 변수를 뜻한다. 모델 추론 과정에서 OOV를 활용하는 능력에서 LLM은 전통적인 ML Model보다 우수한 성능을 보였으며, 이는 사전 학습된 광범위한 지식과 데이터를 프롬프트에 맞게 유연하게 변환하는 기능이 훈련 과정에서 학습하지 않은 새로운 특징에 적응하고 일반화하는 능력에 있어서 유용하다는 것을 보여준다. 본 연구는 OOV가 포함된 상황에서의 모델 추론에서 LLM이 강력한 도구가 될 수 있다는 것을 보여준다.

주요용어 : 테이블형 데이터, Large Language Model, 전통적인 Machine Learning Model, Out of Variable.

[1]02054 대한민국 서울특별시 서울시립대로 163 서울시립대학교 인공지능학과 학부재학생.
E-mail: shwndnjs58@uos.ac.kr
[2]02054 대한민국 서울특별시 서울시립대로 163 서울시립대학교 인공지능학과 학부재학생.
E-mail: sizzflair97@uos.ac.kr
[3]03152 대한민국 서울특별시 종로구 종로1길 42 Crepass Solutions Analytics & Consulting Team.
E-mail: goonzu0207@naver.com
[4](교신저자)03722 대한민국 서울특별시 서대문구 연세로 50 연세대학교 응용통계학과 조교수.
E-mail: kyungwoo.song@gmail.com

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 9-9

설문조사자료에 대한 자연어 문맥 적용 방안 연구*

정재경[1], 김은식[2], 최호식[3]

요 약

설문조사의 응답 데이터는 일반적으로 단일화된 수치로 정량화되어 전체적인 분포와 종합적인 반응으로서 해석된다. 그러나, 이러한 전통적인 방법은 자연어 형태의 질문과 이의 응답에 포함된 문맥(context)을 반영하기 어렵다. 거대 언어 모델은 자연어 상태의 입력을 벡터 공간으로 임베딩하여 주변 단어와의 관계로부터 문맥 정보를 학습한다. 본 연구에서는 거대 언어 모델을 활용하여 질문변수와 응답변수의 자연어 표현에 내포된 문맥을 고려하는 방법을 제안한다. 설문조사의 질문변수와, 긍정('네') 또는 부정('아니요')을 가리키는 응답변수가 일대일로 결합된 질문-응답 쌍으로부터, 다국어에 대해 사전 학습된 거대 언어 모델을 사용해 문장 임베딩 벡터를 추출한다. 고차원의 문장 임베딩 벡터는 t-SNE 차원 축소 기법으로 저차원 공간에 투영해 문맥 정보가 보존된 데이터를 생성한다. 기존의 범주화된 응답변수, 유병률 생성변수, 질문과 응답 쌍으로부터 추출된 임베딩 벡터 변수의 세 가지 데이터 유형이 XGBOOST 모델의 분류 성능에 미치는 정도를 비교함으로써, 자연어 문맥 적용의 영향을 살펴본다. 또한 설문조사자료에 대한 거대 언어 모델의 미세조정을 통해 문맥을 반영하는 방법에 대한 효과를 비교한다. 이의 자연어 문맥 적용 방안들은 실제 정신건강과 관련된 설문조사자료에 적용하여 정신건강 위험군을 분류하는데 효과적임을 실증한다.

주요용어 : 설문조사자료, 거대 언어 모델, 임베딩 백터, t-SNE, XGBOOST, Fine-tuning, Prompt-tuning.

---

*이 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (NO.2022M3J6A1084845).

[1]02504 서울특별시 동대문구 서울시립대로 163, 서울시립대학교 도시빅데이터융합학과 박사과정. E-mail: jaekyeong@uos.ac.kr

[2]02504 서울특별시 동대문구 서울시립대로 163, 서울시립대학교 수학과 학사과정. E-mail: gracely9901@uos.ac.kr

[3](교신저자) 02504 서울특별시 동대문구 서울시립대로 163, 서울시립대학교 도시빅데이터융합학과 교수. E-mail: choi.hosik@uos.ac.kr

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 11-12

흡연 학생 대상 금연 상담의 효과

김순미[1], 박민희[2]

요 약

본 연구의 목적은 END를 활용한 금연 상담이 흡연 학생의 금연 지식, 니코틴 의존도, 금연 자기효능감, 금연 신념, 금연 의사 결정 및 금연변화단계에 미치는 효과를 파악하는 것이다. 본 연구의 대상자는 B시에 소재한 중학생 57명(실험군)과 50명(대조군), 총 107명이었다. 본 연구의 독립변인인 END 프로그램은 동기강화상담 프로그램은 한국건강증진개발원(2014)이 중·고등학교 흡연 학생을 대상으로 2012년 개발하여 개정한「청소년 금연프로그램(END)」을 활용하였으며, 프로그램의 효과는 금연 지식, 니코틴 의존도, 금연 자기효능감, 금연 신념, 금연 의사 결정 및 금연변화단계 척도를 사용하여 평가하였다. 수집된 자료는 SPSS 29.0을 이용하여 백분율, t-test 등을 통해 분석하였으며, 연구 결과는 다음과 같다. 금연 상담 프로그램에 참여한 실험군은 참여하지 않은 대조군보다 금연 지식(t=5.137, p=.009), 니코틴 의존도(t=2.707, p=.029), 금연 자기효능감(t=1.894, p=.030), 금연 신념(t=2.027, p=.044), 금연 의사 결정(t=2.079, p=.013) 및 금연변화단계 점수(t=3.873, p=.000)가 통계적으로 유의한 차이를 보였다. 따라서 END를 활용한 금연 상담은 청소년의 금연을 위한 유용한 중재로 사용될 수 있으리라 사료된다.

주요용어 : 청소년, 흡연, 금연 상담, END.

1. 서론

1.1. 연구의 필요성

흡연 시작 연령이 점점 더 낮아지고 있으며, 저연령에 흡연을 시작할수록 평생 흡연자가 될 위험이 크다. 흡연은 모든 불법 약물의 관문으로써 흡연에서 시작하여 음주, 약물, 마약까지 이르는 중독자가 되기 쉽다. 흡연을 처음 경험한 beginner인 중학생은 금연 동기가 높고, 금연 성공률도 높다. 한정된 금연 상담이라는 자원을 중학생에게 투자하여 금연으로 이끄는 것이 가장 이상적이다. 또한 한국건강증진개발원에서 개발한 END를 활용한 금연 상담의 효과를 확인한 연구가 거의 없으므로 국가정책으로 개발한 프로그램을 적극적으로 활용하여 그 효과를 검증하는 것이 필요하다.

이에 본 연구자는 END를 활용한 금연 상담의 효과를 확인하고, 효과적인 금연을 위한 방법을 제시하고자 한다.

[1]47119 부산광역시 부산진구 화지로 12 부산광역시교육청 교육정책과 학교보건 장학사.
[2]47011 부산광역시 사상구 주례로 47 동서대학교 바이오헬스융합대학 간호학과 부교수.
E-mail: mhpark@gdsu.dongseo.ac.kr

1.2. 연구의 목적

본 연구의 목적은 흡연 학생을 대상으로 한 금연 상담의 효과를 확인하기 위함이며, 구체적인 목적은 다음과 같다.

가. 대상자의 일반적 특성을 파악한다.

나. 대상자의 흡연 특성을 파악한다.

다. 대상자의 금연 상담의 효과를 파악한다.

2. 연구 방법

2.1. 연구 설계

본 연구는 END를 활용한 금연 상담이 중학생의 금연 지식, 니코틴 의존도, 금연 자기효능감, 금연 신념, 금연 의사 결정 및 금연변화단계에 미치는 효과를 확인하기 위한 비동등성 대조군 사전·사후 설계를 이용한 유사 실험 연구이다.

2.2. 연구 대상 및 자료 수집

연구대상자는 B시에 소재한 중학교에서 학교흡연예방 업무담당 교사를 통해 부산광역시교육청의 금연 상담을 의뢰한 학생 중에서 중학생 57명을 실험군으로 하였다. 또한 흡연은 하지만, 금연 상담을 시작하지 않은 학생 중에서 중학생 50명을 대조군으로 하였다. 실험군 대상자 선정 기준은 다음과 같다.

- 금연변화단계의 계획 전 단계에 속한 학생
- 연구 목적을 이해하고 연구 참여에 동의한 학생
- 학부모와 학교흡연예방 업무담당 교사가 연구 참여에 동의한 학생

본 연구의 대상자 수는 Cohen의 표본추출 공식에 따른 표본 수 계산 프로그램인 G*Power 3.1 프로그램으로 산출했으며, 유의수준 α=.05, 중간 크기의 효과 크기 d=0.3, 검정력 1-β=.80을 기준으로 한 집단에 필요한 대상자 수는 102명이었고, 최종 연구대상자는 총 107명으로 최소 표본 수 이상을 충족하였다.

2.3. 연구 도구

2.3.1. END 금연 상담

END 금연 상담은 총 4회기 8개의 세부 주제로 구성되었다. 1회 상담은 100분으로 구성되었으다. 1회기에 2개의 주제를 활용하여 상담하였으며, 세부 주제는 '나는 어떤 사람인가?, 인생에 거름주기, 나의 자아존중감, 나의 가치관, 나의 흡연, 내 몸의 증상들, 나의 자기주장, 나의 금연'이다.

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 13-16

청소년의 행복감이 학업성취만족도에 미치는 영향 : 학업열의와 그릿의 이중매개효과

박공주[1], 고진희[2]

요 약

본 연구는 청소년의 행복감이 학업성취만족도에 미치는 영향을 분석하고 학업열의와 그릿의 매개효과를 검증하고자 한다. 이를 위해 한국청소년정책연구원에서 시행한 제5차년도(2018)의 한국 아동·청소년 패널 조사(KCYPS)자료를 활용하였고, 중학교 1학년 총 2,265명의 자료를 분석하였다. 자료분석은 SPSS 26.0으로 기술통계분석과 상관관계를 실시했으며, Process Macro 4.0을 활용하여 매개효과 분석을 실시하였다. 청소년의 행복감이 학업성취만족도의 관계에서 학업열의와 그릿의 매개효과를 검증한 결과 행복감이 학업성취만족도에 대해 유의미한 영향을 미쳤고, 행복감이 학업성취만족도에 미치는 영향에서 학업열의를 매개하여 검증한 결과 유의미한 정적 매개효과를 미치는 것으로 나타났고 그릿의 관계에 있어서도 유의미한 매개효과를 미치는 것으로 나타났다. 이를 바탕으로 행복감이 학업성취만족도에 미치는 영향을 파악하기 위한 다양한 교육 및 상담프로그램 개발에 이바지해야 할 것이다.

주요용어 : 청소년, 행복감, 학업성취만족도, 학업열의, 그릿.

1. 서론

1.1. 연구의 필요성

한국인의 행복 수준은 매우 낮은 것으로 알려져 있다. 한국인의 행복 수준을 OECD 31개 회원국과 1990년에서 2017년까지의 기간으로 비교 분석한 결과 하위권에 속하는 것으로 보고되었다(박명호, 박찬열, 2019). 특히 한국방정환연구재단이 조사한 한국 어린이·청소년 행복지수에 따르면, 우리나라 청소년들은 타 국가의 청소년들보다 물질적 행복, 보건과 안전, 가족과 친구 관계, 건강 관련 행위 등과 같은 객관적 지표는 중상위권으로 높으면서도 개인의 주관적 행복지수는 최하위권으로 나타났다(박종일, 박찬웅, 서효정, 염유식, 2010). 이러한 현상은 입시 위주의 교육문화 속에서 학업성취가 행복의 중요한 조건으로 여겨지는 사회적 현상에 의한 것으로 여겨진다. 청소년의 행복과 학업성취에 관한 선행연구 결과, 청소년의 행복은 학업 성적, 학업 적응, 목표 달성도와 같은 학업성취도를 예측하는 주요 변인으로 확인되었으며(구재선, 서은국, 2012). 그릿과 학업열의는 청소년의 학업성취도에 영향을 미치는 것으로 나타났다(이미라, 전향신, 2020; 박선숙, 2021). 본 연구는 청소년의 행복과 학업성취도 간의 관계에서 학업열의와 그릿의 매개효과를 확인하여, 청소년 행복증진 프로그램에 기초자료를 제공하기 위해 시도되었다.

[1,2]김해대학교

2. 연구 방법

2.1. 연구모형과 가설

본 연구의 목적은 한국 아동·청소년 패널 조사 2018 제5차년도 제 8기(2022년)를 활용하여 청소년의 행복감이 학업열의와 그릿을 매개하여 학업성취만족도에 미치는 영향을 확인하고자 한다. 본 연구의 구체적인 목적은 다음과 같다.

첫째, 청소년의 행복감은 학업성취만족도에 영향을 미치는가?

둘째, 청소년의 행복감은 학업열의에 영향을 미치는가?

셋째, 청소년의 행복감은 그릿에 영향을 미치는가?

넷째, 청소년의 행복감은 학업열의와 그릿을 매개하여 학업성취만족도에 영향을 미치는가?

2.2. 분석자료 및 연구대상

본 연구는 한국청소년정책연구원에서 시행한 제5차년도(2018)의 한국 아동·청소년 패널 조사(KCYPS)자료를 활용하였다. 원시자료의 대상자는 2018년에 표집된 초등학교 4학년과 중학교 1학년의 2개의 코호트 총 5,197명 중 중학교 1학년 2,590명을 대상으로, 결측치를 제외하면 최종 2,265명을 연구대상자로 선정하였다.

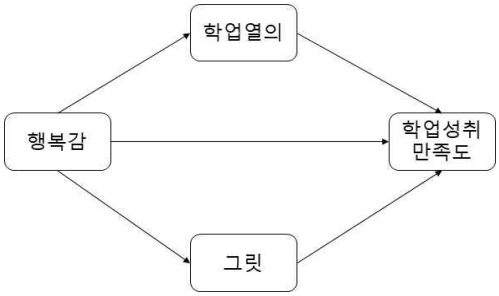


Figure 1. Research model

2.3. 측정도구

1) 독립변수 : 행복감

청소년의 행복감을 측정하기 위하여 육아정책연구소행복지수 관련 문항인 '전반적으로 나는 행복한 사람인가?' 라는 문항을 활용하여 측정하였다. 4점 Likert 척도 '아주 불행한 사람이다', '불행한 사람이다', '행복한 사람이다', '아주 행복한 사람이다'로 점수가 높을수록 행복감 점수가 높음을 의미한다.

2) 종속변수 : 학업성취만족도

청소년의 학업성취만족도를 측정하기 위하여 김지경 외(2010) 한국아동·청소년패널조사 2010의 관련 기존문항을 수정·보완한 문항을 사용하였다. 6점 Likert 척도로 '매우 못함', '못함', '보통', '잘함', '매우 잘함', '잘 모르겠음'으로 점수가 높을수록 학업성위만족도 점수가 높음을 의미한다.

3) 매개변수 : 그릿

청소년의 그릿를 측정하기 위하여 김희명, 황매향(2015)의 Grit척도 8문항 사용하였다. ‘전혀 그렇지 않다’, ‘그렇지 않은 편이다’, ‘그런 편이다’, ‘매우 그렇다’로 점수가 높을수록 그릿점수가 높음을 의미한다.

4) 매개변수 : 학업열의

청소년의 학업열의를 측정하기 위하여 이자영, 이상민(2012)의 한국형학업열의척도 16문항 사용하였다. ‘전혀 그렇지 않다’, ‘그렇지 않은 편이다’, ‘그런 편이다’, ‘매우 그렇다’로 점수가 높을수록 학업열의 점수가 높음을 의미한다.

2.4. 자료 분석방법

연구대상자의 인구사회학적 특성을 확인하기 위하여 SPSS Statistics 26.0으로 기술통계분석과 상관관계를 분석하였으며, SPSS Macro Process 4.0 프로그램을 활용하여 이중매개효과 분석을 실시하였다. SPSS Macro Process는 연구모형의 측정오차를 반영하여 검증이 가능하며 개별 매개효과의 각각의 통계적 검증이 가능하면서도 하나의 모형에서 동시추정을 통해 2개 이상의 매개변수를 검증하므로 이중매개효과 분석에 적합하다(허원무, 2013). 매개효과의 통계적 유의성은 부트스트래핑 방식(5,000회)을 설정하여 검증하였다.

3. 연구결과

Table 1. 대상자의 일반적 특성 (N:2265)

Variable	Categories	Frequency	%
성별	남	1217	53.7%
	여	1048	46.3%
도시규모	대도시	1002	44.5%
	중소도시	901	40.0%
	읍면지역	351	15.6%
지난 학기 전 과목 성적 수준	매우못함	62	2.7%
	못함	351	15.5%
	보통	1072	47.3%
	잘함	696	30.7%
	매우잘함	68	3.0%
	잘 모르겠음	16	0.7%

Table 2. 주요변수의 기술통계 및 변수 간 상관관계

Categories	M	SD	Range	학업성취 만족도	행복감	학업열의	그릿
학업성취 만족도	3.06	.99	1~6	1			
행복감	2.38	.34	1~4	.148***	1		
학업열의	2.26	.46	1~4	.206***	.332***	1	
그릿	2.25	.24	1~4	.165***	.076***	.246***	1

Table 3. 대상자의 행복감이 학업성취만족도의 관계에서 학업열의와 그릿의 매개효과

	Variables	b	SE	t	p	95% CI	
						LLCI	ULCI
Indirect effect	(Constant)						
	행복감→학업열의	0.33	0.11	16.72	<.001	1.59	2.02
	R=.33, F=279.52, P=.001						
	(Constant)						
	행복감→그릿	0.08	0.03	3.61	<.001	0.05	0.17
	R=.25, F=20.11, P=.001						
	(Constant)						
	행복감→학업성취만족도	0.09	0.00	4.26	<.001	0.4	0.10
	학업열의 →학업성취만족도	0.14	0.01	6.85	<.001	0.01	0.03
	그릿 →학업성취만족도	0.13	0.01	5.85	<.001	0.04	0.08
	R=.25, F=50.11, P=.001						
Direct effect	(Constant)						
	행복감 →학업성취만족도	0.15	0.02	7.13	<.001	0.08	0.14
	R=.15, F=50.81, P=.001						

b:regression coefficient

Table 4. Mediating effect verification

변수		Effect	BootSE	95% CI	
				LLCI	ULCI
Total effect	Direct effect + Indirect effect	.11	.02	.08	.14
Direct effect	행복감→학업성취만족도	.07	.02	.10	.09
Indirect effect	행복감→학업열의→학업성취만족도	.04	.01	.02	.05
	행복감→그릿→학업성취만족도	.01	.01	.01	.01

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 17-17

골관절염 여성 노인의 건강관련 삶의 질 영향요인: 국민건강영양조사 자료활용(2016~2020년)

전은미[1], 강세원[2]

요 약

본 연구는 골관절염 여성 노인의 건강관련 삶의 질에 영향을 미치는 요인을 확인하고자 2016~2020년 국민건강영양조사 자료의 골관절염 진단을 받은 여성 노인 1855명을 대상으로 조사하였다. 수집된 자료는 SPSS/WIN 25 프로그램을 이용하여 복합표본에서의 분석방법을 적용하였다. 실수와 백분율, 평균, 표준편차, 집단 간 비교는 교차분석과 일반선형분석을 이용한 t-검정, 분산분석(ANOVA)을 이용하였으며, 사후검정은 Bonferroni 방법으로 분석하였다. 제 변수들과 골관절염 진단 노인여성의 삶의 질 간의 상관관계는 상관계수를, 골관절염 진단 노인여성의 삶의 질에 영향을 미치는 요인은 일반선형분석을 이용한 다중회귀분석을 이용하였다. 연구결과 인구학적 특성에 따른 삶의 질은 연령, 거주지, 결혼상태, 교육수준, 경제활동, 우울, 주관적 건강상태, 활동제한, 스트레스 인지, 흡연, 체질량지수, 산책, 근력운동, 유산소운동에서 통계적으로 유의한 차이가 나타났고, 건강특성에 따른 삶의 질은 연령, 거주지, 결혼상태, 교육수준, 경제활동, 우울, 주관적 건강상태, 활동제한, 스트레스 인지, 흡연, 산책, 근력운동, 유산소운동에서 통계적으로 유의한 차이가 있었다. 대상자의 삶의 질 측정결과 평균0.82±0.18으로 나타났고, 삶의 질은 연령, 주관적 건강상태, 스트레스 인지 정도와 통계적으로 유의한 상관관계가 있는 것으로 나타났다. 건강관련 삶의 질에 영향을 주는 요인으로 유배우자, 가구소득이 높은 경우, 경제활동 유, 연령과 스트레스가 적을수록, 인지된 건강상태가 높을수록 삶의 질이 증가됨을 확인하였고, 모형의 설명력은 42.3%였다. 이상의 결과를 바탕으로 다각적이고 개별적인 간호중재프로그램 개발이 필요하다.

주요용어 : 골관절염, 여성 노인, 삶의 질, 인지된 스트레스, 주관적 건강상태.

[1]35345 대전광역시 서구 배재로 155-40(도마동), 배재대학교 간호학과 교수. E-mail: charminggold@pcu.ac.kr
[2]47011 부산광역시 사상구 주례로 47, 동서대학교 간호학과 부교수. E-mail: nursmile@gdsu.dongseo.ac.kr

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 19-22

온라인 수업을 경험한 간호대학생의 학습몰입, 문제해결능력 및 학업성취도와의 관계

김민영[1]

요 약

본 연구는 온라인 수업을 경험한 간호대학생의 학습몰입, 문제해결능력 및 학업성취도와의 관계를 알아보기 위해 시행되었다. 자료수집은 B시 3개의 대학교 간호학과 대학생 3학년 156명을 대상으로 2023년 4월부터 5월까지 시행되었으며, 수집된 자료는 SPSS program 26.0을 이용하여 기술통계, t-test, ANOVA, pearson's correlation coefficients,으로 분석하였다. 그 결과 학습몰입은 문제해결능력, 학업성취도와 유의한 상관관계를 나타내었고, 문제해결능력은 학업성취도와 유의한 상관관계를 나타내었다. 따라서 학습몰입을 통한 문제해결능력과 학업성취도를 향상시킬 수 있는 방안과 이를 위한 실제적 교육 방안을 모색해야할 것이다.

주요용어 : 온라인 수업, 간호대학생, 학습몰입, 문제해결능력, 학업성취도

1. 서론

1.1. 연구의 필요성

코로나19로 인해 등교에 의한 집합수업 대신 원격수업, 과제물 활용 수업 등 재택수업을 권장하는 「2020년도 1학기 대학 학사 운영 권고안」이 발표되면서(Yoo, 2021) 교육 환경에 많은 변화를 가져왔고, 그 중 하나가 학습자-교수자 간의 수업 방식이 대면에서 비대면으로 바뀐 것이다(Park, 2023). 이로서 비대면 온라인 온라인 수입이 불가피하게 되었고 학습자에게는 능동적이고 적극적인 학습 활동이 요구되고 있지만 비대면 전환은 학습자들과 교수자들에게 많은 어려움을 발생시켰다. 학습자는 양질의 수업을 수강하지 못하는 어려움과 실시간 온라인 수업에 관한 부정적 인식을 포함한 학습자들의 수업 효과성 저하 문제가 나타나게 되었고, 교수자는 학습자를 통제하기 어려운 상황이 발생하므로서 학습자들의 학습 몰입이 떨어지게 되었다(Lee, 2021). 학습 몰입은 개인의 심리적 상태인 내적, 사회적 환경 요인인 외적 요인이 복합적으로 상호작용하기 때문에 코로나19로 인한 불안감과 새로운 학습 환경은 학습자의 학습 몰입에 직접적이고 간접적인 영향을 미치고 있다(Joe, Ju, 2020) 이러한 학습몰입은 그 자체가 동기를 제공하여 학습 시간과 태도에 영향을 주고, 그 결과 학업성취에도 긍정적인 영향을 준다는 보고가 있다(Lee, 2010; Chung, et al., 2010).

간호 대학생은 대상자의 간호 문제를 신속하고 정확하게 중재할 수 있는 문제해결 능력을 길러야 하며(Chaung, 2011), 이를 위해 변화된 온라인 및 교내임상 대체 실습에서 교과목의 실습 시간

[1]47011 부산광역시 사상구 주례로 45, 경남정보대학교 간호학과 조교수. E-mail: 0502young@hanmail.net

을 충족시키고 임상실습과 유사한 사례를 통해 문제해결 능력을 함양할 수 있어야 한다(Ha, Lee, 2021). 이러한 간호학과 학생들의 문제해결능력은 학업성취도와 상관성이 있다는 결과를 보였다(Kim, Ko, Kim, 2016; Gil, 2021) . 학업성취도는 학습몰입과 진로 행동을 통해 실현할 수 있는 수 있고, 그 과정에 대한 객관적인 평가의 기준이 된다(Park, 2019). 간호대학생의 경우, 다양한 이론교육과 임상실습 교육이 병행되어야 하고 학점 이수 및 국가고시 합격 등이 졸업 이후의 진로에도 영향을 미치므로(Han, Kim, 2018) 학업성취도는 중요하게 고려되어야 할 부분이다(Oh, Kang, 2013).

본 연구는 간호대학생의 교내실습 시행 및 임상실습과 이론 수업이 병행되는 상황에서 교육과정 특성상 간호학 교과목에 대한 온라인 강의 한계성을 인식하고(Park, Shin, 2021), 온라인 수업을 경험한 간호대학생들의 학습몰입, 문제해결능력 및 학업성취도와의 관계에 대해 규명함으로써 온라인수업에 대한 간호대학생의 학업 성취도 증진을 위한 기초자료를 제공하고자 한다.

2. 연구 방법

2.1. 연구설계

본 연구는 온라인 수업을 경험한 간호대학생의 학습몰입, 문제해결능력 및 학업성취도의 관계를 조사하기 위한 서술적 조사연구이다.

2.2. 연구 대상

코로나19로 인해 온라인 수업이 시작된 2020년부터 현재까지 B광역시의 3개 대학교에 재학 중인 간호학과 학생 3학년 156명을 대상으로 하였다.

2.3. 연구 도구

1) 학습몰입

박소이(2021)가 Csikszentmihalyi(1990)가 개발한 도구를 수정 보완한 도구를 사용하였으며 총 9문항의 Likert 5점 척도로 박소이(2021)의 연구에서 Cronbach's α=.925였다.

2) 문제해결능력

한국교육개발원(이석재 등, 2003)에서 개발한 생애 능력 측정 도구 중 대학생/성인용 문제해결능력 측정 도구를 사용하였다. 총 45문항, 5점 Likert 척도로 도구의 신뢰도는 개발 당시 Cronbach's α=.94였다[57].

3) 학업성취도

최부기와 전주성(2011)의 측정도구를 참고하여 김희진(2022)이 재구성한 것으로 총 5문항으로 김희진의(2022)의 연구에서 Cronbach's α=.93이였다[58].

2.4. 자료수집 및 절차

자료수집 기간은 2023년 4월 17일부터 5월 10일까지 시행하였다. 연구 대상자에게 연구의 목적을 설명하고 비밀 보장, 익명성, 참여 거부, 참여 중단 가능 들을 설명하고 연구 대상자에게 동의받은 후 자가 보고식 설문지로 조사하였다. 또한 연구 참여의 자발성과 연구 참여에 대한 거절로 인해 받게 되는 불이익은 전혀 없음을 설명서에 명시하였다. 수집한 설문지 및 관련 자료는 연구가 종료되면 폐기할 것이다.

2.5. 자료 분석 방법

본 연구에서 수집된 자료는 SPSS 26.0 프로그램을 이용하여 분석하였으며 구체적인 방법은 다음과 같다.

1) 간호대학생의 일반적 특성은 평균, 표준편차, 백분율 등의 기술 통계량으로 분석하였다.
2) 학습몰입, 문제해결능력, 학업성취도에 대해 평균, 백분율을 분석하였다.
3) 학습몰입, 문제해결능력, 학업성취도 간의 관계를 T-test, ANOVA 등의 분석 방법을 사용하여 그룹 간 차이를 검정하고, Scheffe test를 사용하여 사후분석을 수행하였다.
4) 학습몰입, 문제해결능력, 학업성취도 간의 상관관계를 Pearson's Correlation coefficients로 사용하여 분석하였다.

3. 연구결과

3.1. 대상자의 일반적 특성

대상자의 성별은 여자 136명(87.2%)이며, 연령은 20~24세가 98명(62.8%)였다. 대학생활만족도는 '만족'이 72명(46.2%)으로 가장 많았고, 온라인수강방법은 '편의에 따라'가 86명(55.1%), '시간표에 맞춰서 수강'가 38명(24.4%), '한번에 수강'이 32명(20.5%)이었다. 온라인수강 횟수는 '모든 내용 1회'가 92명(59.0%)으로 가장 많았고 '필요한 부분만'이 40명(25.6%), '반복 학습'이 22명(14.1%)이였다. 온라인수업에 문제점은 '수업 집중도저하'가 73명(46.8%)으로 가장 많았으며 학습자와 '교수의 상호작용 어려움'이 31명(19.9%), '스스로 학습 어려움'이 30명(19.2%), '온라인 접속장애'가 15명(9.6%)로 나타났다.

3.2. 대상자의 학습몰입, 문제해결능력 및 학업성취도

간호대학생의 학습몰입은 5점 만점에 3.47점이었고, 문제해결능력는 5점 만점에 3.60점이었으며, 학업성취도는 4.6점 만점에 3.58점이었다.

3.3 대상자의 일반적 특성에 따른 학습몰입, 문제해결능력 및 학업성취도

대상자의 일반적 특성에 따른 학습몰입은 대학 생활 만족도 (F=19.233, p<. 001), 온라인수업 시 문제점(F=7.658, p<. 001)에서 통계적으로 유의한 차이가 나타났다. 문제해결 능력은 대학 생활 만족도 (F=12.450, p<. 001), 온라인수업 시 문제점(F=3.070, p=0.18)에서 유의한 차이가 나타났다. 학업

성취도는 대학 생활 만족도 (F=20.916, p<. 001), 온라인수업 시 문제점(F=6.897, p<. 001)에서 유의한 차이가 나타났다.

3.4. 대상자의 학습몰입, 문제해결능력과 학업성취도의 상관관계

학습몰입은 문제해결능력(r.612,p<.001), 학업성취도(r.722,p<.001)와 유의한 상관관계를 나타내었다. 또한 문제해결능력은 학업성취도(r.638,p<.001)와 유의한 상관관계를 나타내었다.

References

Chaung, S. K. (2011). Critical Thinking Disposition, Problem Solving Ability, and Clinical Competence in Nursing Students, *The Korean journal of fundamentals of nursing*, 18(1), 71-78.

Chung, A. K., Maeng, M. J., Yi, S. H., Kim, N. Y. (2010). The Effects of Meta-cognition, Problem-Solving Ability, Learning Flow of the College Engineering Students on Academic Achievement, Industry electronics, *Journal of the Institute of Electronics Engineers of Korea*, 47(2), 73-81.

Gil, C. R. (2021). Relationship between self-directed learning ability, learning flow, academic self-efficacy, and academic achievement of nursing students, *Journal of digital convergence*, 19(12), 617-626.

Ha, Y. K., Lee, Y. H. (2021). In COVID-19, Factors Affecting the Problem-solving Ability of Nursing Students Participating in Alternative Clinical Practicum, *The Journal of Learner-Centered Curriculum and Instruction*, 21(2), 989-1006.

Han, S. J., Kim, H. W. (2018). The Relationship of the Subjective Happiness, Ego-resilience and Academic Achievement of Nursing Students : Focusing on the Giver, Taker, Matcher, *Journal of the Korea Convergence Society*, 9(4), 461-467.

Joe, S. S., Ju R. (2020). A Study of Factors Affecting University Students' Learning Flow in Overall Distance Learning Situation: The Moderating Effect of Coronavirus Anxiety, *Journal of Korean Association for Educational Information and Media*, 26(4), 909-934.

Kim, E. H., Ko, Y. J., Kim, S. N. (2016). Effects of a Capstone Nursing Research Course for Nursing Students, *The Journal of Learner-Centered Curriculum and Instruction*, 16(10), 473-492.

Lee, J. E. (2021). Learning Flow, Self-Directedness, Self-Regulated Learning Ability and Learning Achievement of Nursing Students who in Non-Face-To-Face Learning Environment, *Journal of the convergence on culture technology*, 7(4), 511-517.

Lee. S. J. (2010). The Path Analysis of Teacher- Student Relationships, Class Climate, and Learning Flow on Academic Achievement in Elementary Students, *The Journal of Elementary Educatio*n, 23(4), 207-227.

Park, H. J. (2019). The Relationship of Self-regulated Learning and Learning Flow in the Academic Achievement of Medical Student, *The Journal of Humanities and Social science*, 10(4), 917-929.

Park, H. Y. (2023). A Study on the Effect of Academic Self-efficacy on Learning Commitment and Intention to Continue Learning in Non-face-to-face Online Classes: The Mediating Effect of Self-regulated Learning, *The e-business studies*, 24(1), 137-152.

Park, M. M., Shin, J. H. (2021). The Effect of Online Substitution Class Caused by Coronavirus (COVID-19) on the self-directed learning, academic achievement, and online learning satisfaction of nursing students, *Journal of the health care and life science*, 9(1), 77-86.

Yoo E. H. (2021). RESPONSE TO COVID-19 IN REPUBLIC OF KOREA, Ministry of Education.

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 23-23

베이지안 모형과 기계 학습을 이용한 희박한 데이터로부터의 지역별 사망률 산출*

김익한[1], 배현아[2]

요 약

높은 해상도를 가진 자료를 이용하여 지역별 사망률을 산출하면 분모인 인구수와 분자인 사망자 수의 크기가 작은 셀이 다수 존재하여 기저 사망률을 반영하지 못할 가능성이 크다. 선행 연구에 따르면 희박한 자료를 이용하여 지역별 사망률을 산출할 때는 사망자 수가 0인 셀들의 사망자 수를 특정한 수로 대치하거나 모형을 이용하여 해당 셀의 사망자 수를 예측한 뒤 사망률을 산출할 수 있다. 본 연구에서는 우리나라 237개 시군구의 교육 수준에 따른 원인별 사망률 산출 시, 대치 방법과 베이지안 시공간 모형 및 기계 학습 방법을 통하여 예측한 원인별 사망자 수를 이용한 결과를 비교하였다. 분석에는 2005, 2010, 2015년 인구총조사 및 2003-2017년 사망원인통계 자료를 이용하였다. 우리는 연도, 시군구, 교육 수준(중학교 이하, 고등학교, 대학 이상), 10세 연령군(30-39세,...,80세 이상)별 인구수와 자살 사망자 수를 획득하였다. 이 자료를 이용하여 연도, 시군구, 교육 수준별 연령표준화 자살 사망률을 산출하였다. 이때, 자살 사망자 수가 0인 셀은 (1) 관측된 사망자 수를 그대로 이용하거나 (2) 연도-교육수준-도시화 정도에 따른 자살 사망률을 이용하여 대치 혹은 (3) 사망자 수를 0.1로 대치하거나 (4) 베이지안 시공간 모형과 (5) 기계 학습을 이용하여 예측한 자살 사망자 수로 대치한 후 연령표준화 자살 사망률을 산출하고 그 결과를 비교하였다. 결과 비교 시, 교육 수준에 따른 평균제곱근오차(root mean squared error, RMSE), 평균절대오차(mean absolute error, MAE) 및 평균오차(mean error, ME)를 이용하였다. 연구 결과, 사망률이나 사망자 수를 이용한 대치 방법이나 베이지안 시공간 모형 및 기계 학습 방법 모두 관측된 사망자 수를 그대로 이용한 방법보다 인구수가 적은 지역들의 자살률을 비교적 높게 산출하였다. 대치 방법에서는 높은 교육 수준 집단의 자살 사망률이 치환하지 않은 방법보다 높게 산출되는 경향을 보였는데 이는 높은 교육 수준 집단의 자살 사망률이 낮아 사망자 수가 0인 셀이 많기 때문으로 여겨진다. 베이지안 시공간 모형과 기계 학습 이용 시에는 교육 수준이 낮은 집단의 자살 사망률이 높게 산출되는 경향이 나타났다. 이 결과는 모형 이용 시 인구수가 적은 지역은 인접한 지역 등으로부터 정보를 빌려오기 때문으로 여겨진다. 추후 시뮬레이션 연구를 통하여 각 산출 방법이 참값에 얼마나 가까운 값을 산출하는지를 연구할 필요가 있다.

주요용어 : 기계 학습, 데이터 희박성, 베이지안 모형, 지역별 사망률

---

*이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (RS-2023-00212260).

[1](교신저자) 49267 부산광역시 서구 감천로 262, 고신대학교 의과대학 인문사회의학교실 조교수.
E-mail: ikhan.kim@kosin.ac.kr

[2]49267 부산광역시 서구 감천로 262, 고신대학교 의과대학 인문사회의학교실 연구원.
E-mail: qogusdk0115@naver.com

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 25-25

장래가구추계를 위한 가구원수 및 가구유형 구성비 전망 모형 연구*

권태연[1], 이창호[2], 전새봄[3]

요 약

한국의 급격한 사회변동은 저출산, 고령화와 같은 인구문제뿐 아니라 가구분화, 1인가구 증가 등 가구규모 및 유형과도 밀접하게 관련된다. 가구규모 및 유형에 대한 전망은 주택수급, 세금전망 등 다양한 부문과 관계되므로 정확한 예측 및 추계가 요구된다. 한국은 2015년 등록센서스 도입으로 가구추계 기초자료인 센서스가 과거 총조사기반의 5년 단위가 아니라 행정자료기반의 1년 단위로 생성되고 있어, 인구 및 가구구조 변동을 보다 빠르게 파악할 수 있게 되었다. 또한 장래가구추계에 정태적 모형인 가구주율법에 혼인상태 변동을 반영한 준-동태적(semi-dynamic) 방식을 도입한 2012년 이래 가구추계방법은 지속적으로 개선되어 왔다. 그러나 대부분 혼인상태를 기반으로 하는 혼인상태 전이율 및 가구주율에 대한 개선에 초점이 맞추어져 있었으며, 가구원수 및 가구유형의 변동에 대해서는 연구가 미흡하였다. 최근에는 급감하는 혼인에도 불구하고 1인가구는 급증하는 등 혼인과 무관한 가구 형성 및 분화가 많이 발생되고 있어, 한국의 가구변동을 파악 및 전망하기 위한 새로운 가구추계모형이 필요하다. 이에 본 연구에서는 가구원수 및 가구유형 측면에서 한국의 가구 구성비 특성 및 변동을 파악하고, 이러한 추세를 적절하게 반영할 수 있으면서 장기예측시 강건한 전망결과를 작성할 수 있는 구성비 전망모형을 개발하고자 하였다. 본 연구에서는 대표적인 가구추계모형인 2-시점 수정지수모형을 확장 응용한 N-시점 수정지수모형을 개발하고, 우리나라 가구원수 및 가구유형별 구성비에 적용하여 예측한 결과, 본 연구모형이 한국의 장래가구 추계를 위한 구성비 예측에 적절함을 알 수 있었다.

주요용어 : 등록센서스, 가구주율, 수정지수모형, 장기예측.

---

*이 논문은 대한민국 교육부와 한국연구재단의 부분 지원(NRF-2022R1F1A1065520, NRF-2021R1F1A1059513)을 받아 수행되었음.

[1]17035 경기도 용인시 처인구 외대로81, 한국외국어대학교, 국제금융학과. E-mail: tykwon@hufs.ac.kr
[2]04792 서울 성동구 광나루로 228, 데이터웨이, 데이터분석가. E-mail: kkaggung10@naver.com
[3](교신저자) 35349 대전광역시 도안북로 88 목원대학교 사회과학대학 마케팅빅데이터학과 부교수.
E-mail: alwaysns@mokwon.ac.kr

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 27-27

텍스트 데이터와 재정데이터를 이용한 사회정책분야 예산 분석

이충열[1], 황명진[2], 김정학[3], 이지나[4], 이동찬[5], 김기환[6]

요 약

본 연구는 AI, 빅데이터, 전문가 판단을 결합하여 정부의 세부 사업 예산을 분석한 것이다. 정부의 세부 사업을 기존 분류가 아닌 최근 발표된 '2023년 핵심 사회정책 추진계획'의 27개 사회정책 의제 분류를 사용했으며, 아울러 생애주기도 분류 기준으로 사용하였다. 세부 사업을 설명하는 텍스트 데이터의 의미를 파악하고 분류하기 위해 자연어 처리기술을 사용하였으며 2020~2023년 정부의 세부 사업과 예산을 27개 의제에 따라 성공적으로 분류하였다. 분류과정에서 'NKIS', '열린재정'의 공공데이터를 활용하였으며 자연어처리 기술로는 KeyBERT를 사용하였다. 분류결과 27개 의제에 따른 정부 세부 사업 건수 및 예산의 연도별 변화, 27개 의제별 세부 사업의 불균형 정도를 확인할 수 있었다. 아울러 생애주기별 분류 결과 세부 사업과 예산이 누구를 위해 사용되고 있는지도 확인할 수 있었다. 최종 결과작성에서 자연어처리 기술이 많은 부분을 해결해 주었지만 전문가의 지식과 판단이 중요한 역할을 하였다. 연구 결과에 따르면 효율적인 예산 집행, 행정기관 간 협력을 어떻게 해야 하는 지에 관한 판단 근거를 찾을 수 있다. 또한 27개 사회정책 이슈, 생애주기 별로 좀 더 깊이 있는 분야별 연구가 가능할 것으로 기대된다.

주요용어: 재정데이터, 텍스트 데이터, 예산 분석, 열린재정, KeyBERT

[1]고려대학교 경제통계학부 교수. E-mail: cllee@korea.ac.kr
[2]고려대학교 행정전문대학원 교수. E-mail: mojohwang@korea.ac.kr
[3]고려대학교 행정전문대학원 교수. E-mail: trustkjh@korea.ac.kr
[4]고려대학교 사회복지학과 교원. E-mail: jnalee@gmail.com
[5]고려대학교 경제통계대학원 박사과정. E-mail: tozis72@korea.ac.kr
[6](교신저자) 30019 세종시 세종로 2511, 고려대학교 빅데이터사이언스학부 교수. E-mail: korpen@korea.ac.kr

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 29-29

CNN 기반 위성 이미지를 활용한 북한 인구추정

변상영[1], 이충열[2], 김기환[3]

요 약

북한은 공식적인 절차를 통해 인구데이터를 공개하지 않지만, 통계청, 인터넷 검색을 통해 북한의 인구와 관련된 자료를 확인할 수 있다. 이는 1993년, 2008년에 유엔인구기금(UNFPA)의 지원을 받아 실시한 인구총조사를 기반으로 추정된 자료이다. 하지만, 1993년과 2008년의 인구총조사 역시 신뢰성이 떨어진다는 문제점이 있다. 이에 본 연구에서는 2023년 북한의 주간 위성 이미지를 이용하여 북한의 평양과 개성의 현재 인구를 격자 단위로 추정하였다. 연구 결과로 CNN 기반 격자 단위 인구추정 모델을 개발하였다. 이 모형은 우리나라를 대상으로 CNN 모델을 훈련하고, 북한에 적용하는 것을 목표로 한다. 모형은 CNN의 대표적인 알고리즘인 VGG16 모델을 기반으로 전이학습을 하였으며, 주간 위성 이미지에서 나타나는 남북한의 계절 차이를 조정하기 위해 U-net을 활용하여 조정된 이미지를 사용하였다. 또한, 이웃효과(neighboring effects)를 추가하여 모델의 성능을 개선하였다. 모델 적합 결과 우리나라의 4대 광역시의 인구는 실제 인구와 큰 차이 없이 추정되었으며, 북한의 평양과 개성의 인구는 2008년 센서스인구와 유사하게 추정되어 만족스러운 결과를 보여주었다.

주요용어 : 북한 센서스, 주간 위성 이미지, CNN, 격자 단위, U-net

[1]30019 세종시 세종로 2511, 고려대학교 경제통계대학원 박사과정. E-mail: wooaaaaa@korea.ac.kr
[2]30019 세종시 세종로 2511, 고려대학교 경제통계학부 교수. E-mail: cllee@korea.ac.kr
[3](교신저자) 30019 세종시 세종로 2511, 고려대학교 빅데이터사이언스 학부 교수. E-mail: korpen@korea.ac.kr

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 31-31

국제 곡물가격이 양돈용 배합사료 가격에 미치는 비대칭적 영향 분석 -국제 옥수수, 대두박 가격을 중심으로-*

이현선[1], 순병민[2]

요 약

우리나라는 곡물 주요 수입국임에 따라 곡물 수출국 및 국제 시장 상황에 지대한 영향을 받는다. 따라서 시장 안정화를 위해 글로벌 가치사슬(GVC)을 고려하는 것은 중요하다. 이에 따라 본 연구에서는 국제 곡물을 옥수수와 대두박으로 구분하고, 국제 곡물가격이 우리나라 양돈용 배합사료 가격에 미치는 영향을 비선형 ARDL 모형을 통해 분석하였다. 분석 결과, 국제 옥수수 가격은 양돈용 배합사료 가격에 장기와 단기 모두 비대칭적 영향을 보이지 않았으나, 국제 대두박 가격은 단기적으로 양돈용 배합사료 가격에 비대칭적 영향을 주는 것으로 나타났다. 곡물 간 영향 차이를 비교한 결과, 우리나라 양돈용 배합사료 가격에는 국제 옥수수 가격에 비해 국제 대두박 가격이 장기적으로 더욱 큰 영향을 주는 것으로 나타났는데, 이는 대두박이 옥수수보다 상대적으로 양돈 사료에 널리 쓰이기 때문이다. 분석 결과를 종합해보면, 옥수수와 대두박 중 국제 대두박 가격이 우리나라 양돈용 배합사료 가격에 미치는 영향이 단기에서 비대칭성을 보였으며, 국제 옥수수 가격에 비해 국제 대두박 가격이 양돈용 배합사료 가격에 미치는 장기적인 영향이 더욱 큰 것으로 나타났다. 즉, 가격 전이의 비대칭성과 곡물 간 영향 차이를 비교한 결과, 국제 대두박 가격이 양돈용 배합사료 가격에 더욱 큰 영향을 주기 때문에, 국제 대두박 가격 변동을 사전에 파악함으로써 양돈 농가 경영에 미치는 영향을 최소화할 수 있도록 해야할 것이다.

주요용어 : international grain prices, grain import prices, assorted feed price for hog, gloval value chain(GVC), nonlinear auto regression listributed(ARDL) model.

---

*본 논문은 2022년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 연구되었음 (NRF-2022M3J6A1084843).

[1]41438 대한민국 대전시 유성구 대학로99, 충남대학교 농업경제학과 석사과정. E-mail: lhseon_official@naver.com
[2](교신저자)41438 대한민국 대전시 유성구 대학로99, 충남대학교 농업경제학과 조교수. E-mail: soonbm@cnu.ac.kr

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 33-36

Exploring Brain Regions Related to Azheimer's Disease Using Functional Data Analysis Approach on Resting-State fMRI Data*

Ido Ji[1], *Eunjee Lee*[2]

Abstract

Alzheimer's Disease (AD) is a burdensome and incurable neurodegenerative disease, so brain study is needed to understand its pathogenesis. Since functional connectivity (FC) alteration has been shown in among AD patients groups, Resting-State fMRI (RS-fMRI) was used to explore brain regions. However, the previous study overlooked continuity on brain regions in FC analysis. From a study who introduced euclidean distance in FC analysis, we also employed a similar method and obtained FC curves. Since these curves can be considered to be functions, functional data analysis (FDA) was employed to treat this type of data. Especially, functional principal component analysis (FPCA) was used to extract function PC (fPC) scores from each brain region. These scores were used as covariates in logistic regression models with group penalties. Compared to the former approaches, our method showed higher classification rate in normal controls and AD patients (AUC = 0.9). Also, brain regions related to neurodegenerative studies were chosen in our best model. However, there were cases where only one side of the region was chosen. Hence further study is needed on these regions in our future work.

keywords : Alzheimer's Disease, Resting-State fMRI, Functional Connectivity, Euclidean Distance, Functional Data Analysis

1. Introduction

Alzheimer's Disease (AD) represents a significant global health issue for which there is currently no cure or effective treatment, emphasizing the need for intervention development.

---

*This material was based on work partially supported by Institute of Information & communications Technology Planning & Evaluation(IITP) grant (No. 2020-0-01441, Artificial Intelligence Convergence Research Center(Chungnam National University)) This work was also partially supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. NRF-2022M3J6A1084843, No. NRF-2021R1C1C1013936).

[1]MS student, Bio-AI Convergence, Chungnam National University, 99 Daehak-ro, Yuseong-gu, Daejeon, 34134, Korea. E-mail: ido.ji@o.cnu.ac.kr

[2](Corresponding Author) Assistant Professor, the Department of Information Statistics, Chungnam National University, 99 Daehak-ro, Yuseong-gu, Daejeon, 34134, Korea. E-mail: eunjee.cnu@gmail.com

Resting-State functional Magnetic Resonance Imaging (RS-fMRI) is popular in AD research. The method relies on the Blood Oxygen Level Dependent (BOLD) signal, which is a measure of the oxygenation changes in the blood. Functional Connectivity (FC), which is usually computed by Pearson's correlation coefficient of BOLD signals, has been used in uncovering intrinsic brain networks and the pairwise relationships of brain activities. Since alterations in the connectivity among AD and its preclinical groups were observed, it is important to conduct brain research on FC to uncover the underlying mechanisms and contributing factors of AD.

However, many previous studies have focused solely on investigating the FCs discretely, overlooking the fact that the brain is physically contiguous, and therefore, FC can also be considered as a form of continuous data utilizing anatomical distance information. Kelly et al.(2009) study used RS-fMRI and Euclidean distance to analyze the development of FC in the anterior cingulate cortex (ACC) across different age groups. Motivated by the study, Euclidean distance calculations were similarly employed to analyze FC in this work. From this approach, FC curves are obtained which can be considered to be functional data. Functional data analysis (FDA) was applied to this new data and covariates from functional PCA are used in logistic regression models with group penalty.

The purpose of this paper is to explore selected brain regions from the group shrinkage penalties that influence the classification of cognitively normal individuals and patients with AD. Moreover, the performance was compared with logistic models in the cases where traditional FC quantification methods were used as covariates, which was done in the former work (Jung et al. 2020). This comparison could allow us to assess the relative effectiveness of our approach against established methodologies in AD studies.

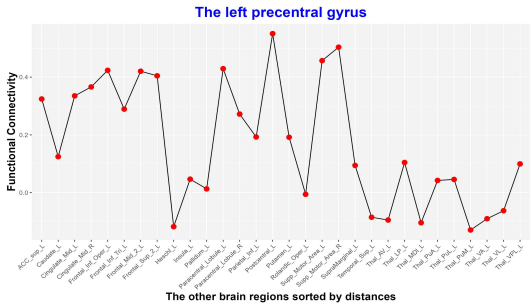


Figure 1. Sorted functional connectivity of the left precentral gyrus by proximity based on Euclidean distances with the other seed regions.

2. Materials and Methodologies

Publicly available RS-fMRI data was downloaded at ADNI (Alzheimer's Disease Neuroimaging Initiative), and data preprocessing was conducted using tools like MATLAB, DPABI, SPM12, and R. Then, applying the Automated anatomical labelling atlas 3 atlas, the

brain was divided into 164 regions to extract BOLD signals. In Figure 1, for example, brain regions are listed from left to right by proximity from the left precentral gyrus, representing the calculated FC values with the other regions. Such curves $f(d)$ were obtained for each patient across the 164 regions. After that, the following B-spline basis expansion to these curves in smoothing. To prevent overfitting, a roughness penalty was used as follows. Here, $f(d) \approx \sum_{i=1}^{n} c_i B_i(d)$ and λ represents a hyperparameter. Figure 2 shows smoothed curves of 30 patients in the same brain region. In this figure, x-axis is the distance, and y-axis is FC values.

$$SSPE_{\lambda} = \sum_{j} [y_j - f(d_j)]^2 + \lambda \int [D^2 f(d)]^2 dt = SSE + \lambda \times PEN(f)$$

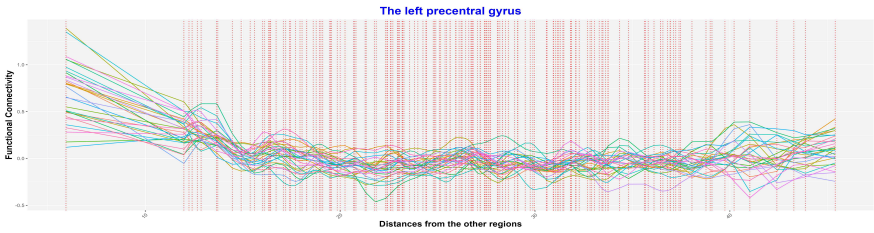


Figure 2. Smoothed curves B-spline basis expansion

For the mth region, the principal characteristics of the smoothed FC curves $f^{(m)}$ are extracted through Functional Principal Component Analysis (fPCA). fPCA aims to identify the principal modes of variation in FC cuves of mth region by resolving subsequent equation:

$$\int C^{(m)}(s,t)\phi_k^{(m)}(t)dt = \lambda_k^{(m)}\phi_k^{(m)}(s),$$

where $C^{(m)}(s,t)$ is the covariance function of $f^{(m)}(\bullet)$, $\phi_k^{(m)}(t)$ and $\lambda_k^{(m)}$ are the k-th eigenfunction and the corresponding eigenvalue. The k-th Functional PC(fPC) score is calculated as

$$\xi_{ik}^{(m)} = \int \phi_k^{(m)} f_i^{(m)}(t)dt,$$

which can be regarded as the low-dimensional data having most of the information of FC curves from mth region. In general, we will assume that there are R_m fPC that explain 90% of the variation, i.e., there are 164 sets of R_m fPCA scores. Hence, a general multivariate dataset can be obtained. Since those dataset has natural grouping information, to take advantage of this and to resolve high-dimensionality problem, logistic regression with group penalty was used in modeling: Group LASSO, Group Exponential LASSO, Group MCP, and Group SCAD.

Also, in order to compare the former work with rlogistic regression, other FC quantification methods were applied to RS-fMRI data: Low-Order Functional Network (LON),

Graph-Theory-Based Measures, and Common Component Analysis (CCA) (Want et al., 2011). In the comparion, the Area Under the Curve (AUC) was used.

3. Results and Discussion

Table 1 shows classification results from all models that fitted in this study. In the results, the model using fPC scores demonstrated superior performance with the Group LASSO penalty, achieving the highest AUC 0.90.

Table 1. Classification results by AUC

Group Penalty	fPC	Penalty	LON	Graph	CCA
Group LASSO	0.90	LASSO	0.65	0.65	0.81
Group Exponetial LASSO	0.72	Elastic-Net	0.61	0.72	0.77
Group MCP	0.81	MCP	0.64	0.74	0.67
Group SCAD	0.83	SCAD	0.59	0.72	0.77

Selected brain regions from the best model were as follows: The left subgenual anterior cingulate cortex, the right caudate nucleus, the middle cingulate and paracingulate gyri, the posterior cingulate gyrus, the left inferior frontal gyrus, the right medial orbital gyrus, the left lateral orbital gyrus, the right inferior parietal gyrus, the right supplementary motor area, the right and left middle temporal gyrus, and the left intralaminar. These brain regions are indicative of their potential involvement in the pathophysiology of AD. For example, the subgenual anterior cingulate cortex showed significant difference in FC analysis among preclinical AD groups. Also, some studies showed that AD patients had the larger caudate nucleus volume than it preclinical stage. However, among the selected brain regions, there were cases where only one region, either on the left or right side, was chosen. Furthermore, the hippocampus, which have known to be a critical brain region in AD research, was not selected. Therefore, further study is needed why these regions were not selected in our model.

References

Jung, J.-H., Ji, S.-J., Zhu, H., Ibrahim, J. G., Fan, Y., & Lee, E. (2020). Penalized logistic regression using functional connectivity as covariates with an application to mild cognitive impairment, *Communications for Statistical Applications and Methods*, 27, 603-624.

Kelly, A. C., Di Martino, A., Uddin, L. Q., Shehzad, Z., Gee, D. G., Reiss, P. T., Margulies, D. S., Castellanos, F. X., & Milham, M. P. (2009). Development of anterior cingulate functional connectivity from late childhood to early adulthood, *Cerebral Cortex*, 19, 640-657.

Wang, H., Banerjee, A., & Boley, D. (2011). Common component analysis for multiple covariance matrices, *in Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 956-964.

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 37-37

Comparison Research for Spatio-Temporal Data Analysis: Methods, Applications, and Implications

Sung Jae Kim[1], *Boseung Choi*[2]

Abstract

The first law of geography states that everything is related to everything else, but things that are closer are more closely related than those that are farther apart. This paper aims to explain various spatio-temporal statistical models for implementing Tobler's first law of geography. It compares and analyzes their strengths and weaknesses to pro-pose a research framework for applying these models to empirical data. Firstly, the pa-per describes the construction of spatial data, which forms the foundation for building spatio-temporal statistical models. It also outlines methods for addressing missing in-formation frequently encountered during the data construction process. To build models using spatio-temporal data, the paper compares Spatial Autoregressive models (SAR), Spatial Error Models (SEM), and Spatial Autoregressive Moving Average (SARMA) models. The model selection method is then applied to suggest the optimal model for the given spatial data. The methods introduced in this study are applied to two empirical datasets. In the first case, the correlation between the concentration of chlorophyll and neurodegenerative disease is investigated. The entire Seoul area was divided into 710 grid cells of 1km size to create spatial grid data. Kriging method was used to predict the concentration of chlorophyll in each region, and GIS information was utilized to adjust the number of senile patients with health conditions. The model comparison results revealed that SEM was the most suitable model for both cases.

Keywords : Spatial correlation, Spatial gridded data, Kriging, SAR, SEM, SARMA.

[1]Graduate Student, Department of Economics Statistics, Korea University, 2511 Sejongro, Jochiwoneup Sejong, South Korea. E-mail: sungkim0704@korea.ac.kr

[2](Corresponding Author)Professor, Department of the Division of Big Data Science, Korea University, 2511 Sejongro, Jochiwoneup Sejong, South Korea. E-mail: cbskust@korea.ac.kr

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 39-39

다양한 사건 현장에서 사후경과시간 추정을 위한 과거 현장온도 예측 모형 연구

정수진[1], 박지은[2], 박성환[3], 이재원[4]

요 약

사후경과시간(postmortem intervals, PMI)은 사람이 사망한 날로부터의 경과 시간을 의미한다. 사후경과시간 추정 방법으로는 시체 특성에 해당되는 체온하강(헨스게법), 시체 강직 발현 여부, 혈액 침하(시반), 위 내용물의 소화상태, 부패(TBS)의 진행 정도 등을 통해 추정할 수 있으며, 또한, 근육의 전기적 흥분도, 눈의 변화, 혈액, 심낭, 뇌척수액의 화학적 변화, 골수의 세포학 변화 등의 실험적 방법이 있다. 하지만 사망 후 오랜 시간 경과되거나 온도·습도 등의 주변 환경적 요인으로 인해 부패가 심해졌을 경우 등에는 일반적인 방법으로 사후경과시간을 추정하기에는 그 정확성이 떨어진다. 이를 해결하기 위해서는 사체 근처에서 발생되는 시식성 곤충의 종류나 성장 단계를 통하여 유효적산온도(accumulated degree day, ADD)를 계산하여 사후경과시간을 추정하는 것이 좀 더 신뢰성이 높다고 알려져 왔다. 유효적산온도는 사건 현장에서 발견된 곤충이 당시의 발생 단계까지 성장하는데 소요되는 시간과 현장 온도, 그리고 곤충이 성장하지 않는 시점의 온도인 기저온도(base temperature)를 통해 계산할 수 있다. 따라서 시체의 부패지수나 유효적산온도에 가장 중요한 요인으로는 현장 온도이기에 사건이 발생하고 난 이후부터의 현장온도를 최대한 정확하게 추정하는 것이 곧 정확한 사후경과시간을 추정할 수 있었다. 지금까지 많은 연구를 통해 기상청 온도를 이용하여 과거 사건현장 온도를 추정하는 연구가 진행되었다. 온도 추정을 위해서는 사건 발생 장소, 주변 기상청 위치, 기상환경(습도, 강우량, 강수량, 풍량, 풍속), 온도 측정 단위, 측정 기간, 예측 기간을 고려하여야 하며 서포트 벡터 머신(Support Vector Machine, SVM) 모델을 통해 예측하였다. 이는 사건 발생 현장의 온도가 주변 기상청 온도와 상관성이 높을 경우, 즉 야외처럼 폐쇄되지 않은 공간일 때 예측 정확도가 높아진다. 하지만 사건 현장은 실외 뿐 아니라 건물 내부와 같은 실내에서도 많이 발생할 수 있으며, 이는 창문 개폐 유무, 냉난방 사용 유무, 공간 크기 등 밀폐도에 따라 주변 기상청 온도와 상관성이 낮아 기존의 연구 모형의 적용이 어려운 실정이다. 따라서 본 연구는 국내 다양한 사건 현장 분석을 통해 유형별 사건 현장을 구현한 후, 직접 실험을 통해 수집된 현장 온도 측정 자료를 수집하였다. 이는 주변 기상청 온도를 기준으로 외부와의 폐쇄 정도를 통해 실내 유형별 과거 현장 대기 온도를 예측할 수 있는 확장된 예측 모형을 마련하고자 한다.

[1]경희의료원 임상의학연구소
[2]고려대학교 통계학과
[3]고려대학교 의과대학 법의학과
[4]고려대학교 통계학과

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 41-41

Long-term stock price manipulation: Evidence from the Korean stock market

Jinyong Kim[1], *Yongsik Kim*[2]

Abstract

We propose four criteria designed to preemptively identify potential stock price manipulations by analyzing common characteristics among stocks suspected of manipulation. These criteria encompass the permanent return ratio, the annual change of permanent price ratio, price informativeness, and the floating share ratio. The permanent return and price ratios focus on price fluctuations primarily influenced by non-fundamental factors rather than the intrinsic value of firms. We utilize price informativeness and the floating share ratio to capture manipulators' strategies to preserve private profit opportunities created by their price impacts, concealed from market investors. Upon applying these criteria to recent suspected cases of price manipulation in Korea, our analysis indicates their effectiveness in early detection of suspected stocks by categorizing them into potentially manipulated groups.

[1]School of Economics, University of Seoul, 163 Seoulsiripdae-ro, Dongdaemun-gu, Seoul 02504, Korea. Email: jinyongkim0409@uos.ac.kr.
[2]School of Economics, University of Seoul, 163 Seoulsiripdae-ro, Dongdaemun-gu, Seoul 02504, Korea. Email: jinyongkim0409@uos.ac.kr.

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 43-46

부도거리(Default to Distance)를 활용한 올슨모형(Ohlson Model) 실증분석

송유인[1], 이용웅[2]

요 약

Ohlson(1995)의 초과이익모형은 기업의 가치를 장부가치 변수, 초과이익 변수 및 기타정보 변수로 설명한다. 본 논문은 Ohlson 모형에서 부도위험이 기업가치에 미치는 영향을 관찰하기 위해 2011년에서 2021년 사이의 코스닥 기업을 대상으로 부도위험을 측정하여 실증분석하였다. 부도위험은 Merton(1974)의 부도거리(default to distance) 모형을 사용하였으며, 부도위험이 높을수록 부도거리는 짧아지고, 부도위험이 낮을수록 부도거리가 길어진다. 전통적인 재무이론인 자본자산결정모형(CAPM)에 따르면 투자 위험도가 높을수록 수익률 또한 높아지기 때문에, 부도거리 변수가 클수록 기업가치는 낮게 측정될 것이다. 하지만, 부도거리와 주식수익률의 관계에 관한 많은 선행연구들에서 부도거리가 주식수익률에 양(+)의 유의성을 보여주는 결과와 음(-)의 유의성을 보여주는 현상 등 상반된 결과가 관찰되고 있다. 본 연구에서는 기업분류별 패널을 구성하여 고정효과모형으로 패널회귀분석하였고, 그 결과 부도거리와 기업가치 간의 관계가 음(-)의 관계를 보여주었다. 이는 전통적인 자본자산결정모형(CAPM)을 지지하는 결과이다. 또한, Ohlson 모형에서 실물옵션변수와 부도거리변수를 함께 사용하였을 때 모두 기업가치를 유의미하게 설명하는 것으로 나타나서 부도거리 변수가 기타정보로서 고려될 수 있음을 제시하였다.

주요용어 : 올슨모형, 실물옵션, 부도위험, 부도거리, 기업가치평가.

1. 서론

Ohlson(1995)의 초과이익모형에 따르면, 기업가치는 장부가치와 초과이익 및 기타정보로 설명된다. 객관적인 회계정보인 장부가치와 초과이익은 많은 선행연구에서 기업가치를 설명하는 유용한 변수로 확인되고 있으나, 기타정보는 아직 명확히 규정되어 있지 않아 기타정보를 찾기 위한 후속연구들이 꾸준히 수행되고 있다. 후속연구에서 기타정보로 사용된 대표적인 변수들에는 배당금, 장부가대비시장가비율, 부채비율, 연구개발비, 애널리스트의 예상순이익 등 주로 회계정보에 기반한 변수들이 많이 사용되었다. 하지만, 해당 변수들을 Ohlson 모형에 적합한 실증분석 결과 대부분 유의미하지 않은 변수로 확인되었거나 연구마다 다른 결과를 보여주었다.

본 연구에서는 Ohlson 모형의 기타정보로 회계정보가 아닌 기업별 부도위험을 고려하였다. 자본자산결정모형(capital asset pricing model; CAPM)에 따르면 투자위험이 높을수록 기대수익률이 높아진다. 따라서 부도위험이 높은 기업의 주식가치가 부도위험이 낮은 기업의 주식가치보다 높게 형

[1]02450 서울특별시 동대문구 이문로 107 한국외국어대학교 대학원, 국제경영학과, 박사과정.
E-mail: yooinsong@gmail.com
[2](교신저자) 17035 경기도 용인시 처인구 모현읍 외대로 81 한국외국어대학교, 국제금융학과, 교수.
E-mail: ywlee@hufs.ac.kr

성될 수 있다. 이를 검증하기 위해 Merton(1974)에서 제안된 부도거리(distance to default; DD)를 통해 부도위험을 측정하였고, Ohlson 모형에 적합하여 패널회귀분석 하였다. 부도거리가 크면 부도위험이 낮은 것을 의미하기 때문에 부도거리와 기업가치는 음(-)의 유의성을 보여줄 것이다.

또한, 부도거리 변수의 강건성 검증을 위해 Song, Lee(2023)에서 제시된 실물옵션가치(real option value; ROV)를 포함한 Ohlson 모형을 사용하였다. Song, Lee(2023)는 기타정보를 회계정보가 아닌 투자자의 기대감으로 정의하고 실물옵션모형(real option valuation)으로 측정하였는데, 이를 Ohlson 모형에 적합하여 실증분석한 결과 실물옵션가치(real option value; ROV)가 Ohlson 모형의 기업가치 설명력을 유의적으로 보완하는 것을 발견하였다.

2. 연구방법

Ohlson(1995)의 초과이익모형은 식(1)과 같이 정리된다.

$$V_t = BV_t + \alpha_1 RI_t + \alpha_2 v_t \tag{1}$$

여기서 V_t는 t시점의 기업가치, BV_t는 t시점의 장부가치, RI_t는 t시점의 초과이익, v_t는 t시점의 기타정보를 의미한다.

Merton(1974)은 부도위험을 부도거리로 측정하는 모형을 제시하였으며, 본 연구에서는 식(2)과 같이 정의하여 사용하였다.

$$DD = \frac{[\ln(V/F) + \mu - 0.5 \times \sigma v^2]}{\sigma v} \tag{2}$$

여기서 V는 자산가치로서, 자기자본의 시장가치+부채의 장부가치(단기부채+1/2×장기부채)이며, F는 부채의 장부가치, μ는 0.06+무위험수익률(국고채 5년 최종호가 수익률), σv는 개별 기업의 1년간 주가 변동성을 의미한다.

Song, Lee(2023)은 투자자의 기대감을 실물옵션가치(real option value; ROV)로 측정하는 모형을 제시하였으며, 식(3)과 같이 측정된다.

$$ROV = S \cdot N(d_1) - \frac{K}{e^{r_f \cdot T}} \cdot N(d_2) \tag{3}$$

여기서 $d_1 = \dfrac{\ln\left(\dfrac{S}{K}\right) + \left(r_f + \dfrac{\sigma^2}{2}\right)T}{\sigma\sqrt{T}}$, $d_2 = d_1 - \sigma\sqrt{T}$이고, S는 당해연도 연평균 주가, K는 주당자본금, r_f는 무위험수익률로서 국고채 5년 최종호가 수익률, σ는 자산가치의 변동성인데 연간일수익율의 표준편차에 $\sqrt{250}$을 곱하여 연간 변동성으로 사용하였고, T는 만기까지의 시간으로 5년으로 설정하였다. 그리고 $N(\cdot)$는 표준정규누적확률분포이다.

이에 따라, 본 연구에서는 Ohlson 모형에서 기타정보로 DD를 적용한 패널회귀모형을 식(4)와 같이 설정하였고, 강건성 검증을 위한 패널회귀모형을 식(5)와 같이 설정하였다.

$$\ln P_{it} = \beta_0 + \beta_1 \ln BPS_{it} + \beta_2 \ln RI_{it} + \beta_3 \ln DD_{it} + \epsilon_{it} \tag{4}$$

$$\ln P_{it} = \beta_0 + \beta_1 \ln BPS_{it} + \beta_2 \ln RIPS_{it} + \beta_3 DD_{it} + \beta_4 \ln ROV_{it} + \beta_5 \ln DPS_{it} + \beta_6 (M/B)_{it} + \beta_7 Leverage_{it} + \beta_8 Beta_{it} + \beta_9 \ln Index + \epsilon_{it} \quad (5)$$

3. 실증분석 결과

본 연구는 2011년에서 2021년 코스닥(KOSDAQ) 기업을 대상으로 실증분석하였다. 또한, 한국거래소의 코스닥시장 업종심사 기준에 따라 우량기업(Bluechip), 중견기업(Midsize), 벤처기업(Venture) 및 기술성장기업(Tech)으로 분류하였고, 각 기업군에서 초과이익의 부호에 따라 각각 초과이익이 양수인 그룹과 음수인 그룹으로 나누어 총 8개의 패널데이터를 구성하였다. 외국기업, 기업인수목적회사(SPAC), 관리종목 및 투자환기종목은 한국거래소에서 업종심사 기준에 따라 분류하지 않으므로 제외하였다. 또한, 일반기업과 재무제표 구성이 현저하게 다른 금융 관련 업종은 제외하였고, 자본잠식기업, 거래정지기업도 제외하여 최종적으로 9,791개의 기업, 연도 데이터를 표본으로 선정하였다.

Table 1. Estimates of fixed effects panel regression in eq.(4)

	RI+				RI-			
	Bluechip	Midsize	Venture	Tech	Bluechip	Midsize	Venture	Tech
BPS	0.5737***	0.7135***	0.7689***	1.0931**	0.6443***	0.5656***	0.6544***	0.5879***
RIPS	0.1133***	0.0634***	0.0817***	-0.0304	-0.0022	-0.0100	0.0027	0.0159
DD	-0.1139***	-0.2588***	-0.2754***	-0.1157	-0.1690**	-0.2664***	-0.2191***	0.0040
R2	0.3425	0.4406	0.3429	0.3822	0.1983	0.2932	0.2808	0.2196

Note 1: BPS and RIPS represent book value per share and residual income per share as in a basic panel regression in eq.(4).
Note 2: DD denotes distance to default considered as additional independent variable in panel regression in eq.(4).
Note 3: *, **, and *** indicate significance at 5%, 1%, and 0.1% level respectively.

Table 2. Estimates of fixed effects panel regression in eq.(5)

	RI+				RI-			
	Bluechip	Midsize	Venture	Tech	Bluechip	Midsize	Venture	Tech
BPS	0.5371***	0.4317***	0.5934***	0.9487***	0.5556***	0.3586***	0.5362***	0.6040***
RIPS	0.0322***	0.0214***	0.0219**	0.0135	0.0128	0.0197**	0.0264**	0.0218
DD	-0.3695***	-0.5612***	-0.3907***	-0.2790	-0.4046***	-0.6629***	-0.4585***	-0.1106
ROV	0.3645***	0.4973***	0.3236***	0.1653	0.3669***	0.5803***	0.3783***	0.2775***
DPS	0.0055	0.0018	0.0054	0.0636	-0.0014	0.0061	0.0134	-
M/B	0.2656***	0.1370***	0.1977***	0.0626***	0.2370***	0.0581***	0.1153***	0.0547***
Leverage	0.0002	-0.0017*	-0.0005	-0.0021	0.0001	-0.0026***	-0.0013	-0.0003
Beta	0.0821***	-0.0171	0.0214	0.0764	0.0491	-0.0179	-0.0429	0.0872
Index	0.1982***	0.2287***	0.2048***	0.0828	0.2252***	0.3180***	0.3530***	0.3762***
R^2	0.8580	0.8115	0.8221	0.9313	0.7919	0.7281	0.7639	0.7733

Note 1: BPS, RIPS, and DD represent book value per share, residual income per share, and distance to default, respectively as in a panel regression in eq.(4).
Note 2: ROV, DPS, M/B, Leverage, Beta, and Index denote real option value, dividend per share, ratio of market value to book value, ratio of total debt to total asset value, systematic risk as a coefficient of CAPM, and log return of KOSDAQ, respectively as in a additional panel regression in eq.(5).
Note 3: *, **, and *** indicate significance at 5%, 1%, and 0.1% level respectively.

Table 1은 Ohlson 모형의 기타정보로 *DD*를 추가하여 패널회귀분석한 결과이다. 모든 기업군에서 *BPS*는 기업가치와 양(+)의 유의성이 있는 것으로 나타났고, *RIPS*는 초과이익이 양수인 기업군에서만 유의미한 것으로 나타났다. *DD*는 Bluechip, Midsize, Venture 기업군에서 모두 음(-)의 방향으로 유의한 결과를 보여주었다. 이는 부도거리가 클수록, 즉 부도위험이 낮을수록 주식가격에 음(-)의 영향을 주었다는 결과이며, 전통적인 재무이론인 자본자산가격결정모형(CAPM)과 일치하는 결과이다.

Table 2는 실물옵션가치 변수 및 기존 연구에서 많이 사용된 설명변수들을 포함한 강건성 분석 결과이다. *DD*는 Bluechip, Midsize, Venture 기업군에서 여전히 유의미한 음의 유의성을 보여주었다. 또한, 기존 Ohlson 모형에 Song, Lee(2023)가 추가한 *ROV*변수 역시 여전히 모든 기업군에서 양의 방향으로 주가를 잘 설명하고 있다. *ROV*와 *DD*변수는 모두 투자자의 심리 내지 행태와 관련된 변수인데, 기존 Ohlson 모형에서 함께 사용할 수 있는 변수로서의 가능성을 보여주었다.

4. 결론

Ohlson 모형은 객관적인 정보인 회계정보에 기반하여 기업가치를 설명하기 때문에 많은 관심을 받아오고 있으나, 기타정보의 존재로 인해 한계점 또한 명확하다. 이에 기타정보를 찾기 위한 많은 연구들이 수행되고 있다. 본 연구에서는 기존 Ohlson 모형의 회계정보 변수가 설명하지 못하고 있는 주식가치 부분을 부도위험 수준이 설명할 수 있는지 실증분석하였다.

Merton(1974)의 부도거리 변수인 *DD*를 접목하여 패널회귀분석한 결과 *DD*와 주식가치의 관계는 유의미한 음(-)의 결과를 보여주었다. 이는 부도위험이 높을수록 주식가치가 높게 형성되는 결과로, 위험이 클수록 주가수익률이 높은 자본자산결정모형(CAPM)을 지지하는 결과이다. 또한 실물옵션가치 변수인 *ROV*를 포함한 실증분석에서도 부도거리는 유의미한 변수로 나타났다. 이는 Ohlson 모형에서 *DD*와 *ROV*가 유의미한 변수로 함께 채택될 수 있는 가능성을 제시한 결과이다.

References

Merton, Robert C. "On the pricing of corporate debt: The risk structure of interest rates." *The Journal of finance* 29.2 (1974): 449-470.

Ohlson, J. A. (1995). Earnings, book values, and dividends in equity valuation. *Contemporary Accounting Research*, 11(2), 661-687.

Song, Y., Lee, Y. W. (2023). A study on the Ohlson model using real options: focusing on Kosdaq companies, *Journal of the Korean Data Analysis Society*, 25(6), 2229-2244.

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 47-47

경제적 비용을 최소화하는 기업부도예측모형의 앙상블 기반 구축

전승유[1], 박찬[2], 양기성[3]

요 약

신용위험관리에서 부실을 정상으로 예측(Type I Error)하는 손실은 정상을 부실로 예측(Type II Error)하는 손실보다 훨씬 크다. 본 연구는 Stacking Ensemble 기법을 활용하여 경제적 성과를 극대화하는 목적으로 학습을 수행하는 기업부도예측모형 알고리즘을 제안한다. 핵심 아이디어는 Base Learner의 예측 결과 Type I Error의 가능성이 있는 Instance만 Meta Leaner에서 재학습을 수행하는 것이다. 대만과 폴란드의 기업부도 데이터를 이용한 실증분석 결과, 제안된 알고리즘은 일반적인 부도예측모형에 비해 예측의 경제적 성과를 강건하게 개선하는 것으로 확인되었다. 하지만 그 반대 급부로 발생하는 통계적 성과의 감소는 매우 미미하였다. 경제적 비용을 고려한 부도예측에 관한 기존 연구들이 주로 개인부도를 다루고 학습의 비용함수를 새롭게 정의하여 문제를 해결하는 반면, 본 연구는 개인부도에 비해 비용의 비대칭성이 더 심한 기업부도 문제를 다룰 뿐 아니라 비용함수를 새롭게 정의하지 않고 Ensemble 기법을 통해 문제를 해결함으로써 실무적 필요성이 크고 적용이 용이한 알고리즘을 제안하였다는 차별성과 기여를 가진다.

주요용어 : Corporate Default Prediction; Cost-sensitive Learning; FinTech; Machine Learning; Stacking Ensemble

[1](제1저자) 02841 서울시 동작구 상도로 369, 숭실대학교 금융기술융합학과 박사과정.
E-mail: 1002228004@soongsil.ac.kr
[2](공동저자) 02841 서울시 동작구 상도로 369, 숭실대학교 금융기술융합학과 박사과정.
E-mail: chanpark@soongsil.ac.kr
[3](교신저자) 02841 서울시 동작구 상도로 369, 숭실대학교 금융학부 교수. E-mail: ksyang@ssu.ac.kr

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 49-52

전공만족도에 따른 학업성취도와 회복탄력성의 관련성 연구 : 인지적 학업성취도와 메타인지전략 중심으로

정회연[1], 조미정[2], 신동혁[3]

요 약

본 연구는 대학생의 전공만족도에 따른 학습자들의 학업성취도(인지된 학업성취도, 메타인지전략)와 회복탄력성이 어떠한 영향을 미치는지 분석하고자 하였다. 이를 위하여 전국대학교에 재학 중인 만18세~만29세 대상의 학생들을 대상으로 설문조사를 실시하였다. 설문조사의 기간은 약 3개월 간 진행하였으며, 설문조사를 통하여 전공만족도, 학업성취도(인지된 학업성취도, 메타인지전략), 회복탄력성의 인과관계를 분석하고자 하였다.

주요용어 : 대학교, 전공만족도, 학업성취도, 메타인지전략, 회복탄력성.

1. 서론

대학에서는 최근 학과간의 벽을 허물어 학생들에게 자유로운 전공 수업을 들을 수 있도록 기회를 주려고 많은 노력을 하고 있으나, 아직 우리나라의 입시교육에 중·고등학생들이 전공에 대한 경험이 적다. 따라서 전공에 대한 이해와 다양한 경험을 중·고등학교 시기에 자신이 하고 싶은 것이 무엇인지 자신의 흥미와 적성 그리고 소질이 어느 분야에 있는지 안다면, 대학에 입학했을 때 전공을 선택할 때 전공에 대한 만족도가 높을 것으로 생각한다. 이에 전공만족도를 높이기 위한 방법으로 중.고등학교 과정에서 자신의 적성과 흥미를 탐색할 수 있는 진로교육 및 직업체험 기회 등 다양한 경험을 바탕으로 전공을 선택할 수 있는 교과과정 편성이 필요하다고 생각한다.

이에 본 연구에서는 전공만족도가 학업성취도와 회복탄력성에 얼마나 많은 영향을 미치는지 관련성을 연구하고자 한다.

[1](1저자 및 교신저자) 13120 경기도 성남시 수정구 성남대로 1342, 가천대학교 교육대학원 글로벌캠퍼스 교육리더십 및 교육공학전공 석사과정. hoeyeon0529@gachon.ac.kr
[2](2저자) 28173 충청북도 청주시 흥덕구 태성탑연로 250, 한국교원대학교 교육정책전문대학원 교육정책정공 박사과정. mjeong37@knue.ac.kr
[3](2저자) 16499 경기도 수원시 영통구 월드컵로 164 홍재관 502호, 아주대의료원 만성뇌혈관질환 바이오뱅크 연구원., dong10s@ajou.ac.kr

2. 연구절차

Table 1. 연구절차

연구 주제 선정
⇩
이론적 배경 및 선행연구 고찰
⇩
분석 대상자 선정
⇩
선행연구의 분석기준 선정
⇩
설문조사 데이터를 활용한 관계성 분석
⇩
연구 결과 해석 및 결론 도출

3. 연구목적

(1) **전공만족도, 학업성취도(인지된 학업성취도, 메타인지전략), 회복탄력성 간의 관계**

(2) **전공만족도, 학업성취도(인지된 학업성취도, 메타인지전략), 회복탄력성의 매개효과**

(3) **전공만족도가 학업성취도(인지된 학업성취도, 메타인지전략), 회복탄력성에 미치는 영향**

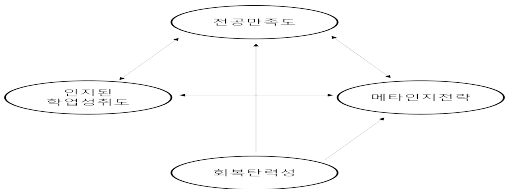


Figure 1. 전공만족도, 인지된 학업성취도, 메타인지전략, 회복탄력성의 매개효과

4. 연구도구 및 대상

전국에 재학중인 대학생들의 전공만족도에 따른 학업성취도와 회복탄력성의 관계성을 분석하기 위해 45명의 재학생들을 대상으로 설문조사를 실시하였다.

1) 연구척도

전공만족도의 척도를 알기 위해 “대학생의 전공선택동기, 전공만족도, 교수-학생 상호작용이 대학생활에 미치는 영향” 연구논문의 전공만족도 설문 문항에서 본 연구에 맞게 전공에 대한 만족도를 측정하기 위하여 하혜숙(2000)과 조원숙(2009)의 연구에서 사용된 설문지를 일반만족, 교과만족, 관계만족, 인식만족의 4개 하위요인 총 22문항으로 구성하였으며, 메타인지전략의 측정은 Pintrich & Garcia(1991)의 학습동기전략검사인MSLQ(Motivated Strategies for Learning Questionnaire)에서 학습

전략에해당하는 50문항 중 메타인지전략을 측정하는 문항인 12문항을 번안하여 사용한 허현경(2011)의 학위논문과 이정민, 윤성혜, 류혜선(2012)의 학술논문과 동일한 문항을 사용하였다. 인지된 학업성취도는 Eom, Wen, & Ashill(2006), 김현지&임진호(2007), Sun,Tsai, Finger, Chen, & Yeh(2008), 김정주(2009)의 연구에서 문항을 추출, 수정하여 구성한 강민석, 박인우(2010)의 측정도구로 사용하였으며, 이 검사도구의 문항은 총 4문항이다. 회복탄력성의 척도를 알기 위해 “대학생의 문제음주에 영향을 미치는 요인: 스트레스, 회복탄력성, 대인관계, 음주문화를 중심으로” 연구논문의 회복탄력성 설문 문항을 보고 본 논문과 연관된 설문문항을 대학생의 회복탄력성을 측정하기 위해 허희정(2022)의 연구에서 사용된 YKRQ-27도구를 사용하였으며, 총 문항은 26으로 구성하였다. 반응양식은 Likert 5점 척도로 하였다.

Table 2. 전공만족도, 학업성취도(인지된 학업성취도, 메타인지전략), 회복탄력성의 신뢰도 계수

구분	하위요인	문항	문항수	Cronbach α
전공만족도	일반만족	1,5,8,14,15,19,21	7문항	.887
	교과만족	2,3,7,9,12,18	6문항	.777
	관계만족	4,10,11,16	4문항	.819
	인식만족	6,13,17,20,22	5문항	.803
	전체			.932
학업성취도	인지된 학업성취도		4문항	.805
	메타인지전략		12문항	.697
회복탄력성			26문항	.707

5. 분석결과 및 해석

1) 전공만족도, 인지된 학업성취도, 메타인지전략, 회복탄력성 간의 영향(매개효과분석)

전공만족도가 인지된 학업성취도에 영향을 미치는지 알아보기 위해 단순선형회귀분석을 실시한 결과, F=34.287(p<.001)으로 본 회귀모형이 적합하다고 할 수 있다. R^2=0.438로 43.8%의 설명력을 나타냈다. 전공만족도는 β=0.662(p<.001)로 나타나 인지된 학업성취도에 유의한 영향을 미친다고 할 수 있으며, β부호가 정(+)적이므로 전공만족도가 증가하면 인지된 학업성취도도 높아지는 것으로 나타났다. 그리고, 전공만족도가 메타인지전략에 영향을 미치는지 알아보기 위해 단순선형회귀분석을 실시하였고, 분석 결과, F=54.158(p<.001)로 본 회귀모형이 적합하다고 할 수 있다. R^2=0.552로 55.2%의 설명력을 나타냈다. 전공만족도는 β=0.743(p<.001)으로 나타나 메타인지전략에 유의한 영향을 미친다고 할 수 있으며, β부호가 정(+)적이므로 전공만족도가 증가하면 메타인지전략는 높아지는 것으로 나타났다. 반면, 전공만족도가 회복탄력성에 영향을 미치는지 알아보기 위해 단순선형회귀분석을 실시하였으나, 분석 결과, F=24.433(p=0.084)로 본 회귀모형이 적합하지 않았다.

Table3. 전공만족도, 인지된 학업성취도, 메타인지전략, 회복탄력성 간의 매개효과 값

구분	변인			비표준화 계수		표준화 계수	t	p-value	R^2
				B	SE	β			
연구목적1	전공만족도	→	인지된 학업성취도	.647	.110	.662	34.287	<.001***	.438
	전공만족도	→	메타인지전략	1.747	.404	1.685	33.414	<.001***	.608
	전공만족도	→	회복탄력성	.252	.142	.182	24.433	.084	.636
연구목적2	인지된 학업성취도	→	전공만족도	0.677	.116	.662	34.827	<.001***	.438
	인지된 학업성취도	→	메타인지전략	1.131	.055	1.066	388.467	<.001***	.948
	인지된 학업성취도	→	회복탄력성	.045	.056	.032	257.097	.424	.948
연구목적3	메타인지전략	→	전공만족도	.716	0.97	.743	54.158	<.001***	.552
	메타인지전략	→	인지된 학업성취도	.801	.039	.850	492.530	<.001***	.958
	메타인지전략	→	회복탄력성	-.0024	.048	-.018	322.719	.619	.958
연구목적4	회복탄력성	→	전공만족도	.326	.097	.453	11.342	.002**	.205
	회복탄력성	→	인지된 학업성취도	.137	.126	.175	6.289	.282	.226
	회복탄력성	→	메타인지전략	.311	.103	.416	4.203	.004**	.231

*p<.05, **p<.01, ***p<.001

References

하혜숙, 김계현(2000), 대학생의 學科(學部)滿足의 要因에 관한 연구, 상담학연구

조원숙(2009).대학생의 전공-흥미 일치도에 따른 전공만족도 및 학업성취도와의 관계, 석사학위논문, 대진대학교 대학원

Pintrich&Garcia(1991), Student Motivation and Self-Regulated Learning : A LISREL Model, The University of Michigan

허현경(2011), 중등학생의 성취목표지향성, 메타인지, 학습전략과 학업성취 간의 구조모형 분석, 이화여대학교 교육대학원

이정민, 윤성혜, 류혜선(2012), 문제중심학습 성과에 대한 팀효능감, 메타인지의 예측력 규명, 아시아교육연구 13권 1호

Eom, S. B., Wen, H. J., & Ashill, N. (2006). The determinants of students' perceived learning outcomes and satisfaction in university online education: An empirical investigation. Decision Sciences Journal of Innovative Education, 4(2), 215-235.

김현지, 임진호. (2007). 초등학교의 교육정보화가교수학습문화 변화에 미치는 효과 분석. 교육공학연구, 23(1), 155-186. 10.17232/KSET.23.1.155

Sun, P., Tsai, R. J., Finger, G., Chen, Y., & Yeh, D. (2008). What drives a successful e-learning? An empirical investigation of the critical factors influencing learner satisfaction. Computers & Education, 50(4), 1183-1201.

김정주. (2009). 원격고등교육에서의 사회적 실재감 측정도구 개발. 고려대학교 박사학위논문

강민석, 박인우(2010). 이러닝 환경에서의 학습자-교수자 상호작용감 요인 모형에 대한 실증적 탐색. 교육공학연구, 26(2), 187-215.

허희정(2022), 간호대학생의 자기주도적 학습, 비판적 사고성향, 임상의사결정능력이임상수행능력에 미치는 영향, 아시아문화학술원, 제13권 제2호, 675-688

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 53-53

부경대 디지털스마트부산 아카데미

노맹석[1]

요 약

‘디지털스마트부산 아카데미’는 부산지역의 현장 수요 맞춤형 SW 전문인재를 키우기 위해 대학과 기업이 대규모 컨소시엄을 꾸려 교육과정을 공동 설계 및 운영하고, 채용까지 연계하는 과학기술정보통신부와 정보통신기획평가원의 ‘2022년 SW전문인재양성사업’입니다. 부경대를 주관으로 대학, 기업 등 12개 기관이 구성한 컨소시엄(책임교수 노맹석, 빅데이터융합전공)으로 4년간 약 78억 원을 지원받아 부산시 선도산업의 첨단화·스마트화를 위해, 디지털금융, 스마트헬스케어, 빅데이터, 스마트선박·항만, 스마트팩토리 5개 교육 분야에서 파이썬, 딥러닝 등 SW 기초과목을 수강한 부경대, 동아대, 동의대 재학생과 졸업생을 대상으로 640시간 기술교육 집중훈련을 통해 4년간 670명 이상의 SW전문인재를 양성합니다. 본 발표에서는 아카데미를 2년간 운영 실적과 성과에 대해서 소개하고자 합니다.

주요용어 : SW 전문인재, 파이썬, 딥러닝, 기술교육 집중훈련.

[1]48513 부산광역시 남구 용소로 45 부경대학교 빅데이터융합전공 교수

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 55-55

지역특성이 생활만족도에 미치는 영향

박현수[1], 이택면[2], 장안식[3]

요 약

이 연구는 생활만족도에 영향을 미치는 지역특성을 살펴보고자 하였다. 개인이 느끼는 생활만족도는 개인의 경제, 문화, 의료, 안전 등 다양한 영역에 대해 인식한 결과이다. 이러한 생활만족도는 개인이 각 영역별로 인식한 자신의 수준에 따라 결정될 것이다. 그러나 개인이 자신의 수준을 인식하는 과정에서 사회적 비교를 통해 그 수준을 인식할 수 있다. 즉 생활에 큰 불편함이 없는 상황에서도 다른 사람들의 모습을 비교하여 자신의 생활만족도 수준을 인식할 수 있을 것이다. 따라서 생활만족도가 자신에 대한 인식이면서 동시에 다른 사람들과의 비교를 통해 판단할 수 있다. 이 연구는 자신이 인식한 생활만족도에 대해 사회적 비교가 반영된 결과라는 점을 경험적으로 검증하고자 한다. 이를 위해 '2023 사회안전지수'의 정량지표 자료를 이용하였다. 이 자료는 사회안전지수를 구성하는 각 세부영역에 대한 지역 거주 주민들의 안전과 생활만족도에 대한 체감도를 측정하기 위해 조사를 실시한 결과이다. 각 시군구별로 성·연령 인구구성비를 고려하여 조사를 시행하였다. 이 자료에서 최소 50표본을 확보한 시군구에 대해서만 분석에 사용하였고, 이에 따라 분석대상은 184개 시군구이다. 이 자료에서 전반적 생활만족도와 경제활동, 생활안전, 건강보건, 그리고 주거환경 영역별로 만족도를 이용하여 분석하였다. 생활만족도에 영향을 미치는 지역의 특성으로는 각 영역별로 지역별 변동지수를 이용하였다. 앞서 설명한 바와 같이 생활만족도가 개인의 영역별 수준뿐만 아니라 사회적 비교에 의해서도 영향을 받는다면 지역 내 영역별 수준의 차이가 큰 지역일수록 개인이 인식하는 전반적인 생활만족도가 낮아질 수 있을 것이다. 이와 같은 연구의 가설을 검증하기 위해 생활만족도에 영향을 미치는 개인 수준과 지역 수준의 영향을 위계적 선형모형(Hierarchical Linear Model; HLM)을 사용하여 살펴보았다. 연구결과로는 전반적 생활만족도에 대해 개인수준에서의 경제활동, 생활안전, 건강보건, 그리고 주거환경 인식이 모두 정적(positive)인 영향을 미치고 있었다. 지역수준에서의 영역별 변동지수 중 주거환경에 대해서만 부적(negative)인 영향을 미치고 있었다. 즉 주거환경에 대한 지역의 인식차이가 큰 지역에서는 전반적인 생활만족도가 상대적으로 낮게 나타났다.

[1]충북대학교 국가위기관리연구소 시민치안연구센터장
[2]한국여성정책연구원 선임연구위원
[3]케이스탯리서치 공공사회정책연구소장

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 57-61

인과 포레스트를 활용한 대졸 청년층의 일과 전공의 일치가 첫 일자리 만족도에 미치는 영향 분석

백예은[1], 정혜원[2]

요 약

본 연구는 대졸 청년층의 일과 전공 일치가 첫 일자리 만족도에 미치는 영향을 탐색하기 위하여 수행되었다. 이를 위해 2019년 대졸자직업이동경로조사(GOMS; Graduates Occupational Mobility Survey) 자료를 활용하였으며 머신러닝 기반의 인과추론방법인 인과 포레스트(casual forest) 기법을 적용하여 대졸 청년층의 개인, 대학, 직장 특성을 공변인으로 설정하고 이를 통제하여 일과 전공 일치가 첫 일자리 만족도에 미치는 보다 정확한 평균 처치효과를 분석하고자 하였다. 더불어 차별적 처치효과를 살펴봄으로써 성별, 전공, 대학의 진로 및 취업 관련 프로그램 참여 여부, 직장 사업체 유형에 따라 일과 전공 일치가 첫 일자리 만족도에 미치는 효과가 다르게 나타나는지 탐색하였다. 그 결과 대졸 청년층의 개인, 대학, 직장 관련 공변인을 통제한 이후에도 일과 전공의 일치 정도가 첫 일자리 만족도에 정적인 영향을 미치는 것으로 나타났다. 더불어 성별, 학생 지원제도, 사업체 유형 등과 같은 특성에 따라 차별적 처치효과가 상이한 것으로 나타났다.

주요용어 : 인과포레스트, 인과추론, 대졸 청년층, 대졸자직업이동경로조사

1. 서론

오늘날 청년세대에게 일자리는 생계유지를 위한 수단일 뿐만 아니라 자신의 적성과 흥미를 발전시키거나 자아를 실현하게 하여 삶의 중요한 의미를 가진다. 특히 대학 졸업 후 처음 갖는 일자리에 대한 만족도를 의미하는 '첫 일자리 만족도'는 향후 개인의 노동시장 이행 과정과 청년층의 진로발달과정에 영향을 미침과 더불어 오늘날 청년세대에서 나타나는 실업 문제의 양상과 깊은 연관이 있어(오윤정, 유희영, 2020; 주영주, 한상윤, 2015), 일자리 만족도에 영향을 미치는 근무 환경이나 직무 특성을 탐색한 선행연구가 활발히 수행되었으며 그중 자신의 전공과 직무가 일치하는 정도에 따라 일자리 만족도가 달라지는 것으로 나타났다(박소영, 김주영, 2020; 심우정, 하지영, 2021; 이경민, 2021; 이다경, 김연후, 박현정, 2021; 최문석, 송일호, 2019). 일자리와 전공의 불일치는 노동시장에서 나타나는 직업불일치(job mismatch)중 하나로, 개인의 일자리 만족도에 부정적인 영향을 미칠 뿐만 아니라 그 자체로도 개인적, 사회적 측면에서 비효율을 야기하는 문제이다. 그러

[1]34134 대전 유성구 대학로 99, 충남대학교 일반대학원 교육학과 교육평가 전공 박사과정.
E-mail : byeunn7@gmail.com
[2](교신저자) 34134 대전 유성구 대학로 99, 충남대학교 일반대학원 교육학과 교육평가 전공 교수.
E-mail : chw7@cnu.ac.kr

나 일과 전공 일치가 일자리 만족도에 미치는 정확한 영향력을 분석하기 위해서는 일과 전공 일치 여부에 따라 구분되는 집단 간의 잠재적 차이로 인한 선택 편의(selection bias)의 문제를 고려해야 한다. 따라서 본 연구에서는 머신러닝 기반의 인과 포레스트(causal forest)(Wager, Athey, 2018) 기법을 통해 대졸 청년층의 개인, 대학, 직장 관련 공변인을 통제하고, 일과 전공 일치 여부가 첫 일자리 만족도에 미치는 평균 처치효과를 분석함으로써 통해 일과 전공 일치도와 일자리 만족 간의 관계를 보다 면밀하게 분석하고자 한다.

연구문제 1. 인과 포레스트를 활용하여 대졸 청년층의 개인, 대학, 직장 특성을 통제한 뒤에도 일과 전공 일치여부에 따라 첫 일자리 만족도에 유의한 차이가 나타나는가?

연구문제 2. 대졸 청년층의 개인, 대학, 직장 특성에 따라 일과 전공 일치 여부가 첫 직장 만족도에 미치는 효과에 차이가 있는가?

2. 연구방법

2.1. 연구대상

본 연구에서는 한국고용정보원이 제공하는 대졸자직업이동경로조사(GOMS: Graduates Occupational Mobility Survey) 자료 중 가장 최근 자료인 2019년 자료를 분석에 활용하였다. 본 연구에서는 대졸 청년층의 전공과 일자리의 일치여부가 일자리 만족도에 미치는 영향을 분석하기 위해서 청년기본법에서 제시한 청년 연령(19세~34세)를 고려하여, 대학 졸업자 중 현재 일자리가 첫 일자리라고 응답하거나 첫 일자리 경험이 있다고 응답한 응답자 중 35세 미만을 대상으로 분석을 실시하였다. 또한 분석에 투입된 공변인이 개인의 성별, 전공, 진로·취업 지원 프로그램의 참여 여부, 정규직 여부 등의 특성으로 구성되어 있어, 결측이 있는 경우 해당 사례를 제거하여 9,647명을 최종적으로 분석에 활용하였다.

2.2. 연구도구

본 연구에서는 처치 변인으로 일과 전공 일치 여부를 설정하였다. 이를 위해 일과 전공 일치도를 묻는‘첫 일자리(현재 일자리)에서 하셨던 일의 내용이 자신의 (편)입학 시 전공(주전공)과 어느 정도 맞았다고 생각하십니까?’문항을 활용하였으며 해당 문항에 ‘잘 맞는다’, ‘매우 잘 맞는다’응답은 1로 코딩하고 그 외의 응답을 0으로 코딩하였다. 그 결과, 일과 전공 일치 집단과 일과 전공의 일치 집단에 해당하는 학생은 4,248명(44%), 불일치 집단은 5,399명(56%)으로 나타났다. 더불어 결과 변인으로 ‘첫 일자리 만족도’로 설정하였으며 ‘첫 일자리(직장)에 대해 전반적으로 얼마나 만족하고 계십니까?’ 문항을 활용하였다. 마지막으로, 본 연구에서는 앞서 선행연구를 통해 일과 전공 일치도와 일자리 만족도에 영향을 미치는 것으로 나타난 변인들을 개인, 학교, 직장 영역으로 구분하여 공변인으로 투입하였다. 개인 변인으로 대졸 청년층의 성별, 전공계열, 부모님의 월 평균 소득, 아버지의 학력, 학자금 대출, 어학 연구, 취업 준비 경험 여부를 활용된 변인을 투입하였다. 또한 학교 변인으로 졸업한 대학의 본, 분교 여부 설립유형, 소재권역, 학생 지원제도 만족도, 진로관련 상담 및 지원제도 만족도, 진로 선택 및 취업 준비 프로그램 참여 여부를 활용하였다. 더불어 사업체 유형, 임금, 근로시간, 정규직 여부를 직장 관련 공변인으로 설정하였다.

2.2. 분석방법

인과 포레스트는 랜덤 포레스트 기법을 기반으로 하는 인과추론 방법 중 하나이다(Wager, Athey, 2018). 이러한 머신러닝 기반의 인과 포레스트는 투입되는 변수가 많고, 변수 간 복잡한 관계가 있을 때 데이터를 기반으로 모형을 유연하게 설정하기 때문에 기존의 전통적인 모수 기반의 인과추론 방법에서 연구자의 모형 설정으로 인하여 발생하는 편의를 줄일 수 있다는 장점이 있다.

인과 포레스트는 앞서 언급한 랜덤 포레스트 기법을 적용하여 공변인(X_i)과 처치 변인(Z_i)의 조건적 종속 평균 $m(x,z)=E[Y_i|X_i,Z_i]$를 통해 결과 모형을, 공변인의 조건적 처치 평균 $e(x)=E[Z_i|X_i]$을 통해 처치 모형을 추청한 뒤 가중 선형 회귀 접근법을 적용하여 조건적 평균 처치효과를 계산한다. 개별 조건적 평균 처치효과 $\tau(x)$는 아래 식 (1)을 통해 계산된다(석유미, 이진실, 2021; Wager, Athey, 2018).

$$\tau(x)=\frac{\Sigma_i\alpha_i(x)(Y_i-\widehat{m}^{-i}(X_i))(Z_i-\hat{e}^{-i}(X_i))}{\Sigma_i\alpha_i(x)(Z_i-\widehat{e^{-i}}(X_i))^2} \quad (1)$$

이를 구체적으로 살펴보면, $\alpha_i(x)$는 특정 사례(응답자) i가 $\tau(x)$를 계산하는데 미치는 공헌도이다. $-i$ 는 머신러닝에서 주로 활용되는 아웃 오브 백 리브원 아웃 방법을 의미하며, $\widehat{m}^{-i}(X_i)$와 $\hat{e}^{-i}(X_i)$는 앞서 언급한 바와 같이 랜덤 포레스트를 통해 추정되는 결과 모형과 처치모형이다. 이를 통해 도출된 조건적 평균 처치효과를 평균 내어 평균 처치효과인 처치집단과 통제집단의 결과변인의 파이 $E(Y_i(Z_i=1)-Y_i(Z_i=0))$을 계산한다. 즉, 본 연구에서는 일과 전공이 일치하는 집단과 일치하지 않는 집단의 평균적인 첫 일자리 만족도 차이가 평균 처치효과에 해당한다. 이에 본 연구에서는 일과 전공 일치 여부가 첫 일자리 만족도에 미치는 평균 처치효과를 추정하고 대졸 청년층의 특성(성별, 전공, 직장 유형, 대학 만족도, 진로 선택 및 취업 준비 프로그램 참여 여부)에 따른 차별적 처치효과를 확인하고자 하였으며 분석을 위해 R 프로그램(Ver. 4.0.2) grf 패키지(Ver. 2.1.0)(Tibshirani et al., 2001)를 사용하였다.

3. 연구결과

3.1. 전공과 일자리 일치 여부가 첫 일자리 만족도에 미치는 평균 처치효과 분석

인과 포레스트를 적용하여 대졸자 청년층의 일과 전공 일치여부가 첫 일자일 만족도에 미치는 평균 처치효과를 분석한 결과, 공변인 통제 전의 일과 전공 일치집단과 불일치 집단의 첫 일자리 만족도는 차이는 0.496으로 통계적으로 유의하게 나타났다. 인과 포레스트를 통해 공변인을 통제한 뒤의 평균 처치효과 추정치는 0.379으로, 일과 전공 일치가 첫 일자리 만족도에 여전히 정적으로 유의한 영향을 미치는 것으로 나타났다.

Table 1. 평균 처치효과 분석결과

	추정치	표준오차	*t*	*p* value
공변인 통제 전	0.496	0.010	24.94 ***	<.001
공변인 통제 후	0.379	0.020	18.897 ***	<.001

***p<.001

3.2. 일과 전공 일치가 첫 일자리 만족도에 미치는 차별적 처치효과 분석

대졸 청년층의 일과 전공 일치가 첫 일자리 만족도에 미치는 차별적 처치효과를 분석한 결과, 개인, 대학, 직장 변인에서 유의하게 나타난 차별적 처치효과를 확인할 수 있었으며 이를 살펴보면 다음과 같다. 우선 개인 변인 중에서 성별에 따라 일과 전공 일치가 첫 일자리 만족도에 미치는 효과가 유의하게 달라지는 것으로 나타났으며(t=2.335, p=.020) 이는 대졸 청년 근로자의 성별이 남성일 때 일과 전공 일치가 첫 일자리 만족도에 미치는 효과가 더 큰 것을 의미한다. 또한 대학 변인 중에서 학생 지원제도 만족도의 수준에 따라 일과 전공 일치의 처치효과가 유의하게 다른 것으로 나타났고(t=-2.279, p=.023), 취업 캠프 참여 여부 또한 유의한 것으로 나타났다(t=2.468, p=.014), 이는 대졸 청년 근로자가 대학에서 제공하는 장학금, 해외 연수 등의 지원 제도에 만족할수록 일과 전공 일치가 첫 직장 만족도에 미치는 효과가 작은 반면, 취업 캠프에 참여한 경험이 있을수록 그 효과가 큰 것으로 해석할 수 있다. 더불어 대졸 청년 근로자의 직장 변인에서 또한 유의한 차별적 처치효과를 확인할 수 있었는데, 첫 직장의 사업체 유형에 따라 통계적으로 유의한 차별적 처치효과가 나타났다. 구체적으로 살펴보면, 공무원이나 군인과 같은 국가 기관에 속한 근로자보다 내국인이 운영하는 민간 회사 또는 개인 사업체에 속한 근로자일수록 일과 전공 일치가 첫 일자리 만족도에 미치는 효과가 작았다(t=-2.255, p=.024).

이와 더불어 대졸 청년 근로자의 졸업 대학에서 제공한 학생 지원제도 만족도와 진로 관련 지원 제도 만족도에 따른 처치효과 분포를 Figure 1과 같이 히트맵(Heatmap)을 통하여 시각화 한 뒤, 그 분포도를 확인하였다. 학생 지원제도 만족도와 진로 관련 지원제도 만족도는 1~5점으로 측정된 척도에 따라 5개 수준으로 구분하였으며 각 만족도에 따른 일과 전공 일치가 첫 직장 만족도에 미치는 처치효과 평균값의 크기가 클수록 진한 색으로 표현하였다. Figure 1에서 확인 할 수 있듯이 장학금, 어학연수와 같은 학생 지원제도에 만족하는 수준이 높을수록 더 옅게 나타난 반면 진로 관련 지원 제도의 경우 수준에 따른 뚜렷한 패턴을 보이지 않았다.

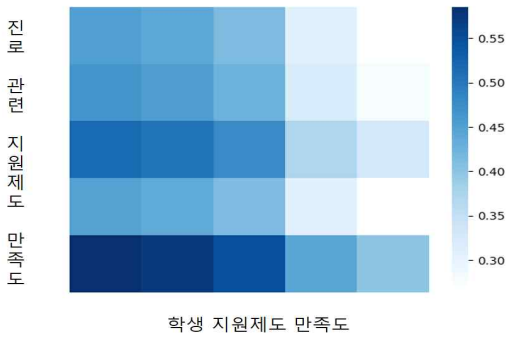


Figure 1. 지원제도 만족도에 따른 처치효과 히트맵

4. 결론

일과 전공의 일치가 첫 직장 만족도에 미치는 평균 처치효과를 분석한 결과, 개인, 대학, 직장 특성을 공변인으로 통제한 이후에도 여전히 일과 전공의 일치여부가 첫 일자리 만족도에 정적으로 유의한 효과를 보였다. 이는 대졸 청년 근로자가 자신의 전공과 첫 직장에서 업무가 일치할수록 첫 직장에서의 만족도가 높음을 의미하며, 일과 전공의 미스매치 문제가 단순히 개인의 시간과 비용을 투자한 대학에서의 전공 교육의 비효율성 문제를 야기할 뿐만 아니라 첫 직장에서 근로자의 주관적인 만족도와 같은 질적인 측면에서의 노동시장 성과에도 부정적인 영향을 미침을 시사한다. 더불어 성별이나 진로 및 취업 지원 프로그램 참여 여부, 첫 직장의 사업체 유형 등 대졸 청년 근로자의 특성에 따라 일과 전공의 일자리 만족도에 대한 효과가 다르게 나타남을 확인하여 두 변인 간 관계를 보다 면밀하게 분석하였다는데에 본 연구의 의의가 있다.

References

박소영, 김주영(2020). 전공-직무일치, 임금, 직장만족 간 구조적 관계의 전공별 차이 분석. *미래교육학연구*, 33(1), 1-23.

석유미, 이진실(2021). 머신러닝 기반의 인과 포레스트 기법을 활용한 처치효과 검증: 교내 동아리활동 참여가 협업능력에 미치는 효과를 중심으로. *조사연구*, 22(4), 55-78.

심우정, 하지영(2021). 대학 교육은 어떻게 일자리 만족도에 영향을 주는가. *미래교육학연구*, 34(2), 155-177.

오윤정, 유희영(2020). 전공만족도와 자기효능감이 대학졸업자의 첫 직장만족도에 미치는 매개효과 탐색. *학습자중심교과교육연구*, 20(21), 323-341.

이경민(2021). 졸업대학 관련 만족도와 전공-직무일치가 현 직장 만족도에 미치는 영향. 관광연구, 36(6), 45-62.

이다경, 김연후, 박현정(2020). 대학 졸업생의 직업만족도에 영향을 미치는 변수 탐색: 벌점회귀모형 sparse group lasso를 활용하여. *아시아교육연구*, 21(4), 1069-1097.

주영주, 한상윤(2015). 대졸 청년층의 대학생활 만족도 및 첫 직장만족도 관련 영향변인에 관한 연구. *교육종합연구*, 13(1), 193-212.

최문석, 송일호(2019). 청년층의 교육 및 전공불일치가 임금과 직장만족도에 미치는 영향. *사회과학연구*, 26(2), 85-102.

Tibshirani, J., Athey S., Friedberg, Rina., Hadad, V., Hirshberg. D., Miner L., Sverdrup E., Wager S., & Wright M. (2021). GRF: Generalized Random Forest. https://CRAN.R-project.org/package=grf.

Wager, S., & Athey, S. (2018). Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests. *Journal of the American Statistical Association*, 113(523), 1228-1242.

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 63-63

Scalable kernel balancing weights in a nationwide observational study of hospital profit status and heart attack outcomes

Kwangho Kim[1], *Bijan A. Niknam*[2], *Jose R. Zubizarreta*[2]

Abstract

Longstanding questions exist about the relationship between profit motives and healthcare treatment and outcomes, which generally must be examined using observational studies with covariate adjustment. Weighting is a general and often-used method for statistical adjustment. Weighting has two objectives: first, to balance covariate distributions, and second, to ensure that the weights have minimal dispersion and thus produce a more stable estimator. A recent, increasingly common approach directly optimizes the weights toward these two objectives. However, this approach has not yet been feasible in large-scale datasets when investigators wish to flexibly balance general basis functions in an extended feature space. For example, many balancing approaches cannot scale to national-level health services research studies. To address this practical problem, we describe a scalable and flexible approach to weighting that integrates a basis expansion in a reproducing kernel Hilbert space with state-of-the-art convex optimization techniques. Specifically, we use the rank-restricted Nystrom method to efficiently compute a kernel basis for balancing in nearly linear time and space, and then use the specialized first-order alternating direction method of multipliers to rapidly find the optimal weights. In an extensive simulation study, we provide new insights into the performance of weighting estimators in large datasets, showing that the proposed approach substantially outperforms others in terms of accuracy and speed. Finally, we use this weighting approach to conduct a national study of the relationship between hospital profit status and heart attack outcomes in a comprehensive dataset of 1.27 million patients. We find that for-profit hospitals use interventional cardiology to treat heart attacks at similar rates as other hospitals, but have higher mortality and readmission rates.

keywords: Causal Inference; Observational Studies; Weighting; Convex Optimization; Propensity Score.

[1]Department of Statistics, Korea University, 145 Anam-ro, Seongbuk-gu, Seoul, 02841, South Korea; email: kwanghk@korea.ac.kr.
[2]Departments of Health Care Policy, Biostatistics, and Statistics, Harvard University, 180 Longwood Avenue, Office 307-D, Boston, MA 02115, United States

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 65-65

동적 마진 할당을 통한 강건한 대조학습

소준혁[1], 임용택[2], 김예원[3], 오창대[4], 송경우[5]

요 약

대조 학습은 표현 학습에서 많이 사용되는 방법 중 하나이다. 의미론상으로 유사한 샘플이 임베딩 공간에서 가까운 곳에 위치하게, 의미론상으로 유사하지 않은 샘플은 임베딩 공간에서 먼 곳에 위치하게 유도하는 방식으로 모델을 학습시킨다. 그러나 대조 학습은 구분이 쉽게 되는 긍정 쌍과 부정 쌍에 의해 모델이 유의미한 학습을 못 하게 되어 기울기 소실 현상이 발생할 수 있다. 이는 모델이 다양한 샘플로 충분히 학습되지 않는 것을 의미하고, 이는 과적합과 같은 문제를 야기할 수 있다. 이 문제를 해결하기 위해, 우리는 동적 혼합 마진 손실 (Dynamic Mixed Margin, DMM)을 제안한다. DMM을 사용하면, 믹스업을 이용해 샘플을 보간하여 쉽게 구분되지 않는 긍정-부정 쌍을 생성한다. 또한, 동적으로 마진을 할당하여 모델의 표현 학습을 개선한다. DMM은 멀리 있는 긍정 쌍은 가깝게 만들고, 가깝고 구분하기 쉬운 긍정 쌍은 멀어지도록 유도하여 과적합을 완화한다. DMM은 플러그 앤 플레이 모듈이고, 따라서 다양한 대조학습 손실과 모델들에 호환 가능한 모델이다. 우리는 이미지 검색, 비디오-텍스트 검색, 다중 모달 감성 분석, 추천 시스템과 같은 다양한 도메인에서 DMM이 베이스라인의 성능을 개선한 것을 보인다. 또한, DMM을 통해 학습된 표현이 실제 환경에서 자주 사용되는 전이 학습과, 모달리티 누락이 발생한 경우 등에서 더욱 강건한 성능을 보이는 것을 확인한다.

주요용어 : 다중 모달 학습, 대조 학습, 강건 학습

---

*이 (성과)는 정부(과학기술정보통신부)의 재원으로 한국연구재단
(NO.2022M3J6A1084845, No. 2021R1F1A1060117)의 지원으로 수행된 연구이며 이에 감사드립니다.
[1]37673 대한민국 경상북도 포항시 남구 청암로 77 포항공과대학교 컴퓨터공학과 석사과정.
Email: goqhadl9298@gmail.com
[2]02054 대한민국 서울특별시 서울시립대로 163 서울시립대학교 인공지능학과 석사과정.
E-mail: yongtaek.lim@uos.ac.kr
[2]02054 대한민국 서울특별시 서울시립대로 163 서울시립대학교 인공지능학과 석사과정.
Email: yeyewon12@uos.ac.kr
[2]02054 대한민국 서울특별시 서울시립대로 163 서울시립대학교 인공지능학과 석사과정.
Email: changdae.oh@uos.ac.kr
[2](교신저자)02054 대한민국 서울특별시 서대문구 연세로 50 연세대학교 응용통계학과 조교수.
E-mail: kyungwoo.song@gmail.com

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 67-67

Forecasting of annual electricity consumption in Vietnam using radial basis function neural network

Bui Thanh Hoa[1], *이근제*[2]

요 약

Electricity consumption forecasting plays an important role in short-term load allocation and long-term planning for new generation and transmission infrastructures in developing countries. Precise forecast results are difficult to obtain due to the economic system's intrinsic complexity. This work proposed a machine learning model based on the radial basis function neural network (RBFNN) for estimating annual electricity consumption in Vietnam from 2024 – 2030. Socioeconomic factors such as population, GDP, values of imports and exports were used as the input for the model. Eight models with different combinations of input variables were built to determine the proper combination for the final forecasting model. The collected dataset was split into the training set (1990 – 2015) and validation set (2016 – 2020) to train and validate the proposed models. For forecasting electricity consumption from 2024 – 2030, the GDP values followed the GDP growth scenario from PDP8, and other factors were obtained by linear extrapolation from 2016 – 2020. The Shapley Additive exPlanations method was used to explain the contribution of each input variable on the forecasting results. The forecasting ability of the final model was tested by comparing the predicted consumptions from 2021 – 2023 with true consumption (2021 - 2022) and predicted consumption from EVN (2023). Additionally, the electricity consumptions predicted by the Power Development Plan VIII (PDP8) were also compared and discussed. The results showed a good agreement between the forecasted consumptions by the RBFNN and the true consumptions, whereas the PDP8 overestimated the electricity consumption in Vietnam.

Keyword : Electricity consumption, radial basis function neural network, Power Development Plan VIII.

[1](교신저자) 46241 부산광역시 금정구 부산대학로63번길 2, 부산대학교 경제학과 석사과정.
E-mail: buithanhhoa@pusan.ac.kr
[2]46241 부산광역시 금정구 부산대학로63번길 2, 부산대학교 경제학과 교수. E-mail: kjlee@pusan.ac.kr

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 69-74

A Copula Based Unsupervised Domain Adaptation for Image Classification*

Seungmin Lee[1], *Kyupil Yeon*[2]

Abstract

In this paper, we propose an unsupervised domain adaptation algorithm for image classification using principal component analysis (PCA) and Gaussian copula function alignment. The method is similar to the CORAL algorithm which aligns the correlation structure between source and target domains, but different in that it applies correlation alignment in copula spaces instead of the original variable spaces. Since a copula function enables us to analyze separately the dependency structure from the marginal distributions, the proposed algorithm is considered to be robust to a severely skewed characteristic of the marginals that can distort the correlation structure among the variables. We compared several feature level domain adaptation algorithms for image classification using office-caltech10 data set, and verified the proposed method showed better classification accuracy in an unsupervised domain adaptation framework.

Keywords: copula function, correlation alignment, domain adaptation, image classification, PCA

1. Introduction

Conventional machine learning assumes that the training and the test data come from the same distribution, but this is often not the case in reality. Domain adaptation (DA) learning focuses on building predictive models for source and target domains with differing distributions. DA often faces limited labeled data in the target domain under which a supervised learning is not easy to utilize. However, using related information from the labeled source domain can reduce the need for extra labeling in the target domain. Theoretical researches in domain adaptation learning, such as the works of Ben-David et al. (2010),

*This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. 2020R1F1A1A01073456). This research was supported by "Regional Innovation Strategy (RIS)" through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (MOE) (2021RIS-004).

[1]Graduate student, Department of Data Science, Hoseo University, Asan, Chungnam, 31499, Korea. E-mail: leeseungmin_22@naver.com

[2](Corresponding Author) Associate Professor, Division of Big Data and AI, Hoseo University, Asan, Chungnam, 31499, Korea. E-mail: kpyeon1@hoseo.edu

Mansour et al. (2009), Germain et al. (2016), Redko et al. (2017), suggest that effective domain adaptation requires high predictive accuracy in the source domain and minimal distributional differences between domains. Based on these insights, various feature level domain adaptation algorithms have been developed.

2. Review of Domain Adaptation and Copula Theory

2.1. Unsupervised domain adaptation

In this study we consider unsupervised domain adaptation in a homogeneous setting, where both source and target domains share the same input and label spaces, but only the source domain is labeled. The main challenge involves training a classifier for the label-lacking target domain while addressing distribution discrepancies. To this end, algorithms are employed to find domain-invariant features in the labeled source domain that are predictive for the unlabeled target domain. CORAL algorithm by Sun et al. (2016) performs a correlation alignment to identify new features for classification, constructs a classifier with the aligned new features of source domain data, and then applies the derived model for predicting the target domain data. Subspace Alignment (SA) algorithm by Fernando et al. (2013) discovers latent features by applying PCA to define a subspace, subsequently aligning the principal axes of the source domain with those of the target domain for domain adaptation. Sun, Saenko (2015) further improved SA by additionally aligning the variances of the latent features and referred it subspace distribution alignment (SDA) algorithm.

2.2. Copula functions

In general, a copula function is just a joint cumulative distribution $P(U_1 \leq u_1, \cdots, U_d \leq u_d)$ of standard uniform random variables $U_1, \cdots, U_d$. The importance of a copula is well represented by Sklar's theorem which states that when we have a d-dimensional random vector $\boldsymbol{X} = (X_1, \cdots, X_d)$ with its cumulative distribution function F_X and the continuous marginal distributions $F_1(x_1), \cdots, F_d(x_d)$, then a d-copula function C uniquely exists such that

$$F_{\boldsymbol{X}}(x_1, \cdots, x_d) = C\big(F_1(x_1), \cdots, F_d(x_d)\big) \tag{1}$$

for all $\boldsymbol{x} = (x_1, \cdots, x_d) \in \mathbb{R}^d$. A copula links together a joint distribution and marginal distributions. The theorem also indicates that given a copula $C: [0,1]^d \rightarrow [0,1]$ and continuous univariate distribution functions $F_1, \cdots, F_d$, the function $F_{\boldsymbol{X}}$ defined in (1) is a joint distribution having its marginals with $F_1, \cdots, F_d$. By differentiating the equation (1) we obtain the multivariate density function expressed with a copula density and marginal density functions as follows.

$$f_{\boldsymbol{X}}(x_1,\cdots,x_d) = c(F_1(x_1),\cdots,F_d(x_d))\prod_{i=1}^{d} f_i(x_i) \tag{2}$$

The Gaussian copula is defined as follows.

$$C(u_1,\cdots,u_d) = \Phi_R\left(\Phi^{-1}(u_1),\cdots,\Phi^{-1}(u_d)\right),$$

where Φ_R is d dimensional Gaussian CDF with a correlation matrix R and Φ^{-1} is the inverse CDF function of the standard univariate normal distribution. With the given Gaussian copula C, marginal CDF F_i and marginal density f_i for $i=1,\cdots,d$, the joint CDF and PDF of X are represented as

$$F(\boldsymbol{x}) = C(F_1(x_1),\cdots,F_d(x_d)) = \Phi_R\left(\Phi^{-1}(F_1(x_1)),\cdots,\Phi^{-1}(F_d(x_d))\right)$$
$$f(\boldsymbol{x}) = \frac{1}{|R|}\exp\left\{-\frac{1}{2}\boldsymbol{z}^T(R^{-1}-I)\boldsymbol{z}\right\}\cdot\prod_{i=1}^{d} f_i(x_i)$$

where $\boldsymbol{z}=\left(\Phi^{-1}(F_1(x_1)),\cdots,\Phi^{-1}(F_d(x_d))\right)^T$.

3. Copula Based Unsupervised Domain Adaptation

We propose a copula-based unsupervised domain adaptation for classification, aligning Gaussian copula correlation matrices to tackle the distortion from skewed marginal distributions. This approach separates dependency modeling from marginals and it mitigates some limitations of the CORAL algorithm when faced with skewed data. It can enhance domain adaptation by focusing on dependency structures using copula functions, regardless of some skewness of marginal distributions.

Algorithm. Copula Alignment for unsupervised domain adaptation

Given a scaled source domain data matrix $X_s=(x_{ij})_{n_s\times p}$ with a corresponding label $\boldsymbol{y}=(y_1,\cdots,y_{n_s})^T$ where $y_i\in\{1,2,\cdots,\ell\}$ and a scaled target domain data matrix $X_t=(x_{ij})_{n_t\times p}$,

① Obtain eigenvalues and eigenvectors for source domain with $d(<p)$ components.

$$R_s \approx P_s E_s P_s^T,$$

where P_s is a eigenvector matrix and E_s is a diagonal matrix with d largest eigenvalues.

② Principal scores $S_s = X_s P_s = (s_{ij}^s)_{n_s\times d}$, $S_t = X_t P_s = (s_{ij}^t)_{n_t\times d}$.

③ Probability integral transform using empirical CDF $F_1^s,\cdots,F_d^s$ and $F_1^t,\cdots,F_d^t$.

$$\boldsymbol{u_j^s} \equiv F_j^s(\boldsymbol{s_j^s}) = \left(F_j^s(s_{1j}^s),F_j^s(s_{2j}^s),\cdots,F_j^s(s_{n_s j}^s)\right)^T,\ \boldsymbol{u_j^t} \equiv F_j^t(\boldsymbol{s_j^t}) = \left(F_j^t(s_{1j}^t),F_j^t(s_{2j}^t),\cdots,F_j^t(s_{n_s j}^t)\right)^T$$

④ Copula features using standard normal scores.

$$\boldsymbol{z_j^s} \equiv \Phi^{-1}(\boldsymbol{u_j^s}) = \left(\Phi^{-1}(F_j^s(s_{1j}^s)),\cdots,\Phi^{-1}(F_j^s(s_{n_s j}^s))\right)^T,\ \boldsymbol{z_j^t} \equiv \Phi^{-1}(\boldsymbol{u_j^t}) = \left(\Phi^{-1}(F_j^t(s_{1j}^t)),\cdots,\Phi^{-1}(F_j^t(s_{n_t j}^t))\right)^T$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function.

⑤ Let $Z_s = (\boldsymbol{z_1^s},\boldsymbol{z_2^s},\cdots,\boldsymbol{z_d^s}) = (z_{ij}^s)_{n_s \times d}$, $Z_t = (\boldsymbol{z_1^t},\boldsymbol{z_2^t},\cdots,\boldsymbol{z_d^t}) = (z_{ij}^t)_{n_t \times d}$ and calculate the correlation coefficient matrice $R_s = \frac{1}{n_s - 1} Z_s^T Z_s$, $R_t = \frac{1}{n_t - 1} Z_t^T Z_t$.

⑥ Adjust the source data so that $Z_s^* = Z_s R_s^{-1/2} R_t^{1/2}$.

⑦ Construct a classification model using Z_s^* and $\boldsymbol{y_s}$, and predict the target data Z_t.

4. Data Analysis

We analyzed the Office-Caltech10 dataset, which consists of images from four domains (Amazon(A), Webcam(W), DSLR(D), Caltech256(C)) processed by the SURF algorithm. Each image in the data set is described by 800 standardized features. Our evaluation involved 12 domain adaptation settings, using a random forest classifier for testing. Six models were compared: non-adaptive 'No_Adapt', 'PCA' based model, and four domain adaptation models ('CORAL', 'SA', 'SDA', our proposed 'Copula_DA'). For each class in the source domain, samples were split into training and testing sets in 7:3 and the process was iterated 20 times. Table 1 shows the average accuracy and standard deviation of models for each DA setting, The accuracy of the best model in each domain adaptation setting is shown in bold font.

We can verify that the proposed method shows better classification performance than other methods in most cases, except for D→A and D→W, where SDA is the best. Interestingly, domain adaptation using correlation alignment has proven to be more effective in copula space than in original space.

5. Conclusions

We proposed a copula based unsupervised domain adaptation algorithm. The main idea is to apply dimension reduction and extract copula features adjusted by aligning Gaussian copula correlation matrices between domains. Since copula functions enable the analysis to separate dependency modeling and marginal distributions, the proposed method is specifically useful in case of highly skewed marginal distributions.

Table 1. Average classification accuracy and standard deviation

Source	Target	No Adapt	CORAL	PCA	SA	SDA	Copula_DA
C	A	0.3156 (0.0208)	0.4487 (0.0160)	0.4190 (0.0241)	0.4314 (0.0146)	0.4223 (0.0140)	**0.4747** (0.0139)
D	A	0.2562 (0.0169)	0.2584 (0.0218)	0.2281 (0.0284)	0.2381 (0.0176)	**0.3114** (0.0150)	0.2858 (0.0169)
W	A	0.2365 (0.0177)	0.2939 (0.0215)	0.2393 (0.0179)	0.2829 (0.0211)	0.3280 (0.0167)	**0.3388** (0.0178)
A	C	0.2803 (0.0135)	0.3762 (0.0106)	0.3534 (0.0170)	**0.4064** (0.0099)	0.3759 (0.0118)	0.4033 (0.0117)
D	C	0.2461 (0.0127)	0.2560 (0.0160)	0.2336 (0.0186)	0.2401 (0.0154)	0.2913 (0.0121)	**0.2957** (0.0127)
W	C	0.2421 (0.0137)	0.2734 (0.0134)	0.2345 (0.0188)	0.2546 (0.0153)	0.2951 (0.0112)	**0.3045** (0.0152)
A	D	0.2385 (0.0277)	0.3111 (0.0254)	0.2846 (0.0359)	0.3443 (0.0310)	0.3211 (0.0234)	**0.3834** (0.0261)
C	D	0.2269 (0.0291)	0.3885 (0.0307)	0.2987 (0.0345)	0.4507 (0.0270)	0.3526 (0.0318)	**0.4700** (0.0271)
W	D	0.5648 (0.0274)	0.6869 (0.0304)	0.7009 (0.0287)	0.7369 (0.0250)	0.7509 (0.0245)	**0.7710** (0.0242)
A	W	0.2237 (0.0245)	0.3101 (0.0187)	0.2973 (0.0264)	0.3447 (0.0205)	0.3202 (0.0189)	**0.3961** (0.0201)
C	W	0.2383 (0.0238)	0.3200 (0.0261)	0.3157 (0.0312)	0.3250 (0.0251)	0.3270 (0.0256)	**0.4406** (0.0221)
D	W	0.5660 (0.0230)	0.6245 (0.0304)	0.6155 (0.0295)	0.5602 (0.0286)	0.7037 (0.0232)	0.6347 (0.0313)

References

Bayestehtashk, A., Shafran, I., Babaeian, A. (2016). Robust speech recognition using multivariate copula models, IEEE International Conference on Acoustics, *Speech and Signal Processing (ICASSP)*, 5890-5894. DOI: 10.1109/ICASSP.2016.7472807

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., Vaughan, J. W. (2010). A theory of learning from different domains, *Machine Learning*, 79, 151-175. DOI: 10.1007/s10994-009-5152-4

Ben-David, S., Blitzer, J., Crammer, K., Pereira, F. (2006). Analysis of representations for domain adaptation, *Proceedings of the 19th International Conference on Neural Information Processing Systems*, 137-144. DOI: 10.7551/mitpress/7503.003.0022

Choi, K. H., Yoon, S. M. (2019). The relationship between international oil price and investor sentiment using copula model, *Journal of the Korean Data Analysis Society*, 21(6), 2961-2974.

Fernando, B., Habrard, A., Sebban, M., Tuytelaars, T. (2013). Unsupervised visual domain adaptation using subspace alignment. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, *Australia*, 2960-2967. DOI: 10.1109/ICCV.2013.368

Germain, P., Habrard, A., Laviolette, F., Morvant, E. (2016). A new PAC-Bayesian perspective on domain adaptation, *Proceedings of the 33rd International Conference on Machine Learning*, 48, 859-868. DOI: https://doi.org/10.48550/arXiv.1506.04573

Liang, C., Zhu, X., Li, Y., Sun, X., Chen, J., Li, J. (2013). Integrating credit and market risk: A factor copula based method, *Procedia Computer Science*, 17, 656-663.

Mansour, Y., Mohri, M., Rostamizadeh, A. (2009). Domain adaptation: Learning bounds and algorithms, Proceedings of The 2nd Annual Conference on Learning Theory.

Redko, I., Habrard, A., Sebban, M. (2017). Theoretical analysis of domain adaptation with optimal transport, *European Conference on Machine Learning and Knowledge Discovery in Databases*, 737-753.

Rhee, B. K. (2021). Asymmetric dependence between Korean sovereign CDS and VKOSPI, *Journal of the Korean Data Analysis Society*, 23(1), 47-60. DOI: https://doi.org/10.37727/jkdas.2021.23.1.47

Song, S., Byun, K. (2021). Value at risk of portfolios using copulas, *Communications for Statistical Applications and Methods*, 28(1), 59-79. DOI: https://doi.org/10.29220/CSAM.2021.28.1.059

Sun, B., Feng, J., Saenko, K. (2016). Return of frustratingly easy domain adaptation, *Proceeding of the AAAI Conference on Artificial Intelligence*, 30, 2058-2065.
DOI: https://doi.org/10.1609/aaai.v30i1.10306

Sun, B., Saenko, K. (2015). Subspace distribution alignment for unsupervised domain adaptation. *In Proceedings of the British Machine Vision Conference*, 24.1-24.10.
DOI: https://dx.doi.org/10.5244/C.29.24

Tran, C. D., Rudovic, O., Pavlovic, V. (2017). Unsupervised domain adaptation with copula models, *IEEE International Workshop on Machine Learning for Signal Processing*, 1-6.
DOI: 10.1109/MLSP.2017.8168131

Yeon, K. (2023). A comparative study on unsupervised domain adaptation using subspace alignment, *Journal of the Korean Data Analysis Society*, 25(1), 161-171. (in Korean).
DOI: https://doi.org/10.37727/jkdas.2022.25.1.161

Yeon, K. (2019). An ensemble approach to domain adaptation in sentiment analysis, *Journal of the Korean Data Analysis Society*, 21(4), 1645-1653.

Zhang, X., Jiang, H. (2019). Application of copula function in financial risk analysis, *Computers and Electrical Engineering*, 77, 376-388. DOI: https://doi.org/10.1016/j.compeleceng.2019.06.011

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 75-75

A Study on Deep Semi-supervised learning method using Data-adaptive Augmentation Technique

Seri Park[1], *Dongha Kim*[2]

Abstract

This study introduces a novel semi-supervised learning approach specifically designed for tabular data, featuring a unique learnable data augmentation technique that preserves the labeled data' information. The approach is mainly motivated by two methods: MixMatch, known as one of the state-of-the-art semi-supervised learning methods in image data, and Neutral AD, a self-supervised learning method for anomaly detection. These inspirations are adapted to tabular data through an innovative loss function comprising three distinct parts: one for labeled data, one for unlabeled data, and another for deterministic contrastive learning. This loss function is pivotal in guiding transformations that produce diverse and informative data augmentations, while preserving the characteristics of the original data. To validate our proposed method, we perform experiments on three tabular datasets, where our method demonstrates remarkable state-of-the-art performance, especially on the two datasets. The results not only show superior test accuracy over several baselines, but also highlight the importance of each components' role by the tuning hyperparameters proposed in the ablation studies.

Keywords: Data augmentation, Deep learning, Tabular data, Semi-supervised learning.

[1]02844 Bomun-ro, 34 Da-gil 2, Seongbuk-gu, Seoul, Korea, Master's Student, Department of Statistics, Sungshin Women's University. E-mail: seri.jpark96@gmail.com

[2](Corresponding author) 02844 Bomun-ro, 34 Da-gil 2, Seongbuk-gu, Seoul, Korea, Assistant Professor, Department of Mathematics, Statistics and Data Science, Sungshin Women's University. E-mail: dongha0718@sungshin.ac.kr

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 77-78

머신러닝 알고리즘을 이용한 서울시 행정동별 상권 활성화 지수 및 폐업리스크 예측

박지호[1], 백예은[2], 최정빈[3], 김동하[4]

요 약

현재 증가하는 개업 상권의 규모만큼 폐업하는 상권 또한 증가하고 있다. 이에 따른 정교한 상권분석이 필요한 시점이다. 현재의 상권분석 시스템들은 단순한 통계 수치만을 보일 뿐 리스크 제공은 하지 않고 있다. 이를 해결하기 위해 본 연구에서는 소상공인 상점 개업 시 폐업 리스크를 예측하는 모형을 개발한다. 이를 위해 개·폐업 데이터(마이 데이터)를 기반으로 생활인구, 행정동별 추정 매출, 임대 시세 데이터를 복합하여 상권 활성화 지수를 예측하는 머신러닝 모델을 개발한다. 해당 모형으로부터 도출된 상권 활성화 지수를 원 데이터와 함께 활용하여 소상공인의 상점 개업 시 폐업 리스크를 예측하는 모형까지 도출한다. 본 연구는 업종, 개업 위치별 폐업리스크 도출 및 업종별 최적의 위치 추천, 행정동별 최적의 개업 업종 추천이 가능하다는 장점이 있다. 따라서 소상공인이나 투자자에게 중요한 비즈니스 결정 및 리스크 관리를 위한 정보를 제공하고, 새로운 경제 지원 정책을 이끌어내 폐업률을 낮추는 데 기여할 수 있을 것이다.

주요용어 : 상권 활성화 지수 예측, 폐업 리스크 예측, 최적 업종 및 위치 추천.

1. 서론

1.1 연구 배경 및 목적

통계청 『전국 사업체 조사』(통계청,「전국사업체조사」, 2020, 2024.01.02, 시도·산업·대표자성별 사업체수('06~)에 따르면 국내 사업체 수는 2010년부터 지속적으로 증가하는 추세를 보이고있다. 전체 사업체 중 소상공인 사업체의 비중은 80% 이상을 차지한다. 하지만, 최근 급격한 최저 임금 인상과 약 3년 동안 이루어진 코로나19 방역 조치에 소상공인들은 사업체 운영에 난항을 겪고 있다. 매년 전국 사업자 수가 증가하지만, 최근 5년간 폐업 사업자 수는 90만 명을 하회하는 것으로 보아 정부의 폐업 대책 마련이 절실해 보인다.

정부가 위탁 운영 중인 소상공인 시장 진흥 공단에서는 폐업 위기 혹은 폐업 직전의 소상공인을 대상으로 희망 리턴 패키지(소상공인시장진흥공단, “희망리턴패키지”, 2023.11.23. “https://www.semas. or.kr/ web/ SUP01/SUP0117/SUP011703.kmdc)를 실시하고 있다. 희망 리턴 패키지는 경영 개

[1]성신여자대학교 수리통계데이터사이언스학부 학사 과정. 20210851@sungshin.ac.kr
[2]성신여자대학교 수리통계데이터사이언스학부 학사 과정. 20210853@sungshin.ac.kr
[3]성신여자대학교 수리통계데이터사이언스학부 학사 과정. 20210900@sungshin.ac.kr
[4]성신여자대학교 수리통계데이터사이언스학부 조교수. dongha0718@sungshin.ac.kr

선을 위한 교육 및 컨설팅 또는, 사업 정리 방향성을 제공하며 철거 비용 지원 및 법률 자문, 채무 조정 서비스 등 다양한 정책을 실행하고 있다. 더불어 폐업 후 재창업을 지원하거나 폐업 사업자에 대한 전직 장려 수당을 지원해 주기도 한다. 하지만, 지원 자격을 충족해야 하고, 경영 개선 컨설팅 지원 인력에는 한계가 있어 직접적인 경영 진단은 5,000건도 채 진행되지 않았다.

성공적인 사업체 운영을 위해 상권 분석은 반드시 선행되어야 한다. 현재 대부분은 상권 분석 서비스들은 단순한 통계 수치만 제공할 뿐 개개인에 맞춘 지표 및 리스크 제공은 불가능하다. 따라서 행정동별 상권 발달 지수를 분석하고 그에 따른 폐업 리스크 계산을 통한 개인별 맞춤 서비스 제공의 필요성이 증대되는 상황이다.

본 연구에서는 서울시 행정동별 분기에 따른 상권 활성화 지수를 예측하고, 그에 따른 행정동별·업종별 폐업 리스크를 예측한다. 머신러닝 및 딥러닝 알고리즘으로 상권 활성화 지수를 예측하여 상권 분석 및 지표 활용 방향을 제시할 것이다. 앞서 계산된 상권 활성화 지수를 활용하여 행정동별 또는 업종별 폐업 리스크를 예측하는 모형을 구축한다. 이 모형을 통해 소상공인에게 개업 방향성을 제공하고, 업종별 위치 추천·행정동별 개업 업종 추천을 이끌어내어 폐업 가능성을 줄이는 것을 근본적인 목표로 삼는다.

본 연구를 통해 상권을 구성하는 다양한 데이터를 종합적으로 다루면서 상권의 형성과 불균형한 폐업 행태를 분석할 수 있다. 상권 활성화 지수 예측치를 통해 미래에 주목받을 명소를 발견 가능하며, 부동산 투자와 지역 경제에 기여할 수 있는 중요한 지표로 간주할 수 있다. 나아가, 계산 가능한 리스크는 소상공인에게는 사업 가능성에 대한 지표로 작용하며, 은행권에서는 사업자의 대출 심사 도구로 유용하게 활용될 수 있다.

1.2 연구 범위 및 방법

본 연구는 공간적 범위를 국내로 한정하여 서울특별시의 폐업 리스크 예측을 목적으로 한다. 국내 사업체 종사자 중 25%가 서울특별시에서 근무하고 있으며, 상권의 인지도가 가장 높고, 다양한 종류의 상권이 분포되어 있다. 따라서, 서울특별시의 상권 분석은 국내 상권 현황을 대표하고, 향후 국내 상권의 동향을 파악하는 데 중요한 지표가 될 수 있다.

예측에는 서울 신용 보증 재단에서 제공한 2019년부터 2023년까지의 약 4년간의 개·폐업 현황 데이터를 기반으로 하며 상권에 영향을 미칠 1) 생활 인구, 2) 추정 매출, 3) 상권 활성화 지수, 4) 임대 시세, 5) 직장 인구 데이터 등을 함께 분석에 활용한다. 본 연구는 인구 및 추정 매출 정보를 활용하여 상권 활성화 지수 내부의 세부 지표(상권 활성화 지수, 매출 지수, 가맹점 지수, 인프라 지수, 인구 지수, 금융 지수) 등을 예측하여 서울시의 상권 분석을 선행한다. 이 예측을 기반으로 개·폐업 현황 데이터에 대한 1년 내의 폐업 여부를 예측하고, 분석 결과를 도축하는 순서로 진행된다.

상권 활성화 지수 예측 모델로는 SVM, XG Boost, Linear Regression, Random Forest, LSTM 등을 활용한다. 각 모델을 상황에 맞추어 최적화시킨 결과 가장 높은 정확도를 모델의 예측 결과를 활용하여 2차 분석에 사용하고자 한다. 2차 분석에서는 XG Boost, Random Forest, Logistic Regression 등의 머신러닝 모델을 사용한다. 최적화 후 가장 예측 정확도가 높은 모델을 활용하여 추천 모델로 발전시키고자 한다.

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 79-79

Explainable Automatic Paper Classification System Using Topic Modeling and SHAP*

Nakyung Shin[1], *Yulhee Lee*[2], *Heesung Moon*[3], *Joonhui Kim*[4], *Hohyun Jung*[5]

Abstract

The advancement of computer and information technology has led to the prolific publication of numerous papers nowadays. As a result, it has become increasingly challenging for individuals to search for and categorize research papers on specific topics. Research abounds on paper classification that plays a crucial role in knowledge management, interdisciplinary collaboration, and paper recommendation systems. Traditional embedding techniques for paper classification have faced limitations in the interpretation of results. In response to this challenge, we propose an explainable paper classification system using topic modeling and XAI(eXplainable Artificial Intelligence) method. The system employs LSA(Latent Sentiment Analysis) to extract the topic assignments from abstracts of the papers. The extracted assignments are then utilized as embedding values for applying a MLP(Multi-Layer Perceptron) classifier. Additionally, we aim to identify which topic significantly influences prediction outcomes using SHAP(SHapley Additive exPlanations), one of the XAI methods. We apply the proposed system to the Web of Science data, consisting of papers in the nanomaterial field. The system outperforms other baseline methods in the perspective of metrics such as accuracy, F1 score, and AUC. Also, we demonstrate the explainability of the system via a case study with noteworthy interpretations.

Keywords: Paper classification, Topic modeling, Shapley value, XAI

---

*This research is supported in part by a National Research Foundation of Korea (NRF) grant funded by the Korean government (no. 2021R1G1A109410312).

[1](02844) Bomun-ro, 34 Da-gil 2, Seongbuk-gu, Seoul, Korea, Master's Student, Department of Statistics, Sungshin Women's University, iamnk3247@gmail.com

[2](51508) 797 Changwon-daero, Seongsan-gu, Changwon, Gyeongsangnam-do, Korea, Senior Researcher, National Nanotechnology Policy Center, Korea Institute of Materials Science, lyh89@kims.re.kr

[3](51508) 797 Changwon-daero, Seongsan-gu, Changwon, Gyeongsangnam-do, Korea, Senior Researcher, National Nanotechnology Policy Center, Korea Institute of Materials Science, hsmoon@kims.re.kr

[4](27740) 1339, Wonjung-ro, Maengdong-myeon, Eumseong-gun, Chungcheongbuk-do, Korea, Associate Research Fellow, Center for R&D Performance Diffusion, Korea Institue of S&T Evaluation and Planning, tedpoll@kistep.re.kr

[5](02844) Bomun-ro, 34 Da-gil 2, Seongbuk-gu, Seoul, Korea, Assistant Professor, Department of Mathematics, Statistics and Data Science, Sungshin Women's University, hhjung@sungshin.ac.kr

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 81-81

A Time-Varying Worker Ability Time-Series Model for Heterogeneous Distributed Computing Systems*

Daejin Kim[1], *Suji Lee*[2], *Hohyun Jung*[3]

Abstract

We propose a novel time-series model to efficiently distribute the task of distributed matrix-vector multiplication, considering workers with varying working rates that may change over time. The model enables us to assign the most suitable load based on workers' abilities to reduce the expected waiting time. The model involves modeling workers' abilities as temporal latent variables, introducing a working rate that follows a log-normal distribution based on the latent variables. To infer model parameters and workers' abilities, we present an algorithm that combines the EM algorithm with a particle method, utilizing Sequential Monte Carlo(SMC) and Forward Filtering Backward Sampling (FFBSa). Monte Carlo simulations validate the effectiveness of the proposed algorithm in estimating workers' abilities and model parameters. Furthermore, numerical simulations demonstrate that load allocation based on estimated workers' abilities reduces expected execution time by 54%, compared to the conventional method.

Keywords : distributed computing, EM algorithm, Sequential Monte Carlo, particle filtering.

---

*This research is supported in part by a National Research Foundation of Korea (NRF) grant funded by the Korean government (no. 2021R1G1A109410312).

[1]Samsung Electronics, Samsung-ro 129, Yeongtong-gu, Suwon, Gyeonggi-do, 16677, Korea.
E-mail: deezay0307@gmail.com

[2]Master's Student, Department of Statistics, Sungshin Women's University, 34 Da-gil, Bomun-ro, Seongbuk-gu, Seoul, 02844, Korea. E-mail: isuji095@gmail.com

[3](Corresponding author) Associate Professor, School of Mathematics, Statistics and Data Science, Sungshin Women's University, 34 Da-gil, Bomun-ro, Seongbuk-gu, Seoul, 02844, Korea.
E-mail: hhjung@sungshin.ac.kr

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 83-83

Keyword Analysis of Twitter data on New Digital Technology through Co-occurrence network and ERGM*

Yoonjin Lee[1], *Hohyun Jung*[2]

Abstract

Digital technologies, such as artificial intelligence, big data, the internet of things, autonomous driving, and blockchain, play a crucial role in society of the Fourth Industrial Revolution. We conducted an analysis of people's demands by examining Korean and English keywords related to digital technologies mentioned on Twitter. Co-occurrence networks are generated to comprehend the interconnection structure of keywords. We observe the emergence of topics related to metaverse, which gained popularity in virtual environments as Covid-19 spreads. We apply Exponential Random Graph Models(ERGM) by utilizing press and research indices as node attributes in Korean keyword co-occurrence networks. The research index positively influences keyword connections, while the press index has a negative effect, suggesting that the media may not accurately capture public trends.

Keywords : Co-occurrence network, ERGM, Social network analysis.

---

*This work is supported by National Research Foundation of Korea (2021R1G1A109410313).

[1]02844 Bomun-ro, 34 Da-gil 2, Seongbuk-gu, Seoul, Korea, Master's Student, Department of Statistics, Sungshin Women's University. E-mail: tiffanyyjlee@gmail.com

[2](corresponding author) 02844 Bomun-ro, 34 Da-gil 2, Seongbuk-gu, Seoul, Korea, Assistant Professor, Department of Mathematics, Statistics and Data Science, Sungshin Women's University. E-mail: hhjung@sungshin.ac.kr

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 85-82

딥러닝 기반의 시계열 분석

박태영[1], **이승한**[2]

요 약

다중 시계열은 시간에 따라 관찰된 다양한 변수들의 데이터를 의미한다. 예를 들어, 주식시장에서 여러 주식의 가격 변동을 연구하고 예측하는 것처럼, 이러한 다중 시계열의 분석은 복잡한 시스템 내에서 여러 요인들이 어떻게 상호작용하는지 이해하는 데 매우 유용하다. 본 논문에서는 이러한 다중 시계열를 예측 분석하기 위해, 데이터 간의 복잡한 관계를 모델링하는 데 효과적인 그래프 인공신경망(Graph Neural Networks, GNN)과 레이블이 없는 데이터로부터 모델이 자율적으로 학습할 수 있게 하는 자기지도학습(Self-Supervised Learning, SSL) 방법을 소개하고, 이를 기반으로 시계열 예측의 정확도와 효율성을 향상시키는 새로운 딥러닝 기반의 분석 방법을 제안한다.

주요용어 : 시계열 예측, 다변량 시계열, 그래프 인공신경망, 자기지도학습.

[1](교신저자) 03722 서울 서대문구 연세로 50, 연세대학교 응용통계학과 교수. E-mail: tpark@yonsei.ac.kr
[2]03722 서울 서대문구 연세로 50, 연세대학교 통계데이터사이언스학과 박사과정.
E-mail: seunghan9612@yonsei.ac.kr

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 87-87

안전한 포트폴리오 최적화 방법

변준영[1]

요 약

최근 유럽 일반 데이터 보호 규정(European General Data Protection Regulation, GDPR) 등과 같은 엄격한 개인정보 규제가 강화됨에 따라, 금융산업에서 역시 개인정보 보호가 시급한 문제로 떠오르고 있다. 로보 어드바이저(Robo-advisor)는 개인화된 투자 전략을 제공하는 대표적인 핀테크(Fintech) 서비스로, 본 연구에서는 로보 어드바이저가 고객의 개인정보를 보호하면서도 최적의 투자 포트폴리오(Portfolio)를 제공할 수 있도록 동형 암호(Homomorphic Encryption)를 이용하여 개인의 위험 회피(Risk Aversion) 정보를 암호화하는 새로운 프레임워크를 제안한다. 특히 동형 암호 친화적인 제약 최적화(Constrained Optimization) 방법을 고안함으로써, 제안된 프레임워크는 공매도 금지 등 포트폴리오에 제약조건이 있는 경우에도 평균-분산(Mean-variance) 최적 포트폴리오를 산출할 수 있다. 실험 결과를 통한 주요 발견은 다음과 같다. (1) 일반적으로 개인정보 보호에 대한 비용(Cost)은 정확도 손실 또는 연산 비효율성으로 나타난다. 제안된 모델은 이러한 비용을 감안할 때 수용 가능한 정확도 손실 수준에서 최적 포트폴리오를 근사할 수 있다. (2) 선택 가능한 자산의 수와 자산 간 상관관계 정도가 정확도 손실에 영향을 미치는 것을 발견하였다. 본 연구의 결과는 특히 금융산업에서 정책 입안자(Policy), 규제자(Regulator)에게 개인정보 강화를 지시할 수 있는 강력한 유인을 제공한다. 향후 본 연구의 결과를 더욱 다양하고 현실적인 상황으로 확장하는 연구를 계획하고 있다.

주요용어 : 동형 암호, 로보 어드바이저, 평균-분산 포트폴리오, 핀테크.

[1]06974 서울특별시 동작구 흑석로 84, 중앙대학교 응용통계학과 교수. E-mail: junyoungb@cau.ac.kr

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 89-89

Ordered probit Bayesian additive regression trees for ordinal data

Jaeyong Lee[1], *Beom Seuk Hwang*[2]

Abstract

Bayesian additive regression trees (BART) is a nonparametric model that is known for its flexibility and strong statistical foundation. To address a robust and flexible approach to analyze ordinal data, we extend BART into an ordered probit regression framework (OPBART). Further, we propose a semiparametric setting for OPBART (semi-OPBART) to model covariates of interest parametrically and confounding variables nonparametrically. We also provide Gibbs sampling procedures to implement the proposed models. In both simulations and real data studies, the proposed models demonstrate superior performance over other competing ordinal models. We also highlight enhanced interpretability of semi-OPBART in terms of inference through marginal effects.

[1]Department of Applied Statistics, Chung-Ang University
[2]Department of Applied Statistics, Chung-Ang University

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 91-91

Censored Experiment for Average Run Length of General Control Chart

Jahan Lim[1], *Sungim Lee*[2]

Abstract

The average run length (ARL), the average number of in-control signals before an out-of-control signal occurs, is a critical measure of the performance of a control chart. Estimating the ARL for various control charts has been studied for a long time, with the Markov chain-based approximation, the integration method, and the Monte Carlo method being the most popular, especially for time-dependent charting statistics. Although computationally expensive, the Monte Carlo method is generic and can be applied to all control charts. In the Monte Carlo approach, it is recommended to choose the maximum run length in advance and to discard the iterations that do not produce an out-of-control signal to accommodate some unusually lengthy runs. However, the discarded runs are also informative for estimating the run length. In this paper, we propose a new Monte Carlo approximation that, unlike existing Monte Carlo methods, retains the iterations that fail to be out of control limits, treats them as Type I censored observations, and uses them in the estimation. This proposal allows us to use a moderate size of the maximum run length, which does not need to be large and to estimate ARL accurately enough with signicantly fewer iterations. Our Monte Carlo method relies on the memoryless expectation of the run length distribution to have a simple and ecient ARL estimator. This assumption is necessary for mean estimation for Type I censored data, and we nd that the exiting ARL approximations for three of the most common control charts, Shewhart, exponentially weighted moving average (EWMA), and cumulative sum (CUSUM) charts, have or also often use the same property. We numerically demonstrate the computational and statistical eciency of our new Monte Carlo method for the four most popular control charts: the mean, median, EWMA, and CUSUM charts.

Keywords: Average run length; control charts; Markov Chain approximation; memoryless property; Monte Carlo method; Type I censoring.

[1]Department of Statistics, Seoul National University, Seoul, Korea.
[2]Department of Statistics, Seoul National University, Seoul, Korea. E-mail: silee@dankook.ac.kr

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 93-96

Evaluation and dynamic prediction of Joint Models for Longitudinal and Interval-Censored Data*

정은정[1], **김양진**[2]

요 약

결합모형(joint model)의 목표는 종단 자료(longitudinal data)와 생존 자료 사이의 관계를 통해 두 확률 변수에 대한 추론 과정을 구하는 것이다. 가장 기본적인 방법은 two-stage 모형으로 종단자료와 생존자료 간의 연관성을 고려하지 않고 각각 개별적인 모형을 사용하는 반면에, 결합모형은 종단자료와 생존자료의 연관성을 고려하여 함께 모델링한다. 이로써 결합모형은 two-stage 모형보다 마커와 사건 발생 위험 사이의 관계를 정확히 추정할 수 있다는 장점이 있다(Wulfsohn and Tsiatis, 1997). 최근에는 이러한 추론 문제 외에도 생존 분석에서 관심있는 사건 발생을 예측하기 위한 시간 가변 공변량의 마커(marker)로써 예측력을 평가하기 위해 결합모형을 활용한 다양한 방법론이 제시되었다 (Rizopoulos, 2011, 2017). 본 논문의 목적은 정확한 사건 발생 시간대신 구간 형태로 관측된 구간 중도절단 자료하에서 경시적 마커와의 연관성을 고려한 결합 회귀모형을 구한 후 이를 이용한 잔여 생존 확률을 추정한다. 또한 이를 마커로 하는 다이나믹 ROC에 기인한 AUC를 통해 여러 가지 마커 형태의 예측력을 비교한다. 본 논문에서 제안된 방법론은 프랑스의 치매 자료에 적용함으로써 노인들의 인지점수를 통해 치매 발생률을 유도한다.

주요용어 : 결합 모형, 동적 예측, 시간 가변 AUC, 구간 중도 절단 자료, 경시적 자료분석, 치매 예측 모형

1. 서론

종단자료 분석을 위해 선형혼합모형(linear mixed effect model)은 가장 자주 적용되는 모형이다.

$$\begin{aligned} y_i(t_{ij}) &= \mu_i(t_{ij}) + \epsilon_{ij} \\ \mu_i(t_{ij}) &= \beta^\top X_{ij} + b_i^\top Z_{ij} \\ b_i \sim N(0, \Sigma_b),\ &\epsilon_{ij} \sim N(0, \sigma_e^2) \end{aligned}$$

$y_i(t_{ij})$는 i번째 개체의 j번째 종단 마커(longitudinal biomarker)이다. 시점 $t_{ij}\,(i=1,...,n,\ j=1,...,n_i)$마다 마커가 측정되어 개체당 n_i개의 마커가 기록되는데, 측정 시 오차 ϵ_{ij}가 발생한다. X_i과 Z_i는 각각 고정 효과 β와 랜덤효과 b_i의 설계 행렬(design matrix)이다. 경시적

*This research is supported by Korean research foundation (NRF-2020R1A2C1A01100755).
[1]08826, 서울특별시 관악구 관악로1 서울대학교 연구공원, 백신연구소(International Vaccine Institute; IVI)
[2]04310 서울특별시 용산구 청파로47길 100(청파동2가) 숙명여자대학교 통계학과. E-mail: yjin@sookmyung.ac.kr

자료와 사건 발생 위험률 간의 관계를 측정하기 위해서 평균값 $\mu_i(t_{ij})$가 생존자료의 위험함수(hazard function)에 공변량으로 적용되며 Cox's 비례 위험모형이 가정 된다.

$$\lambda_i(t|w,\mu_i) = \lambda_0(t)\exp\{\gamma^\top W_i + \eta H(\mu_i(t))\}$$

$\lambda_0(t)$는 시점t 에서의 기저 위험함수(baseline hazard function)이며 W_i는 시간 불변 공변량으로 회귀계수 γ 를 통해 위험률과의 연관관계를 추정하게 된다. η는 종단 공변량(longitudinal covariates)의 효과를 보여주는 회귀계수로 $H(\mu_i(t))$에 대해 다음의 3가지 형태의 $H(\mu_i(t))$를 제안했다(Rizopoulos et al., 2017). (i) $H(\mu_i(t)) = \mu_i(t)$, (ii) 평균과 기울기 $H(\mu_i(t)) = (\mu_i(t), \mu_i'(t))$, (iii) 누적 평균 $H(\mu_i(t)) = \int_0^t \mu_i(s)\,ds$. 여기서, $\eta < 0$이면 마커 $H(\mu_i(t))$값이 클수록 사건 발생 위험이 작아짐을 의미하며, $\eta > 0$이면 위험률이 증가함을 의미하고 η가 0에 가까우면 사건 발생 위험에 대한 마커의 효과가 미미하다고 해석할 수 있다. 관심 있는 모수를 $\theta = (\beta, \Sigma_b, \gamma, \eta, \lambda_0, \sigma_e^2)$라 하고 각 개체에서 $\{O_i = (y_i, L_i, R_i, \delta_i, X_i, Z_i), i = 1, \ldots, n\}$가 관측된다고 할 때, 종단자료와 구간 중도 절단 자료에 대한 결합모형의 관측된 가능도 함수(observed data likelihood)는 다음과 같다.

$$L(\theta) = \prod_{i=1}^{n} f(L_i, R_i, y_i|\theta, O_i)$$

$$= \prod_{i=1}^{n} \int f_T(L_i, R_i|\theta, O_i)\, f_Y(y_i|\theta, O_i)\, f_b(b_i|\theta, O_i)\, db_i$$

여기서,

$$f_T(L_i, R_i|\theta, O_i) = \exp\left\{-\int_0^{L_i} \lambda_0(u)\, e^{(\gamma^\top W_i + \eta H(\mu_i(t)))} du\right\} - \exp\left\{-\int_0^{R_i} \lambda_0(u)\, e^{(\gamma^\top W_i + \eta H(\mu_i(t)))} du\right\}$$

는 구간 중도절단 자료의 확률밀도함수이고,

$$f_Y(y_i|\theta, O_i) = \prod_{j=1}^{n_i} \left(\frac{1}{\sqrt{2\pi\sigma_e^2}}\right) \exp\left\{-\frac{1}{2\sigma_e^2}(y_i(t_{ij}) - \mu_i(t_{ij}))^2\right\}$$

는 종단자료의 확률밀도함수이며 다음은 랜덤효과의 확률밀도함수이다.

$$f_b(b_i|\theta, O_i) = \left(\frac{1}{\sqrt{2\pi}}\right)^{\frac{1}{2}} |\Sigma_b|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} b_i^\top \Sigma_b^{-1} b_i\right)$$

여기서 관심 있는 모수 θ의 MLE를 추정하기 위해 랜덤효과에 대한 적분이 필요하다. 그러나 위 적분은 다차원이며 닫힌 형태로 유도할 수 없으므로 Gauss-Hermite 또는 MCMC (Markov Chain Monte Carlo)와 같은 방법이 적용된다.

2. 다이나믹 예측

추정된 결합 모수의 추정량을 이용하여 다음의 조건부 생존함수 $\pi_i(t|s) = P(T_i > t|T_i > s)$를 새로운 마커로 적용한다. Rizopoulos (2011)은 Empirial 추정량과 모수의 MLE를 이용하는 방법과 사

후 분포을 이용하는 베이지안 방법을 이용하여 조건부 생존 함수를 추정하였다. 예를 들어, 전자의 경우는 다음과 같이 표현되며

$$\hat{\pi}_i^{JM}(t|s) = \frac{\hat{S}_i(t|\hat{\theta},\hat{b}_i)}{\hat{S}_i(s|\hat{\theta},\hat{b}_i)}$$

여기서 $\hat{S}_i(s|\hat{\theta},\hat{b}_i) = \exp\left[-\hat{H}_0(s)\exp\left\{\hat{\gamma}^\top W_i + \hat{\eta}H(\mu_i(s))\right\}\right]$가 된다. 위에서 구한 조건부 확률은 두 시점$(s,t), s<t$ 에 따라 값이 변화하는 종단 마커(longitudinal marker) $M_i(t|s)$ 의 예측력(predictive power)을 평가하기 위해 적용된다. 특히, 마커의 판별력(discriminative power)을 구하기 위해 동적(dynamic) AUC를 유도하며 여기서 민감도와 특이도는 다음과 같다(Kim, 2022).

$$TPR(c,s,t) = \Pr\left\{M_i(s) \geq c \mid T_i^* \in (s,t];\theta\right\}, TNR(c,s,t) = \Pr\left\{M_i(s) < c \mid T_i^* > t;\theta\right\}$$

이를 이용하여, $AUC(s,t) = \Pr\left\{M_i(s) > M_j(s) \mid D_i(s,t)=1, D_j(s,t)=0, T_i^* > s, T_j^* > s\right\}$
여기서 시점 $s(s<t)$ 이전까지 사건을 경험하지 않은 두 사람에 대해, 시점t 이전에 사건이 발생한 사람의 s 시점에서의 마커가 시점t에서도 아직 사건을 경험하지 못한 사람의 s 시점에서의 마커보다 작을 확률을 의미한다. 구간 중도 절단 자료에서 $D_i(s,t) = (T_i \in (s,t], \delta = 1)$ 는 명확하지 않다. 정확한 발생 시점 대신 두 관측 대상의 관측 형태에 따라 다음의 경우로 나뉘게 된다.

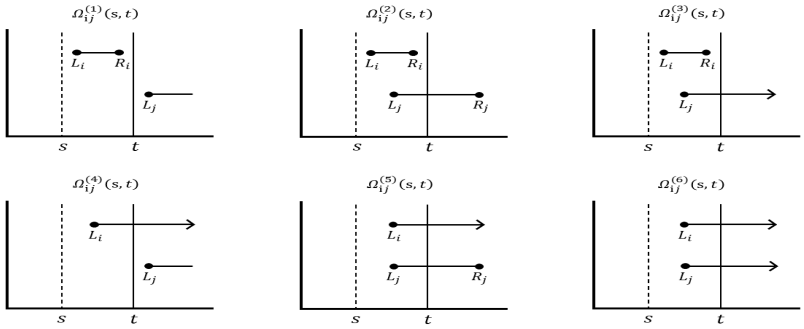


Figure 1. 구간 중도절단 상황에서의 동적 AUC

$$\widehat{AUC}(s,t) = \frac{\sum_{i\neq j}\sum_{m=1}^{6} I\left\{\hat{\pi}_i(t|s) < \hat{\pi}_j(t|s)\right\} \times I\left\{\Omega_{ij}^{(m)}(s,t)\right\} \times \hat{v_{ij}}^{(m)}}{\sum_{i\neq j}\sum_{m=1}^{6} I\left\{\Omega_{ij}^{(m)}(s,t)\right\} \times \hat{v_{ij}}^{(m)}}$$

$\hat{v_{ij}}^{(1)} = 1, \hat{v_{ij}}^{(2)} = r_{jk}, \hat{v_{ij}}^{(3)} = \hat{\pi}_j(t|L_j), \hat{v_{ij}}^{(4)} = \left(1-\hat{\pi}_i(t|L_i)\right), \hat{v_{ij}}^{(5)} = \left(1-\hat{\pi}_j(t|L_j)\right)r_{jk},$
$\hat{v_{ij}}^{(6)} = \left(1-\hat{\pi}_i(t|L_i)\right)\hat{\pi}_j(t|L_j)$ 이다.

3. Paquid 자료 분석

실제 자료 분석에 사용될 자료는 Paquid 자료로, 프랑스 남서부에서 치매에 걸리지 않은 사람

500명을 대상으로 20년간 수집한 전향적 코호트 자료이다. 코호트 연구에서의 관심 사건은 치매이다. 각 개체를 추적조사할 때마다 추적조사 시점 시 나이를 기록하고 치매와 관련된 검사를 수행하여 점수를 기록한다. 검사는 언어 기억력, 인지 기능, 공간 기억력, 신체 의존도와 우울증에 관한 것이다. 자료에는 마지막 추적 조사 시점까지 치매에 걸리지 않은 경우와 치매 진단을 받은 기록이 있는 경우가 있다. 전자는 우중도 절단된 자료이며 후자는 구간 중도 절단된 자료다. 전자의 경우, L_i은 마지막 추적조사 시점이고 R_i은 무한대가 된다. 후자의 경우, 치매 진단을 받지 않은 채 마지막으로 추적조사된 시점이 L_i이 되고 처음으로 치매 진단을 받은 시점이 R_i이 된다. 실제 치매에 걸린 시점은 L_i과 R_i 사이에 위치하며 정확한 발병 시점은 알 수 없다.

본 논문에서의 자료 분석 목적은 크게 두 가지이다. 첫 번째는 종단 마커로써 사용된 언어 기억력 점수 IST (Isaacs Set Test)와 인지 기능점수 MMSE (Mini-Mental State Examination)의 동적 예측 성능을 비교하는 것이고 두 번째는 결합모형(Joint Model; JM), 랜드마킹(Land- marking; LM) 그리고 혼합모형 랜드마킹(Mixed model landmarking; LMmixed)의 동적 예측 성능을 비교하는 것이다.

Table 1. IST와 MMSE에 대한 AUC(s, 84.6) 의 비교

	JM	*LM*	*LMmixed*	*JM*	*LM*	*LMmixed*
	IST			MMSE		
			AUC(70.1, 84.6)			
mean	0.7355	0.3364	0.6818	0.7391	0.4402	0.4305
mean+slope	0.7266	0.3634	0.6817	0.6438	0.4393	0.6435
cumulative	0.7380	0.4211	0.5522	0.7841	0.4914	0.4772
			AUC(80.8, 84.6)			
mean	0.7324	0.5089	0.7006	0.7375	0.4042	0.3856
mean+slope	0.7428	0.5562	0.7005	0.6604	0.4187	0.6826
cumulative	0.7374	0.4568	0.5478	0.7583	0.5829	0.4698
			AUC(84.4, 84.6)			
mean	0.9032	0.3515	0.8011	0.9116	0.3666	0.3118
mean+slope	0.9243	0.2989	0.8047	0.5571	0.3979	0.6964
cumulative	0.9746	0.2247	0.6981	0.8235	0.4366	0.4320

References

Rizopoulos, D. (2011). Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. *Biometrics,* 67, 819 – 829.

Wulfsohn, M.S and Tsiatis, A.A.(1997). A joint model for survival and longitudinal data measured with erorr. *Biometrics,* 53, 330-339.

Rizopoulos, D., Molenberghs, G. and Emmanuel M.E.H. Lesaffre E. (2017). Dynamic predictions with time-dependent covariates in survival analysis using joint modeling and landmarking. *Biometrical journal,* 59, 1261-1276.

Rizopoulos, D. (2012). “JM: An R package for the joitn modeling of longitudinal and time-to-event data”, *Journal of Statistical Software,* 35(9).

Kim,Y-J. (2022). Review for time-dependent ROC analysis under diverse survival models. *Korean journal of applied statistics,* 1, 35-47.

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 97-97

Deep Neural Networks for Semi-parametric Frailty Models*

Hangbin Lee[1], *Il Do Ha*[2], *Youngjo Lee*[3]

Abstract

For prediction of clustered time-to-event data, we propose a new deep neural network-based frailty model (DNN-FM). An advantage of the proposed method is that the joint maximization of the new h-likelihood provides maximum likelihood estimators (MLEs) for fixed parameters and best unbiased predictors (BUPs) for random frailties. Thus, the proposed DNN-FM is trained by using a negative profiled h-likelihood as a loss function, constructed by profiling out the non-parametric baseline hazard. Simulation studies show that the proposed method enhances the prediction performance of the existing methods (e.g. DNN based Cox model) and provides the feature selection using the multi-head attention. A real data analysis shows that the inclusion of subject-specific frailties to the DNN-Cox model helps to improve the risk prediction of the DNN based Cox model.

Keywords : Deep neural network, Frailty model, H-likelihood, Random effect.

*The research was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. RS-2023-00240794)

[1]08826 Dept.. of Statistics, Seoul National University, Seoul, Korea, Postdoctor.
E-mail: hangbin221@gmail.com

[2](Corresponding author) 48513 Dept.. of Statistics & Data Science, Pukyong National University, Busan, Korea, Professor. E-mail: idha1353@pknu.ac.kr

[3]08826 Dept.. of Statistics, Seoul National University, Seoul, Korea, Professor. E-mail: youngjo@snu.ac.kr

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 99-99

Semiparametric accelerated failure time models under unspecified random effect distributions

서병태[1], 하일도[2]

Abstract

Accelerated failure time (AFT) models with random effects have been widely used for analyzing clustered (or correlated) time-to-event data as an alternative to frailty models . In the AFT model, the distribution of the unobserved random effect is conventionally assumed to be parametric, often modeled as a normal distribution. Although it has been known that the model is robust to misspecfied random-effect distribution, in some cases, the impact caused by such misspecification is not negligible. Particularly when our focus extends to quantities associated with random effects, the problem could become worse. In this talk, we propose a semi-parametric maximum likelihood approach in which the random-effect distribution under the AFT models is left completely unspecified. We demonstrate the efficacy of the approach through some numerical studies.

Keywords : Clustered survival data, Frailty model, Nonparametric MLE, Nonparametric mixtures, Semiparametric AFT mdoels.

[1]03063 서울특별시 종로구 성균관로 25-2, 성균관대학교 통계학과 교수. E-mail: seobt@skku.edu
[2]48513 부산광역시 남구 용소로 45, 부경대학교 통계·데이터사이언스학과 교수. E-mail: idha1353@pknu.ac.kr

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 101-101

제로팽창 이변량 음이항 회귀모형에서 산포모수에 대한 가설검정

신지은[1], 장동민[2], 정병철[3]

요 약

본 논문은 제로팽창 이변량 음이항 회귀모형에서 과산포의 존재유무에 대한 가설검정을 다룬다. 현실적으로 카운트 데이터에서는 영(0)의 값이 발생하는 동시에 과산포성을 보이는 경우가 많다. 기존의 이변량 음이항 모형들은 음(-)의 상관을 허용하지 않는다는 한계가 있어 본 논문에서는 음의 상관까지 허용하는 Sarmanov 분포족 기반의 이변량 음이항 모형을 고려한다. 이 분포에서 과산포 유무에 대한 스코어 검정을 유도한 후 모의실험을 통해 유의수준과 검정력을 LR검정과 비교한다. 모의실험 결과, 스코어 검정이 LR검정에 비하여 계산이 간편하다는 장점이 존재하지만 명목유의수준을 다소 과소추정하는 경향을 보인 반면 LR검정은 명목유의수준을 적절히 유지하였다. 마지막으로 호주 건강서베이 자료에 두 검정법을 적용하고 그 결과를 제시한다.

주요용어 : 이변량 데이터, 음이항 회귀모형, 과대산포, 스코어 검정, LR검정

[1]02504 서울시립대학교 동대문구 서울시립대로 163, 서울시립대학교 자연과학대학 통계데이터사이언스학과 석・박사통합과정. E-mail:jieunstat@uos.ac.kr
[2]02504 서울시립대학교 동대문구 서울시립대로 163, 서울시립대학교 자연과학대학 통계데이터사이언스학과 석・박사통합과정. E-mail: jang4530@uos.ac.kr
[3](교신저자) 02504 서울시립대학교 동대문구 서울시립대로 163, 서울시립대학교 자연과학대학 통계데이터사이언스학과 교수. E-mail: bcjung@uos.ac.kr

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 103-103

BiVAE를 활용한 MBTI 기반 OTT 서비스 개인화 추천 시스템

전수영[1]

요 약

현대 사회에서 영화, 드라마, 예능 등 다양한 미디어 콘텐츠를 제공하는 OTT 플랫폼은 사용자의 큰 관심을 끄는 중요한 역할을 하고 있으며, 개인 맞춤형 추천 시스템을 활용하여 사용자 경험을 개선하고 있다. 그러나 현재의 추천 시스템은 사용자의 성향을 충분히 이해하지 못해 정확한 추천에 어려움을 겪고 있다. 본 연구는 이러한 문제를 해결하기 위해 BiVAE(bilateral variational autoencoder)를 활용한 개인화 추천 알고리즘의 사용을 제안하며, 제안한 알고리즘의 성능을 향상시키기 위해 MBTI를 통한 그룹별 적용을 고려한다. BiVAE는 양방향 정보 흐름을 고려하여 데이터를 압축하고 생성하는 확률적인 생성 모델로서, 양방향으로 인코딩과 디코딩을 수행하여 더 풍부한 표현력과 재구성 능력을 제공한다. MBTI를 고려한 방법의 우수성을 보이기 위해 4가지 유형의 데이터를 모델에 적용한 결과, MBTI를 통한 그룹별 BiVAE 알고리즘을 적용했을 때 그렇지 않은 방법보다 OTT 장르 추천 정확도가 우수함을 보였다. 또한, MBTI를 고려했을 때, 기존 추천 시스템과 비교하여 BiVAE를 활용한 개인 추천이 더 높은 정확도를 보였다. 이러한 결과는 사용자의 개인 특성과 성격을 고려한 개인화 추천 시스템이 중요하며, 이를 통해 콘텐츠 제공 업체가 사용자 경험이 향상된 서비스를 개선할 수 있는 새로운 방향을 제시한다.

주요용어 : 미디어 콘텐츠, 성격 기반 개인화 추천 시스템, 개인 특성, MBTI, BiVAE.

[1]30019 세종시 세종로 2511, 고려대학교 응용통계학과 교수. E-mail: scheon@korea.ac.kr

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 105-105

패널자료에서의 결측값 대체방법 활용

이동희[1]

요약

최근 십여 년간 다양한 패널조사가 국내에서 이루어지면서 이와 관련한 연구들이 급증하고 있다. 패널조사를 통해 수집된 자료는 특점 시점에 수집된 횡단면자료와 달리 특정 집단이나 개인의 특성 변화를 시간 흐름에 따라 파악할 수 있도록 하며, 다양한 인과관계에 대한 보다 정확한 관찰 결과를 이끌어 낼 수 있다는 점에서 매우 유용하다. 반면 패널조사는 동일한 대상을 여러 회에 걸쳐 반복적으로 조사하기 때문에 중도탈락으로 인한 결측이 발생하거나 조사문항의 변화로 인해 특정 변수에 대한 지속적인 관찰이 이루어지지 않아 결측이 발생하는 경우가 많다. 본 연구에서는 통계분석 과정에서 발생하는 결측값으로 인한 문제점을 살펴보고, 이를 보완하기 위한 방법들에 대해 제시하고자 한다. 이제까지 통계분야에서 활용해 왔던 결측값 처리 방법 뿐 아니라 최근 제안된 생산적 적대 신경망 모형에 기반한 다중 대체 방법 등에 대해 살펴보고자 한다. 이와 함께 한국미디어패널자료를 이용한 실증분석을 통해 결측값 처리 방법들 간의 차이와 유용성을 비교하여 패널자료 분석에 있어서의 결측값 처리 방법 선택과 활용을 위한 시사점을 제시하고자 한다.

주요용어: 결측값 대체, 다중대체, 단일대체, 생산적 적대 신경망, 선형혼합모형,

[1]16227 경기도 수원시 영통구 광교산로 154-42, 경기대학교 경영학부 교수. E-mail: dhl@kyonggi.ac.kr

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 107-107

자기상관이 존재하는 패널회귀모형에서 회귀계수의 추정에 관한 연구

정병철[1]

요 약

패널회귀모형은 통계학, 의학, 경제, 경영학 등 다양한 학문분야에서 널리 사용되고 있는 분석 모형이다. 본 연구에서는 패널자료에 대하여 개체효과와 자기상관이 동시에 존재하는 패널회귀 모형을 정의하고, 이 모형에서 회귀계수에 대한 ML 추정방법을 제안하였다. 아울러 개체효과 또는 자기상관을 무시하는 경우, 회귀계수의 추정에 어떠한 영향을 미치는지 모의실험을 통하여 알아보았다. 모의실험 결과, 개체효과와 자기상관 등 두 효과가 동시에 존재하는데 만일 둘 중 하나의 효과를 무시하는 경우에도 회귀계수 추정량의 편향(bias)은 0에 가까운 값을 나타내어 어느 한 효과를 무시한다하여도 회귀계수의 추정에는 큰 문제가 없는 것으로 나타났다. 반면 둘 중 어느 한 효과를 무시하는 경우, 회귀계수 추정량의 표준오차를 과소 또는 과대추정하는 결과를 나타내어, 회귀계수에 대한 가설검정에서 명목유의수준을 과소 또는 과대추정하는 결과를 보여주었다. 본 연구에서 고려한 패널회귀모형을 실제 패널자료에 에 적용하여 그 결과를 비교하였다.

주요용어 : 패널회귀모형, 자기상관, 개체효과

[1]02504 서울시 동대무눅 서울시립대로 163(전농동), 서울시립대학교 통계학과 교수. E-mail: bcjung@uos.ac.kr

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 109-109

시뮬레이션 분석을 통한 시스템 수명주기 평가에 관한 연구

엔드하르타 알폰수스 주란토[1], 김종운[2]

요 약

본 연구는 철도 차량 시스템을 대상으로 하여 시뮬레이션 분석을 통한 수명주기 평가에 중점을 둔다. 이를 위해 다양한 측면을 고려하여 철도 차량 시스템의 신뢰성, 가용성, 유지보수성, 안전성, 비용 등을 종합적으로 평가하였다. 연구에서는 특히 철도 차량 시스템의 특성을 반영하기 위해 해당 분야의 실험적 데이터와 운영 경험을 수집하고 분석하였다. 수집된 데이터를 기반으로한 수학적 모델과 시뮬레이션 알고리즘을 개발하여, 철도 차량 시스템의 현실적인 운영 환경을 모방하였다. 수행된 시뮬레이션은 다양한 환경 조건과 고장 시나리오를 고려하여 철도 차량 시스템의 수명주기에 대한 종합적인 평가를 제공하였다. 본 연구는 철도 차량 시스템에 대한 심층적인 이해를 바탕으로, 시뮬레이션 분석을 통한 수명주기 평가가 어떻게 현장에서 유용하게 활용될 수 있는지를 탐구한다. 이를 통해 철도 차량 시스템의 설계, 운영, 및 유지보수 계획에 관한 의사결정에 효과적인 지침을 제공할 것으로 기대된다.

주요용어 : 시뮬레이션, 수명주기 평가, 신뢰성, 가용성, 유지보수성, 비용.

---

*본 연구는 국토교통부/ 국토교통과학기술진흥원의 지원으로 수행되었다. (과제번호 RS-2023-00239464)

[1](교신저자) 13840 경기도 과천시 과천대로7길 33, 디테크타워A동 807호, 네모시스 주식회사, 신뢰성연구팀, 책임연구원. E-mail: alfon@nemosys.kr

[2]13840 경기도 과천시 과천대로7길 33, 디테크타워A동 807호, 네모시스 주식회사, 대표이사.
E-mail: jwkim@nemosys.kr

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 111-111

소프트웨어 신뢰성과 소프트웨어 신뢰성 성장 모형의 연구*

이다혜[1], 장인홍[2], 송광윤[3], 김윤수[4]

요 약

현대 사회에서 소프트웨어 없이는 살아갈 수 없다. 스마트폰과 PC로 업무를 처리할 때, 출퇴근을 위해 교통수단을 이용할 때, 몸이 아파서 찾아간 병원에서 사용하는 의료기기에서도 소프트웨어를 쉽게 찾아볼 수 있다. 실생활에서 밀접하게 사용되는 소프트웨어의 품질과 안전은 어떻게 판단할까? 이 질문에 대한 답을 찾기 위해 소프트웨어 신뢰성(Software reliability) 연구가 수십 년간 진행됐다. 소프트웨어 신뢰성 성장 모형(Software reliability growth model; SRGM)은 가정하는 환경에 따라 고유의 평균값 함수(Mean value function)를 갖는다. 이 평균값 함수는 소프트웨어의 향후 기대 고장 수를 예측하거나 소프트웨어 신뢰성의 척도를 추정할 때 사용되기도 한다. 또한, 평균값 함수를 통해 소프트웨어 출시 시기를 정하는 릴리스 정책(Release policy)을 수립할 수 있고, 순차적 확률비 검정(Sequential probability ratio test; SPRT)을 응용하여 테스트 데이터가 수집되는 시점마다 소프트웨어의 신뢰성을 검정하는 데 사용될 수도 있다. 본 연구에서는 소프트웨어의 오류가 초래한 인명 및 경제적 피해를 끼친 사례를 살펴보고, 소프트웨어의 신뢰성 성장 모형과 이를 기반으로 하는 다양한 소프트웨어 신뢰성 연구를 소개하고자 한다.

주요용어 : 소프트웨어 신뢰성, 소프트웨어 신뢰성 성장 모형, 릴리스 정책, 순차적 확률비 검정.

---

*이 논문은 2023학년도 조선대학교 학술연구비 및 한국연구재단의 지원을 받아 진행된 논문입니다 (NRF-2021R1F1A1048592, NRF-2021R1A6A3A01086716, NRF-2021R1I1A1A01059842).

[1]61452 광주광역시 동구 조선대3길 30, 조선대학교, 컴퓨터통계학과, 박사후연구원.
E-mail: is_hye@chosun.ac.kr

[2](교신저자) 61452 광주광역시 동구 조선대3길 30, 조선대학교, 컴퓨터통계학과, 교수.
E-mail: ihchang@chosun.ac.kr

[3]61452 광주광역시 동구 조선대3길 30, 조선대학교, 컴퓨터통계학과, 조교수.
E-mail: csssig@chosun.ac.kr

[4]61452 광주광역시 동구 조선대3길 30, 조선대학교, 전산통계학과, 박사수료.
E-mail: imk92315@naver.com

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 113-113

무고장 신뢰성 입증 시험방법을 활용한 신뢰성 개선수준 추정방법에 관한 연구*

김효중[1], 박신아[2], 김성준[3]

요 약

오늘날 격화되는 글로벌 경쟁으로 인한 지속적인 기술의 발전과 함께 짧은 개발주기로 인해 높은 신뢰성을 가진 제품의 수명을 측정하는 것이 점차 어려워지고 있다. 특히, 시장에 판매 중인 제품으로부터 확인된 필드 신뢰성 이슈에 대응하기 위해 마련된 개선안을 검증하는데는 매우 촉박한 시간과 자원이 주어지는 경우가 빈번하다. 이로 인해, 신뢰성의 개선여부와 더불어 개선수준을 파악하는 것이 어려운 의사결정의 불확실성에 노출될 수 밖에 없는 실정이다. 이로 인해 과잉설계로 인한 원가상승을 감내 하거나, 개선효과가 없는 대안을 적용하여 품질비용의 절감효과를 누리지 못하는 문제가 발생할 수 있다. 본 연구에서는 개선사양과 동일하거나 유사할 것으로 판단할 수 있는 개선전 사양의 가속수명모형이 주어져 있을 때, 개선사양에 대한 개선여부를 검증하고 나아가 개선수준을 추정할 수 있는 방법을 제안한다. 제안된 방법은 대표적인 원샷 시스템 중 하나인 착화기의 사례에 적용하여 타당성을 검토하였다.

주요용어 : 무고상 신뢰성 입증 시험, 가속열화시험, 가속수명모델.

[1]61452 광주광역시 동구 필문대로309, 조선대학교 산업공학과 박사과정. E-mail: gywnd0107@chosun.kr
[2]61452 광주광역시 동구 필문대로309, 조선대학교 산업공학과 석사과정. E-mail: sina@chosun.kr
[3](교신저자) 61452 광주광역시 동구 필문대로309, 조선대학교 산업공학과 조교수.
E-mail: seongjoon.kim@chosun.ac.kr

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 115-115

한국 재벌 기업집단 지배가족의 경영참여와 내부자본시장을 활용한 과잉투자: 지배가족 유형에 따른 조절효과 검증

문승진[1], 김병곤[2]

요 약

본 연구에서는 한국의 재벌기업집단을 대상으로 지배가족의 경영참여가 내부자본시장을 활용한 과잉투자 문제에 미치는 영향을 분석하였다. 또한 지배가족의 유형에 따라 이러한 영향관계가 달라지는지 확인하였다. 분석기간은 공정거래위원회의 대규모기업집단 지정 기준이 변경된 2002년부터 2022년까지이다. 표본기업은 공정거래위원회에서 발표하는 대규모 기업집단 중 총수가 존재하는 기업집단에 속하는 총 3,902개(기업-연도)의 상장기업이다. 내부자본시장의 가용성을 확인하기 위해 사용된 타계열사현금흐름변수를 측정할 때는 기업집단에 속한 외감법인 총 16,601개(기업-연도)의 데이터도 포함하였다. 분석방법으로는 패널토빗회귀분석법을 사용하였다. 실증분석 결과를 요약하면 다음과 같다. 한국 재벌기업집단에서 내부자본시장이 존재하는 경우 기업의 과잉투자 유인이 증가한다는 것을 알 수 있었다. 그렇지만 기업집단의 총수가 직접 경영에 참여하는 경우에는 내부자본시장을 활용한 과잉투자문제가 감소된다는 것을 확인하였다. 기업집단 내 특정기업에서 과잉투자와 같은 비효율적인 의사결정이 이루어지는 경우 동일 기업집단에 속한 다른 계열사에게도 부정적인 영향을 미칠 수 있으므로 총수는 기업집단 전체의 성과를 고려하여 자신이 직접 경영에 참여하고 있는 기업에 대해 내부자본시장을 활용한 과잉투자문제가 억제되도록 영향력을 행사하는 것으로 이해할 수 있었다. 반면 총수 외의 지배가족이 경영에 참여하는 경우에는 지배가족이 내부자본시장을 활용한 과잉투자문제를 조절한다는 증거는 발견할 수 없었다. 총수 외의 지배가족의 경우 기업집단 내 계열사들이 연관되어 형성되는 내부자본시장을 활용하는 의사결정을 하기 에는 총수와 달리 그 영향력이 제한되고, 기업집단 전체의 성과보다는 자신이 직접 경영에 참여하는 기업의 성과에 초점을 맞추는 경향이 있기 때문으로 이해되었다.

주요용어 : 재벌기업집단, 내부자본시장, 과잉투자, 지배가족, 패널토빗회귀분석

[1]51140 경남 창원시 의창구 창원대학로 20, 창원대학교 대학원 경영학과 박사과정.
E-mail: msj9456@changwon.ac.kr
[2](교신저자) 51140 경남 창원시 의창구 창원대학로 20, 창원대학교 경영대학 경영학과 교수.
E-mail: bgkim@changwon.ac.kr

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 117-117

매출채권 팩토링 이용 기업대상 거래적정성 평가 사례

이연경[1], 김종윤[2]

요 약

매출채권 팩토링 제도는 금융기관들이 기업으로부터 상업어음이나 외상매출증서 등 매출채권을 매입하고 이를 바탕으로 자금을 빌려주는 제도이다. 즉, 판매기업이 구매기업에 제공한 물품 또는 용역의 대가로 취득한 매출채권을 팩터(자금을 빌려주는 금융기관)에 양도하는 대신 팩터로 부터 할인 적용된 자금을 공급 받음으로써, 판매기업이 현금 유동성을 조기 확보할 수 있도록 지원하는 제도이다. 이러한 팩토링은 중소기업의 자금난 해소와 경영안정에 도움이 주고자 최근 상환청구권 없는 팩토링 사업을 공공기관이 팩터로 역할을 수행하기 시작하였으며, 상대적으로 신용도가 높지 않은 중소기업들이 팩토링 제도를 이용하고자 하는 수요가 증가함에 따라, 과거 모뉴엘 사태와 같이 허위 매출채권 발행을 통해 팩토링을 악용하는 사례를 막기 위한 노력이 중요시 되고 있다. 본 발표에서는 팩토링을 이용하는 신청기업(판매기업)과 상환의무가 있는 구매기업간 이상 거래 등을 판단하기 위한 거래적정성 평가 사례를 소개하고자 한다.

주요용어 : 매출채권, 팩토링, 모뉴엘, 거래적정성.

[1]NICE평가정보 정보사업본부 기업공공사업실. E-mail: yklee@nice.co.kr
[2]NICE평가정보 기업부문 부문장.

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 119-119

한국 재벌 기업집단의 내부자본시장과 자본조달순위이론

정민규[1] 문승진[2] 김병곤[3]

요 약

본 연구는 재벌 기업집단 내에 존재하는 내부자본시장이 기업집단 소속 기업의 자본조달 의사결정에 미치는 영향을 분석하였다. 이를 위해 기업집단 내에서 창출되는 현금흐름이 내부금융으로 활용되고 있는지를 분석하고, 내부금융으로 활용된다면 내부자금의 우선조달로 인해 부채에 의한 자금조달이 조절되는지 분석하였다. 분석기간은 2004년부터 2021년까지 재벌 기업집단에 속하고 한국거래소의 유가증권시장과 코스닥시장에 상장된 기업 총 2,352개(기업-연도)를 사용하였다. 재벌 기업집단 내에서 내부금융의 가용성을 나타내는 현금흐름은 기업집단 소속 상장사 외에 비상장 외감법인 총 15,235개(기업-연도)를 포함하여 분석하였다. 분석을 위해 횡단면 자료를 시간적으로 연결한 불균형 패널자료를 형성하고 패널회귀분석법을 사용하여 분석하였다. 실증분석 결과를 요약하면 다음과 같다. 첫째, 재벌 기업집단 내에서 창출되는 현금흐름과 레버리지의 영향관계를 확인한 결과에서는 기업집단에 속한 기업에서 자사가 창출한 현금흐름과 타계열사현금흐름이 증가하면 부채에 의한 자금조달이 감소하는 것으로 나타났다. 이러한 결과는 자본조달순위이론의 측면에서 보면 외부금융 보다 내부금융을 선호하고 재벌 기업집단 내에 존재하는 내부자본시장이 기업의 자본조달 의사결정에 영향을 미치는 것으로 이해할 수 있다. 둘째, Shyam-Sunder and Myers(1999)가 제안한 부채발행과 자금부족의 영향관계를 확인한 결과에서는 기업은 필요자금을 순부채발행을 통해 직접 조달하는 것으로 나타났다. 이러한 결과는 외부금융에 의한 자금조달에서 주식발행보다 부채발행을 선호하는 결과로 해석할 수 있다. 셋째, 부채발행과 내부자본시장의 영향관계를 확인한 결과에서는 내부자본시장은 자금부족의 조절효과를 통해 부채발행을 감소시키는 것으로 나타났다. 이러한 결과는 기업집단 내에 내부자본시장이 존재하는 경우 기업은 자금부족이 발생하면 부채발행에 의한 자금조달보다 우선하여 내부자본시장을 통해 자금을 조달하는 것으로 해석할 수 있다. 분석결과를 종합해 보면, 재벌 기업집단 내에서 가용할 수 있는 현금흐름은 내부자본시장을 형성하고, 내부자본시장은 기업집단 소속기업의 자본조달 의사결정에 영향을 미치는 것으로 이해할 수 있다. 즉 재벌 기업집단 소속 기업은 자본조달 의사결정에서 외부금융보다 내부금융을 선호하고 내부금융으로는 기업의 유보이익 외에 기업집단 내에 형성된 내부자본시장이 활용되고 있는 것으로 이해되었다. 이러한 결과는 한국 재벌 기업집단 내에 내부자본시장이 존재하고 자본조달비용에 기인한 자본조달순위이론(pecking order theory)에 의해 전반적인 자본조달 의사결정이 이루어지는 것으로 해석할 수 있다.

주요용어 : 재벌 기업집단, 내부자본시장, 자본조달순위이론, 자본구조, 패널자료회귀분석

[1]51140 경남 창원시 의창구 창원대학로 20 창원대학교 경영학과 강사. E-mail: 0747@daum.net
[2]51140 경남 창원시 의창구 창원대학로 20 창원대학교 경영학과 박사과정. E-mail: msj9456@changwon.ac.kr
[3](교신저자) 51140 경남 창원시 의창구 창원대학로 20 창원대학교 경영학과 교수. E-mail: bgkim@changwon.ac.kr

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 121-124

한국의 녹색채권 프리미엄은 존재하는가?*

박유현[1], 송철종[2]

요 약

이 연구는 한국의 녹색채권시장을 중심으로 녹색채권 프리미엄이 존재하는지 분석하고자 한다. 녹색채권 프리미엄을 측정하기 위해 기존 연구에서 가장 많이 사용되는 동일 발행사의 녹색채권과 이에 대응되는 일반채권을 매칭시켜 두 채권 간 평균 발행금리 차이를 구하였다. 분석 대상 채권은 2018년 5월부터 2023년 7월까지 발행된 녹색채권과 대응되는 일반채권이다. 또한 2021년 12월에 한국형 녹색분류체계가 발표되었고 2022년부터 금리 상승이 일어난 점을 고려하여 분석 기간을 2018년 5월부터 2021년 12월까지, 2022년 1월부터 2023년 7월까지로 구분하였다. 분석 결과, 전체 기간과 2022년 이후 기간에서 매칭 결과에 따른 발행금리 차이는 녹색채권 디스카운팅의 존재 가능성을 드러내었다. 그러나 채권의 개별 특성과 거시경제 불확실성을 모두 고려한 모형에서는 오히려 녹색채권 프리미엄이 존재하는 것으로 나타났다. 특히, 한국형 녹색분류체계와 같이 녹색채권에 대한 규제가 엄격해진 상황에서 거시경제상황을 고려하는 경우 녹색채권 프리미엄이 존재하는 것을 발견하였다.

주요용어 : 녹색채권, 녹색채권 프리미엄, 녹색채권 디스카운팅, 한국형 녹색분류체계

1. 서론

이 연구는 한국의 회사채와 금융채를 대상으로 녹색채권 시장에 녹색채권 프리미엄이 존재하는지를 분석하고자 한다. 녹색채권 프리미엄이란 녹색채권과 이에 매칭되는 일반채권의 발행 금리를 비교하여 녹색채권의 발행금리가 더 낮은 경우를 의미한다. 즉, 사회적 책임이 요구되는 환경 관련 프로젝트에 투자하는 녹색채권의 경우 일반채권보다 발행금리를 낮춤으로써 발행사의 자본조달비용을 낮출 수 있게 된다. 여기에는 발행사가 환경 관련 프로젝트에 투자한다는 명목으로 녹색채권을 발행하였으나 실제로 조달된 자본을 환경과 무관한 프로젝트에 투자하는 그린워싱의 문제도 존재한다. 따라서 기존의 실증연구에서는 녹색채권 프리미엄의 존재에 관해 일치된 결론을 얻지 못하고 있다. Baker et al. (2018), Gianfrate and Peri (2019), Nanayakkara and Colombage(2019), Zerbib (2019), Huang et al. (2023) 등은 녹색채권 프리미엄의 존재를 지지하는 결과를 보인 반면에 Larcker and Watts (2020)나 Tang and Zhang (2020), Wu (2022) 등은 녹색채권 프리미엄이 존재한다는 증거

*이 논문은 2022년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구(NRF-2022S1A5C2A03093594) 이며, 이 원고는 아이디어를 정리한 초안이므로 인용을 허락하지 않음.
[1]31460 충남 아산시 탕정면 선문로 221번길 70 선문대학교 글로벌지속가능발전경제연구소 연구교수.
Email: yhbak208@sunmoon.ac.kr
[2](교신저자) 31460 충남 아산시 탕정면 선문로 221번길 70, 선문대학교 국제경제통상학과 조교수. E-mail: cjsong5525@sunmoon.ac.kr

를 찾지 못하였다. 심지어 Karpf and Mandel (2018)은 녹색채권의 발행금리가 더 높은 녹색채권 디스카운팅의 존재를 주장하였다. 이에 본 연구에서는 2018년 5월부터 2023년 7월까지 금융위원회의 채권기본정보에 구한 녹색채권과 그에 대응되는 일반채권을 매칭하여 녹색채권 프리미엄의 존재를 분석하고자 한다. 특히, 2021년 12월에 한국형 녹색분류체계가 발표된 점, 2022년부터 그 이전 기간에 비해 이자율이 상승한 점을 반영하여 분석 기간을 2018년 5월부터 2021년 12월, 2022년 1월부터 2023년 7월로 구분하여 녹색채권 프리미엄의 존재를 확인하고자 한다. 또한 2022년부터 이자율이 상승하는 등 대내외 거시경제 불확실성이 커짐에 따라 이를 실증분석에 반영하고자 한다. 따라서 기존 연구에서 많이 사용되는 녹색채권과 일반채권의 매칭 방법을 사용하되 거시경제 상황을 분석에 반영한 점은 기존 연구와의 차별점으로 볼 수 있다.

2. 분석 자료와 매칭 결과

Table 1은 본 연구에서 사용한 한국 녹색채권에 대한 기본 정보를 보여주는데 한국거래소에 상장된 녹색채권의 발행자수, 발행채권수, 발행액을 담고 있다. 녹색채권은 주로 회사채와 금융채로 발행되고 있다. 두 유형의 채권이 전체 녹색채권 발행액의 약 90% 가량을 차지하고 있어 본 연구에서는 이 두 종의 채권을 대상으로 한다.

Table 1. Green bond information

(unit : %, billion won)

	The number of bonds	The number of issuers	Issued amount	Issued amount ratio
Corporate bond	180	58	16,850	0.64
Financial bond	71	18	7,000	0.26
Specific laws bond	20	9	2,420	0.09
Municipal bond	4	3	210	0.01

Note : This table presents green bond information. Each column shows the number of bond and issuer, amount of issue, and issued amount ratio by issuer type in green bond issuance.

본 연구에서 사용한 녹색채권 매칭 방법은 Ehlers and Packer (2017), Zerbib (2019), Wang et al. (2020), Larcker and Watts (2020), Flammer (2021)의 방법론과 유사한데 녹색채권과 발행자가 동일한 일반채권 중에서 잔존만기, 이자지급주기, 신용등급이 동일한 채권을 매칭하였다. Table 2는 녹색채권 매칭 결과를 보여준다.

Table 2. Matching Results

	2018.05-2023.07		2018.05-2021.12	2022.01-2023.07		
	mean	median	mean	mean	median	mean
Green	2.93	3.04	1.94	2.08	4.08	4.01
Non-Green	2.68	2.67	1.99	2.16	3.49	3.20
Difference	0.25**	0.36***	-0.05	-0.08	0.59***	0.81***

Note : This table presents the mean and median interest rate differentials between the green and non-green bonds, along with paired t-tests (mean) and Wilcoxon tests (median) for the statistical significance in differences between the matched sample. Ours match two samples with same issuer, rating, maturity, interest payment cycle, and nearest issue date. Statistical significance at the 1, 5 percent level is indicated by ***, and **.

녹색채권과 이에 매칭된 일반채권의 발행금리의 평균과 중간값에 유의미한 차이가 있는지 t-test와 Wilcoxon test를 하였다. 그 결과, 2022년 1월부터 2023년 7월까지 기간에는 녹색채권 발행금리가 매칭된 일반채권보다 높은 녹색채권 디스카운팅이 나타났다. 다만, 2021년 12월에 한국형 녹색분류체계가 발표되었고 2022년부터 금리가 상승하는 등 대내외 거시경제의 불확실성이 커짐에 따라 이러한 요인들이 녹색채권의 발행금리에 영향을 미칠 수 있다. 따라서 개별 채권의 특성과 거시경제 불확실성을 통제한 분석을 통해 녹색채권 프리미엄의 존재 여부를 확인하고자 한다.

3. 모형 및 분석 결과

채권의 발행금리는 발행사의 자본조달비용과 관련이 있기 때문에 앞서 언급한 대로 개별 채권의 특성과 거시경제 상황을 통제할 필요가 있다. 이를 반영한 분석 모형은 다음과 같다.

$$\begin{aligned} interest\,rate_{it} = r + \beta_1 amount_{it} + \beta_2 maturity_{it} + \beta_3 type_{it} + \beta_4 rating_{it} + \beta_5 RF_t \\ + \beta_6 EER_t + \beta_7 KOSPI_t + \beta_8 VKOSPI_t + \beta_9 BR_t + \beta_{10} move_t + e_{it} \end{aligned}$$

여기서 i는 채권발행사, t는 채권발행시점, $interest\,rate$는 발행금리, r은 상수항이다. $amount$는 채권발행액, $maturity$는 잔존만기, $type$는 회사채 여부, $rating$은 신용등급으로 개별 채권특성을 나타낸다. RF는 무위험수익률, EER은 실효환율, $KOSPI$는 코스피지수, $VKOSPI$는 주식시장 변동성, BR은 회사채 수익률, $move$는 채권변동성지수로 이들은 거시경제상황을 통제하는 변수이다. 여기서 주목하는 변수는 상수항 r인데 이는 채권의 특성과 관계없이 채권발행 시 지불해야 하는 금리로 일종의 기본금리로 볼 수 있다. 따라서 위의 모형을 녹색채권과 매칭된 일반채권에 대해서 회귀분석을 한 후 각각의 상수항 r의 차이를 통해 녹색채권 프리미엄의 존재를 확인할 수 있다. 2022년 1월부터 2023년 7월까지의 분석 결과는 Table 3에 정리하였다.

Table 3에서 눈에 띄는 결과는 다음과 같다. 첫째, 개별 채권의 특성은 발행금리에 유의한 영향을 미치지 못한다. 그래서 녹색채권과 일반채권의 열 (1)의 결과를 보면 기본금리인 상수항만 유의한 값을 보이는데 녹색채권의 상수항이 일반채권의 상수항보다 크다. 이는 Table 2의 매칭결과에서 보듯이 녹색채권 디스카운팅의 존재를 시사한다. 둘째, 거시경제변수를 통제한 상황에서는 녹색채권 프리미엄의 존재 가능성이 나타난다. 녹색채권과 일반채권의 열 (2)와 (3)의 결과는 거시경제변수를 통제한 결과이다. 이 경우에도 기본금리에 해당하는 상수항에 주목해야 하는데 일반채권의 열 (3)의 상수항이 유의하지 않아 열 (2)의 결과에 주목한다. 열 (2)의 결과에서 녹색채권의 기본금리가 일반채권의 기본금리에 비해 약 0.11%가 낮아 녹색채권 프리미엄의 존재를 확인할 수 있다. 아울러 통제변수에 대한 결과는 다음과 같다. 거시경제변수 중 실효환율과 주식시장 수익률이 각각 녹색채권과 일반채권의 발행금리를 유의하게 높이는 것으로 나타났다. 채권시장의 수익률과 변동성은 녹색채권의 발행금리만 유의하게 높이며 무위험수익률은 일반채권의 발행금리만 유의하게 높이는 것으로 나타났다. 또한 주식시장 변동성은 녹색채권의 발행금리는 낮추는 반면에 일반채권의 발행금리는 높이는 것으로 나타났다.

Table 3. Regression Results for bond rate from 2022.01 to 2023.07

	Green bond			Non-Green bond		
	(1)	(2)	(3)	(1)	(2)	(3)
Constant	6.6052*** (9.4629)	0.4413** (1.8716)	0.4280** (1.8017)	4.1766*** (5.3478)	0.5476** (1.7795)	0.4042 (1.297)
amount	-0.1881 (-0.2694)	-0.0009 (-0.0039)	-0.0005 (-0.0019)	0.0373 (0.0478)	-0.0548 (-0.178)	-0.0657 (-0.2107)
maturity	-0.3913 (-0.5605)	-0.1439 (-0.6105)	-0.1418 (-0.5969)	-0.0851 (-0.1089)	-0.1700 (-0.5523)	-0.1504 (-0.4826)
Type	-0.0381 (-0.0546)	0.1361 (0.5770)	0.1284 (0.5404)	-0.0679 (-0.087)	0.2665 (0.8658)	0.2542 (0.8157)
Rating	-0.1003 (-0.1437)	0.0018 (0.0077)	0.0025 (0.0106)	0.1325 (0.1696)	0.1306 (0.4245)	0.1272 (0.408)
RF		0.1408 (0.5971)	0.1302 (0.5481)		0.8621*** (2.8013)	0.8809*** (2.8265)
EER		4.5006*** (19.0861)	5.2366*** (22.0426)		1.6207*** (5.2663)	2.9995*** (9.6237)
KOSPI		8.4312*** (35.755)	6.5685*** (27.6489)		6.9808*** (22.6833)	12.1378*** (38.9436)
VKOSPI			-0.5520** (-2.3237)			1.3650*** (4.3796)
BR		0.8786*** (3.7259)	0.8888*** (3.7413)		0.2841 (0.9231)	0.2991 (0.9597)
move			0.4896** (2.0611)			0.1607 (0.5157)
Adj_R2	0.17	0.91	0.9	0.003	0.85	0.84

Note : This table presents estimation result from January 2022 to July 2023. Numbers in parentheses are t-value. Statistical significance at the 1 and 5 percent level is indicated by *** and **.

References

Baker, M., Bergstresser, D., Serafeim, G., & Wurgler, J. (2018). Financing the response to climate change: The pricing and ownership of US green bonds (No. w25194). National Bureau of Economic Research

Ehlers, T., Packer, F. (2017), "Green bond finance and certification." BIS Quarterly Review September.

Flammer, C. (2021). Corporate green bonds. Journal of financial economics, 142(2), 499-516.

Gianfrate, G., & Peri, M. (2019). The green advantage: Exploring the convenience of issuing green bonds. Journal of cleaner production, 219, 127-135.

Huang, C. Y., Dekker, D., & Christopoulos, D. (2023). Rethinking greenium: A quadratic function of yield spread. Finance Research Letters, 54, 103710.

Karpf, A., & Mandel, A. (2018). The changing value of the 'green'label on the US municipal bond market. Nature Climate Change, 8(2), 161-165.

Larcker, D. F., & Watts, E. M. (2020). Where's the greenium?. Journal of Accounting and Economics, 69(2-3), 101312.

Nanayakkara, M., & Colombage, S. (2019). Do investors in Green Bond market pay a premium? Global evidence. Applied Economics, 51(40), 4425-4437.

Tang, D. Y., & Zhang, Y. (2020). Do shareholders benefit from green bonds?. Journal of Corporate Finance, 61, 101427.

Wang, J., Chen, X., Li, X., Yu, J., & Zhong, R. (2020). The market reaction to green bond issuance: Evidence from China. Pacific-Basin Finance Journal, 60, 101294.

Wu, Y. (2022). Are green bonds priced lower than their conventional peers?. Emerging markets review, 52, 100909.

Zerbib, O. D. (2019). The effect of pro-environmental preferences on bond prices: Evidence from green bonds. Journal of banking & finance, 98, 39-60.

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 125-125

분포동학을 통한 중국 위안화 환율의 안정성 분석 및 결정요인

강효우[1], 박성용[2]

요 약

본 연구는 달러화 대비 중국 위안화 환율의 동태적 비선형 분포동학(Distribution Dynamics)을 이용하여 그의 안정성 및 결정요인을 분석하였다. 구체적으로 환율의 조건부 확률밀도함수가 이자율과 같은 거시경제변수에 영향을 비선형적으로 받도록 설정하여, 매 시점에서 거시경제변수가 환율분포에 미치는 효과를 추정하였다. 이는 시간의 흐름에 따라 확률분포의 형태가 변화될 수 있게 함으로써 동학적 움직임을 역시 관찰하고자 한다. 이러한 분포의 가정에 의해 나타나는 환율의 조건부 확률밀도함수는 단봉(Unimodality)과 쌍봉(Bimodality)의 형태로 나타날 수 있는데, 단봉(Unimodality) 형태가 안정적 균형의 상태로 볼 수 있는 반면, 외환위기나 글로벌 금융위기의 시기에는 불안정적 균형이라고 볼 수 있는 쌍봉(Bimodality)형태가 나타났다. 또한 본 연구는 실증분석을 통해 환율의 분포동학에 대한 결정요인도 분석하였다.

본 연구는 기존 연구들에서 다루어지지 않은 중국 위안화 환율에 대한 동태적 비선형 모형을 설정하여 분석하는데 그 공헌도가 있으며, 분석의 결과에서 나타나는 시사점을 통해 향후 거시경제정책에 도움을 줄 것으로 사료된다.

주요용어 : 중국 위안화, 비선형분포모형, 조건부 확률분포분석, 환율변동 결정요인.

[1]06974 서울특별시 동작구 흑석로 84, 중앙대학교 일반대학원 경제학과 박사과정. E-mail: kxy8166@cau.ac.kr
[2](교신저자) 06974 서울특별시 동작구 흑석로 84, 중앙대학교 경영경제대학 경제학부 교수.
E-mail: sungpark@cau.ac.kr

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 127-127

Cryptocurrencies as Hedges and Safe-havens: A Flexible Semi-parametric Approach

Myeong Jun Kim[1], *Meiling Jin*[2], *Sung Y. Park*[3]

Abstract

In the linear regression models, cryptocurrencies do not exhibit hedging properties against stocks. However, they demonstrate hedging capabilities against bonds. When employing a time-varying regression model, our findings indicate that the hedging characteristics of cryptocurrencies vary across periods for stocks and gold. During the COVID-19 pandemic, when stock and gold market volatilities experienced significant increases, cryptocurrencies did not exhibit hedging properties against either asset. These periods are identified as the period between the end of 2020 and the end of 2022 for stocks and between early 2019 and the end of 2021 for gold. In contrast, the linear and time-varying regression models suggest that cryptocurrencies act as bond hedge assets. The estimated results from the time-varying regression model reveal that when market conditions are exceptionally poor, as indicated by the 0.01 quantile level, more than half of the cryptocurrencies tend to function as safe-havens for assets. Our findings highlight that the relationship between cryptocurrencies and stocks can vary over time, particularly during market shocks, such as the COVID-19 pandemic. While cryptocurrencies initially lost their hedging capability for stocks during the pandemic's turbulent period, they regained their role as hedging instruments as market conditions stabilized. This suggests that the impact of market shocks and the factors driving uncertainty play crucial roles in shaping the correlation between cryptocurrencies and stocks. Furthermore, our analysis reveals intriguing patterns in the relationship between cryptocurrencies and gold. The correlation between cryptocurrencies and stocks influences the correlation between cryptocurrencies and gold. During certain periods, cryptocurrencies exhibit no relationship with gold, indicating a potential diversification benefit. However, the hedging behavior of specific cryptocurrencies can vary significantly, underscoring the need for careful consideration of market conditions and individual characteristics of different coins. Our study contributes to the existing literature by employing semi-parametric methods that capture smoothly changing parameters over an entire sample period. This approach enables us to provide more reliable information for long-term investors and policymakers, particularly when considering the unique characteristics of cryptocurrencies, such as transaction costs and illiquidity.

Keywords: Cryptocurrency; Hedge; Safe-haven; Functional coefficient

---

[1]Division of International Studies, Kongju National University, Gongjudaehak-ro 56, Gongju-si, Chungcheongnam-do, Korea. Email: myeongjun@kongju.ac.kr

[2]School of Economics, Chung-Ang University, 84 Heukseok-Ro, Dongjak-Gu, Seoul, Korea. E-mail: meiling1206@cau.ac.kr.

[3]Corresponding Author: School of Economics, Chung-Ang University, 84 Heukseok-Ro, Dongjak-Gu, Seoul, Korea. E-mail: sungpark@cau.ac.kr.

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 129-132

전 세계 취약점에 대한 통계적 분석: 패치 미적용 서버의 동향*

강병훈[1], 이혜원[2]

요 약

보안 취약점은 컴퓨터 시스템, 소프트웨어, 네트워크 또는 정보 시스템에서 보안을 침해하거나 악용할 수 있는 잠재적인 약점 또는 결함을 가리키며 CVE 는 컴퓨터 및 정보보안 분야에서 사용되는 국제적인 식별체계로, 취약점 정보를 표준화하고 공유하는 데 사용된다. CVE 는 전 세계적으로 계속해서 증가하는 추세지만, 몇 년이 지난 기존 취약점들도 아직도 제대로 대응이 되지 못하고 있는 것이 현재 IT환경에서 큰 문제로 부상하고 있다. 본 논문에서는 최근 5년간(CVE 2018년부터 2023년)의 CVE 개수 통계를 조사하고 Criminal IP 사이버 위협 인텔리전스 검색 엔진을 사용하여 현재 전 세계적으로 아직도 해결되지 않은 취약점이 얼마나 존재하고 있는지 그 내용을 살펴본다.

주요용어 : 취약점, CVE, Criminal IP, 보안패치

1. 서론

보안 취약점은(Vulnerability)은 컴퓨터 시스템, 소프트웨어, 네트워크 또는 정보 시스템에서 보안을 침해하거나 악용할 수 있는 잠재적인 약점 또는 결함을 가리킨다. 이러한 취약점은 해커, 악의적인 공격자, 불법 액세스를 시도하는 자 또는 악성 코드와 같은 보안 위협으로부터 악의적으로 사용될 수 있다.

CVE(CVE - Common Vulnerabilities and Exposures)는 컴퓨터 및 정보 보안 분야에서 사용되는 공통된 식별체계다. CVE는 취약점과 보안 취약점에 대한 정보를 표준화하고 공유하기 위해 만들어진 국제적인 시스템으로, CVE는 취약점을 고유하게 식별하고 관리하는데 그 목적이 있다. 따라서 CVE 에 대한 분석을 진행하는 것만으로 어떤 어플리케이션이 어떤 보안 취약점을 보유하고 있는지 확인할 수 있다(NIST NVD (National Vulnerability Database) https://nvd.nist.gov/general).

본 논문에서는 최근 5년간 (2018년부터 2023년까지) 의 연도별 CVE 개수를 파악하고 연도별 증감 추이를 확인해 봄으로써 얼마나 많은 보안 취약점 발표되고 있는지 그 통계적 동향을 살펴보고, 사이버 위협 인텔리전스 검색엔진인 Criminal IP 를 이용하여 현재 전 세계적으로 지금도 해결되지 않은 취약점이 얼마나 존재하는지 살펴보고 그 위험성에 대해 평가하였다(Criminal IP - Cyber Threat Intelligence Search Engine https://www.criminalip.io).

---

*이 논문은 2022년도 민군협력진흥원의 국방기술상용화지원사업의 재원으로 국가연구개발사업비 지원을 받아 수행된 연구임(No. 22DCIN12).

[1]04782 서울시 성동구 연무장 5가길 7, 더블유동 701~704호 AI SPERA 개발팀 부장 Email: bhkang@aispera.com
[2](교신저자) 04782 서울시 성동구 연무장 5가길 7, 더블유동 701~704호 AI SPERA ASM개발팀 팀장 Email: hwlee@aispera.com

2. 각 국가별 어플리케이션 분석과 통계

최근 5년간 (2018년부터 2023년까지) NIST 에 등록된 공식 CVE 의 개수를 연도별로 파악하고 연도별 증감 추이를 확인해보았다(Figure 1). CVE는 2018년 15,565개로 시작해서 2022년 22,610개, 2023년 29,065개로 5년간 증가하는 추세를 확인할 수 있다. 특히 취약점 총 개수 모수가 늘어남에도 CVSS 그룹의 심각, 높음, 중간의 비율 역시 동일한 분포도로 상승하며 유지되는 추세인 것을 확인할 수 있다. 따라서 심각, 높음, 중간에 해당되는 취약점의 절대 건수 자체도 지속적으로 늘어나고 있음을 알 수 있다(Figure 2).

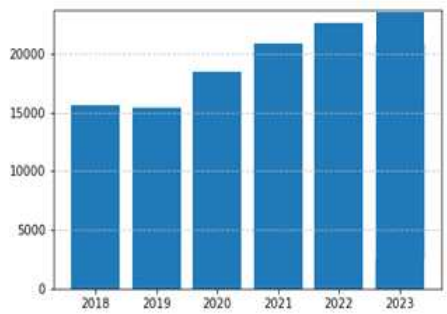


Figure 1. 2018~2023년간의 취약점 개수 추이

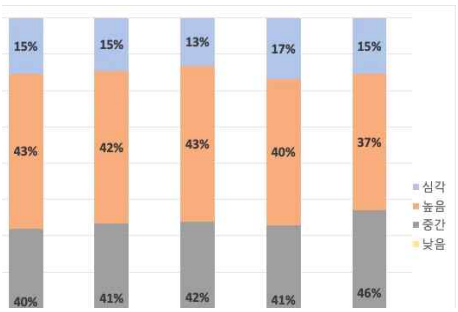


Figure 2. 2018-2023년간의 CVSS 그룹별 취약점 수 추이

Figure 3. TCP/443 포트에서 가동되는 국가 통계, 호스팅 정보 통계, Product 통계

서버 어플리케이션 중에 대중적으로 가장 많이 사용되는 것은 웹 서버로, 특히 TCP/443 포트에서 가동되는 웹서버에 대해 Criminal IP 데이터를 이용해 전 세계의 통계를 내 보았다. United States가 15,991,478개로 가장 많이 나타났으며, Germany가 4,256,087개, China가 2,219,353개로 가장 두드러진 결과를 보여주고 있다. 또한 서버의 ASN_NAME(Autonomous System Numbers)(Autonomous System Numbers https://www.arin.net/resources/guide/asn/)으로 살펴보면 아마존 서버(amazon-02) 가 8,119,224개로 가장 많이 보이고 있으며, akamai-as가 4,102,984개, amazon-aes가 2,711,265개로 3위를 차지하고 있고 Microsoft Azure 서버로 보이는 microsoft-corp-man-as-block 이 4위, GCP 클라우드로 보이는 google-cloud-platform 이 5위를 차지하고 있다. TCP/443 에서 가동되는 어플리케이션으로는 httpd, nginx, amakaihost, apache 등이 차지하고 있다. 이렇게 보면 알 수 있듯이 전 세계에는 TCP/443 서버에서 가동되는 서버는 5천만개 이상 인터넷에 연결된 상태라는 것을 확인할 수 있다. (Figure 3)

3. 아파치 서버 취약점 CVE-2021-25122 의 통계적 분석

웹 서버로 널리 사용되는 아파치 서버의(Usage statistics of web servers https://w3techs.com/technologies/overview/web_server/) 보안 취약점 중 하나인 CVE-2021-25122(CVE-2021-25122 https://nvd.nist.gov/vuln/detail/CVE-2021-25122)는 2021년 3월 1일에 발표된 취약점으로, 웹서버는 클라이언트의 새로운 연결 요청에 응답할 때 Apache Tomcat 버전 10.0.0-M1~10.0.0, 9.0.0.M1~9.0.41 및 8.5.0~8.5.61은 요청 헤더와 제한된 양의 요청 본문이 복제될 수 있다는 문제를 가지고 있다. 이 취약점으로 인하여 웹사이트에 방문하는 사용자는

본인의 요청이 아닌 다른 사람의 요청 결과를 볼 수도 있다. 따라서 본인이 아닌 다른 사람의 중요한 정보나 개인정보 등을 악의적으로 접근할 수 있는 문제를 가지고 있다. 이 취약점을 가진 서버는 신속하게 보안패치를 수행하여 취약점을 없애야 한다.

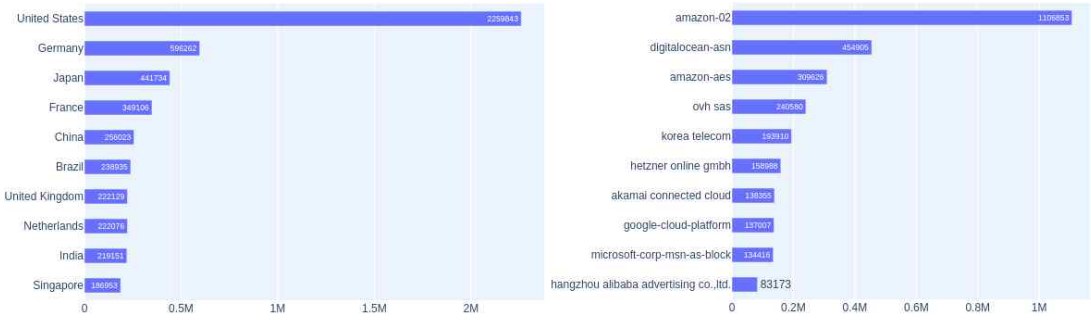


Figure 4. CVE-2021-25122를 보유한 서버의 통계

이 취약점이 등장한 이후, 당시 전 세계의 모든 아파치 서버는 보안 패치 권고를 받으며 신속하게 서버 패치를 진행하였다(Apache Tomcat multiple vulnerabilities https://www.tenable.com/plugins/nessus/150856). 그리고 3년이 지난 지금, 전 세계에서 이 취약점을 아직도 가지고 있는 웹서버가 있는지 조사해 보았다. [Figure 4] 는 이 CVE-2021-25122 취약점을 보유하고 있는 전 세계의 아파치 서버의 통계를 Criminal IP를 통해 추출하여 나타낸 것이다. 총 182,309건의 서버에 보안 취약점

이 있는 것을 알 수 있으며, Brazil가 52,600개로 가장 많이 나타났으며, China가 34,542개, Kazakhstan가 19,910개로 세 번째로 많은 국가에서 취약점이 있음을 보여주고 있다. 또한 ASN Name 통계에서는 v tal가 23228개로 가장 많이 나타났으며, jsc kazakhtelecom가 16825개, hangzhou alibaba advertising co.,ltd.가 12366개를 보여주고 있다. 무엇보다 중요한 점은, CVE-2021-25122 는 취약점이 등장한지 거의 3년이 다 되어가는 문제인데, 아직도 이렇게 18만건이 넘을 정도로 다양한 국가와 다양한 호스팅 서버에서 취약점을 보유한 서버가 가동되고 있다는 점이다. Criminal IP 에서 수집된 취약점 통계를 확인해본 결과, 이것은 단지 아파치 서버의 CVE-2021-25122 문제 에 국한된 것이 아니며 모든 취약점이 비슷한 문제를 안고 있다.

4. 결론

본 보고서는 크게 최근 5년간의 2023년의 취약점 동향과 분야별 취약점 동향에 대해 분석했다. Criminal IP 를 이용하여 취약점의 국가적 통계, 그리고 어플리케이션의 통계를 살펴보며 피해 자산의 규모를 파악했다. NIST 에서 취약점이 공개되어도 보안 패치를 수행하여 취약점을 해결하지 않는 경우가 전 세계적으로 굉장히 많다는 것을 알 수 있으며 이에 대해 보안 문제가 아직도 지속적으로 발생하고 예상할 수 있다.

따라서, 기업은 보다 적극적으로 보안 취약점에 대한 관리를 할 필요가 있으며, 특히 클라우드로 발전하는 요즘 시대는 어떤 서버가 어떤 클라우드에서 가동되고 있는지 관리가 더욱 어려운 때가 되었으므로 OSINT (Open Source Intelligence)(Open-Source Intelligence? https://www.sans.org/blog/what-is-open-source-intelligence/) 기반의 취약점 스캐닝을 통해 취약점 현황 정보를 지속적으로 확보하는 것이 중요하다는 것을 시사할 수 있다.

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 133-136

Determinants of Managerial Pay: The Relative Contribution of Compensation Predictors

Su-In Kim[1], *Jinsuk Heo*[2], *Injoong Kim*[3]

Abstract

Firm characteristics that determine CEO pays are closely interrelated with one another and make the partitioning of variances among correlated multiple predictors difficult. We decompose the interrelated predictors by orthogonalizing the influence of each predictor based on Tonidandel, LeBreton's (2015) relative weight analysis on both normal and crisis period. In the process, we can rank the relative importance of each predictor and investigate its evolution over the economic crisis period. Firm size is the most dominant determinant, occupying over 60% relative weight. Wage discrimination against small company is obvious. ROA contributes 8.7% for the normal period and 10.8% for the crisis, which implies that CEOs' ability to generate profits in crisis is particularly valued high and companies reward managers accordingly. The prolonged good performance is especially valued higher (13.9%) than the short-term performance. Risk and cash flow volatility occupy 3.6% and 1.8%, respectively, and the use of funds, such as capital expenditure and interest payment triggered by leverage occupy only marginal portions. This suggests that firms may lower CEO pays to reserve cash when they face greater risks or new investment opportunities, but the amount of extraction may not be high. In crisis, credit information can potentially outweigh the importance of many other typical predictors.

Keywords : CEO, compensation, relative weight analysis, credit rating, Covid-19

1. Introduction

This paper analyzes the link between the CEO performance and the corporate compensation policy on a representative sample of the exchange traded firms in both the normal and the economic crisis period. Firm characteristics, such as firm size, ROA, cash flow volatility, risk and etc. that influence the CEO pays are closely interrelated and the traditional

[1]Assistant Professor, Accounting Major, College of Business Management, Hongik University, 2639 Sejong-ro, Sejong, 30016, Korea. E-mail: suinkim@hongik.ac.kr
[2]Assistant Professor, Accounting Major, College of Business Management, Hongik University, 2639 Sejong-ro, Sejong, 30016, Korea. E-mail: jinsukheo@hongik.ac.kr
[3](Corresponding Author) Assistant Professor, Finance & Insurance Major, College of Business Management, Hongik University, 2639 Sejong-ro, Sejong, 30016, Korea. E-mail: kij@hongik.ac.kr

statistics, such as correlations and standardized regression weights may yield faulty or misleading information about the variable importance. Relative weight analysis enables us to disentangle the interrelated compensation predictors by partitioning each explanatory variable's unique marginal contribution (Tonidandel, LeBreton, 2015). Consequently, we can rank the pure marginal contribution of each predictor and make more accurate inferences. We extend the traditional compensation model of Fermandes, Ferreira, Matos, Murphy (2013) and Gao, Li's (2015), and examine the relative contribution of constituent compensation predictors and its evolution over the Covid-19 period.

2. Data and Methodology

We analyze CEO compensation policies of both KOSPI and KOSDAQ firms in TS2000 data base over the sample period of 2011-2021. The following compensation model is estimated. (see Fermandes, Ferreira, Matos, Murphy, 2013; Gao, Li, 2015)

$$Ln(CEO\ Pay) = \alpha + \beta_1 Ln(total\ assets) + \beta_2 Performance\ measures + \beta_3 Other\ firm\ characteristics + Industry\ FEs + Year\ FEs + \epsilon \quad (1)$$

The relative weight analysis (RWA) of Tonidandel, LeBreton (2015) is a method to partition explained variance among multiple predictors to better understand the role played by each predictor in a regression equation. The individual weights should add up to the total R^2 and it can be presented in terms of the percent of predictable variances by each predictor.

$$E[Y_A] - E[Y_B] = \{E[X_A] - E[X_B]\}'\beta^* + E[X_A]'(\beta_A - \beta^*) + E[X_B]'(\beta^* - \beta_B) \quad (2)$$

Additionally, to clarify the contribution of credit ratings in isolation from the traditional compensation predictors, we adopt the decomposition method (see Jann, 2008; Elder, Goddeeris, Haider, 2010). Pay differentials can be dividend into the portion explained by the traditional compensation model and the residual part attributable to credit group effect.

3. Empirical Results

In Table 3. we analyze the relative contribution of explanatory variables. It reports the estimates of coefficients together with the relative weights of each predictor over the normal and the Covid-19 period.

Table 3. Relative weight analysis of the CEO compensation model

Panel A. Traditional model						
Predictors	Coefficient	Std. Error	Normal Period		Covid-19 Period	
			RW	Rescaled RW	RW	Rescaled RW
Intercept	5.789*	0.227				
Size	0.328*	0.005	0.2691	0.684	0.2470	0.670
Capex	0.411*	0.074	0.0014	0.004	0.0040	0.011
Cash	0.370*	0.071	0.0054	0.014	0.0081	0.022
CF volatility	6.275*	0.352	0.0069	0.018	0.0070	0.019
RE/TE	0.153*	0.007	0.0545	0.139	0.0519	0.141
Leverage	-0.105*	0.031	0.0039	0.010	0.0022	0.006
Risk	-0.035*	0.012	0.0141	0.036	0.0059	0.016
ROA	1.354*	0.082	0.0342	0.087	0.0399	0.108
Stock return	-0.013	0.008	0.0034	0.009	0.0022	0.006
Sales growth	0.010	0.012	0.0007	0.002	0.0004	0.001
R^2			0.393	1.000	0.369	1.000
Panel B. Augmented with Credit rating						
Predictors	Coefficient	Std. Error	Normal Period		Covid-19 Period	
			RW	Rescaled RW	RW	Rescaled RW
Intercept	6.180*	0.228				
Size	0.319*	0.005	0.2639	0.661	0.2326	0.619
Capex	0.325*	0.074	0.0013	0.003	0.0040	0.011
Cash	0.308*	0.071	0.0056	0.014	0.0085	0.023
CF volatility	6.107*	0.350	0.0064	0.016	0.0071	0.019
RE/TE	0.134*	0.008	0.0489	0.123	0.0431	0.115
Leverage	-0.119*	0.043	0.0101	0.025	0.0054	0.015
Risk	-0.027*	0.007	0.0131	0.033	0.0054	0.014
ROA	0.723*	0.096	0.0254	0.064	0.0271	0.072
Stock return	-0.014	0.009	0.0034	0.008	0.0020	0.005
Sales growth	0.011	0.020	0.0005	0.001	0.0004	0.001
Credit rating	-0.069*	0.006	0.0207	0.052	0.0399	0.106
R^2			0.399	1.000	0.375	1.000

Note: RW stands for the raw relative weight of each predictor in the compensation regression. Within rounding error, raw weights will sum up to R^2. Rescaled RW represents the rescaled percent of predicted variance attributable to each component so that they can sum to 100%. * denotes the statistical significance at 1% level.

The model's explanatory power drops during the crisis period. Knowing traditional firm characteristics such as ROA, leverage, firm size and etc. may not be sufficient to accurately forecast CEO salaries. The credit information seems to compensate the losses in the explanatory power of the traditional compensation model.

Table 4. Decomposition of the compensation difference

	Ln(Pay)	Exponentiated		Ln(Pay)	Exponentiated
High credit	12.565	286,345.6	Mid credit	12.351	231,154.1
Mid credit	12.351	231,154.1	Low credit	11.966	157,240.5
Difference	0.214* (0.017)	1.239* (0.020)	Difference	0.385* (0.026)	1.470* (0.038)
Endowment	0.156* (0.016)	1.169* (0.018)	Endowment	0.235* (0.023)	1.264* (0.029)
Residual	0.058* (0.017)	1.059* (0.018)	Residual	0.151* (0.028)	1.163* (0.032)
Proportion of explained part	0.731		Proportion of explained part	0.609	

Note: Total effect consists of the firm characteristic endowment effect and the residual effect of credit ratings. Bootstrap standard errors are reported in parentheses and * indicates the significance at 1% level.

4. Conclusion

Firm size is the determinant of CEO pays occupying over 60% of the weight. ROA occupies 8.7% weight during the normal period and its portion increases to 10.8% during the Covid-19 period. The ability to generate profits during the economic crisis period is particularly valued high and companies seem to reward managers accordingly. RE/TE occupies 13.9% of the total weight, which implies that managers showing prolonged good performance is likely to be paid higher than the managers showing short-term performance. Risk and cash flow volatility occupies relatively small portion. Firms may lower CEO pays to reserve cash when they face risks but the amount of extraction may not be that great. Similar pattern is observed in the capital expenditure and leverage. Credit rating occupies 5.2% weight and during the crisis, this portion increases to 10.6%. Similar result is derived when we decompose the mean difference of CEO pays. A potential CEO who is afraid of salary cuts in crisis may be better off by considering credit information when searching for a job.

References

Gao, H., Li, K. (2015). A comparison of CEO pay-performance sensitivity in privately-held and public firms, *Journal of Corporate Finance*, 35, 370-388. DOI:10.1016/j.jcorpfin.2015.10.005

Tonidandel, S., LeBreton, J., Johnson, J. (2009). Determining the statistical significance of relative weights, *Psychological Methods*, 14(4), 387-399. DOI:10.1037/a0017735

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 1-1

데이터필로소피 - 계몽사상과 확률론의 만남

김태영[1]

요 약

통계학은 '과학의 문법'이라는 유명한 별칭을 가지고 있거니와 통계학의 문법은 확률이라고 말할 수 있다. 확률을 통한 사고는 학문의 지형과 우리의 사고방식을 어떻게 바꾸어 놓았는지를 추적해 가는 것으로 데이터필로소피(data philosophy)의 장대한 여정을 시작하고자 한다. 데이터필로소피는 데이터사이언스(data science)의 엄청난 파고에 대응하여 데이터를 근원적으로 고찰하는 시도임을 주장한다. 데이터필로소피는 통계학을 위시한 데이터 학문의 역사성과 사상적 배경을 강조하고, 빅데이터라는 레토릭으로 대변되는 데이터가 지배하는 세상에 대한 비판적 사고를 지향한다. 이에, 본 연구는 결정론이 침식되고 우연을 이성의 지배 아래 두기 시작한 지점으로 거슬러 올라가 초기 확률론자(classical probabilists)와 그들이 활동했던 시대상인 계몽주의 및 당대의 사상가들과의 교유(交遊)를 집중 조명한다. 이를 통해 현 시대를 주름잡고 있는 데이터사이언스의 역사적, 사상적 배경에 천착할 수 있는 단초를 제공할 것이다.

주요용어 : 데이터필로소피, 데이터사이언스, 결정론, 확률론, 계몽주의

[1]47340 부산시 부산진구 엄광로 176, 동의대학교 디그니타스교양교육연구소, 연구교수.
E-mail : tkim33@deu.ac.kr

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 139-

표본 분위수 계산 방법에 관한 고찰: 이산형 분포의 경우

김혁주[1]

요 약

본 논문에서는 이산형 분포의 경우에 표본 분위수를 계산하는 방법에 관하여 고찰하였다. 통계학 교재에서 주로 소개하는 두 가지 방법을 모의실험을 통하여 비교하였다. 포아송분포, 이항분포, 기하분포, 이산균등분포의 경우에 대해 모수의 값을 몇 가지로 설정하고, 설정된 각각의 분포에 대해 일정 크기의 확률표본을 추출하는 모의실험을 10,000회씩 실시하였다. 각 회의 모의실험에서는 두 가지 방법 각각에 의하여 계산된 표본 분위수와 모집단 분위수의 차이를 계산하여 두 가지의 기준에 의해 비교하였다. 또한 두 방법에 대해 과소·정확·과대 추정의 확률분포에 대한 동질성검정을 실시하였다.

주요용어 : 동질성검정, 모의실험, 분위수, 십분위수, 평균제곱오차.

1. 서론

확률변수 X의 분포의 제$100p$백분위수($0<p<1$)는 $P(X<\xi_{p)} \le p$, $P(X \le \xi_{p)} \ge p$를 만족하는 값 ξ_p로 정의된다(Hogg and Craig (1978, p.30)). 제$100p$백분위수는 p차 분위수라고도 한다. 모집단의 분위수를 추정하는 데 있어서 표본 자료로부터 계산된 분위수를 사용한다. 그런데 표본 자료가 주어졌을 때 이로부터 분위수를 계산하는 구체적 방법은 한 가지로 통일되어 있지 않다. 통계학 교재에서는 두 가지 방법이 주로 소개되고 있다. Kim (2023)은 연속형 분포의 경우에 이 두 가지 방법을 모의실험을 통하여 비교하였다.

백분위수의 경우를 통하여 두 방법(방법 1과 방법 2로 표기)을 설명한다. 방법 1은, 자료를 작은 값이 앞에 오고 큰 값이 뒤에 오게 순서대로 늘어놓았을 때 적어도 $100p$%의 관측값이 그 값보다 작거나 같으며, 동시에 적어도 $100(1-p)$%의 관측값이 그 값보다 크거나 같게 되는 값을 제$100p$백분위수라고 정의한다. 만약 이 값이 유일하게 결정되지 않으면 그 값들의 평균을 사용하는데, 여기서 평균은 단순평균을 의미한다. 이는 Bhattacharyya와 Johnson (1977), Kim 등(2006)에서 소개하는 방법이며, Walpole (1982)과 Kim 등(2002)에서도 이 방법으로 백분위수를 계산하고 있다.

자료의 순서통계량값들을 $y_1, y_2, \cdots, y_n$으로 나타내자. 방법 2는, $(n+1)p$가 정수이면 $y_{(n+1)p}$를 자료의 제$100p$백분위수로 하고, 정수가 아니면 $y_{[(n+1)p]}$와 $y_{[(n+1)p]+1}$([ ]는 가우스 기호)의 가중평균을 자료의 제$100p$백분위수로 하는 방법이다. 이는 Hogg 등(2019)과 Baek 등(2021)에서 사용하는 방법이다.

[1](54538) 전북 익산시 익산대로 460, 원광대학교 빅데이터·금융통계학부 교수. E-mail: hjkim@wku.ac.kr

제1(D_1), 제2(D_2), …, 제9십분위수(D_9)는 각각 제10, 제20, …, 제90백분위수와 같은 것이다. 본 논문에서 모집단 십분위수는 D_1, D_2,…, D_9로, 표본 십분위수는 $\widehat{D}_1$, $\widehat{D}_2$,…, $\widehat{D}_9$로 나타내겠다. 예를 들어 $n=20$인 경우 방법 1과 방법 2에 의하면 십분위수들이 다음 식에 의해 계산된다. 표본 중위수 $\widehat{D}_5$는 두 방법이 항상 같은 결과를 준다.

$$\widehat{D}_i=\begin{cases}\dfrac{y_{2i}+y_{2i+1}}{2} & (\text{방법 } 1)\\ \dfrac{(10-i)y_{2i}+iy_{2i+1}}{10} & (\text{방법 } 2)\end{cases} \qquad (i=1,2,\cdots,9)$$

본 논문에서는 이산형 분포의 경우 표본 분위수를 계산하는 두 가지 방법에 관하여 고찰하고자 한다. 분위수를 대표하여 9가지의 십분위수의 경우를 살펴본다.

2. 모집단 분위수와의 차이를 바탕으로 한 비교

몇 가지 이산형 분포의 경우에 대해 크기 $n=20$인 확률표본을 추출하는 모의실험을 Minitab ver. 19를 사용하여 10,000회씩 실시하였다. 각 회의 모의실험에서는 방법 1과 방법 2 각각에 의하여 계산된 표본 분위수와 모집단 분위수의 차이를 계산하였다. 두 방법을 비교하는 기준으로 두 가지(기준 1, 기준 2)를 사용하였는데, 기준 1은 10,000회의 모의실험에 걸친 평균제곱오차이고, 기준 2는 10,000회의 모의실험 중 모집단 분위수와의 차이가 상대 방법보다 더 작은 횟수이다.

Table 1부터 Table 5까지는 모의실험 결과를 나타낸 표이다. 각 칸에 있는 숫자는 기준 1의 값이며, 괄호 안의 숫자는 기준 2의 값이다. 예를 들어 평균이 3인 포아송분포인 $Poi(3)$의 경우를 나타낸 Table 1을 보면, 모집단 제1십분위수는 $D_1=1$인데, 기준 1의 값은 방법 1이 0.231, 방법 2가 0.306으로 방법 1이 우세하며, 기준 2의 값은 방법 1이 1,902회, 방법 2가 1,320회로 나와 역시 방법 1이 더 우세한 것으로 나타났다(p-값<0.001). 그 밖의 십분위수를 보면 D_2, D_6, D_8, D_9에서는 방법 1이, D_4, D_7에서는 방법 2가 우세하게 나왔다. 기준 2로 할 때 유의수준 5%에서 유의하게 우세한 쪽은 * 기호로 표시하였다. 대부분의 경우 기준 1에서 우세한 쪽이 기준 2에서도 우세한데, 그렇지 않은 경우도 있다. Table 2는 $Poi(4)$의 경우인데, D_7과 D_9의 경우 기준 1에서는 방법 1이 우세하지만 기준 2에서는 방법 2가 우세하게 나왔다. 이것은 인접한 두 순서통계량의 가중평균으로 D_7과 D_9를 추정하는 것이 단순평균으로 추정하는 것에 비해 모수에 가까운 값을 얻는 빈도는 높았지만, 모수에서 먼 값을 얻을 때는 방법 1에 비해 차이가 상대적으로 컸다는 의미가 되겠다. Table 3, 4, 5는 지면 관계상 각각 이항분포 $B(5,0.3)$, 기하분포 $Geo(0.5)$, 그리고 $\{1,2,3,4,5,6\}$에서의 이산균등분포의 경우에 대해서만 10,000회씩의 모의실험 결과를 나타낸 표이다. 기하분포의 경우 앞쪽 십분위수에서는 방법 2가 우세하다가 뒤쪽 십분위수로 감에 따라 방법 1이 우세한 경향이 있었고, 나머지 분포들의 경우에는 일정한 경향이 없고 몇 분위수인가에 따라 방법 1이 우세한 경우도 있고 방법 2가 우세한 경우도 있었다.

Table 1. Simulation result for $Poi(3)$

	$D_1=1$	$D_2=2$	$D_3=2$	$D_4=2$	$D_5=3$	$D_6=3$	$D_7=4$	$D_8=4$	$D_9=5$
Method 1	.231 (1902*)	.452 (2248*)	.223 (1069)	.434 (181)	.289(0)	.367 (1747*)	.385 (915)	.543 (2711*)	.689 (3653*)
Method 2	.306 (1320)	.543 (221)	.232 (1026)	.416 (1867*)	.289(0)	.384 (507)	.378 (1750*)	.678 (673)	1.058 (1332)

Table 2. Simulation result for $Poi(4)$

	$D_1=2$	$D_2=2$	$D_3=3$	$D_4=3$	$D_5=4$	$D_6=4$	$D_7=5$	$D_8=6$	$D_9=7$
Method 1	.466 (3066*)	.401 (745)	.330 (1606*)	.448 (507)	.365(0)	.497 (2046*)	.440(1400)	.683 (1222)	.897 (2225)
Method 2	.668 (657)	.373 (2181*)	.355 (951)	.432 (1851*)	.365(0)	.521 (562)	.451 (1567*)	.660 (2480*)	.978 (3044*)

Table 3. Simulation result for $B(5,0.3)$

	$D_1=0$	$D_2=1$	$D_3=1$	$D_4=1$	$D_5=1$	$D_6=2$	$D_7=2$	$D_8=2$	$D_9=3$
Method 1	.174(0)	.295 (2120*)	.070 (658*)	.114(75)	.357(0)	.236(71)	.102 (585)	.280 (2047*)	.237 (1002)
Method 2	.127 (1968*)	.377(17)	.082 (250)	.107 (847*)	.357(0)	.223 (1477*)	.107 (569)	.360(93)	.248(1932*)

Table 4. Simulation result for $Geo(0.5)$

	$D_1=1$	$D_2=1$	$D_3=1$	$D_4=1$	$1 \le D_5 \le$	$D_6=2$	$D_7=2$	$D_8=3$	$D_9=4$
Method 1	.00013 (0)	.0025(0)	.028(0)	.160(0)	.301(0)	.214 (621)	.333 (1898*)	.445 (1592)	.890 (2917*)
Method 2	.00010 (1)	.0016 (45*)	.022 (360*)	.149 (1213*)	.301(0)	.211 (1158*)	.382 (367)	.491 (1822*)	1.231 (2424)

Table 5. Simulation result for Discrete Uniform over $\{1,2,3,4,5,6\}$

	$D_1=1$	$D_2=2$	$D_3=2$	$D_4=3$	$3 \le D_5 \le$	$D_6=4$	$D_7=5$	$D_8=5$	$D_9=6$
Method 1	.208(0)	.370 (2022*)	.506 (621)	.424 (1491*)	.506(0)	.426 (1578*)	.507 (651)	.370 (2028*)	.202(0)
Method 2	.147 (2085*)	.427 (869)	.470 (2208*)	.430 (1183)	.506(0)	.433 (1146)	.472 (2191*)	.428 (920)	.142 (2066*)

3. 과소 · 정확 · 과대 추정의 확률분포에 대한 동질성검정

이산형 분포가 연속형 분포와 다른 점 중 하나는 분위수를 정확히(오차가 0이 되도록) 추정할 확률이 존재한다는 점이다. Table 6은 $Poi(3)$ 분포의 경우 10,000회의 모의실험 중 십분위수의 참값을 과소추정, 정확추정, 과대추정한 빈도를 D_1과 D_9의 경우에 대해 나타낸 표이다. 이 절에서 두 방법의 결과는 동일한 모의실험으로부터 얻은 것이 아니고 10,000회씩의 모의실험을 독립적으로 각각 실시하여 얻은 것이다. 동질성검정을 실시한 결과 D_1의 경우 카이제곱 검정통계량의 값이 1.073, p-값은 0.585로서 유의한 차이가 없었으나, D_9의 경우에는 카이제곱 검정통계량의 값이

26.594이고, p-값은 0.001보다 작아서 매우 유의한 차이를 보였다. $Poi(3)$의 경우 십분위수 중 D_9에서만 유의한 차이를 보였다.

동일한 방법으로 몇 가지의 분포에 대해서 동질성검정을 실시한 결과 두 방법이 유의한 차이를 보인 경우는 다음과 같다.

Table 6. Frequency distribution concerning estimation of D_1 and D_9 for $Poi(3)$

	$D_1 = 1$			$D_9 = 5$		
	under-estimation	exact estimation	over-estimation	under-estimation	exact estimation	over-estimation
Method 1	2580	5371	2049	2218	2953	4829
Method 2	2630	5370	2000	2249	2635	5116

(1) **포아송분포**

$Poi(7)$: $D_1 = 1$(p-값<0.001), $D_8 = 9$(p-값=0.008), $D_9 = 10$(p-값<0.001)
$Poi(10)$: $D_1 = 6$(p-값<0.001), $D_2 = 7$(p-값<0.001), $D_3 = 8$(p-값=0.032),
$D_6 = 11$(p-값=0.011), $D_7 = 12$(p-값=0.007), $D_8 = 13$(p-값<0.001),
$D_9 = 14$(p-값<0.001)

(2) **이항분포**

$B(5,0.3)$: $D_9 = 3$(p-값=0.029)
$B(5,0.5)$: $D_8 = 3$(p-값=0.006)
$B(10,0.3)$: $D_1 = 1$(p-값=0.002), $D_9 = 5$(p-값<0.001)
$B(10,0.5)$: $D_1 = 3$(p-값=0.008), $D_9 = 7$(p-값=0.033)

(3) **기하분포**

$Geo(0.3)$: $1 \le D_3 \le 2$(p-값=0.038), $D_7 = 4$(p-값=0.001), $D_8 = 5$(p-값<0.001),
$D_9 = 7$(p-값<0.001)
$Geo(0.5)$: $D_9 = 4$(p-값<0.001)

(4) **이산균등분포**

$DU(\{1,2,3,4,5,6\})$: $D_4 = 3$(p-값=0.022)

References

Baek, J., Son, Y.-S., Jeong J. (2021). Mathematical Statistics, Freedom Academy, Paju.
Bhattacharyya, G. K. and Johnson, R. A. (1977). Statistical Concepts and Methods, Wiley.
Hogg, R. V. and Craig, A. T. (1978). Introduction to Mathematical Statistics (4th edition), Macmillan.
Hogg, R. V., Tanis, E. A., Zimmerman, D. L. (2019). Probability and Statistical Inference (10th edition), Pearson.

Kim, B. H., Choi, K. C., Baek, H. Y., Kim, H. J., Dong, K. H., Park, T. R. and Chang, I. H. (2002). Understanding Statistics, Freedom Academy, Paju.

Kim, H. J. (2023). A study on computing sample quantiles of continuous distributions, Journal of the Korean Data Analysis Society, 25(1), 129-139 (in Korean).

Kim, W. C., Kim, J. J., Park, B. U., Park, S. H., Song, M. S., Lee, S. Y., Lee, Y. J., Jeon, J. W. and Cho, S. (2006). Modern Statistics (4th revised edition), Youngji Publishers, Seoul.

Walpole, R. E. (1982). Introduction to Statistics (3rd edition), Macmillan.

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 145-145

불균형 데이터 분류를 위한 SMOTE 비교연구: 가중치 분포를 중심으로

전병준[1], 엄태웅[2]

요 약

최근 불균형 데이터 문제에서 소수클래스 분류에 대한 문제를 해결하기 위한 다양한 연구가 이어지고 있다. 클래스가 불균형한 자료의 기계학습 과정에서 문제가 발생함에 따라 학습 과정에서 불균형 데이터 처리기법의 필요성이 제기되었다. 클래스가 불균형한 데이터의 밸런스를 조정하는 방법으로 오버샘플링, 언더샘플링과 같은 샘플링 방법이 있다. 언더샘플링은 소수클래스 수에 맞춰 다수클래스 데이터를 제거하는 방법으로 불균형 비율이 높은 자료에서는 유의한 정보의 손실을 야기하므로 본 연구에서는 고려하지 않는다. 오버샘플링의 대표적인 방법으로 SMOTE 등이 있다. 하지만 오버샘플링 방법으로 생성한 데이터가 비현실적인 자료를 생성한다는 점이 지적되었다. 이에 따라 원자료와 유사한 데이터를 생성하기 위해 SMOTE와 후속 연구들에 적용되는 가중치의 분포를 균등분포 대신 비대칭도가 높은 여러 분포를 적용하여 원자료의 근방에서 데이터를 합성하고, 기존의 방법들과 가중치의 분포를 다르게 적용한 방법들의 분류 결과를 비교한다.

주요용어 : Imbalanced data, Class Imbalanced Problem, SMOTE, Over-sampling

[1]48513 부산광역시 용소로 45, 국립부경대학교 통계학과 석사 수료. E-mail: bj2107@pukyong.ac.kr
[2]48513 부산광역시 용소로 45, 국립부경대학교 통계학과 조교수. E-mail: twuhm@pukyong.ac.kr

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 147-150

벤포드의 법칙의 로또복권당첨통계 적용 가능성 연구

이동건[1]

요 약

본 연구는 숫자로 구성된 방대한 데이터에서 임의의 숫자를 뽑을 때, 첫째 자리의 숫자가 고르게 분포되지 않는 현상인 벤포드의 법칙이 로또복권 출현 데이터에 적용될 수 있는지에 대한 가능성을 확인하기 위하여 시행 되었다. 기본적으로 벤포드의 법칙(Benford's Law),은 시험점수나 IQ분포, 사람의 키, 로또당첨번호 등 정규 분포(normal distribution) 및 균일 분포(uniform distribution)를 따르는 데이터에는 적용되지 않는다. 라는 것이 학계의 정설이다. 그러나 실제 로또당첨번호의 통계데이터가 누적 되면 될 수록, 점차 벤포드 법칙과 연관성이 있다고 추정 된다. 이에 본 연구는 2002년 12월 2일부터 2023년 12월 30일까지 약 21년여 간의 역대로또당첨번호의 통계 자료를 수집하여, 회차들을 임의로 선정, 해당 회차의 누적된 표본들과 벤포드의 법칙의 출현 빈도 예측 값을 그래프 및 카이제곱 검정(Chi-squared test), 크레이머-V 계수 (Cramér's V)를 활용하여 각 표본간의 출현 빈도 분석을 통해 벤포드의 법칙의 로또복권당첨통계 적용 가능성을 에 대한 가설을 검증해 보고자 한다. 정확한 분석을 위해 벤포드의 법칙과 임의의 누적 회차 구간 단위를 선정하여 교차분석 해 보았고, 벤포드의 법칙에 로또복권의 번호 수 45개에 맞춘 변수를 적용한 값과 기준점으로 선정한 누적 회차와의 교차분석을 통해 확인 하였다.

주요용어 : 벤포드의 법칙(Benford's Law), 카이제곱 검정(Chi-squared test), 크레이머-V 계수 (Cramér's V), 정규 분포(normal distribution), 균일 분포(uniform distribution), 로또

1. 서론

임의의 방대한 숫자 데이터에서 무작위로 선택한 숫자의 출현 가능성을 볼 때, 모든 숫자의 출현 확률이 동일하다는 것은 기본적인 상식이다. 예를 들어 1부터 9까지의 숫자중 하나를 뽑을 경우, 11.11%로 모두 동일할 것이다. 하지만 Simon Newcomb(1835~1909)은 누적된 데이터에서 임의의 숫자를 뽑을 때, 출현 확률이 동일하지 않은 것을 발견, 자연에서 발견되는 수의 첫째 자리의 숫자가 1일 확률이 가장 높고, 9가 나올 확률은 가장 적게 나오며, 누적된 로그 값을 통해 첫 자리 수 출현 확률을 계산 가능하다는 가설을 1881년 American Journal of Mathmatics에 발표하였으나, 이러한 원리가 당시에는 수학적으로 증명되지 않았기에 학계에서는 제대로 다뤄지지 않고 사장되었다.

이후 미국의 물리학자 Frank Benford(1883~1948)는 1938년 Simon Newcomb의 해당 이론을 공식화하여 강의 넓이, 물리학상수, 분자 중량 등 20여개분야와 2만여개가 넘는 실증 데이터들을 수집 및 분석하여, 다음 공식으로 도출되는 출현 확률과 실증 데이터가 대부분 일치하는 것을 확인 하였다.

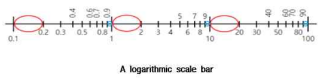
벤포드의 법칙(Benford's law)에서는 어떤 집합에 속한 수들의 첫째 자리가 d (d ∈ {1, ..., 9})가

[1]05553 서울 송파구 올림픽로 212 C동 2207호, 고려대학교 과학기술대학 학부 졸업, Reussirgroup CEO
E-mail : reussirgroup@naver.com

다음의 확률 분포를 따를 때, 해당 집합이 벤포드의 법칙을 따른다고 말한다.

$$P(d) = \log_{10}(d+1) - \log_{10}(d)$$
$$= \log_{10}\left(\frac{d+1}{d}\right)$$
$$= \log_{10}\left(1+\frac{1}{d}\right)$$

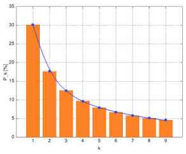
The formula of Benford's law



A logarithmic scale bar

Figure 1. The formula and logarithmic scale of Benford's law

따라서 균일하고 무작위적인 방식으로 임의 숫자 x를 선택하면, 첫자리 숫자가 1이 될 확률이 대략 30.1%이다. 그렇기 때문에 벤포드의 법칙의 첫 자리수의 출현 확률은 아래와 같이 표현된다.



첫자리 숫자	출현 확률
1	0.3010
2	0.1761
3	0.1249
4	0.0969
5	0.0792
6	0.0669
7	0.0580
8	0.0512
9	0.0458
계	1.0000

Figure 2. Probability and distribution of first digit numbers in Benford's law

본 논문에서는 벤포드의 법칙이 적용되지 않는 균일분포(uniform distribution)인 로또당첨번호통계의 적용 가능성에 대해 연구 하였다. 본 논문의 세부 구성은 다음과 같다. 2장에서는 역대 로또당첨번호의 첫 자리 수 출현 빈도와 벤포드의 법칙의 예측 출현 빈도 확률을 통해 생성한 값들과의 시계열 자료 비교 및 검정, 그리고 분석결과를 기술하였다. 3장에서는 결론 도출 및 시사점에 대하여 기술하였다.

2. 연구 방법

본 연구에서는 2002년 12월 2일(1회차)부터 2023년 12월 30일(1100회차)까지 21년간의 역대로또당첨번호 통계 누적자료를 수집하여, 임의의 누적 회차들(100회차 별 누적 당첨번호 자료, 총 11개 항목)을 선정하였다. 해당 회차들의 누적기간동안 발생한 당첨번호들의 첫자리 숫자 별 출현 빈도와 벤포드의 법칙의 값의 간의 시계열자료 비교 및 카이제곱 검정, 크레이머-V계수 분석결과 등을 통해 벤포드 법칙의 로또복권 적용 가능성을 확인해 보고자 한다. 정확한 비교를 위해, 벤포드의 법칙에는 변수(로또번호의 총 개수 제한)를 적용하지 않고, 각 누적 회차의 총계에 대한 단순 첫자리 수의 출현 확률로 자료를 생성하였다.

Table 1에서의 100회차, 500회차, 1100회차의 3개 표본을 선정하여 비교 및 검정 해 보았다.

우선 Figure 3을 보면 시각적인 유사성은 보이지 않는다.

Num / Times	Lotto-100th	Benford-100th	Lotto-200th	Benford-200th	Lotto-300th	Benford-300th	Lotto-400th	Benford-400th	Lotto-500th	Benford-500th	Lotto-600th	Benford-600th	Lotto-700th	Benford-700th	Lotto-800th	Benford-800th	Lotto-900th	Benford-900th	Lotto-1000th	Benford-1000th	Lotto-1100th	Benford-1100th
1	171	211	351	-100th	524	632	717	843	872	1054	1039	1264	1226	1475	1420	1686	1605	1896	1771	2107	1952	2318
2	163	123	319	247	474	370	654	493	842	616	1014	740	1176	863	1324	986	1483	1109	1662	1233	1820	1356
3	186	87	358	175	546	262	692	350	863	437	1034	525	1191	612	1364	699	1548	787	1726	874	1911	962
4	102	68	224	136	326	203	435	271	553	339	666	407	770	475	885	543	994	610	1101	678	1207	746
5	11	55	25	111	44	166	65	222	76	277	98	333	114	388	125	444	138	499	150	554	162	610
6	23	47	35	94	48	140	55	187	70	234	87	281	106	328	125	375	138	421	153	468	174	515
7	18	41	32	81	51	122	65	162	76	203	91	244	114	284	130	325	140	365	157	406	173	447
8	15	36	31	72	49	108	64	143	83	179	98	215	118	251	130	287	140	323	151	358	161	394
9	11	32	25	64	38	96	53	128	65	160	73	192	85	224	97	256	114	289	129	321	140	353
Total	700	700	1400	979	2100	2100	2800	2800	3500	3500	4200	4200	4900	4900	5600	5600	6300	6300	7000	7000	7700	7700

Table 1. A comparison table of the cumulative lotto winning numbers and Benford's law by 100 units

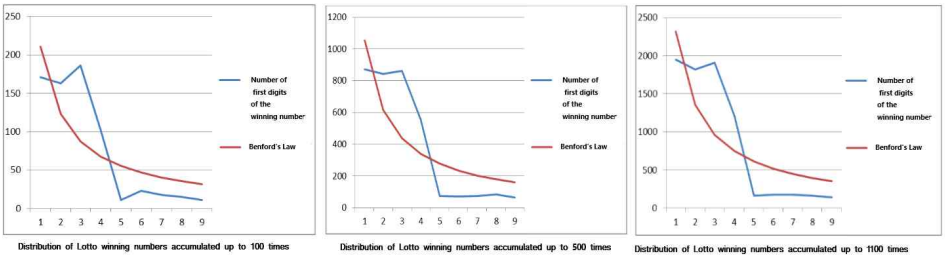


Figure 3. Comparison of Lotto Winning Number Appearance Distribution and Benford Law Graph, which accumulated up to 100th, 500th and 1100th rounds

Contingency Tables

Number	Compare 100th Lotto-100th	Benford-100th	Total
1	171	211	382
2	163	123	286
3	186	87	273
4	102	68	170
5	11	55	66
6	23	47	70
7	18	41	59
8	15	36	51
9	11	32	43
Total	700	700	1400

χ^2 Tests

	Value	df	p
χ^2	118	8	< .001
N	1400		

Nominal

	Value
Phi-coefficient	NaN
Cramer's V	0.290

Contingency Tables

Number	Compare 500th Benford-500th	Lotto-500th	Total
1	1054	872	1926
2	616	842	1458
3	437	863	1300
4	339	553	892
5	277	76	353
6	234	70	304
7	203	76	279
8	179	83	262
9	160	65	225
Total	3499	3500	6999

χ^2 Tests

	Value	df	p
χ^2	579	8	< .001
N	6999		

Nominal

	Value
Phi-coefficient	NaN
Cramer's V	0.288

Contingency Tables

Number	Compare 1100th Benford-1100th	Lotto-1100th	Total
1	2318	1952	4270
2	1356	1820	3176
3	962	1911	2873
4	746	1207	1953
5	610	162	772
6	515	174	689
7	447	173	620
8	394	161	555
9	353	140	493
Total	7701	7700	15401

χ^2 Tests

	Value	df	p
χ^2	1261	8	< .001
N	15401		

Nominal

	Value
Phi-coefficient	NaN
Cramer's V	0.286

Figure 4. Test of Similarity between the 100,500,1100 Lotto Accumulation Attraction Statistics and Benford's Law Values (Chi-square Test and Cramer V Coefficient Analysis Results)

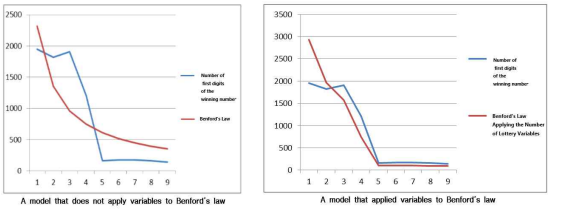
하지만 Figure 4의 분석결과를 보면, 카이제곱 검정의 P-value는 < .001, 크레이머-V 계수는 각각 0.290, 0.288, 0.286으로 P-value가 유의수준 0.05보다 작기 때문에 귀무가설 "역대 로또당첨누적통계와 벤포드의 법칙은 연관성이 없다."을 기각 후, 대립가설인 "역대 로또당첨누적통계와 벤포드의 법칙은 연관성이 있다."을 채택 하였고, 두 변수 간에는 통계적으로 중간 정도의 연관성을 나타낸다는 분석결과를 도출 하였다. 즉, 변수들 사이에 통계적으로 유의미한 관련이 있음을 시사한다.

이번에는 벤포드의 법칙에 로또 번호 45개의 제한 수를 적용한 변수 값들로 분석 하였다. 변수 값을 구하는 방법은 다음과 같다. 1~100까지의 숫자를 기준으로, 벤포드의 법칙을 적용한 첫 자리 수의 예측 출현 빈도 확률 값들을 각각 나누어 구한 후, 나머지 숫자들(46~100)의 전체 출현 확률을 뺀다. 이후 나머지 숫자별 출현 확률들을 더하고, 그 확률값을 45로 나누어 가중값 a를 도출한다. 이후 남아있는 1~45까지의 숫자에 벤포드의 법칙의 비율가중값 b를 환산하여 각 첫 자리 숫자 출현 확률에 더하여, 번호별 최종 확률 값 c를 도출한 후 적용한다. 예를 들어, 벤포드의 법칙에 따라 첫 자리 수 1이 나올 확률인 30.1%를 적용하지 않고, 임의 가중값 a(0.00001~20.999999 사이 값)를 적용하여 임의 확률 값 b(30.10%~50.981000% 사이 값)을 기존 벤포드의 법칙 확률 값에 더해 적용하는 방식으로, 각각의 최종 번호별 확률 값인c(n, n=1…9)를 생성하여 비교 분석 하였다.

Table 1의 1100회차 누적 통계자료와 변수(로또 번호수 45개)를 적용한 벤포드의 법칙 생성 값을 기준으로 비교했을 때, Figure 5를 보면 Figure 3 보다 더 시각적인 유사성을 확인 할 수 있다.

또한 Figure 6의 분석결과를 보면, 카이제곱 검정의 P-value는 < .001, 크레이머-V 계수는 0.165로

P-value가 유의수준 0.05보다 작기 때문에 귀무가설 “역대 로또당첨누적통계와 로또번호개수의 변수를 적용한 벤포드의 법칙은 연관성이 없다.”을 기각 후, 대립가설인 “역대 로또당첨누적통계와 로또번호개수의 변수를 적용한 벤포드의 법칙은 연관성이 있다.”을 채택 하였고, 두 변수 간에는 통계적으로 약한 정도의 연관성을 나타낸다는 분석결과를 도출 하였다. 즉, 변수들 사이에 통계적으로 유의미한 관련이 있음을 시사한다.



A model that does not apply variables to Benford's law　　A model that applied variables to Benford's law

Figure 5. Comparison of models with regular Benford's law and models with variable values (based on the distribution of 1100 cumulative lotto winning numbers)

Contingency Tables

Number	Compare Specific Var Lotto-1100th	BenfordSV-1100th	Total
1	1952	2928	4880
2	1820	1964	3784
3	1911	1571	3482
4	1207	744	1951
5	162	102	264
6	174	102	276
7	173	99	272
8	161	97	258
9	140	93	233
Total	7700	7700	15400

χ^2 Tests

	Value	df	p
χ^2	422	8	< .001
N	15400		

Nominal

	Value
Phi-coefficient	NaN
Cramer's V	0.165

Figure 6. Test of similarity between the value of applying the number of lotto variables to Benford's law and the number of lotto winning statistics. (Chi-square test and Kramer V coefficient analysis results)

3. 결론

위의 결과들을 고려해 볼 때, 벤포드의 법칙은 로또당첨번호에도 적용될 수 있다는 가능성을 시사한다. 본 연구를 계기로 향후 벤포드의 법칙의 효용범위가 확장될 수 있기를 기대한다.

References

Arno berger and Theodore P. Hill, (2011). Benford’s Law Strikes Back: No Simple Explanation in Sight for Mathematical Gem,

Berger, arno; hill, theodore P. (30 June 2020). "The mathematics of Benford’s law: a primer". Stat. *Methods*. 30(3), 779 – 795.

Choi, J. H. (2021). A Study on the Effectiveness of Fraudulent Accounting Investigation Using the Benford Law. *A Criminal Investigation Study*, 7(2), 5-24.

Frank benford (March 1938). "The law of anomalous numbers". *Proc. Am. Philos. Soc.* 78(4), 551 – 572.

Hill, theodore (1995). "A Statistical Derivation of the Significant-Digit Law". Statistical Science..

Miller, steven J. ed. (9 June 2015). Benford’s Law: Theory and Applications. Princeton University Press., 309.

Paul H. kvam, brani vidakovic, Nonparametric Statistics with Applications to Science and Engineering,, 158.

Pimbley, J. M. (2014). "Benford’s Law as a Logarithmic Transformation" (PDF).

Simon newcomb (1881). "Note on the frequency of use of the different digits in natural numbers". *American Journal of Mathematics*. 4 (1/4),39 – 40.

Weisstein, eric W. (7 June 2015). "Benford’s Law". MathWorld, A Wolfram web resource.

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 151-151

A Unified Regularization Paths of L1-penalized SVM models R-package: L1svmpath*

Hyungwoo Kim[1], *Seung Jun Shin*[2]

Abstract

Support vector machine (SVM) is a powerful binary classification tool and has gained massive popularity in many applications due to its high accuracy and flexibility. In this article, we develop an R-package 'L1svmpath' that effectively computes the entire regularization paths for three types of SVM models combined with L1 norm penalty: L1-penalized SVM (Zhu et al., 2003), L1-penalized ROC curve-optimizing SVM (Kim et al., 2021) and L1-penalized fraud detection SVM (Park et al., 2023). The standard SVM finds the decision boundary by maximizing the margin of the classifier to increase the accuracy. The ROC curve-optimizing SVM, referred to as ROCSVM, directly maximizes the area under the ROC curve (AUC) by the hinge loss used for SVM. In fraud detection SVM, the term 'fraud detection' refers to a process identifying and preventing unusual data points in the dataset, and it finds the decision boundary by detecting outliers. Three path algorithms performed by 'L1svmpath' are extremely fast compared to the existing things that solve the optimization problem via linear programming, especially in a large number of samples and variables.

Keywords : Support Vector Machine, ROC curve, Fraud Detection, L1 norm penalty, Regularization Paths

---

*This paper was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (RS-2023-00242528).

[1]Assistant Professor, Department of Statistics and Data Science, Pukyong National University, Busan, 48513, Republic of Korea, E-mail: khw7682@pknu.ac.kr

[2](Corresponding Author) Associate Professor, Department of Statistics, Korea University, Seoul, 02841, Republic of Korea, E-mail: sjshin@korea.ac.kr

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 153-153

Deep neural network based non-crossing multiple quantile regression estimator

Jungmin Shin[1], *Seung Jun Shin*[2], *Sungwan Bang*[3]

Abstract

In this paper, we present the deep neural network based non-crossing multiple quantile regression estimator (DNN-NMQR), which uses a deep neural network structure to solve the multiple quantile regression problem. When estimating multiple quantiles, our method uses the structural characteristics of DNN to improve the estimation accuracy by accommodating a shared strength among different quantiles. In addition, this method effectively addresses the quantile crossing issue by applying the penalization method. To preserve theoretical completeness, we introduce a convolution-type quadratic smoothing loss function to ensure that the objective function remains differentiable across the support. In addition, we briefly discuss the convergence analysis of DNN-NMQR, based on the idea of neural tangent kernel. We verify the proposed method through numerical experiments and real-data analysis.

Keywords : Deep neural network, multiple quantile regression, non-crossing penalization, neural tangent kernel.

[1]Ph.D Candidate Department of the Statistics, Korea University, 145 Anam-ro, Seoungbuk-Gu, 02841, Seoul, Korea. E-mail: jungminshin@korea.ac.kr
[2]Associate Professor, Department of the Statistics, Korea University, 145 Anam-ro, Seoungbuk-Gu, 02841, Seoul, Korea. E-mail: sjshin@korea.ac.kr
[3](Corresponding Author : Professor, Department of Mathematics, Korea Military University, 574 Hwarang-ro, Nowon-Gu, 01805, Seoul, Korea. E-mail: wan1365@gmail.com

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 155-155

Nonparametric Variable Selection for Mixed Model

Yujin Hwang[1], *Jun Song*[2]

Abstract

In response to the challenges posed by high-dimensional datasets incorporating both numerical and categorical variables, this research focuses on developing theories and methods for nonparametric variable selection within mixed models. Our methodology employs b-splines for numerical data and introduces adaptive group lasso, focusing on nonparametric variable selection. Our approach, rooted in the ADMM algorithm, ensures adaptability to diverse relationships, including non-linear ones. We address the intricacies of high-dimensional scenarios, particularly emphasizing cases where the number of categories expands as the sample size grows, such as in document datasets. Ongoing work involves comprehensive simulations and real data analysis. The developed methodology is complemented by non-asymptotic error bounds, applicable in finite sample scenarios, ensuring the practical relevance of our findings. This work contributes a versatile tool for uncovering complex relationships in mixed models, offering a refined approach to variable selection in high-dimensional settings.

Keywords : adaptive group lasso, high-dimensional analysis, nonparametric approach, variable selection

[1]Graduate student, Department of Statistics, Korea University, 145 Anam-ro, Seongbuk-gu, Seoul, 02841, Korea. E-mail: yjh32@korea.ac.kr
[2](Corresponding Author) Associate Professor, Department of Statistics, Korea University, 145 Anam-ro, Seongbuk-gu, Seoul, 02841, Korea. E-mail: junsong@korea.ac.kr

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 157-157

slcm: An R Package for Multiple Latent Class Variables

Youngsun Kim[1], *Hwan Chung*[2]

Abstract

Latent Class Analysis (LCA) is a prominent mixture model utilized for population segmentation. However, beyond simple population segmentation, there are also advanced versions of this model. These versions handle multiple variables that can represent different aspects of individuals or changes over time. For instance, Latent Transition Analysis connects these variables in a sequence over time. Other approaches, like Latent Class Profile Analysis (LCPA) and Joint Latent Class Analysis (JLCA), organize these variables in layers, adding more complexity to the analysis. There are even more intricate models that combine these approaches. Moreover, there are some complicated models combining those several models. In this research, we introduce an R package named 'slcm.' This package is designed to estimate a variety of these complex models. It's built to be user-friendly, making it easier for people to set up and understand the complex latent structures these models use. 'slcm' also includes tools for checking how well your model is working and adjusts for outside influences that might skew your results. This package is a helpful tool for researchers and practitioners who need advanced methods for analyzing population data. It simplifies these sophisticated analyses, making them more accessible and practical for a wide range of studies.

keywords : latent class analysis, R package, EM algorithm, multivariate analysis.

[1]Department of Statistics, Korea University, Seoul, 02841, Korea. E-mail: kim0sun@korea.ac.kr
[2](Corresponding author) Department of Statistics, Korea University, Seoul, 02841, Korea.
E-mail: hwanch@korea.ac.kr

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 159-159

CUDA: Convolutional Unet based Defense Architecture for Adversarial Patch Attack

김희민[1], 김병찬[1], 곽일엽[2]

요 약

심층 신경망의 고도화된 기술적 향상과 함께 작은 왜곡만으로도 심층 신경망을 공격하는 적대적 공격 또한 발전했다. 작은 왜곡만으로도 심층 신경망의 올바른 예측을 뒤엎을 수 있는데 이를 적대적 공격이라 하며 이러한 공격이 자율 주행 상황과 같이 안전이 중요한 분야에 가해질 시 큰 위협이 될 수 있다. 최근 스티커 형태로 이미지에 부착 시 객체 탐지 모델의 오분류를 유도하는 적대적 패치 공격이 현실 세계에 악용되었고 이에 대한 방어 기법의 중요성 또한 대두되었다. 적대적 공격에 대한 방어 기법은 크게 세 가지로 분류되는데 전처리(Pre-processing), 처리중(In-processing), 후처리(Post-processing) 이다. 본 연구에서는 객체 탐지 모델에 이미지가 입력값으로 들어가기 전 단계에서 적대적 패치의 공격을 방어하는 전처리 방어 기법을 제안한다. 적대적 패치는 모든 입력에 대한 예측을 좌우하도록 이미지의 특정 영역에 비정상적으로 큰 영향을 주는 섭동이며, 이는 특징 맵에서도 확인할 수 있다. 즉, 적대적 패치가 부착된 곳에 훨씬 큰 특징 벡터를 갖게되며 이는 곧 예측 결과를 왜곡시키게 된다. 이러한 적대적 패치의 특성을 학습하여 원본 이미지로 복원하는 모델 CUDA: Convolutional Unet based Defense Architecture를 제안한다. CUDA는 인코더-디코더(encoder-decoder) 구조의 모델로, 공격 받은 이미지의 지역적 패턴을 원본 이미지의 분포에 더 가까운 패턴으로 변환함으로써 공격 받은 이미지를 재구성하도록 훈련되었다. 해당 방법론은 타 방법론과는 달리 적대적 패치의 영역을 지역화하고 복원까지 수행하는 One-stage 방법론으로 구현이 용이하다. 뿐만 아니라, 현존하는 다양한 적대적 패치의 특성을 함께 학습하여 적대적 패치의 모양과 개수에 상관없이 강건한 방어가 가능하다. 타 방법론과의 비교 실험에 사용된 객체 탐지 모델은 Yolov5이며, COCO dataset, VisDrone dataset, Argoverse dataset을 사용하여 해당 방어 기법의 높은 성능을 입증하였다.

주요용어 : 적대적 패치 공격, 적대적 공격 방어.

[1]06974 서울특별시 동작구 흑석로 84(흑석동), 중앙대학교 통계학과 석사과정. E-mail: heernink@cau.ac.kr
[1]06974 서울특별시 동작구 흑석로 84(흑석동), 중앙대학교 통계학과 석사과정. E-mail: moch1996@cau.ac.kr
[2](교신저자) 06974 서울특별시 동작구 흑석로 84(흑석동), 중앙대학교 통계학과 교수. E-mail: ikwak2@cau.ac.kr

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 161-161

Developing Robust Performance in Korean Speech Recognition Using JASPER-GRU with Similarity Loss Function*

김병찬[1], 곽일엽[2]

요 약

딥러닝 기술의 지속적인 발전을 통해 음성 데이터에 대한 이해와 음성 처리 분야에서의 성능은 향상되고 있다. 음성데이터를 다루는 분야 중 하나인 음성인식 과제는 이미지 처리에서 주로 사용되는 CNN의 구조와 자연어 처리에서 주로 RNN의 모형을 활용하여 실행된다. 그러나, 음성데이터, 신호데이터의 형태를 포착하기 어려운 특성으로 인하여, 학습하는 데 여러 어려움이 존재한다. 이에 따라, 언급한 음성인식의 모형은 학습데이터에서의 모델 성능과 외부의 테스트 데이터에서의 모델 성능에 차이가 발생하여 보편적인 데이터에 대한 일반화 성능 문제로 이어진다.. 따라서 본 논문에서는 한국어 음성인식 학습 시에 외부데이터에 대한 강건성 성능 확보를 보장하기 위하여 Reconstruction Loss가 적용된 JASPER-GRU 모형을 제안한다. 일반적인 음성인식의 모형학습에 사용되는 CTC Loss 뿐만이 아니라, JASPER 모형의 출력값 시퀀스를 복제하여, 하나의 시퀀스의 일부를 특정한 값으로 변환 후 학습을 진행한다. 이 두 시퀀스는 순환신경망을 통과하여 시간축에 대해 변환이 이루어진 값에 대한 정답을 맞추는 Reconstruction loss를 줄여나가며 일반화 성능을 보정한다. 학습에 사용된 데이터셋은 한국어 상담음성 데이터셋과 자유발화 데이터셋인 KsponSpeech Dataset을 사용하여 실험하였다. 모델의 일반화 성능을 평가하기 위해 학습데이터와는 다른 데이터셋을 활용하여 평가하거나 노이즈를 추가하여 실제와 유사한 환경에서 실험을 진행하였다. 음성인식 혹은 이미지 데이터처리에서 주로 사용되는 모델과 성능을 비교하였고, 비교군 대비 최대 약 26% 의 낮은 CER 지표를 얻을 수 있음을 입증하였다.

*본 논문은 중앙대학교의 지원을 받아서 연구된 것임.
[1]06974 대한민국 서울특별시 동작구 흑석로 84, 중앙대학교 통계학과 석사과정. E-mail: moch1996@cau.ac.kr
[2](교신저자) 06974 대한민국 서울특별시 동작구 흑석로 84, 중앙대학교 통계학과 조교수.
E-mail: ikwak2@cau.ac.kr

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 163-163

MCHeart: Multi-Channel-Based Heart Signal Processing Scheme for Heart Noise Detection Using Deep Learning*

Soyul Han[1], *Il-Youp Kwak*[2]

Abstract

In this study, we constructed a model to predict abnormal cardiac sounds using a diverse set of auscultation data collected from various auscultation positions. Abnormal heart sounds were identified by extracting features such as peak intervals and noise characteristics during systole and diastole. Instead of using raw signal data, we transformed them into log-mel 2D spectrograms, which were employed as input variables for the CNN model. The advancement of our model involves integrating a deep learning architecture with feature extraction techniques based on existing knowledge of cardiac data. Specifically, we propose a multi-channel-based heart signal processing (MCHeart) scheme, which incorporates our proposed features into the deep learning model. Additionally, we introduce the ReLCNN model by applying residual blocks and MHA mechanisms to the LCNN architecture. By adding murmur features with a smoothing function and training the ReLCNN model, the weighted accuracy of the model increased from 79.6% to 83.6%, showing a performance improvement of approximately 4% point compared to the LCNN baseline model.

Keywords : Heart Murmur Detection; Smart Healthcare; Convolutional Neural Network; Multi-head Attention; Deep Learning.

---

*This work was supported by a National Research Foundation of Korea (NRF) grant funded by the Ministry of Science and ICT (RS-2023-00208284, 2020R1C1C1A01013020).

[1]Ph.D. student, Department of Statistics, Chung-Ang University, Seoul, 06974, Republic of Korea. E-mail: soyul5458@cau.ac.kr

[2](Corresponding Author) Professor, Department of Applied Statistics, Chung-Ang University, Seoul 06974, Republic of Korea; E-mail: ikwak2@cau.ac.kr

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 165-165

객체 추적 알고리즘을 활용한 실시간 쓰레기 무단투기 감지 시스템

김희민[1], 김병찬[2], 강민정[3], 남현지[4], 곽일엽[5]

요 약

쓰레기 무단 투기가 자주 발생하는 지역에는 CCTV가 대부분 설치돼 있지만 무단 투기 문제가 해결되지 않고 있다. 현재 시스템은 CCTV 근처를 지나가는 사람이 있으면 경고음을 발생시키는 방식인데 이러한 시스템은 쓰레기를 버리지 않아도 무분별하게 알람이 울리는 문제가 발생하고 실효성이 떨어진다고 판단된다. 이에 따라 CCTV 영상을 통해 실시간으로 사람이 무단 투기하는 장면을 포착할 경우, 경고음이 발생하도록 연구를 진행하였다. 본 연구에서는 yolov5 모델을 이용하여 다양한 종류의 쓰레기와 사람을 학습한 후 두 가지 객체만 탐지(Object Detection)하도록 한다. 쓰레기 무단 투기 장면을 포착하는 방법은 Pose estimation 기술과 객체 추적(Object Tracking) 알고리즘을 사용한다. Pose estimation 기술은 영상에 포함된 인물을 탐지하여 인체 각 부위의 위치를 식별하고 부위를 연결하는 선을 구하는 기술이다. 객체 추적(Object Tracking) 알고리즘은 프레임별 추적을 수행하고 시간의 흐름에 따라 객체가 있는 위치 기록을 남기는 알고리즘이다. 다양한 객체 추적 모델이 존재하는데 ID Switching 문제에 강건한 Deep SORT 모델을 이용하였다. 위 두 가지 기술을 활용하여 손목의 좌푯값과 쓰레기 좌푯값 사이의 거리가 한계점을 넘어가면 투기하는 장면이라고 판단한다. 본 연구를 통해 실시간 모니터링이 가능하며 쓰레기 무단 투기에 대한 예방적 효과를 가져올 수 있다.

주요용어 : Object detection, Object tracking, Pose estimation, Garbage dumping detection

[1]06974 서울특별시 동작구 흑석로 84, 중앙대학교 통계학과 석사과정. E-mail: heernink@cau.ac.kr
[2]06974 서울특별시 동작구 흑석로 84, 중앙대학교 통계학과 석사과정. E-mail: moch1996@cau.ac.kr
[3]06974 서울특별시 동작구 흑석로 84, 중앙대학교 통계학과 석사과정. E-mail: tg00383@cau.ac.kr
[4]06974 서울특별시 동작구 흑석로 84, 중앙대학교 통계학과 석사과정. E-mail: namhj1004@naver.com
[5](교신저자)06974 서울특별시 동작구 흑석로 84, 중앙대학교 통계학과 전임교수. E-mail: ikwack2.cau.ac.kr

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 167-168

웰에이징을 위한 경제교육의 필요성 연구*

안상윤[1], 김설희[2], 김광환[3]

요 약

한국 사회는 갈수록 노인 빈곤율이 높아지고 경제적 양극화가 증폭되고 있는 가운데 노인 경제의 양극화 율도 40%를 넘어서는 등 노인들의 빈곤이 중요한 사회문제로 부각되고 있다. 본 연구는 이와 같은 문제를 해소하고 건강한 사회발전에 기여하기 위해 노인요양복지 사업현장에서 각종 교육을 담당하고 있는 교육자들에게 노인의 경제복지를 향상 목적의 체계적인 교육을 수행하기 위한 방안을 모색하는데 있다. 이 연구는 노인들의 경제력을 높이기 위해 경제교육 프로그램을 효과적으로 기획하고 수행하는데 필요한 기초자료를 제공하기 위한 서술적 조사연구이다. 자료 수집은 한국갤럽을 통해 이루어졌으며, 총 응답자는 37명이었으며 현재 근무분야에서 일한 기간은 평균 16.6년이었다. 수집된 자료는 다음과 같이 분석하였다. 대상자의 일반적 특성은 빈도와 백분율, 경제프로그램 운영중요도와 실행도, 경제교육프로그램 중요도와 지식정도, 재직기간에 따른 업무 중요도 및 수행도 은 평균과 표준편차로 분석하였으며, 재직기간에 따른 경제프로그램 운영의 중요도와 실행도, 경제교육프로그램 중요도와 지식정도, 재직기간에 따른 업무 중요도 및 수행도는 회귀분석을 실시했다. 교육자의 경제교육 관련 영역별 인식을 살펴보기 위해서 경제교육 프로그램 운영 시 각 항목별 중요도 및 수행도에 대해 IPA(Importance-Performance Analysis)를 통해 분석한 결과 '프로그램계획 수립'이 중점 개선영역으로 나타났다. 경제 교육내용에 대한 항목별 중요도 및 지식정도를 분석한 결과 '은퇴 후 경제생활'이 경쟁우위영역으로 나타났고, '정부의 경제지원정책'이 중요도는 높으나 이에 대한 지식정도는 평균 수준에 불과하여 개선이 요구된다고 하겠다. 또한 경제교육 프로그램 제공 및 지도·관리에서 중요도와 수행도 모두 낮아 개선영역으로 나타났다. 이와 같은 결과로 볼 때, 사회적으로 조기 경제교육이 절실하며 노인들에 대한 합리적 투자와 소비에 대한 교육프로그램 개발이 요구된다.

주요용어 : economic education, retirement, poverty ratio of old, investment, consumption.

1. 서론

현대 복지사회에서 은퇴 노인들이 경제력을 기반으로 한 웰에이징 구현은 중요한 복지정책의 과제가 되고 있다. 하지만 한국 사회에서 노인 빈곤율은 OECD 국가 중 최고이며 이에 따른 노인들의 건강악화 및 자살 등은 중요한 사회문제로 부각되고 있다. 은퇴 후를 대비한 실질적이고 체계적인 경제교육이 절실히 필요하다고 하겠다. 은퇴 후에 노인들은 경제교육을 절실히 필요로 하

*본 논문은 한국연구재단의 연구과제로 수행되었음(NRF-2020S1A5C2A04092504).
[1]35365 대전광역시 서구 관저동로 158 건양대학교 병원경영학과교수. E-mail: greahn@konyang.ac.kr
[2]35365 대전광역시 서구 관저동로 158 건양대학교 치위생학과교수. E-mail: ableksh@kycu.ac.kr
[3](교신저자) 35365 대전광역시 서구 관저동로 158 건양대학교 병원경영학과 교수. E-mail: kkh@konyang.ac.kr

고 있는 것으로 인식은 하고 있으나 경제상황에 대한 이해와 경제활동과 관련된 합리적 의사결정 능력을 함양하기 위한 교육제도는 미비한 상황이다. 따라서 이 연구는 노인들과 관계하고 있는 각종 기관에서 노인의 복지와 교육을 담당하고 있는 인력들이 노인들에게 효과적으로 경제교육을 실시할 수 있는 효과적인 방안을 마련하기 위한 기초조사를 하는데 그 의의가 있다.

Reference

Agnes Szabo, Joanne Allen, Fiona Alpass, Christine Stephens, “Loneliness, socio-economic status and quality of life in old age: the moderating role of housing tenure”, *Ageing & Society*, 2019, 39, 998-1021

Ajit Shan, Ravi Bhatt, Sheena MdKenzie, Chris Koen, “Relationship between elderly suicide rates and life expectancy, and markers of socio-economic status and health: multivariae analysis”, *Journal of Chinese Clinical Medicine*, 2009, 4(4), 213-217

Bussarawan Teerawichitchainan, John Knodel, “Economic Status and Old-Age Health in Poverty-Stricken Myanmar”, *Journal of Aging and Health*, 2015, 27(8), 1462-1484

G. Lawrence Atkins, “The Economic Status of the Oldest Old”, *Health and Society*, 1985, 63(2), 395-419

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 169-172

건강증진 및 만성질환관리 교육활동에 대한 웰에이징 교육전문가의 인식 연구*

임효남[1], 황혜정[2], 김광환[3]

요 약

본 연구는 웰에이징 교육전문가를 대상으로 건강증진 및 만성질환관리의 교육활동에 대한 인식을 조사하여 건강증진 및 만성질환 관리 교육프로그램 기획에 필요한 기초자료를 제공하기 위해 시행한 서술적 조사연구이다. 2022년 11월 2일부터 2023년 1월 3일까지 건강증진 및 만성질환관리 관련 웰에이징 교육전문가 125명을 대상으로 구조화된 설문지를 이용해 설문조사를 실시하였다. 건강증진 및 만성질환관리의 교육운영에 필요한 항목의 중요도 및 실행도에 대해 IPA(Importance-Performance Analysis)를 통해 분석한 결과 '교육프로그램 운영을 위한 전문가 인력풀(pool) 구성'이 중점개선영역으로 나타났다. 교육내용에 대한 항목별 중요도 및 지식정도를 분석한 결과 '만성질환 예방 및 관리'가 경쟁우위영역으로 나타났고, 건강증진행위의 항목별 중요도 및 지식정도에서는 '건강진단 및 상담', '일상생활관리'가 중점개선영역으로 이동할 가능성이 있으며, '운동 및 활동'이 경쟁우위영역으로 나타났다. 건강증진 및 만성질환관리 교육이 가장 필요한 시기는 '중년기'가 가장 높았고, 교수학습활동의 구성 요소 중 가장 중요한 것은 '교육내용'이었으며, 적절한 교육방법은 '체험이 동반된 강의'로 나타났다. 건강증진 및 만성질환관리 교육을 기획할 때 이와 같은 결과를 바탕으로 교육대상 및 교육내용을 설정하고 보완하여 체계적이고 통합적인 교육프로그램을 구성하여야 하겠다.

Keyword : Well-aging, Health promotion, Chronic disease management, Education expert, Health promotion behavior.

1. 서론

우리나라는 2025년 초고령 사회로 진입하고, 2060년에는 고령인구의 비율이 전체인구의 약 44%까지 증가할 전망이다(Statistics Korea, 2021). 노인인구가 증가함에 따라 만성질환의 유병률 또한 증가하게 되며 의료기술의 발달로 기대수명이 증가함에 따라 만성질환을 앓고 살아가는 기간이 더 늘어나고 있다(Kim, Lim, 2017). 만성질환은 건강증진행위의 실천과 밀접한 관계가 있기 때문에 생활습관병이라고도 한다(Song, Kim, 2020). 만성질환은 생활 습관 개선 등의 노력을 통해 예방하고 관리하는 능력이 필요한데 이러한 행위가 습관이 되기 위해서는 개인의 노력만으로는 한계가 있으므로 통합적이고 체계적인 교육을 받는 것이 중요하다(Jang, Shin, 2020; Kim, et al., 2017). 이에 본 연구에서는 건강증진 및 만성질환 관리와 관련하여 웰에이징 교육전문가를 대상으로 교육운영, 교

*본 논문은 한국연구재단의 연구과제로 수행되었음(NRF-2020S1A5C2A04092504).
[1]35365 대전광역시 서구 관저동로 158 건양대학교 간호학과 부교수. E-mail: hnlim@konyang.ac.kr
[2]35365 대전광역시 서구 관저동로 158 건양사이버대학교 보건의료복지학과 교수. E-mail: hhj@kycu.ac.kr
[3](교신저자) 35365 대전광역시 서구 관저동로 158 건양대학교 병원경영학과 교수. E-mail: kkh@konyang.ac.kr

육내용, 건강증진행위 내용의 중요성과 지식정도 및 교육활동의 인식정도에 대해 확인하고 건강증진 및 만성질환관리 교육프로그램을 기획하여 한국형 웰에이징 모델 개발을 위한 연구의 기초자료를 제공하고자 하였다.

2. 본론

교육기관과 비교육기관에서 건강증진 및 만성질환관리 교육이 가능한 웰에이징 교육전문가를 대상으로 2022년 11월 2일부터 2023년 1월 3일까지 온라인 설문업체(한국갤럽) 홈페이지에 연구목적과 내용에 대한 모집공고를 게시하고 본 연구의 목적을 충분히 이해하고 동의한 자 125명을 대상으로 구조화된 질문지를 이용한 온라인 조사를 실시하였다. 수집하는 자료는 대상자의 일반적 특성, 웰에이징을 위한 건강증진 및 만성질환관리 교육운영, 교육 내용, 건강증진행위 내용, 교육활동의 인식정도를 확인하였다.

건강증진 및 만성질환 관리를 위한 교육운영 시 필요한 항목의 중요도 및 실행도를 살펴본 결과 중요도가 가장 높은 항목은 ‘체계적인 프로그램의 제공’이었으며, 실행도가 가장 높은 항목은 ‘교육프로그램 계획 수립’이었다. 교육운영 시 필요한 각 항목에 대해 중요도 및 실행도를 교차 분석하여 개선 우선순위를 결정하는 IPA 분석을 시행한 결과, ‘교육프로그램 운영을 위한 전문가 인력풀(pool) 구성’은 중요도는 높으나 실행도가 낮아 중점개선영역으로 나타났으며, ‘체계적인 교육프로그램의 제공’, ‘교육프로그램 지도 및 관리(강사훈련)’, ‘교육프로그램 계획 수립’은 중요도 및 실행도가 모두 높아 경쟁우위영역으로 나타났다. ‘행정기관 및 의료기관과의 협력 체계 구축 및 운영’, ‘교육프로그램 홍보를 위한 정보매체활용 기술 능력(홍보)’는 중요도 및 실행도 모두 낮아 개선영역으로 나타났다

건강증진 및 만성질환 관리를 위한 교육내용의 항목별 중요도 및 지식정도를 살펴보면, ‘만성질환 예방 및 관리’가 중요도 및 지식정도에서 가장 높았다. ‘영양관리’ 및 ‘수면과 휴식’은 중요도가 가장 낮고, ‘영양관리’는 지식정도가 가장 낮았다. 교육내용별 중요도 및 지식정도에 대해 IPA 분석을 시행한 결과, ‘신체활동 및 운동’, ‘만성질환 예방 및 관리’는 중요도 및 지식정도 모두 높아 경쟁우위영역으로 나타났다. ‘안전과 환경’은 지식정도는 높으나 중요도가 낮아 우위영역으로 나타났고, ‘수면과 휴식’, ‘영양관리’는 중요도 및 지식정도 모두 낮아 개선영역으로 나타났다.

건강증진행위의 항목별 중요도 및 지식 정도를 조사한 결과, 건강증진 행위 중 ‘운동 및 활동’이 중요도 및 지식정도가 모두 가장 높았으며, ‘음주 및 흡연’은 중요도 및 지식정도가 모두 가장 낮았다. 건강증진 행위별 중요도 및 지식정도에 대해 IPA 분석을 시행한 결과, ‘건강진단 및 상담’, ‘일상생활관리’는 지식정도는 평균 수준이나 중요도가 높아 중점개선영역으로 이동할 가능성이 있으며, ‘운동 및 활동’은 중요도 및 지식정도가 모두 높아 경쟁우위영역으로 나타났다. ‘영양 및 식생활’은 지식정도는 평균 수준이나 중요도가 낮아 우위영역으로 이동할 가능성이 있으며, ‘음주 및 흡연’은 중요도 및 지식정도 모두 낮아 개선영역으로 나타났다.

건강증진 및 만성질환 관리교육이 웰에이징의 신체적 건강증진에 미치는 효과의 정도를 5점 만점으로 확인한 결과 4.4점이었다. 교육이 필요하다고 생각되는 시기는 ‘중년기(67.2%)’ 응답 비율이 가장 높고, 다음으로 ‘성인기(54.4%)’, ‘장년기(44.0%)’ 등의 순이었다. 교육에 필요한 정보 습득 경로는 ‘교육기관(56.0%)’, ‘대중매체(52.0%)’, ‘의료기관 및 보건기관(50.4%)’ 등의 순으로 나타났다.

교수학습활동 구성 요소는 ‘교육 내용(67.2%)’이 가장 중요하다고 하였으며, 다음으로 ‘교육방법 및 형태(16.0%)’, ‘교육강사(8.0%)’ 등의 순이었다. 적절한 교육 방법은 ‘강의+체험(66.4%)’의 응답 비율이 가장 높고, 다음으로 ‘현장 강의(20.0%)’, ‘사이버 교육(6.4%)’ 등의 순이었다. 교육이 이루어져야 한다고 생각하는 기관을 확인한 결과, ‘교육기관(40.0%)’ 응답 비율이 가장 높고, 다음으로 ‘의료기관 및 보건소(23.2%)’, ‘지역기관(22.4%)’ 등의 순이었다.

교육과정 진행 시, 1회 수업의 적정 수강 인원은 평균 17.4명이었고, 평생 기준, 건강증진 및 만성질환 관리교육의 적정 횟수는 평균 11.6회였으며. 1회 수업의 적정 시간은 평균 4.3시간이었다.

교육전문가 양성 프로그램의 총 적정 시간을 확인한 결과, 평균 23.8시간이며, 교육전문가에게 필요하다고 생각하는 자격 또는 역량을 확인한 결과, ‘관련 지식 함양(22.6%)’과 관련된 응답이 가장 많았으며, 다음으로 ‘관련 자격/면허 여부(17.4%)’, ‘특정 역량 함양(16.4%)’, ‘관련 경력/경험 보유(11.8%)’ 등의 순이었다. 구체적인 자격 또는 역량으로는 ‘건강증진 및 만성질환 관련 지식’, ‘보건 관련 전공자/보건 의료 면허증 소지자’, ‘전문적 지식’, ‘강의 능력’ 등이 언급되었다.

3. 결론

본 연구는 웰에이징 교육전문가를 대상으로 건강증진 및 만성질환관리 교육활동에 대한 인식을 조사함으로써 체계적인 교육안을 개발하는데 기초자료를 제공했다는 것에 의의가 있다. 교육운영을 위해 필요한 항목의 중요도 및 실행도를 교차 분석한 결과 중요도는 높으나 실행도가 낮은 항목은 ‘교육프로그램 운영을 위한 전문가 인력풀 구성’으로 나타났다. 또한 중요도가 가장 높은 항목은 ‘체계적인 프로그램의 제공’이었다. 교육내용에 대한 항목별 중요도 및 지식정도를 분석한 결과 ‘만성질환 예방 및 관리’가 중요도 및 지식정도 모두 가장 높아 경쟁우위영역으로 나타났고, 건강증진행위의 항목별 중요도 및 지식정도에서는 ‘건강진단 및 상담’, ‘일상생활관리’는 지식정도는 평균 수준이나 중요도가 높아 중점개선영역으로 이동할 가능성이 있으며, ‘운동 및 활동’은 중요도 및 지식정도가 모두 높아 경쟁우위영역으로 나타났다. 건강증진 및 만성질환관리 교육이 가장 필요한 시기는 ‘중년기’가 가장 높았고, 성인기, 장년기 순이었다. 교수학습활동의 구성 요소 중 가장 중요한 것은 ‘교육 내용’이었으며, 적절한 교육방법은 ‘체험이 동반된 강의’로 나타났다. 이와 같은 결과를 바탕으로 건강증진 및 만성질환관리 교육프로그램 기획 시 교육요구도에 부합하는 교육내용이 충실한 프로그램이 필요하고, 교육을 운영할 수 있는 교육전문가 양성이 우선시 되어야 함을 확인하였다. 웰에이징을 위한 건강증진 및 만성질환관리 교육은 중년기에 초점을 두되, 성인기로 교육대상을 확대하여 성공적인 노년기를 대비할 수 있도록 하는 것이 요구된다. 향후 통합적이고 체계적인 생애맞춤형 교육프로그램 개발 시 이와같은 사항을 고려하여 개발이 되어야 하겠다.

Reference

Jang, J. H., Shin, Y. S. (2020). Factors Influencing on Health Promoting Behavior of Community-dwelling Older Adults, *Journal of the Korea Academia-Industrialcooperation Society*, 21(2), 460-469.

Kim, M. J., Lim, J. Y. (2017). The Effect of Socioeconomic Status on the Prevalence of Chronic Disease in the Elderly: Focusing on Nutrient Intake, *Health and Social Welfare Review*, 37(4), 125-145.

Kim, S. Y., Kim, M. I., Chang, S. J., Moon, K. J. (2017). Identification and Prediction of Patterns of Health Promoting Behaviors among the Elderly, *Health and Social Welfare Review*, 37(2), 251-286.

Song, H., Kim, H. S. (2020). Convergence Factors Influencing Perceived Health Status, Health Promotion Behavior and Anxiety of Dementia Development in the Elderly Participation in Local Expos on Health-related Quality of Life, *Journal of the Korea Convergence Society*, 11(7), 41-49.

[1] Statistics Korea. Life expectancy and disability adjusted life expectancy, Statistics Korea, c2021 [cited 2021 Dec 15], Available from: https://www.index.go.kr/potal/main/EachDtlPageDetail.do?idx_cd=2758#quick_02; (accessed Mar. 31, 2022)

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 173-173

회복마취간호사의 마취간호 교육프로그램이 잡크래프팅, 임파워먼트, 직무열의에 미치는 효과

유제복[1], 김애숙, 표창옥, 권정희, 유현숙, 이민지, 인우영

요 약

본 연구는 회복마취간호사의 온라인 마취간호 교육프로그램이 잡크래프팅, 임파워먼트, 직무열의에 미치는 효과를 확인하고자 마취간호 교육프로그램을 이수한 회복마취간호사 54명을 대상으로 하였다. 본 연구는 전국의 회복마취간호사 대상으로 Dunabedian(1980)모델을 토대로 온라인 마취간호 교육프로그램을 개발하여 그 효과를 검증하기 위한 비동등성 대조군 전후 설계 연구이다. 전문가 집단을 통해 교육프로그램의 타당성을 검증하였고, 마취통증의학과 교수, 마취간호 강사의 강의와 회복마취간호사의 온라인 참여를 기반으로 1일 4시간 2회 교육을 실시간 적용하였다. 수집된 자료는 SPSS 23.0 프로그램을 이용하여 기술통계량과 Chi-squre test, t-test, Sharpiro-Will test, ANOVA를 이용하였다. 연구결과 온라인 마취간호 교육프로그램에 참여한 대상자의 잡크래프팅(F=5.26, p=.026), 임파워먼트(F=5.67, p=.021)에서 유의한 차이를 나타났다. 이는 마취간호 교육프로그램이 현장에서 일하는 회복마취간호사들의이 개인의 업무를 개념화하고 직무를 완성하기 위해 직무와 관련된 이들과 관계를 맺으면서, 자신의 직무를 의미 있게 변화시키는 행동을 증진시킴을 알 수 있었으며 이를 토대로 실무현장에 필요한 다양한 마취간호 교육프로그램의 개발과 적용을 제언하는 바이다.

주요용어 : 회복마취간호사, 마취간호 교육프로그램, 잡크래프팅, 임파워먼트, 직무열의.

[1]52725 경남 진주시 동진로 33, 경상국립대학교 간호학과 부교수. E-mail: jbyoo@gnu.ac.kr

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 175-175

회복마취간호사의 마취모니터 실습교육이 환자안전에 미치는 효과

유제복[1], 김애숙, 표창옥, 권정희, 유현숙, 이민지, 인우영, 김혜진

요 약

본 연구는 회복마취간호사의 마취모니터 실습교육이 환자안전문화, 조직의사소통, 환자안전관리활동에 미치는 효과를 확인하고자 수행한 연구로 회복마취간호사 56명을 대상으로 실습교육을 시행하면서 자료를 수집하였다. 본 연구는 회복마취간호사 대상으로 마취모니터 실습교육모듈을 개발하고 적용한 후 대상자의 환자안전문화, 조직의사소통, 환자안전관리활동에 미치는 효과를 확인하기 위한 비동등성 대조군 전후 설계 연구이다. 본 프로그램은 Jeffries의 이론적 기틀을 토대로 분석, 설계, 개발 단계에 따라 개발하여 전문가 집단을 통해 마취모니터 실습교육프로그램의 타당성을 검증하였다. 마취통증의학과 교수, 마취간호 강사와 프리셉터 간호사들이 실습교육을 하였고 회복마취간호사의 실습 참여를 기반으로 이론교육 3시간과 실습교육 5시간을 적용하였다. 수집된 자료는 기술통계량과 회귀분석을 이용하였다. 연구결과 실습교육을 받은 실험군은 실습교육을 받지 않은 대조군에 비해 환자안전문화와 조직의사소통의 점수가 더 높았고 실험군의 환자안전문화 점수와 조직의사소통 점수는 사후에 증가하였으며 두 군 간의 유의한 차이가 있었다(t=-2.46, p=017)(t=-2.09, p=.043). 본 연구는 회복마취간호사의 마취모니터를 위한 실습교육 프로그램을 개발하여 실무현장에 적용했다는 점에서 간호학적인 의의가 있다. 또한 본 연구의 결과를 토대로 빠르게 발전하는 다양한 장비 사용을 위해 확장성이 높은 실습교육 프로그램의 개발과 적용을 제언하는 바이다.

주요용어: 회복마취간호사, 실습교육프로그램, 환자안전문화, 조직의사소통, 환자안전관리활동

[1]52725 경남 진주시 동진로 33, 경상국립대학교 간호학과 부교수. E-mail: jbyoo@gnu.ac.kr

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 177-177

Search for R-parity violating supersymettry in pp collisions at center of mass energy 13 TeV in the CMS detector

SungHwan Kim[1], *Seok-Mo Heo*[2], *SeongJun Jung*[3], *YongHo Jeong*[4], *HyeonMin Gwak*[5], *SeongMin Yang*[5], *YongHak Lee*[5]

Abstract

Results are reported from a search for new physics, the Minimal Supersymmetric Standard Model (MSSM) in proton-proton collisions at the center of mass energy . This analysis focused on the signature of a large multiplicity of jets and b-tagged jets without a missing transverse energy requirement. The data sample consists of an integrated luminosity of recorded by the CMS detector in the Large Hadron Collider (LHC). The results are explained in terms of limits for the R-parity violating supersymmetric extension of the Standard Model in the gluino pair production benchmark model where each gluino decays via . The gluino with are excluded at 95% confidence level.

[1]Professor, Department of AppliedStatistics, Konkuk Univeersity, Seoul 05029, South Korea.
Email: shkim1213@konkuk.ac.kr.
[2]professor, Department of Periodontology, Research Institute of Clinical Medicine of Jeonbuk National University, Jeonju, South Korea.
[3]graduate student, IPAI, Seoul National University, Seoul, South Korea.
[4]Ph.D. in physics, AI analytics Team, Mustree, Seoul, South Korea.
[5]graduate students, Department of AppliedStatistics, Konkuk University, Seoul, South Korea.

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 179-179

Automated Technology for Strawberry Size Measurement and Weight Prediction Using AI*

Haejun Jeong[1], *Haejun Moon*[2], *Heejae Kwon*[2], *Yonghak Lee*[2], *Seongmin Yang*[2], *Chanyeong Kim*[2], *Sunghwan Kim*[3]

Abstract

In this study, we propose an automated system for measuring the size of strawberries and predicting their weight using AI technology. The system combines computer vision techniques with LiDAR sensor data to accurately estimate the dimensions of strawberries and infer their weight. By integrating deep learning models, such as HRNet for keypoint detection, and leveraging the capabilities of LiDAR sensors, we minimize human intervention and achieve precise size measurement. The relative errors for the width and height of the strawberries are 3.84% and 5.20%, respectively, with the width exhibiting a lower error rate. The standard deviation for the width and height of the strawberries are 0.28% and 0.26%, this indicates that the individual strawberries had very low error rates in terms of their measurements for the width and height. Weight prediction was performed through regression analysis with width and height estimation. Experimental results demonstrate that our approach enables accurate weight prediction with an accuracy rate of over 90%. This automated technology holds great potential for strawberry harvesting and classification tasks, facilitating the automation of these processes.

Keywords : Deep Learning, Strawberry size, LiDAR, Point Cloud.

---

*This work was supported by the Konkuk University Researcher Fund in 2023 and the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology under Grant NRF 2021R1A4A5032622.

[1]Department of Environmental Health Science, Konkuk University, Seoul 05029, South Korea.
E-mail: joon436@joongdong.hs.kr
[2]Department of Applied Statistics, Konkuk University, Seoul 05029, South Korea.
[3](Corresponding author) Department of Applied Statistics, Konkuk University, Seoul 05029, South Korea.
E-mail: shkim1213@konkuk.ac.kr

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 181-181

선거 여론조사에서 무응답 대체를 통한 결과 예측

성진용[1], 권민수[2], 최승배[3], 강창완[4]

요 약

선거 여론조사를 통해 나타나는 결과에서 후보들의 지지율은 국민의 주요 관심사일 뿐만 아니라 해당 정당의 선거 전략 수립을 위해 중요한 요소로 작용한다. 그러나, 선거 여론조사의 부정확하거나 왜곡된 결과가 나오면 그 영향력 때문에 큰 문제를 일으킬 수 있어, 여론조사에 관한 논란이 자주 발생한다. 여론조사 예측 결과는 실제 선거 결과와 다른 사례들이 다수 나타났으며, 이에 따라 여론조사 관련 기관인 정당이나 조사기관 등에서 선거 예측의 정확성 제고를 위해 노력하고 있다(김정훈 외, 2017). 본 연구에서는 제20대 대통령 선거, 2021년 부산시장 보궐선거, 제8회 전국동시지방선거(부산광역시 남구, 김해시장) 여론조사 데이터를 사용하여 머신러닝(Machine Learning) 알고리즘을 통해 무응답을 대체하여 선거 예측의 정확성을 높이고자 한다. 연구 절차는 다음과 같다: 첫째, 여론조사 문항 중 후보 선택 문항에서 무응답(모르겠다/없다)을 제외한 데이터를 train data로 분리하고 무응답(모르겠다/없다)을 선택한 데이터는 test data로 분리하였다. 둘째, 의사결정나무의 앙상블 모델을 적용하였으며 셋째, train data를 기반으로 학습된 모델을 통해 test data의 무응답을 한 명의 후보를 선택한 응답으로 대체하여 해당 데이터와 train data를 합쳐 후보별 득표율을 산출하고 실제 선거 결과와 비교 분석하였다.

주요 용어 : 여론조사, 무응답 대체, 머신러닝

---

[1]48059 부산광역시 해운대구 센텀북대로 60 ㈜에스티이노베이션, 주임. E-mail : jyseong@stinnovation.co.kr
[2]48059 부산광역시 해운대구 센텀북대로 60 ㈜에스티이노베이션, 대표이사. E-mail : ceo@stinnovation.co.kr
[3]47340 부산광역시 부산진구 가야동 산 24 동의대학교 산업경영·빅데이터공학전공, 교수.
E-mail : csb4851@deu.ac.kr
[4](교신저자) 47340 부산광역시 부산진구 가야동 산 24 동의대학교 산업경영·빅데이터공학전공, 교수.
E-mail : cwkang@deu.ac.kr

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 183-183

파이썬을 활용한 서베이 보고서 자동화에 관한 연구

민근홍[1], 권민수[2], 성진용[3], 심성현[4], 최승배[5], 강창완[6]

요 약

최근 업무 자동화에 대한 관심이 높아지고 있다. 많은 기업이 자동화 프로세스를 구현하여 적용하고 있으며, 이는 특히 데이터 중심의 분야에서 중요하다. 서베이와 같은 분야에서 데이터 수집, 처리, 분석은 시간이 많이 소요되는 작업으로 현재 대부분 수동으로 처리되고 있다. 이는 오류 발생 가능성을 높이고 많은 시간이 소요되는 단점이 있다. 따라서 자동화를 도입하면 업무 효율성과 정확성을 크게 향상할 수 있으며, 품질과 속도를 개선할 수 있다. 이러한 배경에서 본 연구는 서베이에서 수집된 raw data를 효율적으로 분석하고, 분석 결과를 자동으로 문서화 하는 방법을 탐구한다. 이를 위해 파이썬을 사용하여 데이터를 처리하고 빈도분석, 카이제곱 검정(chi-square test), t-검정(t-test), 분산분석(ANOVA), 회귀분석, 로지스틱 회귀분석 등의 다양한 통계 분석을 구현하였으며, 또한 파이썬으로 hwp 문서를 제어하여 통계분석 결과를 보고서 형태로 자동 생성하는 프로세스를 개발하였다. 과정은 다음과 같다. 첫째, raw data의 각 문항에 대하여 적절한 분석 방법을 적용하기 위해 문항을 분석 방법마다 별도로 분리한다. 둘째, 분류된 문항들을 적절한 통계분석 방법론을 사용하여 분석한다. 셋째, 분석 결과를 빈도표, 교차 분할표 등으로 나타내고 각 분석 방법에 맞는 정형화된 해석을 제작하여 나타낸다. 넷째, 표의 결과를 시각화하여 그래프로 나타낸다. 이러한 일련의 과정을 거쳐 하나의 완성된 보고서를 hwp 문서로 변환하여 최종 보고서를 나타낸다. 이 프로세스는 서베이 리포트 작성 시간을 현저히 단축할 뿐만 아니라, 기업과 실무자의 시간적, 경제적 효율성을 증대시킬 것이라 기대한다.

[1]48059 부산광역시 해운대구 센텀북대로 60 ㈜에스티이노베이션, 연구원. E-mail : ghmin@stinnovation.co.kr
[2]48059 부산광역시 해운대구 센텀북대로 60 ㈜에스티이노베이션, 대표이사. E-mail : ceo@stinnovation.co.kr
[3]48059 부산광역시 해운대구 센텀북대로 60 ㈜에스티이노베이션, 주임. E-mail : jyseong@stinnovation.co.kr
[4]47340 부산광역시 부산진구 가야동 산 24 동의대학교 산업경영·빅데이터공학전공, 교수. E-mail : ssh@deu.ac.kr
[5]47340 부산광역시 부산진구 가야동 산 24 동의대학교 산업경영·빅데이터공학전공, 교수. E-mail : csb4851@deu.ac.kr
[6]47340 부산광역시 부산진구 가야동 산 24 동의대학교 산업경영·빅데이터공학전공, 교수. E-mail : cwkang@deu.ac.kr

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 185-185

행사성 사업의 수요 분석 방법론 적용 방안 - TBATS 모형을 적용하여

최영은[1], 윤영학[2], 어승섭[3], 최은철[4]

요 약

본 연구는 지역 주민의 화합, 관광객 유치, 지역 홍보 등을 목적으로 개최되는 행사성 사업의 수요를 정량적으로 분석하는 데 초점을 맞췄다. 행사성 사업의 특성상 방문객 수의 변동성이 크고, 기존 방문객 추세와 행사에 의한 추가 방문객을 구분하기가 쉽지 않다. 이에 본 연구에서는 서울시 생활인구 데이터와 TBATS 모형을 활용하여 행사성 사업의 수요를 분석하는 방안을 제시하였다. 본 연구의 분석 결과, 서울시에서 개최된 행사성 사업은 실제 해당 지역의 방문객 수가 유의미하게 증가하는 것을 보여주었다. 즉 본 연구에서 제안한 연구방법론은 기존의 주로 정성적 연구 방법론에 의존했던 연구들과 비교하여, 향후 정책 수립에 있어 보다 직접적으로 활용 가능한 정량적 자료를 제공할 수 있을 것으로 보인다. 특히, 정례적으로 개최되는 행사성 사업의 수요 예측뿐만 아니라, 사후 평가에도 유용하게 활용될 수 있을 것으로 기대된다. 이는 행사성 사업의 효율성과 효과를 평가하고, 더 나아가 지역 문화 및 경제 발전을 위한 정책 결정에 중요한 기초 자료를 제공할 수 있을 것으로 판단된다.

주요용어 : 행사성 사업, 서울 생활인구, TBATS 모형

[1]06756 서울특별시 서초구 남부순환로340길 57, 서울연구원 부연구위원. Email: hk3327@si.re.kr
[2]06756 서울특별시 서초구 남부순환로340길 57, 서울연구원 연구원. Email: yoonyh@si.re.kr
[3]49111 서울특별시 노원구 공릉로 232, 서울과학기술대학교 에너지융합연구센터, Email: livelab21@nate.com
[4](교신저자) 02841 서울특별시 성북구 안암로 145, 고려대학교 정책대학원 강사. Email: aidster@korea.ac.kr

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 187-187

국제물류주선업의 혁신요인과 혁신이 성과에 미치는 영향

전경숙[1], 김상열[2], 김태훈[3]

요 약

고객 요구의 변화, 비즈니스 환경의 변화, 기술발전 등으로 인해 국제물류주선업의 서비스 범위는 확대되고 복잡해졌으며 글로벌공급망관리, 부가가치서비스, 정보 및 컨설팅에 이르기까지 종합적인 물류 서비스 제공으로 고객의 만족도와 충성도를 높이는데 핵심적인 역할을 하고 있다. 그러나 국내 기업들은 서비스의 질을 개선하기보다는 치열한 가격경쟁을 하고 있으며 국제무대에서의 경쟁력도 약해지고 있다. 국내 물류기업들은 새로운 경제 시장에서 운송의 전통적인 역할을 넘어 경쟁우위 확보를 위한 고객가치 창출에 주력해야 하며 이를 위해서 혁신은 필수적이다. 이에 본 연구에서는 물류 산업의 특성에 초점을 맞추어 국제물류주선업을 대상으로 혁신에 영향을 미치는 요인과 혁신이 성과에 미치는 영향에 대해 연구 하였다. 혁신의 요인을 기술불확실성, 최고경영자 특성, 물류정보시스템, 개인특성으로 보고 혁신의 유형을 서비스혁신과 관리혁신으로 구분하였으며. 물류성과는 효율성과 효과성으로 구분하고 기업의 성과는 재무성과로 측정하였다. 연구결과, 혁신요인들 중 개인특성을 제외한 요인들은 서비스혁신과 관리혁신에 긍정적인 영향을 미치며 서비스혁신은 물류의 효과성, 관리혁신은 물류의 효율성과 효과성에 긍정적인 영향을 미치는 것으로 나타났다. 또한 물류성과인 효율성과 효과성은 기업의 재무적성과에 긍정적인 영향을 미친다는 것을 실증하였다. 본 연구를 통하여 국제물류주선업의 혁신은 물류 효율성과 효과성을 향상하며 기업의 재무성과를 증대하는 데 긍정적인 영향을 미칠 수 있다는 것을 보여주어 기업의 경쟁력 제고에 혁신이 필요하다는 것을 확인하였다.

주요용어: 국제물류주선업 혁신요인, 혁신유형, 물류 효율성, 물류 효과성, 혁신 성과.

[1]46241 부산시 금정구 부산대학로길 63번길2, 부산대학교 국제전문대학원 박사과정. E-mail: barami791004@naver.com
[2](교신저자) 46241 부산시 금정구 부산대학로길 63번길 2, 부산대학교 국제전문대학원 교수. E-mail: ksy@pusan.ac.kr
[3]48434 부산시 남구 수영로 309 경성대학교 경제금융물류학부 교수. E-mail: kdbdc@ks.ac.kr

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 189-192

가상화폐 시장에서의 역허딩 행태는 이성적인 투자 결정의 결과?

남유상[1], 조영민[2]

요 약

역허딩(anti-herding) 행태는 허딩 행태와 다른 움직임을 보여주는 것으로 이성적인 투자 행태로 정의되기도 하지만, 유의적인 특정 움직임을 보여준다는 점에서 비이성적 투자 행태로 정의된다. 하지만, 기관투자자의 허딩 행태와 역허딩 행태는 이성적 판단하에 이루어지는 외견상 허딩 행태로 인정되고, 상대적으로 개인투자자의 허딩 행태는 뚜렷한 신념과 정보의 부재로 인해 비이성적인 투자 행태라고 정의된다. 다만, 역허딩 행태의 경우 일방적으로 비이성적 투자 행태로 단정 짓기 어려운 이유는 역허딩 행태의 발현이 자신의 신념과 시장 흐름에 대한 반응을 추종이 아닌 회피로 선택하는 경우인데, 다름을 선택하는 경우, 개인들의 강력한 신념이 필요하다. 역허딩 행태가 이성적인 투자 결정의 결과임을 확인하기 위해서는 개인투자자의 비중이 높은 가상화폐 시장을 대상으로, 채굴의 개념으로 Pow, PoS, M-other, NM형의 네 가지로 구분하고, 급격한 상승과 하락의 시기를 2017년과 2020년을 기점으로 두 개의 기간을 분리하여, CSSD와 CSAD 방법론을 사용하여 분석했다. 분석 결과 일상적인 허딩 행태와 상승과 하락의 극단에서 보여주는 허딩 행태가 달랐으며, 상승 극단에서는 전체 기간에서 역허딩 행태가 나타났다. 하락 극단과는 달리 상승 극단에서는 나타나는 역허딩 행태는 공포의 회피라기보다는 시장에 관한 판단의 결과로 보인다. 이러한 점에서 역허딩 행태는 단순히 특정한 움직임을 보여주는 것이 아니라, 시장에 대한 자신들의 판단으로 인한 결과로 이성적인 투자 행태일 가능성을 보여준다.

주요용어 : 가상화폐, 허딩 행태, 역허딩 행태, CSSD, CSAD

1. 서론

허딩 행태 연구에서 드러나는 역허딩 행태의 경우 행태의 발생원인에 대해서 뚜렷한 결론을 내리지 못하고 있다. 기관투자자의 경우에는 정보 집합을 바탕으로 투자 결정을 하게 되는데, 기관투자자들이 보유하고 있는 정보 집합의 유사성으로 인해 허딩 행태(Bikhchandani & Sharma, 2001)를 보여주고, 자신들의 능력에 대한 강한 자신감((Effinger et al., 2001; Levy, 2004; Ali et al., 2022)을 바탕으로 자신의 명성과 평판, 그리고 차별성을 위해 기존 전문가들과는 다른 모습을 보여준다고 한다. 하지만, 개인투자자의 경우 기관투자자들이 보유하고 있는 정보 집합을 보유하지 않고, 자신의 능력에 대한 자신감은 과신으로 치부된다.

하지만, 역허딩 행태는 그들이 참여하는 투자자산 시장의 흐름과 같이 움직이는 것이 아니라 다르게 움직이는 것으로 다수의 흐름과 다른 모습을 보여준다는 것은 그들에게 쉬운 판단은 아닐 것

[1]48434 부산시 남구 수영로 309(대연동), 경성대학교 경영학과 초빙교수. E-mail: nysang@ks.ac.kr
[2]48434 부산시 남구 수영로 309(대연동), 경성대학교 경영학과 박사과정생. E-mail: youngminchocho@gmail.net

이다. 그들만의 기준과 능력을 과신하는 기관투자자들을 제외하고 개인투자자의 움직임을 분석하고자 한다면, 가상화폐 시장이 가장 적합하리라 생각된다. 가상화폐 리서치기관인 Diar에 의하면 2020년 1분기 가상화폐 시장의 기관투자자 비중은 26% 정도라고 하는데, 상당수는 비트코인에 한정된 투자이기 때문에 실제 가상화폐 시장의 투자자들은 개인투자자들이라 할 수 있다.

2000년 이후 컴퓨팅능력의 비약적인 발전과 2008년 글로벌 금융위기는 새로운 금융시장을 개척하게 된다. 비트코인이라는 탈중앙화된 블록체인을 바탕으로 하는 가상화폐의 등장 이후, 급속히 성장하였으며, 비트코인 이외의 다양한 알트코인들이 등장하게 된다. 혁신적인 컨셉을 가진 가상화폐의 등장은 투자자에게 상품에 대한 명확한 개념 파악을 어렵게 만들었다. 그로 인해, 시장 내에는 시장의 흐름을 추종하는 허딩 행태가 만연할 것으로 생각되지만, 의외로 허딩 행태는 그렇게 쉽게 드러나지 않고 있다(da Gama Silva et al., 2019; Vidal-Tomas et al. 2019; Stavrouiannis and Babalos, 2019). 가상화폐 시장의 성장 동력은 혁신성이다. 하지만, 그러한 성장 동력으로 인해 시장은 불안정한 모습을 보여주며, 허딩 행태는 그러한 불안정성에서 태어날 것이다. 역허딩 행태가 자신의 판단과 시장에 대한 회피라면, 단순히 유의적인 특정한 모습을 보여준다고 해서 비이성적인 투자 결정이 결과라고 장담하기 어렵다. 이는 상승과 하락이라는 구간의 성격과 함께, 가상화폐 시장에서 발현되는 허딩 행태의 변화추이를 분석하여 역허딩 행태의 특성을 구분하고자 한다.

2. 분석자료와 방법론

본 연구에서 사용된 가상화폐의 자료는 coinmarketcap.com에서 제공하는 2458개 가상화폐의 일별 가격, 거래금액, 시가총액자료이며 분석 기간은 2013년 12월 27일에서 2023년 6월 30일이며, 전체 기간을 (1) 2017-2019, (2) 2020-2022로 구분하여 분석하였다. 채굴형 가상화폐의 경우 PoW, PoS, M-other, NM으로 구분하였다.

$$CSSD_t = \alpha + \beta_u D_t^U + \beta_l D_t^L + \varepsilon_t \tag{1}$$

$$CSAD_t = \alpha + \beta_1 |R_{m,t}| + \beta_2 R_{m,t}^2 + \varepsilon_t \tag{2}$$

$$CSAD_t = \alpha + \beta_1 \left| R_{m,t}^{(+)} \right| + \beta_2 \left| R_{m,t}^{(-)} \right| + \beta_3 R_{m,t}^{2(+)} + \beta_4 R_{m,t}^{2(-)} + \varepsilon_t \tag{3}$$

가상화폐 시장의 허딩 행태를 분석하는 방법론으로 $CSSD_t$와 $CSAD_t$를 통해서 기간별 허딩 행태를 분석하고 극단의 허딩 행태와 비교한다. 식(1)의 D_t^U와 D_t^L은 각각 상승과 하락 극단을 나타내는 더미 변수이다. β_u와 β_l이 유의적인 음수일 때 허딩 행태가 존재한다. 이와 반대로 유의적인 양의 계수가 나타난다면 역허딩 행태라고 할 수 있다. 식(2)에서 분석 기간 전체의 허딩 행태를 분석하고 식(3)에서 상승과 하락 구간의 허딩 행태를 확인한다.

3. 실증분석과 결과

<표 1>과 <표 2>는 각각의 네 종류의 가상화폐 허딩 행태를 분석한 결과이다. 우선 Panel A와 C는 CSAD분석을 통해 상승과 하락 구간의 허딩 행태를 분석한 것이다. Panel C와 Panel D는 CSSD을 이용하여 극단에서의 허딩 행태를 분석한 결과이다. 가장 큰 특징은 기간(1)에서는 수익률이 양인 구간에서는 허딩 행태를 보여주지만, 상승 극단에서는 강력한 역허딩 행태를 보여준다는

것이다. 두 번째, 기간(2)에서는 명확한 허딩 행태를 파악할 수 없으며, 상승 구간에서는 여전히 역허딩 행태가 등장한다는 것이다. CSAD를 통한 기존 연구에서는 대부분 약세구간에서 허딩 행태가 발현되었지만, 기간(2)에서는 역허딩 행태가 발견되기도 한다.

<표 1> PoW(175), PoS(154)

Panel A: PoW(175)의 상승, 하락 $CSAD_t$ 회귀결과

기간	α	β_1	β_2	β_3	β_4	Adj R^2
(1) 2017.01 - 2019.12	0.0657** (61.23)	0.6302** (11.84)	0.1767** (3.17)	-1.8198** (-3.95)	-0.4750 (-1.13)	0.27
(2) 2020.01 - 2022.12	0.0496** (63.82)	0.3980** (8.15)	0.0754* (1.93)	-0.5306 (-0.98)	0.8690** (3.35)	0.23

Panel B: PoW(175)의 $CSSD_t$ 회귀결과

	α	D^U	D^L	Adj R^2
(1) 2017.01 - 2019.12	0.1282** (140.72)	0.2733** (4.31)	0.1068 (1.85)	0.02
(2) 2020.01 - 2022.12	0.1081** (146.85)	0.3267** (4.96)	-0.0802 (-1.64)	0.02

Panel C: PoS(154)의 상승, 하락 $CSAD_t$ 회귀결과

기간	α	β_1	β_2	β_3	β_4	Adj R^2
(1) 2017.01 - 2019.12	0.0660** (58.54)	0.7150** (12.79)	0.2030** (3.38)	-2.4445** (-5.12)	-0.7571 (-1.66)	0.27
(2) 2020.01 - 2022.12	0.0482** (64.68)	0.2715** (5.51)	0.0284 (0.67)	0.6053 (1.10)	1.3672** (4.32)	0.20

Panel D: PoS(154)의 $CSSD_t$ 회귀결과

	α	D^U	D^L	Adj R^2
(1) 2017.01 - 2019.12	0.1279** (132.26)	0.2287** (3.47)	0.1323* (2.17)	0.01
(2) 2020.01 - 2022.12	0.1025** (147.58)	0.3908** (6.36)	-0.1023* (-2.11)	0.04

상승 구간의 역허딩 행태는 추가수익을 포기하는 행태지만, 하락 구간의 역허딩 행태는 손실인식을 거부하는 것이다. 이는 수익의 관점에서 대응하는 것과 손실인식의 관점 차이인데, 상대적으로 상승 극단에서 역허딩 행태를 보여주는 것은 상대적으로 쉬운 결정이지만, 하락 극단에서 역허딩 행태를 보여준다는 것은 맹목적인 신념이 필요하다. 실제 결과를 살펴보면, 네 가지의 가상화폐 중에서 하락 극단에서 허딩 행태를 보여주는 가상화폐는 PoS의 NM에서 살펴볼 수 있는데, 그들은 기간(1)에서 역허딩 행태를 보여주었지만, 기간(2)에서는 허딩 행태를 보여주었다. 역허딩 행태는 자신의 판단에서 시작된다는 것을 고려하면, 상승 구간에서는 수익 포기와는 달리 하락 극단에서 시장과 다른 움직임을 보여준다는 것은 매우 강력한 신념이 필요하다. 문제는 이러한 신념을 형성하기 위한 기본적인 정보 집합을 보유하지 않은 상황이라는 것을 생각해보면, 이는 이성적 투자 결정의 결과물이 가능성은 매우 낮다. 극단적인 상황에서 발생하는 역허딩 행태는 이성적인 판단이나, 지나친 과신의 결과물이라면 이는 최선의 결과물이 아닐지라도 자신이 결정한 이성적인 판단의 결과물이라고 할 수 있다.

<표 2> M-other(288), NM(1899)

Panel A: M-other(228)의 상승, 하락 $CSAD_t$ 회귀결과

기간	α	β_1	β_2	β_3	β_4	Adj R^2
(1) 2017.01 - 2019.12	0.0906** (60.41)	0.4371** (6.87)	0.0767 (1.08)	0.5574 (1.23)	-0.6306 (-1.22)	0.27
(2) 2020.01 - 2022.12	0.0662** (67.91)	0.5482** (8.83)	0.1642** (3.16)	-2.1041** (-3.06)	0.1653 (0.52)	0.16

Panel B: M-other(228)의 $CSSD_t$ 회귀결과

	α	D^U	D^L	Adj R^2
(1) 2017.01 - 2019.12	0.1607** (127.20)	0.5818** (8.04)	0.2749** (3.50)	0.06
(2) 2020.01 - 2022.12	0.1402** (165.02)	0.2560** (3.45)	-0.0456 (-0.75)	0.01

Panel C: NM(1899)의 상승, 하락 $CSAD_t$ 회귀결과

기간	α	β_1	β_2	β_3	β_4	Adj R^2
(1) 2017.01 - 2019.12	0.4352** (79.97)	0.4352** (10.87)	0.1150** (3.04)	-0.8477* (-2.46)	-0.2645 (-0.96)	0.29
(2) 2020.01 - 2022.12	0.0500** (74.00)	0.4468** (10.34)	0.1929** (5.18)	-0.5704 (-1.22)	0.5335* (1.99)	0.28

Panel D: NM(1899)의 $CSSD_t$ 회귀결과

	α	D^U	D^L	Adj R^2
(1) 2017.01 - 2019.12	0.1097** (171.60)	0.2567** (5.82)	0.0269 (0.72)	0.03
(2) 2020.01 - 2022.12	0.0996** (191.88)	0.2944** (6.12)	-0.1280** (-3.75)	0.04

References

Ali, S., Badshah, I., & Demirer, R. (2022). Anti-herding by hedge funds, idiosyncratic volatility and expected returns. *Available at SSRN*, https://dx.doi.org/10.2139/ssrn.4010287.

Babalos, V., & Stavroyiannis, S. (2015). Herding, anti-herding behaviour in metal commodities futures: A novel portfolio-based approach. Applied Economics, 47(46), 4952-4966.Stavroyiannis, S., Babalos, V., Bekiros, S., & Lahmiri, S. (2019). Is anti-herding behavior spurious? *Finance Research Letters*, 29, 379-383.

Effinger, M. R., & Polborn, M. K. (2001). Herding and anti-herding: A model of reputational differentiation. *European Economic Review*, 45(3), 385-403.

Levy, G. (2004). Anti-herding and strategic consultation. *European Economic Review*, 48(3), 503-525.

Lee, C. (2017). An analysis of determinants of purchase of accident coverage using structural equation model, *Journal of the Korean Data Analysis Society*, 19(2), 827-837. (in Korean).

Stavroyiannis, S., Babalos, V., Bekiros, S., & Lahmiri, S. (2019). Is anti-herding behavior spurious? *Finance Research Letters*, 29, 379-383.

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 193-193

수익률 간 거리를 바탕으로 한 군집화 기반 포트폴리오 생성에 관한 연구

최인수[1], 김우창[2]

요 약

본 연구는 계층적 위험 균등 (Hierarchical Risk Parity, HRP) 포트폴리오 최적화에 대한 새로운 접근법을 제안하며, 자산 간의 선형 의존성 외의 의존성을 분석하고 평가하기 위해 여러 가지 통계적 거리 지표를 포함한 실험을 진행하였다. 전통적인 방법들은 주로 자산 간의 선형 관계를 측정하기 위해 상관 계수를 사용하였다. 하지만 이러한 상관 계수는 통계적 성질을 전제로 한다. 이러한 점에서 최적 운송 거리, 정보 거리 등과 같은 다양한 금융 자산 간의 공유된 정보를 정량화하여 분포한 거리를 선형에 국한되지 않고 분석할 수 있다는 장점이 있다. 본 연구에서 사용한 HRP 방법론은 상관 행렬의 계산, 거리 행렬로의 변환, 자산의 계층적 클러스터링, 클러스터 내에서 가중치의 재귀적 할당을 포함한다. 본 연구의 실험 결과는 HRP 방법론을 통해 금융 자산의 수익률에 대한 통계적 분포를 고려한 지표로부터 상호 보완적인 포트폴리오 비중 최적화를 창출할 수 있음을 확인하였다.

주요용어: 정보 이론, 비선형 분석, 군집 분석, 계층적 리스크 패리티 모형

[1]한국과학기술원 산업및시스템공학과 박사과정
[2]한국과학기술원 산업및시스템공학과 교수

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 195-195

A Simple Test for Financial Speculation in the Cryptocurrency Markets

Myeong jun Kim[1], *Meiling Jin*[2], *SungY. Park*[3]

Abstract

This study uses a time-varying coefficient version of the model proposed by Llorente, Michaely, Saar, andWang (2002) and quantifies the level of speculative behavior using a functional-coefficient method to investigate whether speculative trading plays an important role in the cryptocurrency market. Using daily price data for the eight major cryptocurrencies by market capitalization, we find that neither the speculative motive nor the hedging motive dominates the cryptocurrency market over the study period, with the exception of Ethereum, where speculative activity was strong after COVID-19 (towards the end of 2020), and Bitcoin, where the hedging motive dominated towards the end of 2019, despite being temporary. Moreover, the empirical results do not support the argument that speculative activity contributed significantly to the substantial price increase in the cryptocurrency market in the period after COVID-19.

Keywords : Cryptocurrency; Speculation; Hedging; Semi-parametric model; Functional coefficient model.

[1]Division of International Studies Kongju National University, Gongjudaehak-ro 56, Gongju-si, Chungcheongnam-do, Korea. E-mail: myeongjun@kongju.ac.kr
[2]School of Economics, Chung-Ang University, 84 Heukseok-Ro, Dongjak-Gu, Seoul, Korea. E-mail: meiling1206@cau.ac.kr.
[3]Corresponding Author: School of Economics, Chung-Ang University, 84 Heukseok-Ro, Dongjak-Gu, Seoul, Korea. E-mail: sungpark@cau.ac.kr.

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 197-201

원가계산에서 상호배부의 새로운 계산 방법에 관한 연구

남기성[1], 이승호[2], 윤창준[3]

요 약

본 연구는 원가계산에서 간접 부서의 비용 배부에서 많이 사용하는 상호 배부법의 새로운 계산 방법에 관한 연구이다. 기존의 상호 배부법 계산은 연립방정식을 이용하여야 하기에 계산 시간이 오래 걸리고, 역행렬이 존재하지 않을 경우는 계산 자체에 어려움이 있었다. 새로운 계산 방법은 항목의 수가 많은 활동원가 계산 방법에서도 쉽게 사용할 수 있다. 새로운 알고리즘은 상호 배부법의 기본 원리만 정확히 이해하면, 엑셀에서 사용이 가능할 정도로 계산이 비교적 간단하며, 간접 부서와 직접 부서 및 활동의 수에 크게 영향을 받지 않았다. 또한, 전통적인 부서 단위의 원가계산뿐만 아니라, 활동기준원가계산(ABC:Activity-Based Costing)에서도 사용할 수 있다.

주요용어 : 원가계산, 상호 배부법, 배부 테이블, 활동원가

1. 서론

원가(cost)는 제조 물품이나 서비스 분야에서 자원을 얻는 데 소요된 재화나 용역의 가치를 화폐액으로 측정한 것이다(이윤호, 2012). 원가적인 측면에서 병원의 의료비용은 크게 인건비, 재료비, 관리비로 구분된다.

어떤 병원에서 시술한 한 건의 맹장 수술 원가를 계산할 때, 여기에는 의사, 간호사의 인건비와 원무과와 세무 등을 담당하는 재무과 직원의 인건비와 수술 재료와 주사기, X-ray 등 각종 의료 기기 감가 삼각비, 병원 전기료, 수도료, 청소비 등의 관리 비용이 모두 포함되어야 한다. 그리고 의사, 간호사 등의 1년 연봉에서 본 건에 해당하는 인건비는 얼마인가? 등 배분의 문제가 대두된다. 이를 위해서는 간접비의 배분은 매우 중요하다.

선행 연구로서 Hilton 등(2008)은 서비스 부서 비용을 생산 부서에 할당하는 방법에는 직접 방법, 단계적 방법, 상호 방법 등 세 가지 방법이 있지만, 상호 방법은 계산의 어려움 때문에 거의 사용되지 않는다고 하였다. Wallace R. Leese와 Tim Kizirian(2009)는 Excel의 행렬 연산을 사용하여 상호배분 방법이 보다 좋은 할당 방법이 된다고 발표하였다. 이처럼 상호 배부법의 계산은 어려운 과제이었다. 의료분야의 원가계산에 많은 이슈가 있지만, 본 연구에서는 간접비의 배부에서 사용되

[1](교신저자) 06372 서울시 강남구 자곡로 175, 707-905, 前) 한국고용정보원, 선임연구위원.
E-mail : nks@chol.com
[2]07801 서울특별시 강서구 마곡중앙6로 11, 619호, 커넥트메디(주), 대표이사. E-mail : lsho@vlcompany.com
[3]08592 서울시 금천구 가산디지털2로 14, 대륭테크노타운 12차 1405호, 주) 케이원, 이사.
E-mail : changjoon_youn@k-one.co.kr

는 상호 배부법의 계산 방법으로 한정한다.

본 연구의 구성은 1장에서는 연구의 필요성과 배경 등의 설명을 하고, 2장에서는 원가계산에서의 간접비 배분과 상호배부의 새로운 계산 방법을 소개한다. 그리고 3장에서는 제안하는 알고리즘을 이용하여 시뮬레이션을 실시하고, 마지막으로 4장에서는 결론을 다룬다.

2. 배부 테이블을 이용한 새로운 상호 배부법

2.1 원가계산에서 간접비의 배부

일반적으로 간접비에 대한 배부에서 가장 많이 사용하는 방법으로 직접 배부법(direct allocation method), 단계식 배부법(step-down allocation method), 상호 배부법(reciprocal allocation method)이 가장 많이 사용되고 있다.

직접 배부법은 간접 부서 상호 간에 관계를 전혀 고려하지 않고, 간접 부서 원가를 배부하는 방법으로 간접 부서 원가를 다른 간접 부서에는 전혀 배부하지 않고, 주된 부서(직접 부서)에만 배부하는 방식이다. 이는 계산이 간단하고, 해석에서 용이하며, 추적이 가능하다는 장점이 있지만, 간접 부서 상호 간의 관계를 무시하기 때문에 간접 부서 상호 간에 많은 용역을 주고, 받는 경우에는 정확성이 떨어질 수 있다.

단계식 배부법은 간접 부서 상호 간의 용역 수수 관계를 부분적으로 인식하여 간접 부서 원가를 배부하는 방법으로 간접 부서 원가의 배부 순서부터 정한 후, 그 순서에 따라 간접 부서 원가를 다른 간접 부서와 직접 부서에 배부하는 방식이다. 단, 배부가 끝난 간접 부서에는 간접 부서 원가를 배부하지 않는다. 따라서 간접 부서 상호 간의 용역 수수 관계를 일부 인식하여 배부한다는 장점이 있지만, 부서 상호 간에 주고받는 용역을 완전히 인식하는 것은 아니므로 간접 부서 원가의 배부에 부정확성이 상존하고, 나중에 배부되는 간접 부서 원가일수록 정확성이 떨어지고, 원가 배부에 많은 시간이 필요는 단점이 존재한다.

마지막으로 상호 배부법은 간접 부서 상호 간의 용역 수수 관계를 완전히 인식하여 간접 부서 원가 간 용역을 제공한 다른 간접 부서와 주된 부서에 배부하는 방법으로 기존에는 연립방정식의 선형대수의 원리를 이용하는 것으로 알려져 있다. 계산은 가장 정확하지만, 계산이 복잡, 보조부문의 수가 여러 개일 경우 시간과 비용이 많이 소요, 비선형일 때는 오차가 많이 발생, 연립방정식에서 역행렬이 존재하지 않으면 구하기 어려운 것으로 알려져 있다.

2.2 배부 테이블을 이용한 새로운 상호 배부법 알고리즘

본 연구에서 제안하는 상호 배부법(reciprocal allocation method)을 계산하기 위한 배부 테이블을 이용한 새로운 알고리즘은 다음과 같다.

Step 1) 간접 부서와 직접 부서의 초기 비용(Initial Cost)을 읽어 들인다.

- ○ 간접 부서 초기 비용 : InD_i , i는 배부하는 간접 부서($i = 1, 2, \cdots, n$)
- ○ 직접 부서 초기 비용 : DD_j , j는 배부하는 직접 부서($j = 1, 2, \cdots, m$)
  - (검증) $TIC = \sum_{i=1}^{n} InD_i + \sum_{j=1}^{m} DD_j$, TIC는 전체 투입비용(Total Input Cost)

Step 2) FTE(Full Time Equivalent)와 배부기준(Allocation Criteria Probability)을 읽어 들인다.

○ FTE : InA_{ik}, InA_{pq}, k는 배부하는 활동(Activities)($k = 1, 2, \cdots, l$), p는 배부받는 간접 부서($p = 1, 2, \cdots, n$), q는 배부받는 활동($q = 1, 2, \cdots, l$)

- $InA_{i\cdot} = \sum_{k=1}^{l} InA_{1k} = \sum_{k=1}^{l} InA_{2k} = \cdots = \sum_{k=1}^{l} InA_{ik} = 100\%$(배부할 때 k)

- $InA_{i\cdot} = \sum_{q=1}^{l} InA_{1q} = \sum_{q=1}^{l} InA_{2q} = \cdots = \sum_{q=1}^{l} InA_{iq} = 100\%$(배부받을 때 q)

○ 배부기준 : IAP_{ii}(배부할 때), DAP_j(배부받을 때)

- (검증) 배부기준 : $\sum_{i=1}^{n} IAP_{ip} + \sum_{q=1}^{m} DAP_{iq} = 100\%$

Step 3) 배부 테이블을 만든다.

○ 간접 부서 배부 테이블 = 배부기준 × FTE

- $TInPo_{ikpq} = IAP_{ip} \times InA_{pq}$ (2.1)

○ 직접 부서 배부 테이블 $TDPo_{ikj}$

- $TDPo_{ikj} = DAP_{ij} \times DA_j$ (2.2)

Step 4) 1단계($r = 1$) 배부 값을 계산한다.

○ 1단계 간접 부서 배부 값 = 초기 비용 × FTE

- $RInAC_{rik} = InD_i \times InA_{ik}$ (2.3)

○ 1단계 직접 부서 배부 값 = 초기 비용 × FTE

- $RDAC_{rj} = DD_j$ (2.4)

Step 5) 2단계($r = 2$) 배부 값을 계산한다.

○ 2단계 간접 부서 배부 값 = 배부값 × 배부 테이블값

- $RInAC_{rik} = \sum_{i=1}^{n}\sum_{k=1}^{l}(RDAC_{rj} \times TInPo_{ikpq})$ (2.5)

○ 2단계 직접 부서 배부 값 = 배부값 × 배부 테이블값

- $RDAC_{2j} = \sum_{i=1}^{n}\sum_{k=1}^{l}((RInAC_{1ik} \times TDPo_{ikj}) + RDAC_{1j})$ (2.6)

- (검증) $\sum_{i=1}^{n}\sum_{k=1}^{l} RInAC_{2ik} + \sum_{j=1}^{m} RDAC_{2j} = TIC$

Step 6) 간접 부서 배부값의 합을 계산하여 0.01 이상이면, r을 1 증가시키고($r = r + 1$), Step 5를 반복한다. 0.01 이하이면 Step 7을 실행한다.

○ IF $\sum_{i=1}^{n}\sum_{k=1}^{l} RInAC_{rik} \le 0.01$이면 다음 단계로 간다.

○ IF$\sum_{i=1}^{n}\sum_{k=1}^{l} RInAC_{rik} > 0.01$이면 Step 5로 간다.

Step 7) 직접 부서 배부값($RDAC_{rik}$)을 저장한다.

3. 새로운 알고리즘을 이용한 시뮬레이션

간접 부서 4개($n=4$), 직접 부서 5개($m=5$), 간접 부서1의 활동 3개($l=3$), 간접 부서2의 활동 2개, 간접 부서3의 활동 4개, 간접 부서4의 활동 1개로 각각의 비용은 Table 1에서 Table 3과 같다.

Table 1. 원가계산에서 초기 비용

	간접부서, $InD_i(n=4)$				직접부서, $DD_j(m=5)$					합계, TIC
	인사	재무	원무	기획	수술1	CT	MRI	외래	병동	
초기 비용	110,000	160,000	130,000	90,000	320,000	520,000	484,000	780,000	555,000	3,149,000

Table 2. 원가계산에서 FTE

	인사 (InA_{1k}, InA_{1q})		재무 (InA_{2k}, InA_{2q})		원무 (InA_{3k}, InA_{3q})		기획 (InA_{4k}, InA_{4q})		직접 부서 (DAP_j)	
FTE (InA_{ik}, InA_{pq})	급여관리	40.0%	결산관리	30.0%	접수수납	20.0%	예산관리	100.0%	수술1	100.0%
	채용관리	45.0%	세무관리	70.0%	환자관리	40.0%			CT	100.0%
	교육관리	15.0%			재원관리	25.0%			MRI	100.0%
					예약관리	15.0%			외래	100.0%
									병동	100.0%
	합계	100.0%	합계	100.0%	합계	100.0%	합계	100.0%	합계	500.0%

Table 3. 원가계산에서 배부기준

배부기준 IAP_{ip} DAP_{ip}	인사 IAP_{i1}	재무 IAP_{i2}	원무 IAP_{i3}	기획 IAP_{i4}	수술1 DAP_{i1}	CT DAP_{i2}	MRI	외래 DAP_{i4}	병동 DAP_{i5}	합계
인사	0.0%	5.0%	15.0%	10.0%	10.0%	10.0%	15.0%	20.0%	15.0%	100.0%
재무	10.0%	0.0%	5.0%	3.0%	14.0%	20.0%	15.0%	17.0%	16.0%	100.0%
원무	13.0%	7.0%	0.0%	10.0%	5.0%	20.0%	10.0%	20.0%	15.0%	100.0%
기획	10.0%	5.0%	3.0%	0.0%	15.0%	25.0%	10.0%	18.0%	14.0%	100.0%

Table 4. 원가 계산 결과

부서	활동	1 단계	2 단계	3 단계	4 단계	5 단계	6 단계	7 단계	8 단계	9 단계
인사	급여관리	16,760	3,330	938	217	56	14	3	1	0
	채용관리	18,855	3,747	1,056	244	63	15	4	1	0
	교육관리	6,285	1,249	352	81	21	5	1	0	0
재무	결산관리	5,730	1,632	407	99	25	6	2	0	0
	세무관리	13,370	3,807	950	231	58	14	4	1	0
원무	접수수납	5,440	1,621	349	95	22	6	1	0	0
	환자관리	10,880	3,242	698	190	45	11	3	1	0
	재원관리	6,800	2,026	436	118	28	7	2	0	0
	예약관리	4,080	1,216	262	71	17	4	1	0	0
기획		28,800	7,483	1,806	450	112	28	7	2	0
수술1		373,400	385,944	389,066	389,849	390,040	390,088	390,100	390,103	390,104
CT		611,500	632,150	637,562	638,869	639,196	639,277	639,297	639,302	639,303
MRI		546,500	561,250	564,873	565,784	566,007	566,063	566,077	566,080	566,081
외래		871,400	893,651	899,209	900,583	900,923	901,008	901,028	901,034	901,035
병동		629,200	646,653	651,035	652,119	652,387	652,454	652,470	652,475	652,476
합계		3,149,000	3,149,000	3,149,000	3,149,000	3,149,000	3,149,000	3,149,000	3,149,000	3,149,000

Table 4에서와 같이 제안한 알고리즘을 적용한 Step 5의 결과 간접 부서 배부값의 합을 계산하

여 0.01 이상이기에 단계 3으로 Step 5를 반복한다. 그리고 단계 9에서 간접 부서 배부값의 합을 계산하여 0이기에 멈추고 결과값을 저장한다. 최종 배부값은 수술1팀은 390,104원, CT팀은 639,303원, MRI팀은 566,081원, 외래팀은 901,035원, 병동팀은 652,476원 전체 합계는 3,149,000원이다.

4. 결론

본 연구는 원가계산의 간접비 혹은 간접 부서의 활동비 등의 배분에서 가장 정확한 것으로 알려진 상호 배분법의 계산이 복잡하거나, 역행렬의 존재에 영향을 받지 않는 새로운 계산 방법을 제안하는 연구이다.

새로운 알고리즘은 상호 배부법의 기본 원리만 정확히 이해하면, 엑셀에서 사용이 가능할 정도로 계산이 비교적 간단하며, 간접 부서와 직접 부서 및 활동의 수에 크게 영향을 받지 않았다. 또한, 전통적인 부서 단위의 원가계산뿐만 아니라, 활동기준원가계산(ABC:Activity-Based Costing)에서도 사용할 수 있다.

여러 활동에 따라 간접비를 배부하고 각 제품 혹은 서비스별로 활동 소비량에 따라 간접비를 배부함으로써 기존의 전통적인 원가계산방식에서 합리적인 원가 배부를 목적으로 하는 원가계산방식임에도 간접비의 배분이 일부 필요하며, 제안한 방법은 여기서도 사용할 수 있다.

이상영 등(2013)은 관리비의 원가 배부에 있어 가장 이상적인 방법은 상호배부이고, 실제 상호배부 방식으로 원가계산을 수행하는 것은 기술적으로 어려움이 있어 많은 나라에서는 단계적 배부방식으로 관리비를 배부하고 있다고 하고 있으며, 새로운 시스템에 반드시 상호배부를 해야 한다는 제안을 해결할 수 있다.

References

김희정, 정기선, 최성우. (2003). 수술실의 원가배부기준 설정연구. 병원경영학회지. 8(1): 135~164.

신영석 외 (2012). 유형별 상대가치 개선을 위한 의료기관 회계조사 연구, 한국보건사회연구원.

안태식 외 (2011). 진료비용 상대가치점수 개발을 위한 회계조사 연구용역, 서울대학교 경영연구소.

이상영, 신현웅, 황도경, 이해종, 나종익, 우정식, 김진호, 여지영, 이슬기 (2014). 체계적 원가조사를 위한 요양기관 패널제도 도입 연구, 한국보건사회연구원.

이해종 외 (2013). 포괄수가 원가분석 체계 구축방안, 연세대학교, 포괄수가제 발전을 위한 과제 - 원가체계 구축 중심, 건강보험공단 발표자료.

조정화 (2001). 한국 의료기관의 원가계산시스템 분석과 개선(K지역을 중심으로), 전남대학교 석사학위논문.

Hilton, R., Mayer, W., Selto, F. (2008). Cost Management For Strategic Business Decisions. The McGrawHill Companies. 370-397.

Hsiao, W. C., D. B. Yntema, P. Braun, D. L. Dunn, and L. Spencer, (1998). ''Measurement and Analysis of Intraservice Work," JAMA(260) : 2361-2370.

Wallace R. Leese, Tim Kizirian (2009). Using Excel's Matrix Operations to Facilitate Reciprocal Cost Allocations, American Journal of Business Education - December 2009 Volume 2, Number 9.

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 203-203

Optimal Indicator of Death for Using Real-World Cancer Patients' Data From the Healthcare System

Suk-Chan Jang[1], *Sun-hong Kwon*[1], *Serim Min*[1],
Ae-Ryeo Jo[1], *Eui-Kyung Lee*[1], *Jin Hyun Nam*[2]

Abstract

Information on patient's death is a major outcome of health-related research, but it is not always available in claim-based databases. Herein, we suggested the operational definition of death as an optimal indicator of real death and aim to examine its validity and application in patients with cancer. Data of newly diagnosed patients with cancer between 2006 and 2015 from the Korean National Health Insurance Service-National Sample Cohort data were used. Death indicators were operationally defined as follows: 1) in-hospital death 2) case wherein there are no claims within 365 days of the last claim. We estimated true-positive rates (TPR) and false-positive rates (FPR) for real death and operational definition of death in patients with high- (lung, liver, and pancreatic), middle- (stomach, skin, and kidney), and low- (thyroid) mortality cancers. Kaplan-Meier survival curves and log-rank tests were conducted to determine whether real death and operational definition of death rates were consistent. The TPR was 97.08% and the FPR was 0.98% in the high mortality group and the overall TPR and FPR were 96.68% and 1.27%, respectively. We showed that there is no significant difference between the real and operational definition of death in the log-rank test for all types of cancers except for thyroid cancer. Defining deaths operationally using in-hospital death data and periods after the last claim is a robust alternative to identifying mortality in patients with cancer. This optimal indicator of death will promote research using claim-based data lacking death information.

Keywords : Operational definition of death, real-world data, claims data, cancer patients.

[1]School of Pharmacy, Sungkyunkwan University, Suwon, 16419, Korea.
[2](Corresponding author) Division of Big Data Science, Korea University Sejong Campus, Sejong, 30019, Korea. E-mail: jinhnam@korea.ac.kr

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 205-205

Systemic Networks for High-Dimensional Exposures, Mediators, and Outcomes*

Jai Woo Lee[1], *Jiang Gui*[2]

Abstract

Developing methods for identifying associations in high-dimensional data and evaluating the methods to detect these in both statistical simulations and real data applications is an area of growing importance in many domains of biomedical science. Network analysis is becoming increasingly recognized as a vital tool for analyzing high-dimensional biomedical data in order to: 1) understand the complex interaction of factors in a single dataset, 2) enable integration of heterogeneous datasets in order to elucidate the impact of factors from one dataset on the other, and 3) predict outcomes based on our understanding of complex structures of variables within datasets. We developed statistical methods to address these issues and tested them through simulations of generalized cases and applied them to real data. The goals of the thesis were to: (1) develop an unsupervised network method that better identifies patterns of dependency and association of features in the data and apply this method to concentrations of trace elements measured in the human placentas, (2) detect the impact of trace elements on the interacting network of metabolites by developing a data integration method, and (3) predict a health outcome using complex structures of elemental and metabolites as mediators by comparing prediction methods with different grouping strategies. The methods developed here can be applied to analyze complex mechanism of other similarly structured biomedical data.

Keywords : Machine Learning, Probabilistic Graphical Models, Computational Algorithms, Statistical Analysis, Exposomics

---

*This article is financially supported by the 2023 College of Public Policy at Korea University.

[1](Corresponding Author) Department of Big Data Science, College of Public Policy, Korea University, Sejong, Republic of Korea; E-mail: jaiwoolee@korea.ac.kr

[2]Department of Biomedical Data Science, Geisel School of Medicine at Dartmouth College, NH, USA

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 207-207

엔벨롭 모델을 이용한 생체의학 데이터분석

박연희[1]

요 약

Cook et al. (2010)에 의해 처음 소개된 엔벨롭 모델은 다변량 선형 회귀의 맥락에서 회귀 계수를 효율적으로 추정하는 방법입니다. 엔벨롭 모델은 충분한 차원 축소 기술을 사용하여 추정 목표에 무의미한 데이터 부분을 식별하고, 이후의 추정은 무의미한 부분을 제거하기에 더 효율적입니다. 엔벨롭 모델은 여러 분야에 적용되었는대, 그 중 그룹별 엔벨롭 모델 (Park et al. 2017)과 부분 예측 변수 엔벨롭 모델 (Park et al. 2022)이 이 발표에서 다룹니다. 첫째로 그룹별 엔벨롭 모델은 서로 다른 그룹에 대해 서로 다른 회귀 계수와 서로 다른 오류 구조를 허용합니다. 남녀 그룹 간의 유전자 변이와 뇌 영상 형질 간의 관련성이 어떻게 다른지 찾기 위해, 다변량 선형 회귀를 위한 그룹별 엔벨롭 모델이 개발되어 다변량 응답과 공변량 간의 관련성을 수립합니다. 두번째로 COVID-19 환자를 대상으로 사이토카인 기반 생체 표지자를 식별하는데, 사이토카인 기반 생체 표지자와 입원시 질병 상태 및 인구 특성을 포함한 환자의 임상 정보 간의 관련성을 나타냅니다. 연속 및 범주형 예측 변수가 함께 있는 경우 추정 효율성을 달성하기 위해 부분 예측 변수 엔벨롭 모델이 개발되었습니다. 엔벨롭 모델과 부분 최소 제곱 (PLS) 간의 연결을 사용하여 부분 예측 변수 엔벨롭 모델은 응답과 범주형 예측 변수에 대한 연속 예측 변수의 조건부 분포에 대한 PLS 회귀를 고려합니다. 엔벨롭 모델을 이용한 데이터 분석은 모형의 추정에서 효과적임을 보여주며, 이는 결과의 명확한 과학적 해석으로 이어집니다.

[1]University of Wisconsin

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 209-209

코로나 19 시기 여성고용의 특성 분석

오경민[1], 권태구[2]

요 약

본 연구는 코로나19가 여성의 고용에 미친 영향을 분석하고자 시도했다. 실증분석을 위해 2017년 8월에서 2022년 8월까지 6개년도에 걸친 경제활동인구조사 근로형태별 부가조사 자료를 인구통계학적 특성에 기반한 셀(cell) 단위 패널 자료로 변환했고, 추정에는 통상적인 패널분석 모형을 적용했다. 분석결과, 코로나19 시기 3년 간 여성의 고용률이 남성에 비해 더 감소하고 실업자는 증가한 것으로 나타났다. 이전 3년 평균과 비교한 여성고용은 2020년도에 가장 큰 감소를 나타냈지만 그 후 점차 회복한 것으로 나타났다. 실업자의 증가와 함께 비경제활동인구는 감소한 것으로 나타났고, 기혼여성, 초대졸 이상 등의 특성을 지닌 여성 집단 또한 같은 특성을 공유한 남성에 비해 고용은 줄고 실업은 증가한 것으로 나타났다. 과거 대부분의 경제위기 시 남성의 고용이 주로 감소했던 것과 대조적으로 이번 위기에서는 여성 고용의 감소가 두드러지게 관찰됐다. 이는 여성이 주로 종사하는 일자리와 특성 그리고 가정 내 자녀돌봄의 부담 등과 관련되어 있을 수 있지만 본고의 분석에서 확인하지 못한 한계를 지닌다.

주요용어: 코로나19, 여성 고용, 일자리 특성

[1]31253 충청남도 천안시 동남구 충절로 1600 한국기술교육대 인력개발학과 대학원생. E-mail: taegoo@koreatech.ac.kr
[2]31253 충청남도 천안시 동남구 충절로 1600 한국기술교육대 인력개발학과 조교수. E-mail: taegoo@koreatech.ac.kr

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 211-215

가산자료 회귀모형을 이용한 중소기업의 경영활동이 국내 특허등록건수에 미치는 영향 연구
- 신용보증기금 기술평가 기업을 중심으로 -

장홍진[1], 홍승열[2], 신지은[3]

요 약

본 연구는 2019년 1월 1일부터 12월 31일까지 신용보증기금에서 기술자산평가를 받은 중소기업을 대상으로 기업의 경영활동이 특허등록에 미치는 영향을 분석하였다. 분석에는 반응변수인 특허건수가 가산자료라는 특성을 고려하여 포아송 회귀모형, 음이항 회귀모형 그리고 일반화포아송 회귀모형을 활용하였다. 본 연구에 사용된 특허건수는 자료의 분산이 평균보다 큰 과대산포가 존재하였고, 모형의 적합도를 AIC로 평가한 결과 음이항 회귀모형과 일반화포아송 회귀모형이 포아송 회귀모형에 비해 우수한 것으로 나타났다. 음이항 회귀모형과 일반화포아송 회귀모형의 성능은 큰 차이가 나지 않았으나, 일반화포아송 회귀모형이 근소한 차이로 보다 우수했다. 일반화포아송 회귀분석결과 기술개발 조직이 전문화 되어 있으며, 기술개발 고급인력 비율이 높을수록 그리고, 연구개발비용이 많을수록 특허등록에 유의한 영향을 미치는 것으로 나타났다. 한편, 영업현금흐름(EBITDA)은 특허등록과 음의 관계를 보였다.

주요용어: 특허권, 포아송 회귀모형, 음이항 회귀모형, 일반화 포아송 회귀모형.

1. 서론

특허권은 기업이 보유하고 있는 기술에 대해 법률이 배타적 독점권을 부여하는 재산권으로 특허권을 가지고 있는 개인이나 기업의 기술력을 측정할 수 있는 중요한 지표로 활용될 수 있다. 일반적으로 특허는 기업의 경쟁우위 확보와 생존에 결정적인 영향을 미치기 때문에 특허등록은 경쟁기업에 비교 우위를 확보하여 기술사업화 역량을 높일 수 있는 기업의 전략적 선택이라고 할 수 있다.

법적으로 배타적 독점권이 보장된다고 해서 중소기업들이 특허권만을 우선적으로 보유하는 것은 아니다. 술에 대한 대표적인 전유방법인 특허와 영업비밀의 상대적 선호와 관련하여, 기업 규모가 큰 대기업과 첨단 기술 보유 기업(설민수, 2014) 및 매출집중도가 높은 소기업과 벤처기업의 경우에는 특허등록보다는 영업비밀을 선호한다(김상신외, 2016).

반면, 기술기업의 정책자금 지원을 위한 기술평가 시 특허출원등록 건수는 계량화가 쉬울 뿐 아니라, 평가 대상 기업의 기술수준을 가늠할 수 있기 때문에 주요한 평가요소에 해당한다. 따라서,

[1]서울시립대학교
[2]서울시립대학교
[3]신용보증기금

기술력은 있지만, 아직 특허를 등록하지 못하거나 영업비밀을 선호하는 중소기업의 경우, 기술평가 시 기술사업화 역량평가에서 좋은 점수를 받지 못할 우려가 있다. 이러한 배경 속에서 기업의 특허권 확보(등록)는 부가적으로 기업의 수익 및 자금지원의 주요한 수단으로 의미를 지닌다.

특허등록이 기업의 경영성과에 미치는 영향을 실증적으로 분석한 기존 연구는 많으나, 개별기업의 경영성과가 특허등록에 있어서 미치는 영향을 분석한 국내 연구는 상대적으로 부족하다. 본 연구는 특허권을 등록하는데 필요한 경영활동 요인들 간의 인과관계를 실증 분석하여 특허등록에 영향을 미치는 기업의 경영활동을 선별하여 이를 기술평가에 활용하는 정책적 대안을 제시할 뿐만 아니라, 경영전략 등의 사유로 특허권 대신 영업비밀을 선호하는 기술기업의 사업화 역량도 특허권 보유기업과 동일하게 평가할 수 있도록 기술평가모형의 평가방법론에 대한 개선방안을 제시하고자 한다.

2. 연구자료

본 연구에서 활용한 자료는 2019년 한 해 동안 신용보증기금(KODIT, Korea Credit Guarantee Fund)에서 기술자산평가를 통해 보증심사를 받은 중소기업 1,783개이다. 본 연구에서는 특허등록건수와 기업의 재무적 성과를 연계하여 분석하기 위해 연속 2기 이상의 재무제표를 보유한 제조업, 정보통신업, 전문·과학 및 기술서비스업을 영위 중인 939개 기업을 분석대상으로 정의하였다.

본 연구의 목적은 기업의 경영활동이 특허등록에 미치는 영향을 분석하는 것으로 특허등록건수를 반응변수로, 기업의 재무적 또는 비재무적 요소를 설명변수로 구성하다.

특허등록건수는 기술자산평가를 위해 신용조사서에 등록된 건수를 사용하였다. 따라서, 심사관이 등록결정 후 특허청에 3년치 특허료를 납부(설정등록)하여 특허등록번호가 발행된 특허권 수를 의미한다.

<표 1> 변수구성

구분		변수명	비고
반응변수		특허등록건수	2019년 ~ 2020년 중 특허등록 건수
비재무적 요소	기업특성	업종	0 : 제조업 1 : 정보처리업 2 : 전문과학 및 기술서비스업
		총자산(기업규모)	단위 : 백만원
		업력	단위 : 日
		종업원 수	종업원 수
	기술조직	기술개발조직 유형	1 : 기업부설연구소, 2 : 기술개발전담부서 3 : 기타연구조직 4 : 해당없음
		기술개발조직 운영기간	1 : 2년 이상. 0 : 2년 미만
	기술인력	기술직비율	기술직종업원/종업원수
		숙련기술직 비율	숙련기술직종업원/기술직종업원

구분		변수명	비고
재무적요소	유동성	유동비율	유동자산/유동부채
		비유동비율	비유동자산/자기자본
	안정성	부채비율	총부채/자기자본
		차입금의존도	(장기 + 단기) 차입금 / 총자본
	수익성	영업이익률	영업이익/매출액
		ROA	당기순이익/총자산
		매출원가율	매출원가/매출액
	활동성	재고자산 회전율	매출액/재고자산
		매출채권 회전율	매출액/매출채권
	현금흐름	EBITDA	당기순이익 + 이자비용 + 세금과공과 + 감가상각비 + 무형자산 상각비
	성장성	총자산 증가율	(당기 총자산 - 전기 총자산)/전기 총자산
		매출액 증가율	(당기 매출액 - 전기 매출액)/전기 매출액
		영업이익 증가율	(당기 영업이익 - 전기 영업이익)/전기 영업이익
	기술개발활동	연구개발비	(손익계산서) 연구비+경상연구개발비+경상개발비 + (제조원가명세서) 연구비 및 경상개발비
		유형자산 투자액	(당기 유형자산 - 전기 유형자산) + (손익계산서) 감가상각비 + (제조원가명세서) 감가상각비

3. 실증분석

포아송 회귀모형에서는 추정치의 표준오차가 과소추정되어 몇 몇의 설명변수들이 더 유의하게 나타났다. 이론적으로 포아송 회귀모형의 추정모수 분산 $Var(\hat{\beta}) = (\hat{I}_{\beta\beta})^{-1} = (\sum_{i=1}^{n} \mu_i x_i x_i')^{-1}$ 은 음이항 회귀모형의 추정모수 분산 $Var(\hat{\beta}) = (\hat{I}_{\beta\beta})^{-1} = (\sum_{i=1}^{n} \frac{\mu_i x_i x_i'}{1+\tau\mu_i})^{-1}$에 비해 $1+\tau\mu_i$ 배 만큼 작은 값을 가진다. 즉, 과대산포가 있는 자료에서 포아송 회귀모형을 적용할 경우 회귀계수의 표준오차가 과소추정되기 때문에, 실제로는 유의하지 않은 변수임에도 유의한 변수로 추정된다.

실증분석 결과 포아송회귀모형에서는 총자산(기업규모)이 클수록, 업력이 짧을수록, 개술개발조직이 전문화 되어 있을수록, 숙련기술직 비율이 높을수록, 총자산 증가율과 매출액 증가율이 높을수록, 연구개발비가 많을수록, 그리고 유형자산 투자액이 작을수록 특허등록건수에 유의한 영향을 주는 것으로 나타났다. 포아송회귀모형은 표준오차가 과소 추정되기 때문에 다른 두 모형에 비해 유의한 변수가 많았으며, 총자산, 업력, 총자산 증가율, 매출액 증가율, 유형자산 투자액과 같은 변수는 포아송회귀모형에서만 유의한 변수로 추정되었다. 한편, 김진영, 윤우진 (2009)의 연구에서는 종업원 수가 특허등록건수와 음의 관계를 보였지만, 본 연구에서는 유의하지 않았다.

음이항 회귀모형과 일반화포아송 회귀모형은 모두 기술개발조직이 전문화 되어 있을수록, 숙련기술직 비율이 높을수록, 그리고 연구개발비가 많을수록 특허등록에 유의한 것으로 나타났다. 한편

특허등록건수와 음의 관계를 보이는 EBITDA는 포아송 회귀모형과 음이항 회귀모형에서는 유의하지 않았으나, 일반화포아송 회귀모형에서는 유의수준 10% 하에서 유의하였으며(p-value : 0.06), 이는 정성창, 김영환 (2016)의 연구결과와 유사하다.

<표 2> 각 모형별 표준화 회귀계수 추정치

변수			포아송 회귀모형		음이항 회귀모형		일반화포아송 회귀모형	
구분		절편 항	-1.14***	(0.06)	-1.13***	(0.08)	-1.10***	(0.08)
비재무적 요소	기업특성	업종	0.00	(0.06)	0.07	(0.09)	0.09	(0.1)
		총자산(기업규모)	0.08*	(0.03)	0.10	(0.08)	0.86	(0.47)
		업력	-0.13*	(0.06)	-0.11	(0.09)	-0.13	(0.1)
		종업원 수	-0.10	(0.07)	-0.07	(0.11)	-0.02	(0.14)
	기술조직	기술개발조직 유형	-0.50***	(0.07)	-0.48***	(0.1)	-0.47***	(0.1)
		기술개발조직 운영기간	-0.07	(0.06)	-0.02	(0.09)	-0.01	(0.09)
	기술인력	기술직비율	-0.10	(0.06)	-0.07	(0.08)	-0.06	(0.08)
		숙련기술직 비율	0.21***	(0.05)	0.20**	(0.08)	0.20**	(0.08)
재무적요소	유동성	유동비율	0.04	(0.05)	-0.02	(0.08)	-0.03	(0.1)
		비유동비율	0.04	(0.07)	0.02	(0.1)	0.00	(0.11)
	안정성	부채비율	-0.02	(0.08)	-0.05	(0.13)	-0.04	(0.15)
		차입금의존도	-0.06	(0.08)	-0.03	(0.12)	-0.03	(0.11)
	수익성	영업이익률	0.01	(0.05)	0.13	(0.09)	0.19	(0.11)
		ROA	-0.04	(0.06)	-0.11	(0.09)	-0.07	(0.14)
		매출원가율	-0.07	(0.06)	-0.03	(0.1)	-0.05	(0.11)
	활동성	재고자산 회전율	0.06	(0.05)	0.04	(0.08)	0.04	(0.07)
		매출채권 회전율	-0.04	(0.05)	-0.02	(0.07)	-0.02	(0.08)
	현금흐름	EBITDA	0.06	(0.06)	-0.05	(0.11)	-0.37	(0.2)
	성장성	총자산 증가율	0.11*	(0.05)	0.16	(0.08)	0.17	(0.11)
		매출액 증가율	0.12**	(0.04)	0.13	(0.07)	0.14	(0.08)
		영업이익 증가율	-0.01	(0.05)	-0.05	(0.07)	-0.05	(0.08)
	기술특성	연구개발비	0.28***	(0.04)	0.29***	(0.08)	0.30**	(0.09)
		유형자산 투자액	-0.17*	(0.07)	-0.10	(0.09)	-0.10	(0.11)
산포					0.45***°*	(0.06)	0.94***	(0.13)

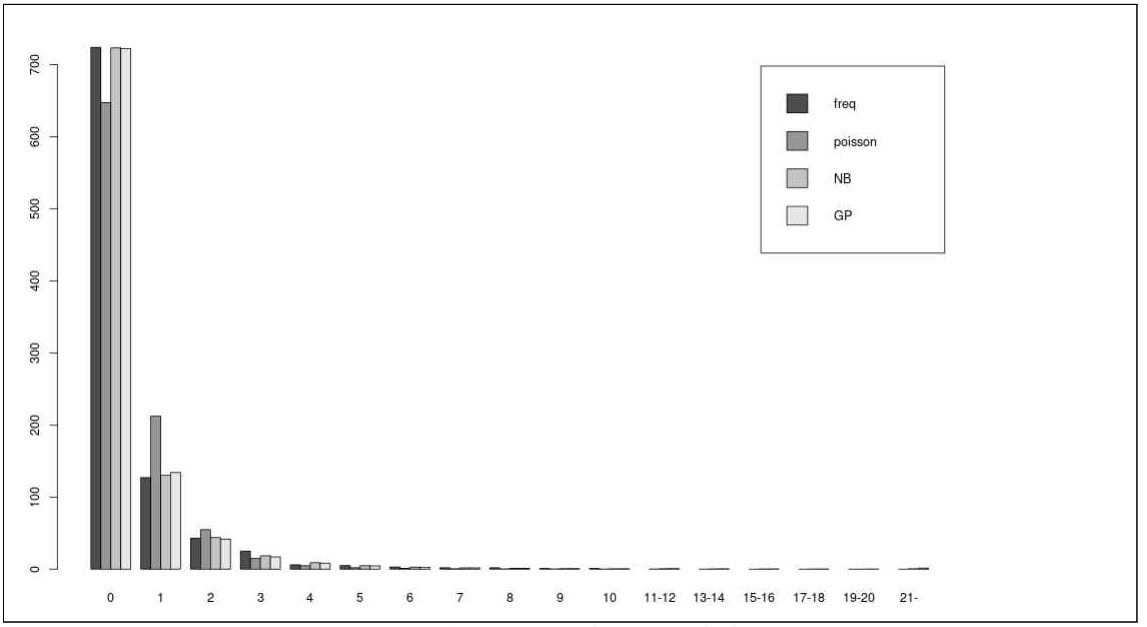
<표 3> 각 모형별 성능지표 비교

구분	포아송 회귀모형	음이항 회귀모형	일반화포아송 회귀모형
log-likelihood	-839.01	-733.04	-732.73
AIC	1726.02	1516.08	1515.48

<표 3>은 각 모형 별 성능 지표인 로그가능도 값과 AIC (Akaike Information Criteria)의 결과이다. 로그가능도 값은 일반화 포아송회귀모형이 근소한 차이로 가장 크고, 음이항 회귀모형, 포아송 회귀모형의 순으로 값이 작아지므로, 일반화 포아송회귀모형의 적합력이 가장 우수하다. AIC 값 역시 포아송 회귀모형이 상대적으로 큰 값을 보이며, 음이항 회귀모형과 일반화포아송 회귀모형은 비슷한 수준이지만, 일반화포아송 회귀모형이 근소한 차이로 작은 값을 가지기 때문에 가장 성능이 좋다. 즉 로그가능도 값과 AIC 통계량 기준 모두 일반화 포아송회귀모형이 다른 두 모형에 비해 성능이 우수하다고 할 수 있다.

* R 프로그램의 glm.nb 함수에 따라 산포모수의 역수를 추정한다.

전체 특허등록건수의 95%를 차지하는 특허등록건수 2건 이하의 각 모형별 예측확률을 보면 음이항 회귀모형과 일반화포아송 회귀모형의 예측확률은 실제 특허등록건수와 유사하게 나타난다. 반면, 포아송 회귀모형의 경우 특허등록건수가 0인 경우 실제보다 작은 값을 보이며, 특허등록건수 1 또는 2인 경우는 실제보다 많은 값을 보인다. 이를 통해 상대도수 혹은 빈도를 추정하는 관점에서도 과대산포를 반영한 음이항 회귀모형과 일반화포아송 회귀 모형의 추정결과가 보다 우수하다고 할 수 있다.



<그림 1> 회귀모형별 특허건수의 예측확률

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 217-221

한국 주택가격의 동태적 변화 연구

고동우[1], 윤성민[2]

요 약

주택은 의식주의 하나로서 삶의 여건을 결정하는 필수재이지만, 주택가격은 거시 경제 여건 등에 따라 그 등락의 폭이 커서 중요한 연구대상이다. 따라서 최근 주택가격의 동태적 변화를 다각적으로 분석한 바, 결론은 다음과 같다. 첫째, 주택가격 상승은 주로 주택대출 증가 및 주택 공급 감소에 기인하는 측면이 있다. 둘째, 최근 팬데믹에 따른 재정지출 및 금융완화가 주택가격 상승에 일조한 것으로 추정된다. 셋째, 주택가격 변동은 관성적 환류효과를 갖는다. 특히, 주택가격 변동 시 그 흐름을 확장시키는 변동성 군집 및 충격의 지속성이 유의한 것으로 판별된다. 넷째, 주택대출은 이전 가격변동의 환류와 함께 단기적으로 수요측면에 양(+)의 영향을 미치며, 주택건설인허가나 미분양주택수와 같은 주택공급 변수는 장기적으로 음(-)의 영향을 확대한다.

주요용어 : 주택가격, 변동성 군집, 주택대출, 주택건설인허가, 미분양주택

1. 이론적 배경 및 선행연구

세계적으로 주택가격은 적지 않은 변동성에 노출되어왔다. 따라서 주택의 경제적 함의에 대한 폭넓은 관심과 함께 주택가격 결정요인이나 그 변동성의 배경에 대하여 다양한 연구가 수행된다. 이 연구는 주택가격 변동요인을 수요와 공급 측면에서 주요 변수를 구성하여 분석하고자 거시경제 분야의 연구에 우선 주목했다. 따라서 금리 등 통화정책이나 신용팽창 여건에 대한 연구를 집중 조사했다(Hofmann, 2001; Mian and Sufi, 2018; 윤성민·강주화, 2018).

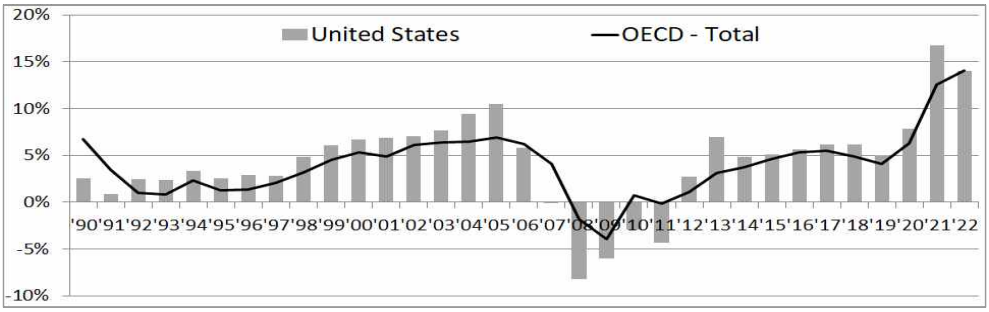


Figure 1. Dynamics of global housing price index. Source: OECD (https://stats.oecd.org/)

2000년대 이후 주택가격은 금융시장과 연관 지점에서 의미 있는 연구 결과들이 보고되고 있다. Hofmann(2001)은 금리와 연계된 주택대출 활용이 레버리지를 활용한 주택구매력을 뒷받침한다고

[1]46241 부산시 금정구 부산대학로 63, 부산대학교 글로벌경제컨설팅전공 박사과정. E-mail: dwko.eco@gmail.com
[2](교신저자) 46241 부산시 금정구 부산대학로 63, 부산대학교 경제학부 교수. E-mail: smyoon@pusan.ac.kr

주장한다. Mian and Sufi(2018)는 최근 40여년간 세계적 신용팽창이 이끄는 가계수요 주도의 경기변동이 지배적이었음을 언급한 바 있다. 특히, 주택시장과 신용공급 사이의 피드백 효과가 의심할 여지없음을 지적한다. 또한 윤성민・강주화(2018)는 2002년부터 2017년까지 15년간 가계부채와 주택가격 및 주요 거시경제 변수 간 동태적 관계를 분석한 바 있다. 가계부채는 주택가격, 유동성, 저금리 등으로부터 영향 받고, 가계부채 변화는 주택가격 변동에 영향을 미친다.

한편, 미시적 재무이론 관점에서 자산가격과 내재가치 차이에 주목한 연구 역시 현재 자산시장 팽창에 시사점을 제시한다(Campbell and Shiller, 1988). 특히, 주택가격에 대한 낙관적 기대는 주택자산에 대한 추가적 가격상승 환류로 이어져 비이성적 과열로 귀결될 수 있다(Shiller, 2006)

2. 표본자료와 분석방법

이 연구의 실증분석 대상은 2006년 1분기부터 2023년 2분기에 이르는 17년 6개월간의 분기별 주택매매가격지수, 주택건설호수(인허가), 주택담보대출잔액 등 시계열 자료(T=70)를 포괄한다. 경제환경 변수인 국내총생산(GDP), 주택수요 측면의 경제활동인구(EP), 주택담보대출(HL), 주택공급 측면 주택건설호수(인허가)(PM), 미분양주택수(UN), 주택거래량(VL)을 적용한다. 시계열에 내재된 계절성(seasonality)을 완화시키기 위해 분석에 사용된 모든 변수는 Census X-12 방법으로 계절 조정한다.

본 연구는 우선, 주택가격을 종속변수로 지정하고 회귀분석을 통해 유효한 결정요인을 식별한다. 또한 GARCH 모형으로 주택가격 변동성의 특징을 살펴 최근 주택시장의 불안정성에 대한 시사점을 구한다. 나아가 그레인저 인과 검정 등을 통해 주요 변수 간 장기 선형관계를 밝혀나가는 가운데 주택수요와 공급 측면을 종합적으로 포괄하는 동태적 상관관계(VECM) 모형을 제안한다. 이 연구는 주택수요 및 공급 변수를 종합하는 가운데 축약된 모형을 제시함에 있어 주된 연구 차별성이 있다.

Table 1. Descriptive statistics

Variables	Definnition	Mean	St.d.	Max	Min
HP	Housing price index(100=Jan,22)	74.5	11.5	101.3	54.6
GDP	GDP (Bill Won)	403,830	59,076	497,214	299,314
EP	Population(Thousand)	26,558.1	1,647.9	29,493.0	23,501.0
HL	Mortgage loans (Bill Won)	1,172,795	422,968	1,871,108	550,330
PM	Housing permit	130,823	52,839	275,302	33,944
UN	Unsold home	68,389	35,765	165,641	13,842
VL	Volume of housing transactions	347,009	93,637	541,520	184,722

3. 분석 결과

우선, 전국 주택매매가격(HP)를 기준으로 회귀분석을 실시했다. 이는 경제전반(GDP)의 상황을 고려하는 가운데 주택수요(EP, HL)와 주택공급(PM, UN) 요인 및 주택경기(VL) 현황을 종합한다. 오차 자기상관을 완화하기 위하여 Prais and Winsten(1954) 추정방법을 적용하여 회귀모형을 추정했다. 주택가격에 영향을 주로 미치는 요인은 주택담보대출(HL) 및 주택거래(VL) 요인으로 나타난다.

실물경제(GDP) 상황의 경우 오히려 역행하는 현황을 보여준다. 주택공급(PM, UN) 요인의 경우 방향성은 이론과 실제에 부합하나, 그 통계적 유의성은 명확히 드러나지 않는다. 그라고 팬데믹에 따른 모의변수(CVD)는 주택가격에 대하여 유의한 양(+)의 영향을 미치는 것으로 확인된다.

나아가 주택가격 수준이 높고 변동을 주도하는 수도권을 포함하여 전국과 수도권 주택가격지수 변동성을 대상으로 AR(1)-GARCH(1,1) 모형을 추정했다. Table 3.은 AR(1)-GARCH(1,1) 모형 개념과 해당 모형을 적용한 분석결과를 보여준다. 주택가격지수 변동(Yt)에 영향을 미치는 설명변수인 전기 주택가격 변동(Yt-1) 계수β0의 통계적 유의성이 확인된다. 즉, 전기의 주택가격 변동이 당기의 주택가격에 영향을 미치는 현상이 실재한다고 판단된다. 나아가 분산방정식에서 전국과 수도권 모두 변동성 충격의 지속성 현상을 나타내는 β1값은 유의하게 나타난다. (ɣ+β1)값은 0.9901로 1에 근사하게 나타나므로 그 충격의 지속성에 대한 안정성 요건에 부합한다.

Table 2. Regression analysis result of basic model

		Dependent variable = HP			HP – Covid dummy		
		Coef.	t-stat	VIF	Coef.	t-stat	VIF
Intercept		-0.6280	-1.023	-	-0.7235	-1.470	-
국내총생산	GDP	-0.1993	-1.165	1.498	-0.2458	-1.837*	1.540
경제활동인구	EP	0.2459	0.734	1.442	0.0949	0.283	1.447
주택담보대출	HL	0.7567	2.879***	1.147	0.7681	3.382***	1.147
주택건설인허가	PM	-0.0037	-0.732	1.152	-0.0039	-0.795	1.152
미분양주택	UN	-0.0073	-1.114	1.260	-0.0063	-1.000	1.302
주택거래량	VL	0.0093	1.714*	1.221	0.0101	1.870*	1.275
팬데믹 더미	CVD	-	-	-	2.365	3.587***	1.101
Ad. R^2			0.519			0.583	
D-W			2.01			1.89	

주: ***, **, * 기호는 각각 1%, 5%, 10% 수준에서 유의하다는 것을 표시한다.

Table 3. Result of GARCH analysis

Mean equation			$Y_t = \alpha_0 + \beta_0 Y_{t-1} + \epsilon_t \ , \quad \epsilon_t \mid I_{t-1} \sim N(0, h_t)$					
Variance equation			$h_t = \alpha_1 + \gamma \epsilon_{t-1}^2 + \beta_1 h_{t-1}$					
			HP (전국)			HP met (수도권)		
			Coef.	SE	z-stat	Coef.	SE	z-stat
Mean equation	α0	Intercept	0.5251	0.2118	2.478**	0.4405	0.3740	1.238
	β0	AR(1)	0.7285	0.0671	10.86***	0.7695	0.0558	13.78***
Variance equation	α1	Intercept	0.0173	0.0187	0.921	0.0361	0.0243	1.482
	ɣ	Resid(-1)2	0.4148	0.1925	2.155**	0.3283	0.2097	1.669*
	β1	GARCH(-1)	0.5753	0.1171	4.910***	0.5966	0.1459	4.087***
Ad. R^2				0.388			0.286	
Log likelihood				-69.61			-88.02	
D-W				2.20			2.50	
Ljung-Box Qs's p				0.97			0.32	
ARCH LM's p				0.97			0.33	

주: ***, **, * 기호는 각각 1%, 5%, 10% 수준에서 유의하다는 것을 표시한다.

다음 Table 4.은 Johansen 공적분 검정 결과를 보여준다. 공적분관계 검정결과 검정통계량에 대해 5% 유의수준 기준으로 공적분 관계가 없다는 귀무가설은 기각된다. 다만, 1개 이하의 공적분 관계가 있다는 귀무가설은 기각되지 않아, 적어도 최소 1개의 공적분 관계가 존재한다고 간주할 수 있다.

Table 4. Johansen 공적분 검정 결과

귀무가설	Eigenvalue	Trace	Critical value(5%)	p-value
None	0.44939	66.4304	47.8561	0.00040
≤1	0.21526	25.8535	29.7970	0.13320
≤2	0.10982	9.3700	15.4947	0.33230
귀무가설	Eigenvalue	max-eigen	Critical value(5%)	p-value
None	0.44939	40.5769	27.5843	0.00060
≤1	0.21526	16.4834	21.1316	0.19790
≤2	0.10982	7.91067	14.2646	0.38780

주택시장의 주요 변수 간 공적분 관계를 확인함에 따라 본 연구는 VECM(vector error correction model, 벡터오차수정모형) 방법을 활용하여 주택가격(HP)과 주택담보대출(HL), 주택건설인허가(PM), 미분양주택(UN) 간 동태적 상관관계를 다음 Table 5. 와 같이 분석했다. 변수 간 장기 균형관계를 보여주는 공적분 관계식에 따르면 주택가격(HP)은 이전 주택가격(HP). 주택건설인허가(PM), 미분양주택(UN)으로부터 유의한 영향(1% 유의수준)을 받는다.한편, 단기 회복력을 나타내는 오차수정계수(ECT)를 보면, 주택가격(HP)이나 주택대출(HL)의 오차수정계수는 통계적으로 유의(1%)하다.

Table 5. 벡터오차수정모형(VECM) 추정결과

종속변수	HP(주택가격)	HL(주택대출)	PM(주택건설)	UN(미분양주택)
Cointegration Equation	1.0000	-0.2581 [-7.8847]***	0.2001 [5.7414]***	0.1256 [5.3854]***
차분종속변수	D(HP)	D(HL)	D(PM)	D(UN)
Error Correction ECT(-1)	-0.1106 [-4.427]***	-0.0395 [-3.330]***	-1.7039 [-3.457]***	1.5838 [3.848]***
D(HP(-1))	0.4085 [4.226]***	-0.1386 [-3.020]***	-1.7228 [-0.903]	3.3238 [2.088]**
D(HL(-1))	0.2254 [1.415]	0.7736 [10.220]***	1.1425 [0.363]	1.4606 [0.556]
D(PM(-1))	0.0263 [3.971]***	0.0173 [5.521]***	-0.0580 [-0.444]	0.0038 [0.035]
D(UN(-1))	0.0131 [1.425]	0.0081 [1.939]*	0.2281 [1.314]	0.0451 [0.311]
C	0.5106 [4.415]***	0.1875 [3.415]***	7.8484 [3.442]***	-7.3595 [-3.864]***
Adj. R2	0.5681	0.6783	0.1496	0.2369
D-W	2.06	2.58	2.00	2.22
Log-likelihood	571.12			

주: [ ] 은 t-통계값으로, ***, **, * 기호는 각각 1%, 5%, 10% 수준에서 유의하다는 것을 표시함

References

윤성민 · 강주화(2018). 방향성 정보전이지수를 이용한 가계부채와 주택가격 사이의 시간가변 파급효과 분석, *주택연구*, 26(4), 151-181.

Hofmann, B. (2001). The determinants of private sector credit in industrialised countries: Do Property Prices Matter?, *BIS Working Papers*, 108.

Mian, A., Sufi, A. (2018). Finance and business cycles: the credit-driven household demand channel, *Journal of Economic Perspectives*, 32(3), 31-58.

Shiller, R. (2006). Long-term perspectives on the current boom in home prices, *The Economists' Voice*, 3(4), 1-11.

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 223-223

프로스포츠 경기의 인근 상권에 대한 경제적 영향: 프로야구 개최 여부에 따른 소비 차이 분석

강미지[1], 문형빈[2]

요 약

프로스포츠 경기가 개최되는 경기장 인근 상권의 매출은 경기의 개최여부에 영향을 받는다. 일반적으로 인근 상권의 매출은 경기가 열릴 때 증가하는 경향을 보이며, 이는 프로스포츠 경기의 개최가 지역경제에 미치는 긍정적인 영향이라고 할 수 있다. 그러나 프로스포츠가 지역경제에 미치는 영향을 정량화한 선행 연구가 부족한 상황이며, 일부 관련 연구의 경우에도 업종과 경기 시점 등의 세부요인을 고려하지는 못하고 있는 실정이다. 따라서 본 연구에서는 프로스포츠 경기가 지역경제에 미치는 영향이 업종, 경기 시점, 소비자 특성 등의 세부 요인에 따라 어떻게 차별되는지를 정량적으로 분석하고자 하였다. 구체적으로, 본 연구는 프로야구를 대상으로 하며 분석 지역은 부산, 분석 기간은 2019년 4월~8월로 하였다. 인근 상권의 매출액을 측정하기 위하여 부산 사직야구장 인근 지역을 정의하고, 해당 지역의 신용카드 지출 데이터를 활용하여 분석하였다. 이때, 세부요인으로 업종과 경기 시점, 소비자의 인구통계적 특성을 구분하여 살펴보았으며, 요인별로 통계적으로 유의미한지 분석하였다. 연구 결과, 부산의 사직야구장에서 경기가 있는 날에는 그렇지 않은 날보다 평균적으로 소비가 증가한다는 것이 식별되었다. 특히 업종별 분석 결과, 편의점, 음/식료품, 기타요식, 할인점/슈퍼마켓, 제과/커피/패스트푸드, 일식/중식/양식 순으로 소비의 유의미한 증가가 확인되었다. 이러한 패턴은 월별로 차이를 보였으며, 특히 6월에는 다른 월과 달리 음/식료품에서 가장 큰 차이를 보였다. 요일별 분석 결과에서는 금요일이 다른 요일과 차이를 보였으며, 소비자의 특성에 다라 분석한 결과, 성별과 연령에 관계없이 편의점 소비의 증가가 관찰되었다. 이러한 연구 결과는 프로야구 경기의 개최 여부가 해당 지역 경제에 미치는 영향을 정량적으로 입증하였다는 데 의의가 있다. 특히, 향후 COVID-19, MERS 등과 같은 예상치 못한 팬데믹 상황에서 프로스포츠 경기 중단 및 축소가 발생하는 경우, 인근 상권에 대한 경제적 손실을 지원하는 정책을 수립하는 과정에서 중요한 근거로 활용될 수 있다는 점에서도 의의가 있다고 할 수 있다.

[1]48513 대한민국 부산광역시 남구 용소로 45 (대연동, 부경대학교대연캠퍼스) 창의관(D15) 907-2호, 부경대학교 자연과학대학 응용수학과 학부재학생. E-mail: meejy100@naver.com
[2]48513 대한민국 부산광역시 남구 용소로 45 (대연동, 부경대학교대연캠퍼스) 창의관(D15) 1118호, 부경대학교 정보융합대학 빅데이터융합전공 조교수. E-mail: hbmoon@pknu.ac.kr

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 225-225

Regression Trees for Zero-Inflated Count Data

Jeonghwan Kim[1], *Hyungjun Cho*[2]

Abstract

The decision tree is a data mining method that divides data into a tree structure for regression or classification. It involves creating a single decision tree by repeatedly partitioning the space concerning a given response variable and making predictions. Due to its characteristic of categorizing in a rectangular form, the decision tree is user-friendly and can analyze data regardless of whether it is linear or nonlinear. This advantage has led to various types of research being conducted. Some studies have also been carried out on count data with zero inflation. However, the existing decision tree methods for zero-inflated count data have the limitation that they are only applicable to data where the zero generation process is distinctly divided into two stages. This presents a drawback in that they cannot clearly explain the process of zero. Therefore, this paper proposes a decision tree to analyze data where the process of zero is categorized into a single form. Additionally, in the proposed tree model, a residual analysis method was implemented, which does not require checking all possible cases when selecting categorical variables. The model's performance was verified through simulation experiments, and the proposed model exhibited the best performance. The model's utility was also confirmed through the application to real data.

Keyword : Decision tree, Zero-Inflated count data, Residual analysis

[1]Korea University(Naval Force Analysis Test Evaluation Group), 145, Anam-ro, Seoungbuk-Gu, 02841, Seoul, Korea, E-mail: kjhsjy0215@korea.ac.kr
[2]Professor, Department of the Statistics, Korea University, 145, Anam-ro, Seoungbuk-Gu, 02841, Seoul, Korea, E-mail: hj4cho@korea.ac.kr

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 227-227

Additive Regression under Low-Rank Structure*

Kwan-Young Bak[1], *Donghwi Nam*[2], *Ja-Yong Koo*[3]

Abstract

This paper reports on our study of a multivariate nonparametric reduced-rank regression method within an additive model framework. The nuclear norm of component functions is penalized to incorporate inherent low-dimensional structure into the estimation process. The proposed penalization scheme introduces sparsity to the singular values of coefficient matrices of basis functions, reducing the rank of the coefficient matrices. The proposed method is implemented with the backfitting and the alternating direction method of multipliers algorithm, an efficient convex optimization algorithm that addresses the complex constrained problem by decomposing it into simpler subproblems. Simulation studies are conducted, and the results are compared with the least squares estimator to demonstrate the effects of considering low-dimensional structure. To demonstrate the practicality of the proposed method, we apply it to the gene expression data.

Keywords: Low-rank structure, Additive model, Reduced-rank regression, Alternating direction method of multipliers algorithm.

---

*The work of K.-Y. Bak was supported by National Research Foundation (NRF) of Korea, RS-2022-00165581. The work of J.-Y. Koo was supported by National Research Foundation (NRF) of Korea, RS-2023-00253020 and RS-2023-00219212.

[1]School of Mathematics, Statistics and Data Science, Sungshin Women's University, 2, Bomun-ro,, 34da-gil, Seongbuk-gu, Seoul 02844 Republic of Korea. kybak@sungshin.ac.kr.

[2]Department of Statistics, Korea University, Anam-ro 145, Seongbuk-Gu, Seoul 02841 Republic of Korea. ndonghwi16@korea.ac.kr

[3](Corresponding author) Department of Statistics, Korea University, Anam-ro 145, Seongbuk-Gu, Seoul 02841 Republic of Korea. jykoo@korea.ac.kr.

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 229-229

Linear quantile regression for doubly-censored data via adaptive loss function*

Seohyeon Park[1], *Yeji Kim*[2], *Sangbum Choi*[3]

Abstract

In many biomedical studies, we frequently encounter doubly-censored data, which includes exact, left-censored, and right-censored observations. To address the challenges posed by the heterogeneity of data, a linear censored quantile regression has been widely explored. This paper introduces a newly modified loss function across various quantiles for the doubly-censored data. The proposed approach considers both fully observed and censored samples, incorporating censoring arguments into the usual check function. Thus, this integration enhances efficiency due to the exploitation of all available information, even in scenarios with high censoring rates. Moreover, this method accommodates both covariate-independent and dependent censoring by utilizing the Beran estimator to estimate the conditional survival probability of censoring. To mitigate issues related to non-convex loss functions leading to local minima, we use a slight variation of the Majorization-Minimization (MM) algorithm to estimate parameters. Inferences are made using a percentile bootstrap approach, and asymptotic properties are established. Extensive numerical studies demonstrate the finite sample performance and German administrative unemployment duration data is applied to illustrate the practical applicability of our method.

Keywords : Beran estimator, Censored quantile regression, Double censoring, Majorization-Minimization algorithm, Survival analysis

---

*This work was in part supported by the National Research Foundation (NRF) of Korea, under grant 2022M3J6A1063595, 2022R1A2C1008514.

[1]Graduate Student, Department of Statistics, Korea University, Anam-ro, Seongbuk-gu, Seoul, 02841, Korea. E-mail: seohyeon20@korea.ac.kr

[2]Graduate Student, Department of Statistics, Korea University, Anam-ro, Seongbuk-gu, Seoul, 02841, Korea. E-mail: yeji_kim@korea.ac.kr

[3](Corresponding Author) Associate Professor, Department of the Statistics, Korea University, Anam-ro, Seongbuk-gu, Seoul, 02841, Korea. E-mail: choisang@korea.ac.kr

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 231-231

Scalable Algorithm for kernel machines via lower rank Linearization

Yukung Kim[1], *Jeongeun Sim*[2], *Seung Jun Shin*[3]

Abstract

The kernel trick has been a canonical and general tool for learning complicated models from data. In the era of bigdata, however, the kernel machine becomes often infeasible due to its prohibitively large computation cost. For example, the kernel support vector machine (SVM) requires $O(n^3)$ flops to learn the machine from the data whose sample size is n. There are various studies to reduce the computational cost of the kernel machine, and Lan et al. (2019) proposed a lower-rank linearization approach to develop a scalable algorithm for the kernel SVM. In this article, we apply the idea of Lan et al. (2019) to various kernel machines, such as kernel ridge regression, kernel quantile regression, kernel logistic regression, and kernel support vector regression. Our numerical experiment shows that the lower-rank linearization approach greatly reduces the computational cost of various kernel machines while preserving prediction accuracy.

Keywords : Kernel Machine, Large-scale Learning, Supervised Learning.

[1]Graduate Student, Department of Statistics, Korea University, Anam-ro, Seongbuk-gu, Seoul, 02841, Korea. E-mail: kdbrud@korea.ac.kr

[21]Graduate Student, Department of Statistics, Korea University, Anam-ro, Seongbuk-gu, Seoul, 02841, Korea. E-mail: sim0348@korea.ac.kr

[3](Corresponding Author) Associate Professor, Department of the Statistics, Korea University, Anam-ro, Seongbuk-gu, Seoul, 02841, Korea. E-mail: sjshin@korea.ac.kr

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 233-233

탄소중립을 위한 수소 생산 및 활용 기술의 미래 트렌드 예측: 거대언어모형 기반 특허분석

김민규[1], 이근우[1], 이주용[2]

요 약

2015년 파리협약 이후, 기후변화 및 탄소중립에 대한 국제적인 대응의 필요성이 대두되고 있다. 전 세계적으로 탄소배출 증가에 가장 큰 요인으로 작용하는 화석연료 소비에 대한 의존도를 줄이고 신재생에너지 비중을 증대하는 정책을 추진하고 있으며, 2020년 유럽연합은 탈탄소화 목표를 달성하는 데 있어 수소의 핵심 우선순위를 강조하였다. 따라서, 본 연구는 국제 등록 특허를 중심으로 2013년부터 2022년까지의 지난 10년간 수소 생산 및 활용 관련 기술의 동향과 잠재적인 기술의 식별을 토픽 모델링을 통해 수행하여 탄소중립을 위한 기술적 시사점을 도출한다. 전통적인 텍스트 마이닝 기법인 LDA와 NMF는 단어의 출현 빈도에만 기반하여 단어 간의 의미있는 상관관계를 반영하지 못하기 때문에, 본 연구는 양방향 임베딩을 통해 문맥을 고려하여 표현을 생성하는 거대언어모형인 BERT를 이용하여 특허 분석 및 토픽 모델링을 시행하였다. 분석 결과, 최근 3년간 기술 개발 추세가 가장 가파르게 증가한 토픽들은 바이오매스의 리그닌을 활용하여 수소를 추출하는 기술과 플라스몬 광촉매, 실리콘 광촉매와 같은 광촉매의 개발을 통해서 태양 에너지를 효율적으로 이용하여 수소를 생산하는 기술이었다. LNG 발전 기술 대상으로 추가적으로 토픽 모델링을 수행한 결과, 수소를 첨가하여 탄소를 저감하고, 효율을 높일 수 있는 수소 혼소 가스터빈 기술에 대한 추세가 최근 가장 가파르게 증가한 것을 도출하였다. 실제로, 이 기술들은 2023년에 가장 활발하게 개발이 이루어지고 있는 수소 생산 및 활용 관련 기술들임을 실증하였다. 따라서 본 연구는 수소 활용 기술에 대한 특허 분석 결과를 바탕으로, 기술 세분화나 토픽 모델링에 주력하는 전통적인 특허 분석 방법론에 거대언어모형을 적용할 경우 미래 트렌드와 혁신 기술을 예측하는 연구에 활용하는 것이 가능함을 실증하였다.

주요용어: 수소, 탄소중립, 토픽모델링, BERTopic

[1]51139 창원시 의창구 창원대학로20 51206호, 창원대학교 산업시스템공학과 기술경영연구실 학부연구생. (Authors were equally contributed). E-mail: rlaalsrbdhqk@gmail.com, inred00@naver.com
[2](교신저자) 창원대학교 산업시스템공학과 조교수. E-mail: jylee@changwon.ac.kr

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 235-235

고객 서비스 산업 특허에서의 패턴 도출 연구 : BERTopic을 활용하여

김채연[1], 이주용[2]

요 약

2000년대 이전까지 제조업으로 중심으로 한 물질적인 가치를 의미하던 서비스의 정의는 2000년대 이후 서비스업의 급진적인 발달을 바탕으로 고객 중심적인 가치로 변화하였다. 변화된 가치는 4차 산업혁명과 코로나19를 차례로 겪으면서 서비스 산업에 디지털 기술의 접목으로 확장되었고, 최근 디지털 전환 기술이 서비스 산업의 신 경영전략으로 대두되면서 기업들에게 디지털 전환은 필수적인 사항이 되어가고 있다. 본 연구는 2000년 1월부터 2023년 9월까지 등록된 고객 서비스 산업 및 디지털 전환과 관련된 국제특허 3029개의 특허를 대상으로 BERT기반 토픽 모델링(BERTopic)을 수행하였다. BERTopic이란 사전 학습된 언어 모델을 기반으로 한 토픽모델링 방법 중 하나로, 전체적인 맥락을 파악하여 토픽을 생성한다는 점에서 기존의 토픽 모델링 방법론과의 차별 점을 가지고 있다. 토픽모델링을 통해 고객 서비스 산업 기술을 10개의 주요 토픽으로 모델링하고 각 토픽들의 연도별 추세를 분석한 결과 2022년 기준 가장 높은 추이를 보이는 토픽은 사용자 중심의 네트워크 서비스 설계(User-centric network service design)이며, 증가 추세가 최근 5년간 가장 가파르게 증가하고 있는 것으로 나타난 토픽은 클라우드 컴퓨팅(Cloud computing)이다. 실제로 사용자 중심의 네트워크 서비스는 인터넷의 발달 이후 꾸준히 발전하고 있는 기술이며, 클라우드 컴퓨팅은 고객 서비스의 디지털 전환을 위해 2023년 가장 집중적으로 개발되고 있는 핵심 기술 중 하나임을 확인하였다. 따라서 본 연구는 특허 분석을 통해 고객 서비스 산업 특허들의 시계열적 추세를 파악하고, 토픽모델링을 통한 기술의 미래 동향 예측 수행의 효과성을 시사한다.

주요용어 : customer service, digital transformation, BERTopic, cloud computing

[1]51139 창원시 의창구 창원대학로20 51206호, 창원대학교 산업시스템공학과 기술경영연구실 학부연구생. E-mail: gap0428@naver.com
[2](교신저자) 창원대학교 산업시스템공학과 조교수. E-mail: jylee@changwon.ac.kr

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 237-237

Artificial Intelligence Techniques for Outcome Prediction in Marketing Strategies and Big Data Analytics for Businesses*

Minho Sun[1], *Seung Woo Kim*[1], *Jai Woo Lee*[2]

Abstract

Machine learning algorithms are innovatively transforming the field of business, attracting deepened interest from researchers. In this project, we review marketing research and develop machine learning methods useful in building marketing strategies. We provide an overview of machine learning methods, and compare them with statistical methods that marketing researchers have traditionally used. Machine learning methods can be used to process large-scale data, providing flexible analysis models and yielding solid predictive performance. We present an integrative conceptual framework to extract insights from large-scale tracking, and network data to represent descriptive, causal, and inferential analyses. Customer purchase journeys with decision-support capabilities can connect the machine learning methods to marketing theories and human insights. The specific applications of machine learning methods in many marketing segments and their contribution for marketing sectors have been validated. The proposed methods can be applied to analyze dynamic mechanism of marketing data with diverse customer features.

Key words: Machine Learning, Business Intelligence, Large-scale data, Marketing, Customer Data

---

*This article is financially supported by the 2023 College of Public Policy at Korea University.
[1]Department of Big Data Science, College of Public Policy, Korea University, Sejong, Republic of Korea.
[2](Corresponding Author) Department of Big Data Science, College of Public Policy, Korea University, Sejong, Republic of Korea; E-mail: jaiwoolee@korea.ac.kr

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 239-239

2차전지 기술 특허분석을 통한 토픽모델링: BERT 모델을 이용하여

신한준[1], 이주용[*]

요 약

2차전지 기술은 현대 생활의 다양한 측면에서 널리 사용되고 있으며 효율적인 에너지 저장의 중요성이 높아지고 있다. 이를 바탕으로 시장 규모가 점점 증가함에 따라 해당 분야에서의 지속적인 기술 개발을 위한 데이터 분석의 중요성이 대두되고 있다. 특허데이터는 일반적으로 기술적 동향을 파악하기 위한 분석 대상으로써 널리 활용되고 있으며 기술 혁신과 경쟁력을 위한 중요한 정보를 제공할 수 있다. BERTopic은 기존의 LDA, LSA와 같은 단어 출현 빈도에 기반한 전통적인 토픽모델링 기법에 비해서 텍스트의 의미론적 유사성을 고려하여 문맥에 맞는 단어와 문장 벡터 표현을 생성함에 있어 매우 우수한 자연어처리 결과를 보여주기에 해당 모델을 이용하여 토픽을 추출하였다. 본 연구에서는 국제특허에 등록된 2차전지 기술 관련 특허 데이터를 크롤링하여 결측치 제거 등의 전처리 과정을 거쳐 총 6218건의 기술 특허 데이터로부터 토픽을 추출하였다. 10개의 클러스터로 이루어진 핵심 토픽을 추출하였으며, 분석 대상 기간인 2013년~2022년 까지의 2차전지 기술의 주요 활용처는 전기차, 무인이동장치, 태양광 패널 등임을 도출하여 BERT를 이용한 토픽 모델링의 유용성을 확인하였다. 또한, 토픽들의 핵심 주제를 선정하고 이에 대한 주요 개념과 동향분석을 제시하여 향후 이차전지 연구에 대한 통찰력을 제공한다.

주요용어 : 2차전지, 토픽모델링, 특허분석, BERT, 태양광 패널.

[1]51139 창원시 의창구 창원대학로20 51206호, 창원대학교 산업시스템공학과 기술경영연구실 학부연구생. E-mail: shj8068@naver.com
[*](교신저자) 창원대학교 산업시스템공학과 조교수. E-mail: jylee@changwon.ac.kr

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 241-241

Identifying topics and future trends of CCUS technology: a BERT-based iterative topic modeling

Jungmin Ahn[1], *Byeongju Park*[1], *Juyong Lee*[2]

Abstract

CCUS refers to Carbon Capture, Utilisation, and Storage technology for achieving carbon neutrality. CCUS technology can be subdivided into various detailed technologies and fields according to the capture, utilisation, and storage methods. Therefore, it is important to choose specific areas to focus on for development and investment. This study leverages the Bertopic algorithm to scrutinise 10 years of patent data from 2013 to 2022 to predict future trends in CCUS technology. Bertopic categorises the 4,988 patents into 10 distinct topics, which are in turn organised into two distinct sub-technology groups to provide a deeper understanding of the technology trajectory. Further analysis is performed on yearly patent trends to facilitate future trend forecasting. The analysis shows that two topics have seen a distinct increase in development over the past three years: efficiency enhancement technology and BECCS (Bioenergy with CCS). Thus, this study performed an additional topic modeling only for the BECCS technology. The analysis finds a significant surge in blue hydrogen among the BECCS topics, empirically and ultimately confirming that the low-carbon hydrogen production will be one of the future trends among the sub-topics of CCUS technology. One of the most heavily invested and demonstrated CCUS technologies in 2023 is hydrogen production, which validates the predictions of this study. In this respect, this study goes beyond traditional patent analysis methodologies such as topic modelling and technology segmentation to demonstrate the applicability of patent analysis for predicting future trends and innovative technologies.

Keywords : CCUS, BECCS, Topic modeling, BERTopic, Hydrogen

[1]Undergraduate researcher, Department of Industrial & Systems Engineering, Changwon National University, 20-1 Changwondaehak-ro, Uichang-gu Seoul, 51139, Korea. (Authors were equally contributed). E-mail: jung101206@gmail.com, qkrqudwn0306@naver.com

[2](Corresponding Author) Assistant Professor, Department of Industrial & Systems Engineering, Changwon National University. E-mail: jylee@changwon.ac.kr

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 243-246

고수익 에어비앤비 분류 모델

이석빈[1], 김세희[2], 김헌주[3]

요 약

에어비앤비는 유휴 공간의 호스트(임대인)와 게스트(임차인)를 P2P 방식으로 연결해 주는 공유경제 플랫폼이다. 에어비앤비를 통해서 호스트는 기간에 덜 구애 받으며 임대 수익을 추구할 수 있고, 게스트는 가격 절감의 효과를 누리며 선택할 수 있는 숙소가 다양해진다. 이러한 에어비앤비의 장점은 관광객, 워킹 홀리데이 노동자가 많은 서호주 지역에도 유익할 것으로 예상되나 이에 대한 선행 연구가 많지 않고 가격도 높다. 서호주 지역의 에어비앤비 시장이 더욱 활성화되기 위해서는 새로 진입하는 호스트들의 에어비앤비 수익성이 보장되어야 한다. 이에 본 연구에서는 고수익, 저수익을 분류하는 모델을 만드는 것을 연구 목표로 한다. 이를 통해 신규 호스트들은 진입 전 자신의 에어비앤비 수익성을 예측하고 고수익 에어비앤비가 되는데 필요한 항목을 고려하여 진입 전략을 설계할 수 있다. 사용된 분류 모델로는 로지스틱 회귀모형, 판별분석, 랜덤 포레스트 알고리즘을 사용하였으며 성능을 비교한 결과 랜덤 포레스트 모델이 가장 성능이 좋은 모델이었다. 또한 고수익, 저수익을 분류하는 가장 중요한 요소는 예약 가능 여부와 관련된 'booking' 변수였다.

주요용어 : 에어비앤비, 서호주, 고수익, 분류기 알고리즘.

1. 서론

2022년에 서호주 관광객은 32,066명으로 코로나 이전 수준으로 복귀하였으나 충분한 숙박 시설이 특정 지역에만 몰려있기 때문에 관광객들이 어려움을 겪을 것으로 예상된다(Visitor statistics, 2023). 이에 호스트가 게스트에게 직접 숙박시설을 제공하는 에어비앤비가 대안이 될 수 있으나 공급이 활성화되기 위해서 숙소를 제공하는 '호스트' 들의 수익성이 보장되어야 한다. 본 연구에서는 퍼스 지역을 중심으로 월평균 이상의 수익을 내는 숙소를 '고수익 숙소'로 정의하고 로지스틱 회귀분석, 판별분석, 랜덤 포레스트 알고리즘을 이용하여 고수익, 저수익 숙소를 나눠주는 분류기 알고리즘을 개발하는 것을 연구 목적으로 한다(Kirasich, Smith, AdlerSadler, 2018). 이를 통해 모델이 다른 지역에도 적용되어 신규 호스트들이 자신의 에어비앤비 흥행 여부를 예측하고 고수익을 내는 중요한 요소를 고려하여 진입 전략을 세우기를 기대한다.

[1]대한민국 포항시 한동로 558, 한동대학교 경영경제학부 학부재학생. E-mail: 21800517@handong.ac.kr
[2]대한민국 포항시 한동로 558, 한동대학교 ICT창업학부 학부재학생. E-mail: kimsehee7414@handong.ac.kr
[3]대한민국 포항시 한동로 558, 한동대학교 창의융합교육원 교수. E-mail: heonjkim@handong.edu

2. 선행연구

다국 내 특수 관광 지역에 대한 에어비앤비 가격 예측은 헤도닉 가격 모형을 적용한 선행 연구 위주로 존재한다. Choi(2016)는 서울, Lee(2019)는 제주시 애월읍, Kim, Kang(2023)은 부산 소재 지역의 가격 예측 모형을 만들었다. 또한 모델에서 일반적으로 침대 수, 욕실 수, 최대 수용인원, 평점, 후기 등을 변수로 사용한 것과 달리 입지 특성을 고려하여 Lee(2019)는 주변 관광지 수, 해변 직선거리, 올레길 직선거리를 변수로 추가하였고 Kim, Kang(2023)은 해안가 더미, 남부 관광지 더미 변수를 추가하였다. 한편, 해외에서는 헤도닉 모델뿐만 아니라 다양한 모델을 적용한 에어비앤비 가격 예측 모델이 존재한다. Pooja et al.(2021)에서 22개국 나라에서의 에어비앤비 예약 예측 변수를 허들 기반 포아송 카운트 회귀 기법으로 추정했을 때, 슈퍼호스트 상태, 호스트 응답 시간 및 게스트와의 커뮤니케이션이 지역과 관계없이 가장 중요한 예측 변수였으며 '즉시 예약의 여부'가 비즈니스 방문객이 이용하는 숙소 예약에 중요했다. 호주 시드니, 멜버른 지역에서 에어비앤비 가격 결정 상위 5개 요인은 숙박 가능 인원, 호스트 응답시간, 객실 유형, 근린 시설 및 리뷰 점수였다. 에어비앤비 숙박 가격 예측과 관련된 연구는 많았으나 고수익과 저수익을 분류하는 연구는 미비하고, 시드니와 멜버른처럼 동호주가 아닌 서호주 지역에 중점을 둔 선행연구는 역시 존재하지 않았다.

3. 실증분석

3.1 데이터 셋

본 연구에서는 Inside Airbnb 사이트에서 22년 12월에 수집된 23년 12월, 1월, 2월 퍼스 성수기 숙소 예약 정보를 이용하였다. 종속변수 'Estimated_monthly_revenue'는 '샌프란시스코 모델'에 근거하여 만들고, 평균 이상(고수익)을 1, 평균 이하(저수익)를 0인 범주형으로 변환하였다 (Marqusee, 2014). 종속 변수를 만드는 데 이용한 'reviews per month', 'price', 'minimum nights', 'beds'는 독립변수로 사용하지 않았다.

3.2 선택

변수의 수가 많았기 때문에 '분류' 하기 앞서서 '선택'을 진행하였다(Kirkos, 2022). 범주형 변수의 경우 크래머 V 상관계수가 높은 것을 제외하는 방식을 이용했고 Figure 1과 같이 'room type', 'bathrooms'의 상관계수가 0.62로 높아서 'bathrooms' 변수를 제거하였다. 수치형 변수의 경우 Figure 2와 같이 PCA를 진행하여서 사람들의 review score와 상관이 있는 PC1, 호스트의 친절함과 관련있는 PC2를 도출하였다. PC3는 즉시 예약 가능한지 나타내는 'instant bookable', 호스트가 숙소를 얼마나 소유하고 있는지 나타내는 'host_listing_count와 같은 예약 관련된 변수와 관련이 있었으며, PC4는 호스트가 실제로 숙소를 소유했는지 암시하는 'availability_365' 변수와 강한 상관관계가 있었다. PC4까지 누적 분산비 0.71로 총 네 가지 성분을 도출하고 각각 'review', 'host', 'booking', 'owing'로 설정하였다. '선택'을 통해 도출된 범주형, 수치형 변수들은 Table 1과 같다.

'분류'의 알고리즘으로 로지스틱 회귀분석, 판별분석, 랜덤 포레스트 알고리즘을 사용하여 성능을 비교하기에 앞서 단계적 선택법을 진행한 결과 'available', 'room_type', 'host_is_super_host', 'host_response_time' 계수가 유의하지 않아서 제거하였다. 이후 세 가지 알고리즘을 사용하여 성능을 비교한 결과 Table 2와 같이 랜덤 포레스트 알고리즘이 정확도, 민감도, 특이도 모두 높았고 분류 성능이 가장 좋았다.

또한 랜덤 포레스트에서 'variable importance plot'을 그려본 결과 'booking' 변수가 고수익, 저수익을 분류하는 데 가장 중요한 변수였다. 'booking' 변수는 host_listing_count와 양의 상관관계에 있었고 'instant_bookable' 변수와 부의 관계가 있었다. 이를 통해 호스트가 소유하고 있는 숙소의 수에 사람들이 반응하여 예약하고, 예약이 치열할수록 고수익 숙소가 된다고 할 수 있다.

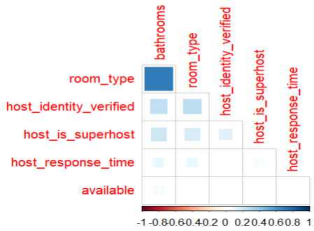


Figure 1. Cramer's V

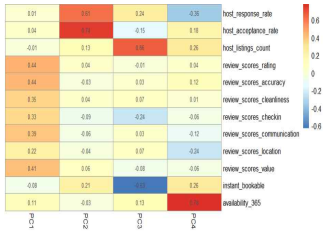


Figure 2. PCA

Table 1. Variables after selection

	variable	type
dependent	Estimated_monthly_revenue	factor
independent	host_response_time	factor
	host_is_superhost	factor
	host_identity_verified	factor
	room_type	factor
	available	factor
	review	numeric
	host	numeric
	owning	numeric
	booking	numeric

Table 2. Performance comparison

	accuracy	sensitivity	specifity
logistic regression	0.6569	0.7778	0.4419
discriminant analysis	0.6946	0.6975	0
random forrest	0.8757	0.8589	0.9270

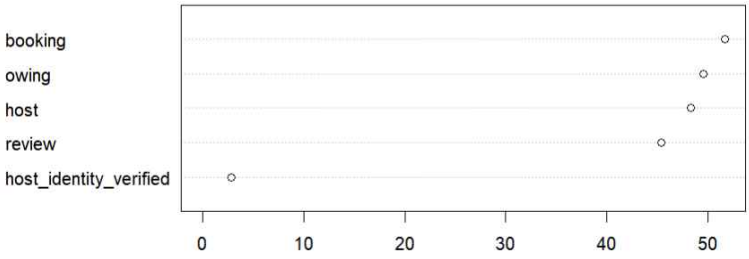


Figure 3.

4. 결론

고수익, 저수익을 분류하는 모델에서는 랜덤 포레스트 알고리즘이 가장 성능이 좋았으며 연구 결과 고수익 숙소가 되기 위해서는 호스트가 많은 숙소를 소유하여 소비자들의 신뢰를 얻고 예약이 치열해야 한다는 결과를 얻었다. 또한 기존 선행 연구와 다르게 에어비앤비 수익을 내는 데 중요했던 'host_is_superhost' 같은 변수들이 고수익을 내는 데 유의하지 않게 나타났으며 PCA 등의 방법론으로 변수의 차원을 축소 시키기도 하였다. 본 연구는 퍼스 지역 성수기로 숙소와 기간을 한정했기 때문에 실제로 상업화시키기 위해서는 각 지역의 입지 조건이나 기간 등 특성을 고려해야 하며 '지역 내에서 월평균 이상의 수익을 올리는 숙소'로 정의한 '고수익 숙소'에 대해 더욱 엄밀한 정의가 필요하며 트렌드에 따라 고수익이 되는 데 유의하게 나타난 변수들이 바뀔 수 있다. 이러한 한계는 후속 연구에 맡기기로 한다.

References

Choi, D. J. (2016). An Analysis of the Determinants of Accommodation Prices in Airbnb, KonKuk University.

Kim, M. J., Kang, S. M. (2023). An Analysis of Pricing Factors for Airbnb Accommodation: Focusing on Busan Metropolitan City, *Journal of Tourism and Leisure Research*, 35(187), 5-25.

Kirasich. K., Smit, T., Sadler, B. (2018). Random Forest vs Logistic Regression: Binary Classification for Heterogeneous Datasets, *SMU Data Science Review*, 1(3).

Krikos, E. (2022). Airbnb listings' performance: determinants and predictive models, *European Journal of Tourism Research*, 30.

Lee, J. Y. (2019). A Study on the Determinants of High Season Accommodation Price by the Characteristics of Airbnb, KonKuk University.

Marqusee, A. (2014). Airbnb and San Fransisco: Descriptive Statistics and Academic Research, San Fransisco Planning Department.

Sengupta, P., Biswas, B., Kumar, A., Shankar, R., Gupta, S. (2012). Examining the predictors of successful Airbnb bookings with Hurdle models: Evidence from Europe, Australia, USA and Asia-Pacific cities, *Journal of Business Research*, 137, 538-554.

Visitor Statistics. (2023).

https://www.tourism.wa.gov.au/Markets-and-research/Latest-tourism-statistics/Pages/Visitor-statistics.aspx#/2023/6/20

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 247-250

4차산업혁명 기술이 여성노동에 미치는 영향*

정예은[1], 홍지훈[2]

요 약

본 연구는 기업활동조사(2017~2019)년 자료를 사용하여 4차산업혁명 기술도입이 여성노동에 미치는 영향을 분석하였다. 균형패널자료로 구축하기 위해 3년간 모두 조사가 되지 않은 기업을 제외하였으며 성향점수매칭과 이중차분법을 사용하여 추정하였다. 본 논문은 기업이 이윤극대화를 위해 상대적 조정비용에 따라 고용을 조절할 것이며 여성이 남성보다 조정비용이 낮기 때문에 오히려 4차산업혁명으로 인한 기술혁신을 도입할 경우 여성고용의 유연성으로 인해 오히려 여성의 고용이 확대 될 것이라는 가설에서부터 출발한다. 상용직근로자를 기능직, 제조업, 비제조업으로 나눠서 분석한 결과 4차산업혁명 도입기업의 경우 비도입기업에 비해 여성의 고용자수가 남성보다 유의미하게 증가함을 확인하였다. 본 연구는 4차 산업혁명기술도입 여부가 고용에 미치는 영향을 여성을 중심으로 연구하였다는 의의가 있으며 이를 통해 상대적으로 불안정한 고용지위를 경험하고 있는 여성고용의 확대를 위한 새로운 경로를 제시하였다는 점에서 시사점을 갖는다.

주요용어 : 이중차분법, 4차산업혁명, 여성고용.

1. 서론

4차 산업혁명은 2016년 세계경제포럼에서 처음 제시되어 현재까지 기업의 빅데이터, 인공지능 등과 같은 기술의 도입을 발전시키고 사회 전반에까지 혁신적인 변화를 불러일으키고 있다. 이러한 4차 산업혁명으로 인한 기술발전은 기술이 노동을 대체함으로써 노동시장에 상이한 충격을 준다는 견해가 다수 존재한다. Blanchard and Katz(1997)는 저숙련 분야는 노동공급이 탄력적이기 때문에 기술혁신으로 인해 고용이 감소할 수 있다고 하였다. 이에 반해 Pissarides(2000)는 기술혁신으로 인해 오히려 생산성이 향상되고 이를 통해 고임금 근로자를 고용할 수 있게 되어 전체적으로 고용이 증가한다고 하였다. 본 연구는 앞선 선행연구에 더하여 여성노동을 중심으로 4차산업혁명을 통한 기술혁신이 고용에 미치는 영향을 확인하고자 한다. 여성은 노동시장에서 비정규직, 임시근로 등을 더 많이 종사함으로 남성보다 이미 불안정한 고용 지위를 경험하고 있다(전기택2020). 또한 코로나19 충격으로 인해 남성보다 여성이 부정적 고용충격을 받았다는 연구도 다수 존재하는데 Baek and Park(2022)은 여성의 경우 코로나19로 인한 양육 부담의 증가로 어린 아이가 있는 기혼 여성의 고용률이 감소하였음을 보고하고 있다. 본 논문은 기업이 이윤극대화를 위해 상대적 조정비용에 따라 고용을 조절할 것이며 여성이 남성보다 조정비용이 낮기 때문에 오히려 4차산업혁

[1]부산대학교 경제통상대학 경제학과 박사수료 정예은 hihi9529@naver.com
[2](교신) 부산대학교 경제통상대학 경제학과 교수 홍지훈 gh9x@pusan.ac.kr

명으로 인한 기술혁신을 도입할 경우 여성의 고용이 확대될 것이라는 가설에서부터 출발한다. 이선경,최창곤(2012)에서는 성별의 차이가 고용조정에 미치는 영향을 분석하였으며 그 결과 여성노동자들이 남자노동자들보다 조정비용이 낮아서 고용 유연성이 크게 나타난다는 결론을 도출하였다.

본 논문은 4차산업혁명 기술이 여성노동에 미치는 영향을 분석하기 위해서 통계청에서 제공하는 기업활동조사(Survey of Business Activities)(2015-2019) 를 사용한다. 기업활동조사는 기업의 전반적인 경영활동에 대한 연간 조사 상용근로자 50인 이상 자본금이 3억원 이상인 기업을 대상으로 하여 2020년 조사 기준 13,429개의 기업이 대상이다. 특히 기업활동조사는 2017년부터 4차 산업혁명 관련 기술 즉, 사물인터넷, 빅데이터, 인공지능, 블록체인, 3D 프린팅 등과 관련된 통계가 추가적으로 집계되었다. 이에 본 연구에서는 기업활동조사 2017~2019년 데이터를 3년간 모두 조사가 되지 않은 기업을 제외하여 균형패널자료로 구축하였다.

본 논문의 세부 구성은 다음과 같다. 2절에서는 본 연구를 위해 적용된 분석방법 및 연구모형을 설명하고 제 3절에서는 본 논문의 연구 결과를 기술한다.

2. 분석방법 및 연구모형

본 연구는 4차산업혁명 기술도입이 여성노동에 미치는 영향을 살펴보고자 4차산업활용기업과 비활용기업을 매칭하여 여성고용에 미치는 영향을 확인하고자 한다. 분석방법과 관련하여 본 연구는 성향점수매칭(PSM)방법과 이중차분법(DID)을 결합하여 분석한다.

먼저 성향점수매칭(PSM)을 활용하여 4차산업활용기업과 비활용기업을 매칭함으로써 두집단을 동질적으로 구성하였다. 그리고 이중차분법(DID)를 통해서 4차산업혁명 기술도입여부에 따라 여성고용에 미치는 차별적 영향을 추정하였다. 성향점수매칭 방법과 이중차이분석 방법을 연계하여 활용하는 이유는 자기선택(self-selection)에 의한 내생성(endogeneity) 문제를 효과적으로 대처하고자 함에 있으며 다수의 연구에서 사용하는 방법이다(오세환 2021)(이승민·신기윤·이정동 2022).

1) 성향점수 추정

본 연구에서 성향점수 추정을 위해 이항 로지스틱 모형을 사용하였으며 식(1)을 통해 도출된다. 개별 기업들의 특성을 대표하는 변수는 매출액, 자산, 상용 근로자수, 비용, 대기업 여부 등을 활용하였으며(Oh Se-Hwan, Baek Hyun-Mi and Lee Sae-Rom, 2016; Chae Ho-Chang, Koh Chang-E and Prybutok, 2014) 선행연구를 참고하여 적용하였다.

$$id4_{i,t} = \beta_0 + \beta_1 * \ln sales_{i,t} + \beta_2 * \ln asset_{i,t} + \beta_3 * \ln emp_{i,t} + \beta_4 * \ln cost_{i,t} + \beta_5 * size_{i,t} + \beta_6 * year_{i,t} + \Sigma\gamma_n * \in dustry_{i,j} \quad (1)$$

$(id4_{i,t}$: 4차산업혁명기술 활용여부 (활용 $=1$, 비활용 $=0)$, $\ln sales_{i,t}$: log(매출액),

$\ln asset_{i,t}$: log(자산규모), $\ln emp_{i,t}$: log(종업원수), $\ln cost_{i,t}$: log(비용),

$size_{i,t}$: 기업규모(대기업 $=1$, 비대기업 $=0)$, $year_{i,t}$: 연도$(2019=1)$,

$\in dustry_{i,j}$: 산업더미(한국표준산업분류))

2) 이중차분법(DID; Difference In Difference)

본 연구는 4차산업혁명기술 도입 여부에 따라 여성고용에 유의한 차이가 있는지를 분석하고자 이중차분법을 사용하였다. 분석모형은 식(2)와 같다.

$$Y_{i,t} = \beta_0 + \beta_1 * id4_{i,t} + \beta_2 * year_{i,t} + \beta_3 (id4_{i,t} + year_{i,t}) + \epsilon_{i,t} \quad (2)$$

($Y_{i,t}$: 전체여성상용직수,기능직상용직수,제조업상용직수,비제조업상용직수,
$id4_{i,t}$: 4차산업혁명기술활용여부(활용 = 1,비활용 = 0)
$year_{i,t}$: 연도더미 (2017 = 0, 2018,2019 = 1))

$id4_{i,t}$는 이중차분모형 내 처치군으로 기업 i가 4차산업혁명을 도입하였을 때 1을 값을 갖는 이항변수이고, $year_{i,t}$은 2017년도에는 4차산업혁명기술을 도입하지 않았지만 2018년도와 2019년도에 새롭게 4차산업혁명기술을 도입한 기업의 경우 1의 값을 갖는 이항변수이다. 기업활동조사가 2017년도부터 4차산업혁명과 관련된 데이터를 구축하였기 때문에 이와 같은 방법으로 연도더미를 구축하였다. 또한 이분산성을 제거하기 위해서 종속변수에 1을 더하여 사용하였다.

3. 분석결과

Table 1는 식(2)에 제시된 모형의 결과이다. 전체상용직을 기능직, 제조업, 비제조업으로 나누어 분석해본 결과 4차산업혁명 도입 기업인 경우 그렇지 않은 기업에 비해 시간이 지남에 따라 기능직상용직 여성종업원수가 9.1% 증가한다. 기능직남성상용직의 경우에는 여성보다 작은 비율인 6.5% 증가한다. 비제조업 상용직을 보더라도 여성이 경우 4차산업혁명 기술도입에 따라 종업원수가 4.5% 증가한다. 모두 1% 수준에서 유의하다.

이러한 결과는 여성의 고용이 남성의 고용보다 조정비용이 낮아 상대적으로 여성고용이 유연하기 때문에 4차산업혁명을 통한 기술혁신 기업의 경우 그렇지 않은 기업에 비해 여성의 고용이 확대될 수 있음을 추론할 수 있다.

Table 1. 4차산업혁명이 여성고용에 미치는 영향

	전체상용직		기능직상용직		제조업상용직		비제조업상용직	
	남성	여성	남성	여성	남성	여성	남성	여성
$id4_{i,t}$	-0.026 (0.031)	0.004 (0.026)	-0.047** (0.020)	-0.063*** (0.017)	-0.063 (0.049)	-0.027 (0.036)	0.005 (0.013)	-0.011 (0.013)
$year_{i,t}$	-0.012*** (0.010)	-0.053*** (0.008)	0.128*** (0.006)	0.099*** (0.005)	-0.355*** (0.016)	-0.216*** (0.012)	0.023*** (0.004)	0.026*** (0.004)
$id4_{i,t} * year_{i,t}$	0.007 (0.031)	-0.010 (0.026)	0.065** (0.021)	0.091*** (0.018)	-0.052 (0.049)	-0.056 (0.036)	0.003 (0.013)	0.045*** (0.014)
cons	2.297*** (0.008)	1.529*** (0.007)	3.187*** (0.005)	2.265*** (0.018)	1.510*** (0.012)	0.920*** (0.009)	4.103*** (0.003)	2.918*** (0.003)
R-squre	0.0035	0.0022	0.0029	0.0031	0.0138	0.0085	0.0022	0.0061
obs	49,114	48278	64,060	63,668	45,572	44,507	64,128	64,128

References

Blanchard, O. L. F. Katz, (1997), What We Know and Do Not Know about the Natural Rate of Unemployment, *Journal of Economic Perspective*, 11, 462-499.

Baek, J. P. (2022). COVID-19, Childcare and Women's Labor Supply. *Korean Economic Review,* 323-345.

Chae, H, C , V. R. (2014), Information Technology Capability and Firm Performance: Contradictory Findings and Their Possible Causes, *MIS Quarterly*, 38(1), 305-326

Pissarides, C. A., (2000), Equilibrium Unemployment Theory, *Cambridge, MA: MIT Press.*

Oh, S.H, H. M.B. (2016), Revisiting the Relationship between Information

전기택, & 배진경. (2020). 코로나 19 의 여성 노동위기 현황과 정책과제. KWDI Brief, (58), 1-9.

이선경, & 최창곤.(2012). 고용 및 임금의 조정비용이 노동시장의 유연성에 미치는 효과. 경제연구, 30(4) 143-159.

오세환. (2021). 4 차 산업혁명 기술 활용이 수출성과에 미친 영향. 무역학회지, 46(2), 323-335.

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 251-251

Analyzing South Korea's Household Finance via Panel Data

최인수[1], 정유진[2], 김도윤[2], 이준용[2], 정용수[2], 김우창[1]

요 약

최근 한국은 소득격차 확대, 저출생 및 고령화와 같은 사회경제적 문제로 인해 지속적인 성장 잠재력 저하의 위험에 직면하고 있다. 이러한 사회적 도전에 대응하기 위해 조세 및 재정정책의 역할이 점점 중요해지고 있지만, 현재까지 가구 단위에서의 조세, 지출, 복지와 관련된 포괄적인 데이터는 상대적으로 부족한 상태다. 이러한 데이터 부족 문제를 해결하기 위해 본 연구는 한국조세재정연구원에서 제공하는 NasTaB 데이터를 중심으로 한국의 국가 재정 통계의 핵심 측면을 분석하였다. 재정패널조사는 가계의 조세부담 및 복지혜택, 그리고 그에 따른 사회경제적 영향을 포괄적으로 파악하기 위한 중요한 도구로 활용되고 있다. 본 연구는 이러한 재정패널조사를 통해 얻어진 데이터를 기반으로, 최근 14차년도 조사에서 특히 주목해야 할 주요 통계적 특성과 경향을 세밀하게 분석하였다. 이를 통해 최근의 한국의 재정 및 경제 상황에 대한 보다 정확하고 깊이 있는 이해를 도모하고자 하였다.

주요용어: 재정패널조사, 금융 데이터 분석, 특성 분석, 가구 단위 분석, 패널 데이터.

[1]한국과학기술원 산업및시스템공학과
[2]경희대학교 산업경영공학과

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 253-257

확장된 기술수용모델을 적용한 UGC 관광정보 플랫폼의 지속적 사용의도에 관한 연구

판밍웨[1], 전재균[2]

요 약

구전 정보는 관광객의 관광상품 구매 결정에 큰 영향을 미친다. 현재 공유 소셜 미디어를 사용하여 관광지 마케팅을 수행하는 것은 매우 일반적이다. 따라서 본 연구는 확장된 기술수용모델(ETAM)을 적용하여 중국 UGC 관광정보 플랫폼인 샤오홍슈에 대한 사용자의 지속적 사용 의도에 영향을 주는 주요 요인들 간의 구조적 관계를 파악하고자 하였다. 분석결과 품질특성인 정보 품질, 시스템 품질과 서비스 품질은 모두 지각된 유용성과 용이성, 즐거움에 정(+)의 영향을 미쳤다. 또한, 지각된 유용성과 용이성, 즐거움은 지속적 사용의도에 모두 정(+)의 영향을 미치는 것으로 나타났다. 이런 연구결과에 다르면 향후 관광정보 플랫폼 기업의 충성고객 확보를 위한 마케팅 전략으로 활용하고자 한다.

주요용어 : 확장된 기술수용모델, 관광정보 플랫폼, 지속적 사용의도

1. 서론

최근 몇 년 동안 경제가 발전함에 따라 사람들의 물질적 생활 수준과 정신적 추구가 끊임없이 향상되었다. 그 중 관광산업은 3차 산업의 중요한 부분으로서 시대적 요구에 부응하는 고속 발전기를 맞이하였다. 이와 함께 인터넷과 스마트폰의 급속한 발전은 뉴미디어 건설을 추진할 수 있는 기반을 마련했고 관광 관련 산업의 융합 발전을 촉진했다. 사용자 생성 콘텐츠(UGC) 플랫폼은 사용자의 전통적인 인터넷 사용 패턴을 깨고 독창적인 공유 모델로 사용자의 열정을 동원하여 소비자의 개인화 요구를 더 잘 충족시킨다. 위의 장점을 바탕으로 UGC 관광플랫폼의 사용자 수는 지속적으로 확대되고 있으며, 그들의 사용자 분석은 수요자의 관점에서 플랫폼 구축의 품질과 대중의존도를 향상하게 시키기 위한 아이디어를 제공하는 데 도움이 될 것인다. 따라서, 본 연구는 중국 모바일플랫폼에서의 관광 정보 탐색 및 수용 과정을 연구하고자 중국의 UGC형 플랫폼인 샤오홍슈를 온라인 구전 관광정보 전달 플랫폼으로 취급하고, UGC형 앱의 품질특성이 기술수용요인을 거쳐 지속적 사용의도에 어떤 영향을 미치는지 확인하고자 한다. 샤오홍슈 플랫폼은 2022년 기준 월간 활성 사용자 수가 2억 6,000만 명으로 중국 내 UGC형 플랫폼 중 선두를 달리고 있다는 점(cbndata, 2022)에서 본 연구 목적에 가장 부합하는 플랫폼이라고 볼 수 있다. 이에 따라 본 연구의 연구목적은 기술수용모델(TAM)에 외부변수를 추가한 확장된 기술수용모델(ETAM)을 적용하여 UGC 관광정보플랫폼인 샤오홍슈의 지속적 사용의도에 영향을 주는 주요 요인들 간의 구조적 관계를 파악하고자 한다.

[1]48513 부산광역시 남구 용소로 45, 부경대학교 일반대학원 박사 수료. E-Mail : 1234567pmy@naver.com
[2]48513 부산광역시 남구 용소로 45, 부경대학교 경영학부 교수. E-Mail : jkjun@pknu.ac.kr

2. 이론적 배경

2.1. UGC(user-generated content)형 플랫폼

UGC는 인터넷 용어로 전체 이름은 User Generated Content 즉 사용자 생성 콘텐츠를 말한다. UGC의 개념은 인터넷 분야에서 시작되었으며, 사용자가 자신의 원본 콘텐츠를 인터넷 플랫폼을 통해 표시하거나 다른 사용자에게 제공하는 방식이다. UGC는 개인화를 주요 특징으로 내세우는 웹2.0 개념과 함께 부상했다. UGC란 일반인들이 자신의 생각, 사진, 동영상 등을 유튜브(YouTube), 블로그(Blog), 위키피디아(Wikipedia) 등과 같은 웹사이트의 특성에 맞게 변형, 창작하여 게시한 콘텐츠를 의미하다(Hermida, Thurman, 2008).

2.2. 확장된 기술수용모델

Davis(1989)에 의해 제안된 기술수용모델은 새로운 정보기술 개발에 따른 사용자의 수용 태도를 예측하고 행동 의도를 이해하기 위해 고안되었다. 기술수용모델의 핵심 변수로는 지각된 유용성과 지각된 사용 용이성을 들 수 있다. 여기에서 지각된 유용성은 새로운 정보기술을 사용하여 자신의 업무성과가 개선될 것이라는 믿는 정도를 의미하며, 지각된 용이성은 새로운 정보기술의 사용이 많은 노력을 요구하지 않을 것이라는 믿는 정도를 의미한다(Davis, 1989). 기술수용모델은 새로운 정보기술 수용과 관련된 연구에 폭넓게 적용되어 왔지만 기술수용에 영향을 미칠 수 있는 다른 변수들의 발굴을 제한했다는 점이 한계로 지적되기도 했다(Bagozzi, 2007). 따라서 초기 기술수용모델은 최근 들어 모델의 설명력을 높이기 위해 정보기술의 종류 및 연구주제에 따라 다양한 외부변수를 발굴하여 신념변수의 선행요인으로 추가한 확장된 기술수용모델이 주된 흐름이 되고 있다. 초기 기술수용 모델 관련된 연구는 주로 실용적인 측면만을 강조하였으나, 최근에는 쾌락적 측면에서의 즐거움을 중시하는 연구들이 증가하고 있다(Lee, 2019). 지각된 즐거움이란 정보기술 사용을 통해 예상되는 성과와는 관계없이 사용 자체가 재미있다고 느끼는 믿음의 정도를 의미한다(Davis, Bagzzi, Warshaw, 1992). 많은 선행 연구에서 지각된 유용성, 지각된 용이성, 지각된 즐거움은 모바일 앱을 포함한 새로운 정보기술의 지속적인 사용과 만족을 유도하는 결정적인 선행요인으로 제시되고 있다(Davis et al., 1992; Novak et al., 2003; Moon, Byun, 2020).

2.3. 수정된 정보시스템 성공모형

DeLone, McLean(1992)은 정보시스템 성공에 영향을 미치는 변수들을 시스템 품질, 정보 품질, 정보시스템 사용, 사용자 만족, 개인적 효과, 조직적 효과로 분류하고, 그들 사이의 인과관계를 설명하기 위한 정보시스템 성공 모델(IS success model)을 제시하였다. DeLone, McLean(2003)은 IT 부서의 서비스 품질(service quality) 영역을 추가한 수정된 정보시스템 성공모델을 제시하였다. 여기서 시스템 품질은 이용자가 평가한 시스템 품질을 의미하며 가용성, 적응성 등이 포함되며, 정보 품질은 시스템이 제공하는 정보의 품질을 나타내며(Kim, Kyung, 2009), 서비스 품질은 정보시스템 부서 및 해당 시스템에서 이용자가 제공받는 품질을 의미한다(Delone, McLean, 2003). 시스템 품질, 정보 품질, 서비스 품질이 정보시스템의 전반적인 성공을 측정하기 위한 중요한 변수들로 알려졌다(DeLone, McLean, 2003). 이에 본 연구에서는 선행연구를 바탕으로 정보시스템의 3가지 차원을 모두 적용하여 샤오홍슈의 기술수용요인에 대한 영향력을 검증하고자 한다.

2.4. 지속적 사용의도

지속적 사용의도는 사용자가 어떤 제품을 사용했거나 혹은 어떤 서비스를 경험해 본 후에도 그 제품을 계속 사용하거나 그 서비스로 변경하는 정도를 가리킨다. 지속적 사용의도는 고객의 실제 행동에 직접 영향을 주는 결정 요소라고 하였다(Garbarino, Johnson, 1999). 따라서 기업이 이익을 계속 확대하려면 소비자의 지속적 사용의도를 중시하고 정확하게 파악해야 한다(Bhattacherjee, 2001). 본 연구에서는 지속적 사용의도를 문설아(2019), Bhattacherjee(2001), 김민정, 이수범(2017)의 연구에 근거하여 샤오홍슈 사용 후, 사용자가 앞으로도 서비스를 중단 없이 사용하며 다른 사람에게 추천의식이 있는 정도로 정의하였다.

3. 연구 방법

3.1. 연구가설

Zhao et al.(2019)의 외식업체의 SNS 품질에 관한 연구에서 정보, 시스템, 서비스 품질은 지각된 용이성과 유용성 모두에 정(+) 영향을 미치는 결과를 얻었다. 장용석(2013)은 프로스포츠구단 웹사이트와 이용의도에 관한 연구에서 시스템 품질, 서비스 품질은 지각된 즐거움에 유의한 긍정적 영향을 미치는 것으로 나타났다. 따라서 선행연구를 바탕으로 가설 1과 가설 2, 가설 3을 설정하였다.

H1: 품질특성은 지각된 용이성에 정(+)의 영향을 미칠 것이다.
H2: 품질특성은 지각된 유용성에 정(+)의 영향을 미칠 것이다.
H3: 품질특성은 지각된 즐거움에 정(+)의 영향을 미칠 것이다.

또한 Gao(2020)는 틱톡의 지속적 사용의도에 관한 연구에서 지각된 용이성과 유용성 모두 지속적 사용의도에 유의한 영향을 미치는 것으로 나타났다. Heo(2016)는 동영상 앱의 지속적 사용의도에 대한 연구에서는 지각된 즐거움이 지속적 사용의도에 유의한 영향을 미치는 것으로 나타났다. 따라서 선행연구를 바탕으로 가설 4과 가설5, 가설 6을 설정하였다.

H4: 지각된 용이성은 지속적 사용의도에 정(+)의 영향을 미칠 것이다.
H5: 지각된 유용성은 지속적 사용의도에 정(+)의 영향을 미칠 것이다.
H6: 지각된 즐거움은 지속적 사용의도에 정(+)의 영향을 미칠 것이다.

3.2. 조사 설계

본 연구에서는 샤오홍슈 서비스를 사용한 경험이 있는 중국인 대상으로 조사연구를 실시하였다. 설문은 중국 인터넷 조사앱 문전성(問卷星)에 이용하여 2023년 9월 6일부터 2023년 9월 15일까지 진행하였다. 온라인 패널에게 총 460부의 설문지를 회수하였다. 그중에 사용하지 못한 응답은 제외하고 회수된 충 454부를 유효 표본으로 확정하고 분석에 사용하였다. 본 연구를 위해 수집된 자료는 SPSS 23.0과 AMOS 23.0 통계 프로그램을 이용하여 빈도분석, 신뢰도분석, 확인적 요인분석 및 구조방정식 모형분석을 실시하였다.

4. 실증 분석

4.1. 응답자의 특성

본 연구 대상인 샤오홍슈 서비스 사용자의 인구통계학적 특성을 이해하기 위하여 성별, 연령, 학력, 평균소득에 대한 빈도분석(frequency analysis)를 분석하였다. 표본에 대한 인구통계학적 특성을 보면, 먼저 성별은 여성이 287명(63.2%), 남성이 167명(36.8%)를 차지하고 있었다. 연령은 20~29세가 215명(47.4%)으로 가장 높은 비율을 차지하였고, 30세~39세 196명(43.2%), 20세 이하 21명(4.6%), 40~49세가 17명(3.7%), 50세 이상이 5명(1.1%)으로 나타났다. 월평균소득은 3,001위안~6,000위안이 158명(34.8%), 3,000위안 이하 이 135명(29.7%), 6,001위안~10,000위안이 87명(19.2%), 10,001위안~15,000위안이 40명(8.8%), 15,000위안 이상이 34명(7.5%) 순으로 나타났다.

Table 1. 가설 검증 결과

가설	경로		비표준화계수	표준화 오차	표준화 계수	t-값	가설채택
H1-1	정보 품질 →	지각된	.218	.061	.200	3.593**	채택
H1-2	시스템 품질 →		.159	.055	.163	2.881**	채택
H1-3	서비스 품질 →	유용성	.306	.060	.283	5.066**	채택
H2-1	정보 품질 →	지각된	.368	.063	.318	5.802**	채택
H2-2	시스템 품질 →		.149	.056	.144	2.648**	채택
H2-3	서비스 품질 →	용이성	.309	.061	.270	5.029**	채택
H3-1	정보 품질 →	지각된	.304	.064	.270	4.740**	채택
H3-2	시스템 품질 →		.134	.058	.132	2.318*	채택
H3-3	서비스 품질 →	즐거움	.253	.062	.227	4.055**	채택
H4	지각된 용성 →	지속적	.184	.039	.236	4.680**	채택
H5	지각된 이성 →		.107	.037	.146	2.914**	채택
H6	지각된 거움 →	사용의도	.283	.040	.375	7.103**	채택

Note: *p<.05, **p<.01

4.2. 신뢰성 및 타당성 분석

내적 일관성 검증을 위하여 Cronbach's α 값을 통하여 신뢰성을 평가한 결과, 모든 변수의 값은 .8 이상으로 적합한 것으로 나타났다. 확인적 요인분석을 실시한 결과, 모형의 적합도 지수는 CMIN=638.395, df=539, p=.002, CMIN/df=1.184(기준: ≤3), GFI=.929, NFI=.940, TLI=.989, CFI=.990, RMSEA=.020를 보여, 모두 기준치에 충족하는 것으로 나타나(Hair et al., 2006), 연구모형은 적합하다고 판단된다. 요인적재값은 756-.886으로 기준치(≥.5)에 적합하고, AVE(Average VarianceExtracted)는 .610 .721 기준치(≥.5)를 충족했으며, CR(Construct Reliability)은 .887~..920 기준치(≥.7)를 충족하여 집중타당성은 확보되었다(Hair et al., 2006). 또한 AVE의 제곱근 값은 .781-.849이고, 요인 간 상관계수 값은 .299-.455로 확인되어 모든 상관계수의 값이 AVE의 제곱근 값에 비하여 작기 때문에 판별타당성 확보되었다고 할 수 있다.

4.3. 가설검증

본 연구에서는 설정한 연구 모형을 검증하기 위하여 구조방정식모델분석을 실시하였다. 먼저, 적합도 검정결과 보면, CMIN=697.121(df=545, p<.01), CMIN/df=1.279, CFI=.985, GFI=.923, TLI=.984, RMSEA=.025로 충분히 수용가능한 적합도 수준인 것으로 나타났다. 또한 본 연구에서 설정한 가설 검정 결과 Table 1에 나타난 것처럼 모든 가설이 채택되었다.

5. 결론

본 연구의 결과는 다음과 같은 시사점을 갖는다. 첫째, 본 연구는 UGC형 관광 정보 플랫폼 품질이 지속적 사용의도에 미치는 영향 관계를 파악함으로써 UGC 관광 정보 플랫폼의 품질 연구에서 후기수용모델의 적용할 가능성에 대한 이론적 기반을 제공하였다. 둘째, 초기 정보시스템모형에서는 전자상거래의 확대로 점차 서비스의 중요성이 부각되면서 서비스 품질을 추가하여 샤오홍슈의 품질특성을 도출함에 따라 기존 선행 연구의 이론적 간격을 보완하였다. 셋째, 지각된 즐거움을 추가함에 따라 기술수용요인의 이론적 확장을 검증하였고 지각된 즐거움은 기술수용요인의 중요한 확장변수임을 확인하였다. 넷째, 정보 품질과 시스템 품질이 지각된 용이성과 지각된 유용성에 유의한 영향력이 있는 변수로 밝혀졌으므로 신뢰 있는 관광 정보를 제공뿐 아니라 편리한 플랫폼도 제공해야 한다는 사실을 알 수 있다. 따라서 기업은 샤오홍슈에서 제공하는 여행 정보를 지속해서 개선하고 감독하여 소비자에게 안전한 서비스를 제공해야 한다. 다섯째, 서비스 품질이 지각된 용이성과 지각된 유용성에 영향을 주는 것으로 밝혀졌으므로 현재 치열한 경쟁 환경 속에서 차별화된 경쟁력을 갖추기 위해서는 새로운 고객가치를 제공해야 한다. 여섯째, 지각된 즐거움은 중요한 핵심 변수인 것으로 밝혀졌다. 이용자 대다수는 20대로 특히 플랫폼의 오락성에 관심이 많기 때문에 기업은 플랫폼의 상호작용과 재미를 높이고 일부 특색 있는 관광 정보를 추가해야 한다. 하지만 위의 시사점에도 불구하고 본 연구는 다음과 같은 한계점이 있다. 본 연구에서는 샤오홍슈 서비스를 사용한 적이 있는 소비자를 대상으로 이루어졌는데 향후 연구에서는 사용자와 비사용자의 비교분석에 관한 연구가 필요할 것으로 사료된다.

참고문헌

Davis, F. D., Bagozzi, R. P., & Warshaw, P. R.(1989). User acceptance of computer technology models. *Journal of Management Science*, 35(8), 982－1003.

Delone, W. H., McLean, E. R.(2003), The DeLone and McLean model of information systems success: A ten-year update, *Journal of Management Information,* 19(4), 9-30.

DeLone, W. H., & McLean, E. R.(2002). Information systems success revisited. In Proceedings of the 35th Annual Hawaii International Conference on System Sciences, *IEEE*, 2966-2976.

Moon, S. A.(2019). Research on the continuous use intention of O2O service in the food industry applying extended technology acceptance model: Focused on the moderating effect of the innovation resistance and the innovativeness. Yeungnam University, Doctoral Thesis.

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 259-259

Educational Python for Big Data Analytics*

Juyong Ko[1], *Donggeun Kim*[1], *Jai Woo Lee*[2]

Abstract

Big data analytics made outstanding achievements in the fields of business, science, public policy, etc. While experts are familiar with the theory and software, it has been challenging for non-experts to use software programs which require skills in programming, mathematical background, and statistical knowledge. In this project, we present a big data analytics tool, an educational Python resource designed for a public willing to learn fundamental ideas in performance-oriented software and utilize data analytics tools in real-world problems. Big data analytics contains an educational application programming interface, educational procedures from data cleaning to machine learning techniques for individuals who learn or develop from scratch. The big data analytics tools paved the way for analyzing simulations on high-dimensional data with step-by-step instructions. Big data analytics tool is an open-source tool which manages and assesses the investigations of high-dimensional data.

Keywords: Big Data Analytics, Educational Python, Machine Learning, Statistics, High-Dimensional Data

---

*This article is financially supported by the 2023 College of Public Policy at Korea University.
[1]Department of Big Data Science, College of Public Policy, Korea University, Sejong, Republic of Korea
[2](Corresponding Author) Department of Big Data Science, College of Public Policy, Korea University, Sejong, Republic of Korea; E-mail: jaiwoolee@korea.ac.kr

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 261-261

청소년의 사회정서역량 유형 분류 및 영향 변인 탐색

백예은[1], 정혜원[2]

요 약

OECD(Organization for Economic Co-operation and Development)에서는 Education and Social Progress 연구를 추진하면서 사회정서역량(social and emotional skills) 발달의 중요성을 직접적으로 언급하며(OECD, 2015). 사회정서역량이란 사회적, 정서적 역량과 관련된 다양한 하위 역량을 포괄하는 역량으로, 자신과 타인의 정서를 이해하는 능력과 더불어 타인과 협력하여 사회적 관계나 문제 해결을 성공적으로 수행하는 역량을 의미한다(김소영, 김현지, 이상수, 2018). 사회정서역량은 다양한 하위 영역으로 구성된 복합적인 특성을 갖는데, 학생 개인마다 하위 역량별로 발달 수준이 상이하게 나타날 수 있다. 이에 본 연구는 중, 고등학생의 사회정서역량 하위 역량(창의성, 그릿, 삶의 만족도, 사회적 위축, 협동심)의 발달 유형에 따른 잠재 프로파일 집단을 분류하고, 집단 분류에 영향을 미치는 변인을 탐색하고자 수행되었다. 이를 위해 중학교 2학년과 고등학교 2학년에 해당되는 한국아동·청소년패널조사 2018의 5차년도 초4 코호트와 중1 코호트 자료를 활용하였다. 이때, 잠재프로파일 분석(latent profile analysis)을 통해 잠재 집단을 분류한 뒤 머신러닝 기법인 XGBoost(extreme Gradient Boosting)를 적용하여 잠재 집단 다중 분류에 높은 기여를 하는 변인을 도출하고 해당 변인을 공변인으로 투입하여 다항 로지스틱 회귀분석을 실시하였다. 주요 결과는 다음과 같다. 첫째, 중, 고등학교급 각각 잠재 집단이 4개, 5개로 나타났는데 하위 역량 간의 균형 있는 발달 양상을 보였으나 잠재 집단 간 수준 차이가 나타난 집단과 더불어 사회적 위축, 협동심 등이 현저하게 낮게 나타나는 등 하위 역량 간 발달 수준이 불균형적인 발달 양상을 보인 집단 또한 도출되었다. 둘째, 잠재 집단 분류에 영향을 미치는 변인을 탐색한 결과, 중, 고등학생 공통적으로 학업에 대한 열의가 높고 교사 및 친구와 긍정적인 관계를 형성한 학생일수록 사회정서역량 발달 수준이 높은 집단과 협동심이 높은 집단에 속할 확률이 높은 것으로 나타났다. 셋째, 중학생의 경우 부모로부터 조언, 구조를 제공받거나 진로 계획 시 경험하는 어려움과 관련된 변인이 집단 분류에 유의한 영향을 미치는 것으로 나타난 반면 고등학생의 경우 부모로부터 자율성을 지지받거나 신체 건강과 관련된 변인이 집단 분류에 유의한 영향을 미치는 것으로 나타났다.

주요용어 : 사회정서역량, 중·고등학생, 잠재프로파일분석, XGBoost

[1]34134 대전 유성구 대학로 99, 충남대학교 일반대학원 교육학과 교육평가 전공 박사과정.
E-mail : byeunn7@gmail.com
[2](교신저자) 34134 대전 유성구 대학로 99, 충남대학교 일반대학원 교육학과 교육평가 전공 교수.
E-mail : chw7@cnu.ac.kr

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 263-263

농산물 라이브커머스 참여농가의 교육 프로그램 개선을 위한 시사점

이정명[1], 이원석[2], 김혜형[3], 이영순[4]

요 약

사회·경제적 환경의 변화에 따라 농산물 소비에서도 구매처가 바뀌는 양상을 보이고 있다. 그 중 라이브커머스는 소비자들이 실시간 방송을 시청하며 상품을 주문하는 전자상거래 방식으로 코로나19 이후에 급속도로 성장한 온라인 판로처이다. 한국농촌경제연구원 이슈보고서에 따르면 라이브커머스는 농식품 마케팅 채널로서 매출 증대, 고객과의 양방향 소통을 통한 밀접한 관계 형성 등 폭넓은 기회를 제공할 가능성이 크다고 하였다. 본 연구는 기존의 농산물 라이브커머스 교육 프로그램의 개선안을 도출하기 위해 교육 수료자를 대상으로 실태조사를 수행하였다. 교육 사업 시행지역인 경기도 화성시, 평택시, 시흥시, 안성시 내 2023년 '농업인 라이브커머스 판매자 육성 사업' 참여자 40명을 조사 대상으로 하였으며, 조사방법은 구조화된 설문지를 이용하여 온라인 설문조사 하였다. 주요 조사내용으로는 농산물 라이브커머스 활용 정도, 교육 요소별 만족도·중요도, 향후 심화교육 수강의향 등을 조사하였다. IPA(Importance-Performance Analysis)를 사용하여 교육 프로그램 개선을 위한 주요 속성을 분석하였다. IPA 격자에 Ⅰ~Ⅳ분면을 구분하기 위하여 전체 중요도와 만족도의 평균값인 4.48, 4.23을 중심축으로 사용하였다. Ⅰ사분면(강점항목)에 속한 속성은 강사의 전문성, 1인 방송 판매자 육성 교육, 교육내용 및 구성, 방송을 위한 시설 및 공간이며, Ⅲ사분면(낮은 우선순위)에 속한 속성은 수업시간의 적절성으로 나타났다. Ⅳ사분면(과잉노력제거)에 속한 속성은 수업장소의 적절성, 입문반 교육 평가로 양성반 선발 후 진행되는 교육방식 등이었다. 농산물 라이브커머스 교육 수강 시 중요한 속성임과 동시에 상대적으로 높은 만족을 느끼는 속성으로서 향후 개선안 도출에 반영되어야 할 연구 결과는 다음과 같다. 첫째, 강사의 전문성이다. 사업 담당자는 교육 진행 업체 선정시 이론과 실습교육에 적합한 우수 강사진 구성을 평가할 필요가 있다. 둘째, 1인 방송 판매자 육성을 목표로 하는 교육이어야 한다. 농업인 스스로가 라이브커머스 판로를 활용하여 농산물을 판매할 수 있는 역량 강화 교육 위주로 이루어져야 할 것이다. 셋째, 교육내용 및 구성이 교육 사업 진행에 가장 중심이 되는 부분이기 때문에, 수준별 개인 역량이 강화될 수 있는 교육자료 제공, 적절한 비중의 이론과 실습 교육 등이 이루어져야 할 것이다. 마지막으로, 실습 교육을 진행하는 각 시군의 농업기술센터에서는 실습을 위한 장비(카메라, 조명 등)가 필요할 것으로 나타났다. 본 결과는 수준별 농산물 라이브커머스 교육 개선안을 도출하는데 참고자료로 활용될 수 있을 것이다.

주요용어 : 농산물 라이브커머스, 유통판로, 농촌지도사업

---

[1]18388 경기도 화성시 병점중앙로 283-33, 경기도농업기술원 농업연구사. E-mail: jmlee@gg.go.kr
[2]18388 경기도 화성시 병점중앙로 283-33, 경기도농업기술원 농업연구관. E-mail: born815@gg.go.kr
[3]18388 경기도 화성시 병점중앙로 283-33, 경기도농업기술원 농업연구사. E-mail: hyeong89@gg.go.kr
[4]18388 경기도 화성시 병점중앙로 283-33, 경기도농업기술원 농업연구관. E-mail: rosesea@gg.go.kr

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 265-265

YOLO를 활용한 3차원 물류 이미지 객체 탐지

김승현[1], 성유민[1], 이성윤[1], 조하늘[1], 김동하[2]

요 약

현재 물류 분야에서는 전자 상거래와 물류 시장의 빠른 성장으로 인해 효율적이고 정확한 상품 객체 탐지가 핵심적인 과제로 주목받고 있다. 기존의 객체 탐지 모델은 정형화된 데이터에 의존하여 물류 이미지의 다양한 각도와 형태에 대한 정확한 예측에 한계가 있다. 이에 따라 3차원 물류 이미지의 다양한 각도와 형태를 고려한 객체 탐지 모델의 개발이 필요하다. 본 연구는 YOLO 객체 탐지 모델을 효과적으로 학습시켜 정확한 예측을 끌어내기 위해 3차원 이미지의 단면을 추출하고 데이터 손실을 최소화하는 방법을 제시한다. 또한, 학습 데이터를 증강하고 원근감을 고려한 쌓임 형태의 평면 이미지로 전환하는 주요 기법을 도입한다. 최종적으로, 초기 학습 데이터와 새로 생성한 데이터셋으로 학습한 모델 간의 성능을 비교하여 물류 형태를 현실적으로 고려한 데이터 기반의 쌓임 형태 설계가 예측 성능을 향상하게 시켰음을 확인한다. 이러한 접근은 단순한 일반화를 넘어 현장의 실제 조건을 반영하여 효율적이고 유연한 물류 예측을 가능케 하며, 향상된 YOLO 모델을 통해 물류 시스템의 효율적인 자동화를 기대한다.

[1]02844 서울 성북구 보문로34다길 2, 성신여자대학교 자연과학대학 수리통계데이터사이언스학부 학부재학생.
[2]02844 서울 성북구 보문로34다길 2, 성신여자대학교 자연과학대학 수리통계데이터사이언스학부 조교수.
E-mail: dongha0718@sungshin.ac.kr

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 267-267

Development of an Machine Learning Model for Advanced Query Response in Bioinformatics for Microbiome Research

Da Som Park[1], Hyo Ri Shin[1], Na Yeun Kim[1], Do Youn Lee[1], Gi Moon Nam[1], Tae Gyun Kim[2], Hyung Take Cho[2], Kang Wook Lee[1] ,Ji Youn Hong[1,3] Jae Kyeom Kim[1,4]

Abstract

Background: The rapidly expanding field of bioinformatics increasingly relies on the effective analysis and interpretation of data. This project aims to develop an machine learning model capable of responding to complex, professional queries from bioinformatics researchers, utilizing a dataset of microbiome-related research papers.

Methods: The project encompasses several critical phases: data review, data cleaning, text analysis, and model development, followed by testing and evaluation. In the data review phase, an in-depth analysis of microbiome-related papers is conducted to understand the data structure, basic statistics, data quality, relational analysis, and text content. The data cleaning process involves addressing missing values, outliers, and duplicates, and standardizing formats to prepare for model training.

During the text analysis phase, key steps include keyword extraction and frequency analysis, topic modeling, and analysis of related concepts and associations within the text.

Results: The development of the AI model incorporates advanced Natural Language Processing (NLP) techniques to process the complex queries typical of bioinformatics researchers. The model is trained on a dataset of microbiome research papers, with a focus on accuracy and depth of analysis for performance evaluation and optimization. The model's performance is rigorously tested and assessed using the F1 score metric, which considers both precision and recall, providing a balanced measure of the model's classification accuracy.

Conclusion: This project demonstrates the feasibility of developing a specialized AI tool for the bioinformatics research community. This tool can respond with high accuracy to intricate research queries, showcasing the potential of advanced text analysis and machine learning techniques in paving the way for more sophisticated, data-driven tools in scientific research.

[*]Department of Food Biotechnology, Korea University, Sejong 30019, Republic of Korea
[1]The Bioinformatix, Gwangmyeong 14348, Republic of Korea
[2]Department of Food Regulatory Science, Korea University, Sejong 30019, Republic of Korea
[3]Department of Behavioral Health and Nutrition, University of Delaware, Newark, DE 19707, USA

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 269-269

Predicting Hypoxia and Estimating the Interactions of Ewe Metabolites Using Machine Learning Techniques*

Sangjin Kim[1], *Chan Gyu Jeon*[1], *Jai Woo Lee*[2]

Abstract

Hypoxia, possibly leading to increased oxidative stress, is a consequence of metabolic alterations and can influence fetal growth as well as lifelong health. While various methods have been utilized to identify associations of ewe metabolite concentration values with an outcome, hypoxia, studies assessing how the coexistence of ewe metabolites impacts the onset of hypoxia are at a nascent stage. Here, we present a statistical approach to identify the associations of ewe metabolites and classify the outcome by including the interactions of ewe metabolites in the network. After using metabolite sub-networks as clusters, we implement lasso for logistic regression to determine the onset of hypoxia. We tested different clustering methods in order to validate the associations of ewe metabolites which may infer the mechanism of significant metabolic processes. We validated, in terms of accuracy and balanced F-score, the performance of classification method to determine whether the onset of hypoxemia is correctly identified. Our study shows that there exist strongly interacting ewe metabolites and that specific metabolites are strongly associated with hypoxia. The proposed approach can be applied to similarly structured metabolite datasets to predict health outcomes.

Keyword s: Machine Learning, Big Data analytics, Statistics, Metabolomics, Hypoxia

---

*This article is financially supported by the 2023 College of Public Policy at Korea University.
[1]Department of Big Data Science, College of Public Policy, Korea University, Sejong, Republic of Korea.
[2](Corresponding Author) Department of Big Data Science, College of Public Policy, Korea University, Sejong, Republic of Korea; E-mail: jaiwoolee@korea.ac.kr

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 271-271

성인 여성의 사용 담배 유형과 우울의 관계 : 제8기 2차년도 국민건강영양조사 자료를 바탕으로*

김상희[1]

요 약

본 연구는 성인 여성의 사용 담배 종류에 따른 흡연유형이 우울과 관련 변인인 수면시간, 스트레스에 미치는 영향을 확인하기 위하여 제8기 2차년도(2020) 국민건강영양조사 자료를 사용한 이차자료 분석연구이다. 대상자는 조사에 참여한 만 19세에서 64세 이하의 성인 여성 2398명 중에서 결측치가 있는 대상자를 제외한 2,190명의 자료를 최종 분석하였다. 수집된 자료의 분석은 SPSS 26.0 프로그램을 이용하여 층화, 군집 및 가중치를 적용한 복합표본설계로 복합표본 빈도분석, 복합표본 교차분석, 복합표본 로지스틱 회귀분석을 시행하였다. 본 연구결과, 성인 여성의 사용 담배 유형에 따른 우울증상 발생확률과 평소 스트레스 인지 정도에 차이가 있었다. 비흡연자에 비해 일반담배 단일흡연자는 우울증상 발생확률이 2.598배 증가하였고 중복흡연자는 5.447배 증가하였다. 비흡연자에 비해 일반담배 단일흡연자는 스트레스를 많이 느낄 확률이 1.650배 증가하였고 중복흡연자는 2.675배 증가하였다. 본 연구는 성인 여성 중복흡연자가 비흡연자와 일반담배 단일흡연자보다 우울증상 발생 및 스트레스를 많이 느낄 확률이 높음을 발견했다. 본 연구는 성인 여성을 대상으로 사용 담배 유형과 우울 및 정신건강의 요인 분석을 통해 사용 담배 유형에 따른 우울 및 스트레스의 차이를 제시하였다는 데 그 의의가 있다.

주요용어 : 흡연, 우울, 스트레스, 수면시간, 중복흡연.

*이 논문은 2022년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 연구되었음 (NRF-2022M3J6A1084843).
[1]34134 대한민국 대전광역시 유성구 대학로99, 충남대학교 문헌정보학과 석사과정. E-mail: shk0594@naver.com

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 273-273

재가장기요양서비스 이용 노인의 삶의 의미 영향요인*

김은지[1], 안정아[2]

요 약

국내 65세 이상 노인인구는 전체 인구의 18.4%로, 향후 계속 증가하여 2025년에는 초고령 사회로 진입될 것으로 전망된다. 국내에서 2008년부터 시행되고 있는 노인장기요양보험 제도는 노인의 건강증진 및 생활안정을 도모하고 그 가족의 부담을 덜어줌으로써 궁극적으로 국민의 삶의 질을 향상시킴을 목적으로 하고 있다. 노인장기요양보험은 재가급여서비스와 시설급여서비스로 나뉘며, 최근 사회적 요구에 따라 재가급여서비스가 더욱 확대될 전망이다. 본 연구에서는 재가급여서비스 확대 전략에 따라 중요성이 강조되고 있는 “재가장기요양서비스” 이용 노인을 대상으로 하여, 이들의 삶의 의미의 수준을 확인하고 이에 대한 영향요인을 파악하고자 하였다. 본 연구의 자료수집은 재가장기요양서비스를 제공하는 경기도 소재 7개 주간보호센터를 이용중인 노인 121명을 대상으로, 2023년 9월 29일부터 10월 15일까지, 일상생활 수행능력, 우울, 사회적 지지와 삶의 의미를 포함한 구조화된 자가보고식 설문지를 이용하여 수집하였다. 수집된 연구자료는 SPSS 29.0 프로그램을 이용하여, 기술통계, independent t-test, one-way ANOVA, Pearson’s correlation coefficient, hierarchical multiple regression을 이용하여 분석하였다. 본 연구 결과, 재가장기요양서비스 이용 노인의 삶의 의미는 70점 만점에 평균 42.95점으로, 중등도 수준으로 나타났다. 일상생활 수행능력은 21점 만점에 평균 8.21점, 우울은 15점 만점에 평균 5.54점, 사회적 지지는 60점 만점에 평균 42.95점이었다. 대상자의 삶의 의미는 일상생활 수행능력(r=-.20, p=.026), 우울(r=-.31, p<.001)과 유의한 음의 상관관계가 있는 것으로 나타났고, 사회적 지지(r=.51, p<.001)와는 유의한 양의 상관관계가 있는 것으로 나타났다. 위계적 다중회귀분석 결과, 인구학적 특성 중 단변량 분석에서 유의성을 나타낸 변수들을 투입하였을 때(Model 1) 설명력은 7%였고, 주요 변수인 일상생활 수행능력, 우울, 사회적 지지 변수를 추가로 투입하였을 때(Model 2) 설명력은 26.7% 증가된 33.7%로 나타났다(F=15.25, p<.001). 대상자의 삶의 의미에 대한 유의한 개별 영향요인은 사회적 지지(β=.45, p<.001), 우울(β=-.16, p=.048) 순으로 나타났다. 본 연구 결과를 토대로, 재가장기요양서비스 이용 노인의 삶의 의미를 향상시키기 위해서는, 노인에 대한 사회적 지지를 높이고, 우울을 낮출 수 있는 전략의 적용이 필요할 것이다. 특히 국가적 차원의 재가장기요양서비스 확대 시행이 예상되는 바, 보다 적극적인 센터재정 지원 및 사회적 관심과 함께, 다양하고 전문적인 사회적 지지 제공 체계의 고려, 노인 우울의 스크리닝 및 이를 중재할 수 있는 맞춤형 프로그램 및 전문기관 연계 체계 적용 등이 고려되어야 할 것이다. 이를 통해 궁극적으로 노인의 삶의 의미 향상을 기대할 수 있을 것으로 사료된다.

주요용어: 재가장기요양서비스, 노인, 삶의 의미, 사회적 지지, 우울

---

*이 논문은 제 1저자 김은지의 석사학위논문의 축약본임.

[1]16499 경기도 수원시 영통구 월드컵로 164 아주대학교병원 간호사, 아주대학교 간호대학 석사과정생.
E-mail: kej3501@ajou.ac.kr

[2](교신저자) 16499 경기도 수원시 영통구 월드컵로 164 아주대학교 간호대학 • 간호과학연구소 부교수.
E-mail: ahnj@ajou.ac.kr

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 275-275

심부전 환자의 질병양상 변화와 완화 의료에 관한 웹기반 가족중심 의사소통 향상 프로그램 개발*

안정아[1], 김경화[2]

요 약

고령화 사회에서 심부전 환자의 유병률은 증가되고 있으며, 심부전은 최근 치료법과 약물 및 의료 기술의 발달에도 완치되기 어려운 만성 질환으로 여겨진다. 따라서 평생에 걸쳐 심부전의 진행 및 증상의 악화와 완화를 반복하는 상태로 살아가게 되는 심부전 환자의 삶에 있어, 환자를 비롯해 가족들은 높은 스트레스 및 부담감과 함께, 나아가 전반적 삶의 질 감소에 영향을 받으며, 이는 다시 환자의 심부전 질환관리에 악영향을 미친다. 최근 심부전 관리 지침에 따르면, 질환 관리의 목표는 환자의 최적의 증상 조절과 삶의 질 향상에 초점을 두며, 의료진은 질병양상의 변화에 따른 치료방향 설정에 있어 환자 및 가족과 함께 적극적으로 논의하고 의사결정하도록 하며, 증상 악화 및 나아가 생애 말기 관련 준비를 돕는 등 전 치료 과정에서 심부전 질병양상 변화에 대해 환자와 가족을 포함해 효과적 의사소통을 시행할 것을 권고하고 있다. 본 연구에서는 국내외 심부전 환자 의사소통 전략 가이드라인들을 토대로 문헌고찰을 통해 의사소통 교육의 이론적 근거와 콘텐츠의 근거를 체계적으로 파악하여, 『심부전 환자의 질병양상 변화와 완화 의료에 관한 웹기반 가족중심 의사소통 프로그램』을 개발하고자 하였다. 본 프로그램의 개발을 위해, 구성 가이드라인을 중심으로 의사소통 교육 콘텐츠와 의사소통 전략 중재 프롬프트 리스트 초안을 작성하였다. 작성된 초안은 전문가(심장내과 전문의 1인, 간호사 3인, 간호대학 교수 1인)의 타당도 검증을 거쳤으며, 검증 후 피드백을 반영하여 최종 중재 프로그램의 내용으로 완성하였다. 프로그램 내용은 소프트웨어 콘텐츠 제작 전문가와 본 연구자의 지속적 협업을 통해 2개의 웹기반 교육 형태(환자 및 보호자용, 의료진용)로 개발 완성하였다. 이는 심부전 환자 및 보호자와 의료진의 참여 편의성과 반복가능성 및 중재 확산을 기대하기 위함이며, 각 군에 특화된 의사소통 중재 및 상호간 의사소통의 모의 경험과 강화가 가능한 기능(VR 콘텐츠)을 포함하여 상호작용이 가능하도록 웹기반 중재 프로그램으로써 개발되었다. 본 연구를 통해 개발된 프로그램은 향후 고령화 사회 질병 궤적에 따른 평생 지속 관리와 이를 위한 효율적 의사소통이 필수적으로 요구되는 심부전 질환 관리를 위해, 환자, 가족 및 의료진을 위한 표준화된 의사소통 가이드 및 웹기반 교육 중재 방안으로써 활용 가능할 것으로 기대된다. 또한 이의 활용을 통한 효과적 의사소통은 심부전 환자의 심리•사회적 지표 향상과 더불어 만성 질환에 대한 통제 자신감과 나아가 가족 상호작용을 증진시킬 수 있을 것으로 사료된다.

주요용어: 심부전, 질병궤적, 완화 의료, 의사소통

---

*이 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2022R1F1A1075049).

[1](교신저자) 16499 경기도 수원시 영통구 월드컵로 164 아주대학교 간호대학•간호과학연구소 부교수. E-mail: ahnj@ajou.ac.kr

[2]16499 경기도 수원시 영통구 월드컵로 164 아주대학교 간호대학 석사수료생. E-mail: summerlove@ajou.ac.kr

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 277-277

간호 대학생의 성공적인 전환 준비를 위한 대학 교육 요구도 조사*

김지혜[1], 이경미[2], 김지영[3]

요 약

간호사의 부족은 세계적인 사회적 문제이며, 신입 간호사의 높은 이직률은 인력 부족의 주요 원인이다. 신입 간호사의 성공적인 전환을 위한 준비는 대학 교육에서부터 필요하다. 본 연구는 간호 대학생의 대학 교육 경험을 바탕으로 간호사로의 성공적인 전환 준비를 위한 교육 요구도를 파악하기 위한 현상학적 연구이다. 자료 수집은 2022년 12월에 시행되었다. 2개의 간호대학에 재학중인 4학년 간호 대학생 14명을 대상으로 2개의 포커스 그룹을 구성하여 반구조적, 개방형 질문지를 이용한 인터뷰를 하였고, 자료는 Colaizzi이 제시한 방법에 따라 분석하였다. 분석 결과, 간호 대학생들은 성공적인 전환 준비를 위해 교육과정, 학습자 역량, 교수자 역량 차원에서 다양한 교육 요구도가 확인되었다. 간호 대학생들은 교육과정 측면에서 '자신, 간호에 대한 인식 교육', '타 학문과의 통합을 위한 교육'을 원하는 것으로 탐색되었다. 학습자 역량 측면에서는 '환자 간호를 위한 역량', '멘탈력', '사회초년생으로서 능력'을 향상시킬 수 있는 교육을 원하는 것으로 파악되었다. 마지막으로 교수자 역량 차원에서는 '선배 간호사로서의 역할'을 원하고, '수업 설계 및 운영 능력'을 갖춘 교수자를 원하는 것으로 나타났다. 본 연구 결과를 바탕으로 대학생에서 간호사로 성공적인 전환을 위해 대학은 교육과정을 설계할 때 교과목, 교육 내용, 방법에 대한 대학생들의 요구를 반영해야 할 것이다.

주요용어 : 간호 대학생, 전환, 질적 연구, 현상학적 연구, 교육 요구도.

---

*이 연구는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. RS-2022-00165947).

[1]55338 전라북도 완주군 삼례읍 삼례로 443. 우석대학교 간호대학 조교수. E-mail: jihye2020@woosuk.ac.kr
[2]31065 충청남도 천안시 동남구 백석대학로 1. 백석대학교 간호학과 조교수. E-mail: linkmi@hanmail.net
[3]22212 인천광역시 미추홀구 인하로 100. 인하대학교 간호학과 조교수. E-mail: jy1223kim@inha.ac.kr

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 279-279

간호대학 졸업생의 간호실무준비도 관련 요인*

김지혜[1], 이경미[2]

요 약

간호대학은 학생들에게 실무에서 필요한 전문 역량의 함양을 목표로 교육하고 있음에도 불구하고, 신입 간호사는 학생에서 간호사로 역할을 전환되는 과정에서 많은 어려움을 호소하고 있다. 간호대학 졸업생의 간호실무준비는 간호사의 성공적인 역할 전환을 위해서 중요하므로 이들의 간호실무준비도를 높이는 전략이 필요하다. 본 연구의 목적은 한국의 간호대학 졸업생의 간호실무준비도를 파악하고, 관련 요인을 규명하는 것이다. 본 연구는 단면적, 서술적 조사연구 설계이다. 연구 대상자는 간호사 면허를 취득한 후 실무 현장에서 근무를 시작하지 않은 간호대학 졸업생이며, 편의표집을 통해 모집하였다. 전국에 위치한 10개 간호대학 졸업생을 대상으로 2023년 4월부터 5월까지, 온라인 설문지를 통해 자료를 수집하였다. 수집된 자료는 SPSS 22.0 프로그램을 이용하여 independent t-test, one-way ANOVA, Pearson's correlation coefficient, hierarchial regression 분석을 하였다. 연구 대상자는 총 178명이며, 여학생이 88.2%이고, 평균 연령은 23.85±1.68이었다. 간호대학 졸업생의 간호실무준비도는 중간 수준이었다. 성별, 연령, 학교지역, 성적에 따른 간호실무준비의 차이는 없었다. 반면, 전공만족도, 수업 만족도, 시뮬레이션실습 만족도, 전반적인 대학교육 만족도에 따라 간호실무준비의 유의한 차이가 있었다. 최종 회귀모델에서 사회적 지능, 임상실습 교육환경, 시뮬레이션실습, 대학교육 만족도는 간호실무준비도의 관련 요인으로 나타났다. 간호대학 졸업생의 간호실무준비를 향상시키기 위해서는 개인의 긍정적인 특성 강화뿐만 아니라 교육적 환경 조성을 위한 노력이 필요하다. 즉, 간호실무준비도를 높이기 위해 사회적 지능을 높이고, 질 높은 임상실습 교육환경 조성과 시뮬레이션 교육 설계가 요구된다.

주요용어 : 간호 대학생, 간호 교육 연구, 사회적 지능

---

*이 연구는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. RS-2022-00165947).

[1]55338 전라북도 완주군 삼례읍 삼례로 443. 우석대학교 간호대학 조교수. E-mail: jihye2020@woosuk.ac.kr
[2]31065 충청남도 천안시 동남구 백석대학로 1. 백석대학교 간호학과 조교수. E-mail: linkmi@hanmail.net

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 281-281

고학년 간호 대학생의 간호실무준비도에 관한 서술적 연구*

김지혜[1], 이경미[2]

요 약

간호 대학생들은 국민의 건강과 안녕을 보호하는 임무를 맡게 될 미래의 중요한 보건의료인이다. 신입 간호사의 조기 사직이 사회적인 문제로 두각되고 있는 현시점에서 고학년 간호 대학생의 실무준비도 확인을 통해 현 한국의 간호대학교육 성과를 평가하고, 추가 교육지원이 필요한 부분을 점검해보는 것이 필요하다. 본 연구는 고학년 간호 대학생의 간호실무준비 정도를 확인하기 위한 횡단적, 서술적 연구이다. 2개의 간호대학에 재학중인 4학년 간호 대학생을 대상으로 2023년 11월에 온라인 설문지를 이용하여 자료 수집을 하였다. 수집된 자료는 SPSS 28.0을 이용하여 서술적 통계, independent t-test, analysis of variance (ANOVA)로 분석하였다. 총 157명(여성 77.7%, 평균 연령 23.12±1.93세)의 자료를 분석한 결과, 전체 간호실무준비 점수는 3.17±0.41점(1~4점 범위)이고, 하위요인별로 보면, 임상문제해결 3.30±0.53, 학습기술 3.34±0.62, 전문직 정체성 3.37±0.49, 역경 3.36±0.69으로 나타났다. 일반적 특성에 따른 전체 간호실무준비도의 차이를 확인한 결과, 성별(t=-0.283, p=.778) 연령(F=0.401, p=.670)에 따른 차이는 없었다. 반면, 4학년 1학기 성적(F=8.162, p<.001), 전공 만족도(F=11.522, p<.001), 교내실습 만족도(F=7.689, p<.001), 시뮬레이션실습 만족도(F=10.756, p<.001), 임상실습만족도(F=4.446, p=.005), 강의 만족도(F=4.422, p=.014)가 높을수록 실무준비도가 높았다. 이와 같은 결과로 볼 때, 대학에서는 교육의 질을 향상시키는 노력으로 간호실무준비도를 향상시킬 수 있을 것이다. 특히, 실무에 대한 준비를 단순히 업무 수행에 대한 역량 뿐 아니라 사회적 지능, 조직적 통찰력과 같은 다차원적 요인이 반영됨을 고려할 때, 대학 교육에서 대인관계역량, 의사소통역량 등을 배양할 수 있는 교육이 필요하다고 하겠다.

주요용어 : 간호 대학생, 실무 준비, 전환, 간호 교육 연구

---

*이 연구는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. RS-2022-00165947).

[1]55338 전라북도 완주군 삼례읍 삼례로 443. 우석대학교 간호대학 조교수. E-mail: jihye2020@woosuk.ac.kr
[2]31065 충청남도 천안시 동남구 백석대학로 1. 백석대학교 간호학과 조교수. E-mail: linkmi@hanmail.net

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 283-283

노인의 신체기능과 삶 만족

박정혜[1]

요 약

본 연구는 증가하고 있는 노인 인구의 삶의 만족을 증가시키기 위하여, 노년기 신체기능 수준과 영양상태가 삶의 만족에 미치는 영향을 확인하기 위한 이차자료 분석연구이다. 본 연구를 위하여 보건복지부와 한국보건사회연구원에서 시행한 2020년 노인실태조사 참여자 10,097명 중 대상자 선정 기준에 적합한 9,279명의 자료를 분석하였다. 대상자 선정 기준은 만 65세 노인으로 설문조사에 다른 사람의 도움 없이 직접 응답하고, 주요 변수에 무응답이나 모르겠다는 응답이 없는 대상자이다. 분석 결과, 신체기능 수준과 영양상태 측면에서 노년기 삶의 만족을 유의하게 증가시키는 요인은 좋은 영양상태(β=-0.14, p<0.001)와 정상체중 유지(β=-0.04, p<0.001), 스스로 일상생활을 수행하는데 필요한 독립적 일상생활수행능력(β=-0.11, p<0.001)과 어려움 없는 기본신체활동(β=-0.06, p<0.001), TV를 보거나 책을 읽는 등 일상생활에 불편함이 없는 시력(β=-0.06, p<0.001), 필요한 영양을 섭취하고 먹고 싶은 것을 먹을 수 있는 불편함 없는 씹기 능력(β=-0.04, p=0.002)이었다. 이러한 결과는 길어진 노년기 만족스러운 삶을 위하여 사회적으로 필요한 복지혜택과 정책에 시사점을 제공한다.

주요용어 : 노인, 신체기능, 삶 만족.

Table 1. Factors associated with life satisfaction

Variables			β	t	SE	p
Physical function	Vision (ref.=Not discomfort)	Discomfort	-0.06	-4.92	0.02	<0.001
	Hearing (ref.=Not discomfort)	Discomfort	0.02	1.23	0.02	0.219
	Chewing (ref.=Not discomfort)	Discomfort	-0.04	-3.17	0.02	0.002
	Muscle strength (ref.=good)	Not good	-0.01	-0.39	0.02	0.695
	Basic physical movement (ref.=not difficult)	Difficult	-0.06	-6.01	0.02	<0.001
	ADLs (ref.=independent)	Dependent	-0.11	-9.00	0.04	<0.001
	IADLs (ref.=independent)	Dependent	-0.02	-1.78	0.03	0.075
Nutritional status	Body mass index (ref.=normal)	Abnormal	0-.04	-4.05	0.01	<0.001
	Nutritional screening initiative (ref.=good)	Risk	-0.14	-12.65	0.02	<0.001
R^2 / Adjusted R^2 / F(p)			0.190 / 0.189 / 114.52(<0.001)			

Note: ref.=reference; Adjusted vaiables: sex, age, education level, living arrangement, house ownership, household income, chronic disease, taking medicine

[1]52725 경남 진주시 동진로 33, 경상국립대학교 간호학과 부교수. E-mail: masternur@gnu.ac.kr

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 285-285

지역사회 1인 가구 노인의 영양 위험

박정혜[1], 강세원[2]

요 약

본 연구는 지역사회에 거주하는 1인 가구 노인의 영양 위험 관련 요인으로 우울과 신체기능을 파악하기 위한 이차자료 분석연구이다. 본 연구를 위하여 보건복지부와 한국보건사회연구원에서 시행한 2020년 노인실태조사 참여자 10,097명 중 만 65세 이상으로 다른 사람의 도움 없이 설문조사에 직접 응답한 1인 가구 노인 2896명을 대상으로 하였다. 자료 분석을 위하여 IBM SPSS/WIN 23.0을 이용하여, 기술통계, independent t-test, 이분형 로지스틱 회귀분석을 수행하였다. 그 결과, 연구대상자의 44.8%가 영양 위험에 있었고, 영양 위험은 우울한 대상자가 2.01배 (p<0.001), 씹기에 제한이 있는 대상자가 1.76배 (p<0.001), 기본신체활동 제한이 있는 대상자가 1.35배 (p=0.009), dependent instumental activities of daily living (IADLs) 대상자가 2.59배 (p<0.001) 영양 위험이 있는 것으로 나타났다. 그러므로 1인 가구 노인의 영양 위험에 대한 screening은 필수이며, 영양 위험이 우울과 의존적 IADLs와 강하게 관련되어 있으므로 이에 대한 문제가 없는지 확인할 필요가 있다.

주요용어 : 노인, 여가, 우울 위험

Table 1. Binary logistic regression of nutritional risk according to depression and physical function

Variables		Nutritional risk			
		B	p	OR	95% CI
Depression (ref.=not risk)	Risk	0.70	<0.001	2.01	1.699-2.389
Physical function					
Eyesight (ref.=good)	Limitation	0.14	0.181	1.15	0.937-1.410
Hearing (ref.=good)	Limitation	-0.11	0.337	0.90	0.716-1.121
Chewing (ref.=good)	Limitation	0.56	<0.001	1.76	1.435-2.146
Repeated chair stands (ref.=good)	Limitation	-0.05	0.623	0.95	0.775-1.165
Basic physical activity (ref.=good)	Limitation	0.30	0.009	1.35	1.078-1.701
ADLs (ref.=independent)	Dependent	-0.16	0.549	0.85	0.499-1.447
IADLs (ref.=independent)	Dependent	0.95	<0.001	2.59	1.847-3.630

*χ^2(p)=460.46(<0.001), Nagelkerke R^2=0.197, Hosmer & Lemeshow χ^2(p)=9.24(0.323)
*ADLs=activities of daily living, IADLs=instrumental activities of daily living
*Adjusted variables: age, educational level, house ownership, annual income, chronic disease, taking medicine, BMI

[1]52725 경남 진주시 동진로 33, 경상국립대학교 간호학과 부교수. E-mail: masternur@gnu.ac.kr
[2]47011 부산시 사상구 주례로 47, 동서대학교 간호학과 부교수. E-mail: swkang75@hotmail.com

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 287-287

노인의 사회교류와 삶 만족

강세원[1], 박정혜[2]

요 약

본 연구의 목적은 1인 가구와 부부가구 노인의 연령별 사회적 교류가 삶의 만족에 미치는 영향을 확인하는 것이다. 연구 대상은 2020년 노인실태조사 참여자 중 다음 기준에 따라 선정한 8188명이다; 1) 자녀나 다른 사람이 응답하지 않고, 본인이 스스로 설문조사에 응답한 대상자, 2) 혼자 살거나 배우자만 함께 거주하는 1~2인 가구 노인이다. 자료 분석을 위해 IBM SPSS/WIN 23.0을 이용하였으며, 연령별 사회적 교류가 삶의 만족에 미치는 영향을 확인하기 위하여 65~74세와 75세 이상 연령군으로 구분하여 분석하였다. 연령군별 빈도와 백분율, 평균과 표준편차, 교차분석, 독립표본 t-검증, Kruskal-Wallis H 검증, 가중치를 적용한 위계적 다중 회귀분석을 사용하여 분석하였다. 그 결과, 65~74세 연령군 노인의 삶의 만족에 가장 큰 영향을 미친 사회적 교류 요인은 친구와의 관계에 대한 만족이었고(β=0.41, p<0.001), 그다음이 자녀와의 관계에 대한 만족이었다(β=0.21, p<0.001). 이러한 결과는 75세 이상 연령군도 동일하였다(β=0.36, p<0.001; β=0.20, p<0.001). 그러나 사회활동과 사회관계 빈도는 두 연령군에서 차이를 보였다. 초기 노년기 노인은 봉사와 같은 사회활동과 활발한 사회관계 빈도가 삶의 만족에 유의한 영향을 미쳤으나, 중기 노년기 이후에는 자녀와 가끔 만나는 것만 유의한 영향을 미쳤다. 이것은 노화의 진행이 사회적 교류에 영향을 미칠 수 있으며, 중기 노년기 이후에는 사회활동과 사회관계의 빈도보다 가깝고 친밀한 사람과 좋은 관계를 유지하고 함께 시간을 보내는 것이 삶의 만족을 증가시키는데 중요함을 보여준다. 따라서, 증가하는 노인 인구의 길어진 노년기 동안, 삶의 만족을 증가시키기 위해 노년의 사회적 교류에 대한 올바로 인식과 준비가 필요하다.

주요용어 : 노인, 사회적 교류, 삶 만족

[1]47011 부산시 사상구 주례로 47, 동서대학교 간호학과 부교수. E-mail: swkang75@hotmail.com
[2]52725 경남 진주시 동진로 33, 경상국립대학교 간호학과 부교수. E-mail: masternur@gnu.ac.kr

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 289-289

A novel model reflecting the realistic distribution of disease spread*

엄은진[1], 최보승[2]

요 약

감염병 확산의 기본적인 구획 모형인 SIR 모형에서 잠복기가 추가된 SEIR 모형을 기반으로 하여 확률적 감염병 확산 모형을 구축할 수 있다. 기존 SEIR 모형은 각 잠복기와 감염기간에 대해 지수분포를 따른다고 가정한다. 그러나 실제 감염병의 잠복기와 감염기간은 더 복잡한 분포를 따를 수 있다. 따라서 지수분포의 가정은 모형의 추정에 편향을 가져올 수 있다. 본 연구에서는 감염병의 잠복기와 감염기간에 보다 현실적인 분포를 가정하기 위해 시간 지연을 반영한 모형을 개발하였다. 구체적으로 두 기간이 감마분포를 따른다고 보았고 이를 지연 시간의 분포로 사용하였다. 모형의 모수 추정을 위해 베이지안 MCMC 알고리즘을 사용하였다. 제안된 방법은 2020년 서울시에서 발생한 코로나19 확진자 자료를 사용하여 모형 적합을 수행했다. 결과적으로 제안된 모형은 기존의 모형 구축 및 추정보다 정확한 추정 결과를 보였다. 다양한 가정에서의 모의실험을 통하여 감마분포를 사용하는 것이 기존의 지수 분포를 가정하는 것보다 정확한 추정 결과를 보였다. 본 연구에서 제시하는 방법론은 실제 상황에 부합하는 모형을 구축하여 향후 감염병 대응 정책 수립에 활용될 수 있을 것으로 예상된다.

주요용어 : 감염병 확산 모형, 베이지안 모형, 비마코비안, 수학적 모델링

---

* 이 논문은 2023년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (RS-2023-00245056).
* 이 논문은 제1저자 엄은진의 박사학위논문의 축약본입니다.
[1](교신저자)30019 세종특별자치시 세종로 2511, 고려대학교 세종캠퍼스 경제통계학과 박사과정.
E-mail: eej0417@korea.ac.kr
[2]30019 세종특별자치시 세종로 2511, 고려대학교 세종캠퍼스 빅데이터사이언스학부 교수.
E-mail: cbskust@korea.ac.kr

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 291-294

라이프로그 데이터를 기반으로 여러 가지 치매 위험도 예측 모델들의 성능 비교 연구*

강소라[1], 조완현[2], 나명환[3]

요 약

본 연구는 광주광역시 서구 건강관리센터에서 수집한 라이프로그 데이터를 기반으로 여러 가지 기계학습 알고리즘들을 사용하여 고령층의 치매 위험도를 예측하고 치매 정도를 완화할 수 있는 바람직한 건강 관리를 방안을 제시하는 것을 목적으로 한다. 먼저, 라이프로그 데이터의 여러 가지 항목 중에서 집중 데이터, H-베타 파워값, 상대좌뇌 파워값, 심장건강, 교감신경 활성도, 부교감 신경 활성도, 자율신경 활성도의 총 7개 항목의 응답 값들을 기반으로 치매 정도를 결정하는 반응변수의 수준별 라벨값을 생성하였다. 그리고 생성된 반응변수의 라벨값들과 라이프로그 데이터를 구성하고 있는 나머지 항목들에 대해서 상관분석을 통해 치매 위험도에 영향을 많이 미치는 중요한 요인들을 도출하였다. 마지막으로 추출된 중요한 요인들을 독립변수로 하고 생성된 반응변수의 라벨값을 종속변수로 하여 세 가지 기계학습 알고리즘을 이용하여 치매 위험도를 예측하였다. 분석 결과로부터 로지스틱 회귀모형이 의학 자료 분석에서 가장 중요하게 생각하는 측도인 AUC (Area under the ROC curve)와 recall이 가장 높게 나타났음을 알 수 있다. 랜덤포레스트 모형과 엑스지부스트 모형은 특이도(specificity) 값이 크게 나타나 정확도(accuracy)가 높은 편이었다. 끝으로 우리는 주어진 결과들을 이용하여 치매 위험도를 예측하는 앱 기반 시스템을 개발하고, 치매 정도를 완화할 수 있는 바람직한 건강 관리를 방안을 제시하는 데 도움이 될 것으로 기대된다.

주요용어 : 치매 위험도 예측모형, 라이프로그 데이터, 기계학습 알고리즘, 예측 정밀도.

1. 서론

우리나라는 평균 수명이 연장됨에 따라 인구 고령화가 급격히 진행되고 있으며, 국제적으로나 국내적으로 치매 발병률도 증가 중이다. 치매환자 수 증가는 치매 관리비용으로 인한 사회·경제적 부담이 증가하고, 치매 환자와 가족의 삶의 질 저하 등의 문제로 이어질 수 있다. World Alzheimer Report에 따르면 전 세계 치매 인구는 2015년에 약 4,680만명이나, 2050년에는 약 3배 증가하여 1억 3,150만명에 이를 것으로 예상된다고 발표하고 있다. 또한, 국내 치매노인 인구는 2012년 약 54

*이 논문은 2023년도 K-Health 국민의료 AI서비스 및 산업생태계 구축사업으로 지원을 받아 수행된 연구임 (H0603-23-1001).
[1](제1저자) 61186, 광주광역시 북구 용봉동 용봉로 77, 전남대학교 수학/통계학과 박사과정.
Email : sc12love@gmail.com
[2]61186 광주광역시 북구 용봉로 77, 전남대학교 통계학과 교수. Email : whcho@chonnam.ac.kr
[3](교신저자) 61186 광주광역시 북구 용봉로 77, 전남대학교 통계학과 교수. Email : nmh@chonnam.ac.kr

만 명에서 2030년에는 약 127만 명, 2050년에는 약 271만 명으로 매 20년마다 약 2배씩 증가할 것으로 추산된다. 따라서 치매 환자가 지속해서 증가할 것으로 예상하는 현 상황에서 치매 환자와 가족들에게 향상된 지원을 제공하기 위한 구체적 방안을 마련하는 것이 필요할 것으로 사료된다.

본 연구에서는 일상생활과 밀접한 연관성을 가지는 라이프로그 빅데이터를 기반으로 사전에 치매 정도를 예측할 수 있는 모형을 개발하고자 한다. 이를 위해서 우리는 라이프로그 데이터로부터 치매 정도를 나타낼 수 있는 반응변수를 생성하였고, 이와 상관성인 높은 설명변수들을 도출하였다. 그리고 이러한 변수들을 토대로 세 가지 기계학습모형들을 구축하고 예측성능을 비교할 수 있는 실험을 시행하였다. 끝으로 유도된 결과들을 사용하여 고령층 노인들의 치매 정도를 알아볼 수 있는 AI 모델을 개발하였다.

2. 라이프로그 데이터의 분석

본 연구에서는 광주광역시 서구 라이프로그 건강관리센터에서 체성분/체수분분석기, 뇌파 및 맥파 측정기 등 다양한 장비들을 활용하여 측정한 459명에 대한 자료들을 제공받아 분석에 사용하였다. 첫째, 체성분/체수분분석기를 통하여 수집한 특성값들은 사람들의 체중, 세포 내외 수분, 단백질, 무기질, 몸통 근육량, 오른팔/왼팔 체수분, 골무기질량, 엉덩이 바깥둘레 등의 체형, 체성분들이다. 둘째, 뇌파 및 맥파 측정기를 통하여 수집한 특성값들은 우뇌/좌뇌 쎄타, 베타, 감마, 델타, 자율신경활성(TP), 부교감신경활성(HF), 교감신경활성(LF), 자율신경균형, 심박변이도, 분당심박수, 스트레스 크기값, 파워 등의 자료이다.

다음으로 각 검사자의 치매 정도를 나타내는 라벨변수를 생성하기 위해서 옴니핏(omnifit) 정보의 집중데이터, H-베타 파워값, 상대좌뇌 파워값, 심장건강(HRV-Index), 교감신경 활성도, 부교감신경 활성도, 자율신경 활성지수들의 일곱 개의 측정값들을 이용하였다. 이때, 이들 각각 특성값들이 정상 임계 구간을 벗어나면 1로 정상 구간 내에 있으면 0의 값을 부여하고 최종적으로 이들 여섯 개의 부여된 값들의 총합으로 새로운 치매 위험군 변수를 만들고 이들값의 분포를 파악하였다. 다음 Figure 1은 459명의 검사자에 대해 새롭게 생성한 치매 위험군변수가 갖는 특성값들의 분포를 보여주고 있다.

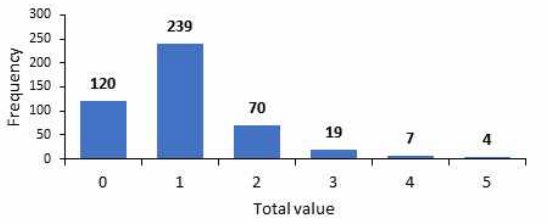
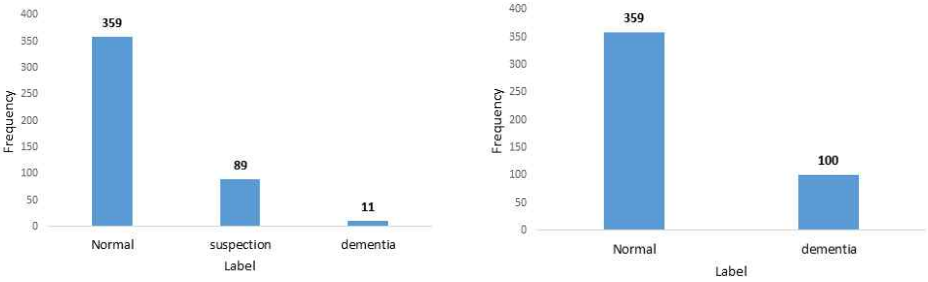


Figure 1. The distribution of the characteristics of the newly created variable

새로운 치매 위험군 변수는 총합이 0~1은 정상, 2~3은 의심군, 4~5는 치매군으로 구성하였으나, 치매 집단이 매우 적은 편이므로 의심군과 치매군의 범주를 합쳐 최종적으로 정상(0)과 치매(1)를 분류하는 이분형 변수를 예측모형에서 사용할 반응변수로 정의하였다. 다음 Figure 2는 최종적으로 생성한 치매 예측 반응변수에 대한 정상과 치매의 각 수준에 대한 관측 도수를 보여주고 있다.



(a) Original categories of dementia risk variable (b) dementia prediction response variable

Figure 2. The category of the new generated dementia prediction response variable

3. 세 가지 기계학습들에 대한 성능 비교

먼저, 우리는 라이프로그 데이터의 여러 가지 항목 중에서 치매 위험도를 판별하는데 영향력이 높은 항목들을 골라내기 위해서 상관분석을 실시하였다. 상관분석 결과에서 상관도가 비교적 높거나 중요한 요인으로 생각되는 변수들은 총 25개이며, 이들은 각각 성별, 나이, 키, 몸무게, 흡연 여부, 가족력 여부, 근골격계 질환, 신경계(파킨슨), 건강상태 양호, 음주량, 수축기 혈압, 이완기 혈압, 맥박, 좌뇌 쎄타, 좌뇌 델타 파워3, 자율신경활성, 부교감신경활성, 교감신경활성, 심박 변이도, 오른팔 체지방량, 왼팔 체지방량, 몸통 체지방량, 오른다리 체지방량, 왼다리 체지방량, 5kHz 파워로 측정한 오른 다리 위상각 등으로 나타났다.

다음으로 치매 위험도를 판별하기 위해서 선형 회귀모델인 로지스틱 회귀분석과 기계학습 알고리즘인 랜덤포레스트와 엑스지부스트 등 3가지 방법을 고려하였다. 첫 번째, 로지스틱 회귀분석(logistic regression)은 독립변수의 선형 결합을 이용하여 사건의 발생 가능성을 예측하는 통계적 기법으로, 식(3.1)을 통해 치매 가능성을 확률로 계산할 수 있다.

$$p(Y=1 \mid \boldsymbol{X}=\boldsymbol{x}) = \frac{1}{1+e^{-\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{x}}}, \quad \boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{x} = \beta_0 + \beta_1 x_1 + \cdots \beta_p x_p \tag{3.1}$$

두 번째, 랜덤포레스트(random forest)는 Breiman(2001)에 의해 제안된 지도학습 알고리즘 중 하나로 분류 문제에서 많이 사용된다. 여러 개의 의사 결정 나무를 앙상블하고 예측 인자가 목적변수에 미치는 영향력을 서로 관련되어 있다는 가정에서 분석하므로, 다양한 요인에 의해 결정되는 목적변수를 예측하기에 적절한 방법이다.

세 번째, 엑스지부스트(extreme gradient boosting, xgboost)는 여러 개의 약한 트리를 만들어 에러를 낮추며, 부스팅을 활용한 앙상블 기법이다. 과적합을 방지하기 위한 가지치기와 의사 결정 나무의 복잡도에 패널티 부과하기 위한 규제항(regularization term)을 도입하여 모델의 복잡도를 규제하면서 예측 정확도가 높은 방법이다. 식 (3.2)에서 $\hat{y_t}$는 예측값, $f_k(x_t)$는 k번째 의사결정나무, Obj는 목적함수, $l(y_t, \hat{y_t})$는 손실함수, $\Omega(f_k)$는 규제항을 의미한다.

$$\hat{y_t} = \sum_{k=1}^{K} f_k(x_t), \quad f_k \in P, \quad Obj = \sum_{t=1}^{n} l(y_t, \hat{y_t}) + \sum_{k=1}^{K} \Omega(f_k), \; \Omega(f_k) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{3.2}$$

위의 세 가지 치매 위험도 예측모형들의 성능을 평가하기 위해서 정확도(accuracy), 민감도(recall), 특이도(specificity), AUC (area under the ROC curve) 측도를 사용하였다. 정확도는 전체 중에서 옳게 분류한 비율, 민감도는 전체 양성 중에서 옳게 분류한 양성 비율, 특이도는 전체 음성 중에서 옳게 분류한 음성 비율을 나타낸다. AUC는 이 측도들을 이용하여 다양한 임계값에서의 모델 분류 성능을 나타낸 그래프로 큰 값일수록 우수한 분류 성능을 가진다. 우리는 의학 자료 분석에서 가장 중요하게 생각하는 측도인 민감도와 AUC를 기준으로 모델들을 비교한다. 또한, 분류 결과표(confusion matrix)를 통해 집단별 분류 결과를 명확히 확인할 수 있다.

마지막으로 주어진 자료를 기반으로 세 가지 기계학습 알고리즘들을 이용하여 치매 위험도 판별의 정확성을 알아보기 위하여 비교 분석하였다. 다음 Table 1은 테스트 데이터에 대한 세 가지 기계학습 모형들의 예측 성능 결과로 네 가지 측도값을 보여주고 있다. 이들 결과로부터 우리는 로지스틱 회귀모형이 가장 우수한 성능을 보이는 것을 알 수 있고, 다음으로 엑스지부스트 모형 그리고 랜덤포레스트 모형 순으로 예측 성능이 우수한 것을 알 수 있다.

Table 1. The diagnosis accuracy of dementia risk for three machine learning methods

Model	Accuracy	Recall	Specificity	AUC
Random forest	0.92	0.69	0.96	0.83
Xgboost	0.91	0.77	0.93	0.85
Logistic regression	0.82	0.85	0.81	0.86

4. 요약 및 결론

본 연구는 광주광역시 서구 건강관리센터에서 수집한 다양한 라이프로그 데이터를 기반으로 여러 가지 기계학습 알고리즘들을 사용하여 고령층의 치매 위험도를 사전에 예측하고 이들의 치매 정도를 완화할 수 있는 바람직한 건강 관리를 방안을 제시하기 위해 연구를 수행하였다.

연구결과로부터 우리는 세 가지 치매 위험도 분류 모형인 로지스틱 회귀, 랜덤포레스트, 엑스지부스트 중에서 로지스틱 회귀모형이 가장 우수한 성능을 보이고, 다음으로 엑스지부스트 모형, 랜덤포레스트 모형 순으로 성능이 우수함을 알 수 있었다.

앞으로 연구수행 방향은 지금까지 연구된 결과를 바탕으로 노인분들이 스스로 치매 정도를 파악할 수 있는 앱 기반 시스템을 개발하는 것이며, 이를 널리 사용할 수 있도록 홍보할 예정이다.

References

Breiman, L. (2001). Random forests, Machine learning, 45(1), 5-32.

Javeed, A., Dallora, A. L., Berglund, J. S., Ali, A., Ali, L., Andernerg, P. (2023). Machine Learning for Dementia Prediction: A Systematic Review and Future Research Directions, *Journal of Medical Systems,* 47, 1-25.

Li, R., Wang, X., Lawler, K., Garg, S., Bai, Q., Alty, J. (2022). Applications of artificial intelligence to aid early detection of dementia: A scoping review on current capabilities and future directions, *Journal of Biomedical Informatics,* 127, 1-13.

Sim, H. D., Jeong, M. R., Lee, M. H., Yang, S. J., (2020). Classification of EEG for Dementia Patientswith One-Dimensional Convolution Neural Network. Journal of the Korean Society of Mechanical Engineers, Volume B, 44(4), 237-244.

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 295-295

Open Computer Vision Software for Healthcare and Urban Mobility Research in the Big Data Era*

Donghyeok Choi[1], *Sangjin Kim*[1], *Jai Woo Lee*[2]

Abstractact

There have been great advancements in the fields of computer vision with the development of data analysis techniques and computational efficiency. In this project, we have built a user-friendly image processing tools using the tkinter libraries in Python by defining functions for each computer vision method and implementing pre-processing, filtering, feature extraction, and clustering on images collected from various fields such as environmental sciences, medical sciences, and mobilities. We have sought to assess the utility of automated image processing software to improve image classifiers even by non-professionals with no coding and no machine learning expertise. Thus, this program can be easily utilized by researchers without programming experience. This program can be used for various purposes by scholars in academic fields such as education, environmental sciences and medicine. This project describes open access, source, and integration to serve various research and education purposes of cutting-edge techniques. These strategies have been implemented in a computer vision software environment and achieved broad and significant impacts.

Keywords : Computer vision, User Interface, Machine Learning, Healthcare, Urban Mobility

---

*This article is financially supported by the 2023 College of Public Policy at Korea University.
[1]Department of Big Data Science, College of Public Policy, Korea University, Sejong, Republic of Korea.
[2](Corresponding Author) Department of Big Data Science, College of Public Policy, Korea University, Sejong, Republic of Korea; E-mail: jaiwoolee@korea.ac.kr

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 297-297

지역사회 거주 불면증 노인에서 수면의 질, 우울, 스트레스 및 인지기능과의 관련성 연구*

한은경[1]

요 약

본 연구의 목적은 지역사회 거주 불면증 노인의 수면 질, 우울, 스트레스 및 인지기능을 조사하고 이들 변수 간의 관련성을 조사하는 것이다. 본 연구는 서술적 조사연구 설계이며 자료 수집은 K도 S시 소재 시니어센터, 복지관에 등록된 노인 가운데 60세 이상이며, 최근 3개월 이상 불면증을 호소하는 111명을 대상으로 하였다. 대상자들은 구조화된 설문지를 통해서 수면의 질(PSQI), 단축형 우울척도(GDSSF-K), 스트레스(PSS)를 측정하였으며 인지기능은 몬트리올 인지평가(Moca-K)를 실시하였다. 자료분석은 SPSS 25.0을 사용하였으며 대상자의 일반적 특성은 빈도와 백분율, 평균과 표준편차로 분석하였다. 대상자의 수면의 질, 우울, 스트레스, 인지기능의 상관관계는 Pearson's correlation coefficient를 산출하였다. 연구결과 수면의 질 평균 점수는 9.41±3.67점, 우울은 5.55±3.78점, 스트레스는 16.62±6.91점, 인지기능 점수는 22.73±5.13점이었다. 상관관계 분석 결과 인지기능은 수면의 질(r=-0.45, p<.001), 우울(r=-0.32, p<.001), 스트레스(r=-0.56, p<.001)와 유의한 음의 상관관계를 나타냈다. 이와 같은 연구결과를 바탕으로 지역사회 거주 불면증 노인에서 수면의 질 저하, 우울, 스트레스, 인지기능 수준과 연관되어 있음을 확인하였다. 그러므로 불면증 노인의 수면, 심리적 증상과 인지기능을 복합적으로 향상시킬 수 있는 프로그램 개발이 필요하다.

주요용어 : 노인, 수면, 우울, 스트레스, 인지기능

---

*이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. RS-2023-00239600).

[1]13135 경기도 성남시 수정구 산성대로 553 박애관 307호, 을지대학교 간호학과 부교수,. E-mail: ekhan@eulji.ac.kr

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 299-299

재무데이터를 이용한 기업부도 예측
(Predicting Corporate Default Using Financial Data)

노시현[1]

요 약

본 연구는 현대 금융 시장의 복잡성과 변동성, 특히 금융 위기 및 경제 불확실성의 증가로 인해 기업 부도 예측의 중요성이 강조됨에 따라, 재무 데이터를 활용한 기업 부도 예측 모델을 개발하고 그 효율성을 검증하는 것을 목적으로 한다. 한국 투자 시장을 배경으로, 경제의 변동성과 금융 위기 상황에서 기업의 재무적 건전성과 부도 위험성 사이의 연관성을 분석하고 예측모형을 탐색하였다. 로지스틱 및 프로빗 모델을 사용하여, 회계 변수를 포함한 다양한 재무 지표를 변수로 하여 분석하였다. 변수들은 건전성, 규모, 성장성, 수익성, 유동성, 현금 흐름, 활동성 등을 고려하여, 각각의 영향력을 평가하여 기업 부도의 가능성을 예측하였다. 코스피와 코스닥 시장에 상장된 기업들을 대상으로 하며, 2000년에서 2022년까지의 데이터를 사용하였다. 부도기업의 정의는 상장 폐지 사유 중 흡수 합병, 물/인적 분할 등을 제외한 순수하게 Kisline 휴폐업 정보에 따라 폐업한 기업을 대상을 기업을 부도로 정의하여 표본을 수집하였다. 연구 결과는 로지스틱 회귀 분석과 프로빗 모델을 통해 도출되었으며, 두 모델 모두 높은 예측 정확도를 보여준다. 특히, 로지스틱 모형은 98.92%의 부도 적중률을, 프로빗 모형은 98.95%의 적중률을 나타냈다. 이 연구는 몇 가지 한계점을 갖고 있다. 첫째, 업종별 차이에 대한 구체적인 분석이 부족하며, 둘째, 거시경제 변수의 고려가 부족하다. 이러한 한계를 극복하기 위해 향후 연구에서는 회계 정보 외의 추가적인 거시경제 및 산업 변수들을 포함시키고, 모델의 일반화 가능성을 강화할 필요가 있다. 이와 같은 개선이 이루어진다면, 부도 예측 모델은 기업의 재무적 건전성을 평가하고 위험을 관리하는 데 더욱 효과적인 도구로 자리매김할 수 있을 것이다.

주요용어 : 회계변수, 로지스틱 모형, 프로빗 모형, 오분류율, 제1종 오류

[1]03063 서울시 종로구 성균관로 25-2, 성균관대학교 일반대학원 퀀트응용경제학과 석사과정. (주)나이스피앤아이 대체투자평가본부 대체투자평가정책실 대체투자평가정책팀 연구원. E-mail: vs6609@naver.com

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 301-301

B형 간염 환자의 간암 발병 예측 모형 연구

서준호[1], 남진현[2]

요 약

최근 의학 분야의 발전은 대량의 환자 데이터와 기계학습 기술의 진보에 힘입어 혁명적인 변화를 겪고 있다. 특히, 기계학습을 이용한 특정 질환의 발병 모형, 질환으로 인한 사망 예측 및 질환의 발병 요인을 탐색하는 연구가 두드러지게 진행되고 있다. 간암은 암 발생률 대비 높은 사망률을 보이면서 그 위험성이 증가하고 있는 주목할 만한 질환 중 하나이며, 간암 발병의 주된 원인 중 하나로는 만성 B형 간염이 큰 비중을 차지하고 있다. 또한, 만성 B형 간염 환자 중에서 간암으로 진행되는 경우가 다양해 B형 간염 환자의 간암 발병 조기 예측과 그 원인을 파악하는 것이 중요하다. 본 연구에서는 건강보험심사평가원에서 제공하는 건강보험 청구자료를 활용하여 B형 간염 환자의 간암 합병증 예측 모형을 비교하였다. 특히 B형 간염에서 간암으로 진행하는 동안 간암에 영향을 주는 기저 질환을 찾기 위해 B형 간염 진단 후 나타나는 모든 진단 상병을 예측 인자로 활용하였다. 이때 진단 상병 외 다른 요인들의 교란 효과를 줄이기 위해 PS 매칭을 활용해 군간 동질성을 확보하였으며, 예측 모형으로 Penalized Logistic Regression, Random Forest, Extreme Gradient Boosting, Support Vector Machines 모형을 사용하였다. 본 연구에서는 B형 간염 환자의 전수 자료를 이용하였기 때문에 자료의 크기 문제와 클래스 간 불균형 문제를 해결하기 위해 부트스트랩과 언더 샘플링을 기반으로 반복적으로 모형을 구축하였으며, 각 반복 내에서 10-fold CV를 통해 각 모형에 활용되는 최적의 모수를 적용하였다. 성능 비교 결과 Random Forest 모형이 가장 좋게 나타났으며, 본 연구에 사용된 진단 상병 외에 시술이나 처방약의 정보를 활용하여 모형의 성능을 향상 시킨다면 간암 합병증을 조기에 예방할 수 있을 것으로 기대된다.

주요용어 : B형 간염, 간암, 기계학습, 발병 예측, 주요 질병 확인

[1]30019 세종특별자치시 세종로 2511, 고려대학교 일반대학원 응용통계학과 석사과정.
E-mail: tjwnsgh0223@korea.ac.kr
[2](교신저자) 세종특별자치시 세종로 2511, 고려대학교 빅데이터사이언스학부 부교수.
E-mail: jinhnam@korea.ac.kr

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 303-

Predicting the customer of cafeteria using unstructured data

Kyeongjun Lee[1]

Abstract

This study aimed to predict the number of meals served in a group cafeteria using machine learning technology. Menu feature variables were created through the Word2Vec technique and clustering, and a stacking ensemble model was constructed using Random Forest, Gradient Boosting, and Cat Boost as sub-models. Results showed that Cat Boost had the best performance, and the ensemble model showed an 8% improvement in performance. The study also found that the date factor had the greatest influence on the number of diners in a cafeteria, followed by menu characteristics and other factors. The implications of the study include the potential for machine learning technology to improve predictive performance and reduce food waste, as well as the removal of subjective elements in menu classification. Limitations of the research include limited data cases and a weak model structure when new menus or foreign words are not included in the learning data. Future studies should aim to address these limitations.

Keywords: Ensemble model, machine learning, food waste, prediction, word embedding.

1. Introduction

Accurately predicting the number of diners in a cafeteria is essential for optimal production and cost management. Inaccurate predictions can lead to problems with ingredient supply and storage, decreased customer satisfaction due to inefficient employees, and loss of profits (Cheng et al., 2003). However, many cafeteria operators rely on subjective experience to make estimations due to budget constraints and a lack of experts (Lim, 2016; Jeon et al., 2019). To address this issue, research is needed to objectively and quantifiably anticipate the number of customers, considering the development of big data and artificial intelligence technology.

Estimating the exact number of diners in a cafeteria is complex due to various factors. Firstly, there is no absolute standard, since the factors affecting the count of diners are diverse and vary depending on the restaurant type, geographical location, and characteristics of the surrounding area. Secondly, individual social activities change over time, making it difficult to

[1]Department of Mathematics and Big Data Science, Kumoh National Institute of Technology, 61 Daehak-ro, Gumi, Gyeongbuk 39177, Korea. E-mail: indra_74@naver.com

update data and models. Finally, since restaurants predominantly manage unstructured data, including text, rather than structured data, researchers require expertise and skills in the preprocessing process (Jeon et al., 2019). Given these challenges, accurately forecasting the count of diners necessitates overcoming them through advanced data analysis and modeling techniques.

Studies on forecasting the number of customers can be categorized into statistical models and machine learning models based on the methodology. In most previous studies, variables such as the menu, date, weather, etc., were identified as major predictors in estimating the volume of meals. Additionally, studies have utilized unstructured data to classify menu ingredients and recipes into the Dewey Decimal Classification (DDC). The Dewey Decimal Classification system was devised in 1873 to organize the collections and catalogs of the Amherst University Library in the United States, and it is currently the most widely used classification system worldwide. However, this approach has limitations, as it involves subjective judgment from the researcher. Therefore, this study aims to minimize the researcher's partial intervention and improve performance by applying the word embedding technique, commonly used in text mining.

The composition of this paper is as follows. First, we will review previous studies related to estimating the count of customers in the cafeteria to derive significant variables and unstructured data to generate derived variables. Next, we will build a predictive model, validate and compare it with other models. Then, we will present the results and implications and discuss the study's limitations.

The results are expected to be used as a basis for estimating the number of diners at cafeterias operated by both public and private sectors. Furthermore, the findings can be utilized as research data for cost reduction in companies and as the basis data for establishing policies to reduce food waste in local governments.

References

Baek, O. H., Kim, M. Y., and Lee, B. H. (2007). Menu satisfaction survey for business and industry foddservice workers - Focused on food preferences by gender, *Journal of The Korean Society of Food Culture*, **22**, 511-519.

Blecher, L., and Yeh, R. J. (2008). Forecasting meal participation in university residential dining facilities, *Journal of Foodservice Business Research*, **11**, 352-362.

Breiman, L. (2001). Random forests, *Machine Learning*, **45**, 5-32.

Brown, G. (2010). Ensemble learning, *Encyclopedia of Machine Learning*, **312**, 15-19.

Cetinkaya, Z., and Erdal, E. (2019). Daily food demand forecast with artificial neural networks: Kirikkale University case, *In 2019 4th International Conference on Computer Science and Engineering*, 1-6.

Cheng, L., Yang, I. S., and Baek, S. H. (2003). Investigation on the performance of the forecasting model in university foodservice, *Journal of Nutrition and Health*, **36**, 966-973.

Dhillon, I. S., and Modha, D. S. (2001). Concept decompositions for large sparse text data using clustering, *Machine Learning*, **42**, 143-175.

Dorogush, A. V., Ershov, V., and Gulin, A. (2018). CatBoost: GradientBoosting with categorical features support, *arXiv preprint arXiv:1810.11363.*

Friedman, J. H. (2001). Greedy function approximation: A GradientBoosting machine, *Annals of Statistics*, 1189-1232.

Huang, A. (2008). Similarity measures for text document clustering, *In Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008)*, Christchurch, New Zealand, 4, 9-56.

Jeon. J., Park, E., and Kwon, O. B. (2019). Predicting the number of people for meals of an institutional foodservice by applying machine learning methods: S city hall case, *Journal of the Korean Dietetic Association*, **25**, 44-58.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space, *arXiv preprint arXiv:1301.3781v3.*

Montgomery, D. C., Peck, E. A., and Vining, G. G. (2021). *Introduction to linear regression analysis*, John Wiley & Sons.

Park, S., and Byun, Y. C. (2021). Improving recommendation accuracy based on machine learning using multi-dimensional features of Word2Vec, *Journal of Korean Institute of Information Technology*, **19**, 9-14.

Park, S. S., and Lee, K. C. (2018). Effective Korean sentiment classification method using word2vec and ensemble classifier, *Journal of Digital Contents Society,* 19, 133-140.

Polikar, R. (2006). Ensemble-based systems in decision making, *IEEE Circuits and Systems Magazine*, **6**, 21-45.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., and Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features, *Advances in Neural Information Processing Systems*, 31.

Ryu, K., and Sanchez, A. (2003). The evaluation of forecasting methods at an institutional foodservice dining facility, *The Journal of Hospitality Financial Management*, 11, 27-45.

Silverman, B. W. (2018). *Density estimation for statistics and data analysis*, Routledge.

Syarif, I., Zaluska, E., Prugel-Bennett, A., and Wills, G. (2012). Application of bagging, boosting and stacking to intrusion detection, *In Machine Learning and Data Mining in Pattern Recognition: 8th International Conference*, **8**, 593-602.

Yoo, J. E. (2015). Random forests, an alternative data mining technique to decision tree, *Journal of Educational Evaluation*, **28**, 427-448.

Zhou, Z. H. (2012). *Ensemble methods: Foundations and algorithms*, CRC press.

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 307-310

2단계 집락 조건부 무관질문모형*

이기성[1], 홍기학[2], 손창균[3], 박근화[4], 홍성준[5]

요 약

본 논문에서는 집락으로 구성되어 있는 모집단으로부터 얻고자 하는 정보가 민감할 때, 덜 민감한 속성과 강요질문을 활용한 확률장치로부터 '예'라고 응답한 사람들에게만 민감한 속성과 무관한 속성을 활용한 무관질문모형을 사용하도록 하는 조건부 모형에 2단계 집락추출법을 적용한 2단계 집락 조건부 무관질문모형을 제안하여 민감한 속성을 효율적으로 추정하였다. 그리고 일정하게 주어진 비용 하에서 분산을 최소화 하는 1단계 추출단위와 2단계 추출단위의 최적값을 도출하였다.

주요용어 : 민감한 속성, 덜 민감한 속성, 무관질문모형, 2단계 집락추출, 조건부 확률화응답모형.

1. 서론

Warner(1965)는 민감한 질문과, 민감한 질문과 배반인 질문을 활용한 확률화응답모형을 제안하여 민감한 속성의 모비율을 추정하였고, Greenberg et al.(1969)은 민감한 질문과 무관한 질문을 활용한 무관질문모형을 제안하였다. Loynes(1976)는 민감한 질문과 강요응답 '예'를 활용한 강요질문모형을 제안하였다. 그리고 Carr, Marascuilo.(1982)는 덜 민감한 속성과 강요질문으로 구성된 확률장치와 Loynes의 강요질문모형을 사용하여 민감한 속성에 대한 모비율을 효율적으로 추정해 내는 조건부 확률화응답모형을 제안하였다.

본 논문에서는 집락으로 구성되어 있는 모집단으로부터 얻고자 하는 정보가 민감할 때, 덜 민감한 속성과 강요질문을 활용한 확률장치로부터 '예'라고 응답한 사람들에게만 민감한 속성과 무관한 속성을 활용한 무관질문모형을 사용하도록 하는 2단계 집락 조건부 무관질문모형을 제안하고자 한다. 그리고 일정하게 주어진 비용 하에서 분산을 최소화 하는 1단계 추출단위와 2단계 추출단위의 최적값을 도출하고자 한다.

*이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (2022R1F1A1063263)

[1]55338 전북 완주군 삼례읍 삼례로 443 우석대학교 아동사회복지학부 교수. E-mail : gisung@woosuk.ac.kr
[2]58245 전남 나주시 건재로 185 동신대학교 컴퓨터학과 교수. E-mail : khhong@dsu.ac.kr
[3]38066 경북 경주시 동대로 123 동국대학교 빅데이터·응용통계학과 교수. E-mail : ckson85@dongguk.ac.kr
[4]07511 서울특별시 강서구 금낭화로 154 한국문화관광연구원 선임전문위원. E-mail : pkhkwen@kcti.re.kr
[5]13822 경기도 과천시 경마공원대로 107 마사회 말산업연구소 연구원. E-mail : hsj8129@naver.com

2. 2단계 집락 조건부 무관질문모형

이 절에서는 Lee, Hong.(2000)의 조건부 무관질문모형에 2단계 집락추출법을 적용하여 민감한 속성에 대한 추정을 할 수 있는 2단계 집락 조건부 무관질문모형을 제안하고자 한다. 모집단이 N개의 집락으로 구성되어 있고, 집락의 크기가 $M_i\,(i=1,2,\cdots,N)$개로 되어 있을 때, 이 모집단에서 집락 n개를 단순임의비복원추출한 다음, 추출된 각 집락으로부터 다시 조사단위 $m_i\,(i=1,2,\cdots,n)$개를 단순임의복원추출하는 2단계 집락추출법을 조건부 무관질문모형에 적용해 보고자 한다. i번째 집락의 첫 번째 단계에서 응답자들 m_i명은 아래와 같은 확률장치로부터 질문 1이 선택되게 되면 '예' 또는 '아니오'라고 응답하며, 질문 2가 선택되게 되면 '예'라고 응답하게 된다.

	질문내용	선택확률
질문 1	당신께서는 덜 민감한 속성 B_i가 있으십니까?	p_i
질문 2	'예'라고 응답하시오.	$1-p_i$

i번째 집락의 첫 번째 단계로부터 응답자들이 '예'라고 응답할 확률 λ_{i1}는 아래와 같다.

$$\lambda_{i1}=p_i\pi_{i1}+(1-p_i)\,.$$

여기서, π_{i1}은 i번째 집락의 덜 민감한 속성 B_i에 대한 모비율이다.

이 때, i번째 집락의 응답자들 m_i명 중에서 m_{i1}명이 '예'라고 응답했다면, $\hat{\lambda}_{i1}=\dfrac{m_{i1}}{m_i}$이 되므로 π_{i1}의 추정량 $\hat{\pi}_{i1}$는 아래와 같다.

$$\hat{\pi}_{i1}=\frac{1}{p_i}[\hat{\lambda}_{i1}-(1-p_i)]=\frac{1}{m_ip_i}[m_{i1}-(1-p_i)\,m_i]\,.$$

두 번째 단계에서는 i번째 집락의 첫 번째 단계에서 '예'라고 응답한 m_{i1}명의 응답자들만을 대상으로 다음과 같은 Greenberg et al.의 무관질문모형의 확률장치를 이용하도록 하여, 선택된 질문에 대해 '예' 또는 '아니오'라고 응답하도록 한다.

	질문내용	선택확률
질문 1	당신께서는 민감한 속성 A가 있으십니까?	p_i
질문 2	당신께서는 무관한 속성 Y가 있으십니까?	$1-p_i$

따라서 i번째 집락의 첫 번째 단계에서 응답자들이 '예'라고 응답했다는 조건하에, 두 번째 단계에서 '예'라고 응답할 확률 λ_{i2}는 다음과 같다.

$$\lambda_{i2}=p_i\frac{\pi_{i2}}{\lambda_{i1}}+(1-p_i)\,\pi_{iy}\,.$$

여기서, π_{i2}는 i번째 집락의 민감한 속성 모비율이고, π_{iy}는 i번째 집락의 무관한 속성 모비율로 사전에 알고 있다고 가정한다.

i번째 집락의 m_{i1}명의 응답자들 중에서 m_{i2}명이 '예'라고 응답했다면, $\hat{\lambda}_{i2}=\frac{m_{i2}}{m_{i1}}$가 되므로 모비율 π_{i2}의 추정량 $\hat{\pi}_{i2}$는 아래와 같다.

$$\hat{\pi}_{i2}=\frac{1}{p_i}\hat{\lambda}_{i1}\left[\hat{\lambda}_{i2}-(1-p_i)\pi_{iy}\right]=\frac{1}{m_ip_i}\left[m_{i2}-(1-p_i)\pi_{iy}m_{i1}\right].$$

한편, 2단계 집락추출법에 의한 민감한 속성 모비율 π_2는 다음과 같이 표현될 수 있다.

$$\pi_2=\frac{1}{M_0}\sum_{i=1}^{N}M_i\pi_{i2},\quad M_0=\sum_{i=1}^{N}M_i.$$

$i\,(i=1,2,\cdots,n)$번째 집락으로부터 응답자들이 단순임의복원추출되었을 때, 이러한 과정을 통해 얻어진 민감한 속성 모비율 π_2의 추정량 $\hat{\pi}_2$는 아래와 같이 정의할 수 있다.

$$\hat{\pi}_2=\frac{N}{nM_0}\sum_{i=1}^{n}M_i\hat{\pi}_{i2}=\frac{1}{n\overline{M}}\sum_{i=1}^{n}M_i\hat{\pi}_{i2},\ \overline{M}=\frac{M_0}{N}.$$

$\hat{\pi}_2$는 π_2의 비편향추정량이며, $\hat{\pi}_2$의 분산은 다음과 같다.

$$V(\hat{\pi}_2)=\frac{N-n}{nN(N-1)}\sum_{i=1}^{N}\left(\frac{M_i\pi_{i2}}{\overline{M}}-\pi_2\right)^2+\frac{1}{nN\overline{M}^2}\sum_{i=1}^{N}M_i^2\frac{1}{m_ip_i^2}\left[\lambda_{i1}(1-p_i)\pi_{iy}\{1-(1-p_i)\pi_{iy}\}-p_i\pi_{i2}\{2(1-p_i)\pi_{iy}+p_i\pi_{i2}-1\}\right].$$

또한, 집락의 크기가 $\overline{M}$로 동일한 N개의 집락으로부터 n개의 집락을 단순임의비복원으로 추출한 다음, 추출된 집락으로부터 다시 m_i개의 조사단위를 단순임의복원추출할 경우, 민감한 속성 모비율 π_2의 추정량 $\hat{\pi}_2$의 분산은 아래와 같다.

$$V(\hat{\pi}_2)=\frac{1}{nN}\sum_{i=1}^{N}(\pi_{i2}-\pi_2)^2+\frac{1}{nN}\sum_{i=1}^{N}\frac{1}{m_ip_i^2}\left[\lambda_{i1}(1-p_i)\pi_{iy}\{1-(1-p_i)\pi_{iy}\}-p_i\pi_{i2}\{2(1-p_i)\pi_{iy}+p_i\pi_{i2}-1\}\right].$$

만약 윗 식에서 m_i가 m으로 동일할 경우, 분산식은 다음과 같이 표현된다.

$$V(\hat{\pi}_2)=\frac{1}{nN}\sum_{i=1}^{N}(\pi_{i2}-\pi_2)^2+\frac{1}{nmN}\sum_{i=1}^{N}\frac{1}{p_i^2}\left[\lambda_{i1}(1-p_i)\pi_{iy}\{1-(1-p_i)\pi_{iy}\}-p_i\pi_{i2}\{2(1-p_i)\pi_{iy}+p_i\pi_{i2}-1\}\right].$$

다음으로, 적절한 표본배분을 위해 최적배분을 고려하여 비용이 일정하게 정해져 있다는 가정하에서 표본의 정도를 최대화 할 수 있는 n과 m의 값을 구해 보도록 하자.

먼저 2단계 추출을 고려하여 비용함수를 다음과 같이 정의하도록 하자.

$$C=c_0+nc_1+nmc_2.$$

여기서, C는 총비용, c_0는 고정비용, c_1은 1차 추출단위 당 비용, c_2는 2차 추출단위 당 비용이다.

비용이 일정하다는 조건 하에서 분산을 최소화 할 수 있도록 Lagrange 승수법으로 m의 최적값 m_0와 n의 최적값 n_0를 구해보면 아래와 같다.

$$m_0 = \sqrt{\frac{\frac{1}{N}\sum_{i=1}^{N}\frac{1}{p_i^2}\left[\lambda_{i1}(1-p_i)\pi_{iy}\{1-(1-p_i)\pi_{iy}\}-p_i\pi_{i2}\{2(1-p_i)\pi_{iy}+p_i\pi_{i2}-1\}\right]}{\frac{1}{N}\sum_{i=1}^{N}(\pi_{i2}-\pi_2)^2}}\ \frac{c_1}{c_2}\ .$$

$$n_0 = (C-c_0)\frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\pi_{i2}-\pi_2)^2/c_1}}{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\pi_{i2}-\pi_2)^2 c_1}+\sqrt{\frac{1}{N}\sum_{i=1}^{N}\frac{1}{p_i^2}\left[\lambda_{i1}(1-p_i)\pi_{iy}\{1-(1-p_i)\pi_{iy}\}-p_i\pi_{i2}\{2(1-p_i)\pi_{iy}+p_i\pi_{i2}-1\}\right]c_2}}\ .$$

따라서 m_0와 n_0의 값을 이용하여 최소분산 $V_{\min}(\hat{\pi}_2)$를 아래와 같이 구할 수 있다.

$$\begin{aligned}
V_{\min}(\hat{\pi}_2) &= \frac{1}{n_0}\left[\frac{1}{N}\sum_{i=1}^{N}(\pi_{i2}-\pi_2)^2+\frac{1}{m_0N}\frac{1}{p_i^2}\{\lambda_{i1}(1-p_i)\pi_{iy}\{1-(1-p_i)\pi_{iy}\}-p_i\pi_{i2}\{2(1-p_i)\pi_{iy}+p_i\pi_{i2}-1\}\}\right] \\
&= \frac{c_1+m_0c_2}{C-c_0}\left[\frac{1}{N}\sum_{i=1}^{N}(\pi_{i2}-\pi_2)^2+\frac{1}{m_0N}\sum_{i=1}^{N}\frac{1}{p_i^2}\{\lambda_{i1}(1-p_i)\pi_{iy}\{1-(1-p_i)\pi_{iy}\}-p_i\pi_{i2}\{2(1-p_i)\pi_{iy}+p_i\pi_{i2}-1\}\}\right] \\
&= \frac{1}{C-c_0}\left(\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\pi_{i2}-\pi_2)^2c_1}+\sqrt{\frac{1}{N}\sum_{i=1}^{N}\frac{1}{p_i^2}\left[\lambda_{i1}(1-p_i)\pi_{iy}\{1-(1-p_i)\pi_{iy}\}-p_i\pi_{i2}\{2(1-p_i)\pi_{iy}+p_i\pi_{i2}-1\}\right]c_2}\right)^2.
\end{aligned}$$

3. 결론

본 논문에서는 집락으로 구성되어 있는 모집단으로부터 얻고자 하는 정보가 민감할 경우, 덜 민감한 속성 B와 강요질문을 활용한 확률장치로부터 '예'라고 응답한 응답자들만을 대상으로 민감한 속성 A와 무관한 속성 Y를 활용한 무관질문모형을 사용하도록 하는 조건부 모형에 2단계 집락추출법을 적용한 2단계 집락 조건부 무관질문모형을 제안하여 민감한 속성을 효율적으로 추정하였다. 그리고 일정하게 주어진 비용 하에서 분산을 최소화 하는 1단계 추출단위와 2단계 추출단위의 최적값을 도출하였다.

참고문헌

Carr, J. W., Marascuilo, L. A. (1982). Optimal randomized response models and methods for hypothesis testing, *Journal of Educational Statistics*, 7, 295-310.

Greenberg, B. G., Abul-Ela, A. A., Simmons, W. R., Horvitz, D. G. (1969). The unrelated question randomized response model : Theoretical framework, *Journal of the American Statistical Association*, 64, 520-539.

Lee, G. S., Hong, K. H. (2000). A conditional unrelated question randomized response model, *Korean Journal of Computational and Applied Mathematics*, 8(1), 253-260.

Loynes, R. M. (1976). Asymptotically optimal randomized response procedures, *Journal of the American Statistical Association*, 71, 924-928.

Warner, S. L. (1965). Randomized response ; A survey technique for eliminating evasive answer bias, *Journal of the American Statistical Association*, 60, 63-69.

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 311-314

주파수 변화에 따른 IMU 센서 민감도 분석

이부건[1], 최원희[2], 임정현[3], 윤상후[4]

요 약

IMU(Inerial Measurement Unit)는 관성측정장치로 가속도 센서(Acceleration Sensor), 각속도 센서(Gyroscope)로 이루어져 있으며(6축) 일부는 지자기 센서(Magnetometer)도 포함한다(9축). 물체에 작동하는 가속력, 진동력, 충격력 등을 감지하여 움직임의 변화에 따른 가속도 변화를 순간적으로 감지한다. 가속도를 적분하면 물체의 진행 방향에 대한 속도를 계산하여 물체의 위치를 알 수 있다. 센서의 주파수가 높으면 오차가 작지만 주파수가 낮으면 오차가 커진다. 센서에서 수집된 가속도 값을 적분하므로 시간에 따른 오차가 누적되는 단점이 있어 주파수 변화에 따른 오차의 민감도를 알기 위해 256Hz를 기준으로 128Hz, 64Hz, 32Hz 주파수를 변화하며 분석하였다. 연구에 사용된 자료는 Madgwick의 “3D tracking with IMU”로 계단과 복도의 이동자료이다. 분석결과 가속도 센서는 주파수 변화에 따른 민감도가 낮지만, 각속도 센서는 주파수 변화에 따른 민감도가 높았다. 그리고 라이프로그 데이터를 이용하여 시각화를 하였고 비록 정확하게 시각화가 되지 않았지만 어느정도 위치가 잡혔고 높은 주파수를 사용하면 좀 더 정확하게 위치를 알 수 있다.

주요용어 : 가속도, 각속도, 주파수, 지자기, IMU.

1. 서론

IMU(Inertial Measurement Unit)는 물체의 기울어진 각도를 측정하는 관측측정장치로 가속도 센서(Acceleration sensor), 각속도 센서(Gyroscope sensor)로 이루어진 6축 센서와 지자기 센서(Magnetometor sensor)도 포함한 9축 센서가 있다(Song et al., 2018). 가속도 센서는 가속도를 측정하는 센서로 물체에 동적 힘(가속력, 진동력, 충격력)이 발생했을 때 움직임의 변화에 따른 가속도의 변화(단위:m/s^2)를 감지하는 센서이다. 각속도 센서는 대표적인 관성 센서로, 물체의 회전변화량인 각속도(단위:rad/s)를 측정하는 센서이다. 지자기 센서는 지구 자기장의 세기(자기속선)와 방향(자북을 기준으로 틀어진 각도)을 측정하는 센서이다.

가속도계는 선형 가속도를 측정하고, 각속도계는 회전 가속도를 측정한다. 그러므로 이 두 센서의 값을 적분하면 속도와 각도가 구해진다. 위치와 방향은 속도와 각도를 다시 적분하면 구할 수 있으므로 센서만으로 위치가 추정된다. 하지만, 적분 과정에서 오차가 누적되기 때문에 정확도가

[1]38453 경상북도 경산시 진량읍 대구대로 201, 대구대학교 통계학과. 학부생 E_mail:leebugun@naver.com
[2]38453 경상북도 경산시 진량읍 대구대로 201, 대구대학교 통계학과. 학부생 E_mail:coh0208@naver.com
[3]42734 대구광역시 달서구 송현로 205, 위니텍 연구기획팀 팀장 E-mail: cheerupm2@naver.com
[4](교신저자)38453 경상북도 경산시 진량읍 대구대로 201, 대구대학교 빅데이터학부 부교수.
E-mail: statstar@daegu.ac.kr

저하되는 문제가 발생한다.

Madgwick, Harrison, Vaidyanathan(2011)은 256Hz 주파수의 IMU 센서를 이용한 계단과 복도 이동 자료를 수집하여 경사하강 알고리즘으로 복원하였다. 본 연구는 적분 과정에서 발생하는 오차의 정도가 얼마인지를 파악하기 위해 256Hz 주파수를 128Hz, 64Hz, 32Hz 주파수로 변환하여 연구를 수행하였다. 가속도계와 각속도계의 6축의 유사 비율을 계산하여 운동의 강도에 따른 오차의 민감도를 확인하였다. 마지막으로 Chung 등(2021)이 32Hz로 수집한 라이프로그 데이터를 이용하여 실험자의 이동경로를 시각화해보았다.

2. 연구방법론

가속도를 이용하여 위치를 파악하기 위해서는 가속도 크기(accelerometer magnitude) 계산이 필요하다. 가속도의 크기는 다음 식으로 계산된다.

$$magnitude = \sqrt{a_x^2 + a_y^2 + a_z^2},$$

여기서 a_x, a_y, a_z는 가속도계의 x,y,z,축 값을 의미한다. 이렇게 구해진 가속도 크기는 버터워스 필터(butterworth filter)를 이용하여 로우 필터(low filter)와 하이 필터(high filter)로 노이즈와 이상치를 제거한다. 노이즈와 이상치가 제거된 값을 적분하면 속도가 구해진다.

물체의 위치에 방향 전환을 반영하기 위해 각속도계에 쿼터니언(quaternion)의 오일러각 변환 방식을 적용한다. 오일러 각은 물체의 방향을 3차원공간에서 세 개의 회전으로 표현하는 방법으로 다음 식으로 정의된다.

$$\alpha(\theta_x, \theta_y, \theta_z), \beta(\phi), \gamma(\psi),$$

여기서 α는 x축을 기준으로 한 Roll회전, β는 y축을 기준으로 한 Pitch회전, γ는 z축을 기준으로 한 Yaw회전을 의미한다. 오일러 각은 직관적이지만 짐벌락(Gimbal Lock)과 같은 문제가 발생 할 수 있어 쿼터니언으로 변환하여 안정적인 회전을 얻는다. 쿼터니언 식은 다음과 같다.

$$q = a + bi + cj + dk,$$

여기서 a는 스칼라 부분(실수)이고, bi, cj, dk는 벡터부분(허수)이다. 가속도계로 구한 속도에 적분한 후 각속도계 이용한 쿼터니언을 반영하면 물체의 위치가 구해진다.

3. 연구결과

Madgwick, Harrison, Vaidyanathan(2011) 계단복도 이동데이터(256Hz)를 시각화 하면 Figure 1이다.

계단 오르는 모습과 복도 걷는 모습이 그리고 층 중간 회전하는 모습도 잘 시각화되어 있다.

주파수를 256Hz에서 128Hz, 64Hz, 32Hz로 변환하였을 때 가속도계 및 각속도계의 6축 센서의 유사정도를 행동상태별로 요약하면 Table1이다. 멈춰있는 상태에서는 가속도계의 모든 주파수에서 98%이상의 설명력을 보여주었다. 하지만 각속도계에서는 78.67%(128Hz), 67.99%(64Hz),

59.03%(32Hz)로 주파수 변화에 따른 오차의 차이가 큰 편이다. 움직이는 상태에서는 256Hz에 비해 128Hz에서 89.64%, 64Hz에서 78.20%, 32Hz에서 60.69%로 멈춰있는 상태보다 가속도계의 유사성이 낮은편이다. 각속도계 역시 멈춰있는 상태에 비해 움직이는 상태에서 유사성이 낮았다.

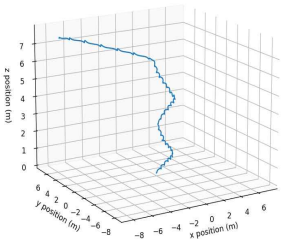
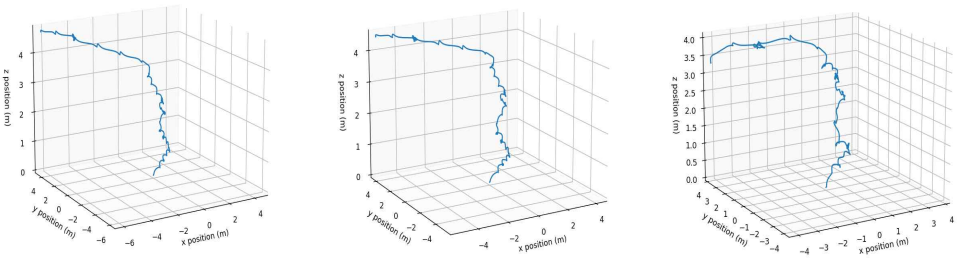


Figure 1. 256Hz data visualization using “3D tracking with IMU”

Table 1. The overlap ratio for each frequency and motion state

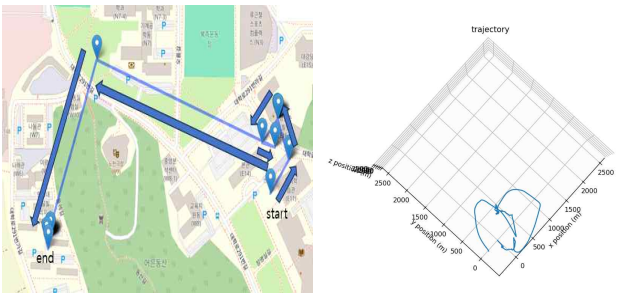
motion state	IMU 6-DOF	256Hz vs 128Hz	256Hz vs 64Hz	256Hz vs 32Hz	128Hz vs 64Hz	128Hz vs 62Hz	128Hz vs 32Hz
stop	accleration_x	99.74%	99.62%	99.52%	99.77%	99.65%	99.77%
	accleration_y	99.70%	99.54%	99.33%	99.69%	99.45%	99.50%
	accleration_z	99.55%	99.29%	98.86%	99.48%	98.90%	99.19%
	sub total	**97.02%**	**96.68%**	**96.26%**	**98.80%**	**98.40%**	**98.83%**
	gyroscope_x	90.23%	81.90%	71.52%	86.09%	76.68%	74.72%
	gyroscope_y	82.37%	83.85%	76.09%	87.60%	80.10%	89.26%
	gyroscope_z	98.31%	93.96%	93.56%	94.19%	93.97%	98.26%
	sub total	**78.67%**	**67.99%**	**59.03%**	**80.26%**	**66.94%**	**65.01%**
move	accleration_x	93.79%	77.68%	61.16%	80.25%	63.26%	69.94%
	accleration_y	95.86%	89.04%	73.42%	91.31%	75.16%	77.02%
	accleration_z	85.19%	67.68%	52.49%	71.85%	54.78%	65.42%
	sub total	**89.64%**	**78.20%**	**60.69%**	**81.90%**	**65.62%**	**70.03%**
	gyroscope_x	92.79%	84.71%	67.00%	86.11%	69.30%	71.43%
	gyroscope_y	84.70%	67.48%	46.86%	71.43%	49.37%	57.50%
	gyroscope_z	92.78%	83.80%	70.76%	85.80%	71.71%	75.32%
	sub total	**75.40%**	**56.68%**	**39.20%**	**64.66%**	**39.20%**	**48.37%**



(a) 128Hz (b) 64Hz (c) 32Hz

Figure 2. 128Hz, 64Hz, 32Hz data visualization using “3D tracking with IMU”

계단복도 이동데이터의 가속도계와 각속도에서 구한 위치(x,y,z)를 시각화하면 Figure 2이다. 누적 오차를 구하기 위해 256Hz를 기준으로 각 주파수와의 거리 차의 평균제곱거리를 구했다. 그 결과 126Hz는 41.9, 64Hz는 46.7, 32Hz는 52.9로 주파수가 낮아질수록 위치의 정확도가 낮아지고 있다.Chung 등(2021)이 32Hz로 수집한 라이프로그 데이터에서 GPS와 IMU를 이용하여 시각화 하면 Figure 3이다. GPS 관측자료는 분 단위로 관측되어 실험자의 이동이 직선으로 표현되어 실험자가 정확히 어떤 경로로 움직였는지 알기 어렵다. IMU를 이용한 이동경로는 이동방향 및 이동위치가 시각화 되었으나 낮은 주파수로 인해 이동 위치의 정확성이 낮았다.



(a) GPS (c) IMU

Figure 3. The visualization of GPS and IMU data.

4. 결론

본 연구는 IMU의 주파수에 따른 위치추적의 민감도가 얼마나 되는지 정량적으로 평가한 연구이다. 높은 주파수는 정확성이 높지만 데이터 양이 많아 저장공간을 많이 차지하고 낮은 주파수는 정확성이 떨어지지만 용량이 적어 저장공간을 많이 차지한다. Madgwick, Harrison, Vaidyanathan(2011)의 계단복도 이동 데이터(256Hz)를 기준으로 128Hz, 64Hz, 32Hz를 비교한 결과 멈춰있는 상태에서는 가속도의 유사성이 95%이상 유지되지만 각속도계의 유사성은 60% 수준으로 물체의 속도와 이동거리는 잘 추적하지만 방향을 추적하는데 한계가 있다. 움직이는 상태에서는 주파수가 낮아질수록 가속도계의 유사성이 89.64%(126Hz), 78.20%(64Hz), 60.69%(32Hz)로 낮아지고 각속도계의 유사성이 80%가 되지 않는다. 즉, 운동량이 증가할수록 주파수에 따른 물체의 속도, 이동거리, 그리고 방향 추적이 어려워진다. 따라서 IMU를 이용하여 위치를 추적하기 위해서는 되도록 높은 주파수가 필요하다.

Reference

Chung, S., Jeong, C. Y., Lim, J. M., Lim, J., Noh, K. J., Kim, G., & Jeong, H. (2022). Real world multimodal lifelog dataset for human behavior study. ETRI Journal, 44(3), 426-437.

Madgwick, S. O., Harrison, A. J., & Vaidyanathan, R. (2011, June). Estimation of IMU and MARG orientation using a gradient descent algorithm. In 2011 IEEE international conference on rehabilitation robotics (pp. 1-7). IEEE.

Song, J. W., Liu, Y., Yang, H. S., Lee, K. B., & Lee, J. M. (2018). Three-dimensional pedestrian position estimation algorithm using waist-mounted IMU sensor. J. Inst. Control Robot. Syst., 24, 453-459.

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 315-315

Comparing Scan Statistics For Zero-Inflated Spatial Count Data: A Case Study Of Arson Data

Jiwon Lee[1], *Yejin Kim*[2], *Jiae Park*[3], *Yujin Oh*[4], *Donghwan Lee*[5]

Abstract

In the analysis of count data such as crime and disease incidence, the excess of zero counts are commonly observed so that the conventional Poisson model has a limit to explain the data. To identify significant areas of event occurrence, we consider the Scan-Poisson statistic, based on the Poisson model, and the Scan-ZIP (Zero-Inflated Poisson) statistic, which accounts for zero inflation. By conducting the simulation studies, we compare the results of both approaches. We demonstrate that the Scan-ZIP statistic is more effective in identifying clusters in zero-inflated spatial data. Also, the results show that Scan-Poisson statistic steadily deteriorates as the number of zeros increases, producing biased inferences. To illustrate the usage, we applied these methods to arson incident data from 426 administrative districts in Seoul (2012-2021) to detect significant areas of arson risk.

Keywords: spatial scan statistics, zero-inflation, multiple cluster detection, arson data, Seoul

[1]Graduate Student, Department of Statistics, Ewha Womans University, 52, Ewhayeodae-gil, Seodaemun-gu, Seoul, 03760, Korea. E-mail : whitebean@ewhain.net
[2]Graduate Student, Department of Statistics, Ewha Womans University, 52, Ewhayeodae-gil, Seodaemun-gu, Seoul, 03760, Korea. E-mail : yegenuine@ewhain.net
[3]Graduate Student, Department of Statistics, Ewha Womans University, 52, Ewhayeodae-gil, Seodaemun-gu, Seoul, 03760, Korea. E-mail : wldopie@ewhain.net
[4]Graduate Student, Department of Statistics, Ewha Womans University, 52, Ewhayeodae-gil, Seodaemun-gu, Seoul, 03760, Korea. E-mail : yujin5o@ewhain.net
[5](Corresponding Author) Associate Professor, Department of Statistics, Ewha Womans University, 52, Ewhayeodae-gil, Seodaemun-gu, Seoul, 03760, Korea. E-mail : donghwan.lee@ewha.ac.kr

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 317-317

CLIP 기반 모델을 활용한 다중 라벨 Zero-shot 분류

정승현[1], 티머만 크리스토프[2], 김창우[3], 이우진[4]

요 약

CLIP(Contrastive Language – Image Pretraining) 모델은 비전 기반 검색, 제품 추천, 자동 이미지 캡션 생성 등 다양한 분야에서 활용되며, 특히 이미지와 텍스트 간의 관계를 학습하여 학습되지 않은 데이터에 대한 Zero-shot 분류 문제에서 우수한 성능을 보인다. 이러한 모델을 학습하는 과정에서 실제 세계 데이터셋은 하나 이상의 라벨을 가지는 다중 라벨 데이터셋으로, 종종 클래스 또는 레이블 간의 불균형이 나타날 수 있다. 이러한 불균형은 다중 라벨 Zero-shot 분류 문제에서 성능 저하의 원인이 될 수 있습니다. 본 논문에서는 CLIP 모델을 사용하여 다중 라벨 Zero-shot 분류 문제를 해결하는 새로운 방법을 제안한다. 본 논문에서 사용하는 CLIP 모델은 이미지와 텍스트 모두를 입력값으로 받기 때문에, 두 입력값에 대해서 데이터셋의 불균형을 줄여 성능을 개선하였다. 이미지 데이터의 불균형을 해결하기 위해 Diffusion 모델이나 Mixup과 같은 이미지 증강 기법을 사용한다. 이를 통해 소수 클래스의 데이터를 더 많이 생성하여 클래스 간의 균형을 유지할 수 있다. 또한, 텍스트 데이터의 불균형을 해결하기 위해 손실 함수에서 소수 레이블에 가중치를 추가하는 pos-weight 방법을 활용한다. 이런 방식으로 소수 레이블에 더 높은 가중치를 부여하여 모델이 각 레이블을 더욱 균등하게 고려하도록 한다. 다양한 실제 세계 데이터셋에 대한 실험 결과를 통해 CLIP 모델이 다중 라벨 Zero-shot 분류 문제에 대한 우수한 성능을 실증적으로 입증한다. 이 연구는 CLIP 모델의 활용 가능성을 확장하며, 다중 라벨 Zero-shot 분류 문제에 대한 해결책을 제시한다. 마지막으로 실험 결과를 통해 모델의 한계와 가능성에 대한 논의를 진행한다.

주요 용어 : CLIP, 다중 라벨 Zero-shot 분류, Diffusion 모델, Mixup, pos-weight

[1]04620 서울특별시 중구 필동로1길 30, 동국대학교 전자전기공학부 학사과정. E-mail: alzkdpf23@dgu.ac.kr
[2]04620 서울특별시 중구 필동로1길 30, 동국대학교 컴퓨터· AI학과 석사과정. E-mail: christti@dgu.ac.kr
[3]04620 서울특별시 중구 필동로1길 30, 동국대학교 컴퓨터· AI학과 석사과정. E-mail: note8335@dgu.ac.kr
[4](교신저자) 04620 서울특별시 중구 필동로1길 30, 동국대학교 컴퓨터· AI학과 교수. E-mail: wj926@dgu.ac.kr

Proceedings of the Korean Data Analysis Society
January 25-26, 2024, pp. 319-319

차종별 교통사고 수를 이용한 사회적 거리 두기 영향 분석

최원희[1], 이부건[2], 윤상후[3]

요 약

전국적으로 코로나19 바이러스는 사회 전반에 심각한 영향을 미쳤다. 사회적 거리 두기 정책으로 인해 외출하는 사람들이 줄어들고 온라인을 통한 회의, 쇼핑, 교육과 같은 비대면 활동량이 증가했다. 본 연구는 코로나19 바이러스가 산업에 미치는 영향을 분석하기 위해 경기도에서 발생한 차종별 교통사고 수를 인과 영향 모형으로 분석하였다. 교통량은 산업별로 다른 영향을 미치고 여러 사건, 정책에 민감하게 반응한다. 이를 이용하여 사회적 거리 두기 영향을 간접적으로 분석하기 위해 2륜 오토바이, 화물차, 개인형 이동 수단(PM) 등 측정이 어려운 교통량 자료들을 포함하고 있는 교통사고 자료를 이용했다. 첫 거리 두기 정책 기간(2020-02-29 ~ 2020-05-05) 차종별 교통사고 수를 비교한 결과 승용차와 승합차의 교통사고 수는 감소했고 2륜 오토바이, 화물차, 개인형 이동 수단(PM)의 교통사고 수는 증가했다. 이는 사회적 거리 두기 정책 3단계 시행에 따라 재택근무, 사적 모임 금지 등으로 승용차와 승합차의 교통량이 줄어들었으나, 물류와 배송이 증가하여 2륜 오토바이, 화물차의 교통사고가 증가함을 의미한다. 2차(2020-08-15 ~ 2020-09-21), 3차(2020-11-23 ~ 2021-02-15), 4차(2021-07-12 ~ 2021-10-31) 유행 기간에서 차종별 교통사고 수는 사회적 거리 두기 정책과 시기별 특성에 맞춰서 변했다. 코로나19 바이러스는 산업 전반에 영향을 미치며, 교통량 역시 이에 영향을 받는다. 다양한 상황에서 이러한 분석을 통해 피해를 최소화할 수 있다.

주요용어 : 교통사고, 사회적 거리 두기, 인과 영향 모형, 코로나19 바이러스.

[1]38453 대한민국 경상북도 경산시 진량읍 대구대로 201, 대구대학교 과학생명융합대학 통계학과 학부생.
E-mail: coh0208@naver.com
[2]38453 대한민국 경상북도 경산시 진량읍 대구대로 201, 대구대학교 과학생명융합대학 통계학과 학부생.
E-mail: leebugun@naver.com
[3](교신저자) 38453 대한민국 경상북도 경산시 진량읍 대구대로 201, 대구대학교 통계학과 부교수.
E-mail: statstar@daegu.ac.kr

trendresearch
Technology for Market Research

트랜드리서치는?

국내&해외 마케팅 및 사회/여론 조사 경력이 20년 넘는 전문가가 직접 연구에 참여해 통찰력 있는 INSIGHT 제공
트랜드리서치의 약 72만 ACCESS PANEL을 통한 정확한 조사 수행이 가능함

마케팅 전문가와
사회조사 전문가의 결합

자사 패널을 직접 보유
(720,000명)

별도의 정성조사 전문가 활용

주요 서비스

온라인/모바일 조사

전문인력을 통해 국내 온라인 및
모바일 조사를 진행하고 있으며
Quick survey, Mobile Diary
등의 조사 수행

마케팅/소비자 조사

마케팅 의사 결정을 위한 시장
및 소비자 정보 제공

학술/연구/석박사 논문 조사

학술/연구/석박사 논문의 경우
사회기여 차원으로 일반 조사와는
차별화된 단가(다운된 단가)로 제공

T-Panel 현황

해외조사
해외 제휴 네트워크를 활용하여
신속한 조사 진행이 가능합니다.

대표성이 있습니다
샘플링을 통해 추출된 표본들에게만
이메일, 문자, 알림톡을 발송하여
조사를 진행합니다.

특정 포탈 사이트 회원이 아닙니다
조사참여 횟수, 조사 참여일,
참여조사 종류 및 응답의
성실도를 체계적으로
관리합니다.

분석 서비스
조사 결과의 통계 검정
결과를 제공하여 분석의
신뢰성을 제고합니다.

서울 서초구 반포대로24길 77 상원빌딩 3층 | TEL 02.3442.0604 | FAX 070.4275.5529 | EMAIL cu@trendmr.com

데이타솔루션 | 前 SPSS Korea

데이타솔루션은 Digital Transformation Journey를 함께 하는 Cloud와 Data 기반의 Digital Innovation Provider입니다. 데이타솔루션은 Data와 Application Lifecycle 전 과정에 최적화된 Total Service와 솔루션을 보유하고 있으며, 특히 Cloud, IT Modernization, Analytics, Intelligence Service를 통해 고객에게 새로운 Insight와 차별화된 Value를 제공하고, 고객의 비즈니스 성공을 위한 Digital Innovation을 함께 합니다.

주요 분석 사업 영역

SPSS Korea를 전신으로 한 Data & Analytics 부문에서는 통계소프트웨어/ML/AI 솔루션 공급 및 데이터분석 컨설팅 사업을 수행하며, 고객사의 데이터 기반 의사결정을 지원하고 있습니다. 30여 년간 제조, 금융, 유통, 통신 등 다양한 산업 군에서 경험 및 지속적인 연구개발을 통해 통계분석에서 ML/AI로 진화하는 시장의 변화를 이끌고 있습니다.

Analytics Consulting	MAaaS (Managed Analytics as a Service)	Intelligence Solution	AI/BigData Education
· 수요 예측 · 리스크 예측 · 상품 추천 · 품질 분석 및 예지 보전 · 이미지 분석 · 텍스트 분석 · 고객 분석	· MAaaS PASP · PoC · Pro · Mainteanace · MAaaS by AI Model	· KoreaPlus Statistics (Embaded on SPSS) · KNIME · IBM CP4D · Market Mind · Brightics AI · AI.NER	· 빅데이터러닝센터 · SPSS/Python/ R 교육 연간 80여개 온/오프라인 강의제공

데이타솔루션 분석 사례

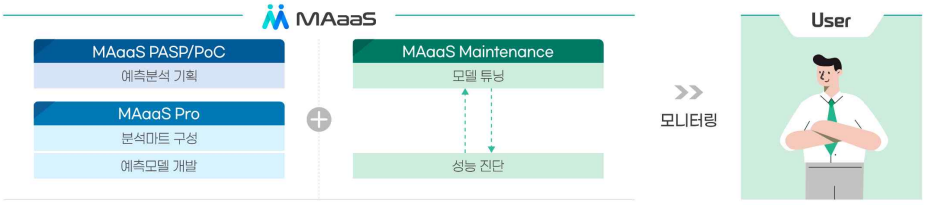
수요 예측	제조 | 수요예측 분석 시스템 구축 유통 | 판매량 예측 모델 개발 물류 | 물동량 예측 시스템 구축 공공 | 열 수요 예측 모델 개발	텍스트 분석	통신 | VoC 분석 금융 | 텍스트 분석을 활용한 평판 분석 유통 | 텍스트 분석을 통한 마케팅 증권 | 챗봇 메시지 분석
상품 추천	유통 | AI 상품 추천 프로그램 구축 유통 | 온라인 쇼핑몰 개인화 상품 추천 모델 개발 유통 | 개인화 추천 서비스 모델 구축 금융 | 고객관리 시스템을 통한 상품 추천	품질 분석 및 예지보전	제조 | 취약 부위 예측 모델 개발 제조 | 특정 공정이상 예측 분석 제조 | 공정 최적화 및 설비 고장 모델 개발 화학 | 수율 예측 시뮬레이터 개발
이미지 분석	제조 | 이미지 데이터 기반의 양불 판정 제조 | 비전 검사를 통한 양불 판정		

AI 예측서비스, 이제 구독하세요!

No 플랫폼 | No 데이터수집 | No 데이터사이언티스트 | No 유지관리

MAaaS | Managed Analytics as a Service

MAaaS는 AI 예측모델 기획부터 유지관리까지의 Data Life-cycle에서 필요한 컨설팅을 제공하는 데이타솔루션의 AI 예측분석 토탈 서비스입니다. 이제는 분석 예측 결과를 구독으로 제공받으세요.



[문의] Analytics영업팀 MAaaS@datasolution.kr 02-3467-7238 서울 강남구 언주로 620, 현대인텔렉스 10층, www.datasolution.kr

DATAsolution

지디에스컨설팅그룹

Company : 회사소개 및 연혁

2003년 이래, 데이터 가공/분석 및 유통의 신시장 개척에 주력하였습니다.

회사명	㈜지디에스컨설팅그룹
설립일	2015년 2월
대표이사	김은석
업종	데이터베이스, 소프트웨어 자문, 개발 및 공급, 전자상거래업
주소	서울 성동구 광나루로 228, 9층 (성수동2가, 렉스모터스)
연락처	Tel: 02-2135-8895(대표) Fax: 02-2135-8896

2018.07
고려대 '데이터 사이언스 센터' 업무 제휴 (교육 및 인턴쉽)

2017.07
KB국민카드 '빅데이터 센터' 업무 제휴 (카드 데이터 상품개발)

2016.09
KT '빅데이터 지원단' 업무 제휴 (빅데이터 가공, 중개, 서비스 분야)

2016.08
데이터 거래 중개 사업 개시 (미래창조과학부 ICT 기금사업)
부설연구소 '한국공공정보개발원' 설립 (산기협 2016113799호)

2016.07
빅데이터 분석 사업 개시 (지속가능한 빅데이터 사업, 경기도)

2015.04
국가 DB 구축 사업 개시 (유동인구 실사 DB 구축, NIA)

2015.02
㈜지디에스컨설팅그룹 설립
(㈜지디에스케이 內 컨설팅 사업본부의 법인 분할)

2004.10
벤처기업 등록 (041134221-1-01033호)

2003.02
㈜지디에스케이 설립 (대표이사 김은석 外)

Business : 사업 영역

데이터 활용 및 유통의 일관된 서비스 제공!!

사업영역	빅데이터 분석 컨설팅	데이터 가공	빅데이터 플랫폼 구축	데이터 판매/유통
		자세히 보기	자세히 보기	자세히 보기
상품 및 서비스	· 데이터 관련 ISP 및 계획수립 · 주요 빅데이터 (통신, 카드) 공급 및 활용 · 공공 데이터 활용 서비스 · 데이터 분석 및 시각화 서비스	· 공공 개방 정보의 정제, 가공 및 제공 · 통신, 카드 등 빅데이터 정제, 결합 및 제공 · 비정형 데이터 가공 및 제공	· 통계분석 도구 (알트릭스) 판매 · 데이터 융복합 및 통합 시스템 구축 · 빅데이터 분석 시스템 구축	· 데이터 사용영역별 데이터 판매 · 업무 영역별 맞춤형 데이터 구축 및 제공 · 주제별 맞춤형 데이터 제공
사업사례	· 국방기술품질원, 빅데이터 시범사업 ('17~) · 경기도, 지속가능한 빅데이터 분석 ('16) · 서울시, GRDP 분석 모델링 구축 ('16) · 건강보험공단, 빅데이터노인점환모델구축('14) · 서울시, 유동인구 DB 구축 및 모델링 ('15)	· 삼성전자, 법인영업 데이터 제공 ('15~) · 한샘, 가구판매 고객 데이터 제공 ('15~) · 코웨이, 고객정보 관리 데이터 제공 ('06~) · KT, 지사 관리 데이터 제공 ('10~) · 기업은행, 영업점 관리 데이터 제공 ('10~)	· 창원시, 빅데이터 분석시스템 구축 ('17) · KT, AI 조류독감 시각화 시스템 구축 ('16) · 경상북도, 통계정보시스템 구축 ('15) · KT, 지사 관리 데이터 제공 ('10~현재) · 서울시, 도시통계 시스템 구축 ('14, '15)	· 한국정보화진흥원, 데이터 거래 중개 플랫폼 구축 ('16) · 한국정보화진흥원, 소상공인 데이터 제공 플랫폼 ('17) · 경상북도, 서비스 인구구축을 위한 데이터('16) · 코웨이, 고객정보 관리를 위한 데이터('16)외 다수