PROCEEDINGS OF THE KOREAN DATA ANALYSIS SOCIETY

January 26-27, 2023

at

Haeundae Centum Hotel

Haeundae-gu, Busan, Korea

(http://www.ecentumhotel.com/)

THE KOREAN DATA ANALYSIS SOCIETY Founded 1998

PROCEEDINGS OF THE KOREAN DATA ANALYSIS SOCIETY January 26, 2023

한국자료분석학회

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

문서번호 : KDAS-2301000

시행일자 : 2022. 1. 26.

(경유)

수 신 : 전 회원

참 조 :

선결			지시		
접수	일자				
	시간		결재		
	번호				
담당자					

제목: 한국자료분석학회 2022년 동계 학술논문발표대회

---

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

2. 본 학회 2022년도 동계 학술논문발표대회를 다음과 같이 개최함을 알려 드리오니 많은 참여 바랍니다.

   1) 개최일자 : 2023년 1월 26일(목) ~ 27일(금)

   2) 장 소 : 해운대 센텀 호텔(부산)

한 국 자 료 분 석 학 회 장 한 상 태

한국자료분석학회 2022년도 동계 학술논문발표대회 준비위원

조직위원회

위원장	한 상 태	호서대학교 빅데이터AI학과
부위원장	김 기 환	고려대학교 빅데이터사이언스학부
	박 승 열	㈜ 케이스탯리서치 회장
	박 찬 근	한국해양대학교 데이터사이언스전공
	윤 성 민	부산대학교 경제학부
	이 영 섭	동국대학교 통계학과
	이 학 배	연세대학교 응용통계학과
	홍 재 범	부경대학교 경영학부
위원	김 용 대	서울대학교 통계학과
	이 동 희	경기대학교 경영학부
	이 성 건	성신여자대학교 수리통계데이터사이언스학부
	진 서 훈	고려대학교 빅데이터사이언스학부
	허 태 영	충북대학교 정보통계학과
편집위원장	강 현 철	호서대학교 빅데이터AI학과
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	김 지 수	가천대학교 간호학과
	신 승 준	고려대학교 통계학과
	이 성 건	성신여자대학교 수리통계데이터사이언스학부
	제 상 영	고려대학교 경제통계학과
	조 장 식	경성대학교 정보통계학과
위원	곽 일 엽	중앙대학교 통계학과
	김 성 환	건국대학교 응용통계학과
	김 양 진	숙명여자대학교 통계학과
	김 은 석	㈜ GDS 컨설팅그룹 대표이사
	김 지 연	㈜ 케이스탯리서치 대표이사
	김 지 희	강원대학교 응급구조학과
	김 태 훈	경성대학교 경제금융물류학부
	박 만 식	성신여자대학교 수리통계데이터사이언스학부

한국자료분석학회
2022년도 동계 학술논문발표대회

- 일시 : 2023년 1월 26일 ~ 27일
- 장소 : 부산 센텀호텔 (부산 해운대구 센텀3로 20)
- 주최/주관 : 한국자료분석학회
- 후원: 한국연구재단, (주)케이스탯리서치, (주)GDS컨설팅, 나이스평가정보, 베가스, SAS 코리아, 데이터솔루션, 정보통신정책연구원

1월 26일 목요일

17:00 - 19:00 편집위원회 (금루: 부산시 해운대구 센텀메디컬센터 3층)

17:00 - 19:00 이사회 (금루: 부산시 해운대구 센텀메디컬센터 3층)

1월 27일 금요일

09:00 - 18:00 포스터발표 (18층 에메랄드룸)

09:30 - 11:00 논문발표 A1 - A3 (18층 에메랄드,제이드,사파이어룸)

11:10 - 12:10 튜토리얼 (18층 에메랄드룸) 좌장: 이성건교수(성신여대)

- 제목 : 사회과학 분야에서 척도구성과 활용:
  감성통제이론의 어의차이척도를 중심으로
- 강사 : 이명진교수 (고려대학교 사회학과)

12:10 - 12:30 총회 (18층 에메랄드룸)

12:30 - 13:40 중식시간 (개별)

13:40 - 14:40 초청 강연 (18층 에메랄드룸) 좌장: 조형준교수(고려대)

- 제목 : 평균낼 수 없는 것들의 통계
- 강사 : 정성규교수 (서울대학교 통계학과)

14:50 - 16:20 논문발표 B1 - B3 (18층 에메랄드,제이드,사파이어룸)

16:30 - 18:00 논문발표 C1 - C3 (18층 에메랄드,제이드,사파이어룸)

논문발표 A1 (18층 에메랄드룸) **기계학습** 좌장 : 최호식(서울시립대학교)

09:30 | 11:00

1. RPKG: Robust Prompt Learning with Knowledge Graph ····· 1
김예원, 임용택*, 윤도경, 송경우(서울시립대학교 인공지능학과)

2. Spatial-Temporal Transformer for Water Level Forecasting ····· 3
홍성철*, 전종준(서울시립대학교 통계학과)

3. Growth Model for Scientific Collaboration Hypergraphs ····· 5
정호현*(성신여자대학교 수리통계데이터사이언스학부),
김성호(KAIST 수리과학과), Frederick Kin Hing Phoa(Academia Sinica)

4. Censored data에 대한 토빗과 헤크만 모형의 활용 ····· 7
김정환*, 조형준(고려대학교 통계학과)

5. 비확률 표본을 이용한 확률화응답 모형의 보정 추정 ····· 9
손창균*(동국대학교), 홍기학(동신대학교), 이기성(우석대학교)

논문발표 A2 (18층 제이드룸) **교육/알고리즘** 좌장 : 양정호(성균관대학교)

09:30 | 11:00

1. 은닉마코프모델을 이용한 적응형 학습 경로 생성에 관한 연구 ····· 13
최현희*(라이브데이터), 이윤지(중앙대학교 응용통계학과),
이하윤(중앙대학교 소프트웨어학과)

2. 사회연결망분석을 활용한 코로나와 교육에 관한 최근 연구동향 분석: KCI 2020-2022 학술논문의 NodeXL 분석 및 시각화 ····· 17
임수현*, 김수경, 김지선, 손현규, 양정호(성균관대학교 교육학과)

3. 중국 데이터 조사 현황 및 개선방안 ····· 21
뢰예*, 진선순, 김화, 소성, 손정옥, 마혜정, 양정호(성균관대학교)

4. 교원의 역량개발을 위한 비형식 학습 참여 효과에 관한 연구: piaac의 한국,미국, 핀란드 데이터 비교 ····· 23
최정윤*, 양정호(성균관대학교)

논문발표 A3 (18층 사파이어룸) **간호** 좌장 : 박공주(김해대학교)

09:30 | 11:00

1. 다문화 학령기아동의 자아존중감 모형 ····· 27
김미경*(대동대학교), 박공주(김해대학교)
2. 간호대학생의 코로나19 위험인식, 수업참여도 및 학습만족도에 관한 연구 ····· 33
김민영*(경남정보대학교)
3. 자궁경부암 진단 여성의 삶의 질 영향요인 ····· 39
전은미*(배재대학교)
4. 청소년 위험음주 영향 요인 ····· 41
박민희*(동서대학교)

논문발표 B1 (18층 에메랄드룸) **통계** 좌장 : 정윤서(고려대학교)

14:50 | 16:20

1. Locally Averaging Composite Quantile Regression ····· 45
김수인*(고려대학교 통계학과), 조윤영(카카오게임즈), 정윤서(고려대학교 통계학과)
2.The use of Truncated Loss on Multi-armed outcome weighted learning ····· 47
심정은*, 신승준(고려대학교 통계학과)
3.Ensemble Teacher: Semi-Supervised Semantic Segmentation with Teacher's Cross-Pseudo Supervision ····· 49
곽현민, 김성환*(건국대학교 응용통계학과)
4. 형상점 자료의 변환법 및 분류모형의 비교 ····· 51
김준홍*, 최용석(부산대학교 통계학과)
5. 표본 자료의 분위수 계산에 관한 고찰: 몇 가지 연속형 분포의 경우 ····· 55
김혁주*(원광대학교 빅데이터금융통계학부)

논문발표 B2 (18층 제이드룸) **인공지능/머신러닝** 좌장 : 김기환(고려대학교)

14:50 | 16:20

1. 설명 가능한 인공지능(XAI) 기반 토지 피복 구성에 따른 지표면 온도의 영향분석 ····· 59
이동찬*, 변상영, 김기환(고려대학교 경제통계학과)

2. XAI 기법을 이용한 보이스피싱 이상거래 탐지 모델 연구 ····· 61
김주미*, 정여진(국민대학교, 데이터사이언스학과)

3. 앙상블 비모수적 변수 선택 방법 ····· 67
신웅섭*, 조형준(고려대학교 통계학과)

4. 머신러닝 기법을 활용한 소프트웨어 신뢰성 모형 비교에 관한 연구 ····· 69
김윤수*, 송광윤, 이다혜, 장인홍(조선대학교)

5. 머신러닝을 활용한 부도 예측 모형 연구: 소매 익스포저를 중심으로 ····· 71
조용복(동아대학교 경영정보학과), 박진수*(동아대학교 디지털금융학과)

논문발표 B3 (18층 사파이어룸) **경영/경제** 좌장 : 김태훈(경성대학교)

14:50 | 16:20

1. 영구적·일시적 현금흐름 뉴스와 할인율 뉴스에 따른 모멘텀 수익률 분석 ····· 75
오명*(부경대학교)

2. 환환경정책이 한국 제조업의 기업 성과에 미치는 영향 분석 ····· 79
이진희(국회예산정책처), 송철종*(선문대학교 국제경제통상학과)

3. 한국 가계부채의 형성요인에 대한 실증분석 ····· 83
고동우*(부산대학교 글로벌경제컨설팅학과), 윤성민(부산대학교 경제학과)

4. 항만 개발과 운영이 미치는 사회적 영향에 관한 연구 ····· 89
박호*(군산대학교 국제물류학과), 김상열, 장현미(부산대학교 국제전문대학원),
김태훈(경성대학교 경제금융물류학부)

5. 비대칭 마진 SVM을 이용한 기업부실 예측모형의 범주 불균형 문제 해결 ····· 91
조성임*, 김명종(부산대학교 경영학과)

논문발표 C1 (18층 에메랄드룸) **경영통계융합** 좌장 : 강현철(호서대학교)

16:30 | 18:00

1. 통계용어 표준화 및 체계적 구축방안에 대한 연구 ····· 95
한상태*(호서대학교 빅데이터AI학과), 장원철(서울대학교 통계학과),
여인권 (숙명여자대학교 통계학과), 임종호(연세대학교 응용통계학과)

2. 안전에 대한 객관적 지표와 주관적 인식 차이: 사회안전지수를 중심으로 ····· 97
박현수(충북대학교 국가위기관리연구소), 장안식*(케이스탯컨설팅)

3. 가우시안 필터를 활용한 주가예측의 최적화 ····· 101
김명종, 권성훈*(부산대학교 경영대학)

4. 기하평균 정확도 최적화 기반의 부스팅 모델의 범주 불균형
데이터 분류/예측 문제 적용 ····· 107
김명종, 안재현*(부산대학교 경영대학)

5. 서울서베이 표본추출틀 구축을 위한 방법론 연구 ····· 113
강현철*(호서대학교 빅데이터AI학과), 조수정, 이동수, 이은혜, 최성용(서울특별시)

논문발표 C2 (18층 제이드룸) **통계** 좌장 : 김정연(서울시립대학교)

16:30 | 18:00

1. 결정계수를 이용한 선형혼합모형의 선택 ····· 117
이용희*(서울시립대학교, 통계학과)

2. Quantile Estimation for Encrypted Data ····· 123
박민제, 김재선, 신성철(크립토랩), 최호식*(서울시립대학교 도시빅데이터융합학과)

3. A study on multiple imputation modeling for Korean EAPS ····· 125
박민정(통계청), 배윤종(한미약품), 김정연*(서울시립대학교 인공지능학과)

4. Development of Demand Response Transit mode choice models using
Quantum Computing ····· 127
박철규*, 최호식(서울시립대 도시빅데이터융합학과)

논문발표 C3 (18층 사파이어룸) **의료/환경/통계** 좌장 : 한진태(경성대학교)

16:30 | 18:00

1. 슬링을 이용한 열린사슬운동과 닫힌사슬운동의 턱당기기 동작 시 목굽힘근 두께 비교 ····· 129
이주훈*(경성대학교 일반대학원), 한진태(경성대학교 물리치료과)

2. 전방머리자세 환자의 목기능, 통증과 목빗근 두께가 삶의 질에 미치는 영향 ····· 133
강민지*, 한진태(경성대학교 물리치료과)

3. 발가락 사용유무가 능동적 발등굽힘 가동범위 및 하지근활성도에 미치는 영향 ····· 137
김인균*(서울스카이재활의학과), 최수홍(부산대학교병원 재활의학과), 이상열(경성대학교 물리치료학과)

4. 키워드 네트워크 분석을 통한 심부 디지털 트윈 국내외 공공 연구기관 논문 분석 ····· 141
안은영*, 김성용(한국지질자원연구원)

5. 비가 온 후의 토양수분 변화 고찰 및 예측 모델 개발 ····· 143
강소라*(전남대 수학/통계학과), 조완현(전남대 통계학과), 나명환(전남대 통계학과)

포스터발표 (18층 에메랄드룸) 09:00-18:00

1. 한국 노인의 건강정보이해력 영향요인 분석 ····· 147
김은경*(군산대학교 간호학과)
2. 지역사회 노인의 여가생활 ····· 149
박정혜*(경상국립대학교 간호학과)
3. 노인의 우울 위험 ····· 153
박정혜*(경상국립대학교 간호학과)
4. 여대생의 신체감시성이 섭식절제행동에 미치는 영향: 신체수치심의 매개효과 ····· 157
김문정*(부경대학교), 오은정(춘해보건대학교)
5. 간호간병통합서비스 병동 간호사의 환자중심간호 영향요인에 대한 연구 ····· 161
김미래*, 안정아(아주대학교)
6. 심부전 환자와 의료진 간 질병양상 변화 및 완화 의료에 대한
의사소통 요구도의 차이 ····· 163
김경화*(아주대학교), 안정아(아주대학교)
7. 간호대학생을 위한 당뇨병 교육 프로그램:체계적 문헌고찰 ····· 165
김경아*(가천대학교), 안정아(아주대학교)
8. 임상간호사 전향적 코호트 연구동향 분석: 텍스트 네트워크 및 토픽모델링 ····· 167
이은교*(아주대학교 간호과학연구소), 박성희, 김광현(아주대의료원, 아주대학교 간호학과),
이영진*(아주대학교 간호대학, 간호과학연구소)
9. Testing for House Price Bubble in Seoul ····· 169
심성훈*(부경대학교 경제학과)
10. 함수적 선형 모형을 활용한 아파트 매매 지수 예측 ····· 173
김희진*, 이은지(충남대학교)
11. 기계학습을 이용한 대학출원 특허의 소멸예측 ····· 175
김용우*(한양대학교 기술경영학과), 우상미(연세대학교 행정학과)
12. 층화추출법에 의한 조건부 Quatember 확률화응답모형 ····· 181
이기성*(우석대학교 아동사회복지학부), 홍기학(동신대학교 컴퓨터학과), 손창균(동국대학교
빅데이터응용통계학과), 박근화(한국문화관광연구원), 홍성준(마사회 말산업연구소)
13. 토픽모델링을 기반으로 한 디지털 신기술에 대한 트위터 데이터의 키워드 분석 ····· 185
이윤진*(성신여자대학교), 정호현(성신여자대학교 수리통계데이터사이언스학부)
14. 토픽모델링과 시계열 분석을 활용한 국내외 독과점 산업 동향 분석 ····· 189
이수지*(성신여자대학교), 정호현(성신여자대학교 수리통계데이터사이언스학부)
15. 함수적 회귀모형을 활용한 알츠하이머 관련 유전자 변이 탐지 연구 ····· 193
이윤석*, 이은지(충남대학교 통계학과)

포스터발표 (18층 에메랄드룸) 09:00-18:00

16. 아파트 주거만족도에 영향을 미치는 주거환경 요인에 대한 메타분석 ····· 195
김명인*, 윤상복, 이현경(동의대학교 도시공학과)
17. 인공신경망을 활용한 미세먼지 예측 모델 기초 연구 ····· 199
이재현*(동의대학교 대학원 도시공학과), 이현경, 윤상복(동의대학교 건설공학부 도시공학전공)
18. 한국 인공지능 학술논문 저자들 연결 네트워크 분석 ····· 203
양희건(연세대학교), 양희성(국군수도병원), 양희창*(Open Future Labs)
19. 모집단이 부분층화 가능한 경우의 비추정치 ····· 207
손경호*(케이스탯리서치)
20. 서울시 전기차 충전소 입지 선정 최적화 ····· 209
박진아*, 김혁구, 전종준(서울시립대학교)
21. 초정밀 금형 프레스 공정 데이터 분석을 통한 이상 탐지
AI 모델 개발 및 성능 평가 ····· 213
조한별*,설현욱, 정현진, 백범수, 김영순(경상국립대학교)
22. 장애인 거주시설 물리치료사들의 업무 만족도와 시간 외 근무시간 정도 ····· 215
노효련*(강원대학교)
23. 만성 뇌졸중 환자의 마비측과 보행기능 및 속도 ····· 219
김찬우, 노효련*(강원대학교)
24. 농식품 소비자 패널을 활용한 포도 소비행태 분석 - 수도권 소비자를 중심으로 ····· 223
이정명, 박건환, 김혜형, 이영순(경기도농업기술원)
25. Product spacing estimation of exponential logistic distribution under multiply
type hybrid censoring ····· 225
이경준(금오공과대학교), 박찬근*(한국해양대학교)
26. 엘라스틱 넷 벌점항을 적용한 비선형 지지벡터기계의 변수선택법 ····· 229
황호석*(서울시립대 통계학과), 최호식(서울시립대학교 도시빅데이터융합학과),
박창이(서울시립대학교 통계학과)
27. 다단집락추출에서의 분산 추정량의 비교 ····· 235
김유미*(한신대학교)
28. Text To Table을 적용한 Robotic Process Automation 사례 연구 ····· 243
이준서, 염승민, 이다운(서울시립대 융합전공학부),
백송희(서울시립대 도시공학과), 최호식(서울시립대 도시빅데이터융합학과)
29. 정규화 특이값 분해 기반 추천 시스템 모델 제안 및 비교 연구 ····· 245
김재훈*, 이도현, 이상헌, 김도윤, 임광묵, 육현섭, 서문민건, 홍석수, 허태영(충북대학교)
30. 방한 외래관광객의 관광지 연관분석을 통한 한국관광 개선방안 ····· 251
오유진*, 윤상복, 이현경(동의대학교 도시공학과)

튜토리얼

(1) 발표자 : 이명진(고려대학교 사회학과)

(2) 제목 : 감성조절이론과 어의차이척도

사회정체성(social identity) 개념의 가장 중요한 전제는 사람들은 하나의 대상을 특정한 사회적 집단에 소속된 하나의 구성원으로 파악한다는 것이다. 그리고 사회적으로 규정된 집단적인 가치나 문화적 가치에 견주어 그 대상의 사회적 역할을 정형화해간다는 것이다. 이러한 사회정체성 개념과 판단기준이 하나의 이론으로 형성된 것은 오스굿을 비롯한 일련의 사회심리학자들의 연구성과에 기인하고 있다. Osgood 등은 감성의 보편적인 세 가지 차원을 증명하기 위한 기초작업을 하였다. 측정의 방법은 모든 문화에 존재하는 개념들의 리스트를 제공하고 응답자에게 각 개념에 대한 형용사를 응답하도록 하였다. 그리고 각 형용사에 대한 정반대 의미의 형용사를 제시하였다. 이들은 이러한 정반대의 형용사 묶음들은 사회정체성 측정을 위한 의미분화척도(semantic differential scale)로 구성되었다. 그리고 인물이나 사물을 포함한 다양한 형태의 사회정체성을 평가하기 위해서 척도로 사용하였다. 응답자들은 제시된 각 개념에 대하여 등간척도로 구성된 의미분화척도를 이용하여 평가하게 된다.

이러한 과정을 통해서 구성된 자료를 기반으로, 이들은 사회별로 평균을 산출하고, 비교하여 범문화적 상관관계를 계산하였다. 이러한 상관관계를 요인분석(factor analysis)한 결과 척도가 주로 세 차원으로 구분되었다. 이러한 세 차원은 각각 평가(evaluation), 권력(potency), 활동(activity)으로 구분할 수 있다. 평가의 차원은 도덕, 미적 감각, 기능, 쾌락, 또는 다른 표준의 판단에 대한 찬성 또는 반대의 느낌과 관련된다. 권력의 차원은 물질적인 크기, 강제성, 힘, 사회적 권력, 중요성 등의 평가로 그 실체의 영향력과 관련된다. 활동의 차원은 행동의 경향, 속도, 지각이 있는 자극, 연령, 민감성 등의 판단으로 설명될 수 있는 실제의 자발성과 관련된다. 이 세 가지 차원은 모두 사회적 연구에 있어서 중요하다. 평가의 차원은 가치와 태도에 상응하고, 권력의 차원은 행동의 영향과 역할의 힘을 나타내며, 활동의 차원은 행동에서 개인의 경향과 정서적 활동의 수준을 평가한다.

이렇게 감성조절이론(Affective Control Theory)은 기본적으로 두 가지 형태의 가정에 근거하고 있다. 가장 기본적인 가정은 사회정체성을 통해 한 사회 전체의 전반적인 문화에 접근할 수 있다는 것이다. 즉 사람들이 다양한 사회적 행위(social behaviors) 그리고 사회적 상황(social setting)에 부여한 감성적 의미를 연구함으로써, 그 사회의 특정 역할의 점유자와 이들의 행동양식에 대한 규범적인(normative) 반응을 예측할 수 있다는 것이다. 두 번째 가정은 사람들은 어디에서나 일반적인 시각을 공유하지만, 감성의 특정한 패턴은 사회마다 다를 수 있다는 것이다. 따라서 범문화적으로 존재하는 감성의 공통된 측면들을 토대로 각 문화의 특성을 살펴보면, 각 문화의 차이점들을 효과적으로 보여줄 수 있다는 것이다.

초청강연

(1) **발표자 : 정성규(서울대학교 통계학과)**

(2) **제목 : 평균 낼 수 없는 것들의 통계**

통계의 기본적인 도구인 평균을 중심으로 통계학의 과거와 현재, 미래를 돌아본다. 평균이 통계학의 핵심 도구가 된 이유부터, 키의 평균, 점수의 평균, 평균 인간 그리고 산술적으로 평균낼 수 없는 것 등을 예로 삼는다. 구성비 (compositional) 데이터, 분포들의 데이터, 방향과 3D 모형들의 데이터, 확산텐서 데이터 등 흔히 비유클리드 데이터라 불리는 데이터에서 패턴을 파악하고 요약할 때, 간단한 산술평균으로는 부족하다. 산술평균의 일반화된 개념인 프레쉐평균과 최소제곱추정법의 일반화된 개념은 일반화 프레쉐평균을 소개하고, 그에 대한 통계 이론의 전개를 통해 현대 통계학의 도전과 성취를 알아본다.

Proceedings of the Korean Data Analysis Society
January 26-27, 2023, pp. 1-1

RPLKG: Robust Prompt Learning with Knowledge Graph*

김예원[1], 임용택[2], 윤도경[3], 송경우[4]

요 약

사전 훈련된 대규모 언어모델은 다른 데이터셋에 쉽게 전이 가능하고, 학습에 사용되지 않은 데이터에 일반화할 수 있다는 사실이 알려져있다. 최근 CLIP과 같은 멀티모달 사전 훈련 모델은 다양한 세팅의 실험에서 기존 모델들에 비해 상당한 성능 향상을 보였다. 하지만 라벨링된 데이터셋이 부족할 경우, 새로운 데이터셋이나 도메인에 대해 일반화시키는 것은 여전히 어려운 문제이다. Few-shot 세팅에서의 일반화 성능 향상을 위해 프롬프트 학습, 어댑터 와 같은 다양한 연구들이 진행되어 왔다. 그러나 이러한 방법론들은 해석할 수 없고, 사용하기 위해 높은 시/공간적 복잡도를 필요로 한다. 본 연구에서는 이러한 문제를 해결하고자 지식 그래프를 활용한 강건한 프롬프트 학습 (RPLKG)라는 새로운 방법론을 제시한다. 본 방법론은 지식 그래프에 기반하여 다양하고 해석 가능한 프롬프트의 집합을 생성한다. 그 뒤, 이미지 정보에 기반해 gumbel-softmax를 이용하여 프롬프트를 선택하는 과정을 최적화한다. RPLKG는 데이터셋에 따라서 프롬프트의 집합 중 최적의, 해석 가능한 프롬프트를 선택한다. RPLKG는 기존의 프롬프트 학습을 세 가지 측면에서 개선하였다. 1)해석 가능하고, 2)적은 비용으로 3)사전 지식을 사용할 수 있는 프롬프트를 생성하였다. 본 연구에서는 RPLKG의 성능을 검증하기 위해 few-shot 세팅과, 도메인 일반화 세팅에서 실험을 진행하였다. 모든 데이터셋에 대해 두 가지 세팅에서, RPLKG는 기존의 zero-shot 및 few-shot 기반 방법론들에 비해, 큰 성능 향상을 이루었다.

주요용어 : 멀티모달 학습, few-shot 학습, 지식 그래프.

---

*이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. 2021R1F1A1060117, No. 2022M3J6A1084845).

[1]02054 대한민국 서울특별시 서울시립대로 163 서울시립대학교 인공지능학과 석사과정.
E-mail: yeyewon12@uos.ac.kr

[2]02054 대한민국 서울특별시 서울시립대로 163 서울시립대학교 인공지능학과 석사과정.
E-mail: yongtaek.lim@uos.ac.kr

[3]02054 대한민국 서울특별시 서울시립대로 163 서울시립대학교 인공지능학과 학부재학생.
E-mail: dokyung36@uos.ac.kr

[4](교신저자)02054 대한민국 서울특별시 서울시립대로 163 서울시립대학교 인공지능학과 조교수.
E-mail: kyungwoo.song@uos.ac.kr

Proceedings of the Korean Data Analysis Society
January 26-27, 2023, pp. 3-3

Spatial-Temporal Transformer for Water Level Forecasting

Sungchul Hong[1], *Jong-June Jeon*[2]

Abstract

Forecasting the water level of the Han river is important to control the traffic and avoid natural disasters. There are many variables related to the Han river and they are intricately connected. In this work, we propose a novel transformer that exploits the causal relationship based on the prior knowledge among the variables and forecasts the four bridges of the Han river: Cheongdam, Jamsu, Hangang, and Haengju. Our proposed model considers both spatial and temporal causation by formalizing the causal structure as a chain graph and using masking methods. Due to this approach, we can have the explainability that consists to prior knowledge. Additionally, we propose a novel calibration method and loss function for high accuracy of extreme risk. In real data analysis, we use the Han river dataset from 2016 to 2021, and compare the proposed model with deep learning models.

Keywords : Water Level Prediction, Probabilistic forecast, Multivariate Time Series Forecasting, Explainable AI, Calibration, Transformer.

[1]02504 Graduate Student, Department of Statistics, University of Seoul, Seoulsiripdaero 163, Korea. E-mail: shong@uos.ac.kr

[2](Corresponding Author) 02504 Associate Professor, Department of Statistics, University of Seoul, Seoulsiripdaero 163, Korea. E-mail: jj.jeon@uos.ac.kr

Proceedings of the Korean Data Analysis Society
January 26-27, 2023, pp. 5-5

Growth Model for Scientific Collaboration Hypergraphs*

Hohyun Jung[1], *Sung-Ho Kim*[2], *Frederick Kin Hing Phoa*[3]

Abstract

A hypergraph is useful to express the relationship between two or more nodes. Real hypergraph data are typically weighted. We propose a weighted evolving hypergraph model that considers preferential attachment. The model allows variability on the two basic components of the evolving hypergraph: the number and the size of the hyperedges to be connected. Under the mild distributional conditions on the two varying quantities, we derive the exact degree distribution that asymptotically follows a power-law distribution. We find that the limiting power-law exponent is affected by the distribution of hyperedge sizes. The distribution of the number of hyperedges to be connected has a considerable impact on a small-degree range in which non-power-law behavior is frequently observed in real data. Moreover, we argue that the degree distribution of the model can be expressed as a mixture of the degree distributions with a fixed number of hyperedges to be connected. The validity and usefulness of the model are explained with interpretations via simulation study and real data analysis.

Keywords : Hypergraph, Preferential attachment, Network growth, Scientific Collaboration.

---

*This research was supported by the National Research Foundation of Korea (2021R1G1A1094103).

[1](Corresponding Author) Assistant 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

[2]Professor Emeritus, Departmet of Mathematical Sciences, KAIST, 291 Daehak-ro, Yuseong-gu, Daejeon 34141, Korea

[3]Research Fellow, Institute of Statistical Science, Academia Sinica, 128 Academia Rd., Section 2, NanKang, Taipei City 11529, Taiwan

Proceedings of the Korean Data Analysis Society
January 26-27, 2023, pp. 7-7

Censored data에 대한 토빗과 헤크만 모형의 활용

김정환[1], **조형준**[2]

요 약

토빗과 헤크만 모형은 censored data의 분석에 활용되는데, 이때 censored data란 자료의 형태가 일정한 부분에서는 관측값이 양의 확률로 다수가 분포, 다른 부분에서는 연속형 형태로 분포되어있거나 자료가 특정지점에서 절단되어 있는 것을 의미한다. 이처럼 절단되어 있는 자료의 경우, 일반적인 선형회귀모형을 사용하게 되면 문제점이 발생하기 때문에 토빗 모형과 헤크만 모형을 사용하지만 두 모형에는 차이점이 존재함에도 불구하고 명확하게 고려되지 않고 종종 혼용된 채로 사용되어진다. 따라서 여기서는 censored data의 형태를 세분화하여 시뮬레이션 자료를 통해 모형의 적합성을 먼저 비교하였으며, 이후 실제 자료를 사용하여 두 모형을 적합하였다. 그 결과, 자료의 절단에 영향을 주는 잠재변수가 없는경우에는, 토빗과 헤크만 모형 모두 잘 적합되나 토빗 모형이 참값에 더 근접하게 적합되면서도 간결하다는 것을 확인할 수 있었다. 하지만 자료의 절단에 영향을 주는 잠재변수가 존재한다면, 헤크만 모형만 적합이 잘 되는 것을 확인할 수 있었다.

주요용어 : censored data, 토빗 모형, 헤크만 모형.

[1]02841 서울특별시 성북구 안암로 145, 고려대학교 통계학과 박사과정. E-mail: kjhsjy0215@korea.ac.kr
[2](교신저자) 02841 서울특별시 성북구 안암로 145, 고려대학교 통계학과 교수. E-mail: hj4cho@korea.ac.kr

Proceedings of the Korean Data Analysis Society
January 26-27, 2023, pp. 9-12

비확률 표본을 이용한 확률화응답 모형의 보정 추정

손창균[1], 홍기학[2], 이기성[3]

요 약

본 논문에서는 민감한 속성에 대한 모비율 추정에서 간접 응답 방법인 확률화응답 모형을 활용할 수 있다. 한편 추정량의 분산 감소 및 무응답 편향을 축소하기 위해 민감한 속성과 관련이 있는 보조변수를 이용한 보정 추정과정을 기존의 확률 표본에 기반한 보정 추정방법을 비확률 표본에 기반한 LASSO 보정 추정 방법을 제안하였다.

주요용어 : 비확률 표본, 확률화응답모형, Warner모형, LASSO 보정 추정.

1. 서론

응답자의 신분노출을 꺼리는 개인적으로 민감한 속성에 대한 모집단 비율을 추정하기 위해 Warner(1965)는 간접응답 방법으로 확률화응답모형(randomized response model)을 처음으로 제안하였다. 한편 Deville and Sarnadal(1992)은 추정량의 분산 감소와 편향을 감소시키기 위해 표본설계당시의 설계가중치를 특정한 거리함수(distance function)를 최소로 하는 보정 추정(calibration estimation) 방법을 제안하였다. 이때 활용한 보조 정보는 관심변수와 강한 상관관계가 있는 보조변수에 대한 모집단 정보를 이용하였으며, 결과적으로 보정 추정량은 점근적으로 일반화회귀추정량(generalized regression estimator)으로 수렴함을 보였다.

본 논문에서는 확률화응답 추정량에 대한 보정 추정과정에서 활용되는 보조변수가 비확률 표본으로 수집된 경우 접근적으로 추정량의 강건성을 유지할 수 있는 LASSO 회귀 추정방법을 적용하여 보정된 확률화응답 추정량을 제안하고자 한다.

2. Warner의 확률화응답모형에 대한 보정추정

2.1. Warner의 확률화응답 모형

크기가 N인 모집단으로부터 크기가 n인 단순임의 복원추출에 의해 추출된 응답자들이 설문1을 선택활 확률이 P이고 설문2를 선택할 확률이 $1-P$인 다음과 같은 확률장치에 의해 선택한 설문에 대해 '예' 또는 '아니오'로 응답한다.

[1](교신저자) 38066 경북 경주시 동대로 123 동국대학교 빅데이터·응용통계학과 교수,
E-mail: ckson85@dongguk.ac.kr
[2]58245 전남 나주시 건재로 185 동신대학교 컴퓨터학과 교수. E-mail: khhong@dsu.ac.kr
[3]55338 전북 완주군 삼례읍 삼례로 443 우석대학교 아동사회복지학부 교수. E-mail: gisung@woosuk.ac.kr

설문1 : 나는 민감한 그룹 A 에 속한다.
설문2 : 나는 민감한 그룹 A 에 속하지 않는다.

이러한 확률화응답 모형으로부터 응답자가 '예'라고 응답할 확률은 $\phi = P\pi + (1-P)(1-\pi)$이며, 여기서 π는 민감한 그룹에 속하는 모집단 비율이다. n명의 응답자 중에서 '예' 라고 응답한 사람을 1, 그렇지 않으면 0이라 하면, $\sum_{i=1}^{n} y_i$는 $B(n,\phi)$ 인 이항분포를 따른다. 따라서 '예'라고 응답한 확률 λ의 추정치 $\hat{\phi} = \sum_{i=1}^{n} y_i/n$이므로 모비율 π의 최우추정량 $\hat{\pi}$는 다음과 같다.

$$\hat{\pi} = \frac{\sum_{i=1}^{n} y_i}{n(2P-1)} + \frac{P-1}{2P-1}, \; P \neq 1/2. \tag{2.1}$$

추정량 $\hat{\pi}$의 분산은 다음과 같다.

$$V(\hat{\pi}) = \frac{\pi(1-\pi)}{n} + \frac{P(1-P)}{n(2P-1)^2}. \tag{2.2}$$

2.2. Warner의 확률화응답 모형에 대한 보정 추정

크기가 N인 모집단으로부터 단위 i가 표본에 포함될 확률은 v_i라 하고, 단위 i와 j가 동시에 표본에 포함될 확률을 v_{ij}라 하고, 관심변수를 y_i, 보조변수를 x_i라 하자. 표본설계 s하에서 단위 $i \in s$에 대해 (y_i, x_i)를 관찰하게 된다. 한편 보조변수 x_i의 모집단 평균 μ_x는 알려져 있다고 가정한다.

그러면 식(2.1)에 대해 기지의 보조정보를 이용하여 기존의 설계 가중치 $d_i = 1/v_i$를 조정하는 것이다. 이때 새로운 가중치를 w_i라 하면, 다음과 같은 보정 방정식을 만족한다.

$$\sum_s w_i x_i = \tau_x \tag{2.3}$$

새로운 보정 가중치 w_i는 기존의 설계가중치 d_i와 기대 거리함수 $E\left[\sum_s G(w_i, d_i)\right]$를 최소로 하며, 카이제곱 거리함수에 대해 새로운 가중치는 다음과 같다.

$$w_i = d_i(1 + x_i'\lambda) \tag{2.4}$$

여기서 라그랑지 승수 $\lambda = T_s^{-1}(\tau_x - \hat{\tau}_{xHT})$ 이며, $T_s = \sum_s d_k x_k x_k'$ 이다.

결과적으로 새로운 가중치 w_i를 적용한 관심변수 y에 대한 모집단 총합 τ_y의 보정 추정량은 식(2.5)와 같다.

$$\hat{\tau}_{ycal} = \sum_s w_i y_i = \hat{\tau}_{yHT} + (\tau_x - \hat{\tau}_{xHT})' \widehat{B}_s \tag{2.5}$$

식(2.5)로 부터 $\hat{\tau}_{xHT}=\sum_s d_k x_k$는 보조변수 x의 모집단 총합에 대한 HT추정량이이며, $\hat{B}_s= T_s^{-1}\sum_s d_k x_k y_k$로 다중회귀계수의 가중최소제곱 추정량이다.

그러면 식(2.1)의 확률화 응답 추정량의 보정 추정식은 다음과 같이 도출할 수 있다.

$$\hat{\pi}_{cal}=\frac{\hat{\tau}_{yHT}+(\tau_x-\hat{\tau}_{xHT})\hat{B}_s}{n(2P-1)}+\frac{P-1}{2P-1},\ P\neq 1/2 \tag{2.6}$$

3. 확률화 응답 추정량의 LASSO 보정추정

3.1 LASSO 회귀 방법

다공선성(Multi-colinearity) 문제가 발생하는 고차원 자료에서는 예측결과가 불안정하고 분산이 크기 때문에 오차가 크게 발생한다. 이를 해결하기 위해 Hoerl and Kennard(1970)는 다공선성 문제에서 선형회귀계수의 크기를 축소시켜 모형의 편향을 허용함으로써 예측 정밀도(prediction precision)를 향상시킬 수 있는 ridge 회귀를 제안하였고, ridge 회귀계수 값이 0이 될 수 없기 때문에 변수선택이 불가능하다는 단점을 극복하기 위해 Tibshirani(1996)에 의해 제안된 LASSO 회귀에 대한 목적식은 다음과 같다.

$$\hat{\beta}_{LASSO}=\arg\min\{\sum_{i=1}^{n}(y_i-\beta_0-\sum_{j=1}^{p}x_{ij}\beta_j)^2+\delta\sum_{j-1}^{p}|\beta|\}$$

3.2 모형보조 LASSO 보정 추정량

분석표본 s_A와 벤치마크 표본 s_B를 고려하고, 분석표본에 대해 크기 n_A로서 추출설계 A에 의해 분균등 확률 추출하며, 이때 단위 i의 추출확률은 v_i^A, 단위i와 단위j의 동시추출확률을 v_{ij}^A라 하면, 단위 i의 설계가중치는 $d_i^A=1/v_i^A$이며, 설계가중치 벡터를 d^A, 설계가중 행렬 D^A, 보조변수 행렬은 X^A로 정의한다. 한편 벤치마크 표본의 크기는 n_B, 표본설계 B 하에서 불균등확률추출을 가정하고, 이때 단우 i의 추출확률은 v_i^B, 동시추출확률은 v_{ij}^B, 단위 i의 설계가중치 d^B, 설계 가중행렬 D^B, 보조변수 X^B로 정의한다. 이와 더불어 초모집단 모형(super population model)을 다음과 같이 정의하고, 모집단에서 보조변수 X의 범위는 분석표본과 벤치마크표본 둘 다에서 0이 아닌 확률로 관찰된다.

$$E_\xi(y_i|x_i)=\mu(x_i,\beta),\ \ V_\xi(y_i|x_i)=k_i^2\sigma^2$$

그러면 모형보조 LASSO 보정추정량은 LASSO 회귀로부터 회귀계수 $\hat{\beta}$를 추정하고, 다음으로 분석표본과 벤치마크 표본으로부터 각각 $\hat{\mu}_i=\mu(x_i^A,\hat{\beta})$, $\hat{\mu}_i=\mu(x_i^B,\hat{\beta})$를 구한다. 다음 단계에서는 카이제곱거리 함수를 이용하여 $T^M=(\sum_{s_B}d_i^B,\sum_{s_b}d_i^B\hat{\mu}_i)$와 $M=(1^A,\hat{\mu}_i)$응 구하여, 이로부터 모형보조 가중치를 다음과 같이 구한다.

$$w^L=d_A+D^AM(M'D^AM)^{-1}(T^M-d^{A\prime}M)'$$

최종적으로 관심변수 y의 모집단 총합에 대한 LASSO 보정추정량은 다음과 같다.

$$\hat{\tau}_y^L = (w^L)'y = d^{A'}y + \left(\sum_{s_B} d_i^B \hat{\mu}_i - \sum_{s_A} d_i^A \hat{\mu}_i\right)\hat{B}^M$$

3.2 Warrner의 확률황 응답 모형에 대한 LASSO 보정 추정

Warner의 확률화 응답모형을 적용한 민감한 속성에 대한 모집단 비율 π에 대해 LASSO 모형 보정 추정량은 다음과 같이 도출할 수 있다.

$$\hat{\pi}^L = \frac{\hat{\tau}_{yHT} + (\sum_{i=1}^{n_B} d_i^B \hat{\mu}_i - \sum_{i=1}^{n_A} d_i^A \hat{\mu}_i)\hat{B}^M}{n(2P-1)} + \frac{P-1}{2P-1}.$$

4. 결론

본 논문에서는 민감한 속성에 대한 조사에서 개인이 사생활을 보호할 수 있는 간접응답 방법인 확률화응답 모형의 LASSO 보정추정과정에 대해 제안하였다.

Reference

Deville, J. C. and Sarndal, C. E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87: 376 – 382

Elliott, M.R. (2009). Combining data from probability and nonprobability samples using pseudo-weights. *Survey Practice,* 2(6).

Elliott, M.R., and Valliant, R. (2017). Inference for non-probability samples. *Statistical Science,* 32, 249- 264.

Hoerl, A. and Kennard, R. (1970) Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics, 12, 55-67.

Singh, A.C., and Mohl, C.A. (1996). Understanding calibration estimators in survey sampling. *Survey Methodology*, 22: 107 – 115.

Tibshirani, R. (1996), Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society, Series B (Methodological),* 267- 288.

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 26-27, 2023, pp. 13-16

은닉 마르코프 모델을 이용한 적응형 학습경로 생성에 관한 연구

최현희[1], 이윤지[2], 이하윤[3]

요 약

전 세계적인 COVID19의 영향으로, 비대면 학습에 대한 활용도가 높아지고 있다. 비대면 학습을 위해서는 효율성을 토대로 한 학습경로의 생성이 중요하다. 본 연구는 학생들이 수업에서 마주칠 수 있는 경험의 경로, 즉, knowledge concept(*KC*) 을 순서대로 배열한 연결망으로 알려진 학습경로를 만드는 방법으로 hidden Markov model(HMM)을 소개하고, HMM을 활용하여 경로 예측의 정확성을 향상하는 것을 목적으로 한다. 정확한 학습경로의 생성을 위하여 HMM 수행 전 Elastic Net, LASSO, random forest를 이용한 변수 선택(variable selection) 기법을 활용한다. 마지막으로, HMM을 이용해서 *KC*들의 그룹으로 만들어진 상위 개념을 위주로 만들어진 경로와 다른 후보 경로들과의 비교를 통하여 *KC*들의 선후관계가 HMM을 활용하여 제대로 추출되었음을 입증한다. 실험을 위하여 AI-hub(https://aihub.or.kr/)에서 공유한 수학 학습데이터를 이용한 결과 Elastic Net 등의 변수 선택 방법으로 관계 쌍을 추려냈을 때의 HMM의 성능이 월등히 좋아졌고, 상위 개념으로 이루어진 학습경로를 평가할 때 HMM을 활용한 경우가 다른 모형에 비해 모형정확도 면에서 매우 훌륭하였다.

주요용어 : hidden Markov model(HMM), 학습경로, 변수 선택(variable selection), knowledge concept(*KC*).

1. 서론

학습경로 관련한 연구들은 주어진 커리큘럼에 그대로 의존하기보다는 확률 개념을 도입하여 입력 데이터에 기반한 모형화를 시도한다. 참고로, Mandalapu, Gong, Chen(2021)은 변수 선택에 있어 기존의 변수 선택 방법과 심층 모델을 이용한 변수 선택 결과를 비교할 때, 기존의 변수 선택 방법을 활용하는 것이 더 훌륭한 결과를 가져옴을 언급하였다.

본 논문에서는 HMM을 소개한 후, 학생들의 수학 문제 풀이 데이터에 HMM을 적용한다. 변수 선택 기법을 이용하여 *KC*의 후보를 먼저 골라낸 후 HMM을 적용하였을 때, HMM만을 적용한 경우보다 정확도가 약 20% 향상된 것을 확인하였다. 생성된 학습경로의 정확성 비교의 견지에서 *KC*들을 상위 개념의 콘셉트라는 이름으로 나누었을 때, HMM에 기반하여 생성한 모델은 전체 가능한 학습경로에서 상위 5%이내의 정확도를 보였다.

[1](교신저자) 06633 서초대로54길 29-30, 라이브데이터(주) 수석연구원. Email: choihh@laivdata.com
[2]06974 흑석로 84, 중앙대학교 응용통계학과 학부재학생. E-mail: pioneeryjlee@laivdata.com
[3]06974 흑석로 84, 중앙대학교 소프트웨어학부 학부재학생. E-mail: hayun0406@laivdata.com

2. HMM

HMM은 은닉된 상태(hidden state)와 관측상태(observable state)가 동시에 존재하는 시계열 데이터에 적용하는 모형으로, 데이터의 모수인 초기확률(initial probability), 전이확률(transition probability), 방출확률(emission probability)을 예측하거나, 관측값으로부터 은닉상태의 값을 예측하는 것, 또는 모수 확률값들이 주어졌을 때 관측 시계열의 확률을 구하는 3가지 문제에 적용된다(Rabinar, 1989).

HMM의 이해를 도우려고 문제 풀이 예를 설명하고자 한다. t=1에서의 학습이 시작되기 전 이해 확률(모름(unlearned)/습득(learned))을 초기확률이라고 하고, 시간이 가면서 생성되는 모름→모름, 모름→습득, 습득→모름, 습득→습득과 같은 은닉상태 변경 확률을 전이확률이라고 한다. 이해 정도에 따라 시험에서 정답/오답의 답안(observable state)이 달리 나타날 수 있는데, 각 답안이 관찰될 수 있는 확률을 방출확률이라고 한다.

$\theta=(\pi,A,B)$의 예측값은 EM 알고리즘(Expectation Maximization algorithm)을 수행하여 얻는다. Expectation 단계에서는 입력된 θ값을 기준으로 은닉상태에 대한 분포를 얻고, maximization 단계에서는 E 단계에서 얻은 은닉상태 분포를 최대화하는 갱신된 θ값을 얻는다. 즉. θ의 정의와 더불어 EM의 과정을 포함한 모형을 HMM이라 한다(Zucchini, MacDonald, Langrock, 2016).

EM 알고리즘을 시행하면서 국부 최적값(local optimum)에 빠지지 않고 $\theta=(\pi,A,B)$를 구하려면 초깃값이 필요하다. 초깃값을 결정하는 방법에는 데이터의 특성을 반영하는 것, 임의의 값으로 설정하는 것, 추측에 의한 것으로 3가지가 있다.

3. 자료 및 실험

3.1. 자료

3.2. 실험

본 연구의 목적은 Figure 1에서 확인하듯이 공개된 AI hub 데이터를 이용하여 HMM을 이용한 학습경로를 만들고, HMM을 사용하여 생성한 학습경로의 평가 방법을 생성하는 것이다.

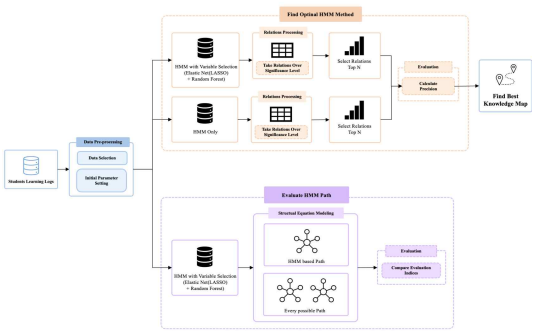


Figure 1. Research process

4. 실험 결과

1) 초깃값 계산

HMM을 수행하여 모수 θ를 추정하려면 θ의 초깃값을 설정해야 한다. θ에 대한 초깃값이 의미 없이 설정될 경우, 수행시간이 느려질 뿐만 아니라 모형의 정확성이 뛰어나지 않을 수 있다(Sá et al., 2008). 그러므로 HMM을 이용하기 전 KC 쌍(i번째 KC 와 j번째 KC, $KC(i)$와 $KC(j)$)에 대해 Table 2와 같이 초깃값 $\theta=(\pi,A,B)$를 설정하였다.

Table 1. Starting values of initial parameters

	π	A	B
계산식	$\left[\frac{count(0)}{n}, \frac{count(1)}{n}\right]$	$\begin{bmatrix} \frac{count(0\rightarrow 0)}{n} & \frac{count(0\rightarrow 1)}{n} \\ \frac{count(1\rightarrow 0)}{n} & \frac{count(1\rightarrow 1)}{n} \end{bmatrix}$	$\begin{bmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{bmatrix}$
초깃값			

*n: 학생 수

2) Elastic Net(LASSO)+ RF 여부에 따른 정확도 비교

Table 2. HMM with candidates .vs. HMM only

	HMM만 수행한 경우 (비교모형과 동일한 수 추출, top N)	Elastic Net(LASSO)+RF 후보를 뽑은 후 HMM을 수행한 경우
Precision		

3) 지식맵의 생성

4) HMM을 이용하여 생성한 경로의 정확도

5. 결론

본 논문에서는 기계 학습을 위하여 공개된 AI-hub 데이터를 이용하여 HMM을 수행하고, 이를 이용하여 학습경로를 생성하는 작업을 하였다.

전이확률에 따른 상위 n개(Top n), Elastic Net(LASSO)+RF를 후보로 하여 HMM의 성능을 확인한 결과 Elastic Net(LASSO)로 후보 관계를 추출한 후 HMM을 적용하였을 때 더욱 높은 정확도를 보였다. 이를 통하여 KC간 선후관계를 확인하기 위해서는 통계적으로 관련이 높을 것으로 판단되는 KC관계 쌍을 먼저 선택한 후 방향성을 확인하는 것이 필요하다는 결론을 얻을 수 있다.

또한, KC그룹으로 이루어진 상위 개념에 대한 학습경로를 만드는 경우, 제안한 방법론에 따르면, 가능한 모든 경로와 HMM에 기반에 기반한 경로를 비교했을 때, HMM에 기반한 경로가 적합도 면에서 가능한 모든 경로에 대해 상위 5% 이내의 정확도를 따르는 것을 확인하였다. 즉, HMM으로 문제 풀이 데이터에 대한 선후관계를 판단한 경우가 다른 어떤 모형에 기반하여 선후관계를 파악하였을 때에 비해 높은 신뢰도를 가지는 것을 확인한 것이다.

References

Mandalapu, V., Gong, J., & Chen, L. (2021). Do we need to go deep? knowledge tracing with big data. arXiv preprint arXiv:2101.08349.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2), 257-286.

Zucchini, W., MacDonald, I. L., Langrock, R. (2016), Hidden Markov Models for Time Series: An Introduction Using R, Second Edition. Chapman & Hall/CRC

Proceedings of the Korean Data Analysis Society
January 26-27, 2023, pp. 17-20

사회연결망분석을 활용한 코로나와 교육에 관한 최근 연구동향 분석: KCI 2020-2022 학술논문의 NodeXL 분석 및 시각화*

임수현[1], 김수경[2], 김지선[3], 손현규[4], 양정호[5]

요 약

코로나 발생이후 교육분야에서 상당한 영향을 미치고 있어, 본 연구에서는 학술지에 게재된 학술논문을 중심으로 키워드 분석하였다. 코로나와 교육을 키워드로 해서 추출된 KCI 자료를 사회연결망 분석방법을 이용하여 분석한 결과, 지난 3년간 코로나와 교육관련 연구가 증가하였으며, 다양한 키워드를 중심으로 관련 학술논문이 게재된 것으로 나타났다. 코로나 발생이후 3년이란 짧은 기간에도 불구하고, 코로나와 교육의 관련성을 살펴본 논문이 급증한 것은 코로나가 사회 전반에 미친 영향이 상당한 것을 알 수 있다. 향후 사회연결망분석을 활용해서 코로나와 교육의 키워드 관계, 각 학문분야마다 특징적인 키워드 관계, 저자와의 관계 등을 추가로 분석할 필요가 있다.

주요용어 : 코로나, 비대면교육, 사회연결망분석, 연구동향.

1. 서론

학문연구는 사회변화를 일정부분 반영할 수밖에 없는 것이 일반적인 추세이다. 특히 코로나 시작 후 연구자들의 관심도 코로나에 집중되는 경향을 보이고 있다. 교육분야에서도 변화에 민감하기 때문에 코로나와 교육과 관련된 연구동향과 연구추이를 살펴볼 필요가 있다. 우리나라 교육의 역사에서 전례 없는 학교폐쇄, 비대면전면수업 진행 등 교육변화 따른 연구를 세밀히 분석하는 것이 중요해보여, 본 연구에서는 우리나라의 핵심 학술지 색인인 KCI 통해 코로나와 교육 연구동향 분석하려고 한다.

KCI를 통해 2020년과 2021년의 연구 주요 키워드를 비교해 보면, 코로나가 한국에 전파되기 시작하면서 지난 3년 동안 전체 학술지 게재된 연구물에서 COVID-19이 가장 많은 비중을 차지한 것으로 나타났다. 이 중에서 교육분야에 한정해서 살펴보면, 코로나가 상위권 연구주제로 나타났으며 이를 통해 코로나는 전체 사회 뿐만 아니라 교육현장에서도 중요한 이슈로 자리매김하고 있음을 알 수 있다.

[1]03063 서울시 종로구 성균관로 25-2, 성균관대학교 교육학과 박사과정. E-mail: shyunr@skku.edu
[2]03063 서울시 종로구 성균관로 25-2, 성균관대학교 교육학과 학사과정. E-mail: rlatnrd2002@gmail.com
[3]03063 서울시 종로구 성균관로 25-2, 성균관대학교 교육학과 학사과정. E-mail: kjsun031030@gmail.com
[4]03063 서울시 종로구 성균관로 25-2, 성균관대학교 교육학과 학사과정. E-mail: sdh10077@naver.com
[5](교신저자) 03063 서울시 종로구 성균관로 25-2, 성균관대학교 교육학과 교수. E-mail: jyang@skku.edu

2020		2021	
전체	교육	전체	교육
COVID-19	대학생	COVID-19	대학생
인공지능	연구 동향	Depression	코로나19
Depression	코로나19	인공지능	연구동향
대학생	예비유아교사	대학생	유아교사
China	초등학생	우울	초등학생
우울	유아교사	자기효능감	자아존중감
Vietnam	유아	딥 러닝	유아
Deep Learnig	자아존중감	Deep Learnig	질적연구
빅데이터	간호대학생	노인	예비유아교사
Korea	질적 연구	빅데이터	매개효과

Figure 1. 연구 주요 키워드 추이 비교

코로나로 인해 한국의 교육현장에서는 혁명적 변화와 혼란이 발생하였는데 지금까지 전쟁속에서도 문을 한번도 닫지 않았던 학교가 갑자기 멈춘 첫 사례였으며 교실에서 주로 수업이 이루어지던 교육현장에 예상치 못했던 온라인 수업의 확대에 따라 교사와 학생 뿐만 아니라 학부모 또한 교육 현장에 대한 우려가 넘쳐났다. 즉 온라인으로 진행되는 수업을 통해 학생들이 제대로 학습할 수 있는 여건이 마련된 것인지에 관한 고민이 많이 나타났다고 볼 수 있다.

다른 학문분야에서도 코로나 관련 연구가 늘어났으며, 국가별 연구주제 주력도 순위를 봤을 때 한국은 13위에 자리 잡고 있으며 의약·바이오 면에서는 합병증, 사회·경제활동 부문에서는 경제와 심리적 영향에 관련해 많은 연구가 이뤄지고 있음을 알 수 있다. 이 자료들을 바탕으로 코로나와 사회 각 분야, 의학분야 등 모든 분야에서 연구주제로 코로나가 등장했으며 2020년 코로나 발생후, 관련 연구분야 동향을 고급 통계방법을 활용해 파악하는 것이 더욱 중요한 시점이 되었다고 생각한다.

2. 연구방법

2.1 연구자료 수집 및 데이터 전처리 과정

본 연구는 KCI 한국 학술지 인용 색인(https://www.kci.go.kr)에서 1단계로 2020년부터 2022년까지의 논문 중 코로나와 covid를 키워드로 한 논문을 추출하였다. 2단계로 같은 기간 동안 출판된 논문 중 교육, 수업, 학교, 학습, 러닝, 강의, 교실, 학생, 교사, 학부모, 강사, 교수, 교원 총 13개의 키워드를 포함하는 논문에 대한 자료를 수집하였다. 총 35,000개 이상의 논문 자료가 검색되었고, 교육 외에도 보건, 경영 등의 다양한 분야에서 논문이 2020년, 2021년 그리고 2022년 11월 20일까지 출판되었음이 나타났다. 전체 논문 표본 중 중복되는 것과 저자명이 다른 경우 등을 제외하여 최종 14,567개의 코로나 관련 논문을 수집하였다.

2.2 자료 분석 방법

연도별 논문 추이부터 저자별, 학술지별 논문 추이를 분석하였다. 저자별, 학술지별 논문 추이의 경우 출판물 개수가 많은 순으로 정렬한 뒤 상위 20개의 저자 및 학술지를 대상으로 수치를 비교분석했다. 또한, 학문분야에 따라 논문을 분류하고 이를 연도별로 세부 분류하여 각 연도에 특정 분야에서 코로나 관련 연구 및 논문이 얼마나 이루어졌는지 분석하였다.

워드클라우드를 활용하여 대상 논문들의 키워드 노출 빈도수를 가시적으로 확인하였고 키워드 분석을 위해 NodeXL 사회연결망 분석(SNA, Social Network Analysis)을 진행하였다. 이를 통해 중심 키워드 코로나와 다른 키워드들 간의 상호연결 관계를 확인하였다.

3. 코로나와 교육 키워드 사회연결망 분석결과

3.1 코로나 연구논문 추이

코로나19와 관련하여 작성된 연구논문의 연도별 추이는 국내에서 본격적으로 코로나가 전파되기 시작한 2020년에는 2,838편의 연구논문이 작성된 것으로 나타났다. 그리고 코로나19의 확산이 본격적으로 이루어진 시기인 2021년과 2022년에는 각각 6,559편, 5,170편으로 코로나 초기인 2020년에 비해 연구논문의 수가 2~3배 가량 증가하였다.

이 시기에 200여개의 코로나 관련 논문 게재 학술지를 통해 연구자 진행되었으며, 가장 많은 논문이 실린 학술지 1위는 Journal of Korean Medical Science로 나타났으며, 이 학술지는 314편의 논물을 게재하였다. 2위는 284편을 게재한 학습자 중심 교과교육연구로, 교육학 분야에서도 코로나 연구가 활발히 이루어졌다. 또한, 2020년에는 100편 미만이었던 코로나 논문이 2021년에는 약 250편, 2022년에는 약 180편으로 급증한 추이를 보이고 있다.

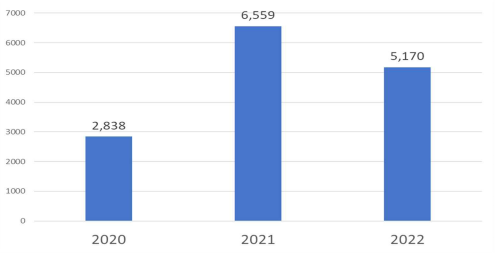


Figure 3. 코로나 최다 논문게재 학술지별 추이

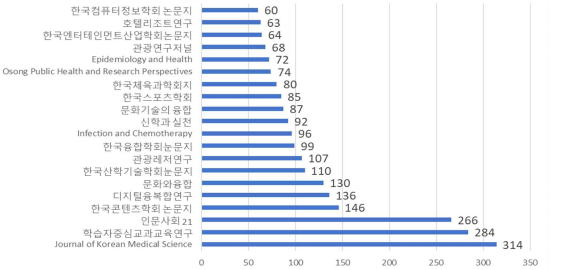


Figure 2. 코로나 논문 연도별 추이

3.2 코로나와 교육 논문 키워드 SNA 분석

코로나와 교육을 통해 추출된 학술논문을 키워드 중심으로 살펴보면, 전체 키워드 중에서 코로나19나 COVID-19처럼 코로나 관련된 키워드 빈도가 압도적으로 높게 나타났다. 온라인 수업, 비대면, 원격수업처럼 교육관련 관련된 키워드도 전체 중에서 10위안에 드는 것으로 나타났다.

코로나 논문을 종합하여 분석하여 키워드 사이의 관계를 노드와 링크로 표현하면, '코로나'를 중심으로 다양한 키워드들이 연결망을 구성하고 있었다. 키워드 중심으로 사회연결망 분석을 한 결과는 노드 연결정도(Degree) 평균은 5.185, 네트워크 밀도(Density)는 0.0002, 노드간 평균 거리(Distance)는 4.240으로 코로나와 교육의 키워드간에 긴밀한 연결관계를 보이고 있는 것을 알 수 있었다.

Table 1. 코로나 관련 상위 30개 핵심 키워드 분석결과

no.	키워드	N	%
전체	전체	147865	100.000
1	코로나19	9150	6.188
2	COVID-19	8293	5.608
3	팬데믹	1018	0.688
4	SARS-CoV-2	986	0.667
5	포스트 코로나	643	0.435
6	Coronavirus	628	0.425
7	메타버스	447	0.302
8	감염병	437	0.296
9	코로나19 팬데믹	424	0.287
10	온라인 수업	362	0.245
11	비대면	361	0.244
12	원격수업	323	0.218
13	대학생	315	0.213
14	빅데이터	298	0.202
15	비대면 수업	280	0.189

no.	키워드	N	%
전체	전체	147865	100.000
16	인공지능	254	0.172
17	간호대학생	211	0.143
18	우울	200	0.135
19	뉴노멀	199	0.135
20	전염병	187	0.126
21	사회적 거리두기	176	0.119
22	언택트	162	0.110
23	4차 산업혁명	156	0.106
24	노인	152	0.103
25	가상현실	142	0.096
26	코로나 팬데믹	140	0.095
27	스트레스	138	0.093
28	재택근무	137	0.093
29	불안	135	0.091
30	재난	132	0.089

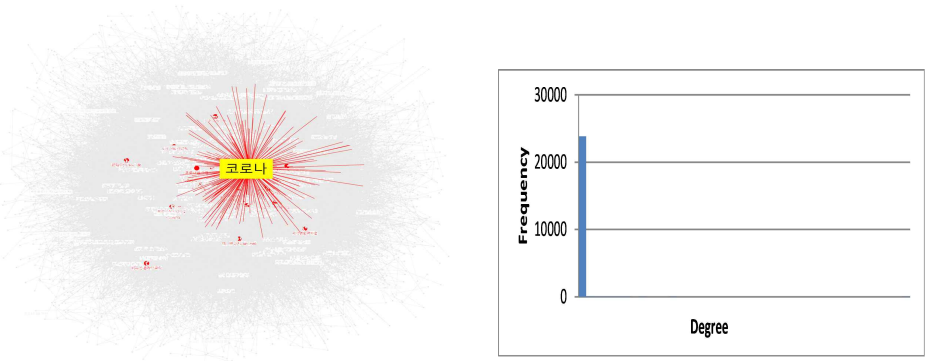


Figure 4. 코로나 논문 키워드 사회연결망 분석

References

한진희, 한가영. (2021). 텍스트 마이닝을 활용한 코로나19 이후 온라인수업 관련 연구 동향 분석. 영어영문학연구, 47(3), 171-194.

Cho, J. (2020). Analysis of open access status of domestic author's papers published in international journals: Based on highly cited papers. *Journal of the Korean Society for Library and Information Science, 54*(1), 325-341.

Kim, J. Y., Kim, H. S., Shim, W. S. (2020). A Study on open peer review perception of Korean authors in a mega OA journal. *Journal of the Korean Society for Information Management, 37*(4), 131-150.

Proceedings of the Korean Data Analysis Society
January 26-27, 2023, pp. 21-22

중국 데이터 조사 현황 및 개선방안

뢰에[1], 진선순[2], 김화[3], 소성[4], 손정옥[5], 양정호[6]

요 약

본 연구는 중국에서 수집되어 공개된 데이터 현황을 조사하기 위해 자료 수집법을 이용하여 중국 데이터의 구축 현황 및 특징을 분석하였다. 주요 연구 결과는 첫째, 중국 데이터 접근하는 방법은 크게 3가지가 있다. 중국 정부 공공 데이터, 중국 대학별 개방형 연구 데이터 플랫폼, 그리고 민간 영리성 데이터 분석 벤처 기업을 통해 중국 데이터 구축 현황 파악이 가능하다. 2015년 8월 중국 국무원은 <빅데이터 발전 촉진 행동요강>를 발표하였다. 이에 따라 중국 지역별 정부 공공 데이터 오픈하였다. 상하이는 2012년 상반기에 첫 번째 중국 정부 데이터 공개 플랫폼을 출범한 이후 중국 정부는 50개 이상의 중앙 부처, 위원회, 지자체 차원의 플랫폼을 연속적으로 출범시켰다. 이를 통해서 각 지역별 법적, 정치적, 기술적, 교육적, 사회적, 경제적 등 다양한 공공 데이터를 정부 공공 데이터 플랫폼을 통해서 볼 수 있다. 뿐만 아니라, 중국 대학은 개방형 연구 데이터 플랫폼을 구축하기 시작하였다. 본 연구 자료를 수집하는 동안 2022년 QS대학 순위 중 중국 전 50개 대학의 개방형 연구 데이터 플랫폼을 살펴본 결과, 북경대학교, 복단대학교, 인민대학교의 연구 데이터 플랫폼 이용자수가 제일 많고 데이터 종류 다양한 것으로 확인할 수 있다. 중국 데이터 접근할 수 있는 3번째 방법은 민간 영리성 데이터 분석 벤처 기업이다. 이를 통해서 원하는 데이터 구성 설계, 수집 및 분석 등 다양한 데이터 분석 작업을 수행할 수 있다. 둘째, 각 접근 방식에 따라 데이터의 현황 및 특징이 다르다. 본 연구 논문은 개방형 데이터 평가 지표의 4가지 측면인 데이터의 구축배경 및 이용환경, 데이터의 질적 측면, 데이터의 사용, 데이터의 영향을 통해서 중국 데이터를 분석하였다. 우선, 중국 정부 공공 데이터의 특징이 다음과 같다. 중국 정부 공공 데이터는 정부의 주도하에 공개된 데이터이기 때문에 넓은 범위를 가지고 있지만, 연구자가 쉽게 다운로드하거나 접근이 쉽지 않다. 데이터의 질은 상대적으로 기초적인 자료이거나 대부분 일반 기술통계 데이터가 제일 많이 차지하고 있다. 오픈된 데이터는 업데이트 주기가 대부분 년 단위로 이루어지고 있다. 오래 동안 업데이트 안 하는 경우도 종종 발생하고 있다. 중국 대학의 개방형 연구 데이터 플랫폼은 각 대학학술정보관 주도하기 때문에 해당 대학 주요 연구 분야의 연구 데이터가 가장 많이 구축되어 있다. 연구 분야별 데이터 다양하지만 조사 시기를 잘 살펴보면 90년대 초반 데이터 제일 많이 공개하였다. 최근 5년간 조사된 데이터를 공개한 경우가 거의 없는 것으로 확인할 수 있다. 바로 다운로드해서 사용할 수 있는 데이터는 많지 않는 반면에 대부분 데이터는 연구자가 직접 이메일 또는 사이트를 통해 신청해야 사용 가능하다. 민간 영리성 데이터 분석 벤처 기업을 통해서 접근할 수 있는 개방형 데이터는

[1]03063 서울시 종로구 성균관로 25-2, 성균관대학교 교육학과 박사수료. E-mail: leirui5191@naver.com
[2]03063 서울시 종로구 성균관로 25-2, 성균관대학교 교육학과 박사수료. E-mail: jinxianshun602@126.com
[3]03063 서울시 종로구 성균관로 25-2, 성균관대학교 교육학과 석사수료. E-mail: jinhua9811@163.com
[4]03063 서울시 종로구 성균관로 25-2, 성균관대학교 교육학과 석사과정. E-mail: sucheng960716@gmail.com
[5]03063 서울시 종로구 성균관로 25-2, 성균관대학교 교육학과 석사과정. E-mail: sunjingyu1999@naver.com
[6](교신저자) 03063 서울시 종로구 성균관로 25-2, 성균관대학교 교육학과 교수. E-mail: jyang@skku.edu

대부분 정부 공공 데이터 및 대학 개방형 연구 데이터 플랫폼에도 다운로드 할 수 있는 데이터다. 이를 통해서 학자가 개인적으로 얻을 수 있는 데이터 많지 않으며, 민간 영리목적 데이터 분석 벤처 기업은 주로 연구소 혹은 기업과 협력해서 데이터 분석하기 때문에 자신들이 가지고 있는 데이터 공개하지 않는 경향을 보이고 있다. 대부분 통계 결과물만 사이트를 통해서 살펴볼 수 있다. 셋째, 중국 데이터 구축 및 이용 관련 전체적인 문제점은 다음과 같다. 공개된 데이터가 오래되기 때문에 연구 실용성이 떨어지고 있다. 학자로서 데이터의 접근성 높지 않고 쉽게 이용하지 못하는 부분도 나타난다. 중국 내에 개방형 데이터에 관련 법적 근거도 부족하며, 데이터를 체계적으로 관리하거나 주기적인 업데이트가 필요해 보인다. 끝으로 본 연구 결과에 기초해서 중국 데이터를 이용하여 후속연구를 위한 시사점을 다음 같다. 우선, 중국 데이터를 이용해서 더 많은 연구가 이루어지려면, 법적 근거의 마련이 필요하다. 법을 통해서 데이터를 보호하고 관리하게 되면 공개할 수 있는 데이터를 많이 증가할 것으로 생각한다. 또 지속적인 연구를 수행하기 위해 이미 공개된 데이터는 주기적 업데이트 필요하다. 기술통계의 결과만 공개하는 것뿐만 아니라 데이터 수집의 설계, 구성 등 내용도 동시에 공개하게 되면 더 활발한 연구 활동을 이루어질 수 있다. 불필요한 데이터 자원의 낭비를 피하기 위해 비공개 된 조사 데이터의 조사 내용만 알려지기만 해도 학자한데 큰 도움이 될 것이다. 중국 데이터에 관련 된 연구는 중국 내에서 많은 관심을 받고 있으며 최근에 관련 연구 결과물이 늘어나고 있다. 특히 정부 공공 데이터 중심으로 이루어진 연구물 최근 5년간 많이 증가했다. 앞으로 중국에서도 미래연구의 발전 중시하며 데이터에 관련 대응법 등 더 많은 요구들이 늘어날 것으로 예상된다.

주요용어 : 중국 데이터, 중국 데이터 현황, 데이터 분석, 중국 데이터 특징.

Proceedings of the Korean Data Analysis Society
January 26-27, 2023, pp. 23-26

교원의 역량개발을 위한 비형식 학습 참여 효과에 관한 연구 : PIAAC의 한국, 미국, 핀란드 데이터 비교

최정윤[1], 양정호[2]

요 약

본 연구에서는 OECD 국제성인역량조사(The Program for International Assessment of Adult Competencies, PIAAC) 2012년 1차 조사의 한국, 미국, 핀란드 데이터를 이용하여 한국 교원의 역량을 국제적인 수준에서 비교하고, 교원의 비형식 학습의 참여가 교원의 역량에 미치는 영향을 분석하고자 한다. 19세~65세 성인 중 산업별 직업분류 문항에서 '교육/Education'을 선택한 응답자를 연구 대상자로 선정한 후 가중치를 적용하였다. 교원의 역량인 언어능력 점수를 분석하기 위해 기술통계, t검정, F검정을 하였으며, 비형식 학습 참여의 영향 효과를 분석하기 위해 다중회귀분석을 하였다. 분석 결과, 비형식 학습 참여 여부의 영향 효과는 학습 종류에 따라 국가마다 다른 결과를 나타냈으나, 비형식 학습 참여 수는 정(+)적인 효과를 나타내 비형식 학습 참여 수가 증가할수록 언어능력 점수가 높아지는 것으로 나타났다.

주요용어 : 교원역량, 비형식학습, PIAAC, 국제비교, 회귀분석.

1. 서론

1.1. 연구의 배경 및 목적

최근 OECD의 국제성인역량조사(The Program for International Assessment of Adult Competencies, 이하 PIAAC)의 결과를 보면, 우리나라 교원의 능력은 OECD 평균을 밑도는 수준으로 나타났다. 한국 데이터 분석 결과를 보면 교원의 언어능력은 전체 27개 국가 중 12위, 수리력은 14위이며 컴퓨터 기반 문제해결력은 15위로 나타났다. 전반적으로 우리나라 교원의 국제적 역량은 조사 국가 중 평균에서 평균 이하에 머무르는 것으로 나타났다. 반면, 만 15세 학생을 대상으로 한 OECD의 PISA 2018년 조사 결과에서 우리나라 학생의 국제성취도가 읽기 2-7위, 수학 1-4위인 점을 고려한다면 우리나라 교원의 능력은 학생의 국제적 성취에 비해 뒤처진다는 것을 알 수 있다(양정호, 2016; 한국교육과정평가원, 2019).

교원의 역량 및 전문성 증진이 중요함에도 교원의 역량 개발을 위한 학습 참여의 효과를 분석하고 우리나라 교원의 역량 수준을 국제적인 수준에서 파악할 수 있는 연구를 찾아보기 힘든 상황이다. 따라서 기존의 연구가 지닌 한계를 보완하여 교원의 비형식 학습 참여와 그 효과를 국제적

[1](교신저자) 03063 서울특별시 종로구 성균관로 25-2 성균관대학교 교육학과 박사과정 수료.
E-mail: 81jychoi@naver.com
[2]03063 서울특별시 종로구 성균관로 25-2 성균관대학교 교육학과 교수. E-mail: jyang@skku.edu

으로 비교한 연구가 필요하다 하겠다.

본 연구에서는 2012년 PIAAC 1차 자료 중 한국, 미국, 핀란드의 데이터를 사용하여 교원의 역량 개발을 위한 비형식 학습 참여의 효과성을 국제적인 수준에서 비교 분석하고자 하였다.

1.2. 연구 문제

연구 문제 1. 교원의 비형식 학습 참여는 교원의 역량 개발에 효과를 미치는가?

연구 문제 2. 한국, 미국, 핀란드 국가별 교원의 비형식 학습 참여가 교원 역량에 미치는 영향에 차이가 있는가?

2. 분석 방법

2.1. 분석 자료 및 대상

본 연구에서는 2011-2012년 PIACC 1차 조사 참여국 중 한국, 미국, 핀란드의 데이터를 선정하여 각국의 19세~65세 성인 중 산업별 직업분류 문항에서 '교육/Education'을 선택한 응답자를 연구 대상자로 설정하였다. 한국 314명, 미국 341명, 핀란드 308명을 최종 분석 대상으로 선정한 후 가중치를 적용한 최종 사례수는 한국 1601080명, 미국 11148752명, 핀란드 182532명이다.

2.1. 분석 자료 및 대상

자료의 처리와 분석을 위하여 SPSS 28.0 프로그램을 사용하여 독립표본 t검정(T-test), 일원배치 분산분석(ANOVA), 다중회귀분석(Multi-Regression)을 실시하였다. 종속변수인 교원의 언어능력은 각 10개의 Plausible Values(이하 PVs)평균값으로 산정하였다. 교원의 언어능력 점수 분석을 위한 사용 변수 범주를 크게 개인 배경 변수, 직장 배경 변수, 비형식 학습 변수로 구성하였다. 하위 요인에는, 개인 배경 변수로 성별, 연령대, 최종학력, 부모 학력을, 직장 배경 변수로 고용주 학습비용 지원, 고용상태, 현재 상태, 월급을, 비형식 학습 변수로 원격 e러닝, OJT, 세미나 및 워크숍, 강의 및 개인교습, 비형식 학습 참여 수로 선정하였다. 분석대상자의 일반적인 특징은 Table 1과 같다.

3. 분석 결과

교원의 배경 변수가 교원의 언어능력 점수에 미치는 영향을 분석하기 위해 다중회귀분석을 실시한 결과를 Table 2에 제시하였다. 개인 배경과 직장 배경 변수의 결과를 생략한 비형식 학습 배경 변수의 다중회귀분석 결과는 다음과 같다.

비형식 학습 참여 여부 변수를 투입한 결과, 한국은 비형식 학습 참여 여부에 따라 원격 e러닝 3.587, OJT 1.641, 세미나 및 워크숍 1.498, 강의 및 개인교습 .553씩 높아지는 것으로 나타났다. 비형식 학습 참여 개수는 0개일 때 보다 1개씩 늘어날수록 1.796씩 높은 것으로 나타났다. 미국은 비형식 학습 참여 여부에 따라 참여한 경우 원격 e러닝 -3.463, OJT -1.122 차이를 보이며, 세미나 및 워크숍 4.990, 강의 및 개인교습 12.421 높은 것으로 나타났다.

비형식 학습 참여 개수는 0개일 때 보다 1개씩 늘어날수록 1.119씩 높은 것으로 나타났다. 핀란

드는 e러닝 1.437, OJT 6.241, 세미나 및 워크숍 10.878, 강의 및 개인교습 4.084 높은 것으로 나타났다. 비형식 학습 참여 개수는 0개일 때 보다 1개씩 늘어날수록 5.932씩 높은 것으로 나타났다.

Table 1. 비형식 학습 변수의 기술통계치

변수		한국(N=1601080)		미국(N=11148752)		핀란드(N=182532)	
		사례수	비율(%)	사례수	비율(%)	사례수	비율(%)
원격e러닝	비참여	918044	57.3	8464104	75.9	132076	72.4
	참여	683036	42.7	2684648	24.1	50456	27.6
OJT	비참여	833855	52.1	4129491	37.0	53530	29.3
	참여	767225	47.9	7019261	63.0	129002	70.7
세미나 및 워크숍	비참여	837363	52.3	3709527	33.3	106253	58.2
	참여	763717	47.7	7439224	66.7	76279	41.8
강의 및 개인교습	비참여	1153040	72.0	10030317	90.0	141943	77.8
	참여	448040	28.0	1118435	10.0	40589	22.2
비형식 학습 참여수	0개	333664	20.8	2168702	19.5	27980	15.3
	1개	454943	28.4	2274835	20.4	56837	31.1
	2개	371972	23.2	4483698	40.2	61341	33.6
	3개	298875	18.7	2095998	18.8	28687	15.7
	4개	141627	8.8	125519	1.1	7687	4.2

Table 2. 한국, 미국, 핀란드 회귀모형 비교(언어능력)

		한국				미국				핀란드			
		모형1		모형2		모형1		모형2		모형1		모형2	
		B	SE	B	SE	B	SE	B	SE	B	SE	B	SE
상수		281.248***	.118	281.035***	.118	268.392***	.057	268.680***	.057	309.700***	.420	309.668***	.421
원격e러닝 (비참여*)	참여	3.587***	.056			-3.463***	.026			1.437***	.183		
OJT (비참여*)	참여	1.641***	.056			-1.122***	.025			6.241***	.195		
세미나 및 워크샵 (비참여*)	참여	1.498***	.054			4.990***	.026			10.878***	.184		
강의 및 개인교습 (비참여*)	참여	.553***	.059			12.421***	.036			4.084***	.200		
비형식학습 참여 수				1.796***	.022			1.119***	.012			5.932***	.085
F		42972.389***		51636.552***		199587.074***		229711.334***		5417.132***		6438.060***	
R2		.378		.377		.330		.316		.400		.394	
수정된 R2		.378		.377		.330		.316		.400		.394	

*더미변수 생성 시 기준집단(reference group)
*p≤.05, **p≤.01, ***p≤.001

4. 결론

분석 결과, 한국과 미국의 경우 e러닝, OJT, 세미나 및 워크숍, 강의 및 개인교습을 참여했을 때 언어능력 점수가 높아지는 것으로 나타났다. 미국은 세미나 및 워크숍과 강의 및 개인교습을 참여했을 때 언어능력 점수가 높아지는 것으로 나타났으나, 원격 e러닝과 OJT를 참여했을 때 점수가 낮아지는 것으로 나타났다. 비형식 학습 참여 수를 투입한 결과를 보면 한국, 미국, 핀란드 세 나라 모두 비형식 학습 참여 수가 증가할수록 참여 수가 0개일 때 보다 언어능력과 수리력 점수가 증가하는 것으로 나타났다.

본 연구의 결과를 보면, 교원의 비형식 학습 참여는 교원의 언어능력 향상에 긍정적인 영향을 미치는 것으로 나타났다. 또한, 비형식 학습 참여 수가 증가할수록 교원의 역량점수가 높아지는 것으로 나타났다. 이는 교원의 연수, 멘토링 및 현장실습 학습 등 비형식 학습 참여가 교원의 전문성 증진에 긍정적인 역할을 한다고 볼 수 있다. 따라서 교원의 비형식 학습 참여를 이끌어내고 교원이 전문성 개발 활동에 적극적으로 참여할 수 있도록 학교와 정부 차원의 지원이 필요하다 하겠다.

References

양정호 (2016). OECD 교원의 역량 국제비교 분석: PIACC 언어능력, 수리력, 문제해결력 순위 비교. *한국교원교육학회 학술대회자료집*, 2016(11), 457-460.

최윤정 (2015a). 성향점수매칭을 활용한 평생학습 참여의 역량 효과 분석: 국제성인역량조사(PIAAC)를 중심으로. *평생교육학연구*, 21(4), 139-167.

한국교육과정평가원 (2019). *「PISA 2018」 결과 발표 별첨 자료*. 한국교육과정평가원 웹 사이트에서 https://kice.re.kr/boardCnts/view.do?m=050102&boardID=10024&viewBoardID=10024&boardSeq=5048324&lev=0&statusYN=W&page=1에서 2020년 5월 13일 인출.

한국직업능력개발원(2013). 한국인의 역량, 학습과 일: *성인역량조사(PIAAC) 보고서*. 서울: 교육부.

한국교육개발원 (2020). *평생학습개인실태조사 (통계정보보고서)*. 통계청 웹 사이트https://www.kostat.go.kr/portal/korea/kor_pi/8/7/index.board?bmode=read&aSeq=387566에서 2022년 9월 7일 인출.

Hanushek, E. A., Piopiunik, M., & Wiederhold, S. (2019). The value of smarter teachers: International evidence on teacher cognitive skills and student performance. *Journal of Human Resources*, 54(4), 857-899.

Meroni, E. C., Vera-Toscano, E., & Costa, P. (2015). Can low skill teachers make good students? Empirical evidence from PIAAC and PISA. *Journal of Policy Modeling*, 37, 308-323.

OECD (2010). PIAAC background questionnaire MS version 2.1 d.d. 15-12-2010. Retrieved January 10, 2018 from http://www.oecd.org/edu/48442549.pdf.

OECD (2011). OECD at a Glance. Paris: OECD Publishing.

Proceedings of the Korean Data Analysis Society
January 26-27, 2023, pp. 27-31

다문화 학령기 아동의 자아존중감 모형

김미경[1], 박공주[2]

요 약

본 연구는 한국청소년정책연구원 2019 다문화청소년패널조사 2기 자료(MAPS 2019)의 원시자료를 이용한 이차분석 연구이며, Maslow (1987)의 욕구단계이론을 기반으로 다문화 학령기 아동의 자아존중감에 대한 가설적 모형을 구축하여 모형의 적합도를 검증하고, 다문화 학령기 아동의 자아존중감에 영향을 미치는 요인의 경로와 직·간접 효과를 파악하고자 시도되었다. 수집된 자료는 SPSS/Win와 Amos 22.0 프로그램을 이용하여 분석하였다. 연구결과 자아존중감에서 건강평가(β=.186, p<.001), 친구의 지지(β=.062, p=.003), 사회적 역량(β=.388, p<.001), 학교생활(β=.118, p<.001)은 유의한 직접효과와 총효과가 있었다. 건강평가(β=.014, p<.001), 친구의 지지(β=.029, p<.001), 사회적 역량(β=.025, p<.001)은 학교생활을 통하여 자아존중감에 영향을 미치는 간접효과가 있었다. 학교생활에서 건강평가(β=.121, p<.001), 친구의 지지(β=.248, p<.001), 사회적 역량(β=.213, p<.001)은 유의한 직접효과와 총효과가 있었다.

주요용어 : 다문화, 학령기, 아동, 자아존중감.

1. 서론

1.1. 연구의 필요성

최근 우리나라는 점차 다문화 가정이 증가되고 있다. 2021년 다문화 인구동태 통계에 의하면, 다문화 혼인은 전체 혼인 비중의 7.2%를 차지하며, 전체 출생 중 다문화 출생은 5.5%이다(Kosis, 2022). 학령기 아동은 다양한 신체적 변화 및 상황과 환경에 대해 현실적인 이해를 시작하는 시기로 자아성장에 중요한 시기이며, 청소년기와 성인기의 발달과정 영향을 미치는 중요한 시기이다(Kang, et. al., 2007; Han, Choi, Park, 2004; Lee, Park, 2021). 이 시기에 자신에 대한 긍정적 평가 혹은 부정적 평가는 자아존중감에 영향을 미치며(Han, Choi, Park, 2004; Shin, 2015), 자아존중감은 삶의 질에 영향을 미치는 중요한 요소가 된다(Baik, 2012). 또한 또래집단인 친구들과의 관계에 많은 영향을 받는 시기이며. 소속감과 동질감 및 사회적 관계가 자아존중감에 영향을 미치며(Jang, Jung, 2013; Yang, Kim, 2015), 학교생활 적응을 잘 할수록 자아존중감이 높다(Son, Seo, 2022). 다문화 가정의 학령기 자녀들이 학교생활을 할 때 다문화의 특수한 환경으로 인해 일반 학생들보다 학교생활에 잘 적응하지 못하는 경향이 있고, 그러한 환경과 상황으로 인해 자아존중감이 낮을 수 있다

[1](제1저자) 46270 부산시 금정구 부곡동로 27번길 88, 대동대학교 간호학과 부교수.
E-mail: joykim2008@naver.com
[2](교신저자) 50811 경남 김해시 삼안로 112번길 198, 김해대학교 간호학과 조교수. E-mail: conti52@hanmail.netr

(Erikson, 1968; Park, Song, 2010). 다문화 청소년은 일반 청소년에 비해 학업 중단율이 높게 나타났으며(Ministry of Education, 2020), 학습 및 학교생활과 친구들과의 관계 등에 잘 적응하지 못하는 결과가 나타났다(Kim, 2018; Son, Seo, 2022). 또한 학교폭력은 학교생활 적응의 어려움을 가져오고, 자아존중감에 영향을 미칠 수 있다(Oh, 2010).

본 연구는 Maslow의 욕구단계이론을 다문화 학령기 아동의 자아존중감을 설명하는 이론적 기틀로 설정하고, 기본적 욕구단계인 생리적 욕구단계와 안전 욕구를 외생변수로 설정하였다(Lee, 2020; Maslow, 1943). 소속과 애정의 욕구는 내생변수 중 과정변수로 설정하여 학령기 아동이 소속되어 있는 학교에서의 생활을 통해 확인하고자 하였다. 학교생활에 잘 적응할수록 자아만족감을 높일 수 있어(Kim, 2018; Son, Seo, 2022) 학교생활을 과정변수인 매개변수로 설정하였다. 자아존중 욕구는 내생변수 중 결과변수로 자아존중감을 설정하였다(Kim, 2019; Maslow, 1987). 따라서 본 연구는 점차 증가하는 다문화 학령기 아동의 자아존중감에 영향을 미치는 요인을 매슬로우의 욕구이론을 토대로 직접적, 간접적 요인을 확인하고, 그 결과를 근거로 다문화 학령기 아동의 자아존중감을 높이는 간호중재 프로그램 개발의 기초자료를 제공하고자 한다.

1.2. 연구의 목적

본 연구의 목적은 Maslow (1987)의 욕구단계이론을 기반으로 다문화 학령기 아동의 자아존중감에 대한 가설적 모형을 구축하고, 모형의 적합도를 검증하고 다문화 학령기 아동의 자아존중감에 영향을 미치는 요인의 경로와 직·간접 효과를 파악하고자 한다.

1.3. 연구의 개념적 기틀 및 가설적 모형

본 연구는 다문화 학령기 아동의 자아존중감을 설명하는 모형을 구축하기 위하여 Maslow (1987)의 욕구단계이론과 선행연구를 근거로 연구의 개념적 기틀을 구축하였다(Figure 1, 2).

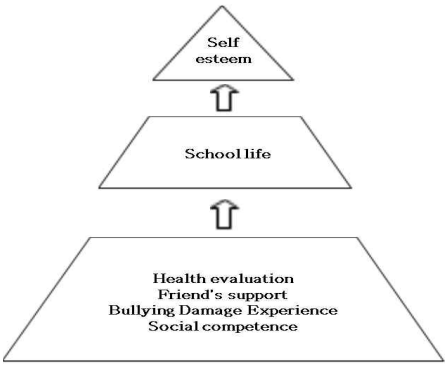


Figure 1. Theoretical framework

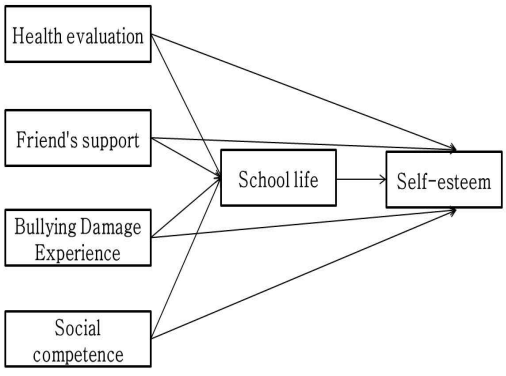


Figure 2. Hypothetical model.

2. 연구방법

2.1. 연구 설계

본 연구는 한국청소년정책연구원 2019 다문화청소년패널조사 2기 자료(MAPS 2019)의 원시자료를 이용하여 다문화 학령기 아동의 자아존중감에 영향을 미치는 요인들을 종합하여 가설적 모형을 구축하고 모형의 적합도를 판단하여 변수 간의 관계를 검증하는 2차 자료분석 연구이다.

2.2. 연구대상 및 자료 수집 방법

본 연구는 한국청소년정책연구원 2019 다문화청소년패널조사 2기 자료(MAPS 2019)를 활용하였다. MAPS 2기는 다문화청소년의 성장과 발달과정에 대한 기초자료를 구축하는 목적에서 2019년을 기준으로 초등학교 4학년으로 연구대상자 표집방법은 확률비례추출법으로 추출하였다. 본 연구는 원시자료를 이용하는 2차 분석 자료이므로 원자료에서 표본으로 선출된 대상자수는 2,246명으로 구조방정식 모형에서의 표본크기에 최소 권장수준을 만족한다. 본 연구는 질병관리본부의 청소년 건강행태온라인조사 원시자료 활용 규정에 의거하여 연구자가 자료 이용을 승인받아 사용하였다.

2.3. 자료분석 방법

수집된 자료는 SPSS/Win와 Amos 22.0 프로그램을 이용하여 분석하였다. 적합도 검증은 카이검증, χ2/df, 기초적합지수(Goodness of Fit Index, GFI), 조정적합지수 (Adjusted Goodness of Fit Index, AGFI), 근사오차 평균자승의 이중근(Root Mean Square Error of Approximatiom, RMSEA), 비교적합지수(Comparative Fit Index, CFI), 표준적합지수(Turker-Lewis Index, TLI), 원소간평균차이(Root Mean Square Residual, RMSR), 잔차평균자승이중근(Root Mean Square Residual, RMR)를 구하였다. 가설적 모형의 경로에 대한 유의성 검증은 Standardized estimate, Critical Ratio(C.R.), p 값을 이용하고, 내생변수에 대한 설명력은 Squared Multiple Correlation (SMC)를 이용하였다.

3. 연구결과

3.1. 가설적 모형의 적합도 검증

본 연구에서 적합도 지수를 분석한 결과 χ2=.25 (p>.874), df(자유도)=1, RMSEA=.01, RMR=.01, GFI=1.0, AGFI=.1.0, TLI=1.0, CFI=1.0이었다. χ2은 p>.05이상이면 모형이 표본자료에 적합하다고 평가할 수 있다(Table 2).

3.2. 가설적 모형의 모수추정

가설모형에서 통계적으로 유의하게 나타난 경로를 살펴보면, 자아존중감에 유의하게 영향을 준 경로는 건강평가(β=.186, p<.001), 친구의 지지(β=.062, p=.003), 사회적 역량(β=.388, p<.001), 학교생활(β=.118, p<.001)으로 설명력은 22%였다. 학교생활에 유의하게 영향을 준 경로는 건강평가(β=.121, p<.001), 친구의 지지(β=.248, p<.001), 사회적 역량(β=.213, p<.001)으로 설명력은 36%이였다(Table 2).

3.3. 가설적 모형의 효과분석

자아존중감에서 건강평가(β=.186, p<.001), 친구의 지지(β=.062, p=.003), 사회적 역량(β=.388, p<.001), 학교생활(β=.118, p<.001)은 유의한 직접효과와 총효과가 있었다. 건강평가(β=.014, p<.001), 친구의 지지(β=.029, p<.001), 사회적 역량(β=.025, p<.001)은 학교생활을 통하여 자아존중감에 영향을 미치는 간접효과가 있었다. 학교생활에서 건강평가(β=.121, p<.001), 친구의 지지(β=.248, p<.001), 사회적 역량(β=.213, p<.001)은 직접효과와 총효과가 유의하게 나타났다(Table 2). 최종적인 가설모형 경로는 Figure 2와 같다.

Table 2. Standardized Direct, Indirect, and Total Effects for the Final Model (N=2,246)

Endogenous variables	Exogenous variables	β	S.E	C.R	p	Direct effect (p)	Indirect effect (p)	Total effect (p)	SMC*
Self-esteem	Health evaluation	.186	.032	9.558	<.001	.186(<.001)	.014(<.001)	.186(<.001)	.22
	Friend's support	.062	.015	3.010	.003	.062(.003)	029(<.001)	.062(.003)	
	Social competence	.388	.022	17.572	<.001	.388(<.001)	.025(<.001)	.388(<.001)	
	School life	.118	.015	6.210	<.001	.118(<.001)	-	.118(<.001)	
School life	Health evaluation	.121	.032	9.558	<.001	.121(<.001)	-	.121(<.001)	.36
	Friend's support	.248	.020	10.992	<.001	.248(<.001)	-	.248(<.001)	
	Bullying Damage Experience	.032	.019	1.633	.103.	.032(.103)		.032(.103)	
	Social competence	.213	.031	8.772	<.001	.213(<.001)	-	.213(<.001)	
Fit indices for hypothetical model		χ2=.025, df=1, p>.874, GFI=.1.0, AGFI=.1.0, CFI=.1.0, AGFI=.1.0, RMR=.01, RMSEA=.01							

C.R. = Critical ratio, *SMC=squared multiple correlation

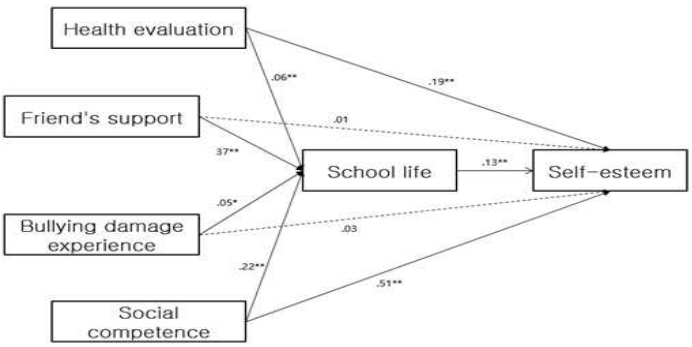


Figure 3. Path coefficients of modified model

References

Han, S. C., Choi, A. M., Park, S. H. (2004). Adolescent psychology, Seoul: Yangsuewon Publishing Co.

Kang, K. A., Kim, S. J., Song, M. K., Sim, S. Y., Kim, S. H. (2007). Development of an instrument to measure meaning of life(MOL) for higher grade primary school children. *Journal of Korean Academy Child Health Nursing*, 13(3), 308-317. (in Korean).

Kosis (2022). https://kosis.kr/index/index.do

Lee, M. S., Park, S. C. (2021). The Effect of Multicultural Adolescents' School Life Satisfaction on Career Attitudes - Moderating Effects of Social Withdrawa. *Journal of the Korean Data Analysis Society*, 23(2), 829-843. (in Korean). https://doi.org/10.37727/jkdas.2021.23.2.829

Shin, H. G. (2015). Influencing Factors of Quality of Life, Self-esteem of Higher Grade Primary School Children. *Journal of the Korean Data Analysis Society*, 17(1), 485-499. (in Korean).

Baik, B. R (2012). A Study on the Effect of Self-esteem, Sociality, School resilience to create the social capital of adolescence. Hanyoung Theological University Ph.D. thesis. (in Korean).

Yang, Y. M., Kim, J. S. (2015). Relationship between Children's Self Esteem, Sense of Community and Multicultural Receptivity. *Journal of Youth Welfare*, 17(4), 309-328. (in Korean).

Proceedings of the Korean Data Analysis Society
January 26-27, 2023, pp. 33-37

간호대학생의 코로나19 위험인식, 수업참여도 및 학습만족도에 관한 연구*

김민영[1]

요 약

본 연구는 간호대학생의 코로나19 위험인식과 수업참여도가 학습만족도에 미치는 영향을 알아보기 위해 시행되었다. 자료수집은 B시 5개 대학교 간호학과 대학생 168명을 대상으로 2021년 4월부터 5월까지 시행되었으며, 수집된 자료는 SPSS program 26.0을 이용하여 기술통계, t-test, ANOVA, pearson's correlation coefficients, multiple regression analysis로 분석하였다. 그 결과 코로나19 위험인식은 수업참여도(r=.258, p<.001) 및 학습만족도(r=269, p<.001)와 유의한 정적 상관관계를 보였고, 수업참여도는 학습만족도(r=.505, p<.001)와 정적 상관관계를 보였다. 대상자의 학습만족도에 영향을 미치는 요인으로 코로나19 위험인식(β=.140, p=.041)과 수업참여도(β=.444, p<.001)가 나타났다. 따라서 코로나19에 대한 객관적이고 정확한 정보를 통해 코로나19 위험인식을 올바르게 증가시킬 수 있는 방안과 수업참여도 향상을 위한 실제적 교육 방안을 모색해야할 것이다.

주요용어 : 간호대학생, 코로나19, 위험인식, 수업참여도, 학습만족도.

1.서론

1.1 연구의 필요성

2019년 코로나바이러스감염증-19(코로나19)의 첫 인체 감염이 확인되면서, 현재 세계 인구의 658,080,362명이 감염되었고(WHO, 2022. 12) 국내 누적환자도 28,534,558명에 이르고 있으며(Korea Disease Control and Prevention Agency, 2022. 12) 세계적 대유행은 계속되고 있다.

이러한 지속적인 감염병의 확산은 사회적 불안과 공포로 나타날 수 있고(Park, 2020) 이는 개개인의 위험인식에 대한 중요한 작용의 하나가 될 수 있으며(Kasperson et al., 1998), 코로나19와 같이 불확실성이 큰 질병일수록 위험인식과 두려움이 증가하는 것으로 나타났다(Park, et al., 2021). 코로나19에 대한 위험인식은 그 예방행위에 유의한 영향을 미치고 있다는 결과(Han, Lee, 2021)를 볼 때 간호대학생들의 감염예방 측면에서도 이해할 필요가 있을 뿐 아니라, 수업운영에도 다양한 변수로 작용할 수 있으므로 간호대학생의 코로나19 위험인식을 파악해야할 필요성이 있다.

코로나19는 대학교육에도 영향을 주어 교육방식의 변화를 가져왔다. 2020년부터 정부 지침의 일환으로 시행된 비대면수업(온라인수업)으로 인해 간호대학생은 실무 및 임상실습 교육을 실시함에 있어 실습기관의 실습현장 제공 등이 불가피해지면서 임상실습이 대면으로 이루어지지 못하였고,

*본 논문은 2022년도 경남정보대학교 간호학과 발전기금 교수연구지원비 지원을 받아 수행된 연구임.
[1]47011 부산광역시 사상구 주례로 45, 경남정보대학교 간호학과 조교수. E-mail: 0502young@hanmail.net

갑작스러운 상황에서 간호교육자들의 대체 콘텐츠가 충분히 준비되지 않아 상호간 여러 문제들이 발생하였다. 간호대학생은 이러한 혼란스러운 교육상황에서 실무와 연계된 교과목을 비롯해 여러 교과목을 비대면수업으로 진행하게 되었고 이는 교과목에 대한 정보공유와 질의응답 등 학습자와 교수자의 상호작용이 어려워 수업참여도에 부정적 영향을 주면서(Min et al., 2014) 학습만족도에도 영향을 미친다는 보고가 있었다(Gill, Andersen, Hillsmann, 2019). 수업참여도는 학습자가 행동적, 정서적 및 인지적 측면에서 자발적이고 능동적으로 수업 활동에 개입하는 것으로 학습만족도에 영향을 준다(Kim, 2017). 또한 학습만족도는 학습과정에서 교육자와 학습자가 만들어낸 결과물에 대한 반영일 뿐 아니라 비대면 수업에서도 학습의 지속성에 영향을 미치는 중요한 요인이므로(Chiu et al., 2007) 수업참여도와 학습만족도는 코로나19 상황에서 연구되어야할 과제이다.

이에 본 연구에서는 간호대학생의 코로나19의 위험인식을 알아보고 수업참여도와 학습만족도의 관련성 및 영향요인을 파악하여 향후 간호대학생들의 수업참여도와 학습만족도를 향상시키기 위한 기초자료로 활용하기 위해 시행하였다.

1.2. 연구 목적

본 연구는 간호대학생의 코로나19 위험인식, 수업참여도, 학습만족도를 살펴보고 간호대학생의 학습만족도에 관한 영향요인을 파악하는 것을 목적으로 한다.

2. 연구방법

2.1. 연구설계

본 연구는 간호대학생의 코로나19 위험인식, 수업참여도 및 학습만족도를 알아보고 그 영향요인을 알아보기 위한 서술적 조사연구이다.

2.2. 연구대상

본 연구의 대상자는 B시 5개 대학교에 재학중인 간호학과 대학생 2~4학년이다. 연구대상자의 수는 G*Power analysis program 3.1을 이용하여 최소 표본수는 152명으로 산출되었으나 탈락률 20%를 고려하여 180명에게 설문지를 배포하였고, 수거된 설문지 중 응답이 불충분한 자료 12부를 제외한 총 168부를 자료 분석에 이용하였다.

2.3. 자료수집 및 윤리적 고려

본 연구는 2021년 4월~5월까지 B시에 소재한 K대학교의 간호대학생을 대상으로 자가보고식 설문지를 사용하여 진행하였다. 자료수집 전에 본 연구 방법 및 목적을 충분히 설명하고, 대상자의 자발적 참여가 가능하며 응답한 내용은 모두 무기명 처리되어 대상자의 비밀 보장이 원칙이라는 내용을 설명하고 연구참여 동의서를 받았다. 연구대상자에게 설문지를 직접 배부하였으며 작성시간 약 20분 정도가 소요되었고, 설문에 응답한 대상자에게는 소정의 선물을 제공하였다. 또한 연구참여 의사 철회가 가능하고 모든 자료는 본 연구의 목적으로만 사용되고 연구 종료 후 폐기됨을 설명하였다.

2.4 자료분석

본 연구는 SPSS program 26.0을 사용하여 기술통계, t-test, ANOVA, pearson's correlation coefficients, multiple regression analysis로 분석하였다.

3. 연구결과

3.1. 대상자의 코로나19 위험인식, 수업참여도, 학습만족도 정도

본 연구대상자의 코로나19 위험인식 정도는 4.66±0.98점, 수업참여도는 3.43±0.59점, 학습만족도는 3.74±0.56점으로 나타났다.

3.2. 대상자의 일반적 특성에 따른 코로나19 위험인식, 수업참여도, 학습만족도

대상자의 일반적 특성에 따른 코로나19 위험인식은 유의한 차이가 나타나지 않았고, 수업참여도에서는 학년(F=4.862, p=.009)에서 유의한 차이가 있었는데 사후분석 결과 2학년보다 3학년이 유의하게 높게 나타났다. 대상자의 일반적 특성에 따른 학습만족도는 학년(F=6.965, p<.001)이 유의하게 나타났으며, 사후분석 결과 3~4학년이 2학년보다 유의하게 높게 나타났다.

3.3. 대상자의 코로나19 위험인식, 수업참여도, 학습만족도의 상관관계

코로나19 위험인식은 수업참여도(r=.258, p<.001) 및 학습만족도(r=269, p<.001)와 유의한 정적 상관관계로 나타났고, 수업참여도는 학습만족도(r=.505, p<.001)와 정적 상관관계를 보였다(Table 1).

Table 1. Correlation of COVID 19 Risk Perception, Class participation and Learning Satisfaction (N=168)

Variables	COVID 19 Risk Perception r(p)	Class participation r(p)	Learning Satisfaction r(p)
COVID 19 Risk Perception	1		
Class participation	.258**	1	
Learning Satisfaction	.269**	.505**	1

p**<0.01

3.4. 대상자의 학습만족도에 미치는 영향요인

대상자의 학습만족도에 미치는 영향요인을 알아보기 위해 다중회귀분석을 실시한 결과 Table 2와 같이 나타났다. 일반적 특성에서 유의한 결과를 보인 학년 및 코로나19 위험인식, 수업참여도, 학습만족도를 분석하였으며 그 결과 학습만족도에 영향을 미치는 요인으로 코로나19 위험인식(β=.140, p=.041)과 수업참여도(β=.444, p<.001)로 나타났다. 이들 변인에 의한 설명력은 27.9%(F=22.54, p<.001, R^2=.292, Adj-R^2=.279)이었다. 분석 전 잔차들 간 상관이 있는지를 알아보기 위해 더빈-왓슨 통계량을 살펴본 결과 1.745로 잔차들 간 상관은 없었으며, 독립변수들 간 다중공선성을 알아보기 위해 공차한계와 분산팽창지수(VIF)를 살펴본 결과 공차한계는 .906~.930로 .10 이상으로 나타났고, 분산팽창지수(VIF)는 1.043~1.104로 10 이하의 결과를 보여 독립변수들 간 다중공선성은 없었다.

Table 2. Influencing factor in Learning Satisfaction (N=168)

Variable	Quality of life						
	B	SE	β	t	p	Tolerance	VIF
Constant	1.749	0.254		6.890	.000		
Grade	.101	.052	.130	1.937	.054	.959	1.043
COVID 19 Risk Perception	.081	.039	.140	2.059	.041	.930	1.076
Class participation	.416	.065	.444	6.428	.000	.906	1.104

Durbin-Watson=1.745, F=22.540, p<.001, R2=.292, Adj-R2=.279
Tolerance=.906~.930, VIF=1.043~1.104

4. 논의

본 연구의 논의는 다음과 같다.

간호대학생의 코로나19 위험인식은 7점 만점에 4.66점으로 중간 수준으로 간호대학생을 대상으로 측정한 Kim, Cheon, Yoo(2021)의 3.18(7점 만점 환산)점보다 높았는데 이는 코로나19의 확산 기간이 길어지고 누적확진자수가 증가함에 따라 위험인식도 높아진 것으로 생각되며 적절한 위험인식을 통하여 코로나19에 대응하는 긍정적 요소로 작용할 수 있도록 해야 할 것이다. 대상자의 수업참여도는 5점 만점에 3.43점으로 간호대학생을 대상으로 한 Lee(2021)의 연구에서의 4.18점(5점 만점)보다 낮게 나타났는데 이는 코로나19 상황상 비대면 수업이 생기면서 기존 실습 수업 등의 대면수업에 비해 참여도가 낮아진 것으로 판단된다. 대상자의 학습만족도는 5점 만점에 3.75점으로 간호대학생을 대상으로 비대면 혼합수업 학습만족도를 조사한 Park(2022)의 연구에서 3.98점(5점 만점)보다 낮게 나타났는데 이는 코로나19로 인한 갑작스러운 수업방식의 변화로 어려움이 생기고 적응의 전환기로 인한 것으로 추론되며 향후 비대면 수업과 관련된 시설, 사용자 편의성 등을 고려하여 학습만족도 향상을 위한 수업 운영 개선이 필요하다고 생각된다.

대상자의 일반적 특성에 따른 코로나19 위험인식과 수업참여도 및 학습만족도를 분석한 결과 수업참여도와 학습만족도에서 각각 학년과 유의한 차이를 나타내었다. 수업참여도에서는 2학년이 3학년보다 높은 것으로 나타났는데 이는 3학년 학생들은 팀별 학습, 개인준비도 확인시험(IRAT), 그룹준비도 확인시험(GRAT) 등의 수업이 많아지면서 오히려 부담감이 증가하여 수업참여도가 낮은 것으로 판단되므로, 팀원 간의 상호작용을 강조하고 학습자와 교수자의 적극적 소통이 이루어져야할 것이다. 대상자의 일반적 특성에 따른 학습만족도는 학년에서 3~4학년이 2학년보다 높게 나타났다. 이는 Kim(2021)의 연구에서 간호대학생의 학년별 학습만족도가 4학년이 2~3학년보다 낮게 나타난 결과와 상이한 결과로, 이는 이미 수업에 익숙해있던 3~4학년이 코로나19라는 특수한 상황에서 변화된 수업 방식에 대한 적응속도가 상대적으로 빨라서 나타난 결과라 사료된다.

대상자의 코로나19 위험인식, 수업참여도 및 학습만족도의 상관관계를 살펴본 결과, 코로나19 위험인식은 수업참여도(r=.258, p<.001) 및 학습만족도(r=269, p<.001)와 유의한 정적 상관관계로 나타났고, 수업참여도는 학습만족도(r=.505, p<.001)와 정적 상관관계를 보였다. 코로나19 위험인식이 높을수록 수업참여도와 학습만족도가 높게 나타난 결과로 볼 때 올바른 위험인식을 갖출 수 있도록 정확한 정보를 전달할 수 있는 체계가 객관적으로 마련되어야할 것이다. 수업참여도는 학업만족도

와 상관성이 있으며, 특히 비대면 수업에서 수업참여는 Kim(2021)의 연구에서 학업만족도와 유의한 상관관계가 있다는 결과와 일치한다. 따라서 간호대학생의 수업참여도를 높이기 위해 교수자와 학습자의 상호작용을 높이고 다양한 토론 및 평가방법을 활용하여 학업만족도를 높이기 위해 노력해야할 것이다.

대상자의 학습만족도에 영향을 미치는 요인을 파악하기 위해 다중회귀분석을 실시한 결과, 코로나19 위험인식과 수업참여도가 영향요인으로 나타났다. 따라서 간호대학생의 학습만족도를 향상시키기 위해서 코로나19 위험인식을 올바르게 함양하고 수업참여도를 높여야 할 것이다. 이를 위해 대학에서는 코로나19에 대한 빠르고 정확한 정보 서비스를 제공하여 효과적으로 코로나19에 대한 위험인식을 향상시키고, 다양한 형태의 수업에 대한 변화와 소통에 민감하게 반응하여 시스템을 구축해 나가야할 것이다.

5. 결론 및 제언

본 연구의 결과에 따라, 코로나19에 대한 객관적이고 정확한 정보를 통해 코로나19 위험인식을 올바르게 증가시킬수 있는 방안과 수업참여도향상을 위한 실제적 교육 방안을 모색해야할 것이다. 또한 코로나19 상황에서 간호대학생의 학습만족도에 영향을 줄 수 있는 다양한 요인들을 반영한 연구가 추후 시행되어야할 것이며, 신종감염병에 대한 위험인식수준을 높이고 예방하는 구체적인 방안을 고려한 연구를 제언한다.

Proceedings of the Korean Data Analysis Society
January 26-27, 2023, pp. 39-39

자궁경부암 진단 여성의 삶의 질 영향요인

전은미[1]

요 약

본 연구는 자궁경부암 진단 여성의 삶의 질에 영향을 미치는 요인을 확인하고자 2016~2020년 국민건강영양조사 자료의 자궁경부암 진단 여성 151명을 대상으로 조사하였다. 수집된 자료는 SPSS/WIN 25 프로그램을 이용하여 실수와 백분율, 평균, 표준편차, 집단 간 비교는 교차분석과 일반선형분석을 이용한 t-test, ANOVA를 이용하였으며, 사후검정은 Bonferroni 방법으로 분석하였다. 자궁경부암 진단 여성의 삶의 질에 영향을 미치는 요인은 일반선형분석을 이용한 다중회귀분석을 이용할 예정이다. 연구결과 인구학적 특성에 따른 삶의 질은 연령(F=5.21, p=.001), 결혼상태(F=16.56, p=.001), 교육수준(F=4.97, p=.002), 경제활동(t=6.00, p=.015)에서 통계적으로 유의한 차이가 있는 것으로 나타났고, 건강특성에 따른 삶의 질은 산책여부(F=164.8, p=.001), 근력운동(F=149.9, p=.001), 유산소운동(F=294.0, p=.001), 건강검진(F=5.28, p=.022), 우울(F=-193.3, p=.001), 피임약 복용(F=52.9, p=.001), 음주(F=210.9, p=.001), 흡연(F=3.11, p=.045)과 수면시간(F=35.3, p=.001)에서 통계적으로 유의한 차이가 있었다. 대상자의 삶의 질 측정결과 EQ-5D Index는 1점 만점에 0.91±0.14점, 주관적 건강상태는 5점 만점에 3.31±0.87점, 스트레스 인지정도는 4점 만점에 2.92±0.78점, 체질량지수는 평균 24.14±3.33점이었다. 삶의 질은 주관적 건강상태(r=-0.562, p=0.001), 체질량지수(r=-0.174, p=.034)와 통계적으로 유의한 상관관계가 있는 것으로 나타났다.

주요용어 : 자궁경부암, 삶의 질, 주관적 건강상태, 체질량지수.

[1]35345 대전광역시 서구 배재로 155-40(도마동), 배재대학교 간호학과 교수. E-mail: charminggold@pcu.ac.kr

Proceedings of the Korean Data Analysis Society
January 26-27, 2023, pp. 41-43

청소년 위험 음주 영향 요인

박민희[1]

주요용어 : 청소년, 음주, 위험 음주.

1. 서론

1.1. 연구의 목적

본 연구의 목적은 청소년 위험 음주에 영향을 미치는 요인을 파악하기 위함이며, 구체적인 목적은 다음과 같다.

가. 대상자의 일반적 특성과 심리 특성을 파악한다.
나. 대상자의 음주 및 위험 음주 특성을 파악한다.
다. 대상자의 일반적 특성과 심리 특성에 따른 위험 음주의 차이를 파악한다.
라. 대상자의 음주 특성에 따른 위험 음주의 차이를 파악한다.
마. 대상자의 위험 음주 영향 요인을 파악한다.

2. 연구 방법

2.1. 연구 설계

본 연구는 부산광역시교육청의 2022년 학생 흡연, 음주, 약물 오·남용 실태의 원시자료를 이용하여(Busan Metropolitan City Office of Education, 2022), 청소년 위험 음주에 영향을 주는 요인을 파악하기 위한 2차 분석연구이다.

2.2. 연구 대상 및 자료 수집

연구대상자는 부산광역시교육청이 시행한 2022년 학생 흡연, 음주 약물 오·남용 실태 설문에 참여한 학생 중 음주 경험이 있는 5,886명이다.

위험 음주 청소년은 '최근 30일 동안 1회 평균 음주량이 남학생은 소주 5잔(맥주 3병) 이상, 여학생은 소주 3잔(맥주 2병) 이상'인 대상자이다.

[1]47011 부산광역시 사상구 주례로 47 동서대학교 바이오헬스융합대학 간호학과 조교수.
E-mail: mhpark@gdsu.dongseo.ac.kr

3. 연구 결과

3.1. 일반적 특성

Table 1. General factors

Variables		Range	n (%) or mean (SD)			χ^2 or F	p
			Drinking experience	Current drinking	Risk drinking		
Total	-		4,548(77.2)	664(11.3)	674(11.5)	-	-
Gender	boys	-	2,838(77.0)	449(12.2)	401(10.9)	9.998	.007
	girls	-	1,710(77.8)	215(9.8)	273(12.4)		
School level	middle	-	2,097(84.0)	287(11.5)	113(4.5)	264.331	< .001
	general High	-	1,837(74.9)	272(11.1)	344(14.0)		
	specialized high	-	614(65.6)	105(11.2)	217(23.2)		
Living arrangements	with family	-	4,385(77.9)	637(11.3)	610(10.8)	49.794	< .001
	without a family	-	163(64.2)	27(10.6)	64(25.2)		
Subjective academic achievement	1~5		3.00(1.195)a	3.15(1.2.34)b	2.85(1.262)c	6.414	< .001 b > a > c

3.2. 심리 특성

Table 2. Psychological factors

Variables	Range	n (%) or mean (SD)			χ^2 or F	p
		Drinking experience	Current drinking	Risk drinking		
Subject health recognition	1~5	3.86(0.988)a	3.95(0.963)b	3.80(1.080)c	4.012	.018 a, b > c
Subject well-being	1~5	3.74(1.013)a	3.83(1.012)b	3.60(1.160)c	8.780	< .001 a, b > c
Stress recognition	1~5	3.28(0.949)a	3.23(1.021)b	3.41(1.090)c	6.184	.002 a, b < c
loneliness	1~5	2.52(1.078)a	2.59(1.144)b	2.81(1.218)c	20.423	< .001 a, b < c
Anxiety	0~21	3.09(4.267)a	3.31(4.312)b	4.05(5.343)c	13.902	< .001 a, b < c
Suicidal ideation	yes	556(12.2)	106(16.0)	154(22.8)	58.243	< .001
	no	3,992(87.8)	558(84.0)	520(77.2)		

3.3. 음주, 흡연, 약물 오·남용 특성

Table 3. Drinking factors

Variables		n (%) or mean (SD)			χ^2 or F	p
		Drinking experience	Current drinking	Risk drinking		
Reason for drinking	friends, senior	63(1.4)	9(1.4)	12(1.8)	311.292	< .001
	family	2,161(47.5)	293(44.1)	188(27.9)		
	curiosity	1,318(29.0)	140(21.1)	143(21.2)		
	stress	51(1.1)	16(2.4)	42(6.2)		
	special day	340(7.5)	54(8.1)	67(9.9)		
	no reason	615(13.5)	152(22.9)	222(32.9)		
Ease of purchasing alcohol	yes	0(0.0)	70(10.5)	343(50.9)	2344.371	< .001
	no	4,548(100.0)	594(89.5)	331(49.1)		
Storage of alcohol at home	yes	2,554(56.2)	425(64.0)	423(62.8)	22.318	< .001
	no	1,994(43.8)	239(36.0)	251(37.2)		
Allowed to drinking	yes	733(16.1)	146(22.0)	121(18.0)	14.662	.001
	no	3,815(83.9)	518(78.0)	533(82.0)		
Exposure to drinking advertising	yes	2,331(51.3)	400(60.2)	351(52.1)	18.7888	.001
	no	2,217(48.7)	264(39.8)	323(47.9)		
Drinking prevention education	yes	2,247(49.4)	371(55.9)	314(46.4)	12.861	.002
	no	2,301(50.6)	293(44.1)	360(53.4)		
Willingness to no drinking in adolescents	1~5	3.88(1.105)a	3.02(1.232)b	2.55(1.259)c	504.092	< .001 a > b > c
Willingness to no drinking for life	1~5	2.64(1.439)a	2.15(1.305)b	1.96(1.264)c	92.060	< .001 a > b > c
Refusal to drinking	1~5	3.80(1.075)a	3.08(1.179)b	2.66(1.267)c	385.177	< .001 a > b > c
Current smoking	yes	105(2.3)	39(5.9)	225(33.4)	964.808	< .001
	no	4,443(97.7)	625(94.1)	449(66.6)		
Liquid e-cigarette using	yes	64(1.4)	31(4.7)	166(24.6)	747.088	< .001
	no	4,484(98.6)	633(95.3)	508(75.4)		
Heating e-cigarette using	yes	25(0.5)	17(2.6)	87(12.9)	418.706	< .001
	no	4,523(99.5)	647(97.4)	587(87.1)		
Habitual drug using	yes	117(2.6)	32(4.8)	56(8.3)	59.287	< .001
	no	4,431(97.4)	632(95.2)	618(91.7)		
Weight loss drug experience	yes	164(3.6)	26(3.9)	6.9(10.9)	61.783	< .001
	no	4,384(96.4)	638(96.1)	605(89.8)		
Concentration-enhancing drug experience	yes	37(0.8)	12(1.8)	21(3.1)	28.908	< .001
	no	4,511(99.2)	652(98.2)	653(96.9)		
To get rid of a Hangover experience	yes	30(0.7)	9(1.4)	60(8.9)	241.644	< .001
	no	4,518(99.3)	655(98.6)	614(91.1)		
Illegal drugs experience	yes	15(0.3)	8(1.2)	25(3.7)	84.283	< .001
	no	4,533(99.7)	656(98.8)	649(96.3)		
Drug dependence	yes	21(0.5)	10(1.5)	15(2.2)	28.617	< .001
	no	4,527(99.5)	654(98.5)	659(97.8)		

Proceedings of the Korean Data Analysis Society
January 26-27, 2023, pp. 45-45

Locally Averaging Quantile Regression*

Suin Kim[1], *Yoonyoung Jo*[2], *Yoonsuh Jung*[3]

Abstract

Quantile regression (QR) is a robust regression method that gives the overall distributional information of the data. The composite quantile regression (CQR) finds an efficient estimator by combining multiple quantile estimators. Therefore CQR targets the conditional mean even if it uses quantile estimates. The target of CQR can be a conditional median when errors are symmetric. In this work, we generalize the composite quantile regression to other quantiles. The suggested method averages some estimated quantiles around the target quantile. We call it a local averaging quantile regression (LAQR). Determining the range of quantiles to combine is a crucial task. We suggest an empirical rule for selecting an appropriate range of quantiles from extensive simulation studies. Simulation results show that LAQR outperforms original QR in most cases. We explain that the suggested rule is desirable even when the errors are heteroscedastic and nonlinear.

Keywords: locally averaging quantile regression, bandwidth selection, quantile regression.

---

*This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korean government(MSIT)(No. 2022M3J6A1063595).

[1]Ph.D. student, Department of Statistics, Korea University.
[2]Data Analysis TF, Kakao Games Corp, Korea.
[3](Corresponding Author) Associate Professor, Department of Statistics, Korea University.

Proceedings of the Korean Data Analysis Society
January 26-27, 2023, pp. 47-47

The use of Truncated Loss on Multi-armed outcome weighted learning*

Jeongeun Sim[1], *Seung Jun Shin*[2]

Abstract

Studies on allocating appropriate treatment with patients' heterogeneous responses to treatment have been actively explored. An individualized treatment rule (ITR) is a probabilistic model that assigns treatment to maximize potential outcomes in consideration of personal conditions. Outcome weighted learning that casts ITR into a standard binary classification is one of the canonical approaches for finding optimal ITR. In this article, we propose to employ a truncated loss for the multi-armed ITR problem. Toward this, we first extend Fisher consistency of the truncated loss function and then develop an efficient algorithm to optimize it by iteratively solving a sequence of weighted SVMs. The simulation study shows that promising performance of the proposed method.

Keywords : Fisher Consistency, Individualized Treatment Rule, Multi-Armed Bandit, Weighted Support Vector Machine.

*This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korean government(MSIT)(No. 2022M3J6A1063595).

[1]Graduate Student, Department of Statistics, Korea University, Anam-ro, Seongbuk-gu, Seoul, 02841, Korea. E-mail: sim0348@korea.ac.kr

[2](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 26-27, 2023, pp. 49-49

Ensemble Teacher: Semi-Supervised Semantic Segmentation with Teacher's Cross-Pseudo Supervision

Hyunmin Gwak[1], *SungHwan Kim*[2]

Abstract

Semi-supervised learning utilizing both labeled and unlabeled data has recently made significant progress; however, semi-supervised semantic segmentation research still has some problems that need to be solved. In this work, semi-supervised semantic segmentation, which requires more annotation work than other tasks such as classification, was investigated because labels are required on a pixel-by-pixel basis. Consistency regularization based on input image perturbation, feature perturbation and network perturbation in the semi-supervised learning approach showed significant performance improvement. In particular, consistency regularization through network perturbation showed excellent performance in semi-supervised semantic segmentation. However, important challenges remain. Due to the bias problem that pseudo labels used in consistency regularization can have, inaccurate pseudo labels for unlabeled data can have a negative effect on training, which causes more serious problems in semi-supervised segmentation tasks. This study mitigates the problem of pseudo-label bias by expanding the network perturbation-based method, which showed excellent results through a simple ensemble strategy. In addition, this simple method shows better prediction performance than the parallel ensemble structure, and shows the best performance among other consistency regularization based methods. The proposed method can be practically applied to the hands-on data science domain that hardly obtains sufficient class-labels (e.g. medicine, manufacturing, etcs)

Keywords : deep learning, semantic segmentation, semi-supervised learning, consistency training, pseudo label.

[1]Graduate student, Department of Applied Statistics, Konkuk University, 120, Neungdong-ro, Gwangjin-gu, Seoul, Republic of Korea. E-mail: datajobkhm@gmail.com
[2](Corresponding Author) Associate Professor, Department of Applied Statistics, Konkuk University, 120, Neungdong-ro, Gwangjin-gu, Seoul, Republic of Korea. E-mail: shkim1213@konkuk.ac.kr

Proceedings of the Korean Data Analysis Society
January 26-27, 2023, pp. 51-54

형상점 자료의 변환법 및 분류모형의 비교

김준홍[1], 최용석[2]

요 약

Dryden과 Mardia(2016)는 형상분석(shape analysis)에 대하여 기하적 공간상에서 형상점들(landmarks)에 의해 나타낸 개체들(objects)의 형상(shape)을 측정 및 기술하며 이를 비교하는 분석이라고 언급하고 있다. 여기서 형상은 어느 곳에서나 다양하게 존재하는 개체들을 표현할 때 주로 사용된다. 대표적인 예시로는 병원의 x-ray 자료를 통해 병을 진단하거나, 단백질 구조의 형상을 통해 특정 단백질을 식별하는 과정 등이 있고, 이러한 과정에서 여러 분류(classification) 모형들이 활용된다. 기본적으로 형상점 자료는 배열의 형태를 가지고 있는데, 이는 그대로 분류모형에 적용하기 어려움이 있는 형태이다. 본 연구는 형상점 자료를 분류모형에 적용하기 위한 두 가지 방법을 제시한다. 첫 번째 방법은 GPA(Generalized Procrustes Analysis) 적합을 적용한 후 리만거리, 중심크기 및 형상주성분분석을 통한 형상 PC 점수를 얻어내어 이를 변수로 선택하는 방법이고, 두 번째 방법은 각 표본의 형태행렬(configuration matrix)을 벡터화하는 과정을 진행하여 좌푯값들을 모두 변수로써 사용하는 방법이다. 또한, Convex Hull 알고리즘을 통해 형상의 가장 바깥쪽을 구성하는 형상점만을 추출하는 방법을 통해 형상점의 수를 축소하는 방법을 도입한다. 이렇게 새롭게 얻어진 자료들을 랜덤 포레스트, 로지스틱 회귀분석, 그리고 SVM 등 3가지의 분류모형에 투입하여 자료별 오분류율을 비교한다.

주요용어 : 형상분석, GPA(일반화프로크러스티즈분석), 형상점, Convex Hull, 분류모형.

1. 서론

최근 이미지 및 형상에 관련된 분석이 활발하게 진행되고 있고, 그중 이미지 분석의 경우 딥러닝 기법을 통한 분류기법이 적극적으로 활용되고 있다(Lee, Byeon, Kim, 2022; Kim, Cho, Na, Chun, 2021). 본 연구에서는 이러한 이미지 분석이 아닌 형상분석에 대하여 진행하려고 한다. 형상분석(shaep analysis)은 기하적 공간상에서 형상점(landmark)들에 의해 나타낸 개체들(objects)의 형상(shape)을 측정 및 기술하며 이를 비교하는 분석이다(Dryden, Mardia, 2016). 여기서 형상은 어느 곳에서나 다양하게 존재하는 개체들을 표현할 때 주로 사용된다. 예시로는 병원의 x-ray 자료를 통해 병을 진단하거나, 단백질 구조의 형상을 통해 특정 단백질을 식별하는 과정 등이 있다. 혹은 3D로 존재하는 얼굴의 형상 데이터를 기반으로 병을 진단하는 연구 또한 이러한 예시가 된다(Kim, Kim, Lee, Lee, Chi, Kim, 2008; Kim, 2009). 기본적으로 형상점 자료는 배열의 형태를 가지고 있는데, 이

[1]46241 부산광역시 금정구 부산대학로 63번길 2, 부산대학교 통계학과 석사과정. E-mail: wnsghd4685@naver.com
[2](교신저자) 46241 부산광역시 금정구 부산대학로 63번길 2, 부산대학교 통계학과 교수.
E-mail: yschoi@pusan.ac.kr

는 그대로 분류모형에 적용하기 어려움이 있는 형태이다. 본 연구는 이러한 어려움을 가진 형상점 자료를 변환하는 방법에 대해 기존의 방법인 형상 특성변수의 변수화 방법과 새롭게 제시하는 방법인 좌표의 변수화 방법을 제시한다. 또한, 형상점이 많은 경우에 대하여 Convex Hull 알고리즘을 적용하여 형상점을 선택하여 수를 조절한다.

본 논문에서는 위의 제시된 방법들을 랜덤 포레스트, SVM, 로지스틱 회귀모형 등 세 가지의 분류모형에 적용하였다. 2절에서는 본 연구를 위해 제시하는 모형들 및 방법론들에 관하여 기술하였고, 3절에서는 이를 실제 형상점 자료에 투입하여 결과를 비교하였다.

2. 이론 설계

기본적으로 형상점 자료는 k개의 형상점과 m차원으로 구성된 $k \times m$ 형태행렬(configuration matrix)이 표본 개수 n개만큼 배열로서 합쳐진 자료형태이며, 이중 m은 형상 자료 특성상 2차원 혹은 3차원으로 나타내기 때문에 2 혹은 3의 값이다. 여기서 i번째 형태행렬 X_i에 대하여 아래와 같이 표현할 수 있다. 여기서 $x_{jl}^{(i)}$는 i번째 형태행렬의 j번째 형상점을 나타내는 행과 l번째 차원을 나타내는 열의 원소를 뜻한다.

$$X_i = \begin{pmatrix} x_{11}^{(i)} & x_{12}^{(i)} & \cdots & x_{1m}^{(i)} \\ \vdots & \vdots & \ddots & \vdots \\ x_{k1}^{(i)} & x_{k2}^{(i)} & \cdots & x_{km}^{(i)} \end{pmatrix}, \; i = 1, \cdots, n$$

다음과 같은 배열 형태의 형상점 자료들에 대하여 두 가지 방법을 통해 행렬 형태로 변형한다.

첫 번째 방법은 Choi(2021)가 제시한 방법으로, GPA 적합 후 형상특성변수들을 추출하여 이를 분류모형에 투입하는 방법이다. 여기서 GPA 적합은 형상표본들에 대하여 위치, 척도 및 회전 등 세 가지 요소를 제거하는 적합 방법으로, 이렇게 세 요소를 제거한 표본들에 대하여 각각 리만거리, 중심크기 및 형상 PCA를 통한 PC 점수를 추출한다.

두 번째 방법은 n개의 형태행렬을 벡터화하는 방법으로, 앞에서 제시한 i번째 형태행렬 X_i를 $1 \times km$ 크기의 벡터 Y_i로 벡터화한다. 이후 최종적으로 n개의 표본들을 포함하는 행렬 Y를 생성할 수 있으며, 이는 각각의 좌푯값들이 변수로 변환된 행렬로 재구성된 자료이다. 본 연구에서는 기본 자료와 중심화 자료에 대하여 위의 과정을 진행한다.

$$Y_i = \left(x_{11}^{(i)} \; \cdots \; x_{k1}^{(i)} | \; \cdots \; | \, x_{1m}^{(i)} \; \cdots \; x_{km}^{(i)} \right), \; Y = \left(Y_1 \cdots \, Y_i \cdots \, Y_n \right)^T, \; i = 1, \cdots, n$$

위의 형태행렬의 벡터화 방법은 모든 좌푯값을 변수로써 사용할 수 있다는 장점이 있지만, 형상점의 개수가 많아질 경우 너무 많은 변수가 생성될 수 있다는 점이다. 따라서 형상을 형성할 때 중요하다고 판단되는 형상점만을 선택하는 방법을 사용하여 형상점의 수를 줄일 것이다. 본 연구에서 제시하는 방법은 Convex Hull 알고리즘으로, 이를 통해 2차원 좌표 평면에서 형상을 구성하는 가장 바깥쪽의 형상점을 추출한다(Eddy, W. F, 1977). 본 연구의 전체적인 방향에 대한 흐름도는 Figure 1과 같다.

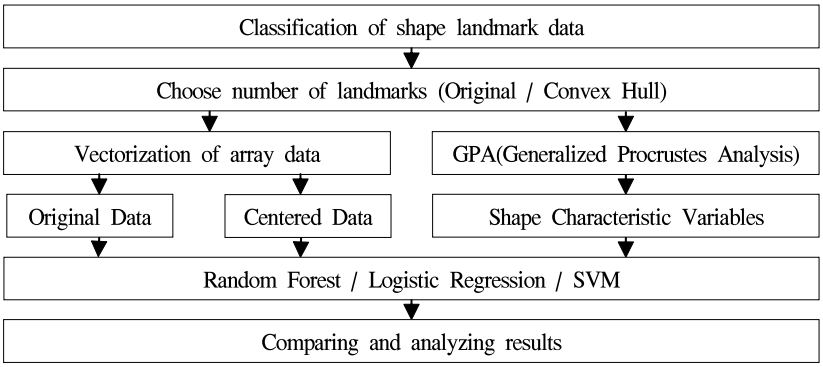


Figure 1. Analysis Procedure

3. 실증분석 및 해석

본 연구에서 사용하는 자료는 Ahn 외(2019)가 제시한 골격성 Class III 기형(하악이 나온 보통 주걱턱) 진단을 받은 성인 환자 100명을 대상으로 한 형상점 자료이다. 또한, 각각 40명의 여성 환자 군집(C_1)과 60명의 남성 환자 군집(C_2)으로 구성되어 있으며, CBCT(cone-beam computed tomography) 검사를 통해 200개의 3차원 형상점 자료로 구축되어 있다. 본 연구에서 사용하는 3D 자료에 대한 2차원 좌표평면 그림은 Figure 2와 같으며, 이중 위는 형상점이 200개, 아래는 Convex Hull 알고리즘을 통해 형상점의 수를 24개로 축소한 자료이다.

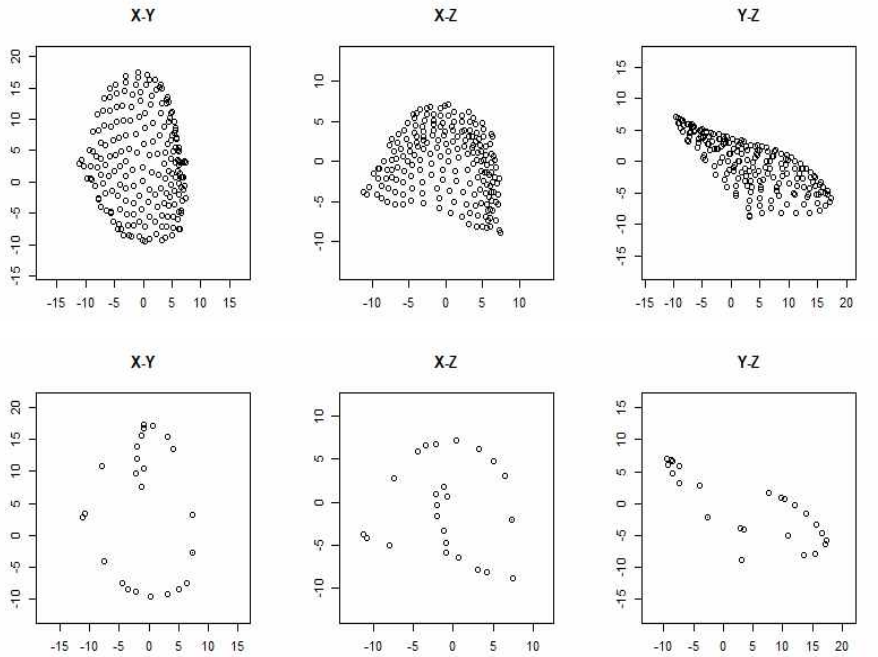


Figure 2. original shape data and revised data

Table 1. Reesults of classfication analysis (misclassification rate)

# of landmarks	Methods		Logistic Regression	Random Forest	SVM
Original (200)	Vectorization	Original Data	29.00%	32.00%	25.00%
		Centered Data	27.00%	36.00%	22.00%
	GPA	Shape Characteristic Variables	26.00%	22.00%	18.00%
Convex Hull (24)	Vectorization	Original Data	23.00%	34.00%	27.00%
		Centered Data	23.00%	26.00%	27.00%
	GPA	Shape Characteristic Variables	26.00%	27.00%	39.00%

Table 1은 Figure 1의 연구흐름도에 따른 분석 결과로, 각 방법들에 따른 오분류율을 나타낸 표이다. 전체적으로 로지스틱 회귀분석에서는 비슷한 결과를 보였으며 랜덤 포레스트와 SVM은 제법 편차가 나타났다. 원 데이터의 GPA를 통한 형상 특성변수의 SVM이 가장 좋은 결과를 보였으며, 반대로 Convex Hull 알고리즘을 통해 형상점을 감소시킨 경우에 대해서는 GPA를 통한 형상 특성변수의 SVM이 가장 좋지 못한 결과를 보였다. 또한, 전체적으로 SVM의 경우 원 자료에서 좋은 결과를 보인 반면, 랜덤 포레스트와 로지스틱 회귀분석의 경우 Convex Hull이 좋은 결과를 나타내었다.

References

Ahn, M. S., Shin, S. M., Choi, Y. S., Wu, T. J., Yamaguchi, T., Maki, K. and Kim, Y. I. (2019). Analysis of the relationship between the morphology of the palate and facial skeletal patterns in Class III malocclusion using structural equation modelling, *ORTHODONTICS & CRANIOFACIAL RESEARCH*, 22(2), 87-92.

Choi, Y. S. (2021), *Multivariate Statistical Shape Analysis with R*, Pusan National University Press.

Dryden, I.L. and Mardia, K.V. (2016). *Statistical Shape Analysis: with applications in R*, (Vol. 995). John Wiley & Sons.

Kim, J, K., Cho, W, H., Na, M, H., Jeon, M, H. (2019). Development of Automatic Classification System of Vegetables by Image Processing and Deep Learning. *Journal of The Korean Data Analysis Society*, 21(1), 63-73. (in Korean).

Kim, K. K.,, Kim, J, W., Lee, Y, T., Lee, I, S., Chi, G, Y., Kim, K, C. (2008), Study on the Statistical Methods for the 3D Facial Data in the HyungSang Clinical Medicine - Focused on the Dam(gall bladder) and BangKwang(urinary bladder) Body -. *Journal of The Korean Data Analysis Society*, 10(3), 1327-1337. (in Korean).

Kim, K.. K. (2009). Statistical Methods for Improving the Correct Rate in the HyungSang Clinical Medicine using the 3D Facial Data. *Journal of The Korean Data Analysis Society*, 11(6), 2981-2994. (in Korean).

Lee, D, C., Byeon, S, Y., Kim, K, H. (2022). An Inspection of CNN Model for Citrus Canker Image Classification Based on XAI: Grad-CAM. *Journal of The Korean Data Analysis Society*, 24(6), 2133-2142. (in Korean).

William F. Eddy. (1977). A New Convex Hull Algorithm for Planar Sets. *ACM Trans. Math. Softw.* 3(4), 398 – 403. https://doi.org/10.1145/355759.355766

Proceedings of the Korean Data Analysis Society
January 26-27, 2023, pp. 55-58

표본 자료의 분위수 계산에 관한 고찰: 몇 가지 연속형 분포의 경우

김혁주[1]

요 약

표본 자료가 있을 때 이 자료의 분위수를 계산함으로써 모집단의 분위수를 추정한다. 통계학 교재에서 주로 소개하는 표본 분위수 계산 방법은 두 가지가 있다. 본 연구에서는 몇 가지 연속형 분포의 경우에 표본 분위수를 계산하는 두 가지 방법을 모의실험을 통하여 비교하였다. 균일분포, 정규분포, 지수분포, 카이제곱분포, 베타분포의 경우에 대해 모수의 값을 몇 가지로 설정하고, 설정된 각각의 분포에 대해 크기 20, 30, 40, 50인 확률표본을 추출하는 모의실험을 10,000회씩 실시하였다. 각 회의 모의실험에서는 두 가지 방법 각각에 의하여 계산된 표본 분위수와 모집단 분위수의 차이를 계산하였다. 두 방법을 비교하는 기준으로 두 가지를 사용하였는데, 한 가지는 10,000회의 모의실험에 걸친 평균제곱오차이고, 다른 한 가지는 10,000회의 모의실험 중 모집단 분위수와의 차이가 상대 방법보다 더 작은 횟수이다.

주요용어 : 모의실험, 분위수, 사분위수, 십분위수, 평균제곱오차.

1. 서론

표본 자료(이하 '자료'라 칭한다)가 있을 때 이 자료의 분위수를 계산함으로써 모집단의 분위수를 추정하게 된다. 모집단 분위수의 개념은 한 가지이지만, 자료로부터 실제로 표본 분위수를 계산하는 방법은 한 가지로 통일되어 있지 않다. 통계학 교재에서 주로 소개하는 표본 분위수 계산 방법은 두 가지가 있다. 본 연구에서는 몇 가지 연속형 분포의 경우에 표본 분위수를 계산하는 두 가지 방법을 모의실험을 통하여 비교하고자 한다. 두 방법을 방법 1과 방법 2로 표기한다.

2. 표본 분위수를 계산하는 두 가지 방법

백분위수의 경우를 통하여 두 방법을 설명한다. 먼저 방법 1은 Bhattacharyya와 Johnson (1977), Kim 등(2006)에서 사용하는 방법이다. 여기에는 자료의 백분위수가 다음과 같이 정의되어 있다. 자료를 작은 값부터 큰 값까지 순서대로 늘어놓았을 때 적어도 p%의 관측값이 그 값보다 작거나 같으며, 동시에 적어도 $(100-p)$%의 관측값이 그 값보다 크거나 같게 되는 값(이 값이 유일하게 결정되지 않을 때는 그 값들의 평균을 사용함. 여기서 평균은 단순평균을 의미)을 제p백분위수라고 정의하였다. Walpole (1982)과 Kim 등(2002)에서도 이 방법으로 백분위수를 계산하고 있다.

제1(Q_1), 제2(Q_2), 제3사분위수(Q_3)는 각각 제25, 제50, 제75백분위수와 같은 것이며, 제1(D_1), 제

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

2(D_2), …, 제9십분위수(D_9)는 각각 제10, 제20, …, 제90백분위수와 같은 것이다. Q_2는 중앙값(median)이라고도 한다. 본 논문에서 모집단 사분위수는 Q_1, Q_2, Q_3로, 표본 사분위수는 $\widehat{Q}_1$, $\widehat{Q}_2$, $\widehat{Q}_3$로 나타내겠으며, 십분위수도 같은 방식으로 나타내겠다.

$y_1, y_2, \cdots, y_n$을 자료의 순서통계량값들이라 하자. 방법 2는 Hogg 등(2019)과 Baek 등(2021)에서 사용하는 방법으로, $(n+1)p$가 정수이면 $y_{(n+1)p}$를 자료의 제p백분위수로 하고, 정수가 아니면 $y_{[(n+1)p]}$와 $y_{[(n+1)p]+1}$의 가중평균을 자료의 제p백분위수로 하는 방법이다. [ ]는 가우스 기호이다. 예를 들어 $n=20$인 경우와 $n=50$인 경우 방법 1과 방법 2에 의한 사분위수와 십분위수 계산은 Table 1, Table 2와 같다. 중앙값은 두 방법이 항상 같은 결과를 주므로 생략하였다.

Table 1. Quartiles and deciles of sample data of size $n=20$

	$\widehat{Q}_1$	$\widehat{Q}_3$	$\widehat{D}_1$	$\widehat{D}_2$	$\widehat{D}_3$	$\widehat{D}_4$	$\widehat{D}_6$	$\widehat{D}_7$	$\widehat{D}_8$	$\widehat{D}_9$
Method 1	$\frac{y_5+y_6}{2}$	$\frac{y_{15}+y_{16}}{2}$	$\frac{y_2+y_3}{2}$	$\frac{y_4+y_5}{2}$	$\frac{y_6+y_7}{2}$	$\frac{y_8+y_9}{2}$	$\frac{y_{12}+y_{13}}{2}$	$\frac{y_{14}+y_{15}}{2}$	$\frac{y_{16}+y_{17}}{2}$	$\frac{y_{18}+y_{19}}{2}$
Method 2	$\frac{3y_5+y_6}{4}$	$\frac{y_{15}+3y_{1}}{4}$	$\frac{9y_2+y_3}{10}$	$\frac{8y_4+2y_5}{10}$	$\frac{7y_6+3y_7}{10}$	$\frac{6y_8+4y_9}{10}$	$\frac{4y_{12}+6y}{10}$	$\frac{3y_{14}+7y}{10}$	$\frac{2y_{16}+8y}{10}$	$\frac{y_{18}+9y_{1}}{10}$

Table 2. Quartiles and deciles of sample data of size $n=50$

	$\widehat{Q}_1$	$\widehat{Q}_3$	$\widehat{D}_1$	$\widehat{D}_2$	$\widehat{D}_3$	$\widehat{D}_4$	$\widehat{D}_6$	$\widehat{D}_7$	$\widehat{D}_8$	$\widehat{D}_9$
Method 1	y_{13}	y_{38}	$\frac{y_5+y_6}{2}$	$\frac{y_{10}+y_{11}}{2}$	$\frac{y_{15}+y_{16}}{2}$	$\frac{y_{20}+y_{21}}{2}$	$\frac{y_{30}+y_{31}}{2}$	$\frac{y_{35}+y_{36}}{2}$	$\frac{y_{40}+y_{41}}{2}$	$\frac{y_{45}+y_{46}}{2}$
Method 2	$\frac{y_{12}+3y_{1}}{4}$	$\frac{3y_{38}+y_{3}}{4}$	$\frac{9y_5+y_6}{10}$	$\frac{8y_{10}+2y}{10}$	$\frac{7y_{15}+3y}{10}$	$\frac{6y_{20}+4y}{10}$	$\frac{4y_{30}+6y}{10}$	$\frac{3y_{35}+7y}{10}$	$\frac{2y_{40}+8y}{10}$	$\frac{y_{45}+9y_{4}}{10}$

3. 모의실험을 통한 비교

연속형 분포인 균일분포, 정규분포, 지수분포, 카이제곱분포, 베타분포의 경우에 대해 모수의 값을 몇 가지로 설정하고, 설정된 각각의 분포에 대해 크기 20, 30, 40, 50인 확률표본을 추출하는 모의실험을 10,000회씩 실시하였다. 각 회의 모의실험에서는 방법 1과 방법 2 각각에 의하여 계산된 표본 분위수와 모집단 분위수의 차이를 계산하였다. 두 방법을 비교하는 기준으로 두 가지를 사용하였는데, 한 가지는 10,000회의 모의실험에 걸친 평균제곱오차(이하 기준 1이라 칭한다)이고, 다른 한 가지는 10,000회의 모의실험 중 모집단 분위수와의 차이가 상대 방법보다 더 작은 횟수(이하 기준 2라 칭한다)이다. 모의실험은 Minitab ver. 19를 사용하여 실시하였다.

Table 3부터 Table 7까지는 모의실험 결과를 나타낸 표이다. 각 칸에 있는 숫자는 기준 1의 값이며, 괄호 안의 숫자는 기준 2의 값이다. 예를 들어 $U(0,1)$의 경우를 나타낸 Table 3을 보면, 모집단 제1사분위수는 $Q_1=0.25$인데, $n=20$인 경우 기준 1의 값은 방법 1이 0.0084, 방법 2가 0.0082로 방법 2가 약간 우세하지만, 기준 2의 값은 방법 1이 5,118회, 방법 2가 4,882회로 나와 방법 1이 더 우세한 것으로 나타났다(p-값=0.009). 이것은 균일분포의 경우 인접한 두 순서통계량의 단순평균으로 제1사분위수를 추정하는 것이 가중평균으로 추정하는 것에 비해 모수에 가까운 값을 얻는 빈도

는 높았지만, 모수에서 먼 값을 얻을 때는 방법 2에 비해 차이가 상대적으로 컸다는 의미로 볼 수 있겠다. 분포의 중심에서 먼 D_1, D_2, D_8, D_9의 경우에도 유사한 현상이 나타났고, 분포의 중심에 가까운 분위수들의 경우에는 두 방법의 차이가 크지 않았다. $n=20$인 경우만 보였지만, $n=30$, $n=40$, $n=50$인 경우도 크게 다르지 않은 양상으로 나왔다.

Table 4, 5, 6, 7은 각각 정규분포 $N(0,1)$ $(n=30)$, 지수분포 $\mathrm{Exp}(3)$ $(n=20)$, 카이제곱분포 $\chi^2(4)$ $(n=20)$, 베타분포 $Be(4,4)$ $(n=30)$의 경우 10,000회씩의 모의실험 결과를 나타낸 표이다. 대체적으로 기준 1에서는 분포에 따라, 그리고 몇 분위수냐에 따라 방법 1이 우세한 경우도 있고 방법 2가 우세한 경우도 있었지만, 기준 2에서는 대부분의 경우 방법 1이 우세하게 나왔다.

Table 4에서 주목할 점은 다음과 같다. 정규분포이고 $n=30$인 경우 모든 십분위수에서 방법 1이 방법 2보다 우세하게 나왔다. 분포의 중심에서 먼 십분위수의 경우에는 확연한 우위를 보였고, 분포의 중심에 가까운 십분위수의 경우에도 미세하나마 우위를 보였다. 그런데 두 사분위수의 경우에는 기준 1에서 방법 2가 우위를 보였다. 그 이유는 다음과 같이 생각된다. 표본 크기 n이 10의 짝수배인 경우에는 방법 1과 방법 2 공히 모든 표본 사분위수와 표본 십분위수가 인접한 두 순서통계량의 단순평균 또는 가중평균으로 이루어진다. 반면 n이 10의 홀수배(30, 50 등)인 경우에는 방법 1에 의한 표본 사분위수가 두 순서통계량의 평균이 아닌 하나의 순서통계량으로 이루어진다. 정규분포와 같은 대칭 분포의 경우에 이것이 평균제곱오차를 크게 하는 요인이 되는 것으로 추측된다. $n=50$인 경우에도 같은 현상이 나타났으며, 역시 대칭 분포인 $Be(4,4)$의 경우(Table 7 참조)에도 유사한 현상이 관찰되었다(소수점 이하 4자리로 반올림하면 두 방법이 모두 0.0018로 표시되었지만 방법 2의 평균제곱오차가 미세하게 더 작았다).

Table 3. Simulation result for $U(0,1)$ $(n=20)$

	$Q_1=0.2$	$Q_3=0.7$	$D_1=0.1$	$D_2=0.2$	$D_3=0.3$	$D_4=0.4$	$D_6=0.6$	$D_7=0.7$	$D_8=0.8$	$D_9=0.9$
Method 1	.0084 (5118)	.0084 (5021)	.0047 (5184)	.0074 (5152)	.0092 (5086)	.0103 (5006)	.0105 (5045)	.0092 (5017)	.0073 (5062)	.0046 (5050)
Method 2	.0082 (4882)	.0081 (4979)	.0040 (4816)	.0070 (4848)	.0091 (4914)	.0102 (4994)	.0105 (4955)	.0090 (4983)	.0069 (4938)	.0039 (4950)

Table 4. Simulation result for $N(0,1)$ $(n=30)$

	$Q_1=-0.$	$Q_3=0.67$	$D_1=-1.$	$D_2=-0.$	$D_3=-0.$	$D_4=-0.$	$D_6=0.2$	$D_7=0.5$	$D_8=0.8$	$D_9=1.2$
Method 1	.0608 (4995)	.0621 (5160)	.0856 (5239)	.0640 (5081)	.0546 (5025)	.0513 (5007)	.0515 (5085)	.0554 (5169)	.0642 (5112)	.0882 (5222)
Method 2	.0599 (5005)	.0614 (4840)	.0997 (4761)	.0674 (4919)	.0554 (4975)	.0514 (4993)	.0517 (4915)	.0564 (4831)	.0673 (4888)	.1036 (4778)

Table 5. Simulation result for $\mathrm{Exp}(3)$ $(n=20)$

	$Q_1=0.86$	$Q_3=4.15$	$D_1=0.31$	$D_2=0.66$	$D_3=1.07$	$D_4=1.53$	$D_6=2.7$	$D_7=3.6$	$D_8=4.8$	$D_9=6.9$
Method 1	.1650 (5032)	1.273 (5201)	.0620 (5095)	.1259 (4990)	.2063 (4958)	.3076 (4926)	.6551 (5110)	1.000 (5201)	1.670 (5227)	3.638 (5475)
Method 2	.1521 (4968)	1.423 (4799)	.0491 (4905)	.1114 (5010)	.1933 (5042)	.2986 (5074)	.6773 (4890)	1.089 (4799)	1.982 (4773)	5.373 (4525)

Table 6. Simulation result for $\chi^2(4)$ $(n=20)$

	$Q_1=1.9$	$Q_3=5.3$	$D_1=1.0$	$D_2=1.6$	$D_3=2.1$	$D_4=2.7$	$D_6=4.0$	$D_7=4.8$	$D_8=5.9$	$D_9=7.7$
Method 1	.2712 (5027)	1.062 (5311)	.1751 (5250)	.2376 (5122)	.3077 (4990)	.3838 (4935)	.6394 (5131)	.8750 (5279)	1.334 (5356)	2.609 (5475)
Method 2	.2607 (4973)	1.180 (4689)	.1679 (4750)	.2298 (4878)	.2989 (5010)	.3772 (5065)	.6562 (4869)	.9403 (4721)	1.567 (4644)	3.767 (4525)

Table 7. Simulation result for $Be(4,4)$ $(n=30)$

	$Q_1=0.3$	$Q_3=0.6$	$D_1=0.2$	$D_2=0.3$	$D_3=0.4$	$D_4=0.4$	$D_6=0.54$	$D_7=0.59$	$D_8=0.65$	$D_9=0.72$
Method 1	.0018 (5144)	.0018 (5050)	.0020 (5342)	.0018 (5228)	.0017 (5104)	.0016 (5075)	.0017 (4983)	.0017 (5053)	.0018 (5113)	.0021 (5263)
Method 2	.0018 (4856)	.0018 (4950)	.0023 (4658)	.0018 (4772)	.0017 (4896)	.0016 (4925)	.0017 (5017)	.0018 (4947)	.0019 (4887)	.0023 (4737)

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., 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, 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 26-27, 2023, pp. 59-59

설명 가능한 인공지능(XAI) 기반 토지 피복 구성에 따른 지표면 온도의 영향분석

이동찬[1], 변상영[2], 김기환[3]

요 약

최근 인공지능 아키텍처의 발전과 하드웨어의 성능이 급속도로 좋아지면서 다양한 인공지능 모델들이 주목을 받고 그 유용성이 증가하고 있다. 그러나 성능의 발전 속도와는 달리, 모델의 설명 능력은 진척이 더딘 실정이다. 갈수록 복잡해지고 있는 머신러닝(machine learning) 모델의 의사결정 분기점은 기하급수적으로 늘어나고 있으며, 이는 연구자의 모델 해석 용이성을 저해시키는 원인이 된다. 설명 가능한 인공지능(eXplainable Artificial Intelligence, XAI)은 위와 같은 복잡성을 해소하기 위해 등장하였으며, 모델의 블랙박스를 연구자가 이해 가능한 수준으로 분해하여 해석 가능성 및 신뢰도 제고에 도움을 준다. 본 연구는 서울특별시 셀(cell) 단위의 토지 피복 데이터를 이용하여 특정 공간 단위의 지표면 온도를 추정한다. 고차원 데이터에 대한 기존 통계모형의 한계를 살펴보았으며, 머신러닝 모델을 이용하여 추정 결과를 비교하였다. 사용한 모형으로는 Lasso(Least absolute shrinkage and selection operator) 회귀, 랜덤포레스트 회귀(random forest regression), XGBoost(eXtreme Gradient Boosting)가 있으며, XGBoost 모델의 적합 결과를 바탕으로 SHAP(SHapley Additive exPlanation)을 통한 XAI 점검을 시행하였다.

주요용어 : 머신러닝, 설명 가능한 인공지능, SHAP, XGBoost.

[1]30019 세종특별자치시 세종로 2511, 고려대학교 일반대학원 경제통계학과 국가통계전공 박사과정. E-mail: tozjs72@korea.ac.kr
[2]30019 세종특별자치시 세종로 2511, 고려대학교 일반대학원 경제통계학과 국가통계전공 박사과정. E-mail: wooaaaaa@korea.ac.kr
[3]30019 세종특별자치시 세종로 2511, 고려대학교 경제통계학부 국가통계전공 교수. E-mail: korpen@korea.ac.kr

Proceedings of the Korean Data Analysis Society
January 26-27, 2023, pp. 61-66

XAI 기법을 이용한 보이스피싱 이상거래 탐지 모델 연구*

김주미[1], 정여진[2]

요 약

최근 전자통신수단을 이용한 피싱(Phising)사기가 급증함에 따라 금융대출 피해액이 증가하고 있다. 이에 금융사에서는 FDS 시스템을 통해 신종 사기 수법을 빠르게 파악하고 사전에 예방 할 수 있는 방법을 모색하기 위해 축적된 고객 관련 각종 데이터를 활용하려는 시도를 강화하고 있다. 기존의 시도들은 주로 사기적발 예측에만 중점을 두었으나 이는 피싱 범죄의 전후 과정을 해석하는데 어려움이 존재하는 것은 물론, 모형의 2종 오류인 정상 고객의 사기 의심으로 대출 거절 발생 시 고객에게 제대로 설명하지 못하는 문제점도 발생했다. 본 연구에서는 대표적 불균형 데이터인 피싱사기 데이터의 효과적 학습을 위하여 오버샘플링을 진행하였으며 XGBoost 기법을 적용하여 모형을 구현하였을 때 F1 Score가 0.8로 가장 높은 ADASYN 샘플링 기법을 선택하였다. 최종적으로 피싱사기 예측에 영향을 미치는 주요 요소들을 파악하기 위하여 대표적인 설명가능한 인공지능(XAI) 기법 중 하나인 SHAP를 활용하여 예측 모형을 해석하였고 이를 통해 요인들의 모델에 대한 전반적인 영향도 뿐만 아니라, 각 개별 고객에 대해서도 어떤 요소가 사기 예측 결정에 영향을 미쳤는지 파악할 수 있었다.

주요용어 : 피싱(Phishing), FDS(Fraud Detect System), XGBoost(eXtreme Gradient Boosting), SHAP(SHapley Addictive exPlanation), 설명가능한 인공지능(eXplainable AI),

1. 서론

피싱(Phising) 사기란 개인정보(Private Data)와 낚시(Fishing)를 합성한 신조어로 전자통신수단을 이용한 비대면거래를 통해 타인의 재산을 편취 하는 금융 분야에서 발생하는 일종의 특수 사기 범죄를 뜻한다. 주요 사기유형으로는 지인을 사칭하여 메신저상에서 송금을 요구하거나 인터넷뱅킹 혹은 신용카드 정보 취득 후 ARS를 통한 카드론 편취, 금융회사나 금감원 명의의 허위 문자메시지로 기망 한 후 피싱사이트로 유도하고 예금을 가로채는 등 그 수법도 날로 다양해지고 있다.

금융감독원 보도자료에서 21년 피싱 피해 금액은 총 1,682억원으로 나타났으며 코로나19로 인한 사기 활동 위축으로 전년비 감소(△28.5%)하였으나 백신접종이나 재난지원금 지원 등 국민적 관심도가 높은 주제로 개인정보 입력을 유도하는 신종 사기 수법도 발생하고 있어 지속적인 피해 예방 노력이 필요하다. 이에 이상금융거래를 탐지하는 금융회사 FDS(Fraud Detect System)는 AI기술을 통해 사기 거래패턴을 분석하여 의심거래를 정확히 포착하여 대응력이 제고될 수 있도록 고도화가

*본 논문은 한국대학교의 지원을 받아서 연구된 것임(2021a12345-7).
[1]12345 대한민국 홍길동시 미래4길, 한국대학교 미래학과 교수. E-mail: abc1234@ccc.ac.kr
[2](교신저자) 12345 대한민국 홍길동시 미래4길, 한국대학교 미래학과 부교수. E-mail: def5678@ccc.ac.kr

진행되고 있다. 금융산업은 특히 데이터 분석 방법과 머신러닝 알고리즘이 고도화됨에 따라 예측력 향상을 위한 다양한 시도가 진행되었다(Zhen, 2019). 주요 내용으로는 부스팅 기반의 머신러닝 모델이나 딥러닝 같은 복잡한 모형을 적용하는 방법들이 존재한다(Zheng 등, 2018). 하지만 인공지능 기반 모델이 복잡해질수록 모델의 내부를 알 수 없다는 단점이 존재한다(Qui 와 Choi, 2019, Jang 등, 2020). 따라서 머신러닝 모델 내부의 알고리즘을 통해 의사 결정되는 과정을 정확하게 확인할 수 없고 모델을 통해 도출된 결과를 완전히 신뢰하는 일은 쉽지 않다. 특히 정상적인 예금인출이나 카드론 요청마저도 FDS에 의해 이상거래로 탐지되어 지급이 거부 될 경우 고객에게 불편을 주게 되고 거절 사유를 제대로 설명하지 못한다면 그 결과 금융기관의 신뢰도는 떨어지게 될 것이다. 따라서 어떤 요인들에 의해 모형이 작동되었는지를 확인하고 살펴볼 필요가 있다.

이에 본 논문에서는 피싱사기 예측 모델의 성능 개선 연구에 그치지 않고 설명가능한 인공지능(eXplainable AI)기법을 적용해 피싱사기를 예측하는데 영향을 주는 주요 요인은 무엇인지 확인하고 결과 피드백을 통하여 최종적으로 개선된 모형을 생성하고자 한다.

2. 연구 방법론

2.1 방법론

2.1.1 XGBoost(eXtreme Gradient Boosting)

부스팅 알고리즘은 여러 개의 저성능 분류기를 결합하여 더욱 고성능을 내도록 만드는 기법이다. XGBoost는 이런 부스팅 알고리즘의 하나로 캐글(Kaggle) 등 다양한 데이터 분석 대회에서 뛰어난 성적을 나타내고 있으며 수치 데이터 예측 모델로 다양하게 활용되고 있다. XGBoost는 새로운 분류 트리를 만들어 내기 위해 탐욕 알고리즘(greedy algorithm)을 사용하여 내부에 생성된 다양한 모델들의 성능을 보완하는 가중치를 탐색한다. 이러한 가중치는 CART라 불리는 앙상블 모델을 사용하며 모든 최종 노드들이 최종 스코어를 계산하는데 사용된다. XGBoost는 이렇게 구해진 분류기들과 가중치를 이용하여 더욱 정확한 예측을 가능하게 하며 내부 하위 모델의 앙상블을 통해서 과적합 문제를 해결할 수 있고 동시에 병렬 처리를 통해서 학습 시간을 단축할 수 있다는 장점이 있다.

2.1.2 SHAP(SHapley Additive exPlanations)

SHAP 기법은 게임 이론에 등장하는 Shapley Value를 이용하는 기법이다. Shapley Value는 전체 성과를 만들어 내는 데 변수별로 어느 정도 기여했는지를 수치로 나타낸 값이다. 여기서 특정 변수의 기여도는 모든 변수를 조합한 경우 계산된 성과에서 해당 변수를 제외한 나머지 변수들을 조합한 경우 계산되는 성과를 빼 차이를 계산함으로써 측정할 수 있다. 즉, 전체 성과가 특정 특성을 제외하면 그 결과가 어떻게 변화하는지 정도를 계산해 내어 해당 특성의 기여도로 계산하는 방법이다.

SHAP 기법은 예측 모델에 사용되는 특성들이 예측에 미치는 평균 영향도를 계산해 준다. 이때 SHAP는 양의 값뿐만 아니라 음의 값을 가질 수도 있으며 이는 해당 특성이 모델 예측에 부(-)의 영향을 미쳤다고 판단할 수 있다. 또한 SHAP 기법은 지역적인 설명(local interpretation)도 가능하다.

각각의 사례에 대해 어떤 특성이 예측에 긍정적 또는 부정적인 영향을 주었는지 나타내 줌으로써 사례별로 예측 결과에 대한 설명이 가능하다는 장점이 있다.

3. 실증분석 결과

3.1 분석 자료

본 연구에서 사용한 자료는 개인 식별키가 삭제 처리된 국내 S 신용카드사의 카드론 대출 데이터이며 분석 기간은 2022년 1월부터 2022년 10월까지이다. 기간 중 데이터는 총 1,054,452건 이며 고객 확인을 통해 확정된 551건의 사기 데이터가 포함되어 있다. 사기 예측에 유의하다고 판단되는 총 71개의 변수를 사용하였으며 범주형데이터는 One Hot 인코딩, 수치형 데이터는 정규화 하여 사용하였다.

Table 1. 모형 구축에 사용한 변수 카테고리

Classification	변수 수	변수 비중(%)	Feature
승인정보	2	3	승인금액, 신청경로
원장정보	2	3	신용등급, 나이, 성별
회원/카드 프로파일	36	51	사용패턴 변수 가공
외부신용정보	30	42	대출기관수, 대출 잔액 등
온라인접속정보	1	1	기존 사용PC/모바일기기 여부 등

3.2 모델 정확도

451,909건의 Test 데이터를 제외하고 비대칭 데이터의 단점을 보완할 수 있는 Oversampling 기법을 통해 Train 데이터를 구성하여 데이터 비대칭에 대한 한계를 해결하였다. Table 2는 XGBoost 방법론을 활용하여 각기 다른 샘플링 방법으로 구성된 Train 데이터에 학습 한 후, 테스트 데이터에 성능평가 해 본 결과이다. SMOTE 방법론을 사용한 샘플에서 F1 Score가 0.80으로가장 높게 나타나 피싱사고 예측 모델로 확정하였다.

Table 2. 샘플링 방법에 따른 XGBoost 성능 비교

Method	Training Set		Test Set		Detection Performance			
	Non Fraud	Fraud	Non Fraud	Fraud	Auc	Pre-cision	Recall	F1 Score
ADASYN	560,159	140,124	451,909	177	0.99	0.77	0.81	0.79
SMOTE	559,982	139,989	451,909	177	0.99	0.78	0.82	0.80
Random OVER	559,982	139,989	451,909	177	0.99	0.64	0.81	0.71

3.2 모델 해석

Figure 1은 SHAP 기법을 적용하여 얻은 피싱사기 적발 모형의 요인별 영향력을 나타낸 그래프이다. 그래프를 살펴보면 카드론 개설 가능 지수((KCB_CDLN_OP_PSB_IDXT_SCORE)가 가장 큰 영향을 주는 변수임을 확인함과 동시에 이 값이 낮을수록 사기에 의한 카드론 대출로 간주될 가능성이 높아 왼쪽 영역에서 붉은색으로 나타났다. 이 변수는 카드론을 필요로 할 가능성이 높은 사

람에게 큰 값을 부여하는 변수라는 점을 감안할 때 금리가 비싼 카드론을 굳이 필요도 없는 사람이 신청했다는 점에서 피싱사기 예측에 양의 영향을 미치는 것으로 판단된다. 또한 승인금액(APR_AM)이 클수록, 금융신청을 ARS가 아닌 경로로 했을수록(APLC_PH|8), 1시간 이내 현금서비스 승인금액 합계(AH_CARD_NO1_DLNG_APR_AM)가 클수록, 최근 6개월내 현금서비스 이용건수 대비 최근 7일 이용 건수 비율(MMS6_CARD_NO2_TOBZ_D7_CT_RT)이 낮을수록, 금융신청을 모바일앱으로 하지 않은 경우(APLC_PH|H) 피싱사기로 예측될 가능성이 높은 것으로 나타났다. 또한 신용스코어(KCB_CRD_SCORE)가 높을수록, 직업군이 VIP 혹은 우량직장이나 안정직업(CST_JOB_CLSF_C|9)일수록 카드론을 사용하는 것이 일반적이지 않으므로 피싱사기로 예측될 가능성이 높게 예측되는 것을 확인할 수 있었다.

또한 SHAP 기법은 Figure. 2 처럼 개별 사례에 대해서도 왜 그런 결정을 했는지 변수의 영향력을 출력해 준다. 이 고객의 경우 카드론 개설 가능 지수가 높았던 것은 피싱사기 가능성이 낮은 것으로, 요청한 카드론의 금액이 큰 것은 피싱사기 가능성을 높게 평가한 것으로 나타났다.

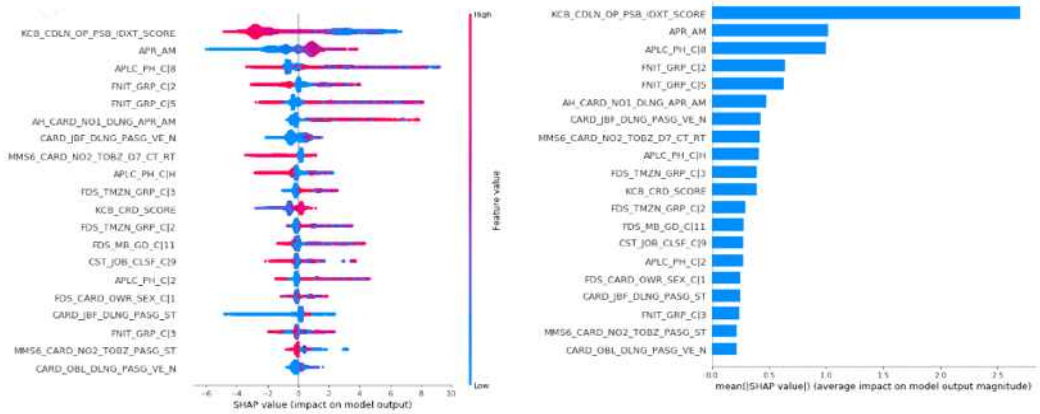


Figure 1. SHAP Summary Plot

Figure 2. Feature Influence for a sample instance

3.3 모델 재구축

위의 결과로 판단해 볼 때 예측기여도가 높은 몇 개의 변수만으로도 알고리즘을 설명하는데 무리가 없다면 처음부터 해당 변수들만으로 예측할 수 있는 모형을 만드는 것도 하나의 XAI 방법론 적용의 좋은 피드백의 예가 될 수 있다. 피싱사기를 예측하기 위해 고려했던 변수는 모두 71개 였으며 이를 통해 얻은 최적 결과는 F1 Score 0.80이었다. Table 3은 XAI를 통해 인사이트를 얻은 변수 20개만으로 다시 XGBoost 모델을 구축하여 기존의 성능을 비교한 표이다. F1 Score 기준 0.76은 기존 최적 성능을 보인 모형이 71개의 변수를 모두 활용했다는 점을 감안하여 볼 때 별도의 하이퍼파라미터의 최적 조합을 찾는 노력 없이 예측력을 높일 수 있음을 확인할 수 있었다.

Table 3. XGBoost 성능 비교

Model	Variable	Training Set		Test Set		Detection Performance			
		Non Fraud	Fraud	Non Fraud	Fraud	Auc	Pre-cision	Re-call	F1 Score
XGBoost	71개	560,159	140,124	451,909	177	0.99	0.78	0.82	0.80
XAI-XGBoost	20개	559,982	139,989	451,909	177	0.99	0.72	0.81	0.76

4. 결론 및 향후 연구

본 연구는 학술적으로 실제 피싱 데이터에 대해서 예측 모형을 만들고 XAI 기법을 적용했다는 점에서 의의가 있다. 특히 SHAP기법을 적용함으로써 특성들의 전반적인 영향력뿐만 아니라 각각의 사례에 대해서도 왜 그러한 예측 결과를 산출했는지에 관해 설명할 수 있다는 것을 확인했다. 기존 방법론들은 모델의 설명력을 위해서 의사결정나무나 회귀분석 등을 사용하여 예측력이 떨어지는 모델을 사용하곤 했으나, 본 연구에서는 정형데이터에서 가장 높은 성능을 가지는 XGBoost를 적용하면서도 XAI 기법을 통해 설명력을 부여할 수 있었다는 점에서 의의를 갖는다.

향후 본 연구에서 제안한 기법을 발전시켜 XAI를 통한 변수 선택 방법론 등을 추가로 연구하여 가장 큰 영향력을 가지는 소수의 변수 조합만으로도 더욱 정확한 예측력을 가지는 설명가능한 인공지능 모델을 구성할 계획이다.

References

Lundberg, S. M. Lee S. (2017) A Unified Approach to Interpreting Model Predictions, In Advanceds Conference on Neural Information Processing Systems, 4765-4774, *Curran Associates*, Inc.

Ge, Dingling, et al. (2020). Credit card fraud detection using lightgbm model. International Conference on E-Commerce and Internet Technology(ECIT), pp. 232-236, *IEEE*

Qiu, X. and Choi, P. (2019). A study on discrimination in mortgage lending in the United States: A revisit by random forest method. *Journal of the Korean Data & Information Science Society*, 30, 261-270.

Zhen, W. (2019). Machine learning applications in finance: Some case studies, Ph. D.'s thesis, Department of Computing, Imperial College London.

Zheng, X., Zhu, M., Li, Q., Chen, C. and Tan, Y. (2019). FinBrain: When finance meets AI 2.0. *Frontiers of Information Technology & Electronic Engineering*, 20, 914-924

Bang, S. Jaeoh, Kim, J. (2020). Sampling Method Using Gaussian Mixture Clustering for Classification Analysis of Imbalanced Data, *Journal of the Korean Data Analysis Society*, 22(2), 565-574. (in Korean). https://doi.org/10.37727/jkdas.2020.22.2.565

Cho, Y. Cho, D. Choi, B. (2022). Applications of the classification algorithm for unbalanced time series data:Focusing on the corporate default model, *Journal of the Korean Data Analysis Society*, 24(2), 639-651. (in Korean). https://doi.org/10.37727/jkdas.2022.24.2.639

Choi, S. Lee, S. Kim, M. (2022). Analysis of the degree of damage according to typhoon information using the influential sphere, *Journal of the Korean Data Analysis Society*, 24(3), 983-993. (in Korean). https://doi.org/10.37727/jkdas.2022.24.3.983

Kim, H. (2020). The Prediction of PM2.5 in Seoul through XGBoost ensemble, *Journal of the Korean Data Analysis Society*, 22(4), 1661-1671. (in Korean). https://doi.org/10.37727/jkdas.2020.22.4.1661

Study on Phishing Fraud Detection Model Using XAI

Ju Mi Kim[1], *Yeo Jin Chung*[2]

Abstract

With the rapidly growing number of phishing scams using electronic communication devices, the amount of financial loan losses has grown steadily. Therefore, financial institutions are increasing their efforts to utilize accumulated customer-related data in order to quickly identify new fraud methods through the FDS (Fraud Detection System) and find ways to prevent in advance. Existing attempts have mainly focused on predicting fraud detection and there are difficulties in not only interpreting pre- and post-process of phishing fraud, but also explaining to customers properly when the loans are declined due to suspicion of fraud, which is a type II error. In this study, oversampling is used to train phishing dataset that represents imbalance data effectively, and ADASYN sampling method, which has the highest F1 score of 0.8 when XGBoost is proposed in order to develop a more accurate prediction model, is selected. Finally, the prediction model is analyzed using SHAP (SHapley Additive exPlanation), which is one of the popular Explainable AI (XAI) techniques in order to identify the major factors affecting the prediction of phishing fraud and through this, it is also possible to identify which factors influenced the fraud prediction decision for each individual customer, as well as the overall impact of the factors on the model.

Keywords : Phishing, FDS, XGBoost SHAP, eXplainable AI.

[1]Ph. D.student Department of Data Science, Kookmin University, Jeongneung-ro77, Seoul, Korea. E-mail: jumi.kim@kookmin.ac.kr

[2](Corresponding Author) Professor, Department of Data Science, Kookmin University, Jeongneung-ro77, Seoul, Korea. E-mail: ychung@kookmin.ac.kr

Proceedings of the Korean Data Analysis Society
January 26-27, 2023, pp. 67-67

앙상블 비모수적 변수 선택 방법

신웅섭[1], **조형준**[2]

요 약

앙상블 기법은 다중 학습 알고리즘을 이용한 방법으로 높은 예측 성능을 보인다. 앙상블 기법의 대표적인 방법론으로 랜덤포레스트와 그래디언트 부스팅이 있으며, 해당 방법론들은 분류 문제 뿐만 회귀분석에서도 사용 가능하다는 장점이 있다. 그러나 이러한 앙상블 기법은 블랙박스(black-box)적인 특성으로 인해 모델에 대한 해석 필요한 경우에는 선호되지 않는다. 이런 부족한 설명력을 보완하기 위해 앙상블 기법에서는 변수중요도라는 척도를 제시해주지만 설명변수들과의 상대적인 수치일 뿐, 해당 변수의 유의성까지 보장해주지 않는다는 한계가 있다. 우리는 앙상블 기법에서의 변수중요도를 활용하여 변수의 유의성을 검정하는 비모수적 변수 선택 방법을 제안한다. 모의실험 결과, 앙상블 비모수적 변수 선택 방법은 설명변수의 종류 및 모형에 상관없이 기존 회귀모형 변수 선택 방법들보다 우수한 성능을 보였다. 또한, 사례 분석을 통해 렌덤포레스트 비모수적 변수 선택 방법이 설명변수들간 강한 상관관계가 존재하는 경우에도 유의하다고 판단되는 변수를 올바르게 선택한다는 것을 확인하였다.

주요용어 : 앙상블 기법, 랜덤포레스트, 그래디언트 부스팅, 유의성 검정.

[1]02841 서울특별시 성북구 안암로 145, 고려대학교 통계학과 석사과정. E-mail: triple0112@naver.com
[2]02841 서울특별시 성북구 안암로 145, 고려대학교 통계학과 교수. E-mail: hj4cho@korea.ac.kr

Proceedings of the Korean Data Analysis Society
January 26-27, 2023, pp. 69-69

머신러닝 기법을 활용한 소프트웨어 신뢰성 모형 비교에 관한 연구*

김윤수[1], 송광윤[2], 이다혜[3], 장인홍[4]

요 약

소프트웨어는 매우 다양한 분야에서 활용된다. 다양한 기능을 수행하는 소프트웨어가 고장나게 된다면 사회·경제적으로 큰 손실을 불러일으킬 것이다. 따라서, 소프트웨어 신뢰성의 향상을 위해 비동질 포아송 과정을 가정한 소프트웨어 신뢰성 모형이 연구되었고, Goel과 Okumoto로부터 시작된 소프트웨어 신뢰성 모형은 여러 조건을 반영한 소프트웨어 신뢰성 모형으로까지 확장되어 개발되고 있다. 하지만 여러 조건에 의존한 소프트웨어 신뢰성 모형은 특수한 경우에만 잘 적용되는 경우가 많아 조건을 만족하지 못하면 좋은 결과를 보여주지 못하는 경우가 존재한다. 따라서, 본 연구에서는 특수한 상황만을 고려하는 것이 아닌 일반적인 상황에 잘 적용시킬 수 있는 모형을 제안하였다. 제안한 모형은 데이터에 의존하는 머신러닝 기법들을 활용한 모형으로 머신러닝의 대표적인 기법인 랜덤포레스트, 서포트벡터머신, 심층신경망, 순환신경망과 기존의 소프트웨어 신뢰성 모형을 비교 분석하였으며, 5가지의 criteria 비교를 통해 제안하는 모형 중 신경망 모형이 가장 적합하였다. 해당 결과를 통해 통계적 가정을 기반한 소프트웨어 신뢰성 모형보다 데이터에 의존하여 학습시킨 소프트웨어 신뢰성 모형의 우수함을 보였다.

주요용어 : 소프트웨어 신뢰성, 랜덤포레스트, 서포트벡터머신, 심층신경망, 순환신경망.

---

*이 논문은 2021년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (NRF-2021R1F1A1048592, NRF-2021R1A6A3A01086716, NRF-2021R1I1A1A01059842).

[1]61452 광주광역시 동구 필문대로 309, 조선대학교 대학원 전산통계학과 박사수료.
E-mail: imk92315@naver.com

[2]61452 광주광역시 동구 필문대로 309, 조선대학교 컴퓨터통계학과 연구교수. E-mail: skyjacker152@naver.com

[3]61452 광주광역시 동구 필문대로 309, 조선대학교 컴퓨터통계학과 박사후연구원(Post-Doc).
E-mail: ec4869@naver.com

[4](교신저자) 61452 광주광역시 동구 필문대로 309, 조선대학교 컴퓨터통계학과 교수.
E-mail: ihchang@chosun.ac.kr

Proceedings of the Korean Data Analysis Society
January 26-27, 2023, pp. 71-74

머신러닝을 활용한 부도 예측 모형 연구: 소매 익스포져를 중심으로

조용복[1], **박진수**[2]

요 약

코로나19 팬데믹 이후 장기간 공급된 유동성은 인플레이션 급등으로 이어졌고, 이에 대응한 금리 인상과 함께 유동성 회수가 시작되면서 금융기관의 건전성 관리가 화두로 여겨지고 있다. 하지만 기존의 연구 대부분은 재무제표 정보를 기반으로 한 법인을 위주로 이루어지고 있는 바, 본 연구에서는 가계 및 중소기업을 포함하는 소매 익스포져를 중심으로 다루고자 한다.

한편, 부도 예측 모형에 필요한 데이터는 부도 데이터에 비해 정상 데이터가 압도적으로 많은 특성을 가지고 있다. 따라서, 부도 예측 모형을 구축 시 데이터 불균형 문제를 반드시 고려해야 한다. 본 연구는 부도 예측 모형 구축 시, 불균형 데이터 문제를 해결하기 위해 적용하는 방안의 적용 유무 및 적용 방안에 따른 모형의 성능 비교를 중심으로 진행하였다.

실증분석을 위해 A 은행의 소매 익스포져를 대상으로 부도 모형을 구축하였고, 모형의 예측 성능을 비교·분석하였다. 실험 결과 첫째, 불균형이 존재하는 원천 데이터를 활용한 경우 정밀도가 가장 높은 반면, 민감도가 가장 낮음을 확인하였다. 둘째, Oversampling 기법을 적용하여 데이터 불균형 문제를 해결한 경우, 모형의 민감도는 상승하였으나 동시에 정밀도가 하락하였다. 셋째, 모형의 Classification Threshold를 조정할 경우, Oversampling 기법을 적용한 경우와 유사하게 모형의 민감도는 상승하는 반면 정밀도는 하락하였다. 다만, 민감도와 정밀도를 종합적으로 고려한 지표인 F-1 Score를 기준으로 가장 우수한 성능을 나타내었다.

주요용어 : 소매 익스포져, 부도 예측 모형, 불균형 데이터, SMOTE, GHOST.

1. 서론

코로나19 팬데믹 이후, 전 세계는 금융 안정을 위해 막대한 유동성을 공급하였다. 대한민국의 금융당국 또한 완화적인 통화정책 기조와 함께 다양한 금융 지원 방안을 마련하였고, 이로 인해 대출 시장은 큰 폭으로 성장하였다. 하지만 장기간 공급된 유동성은 러시아-우크라이나 전쟁 등 다양한 원인과 맞물려 인플레이션으로 이어졌고, 이에 대응하기 위한 금리 인상 및 유동성 회수가 시작되면서 금융기관의 건전성 관리가 그 어느 때보다 주목을 받고 있는 상황이다.

그러므로 부도 예측 모형을 통해 부도가 예상되는 차주를 선별하고, 선제적 조치를 취하는 것이 중요한 시점이지만 부도 예측 모형 연구의 대부분은 재무제표 정보를 기반으로 한 법인을 위주로 이루어지고 있다. 따라서, 본 연구에서는 Basel 기준의 소매 익스포져 중 일반 소매 익스포져와 소

[1](교신저자) 49236 부산광역시 서구 구덕로 225, 동아대학교 경영정보학과 교수.
E-mail: ybcho@dau.ac.kr
[2]49236 부산광역시 서구 구덕로 225, 동아대학교 특수대학원 디지털금융학과 석사과정.
E-mail : 2180968@donga.ac.kr

매 중소기업 익스포져를 중심으로 하는 부도 예측 모형을 다루고자 한다.

한편, 부도 예측 모형에 필요한 데이터는 부도 데이터에 비해 정상 데이터가 압도적으로 많은 특성을 가지고 있다. 이러한 데이터의 불균형 문제를 해결하지 않고 모형을 구축할 경우, 다수의 관측치를 가지고 있는 정상 데이터에 관련된 특성 변수의 영향에 편중되면서 모형의 예측 성능이 저하될 수 있다. 따라서, 본 연구는 데이터의 불균형 문제를 해결하기 위해 적용하는 방안의 적용 유무 및 적용 방안에 따른 모형의 성능을 비교하는 방향으로 진행하고자 한다.

본 논문의 구성은 다음과 같다. 먼저 2장에서는 데이터 및 연구 방법에 대해 간략히 소개한다. 이어서 3장에서는 실증분석에 따른 실험 결과를 비교하고, 마지막으로 4장에서 결론을 정리한다.

2. 데이터 및 연구 방법

2.1 연구 데이터

부도 예측 모형 구축에 앞서, 목적 변수인 부도에 대한 정의가 필요하며 본 연구에서는 Basel Ⅲ 기준의 부도 정의를 활용하였다. 실증분석을 위해 A 은행의 2019년 12월 말, 2020년 12월 말, 2021년 11월 말 기준 소매 일반 익스포져와 소매 중소기업 익스포져 데이터를 수집하였다.

기준 시점으로부터 향후 1년간 부도 여부를 관측하였으며, 관측 기간 내 한 번이라도 부도를 경험한 차주는 이후 부도 사유가 해제된 경우에도 부도 차주로 인식하였다.

부도를 예측하기 위한 특성 변수를 구성을 위해, 신용 관련 정보와 은행 거래실적 관련 정보를 활용하였다. 먼저, KCB(Korea Credit Bureau)에서 A 은행에 제공하고 있는 440개의 속성 중 2019년 12월 말 기준으로 존재하는 260개의 속성을 기초 변수로 선택하였다. 이어서 결측치 및 이상치 제거 등의 전처리를 실시하였고, 속성을 총 4개의 그룹(대출, 카드, 상환능력, 불량)으로 구분한 뒤 그룹별로 LASSO 회귀를 활용하여 특성 변수를 선택하였다. 또한 변수 간 다중공선성 문제를 해결하기 위해 VIF Test와 상관관계 분석을 실시하여 총 26개의 신용 관련 특성 변수를 선정하였다. 여기에 총 8개의 은행 거래실적 관련 정보를 추가하여, 최종적으로 총 34개의 특성 변수를 선정하였다.

모형 설정을 위한 학습 데이터로 2019년 12월 말, 2020년 12월 말 데이터를 사용하였고 모형 성능 평가를 위한 시험 데이터로 2021년 11월 말 데이터를 사용하였다.

2.2 연구 절차 및 방법

2.1절에서 언급한 바와 같이 수집한 데이터에 대한 전처리와 함께 특성 변수를 구성한 뒤, 학습 데이터와 시험 데이터를 구분한다.

이후 Decision Tree 계열 알고리즘의 대표적인 기법인 Random Forest, Gradient Boost, XGBoost를 사용하여 각 모형의 성능을 측정한다. 이때, 데이터 불균형이 존재하는 원천 데이터를 그대로 활용한 경우와 Oversampling 기법 중 SMOTE를 적용한 경우, Classification Threshold 조정을 위해 GHOST를 적용한 경우 모형의 성능 차이에 대해 비교 분석을 실시한다.

2.3 성능평가 지표

일반적으로 분류 모형의 성능은 혼동행렬(Confusion Matrix)을 통해 TN(True Negative), TP(True Positive), FN(False Negative), FP(False Positive)의 수치를 바탕으로 한 정확도(Accuracy), 정밀도(Precision), 민감도(Sensitivity) 등과 같은 지표를 통해 평가한다.

본 연구는 불균형 데이터를 사용한 분류 모형의 성능 비교를 중심으로 이루어지므로, 실제 부도 데이터를 부도로 예측하는 비율인 민감도와 부도로 예측한 데이터 중 실제 부도 데이터의 비율인 정밀도를 성능평가 지표로 활용한다. 한편, 민감도와 정밀도는 상충 관계를 가지고 있으므로, 민감도와 정밀도를 종합적으로 고려할 수 있는 F-1 Score 또한 성능평가 지표로 활용하고자 한다.

3. 실험 결과

Table 1은 불균형이 존재하는 원천 데이터를 활용한 경우, Oversampling 기법을 통해 데이터 불균형을 해소한 경우, Classification Threshold를 조정한 경우의 모형 성능을 요약한 결과이다. 머신러닝 알고리즘에 따라 성능의 차이는 일부 존재하나, 전반적인 실험 결과는 다음과 같다.

첫째, 불균형이 존재하는 원천 데이터를 활용한 경우 정밀도가 가장 높은 반면, 민감도는 가장 낮은 모습을 나타내었다. Random Forest와 XGBoost는 민감도 및 정밀도 관점에서 유사한 성능을 보이고 있으며, 민감도와 정밀도를 모두 감안한 F-1 Score 관점에서는 Gradient Boost가 가장 우수한 성능을 보이고 있다.

둘째, Oversampling 기법 중 SMOTE를 적용하여 데이터 불균형을 해소한 경우 민감도가 상승하는 반면 정밀도는 하락하는 모습을 나타내었다. F-1 Score 관점에서 볼 때, Gradient Boost는 원천 데이터를 활용한 경우보다 성능이 소폭 하락하였으나, Random Forest 및 XGBoost의 경우 성능이 개선되었다.

셋째, Classification Threshold 조정을 위해 GHOST를 적용한 경우, 데이터 불균형을 해소한 경우와 유사하게 민감도가 상승하며 정밀도가 하락하였으나, F-1 Score 관점에서 가장 우수한 성능을 보이고 있다. 즉, 민감도와 정밀도가 가장 균형을 이루고 있는 실험 결과라고 할 수 있다.

Table 1. Summary of the model performance

		Sensitivity	Precision	F-1 Score	TN	FP	FN	TP
Random Forest	Raw Data	0.0677	0.6049	0.1217	152,565	96	2,025	147
	SMOTE	0.1289	0.2454	0.1690	151,800	861	1,892	280
	GHOST	0.1215	0.4415	0.1906	152,327	334	1,908	264
Gradient Boost	Raw Data	0.1105	0.3974	0.1729	152,297	364	1,932	240
	SMOTE	0.4811	0.1042	0.1713	143,679	8,982	1,127	1,045
	GHOST	0.3218	0.2063	0.2514	149,972	2,689	1,473	699
XGBoost	Raw Data	0.0792	0.5890	0.1396	152,541	120	2,000	172
	SMOTE	0.5009	0.1036	0.1717	143,245	9,416	1,084	1,088
	GHOST	0.2818	0.2363	0.2570	150,683	1,978	1,560	612

4. 결론

본 연구는 머신러닝 알고리즘을 통한 소매 익스포져 부도 예측 모형 구축 시, 데이터의 불균형 문제 해결을 위한 Oversampling 기법 적용과 Classification Threshold 조정에 따른 모형의 성능 차이를 중심으로 점검하였다.

먼저, 불균형이 존재하는 원천 데이터를 그대로 활용할 경우 모형의 정밀도는 높은 반면 민감도는 낮은 모습을 발견하였다. 정밀도를 기준으로는 Random Forest가 가장 우수한 성능을 보였고, XGBoost의 경우 Random Forest에 비해 정밀도가 소폭 낮으나 F-1 Score 기준으로는 가장 우수한 성능을 보였다.

Oversampling 기법을 통해 데이터 불균형 문제를 해결한 경우, 모형의 민감도는 상승하였으나 정밀도가 하락함을 확인하였다. F-1 Score를 기준으로 볼 때, Gradient Boost는 원천 데이터를 활용할 때에 비해 오히려 성능이 하락하였으나 Random Forest 및 XGBoost의 경우 모형의 성능이 개선되었다.

Classification Threshold를 조정한 경우, Oversampling 기법을 활용한 경우와 유사하게 모형의 민감도가 상승하고 정밀도가 하락하였으나, F-1 Score를 기준으로는 가장 우수한 성능을 보였다. 특히, 세 가지 머신러닝 알고리즘 중 XGBoost가 가장 우수한 성능을 보였다.

본 연구는 데이터 불균형 문제를 해결하기 위한 두 가지 방안의 적용 여부에 따른 모형의 성능 비교를 중심으로만 진행되었다는 한계점을 갖고 있는 바, 향후 특성 변수 선택에 따른 모형 성능 비교 등을 후속 연구로 진행하고자 한다.

References

Burez, J., Van den Poel, D. (2009). Handling class imbalance in customer churn prediction, *Expert Systems with Applications*, 36(3), 4626-4636. https://doi.org/10.1016/j.eswa.2008.05.027

Cho, Y., Cho, D., Choi, B. (2022). Applications of the classification algorithm for unbalanced time series data: Focusing on the corporate default model, *Journal of the Korean Data Analysis Society*, 24(2), 639-651. (in Korean). https://doi.org/10.37727/jkdas.2022.24.2.639

Esposito, C., Landrum, G., Schneider, N., Stiefl, N., Riniker, S., (2021). GHOST: Adjusting the decision threshold to handle imbalanced data in machine learing, *J. Chem. Inf. Model.* 2021, 61, 6, 2623 - 2640. https://doi.org/10.1021/acs.jcim.1c00160

Proceedings of the Korean Data Analysis Society
January 26-27, 2023, pp. 75-78

영구적·일시적 현금흐름 뉴스와 할인율 뉴스에 따른 모멘텀 수익률 분석

오명[1]

요 약

본 논문에서는 로그-선형 구조형 벡터자기회귀 모형을 이용하여 예기치 못한 시장 초과수익률을 영구적인 현금흐름 뉴스, 일시적인 현금흐름 뉴스 및 할인율 뉴스로 분해하였고, 이러한 뉴스 상황에 따른 모멘텀 전략의 수익성을 분석하였다. 실증 분석한 결과는 다음과 같다. 첫째, 미국 주식시장에서는 모멘텀 현상이 발생한다는 것을 알 수 있다. 둘째, 영구적인 현금흐름 뉴스와 일시적인 현금흐름 뉴스 상황 하에서 모두 모멘텀 수익성이 존재한다는 것을 보여 주었다. 이는 Celiker, Kayacetin, Kumar, and Sonaer(2016)의 결과와 일관성을 가진다. 셋째, 할인율 뉴스 상황 하에서도 모멘텀 현상이 뚜렷하게 나타난다는 것을 알 수 있다. 마지막으로, 전체 시장이 상승기일 때 모멘텀 수익성이 더 강하게 나타난다.

주요용어 : 모멘텀 수익률, 로그-선형 구조형 벡터자기회귀 모형, 영구적인 현금흐름 뉴스, 일시적인 현금흐름 뉴스, 할인율 뉴스.

1. 서론

모멘텀 효과에 기초한 투자전략의 성과를 다룬 기존 연구들을 살펴보면, Jegadeesh, Titman(1993)은 미국 주식시장에서 승자주를 매수하고 패자주를 매도하는 무비용 포트폴리오를 구성하면 유의적인 초과수익이 발생함을 보여주었다. Rouwenhorst(1998)와 Griffin, Ji, Martin(2003)은 글로벌 주식시장을 대상으로 모멘텀 수익성을 분석한 결과, 다수의 국가에서 모멘텀 수익률이 존재함을 보여주었고 모멘텀 현상은 특정 주식시장에 국한된 현상이 아니라고 주장하였다. 한편, 모멘텀 투자 성과의 원인을 규명하고자 하는 연구가 활발히 진행되었다. Daniel, Titman(1999)은 모멘텀 현상이 성장주 기업에서 뚜렷하게 나타난다고 보고하였다. Hong, Lim, Stein(2000)은 재무분석가 수가 적은 기업의 모멘텀 수익성이 더 높다고 하였다.

모멘텀 수익성이 시장 상황에 기인한다는 연구 결과도 있다. Cooper, Gutierrez, Hameed(2004)는 시장 상황을 고려하여 미국 주식시장에서의 모멘텀 수익성을 분석하였다. 그들은 시장이 상승기일 때 투자자들은 과신하는 경향을 보이며, 이로 인해 모멘텀 현상이 발생한다는 것이다. Celiker, Kayacetin, Kumar, and Sonaer(2016)는 현금흐름 뉴스와 할인율 뉴스 상황에 따라 모멘텀 효과가 다르게 나타난다는 것을 보여주었다. 그들은 Campbell(1991)에 따라 로그-선형 모형에 벡터자기회귀 모형을 결합시킨 로그-선형 벡터자기회귀 모형을 이용하여 예기치 못한 시장 초과수익률을 현금흐름 뉴스와 할인율 뉴스로 분해하여 각 뉴스 상황 하에서의 모멘텀 수익성을 분석하였다. 그들은 미국 주식시장에서는 현금흐름 뉴스 상황 하에서 모멘텀 수익성이 강하게 나타난다는 것을 보여

[1]48513 부산광역시 남구 용소로 45, 부경대학교 글로벌자율전공학부 조교수. E-mail: wuming0530@pknu.ac.kr

주었다. 이는 미국 주식시장일 경우 분산된 소유권 구조를 가지기에 정보비대칭 현상이 심하고 안정적인 배당정책을 실시하여 배당의 신호효과가 강하다. 또한 시장을 통한 자본조달 시스템이기에 부채 비율이 상대적으로 낮아 투자자들은 현금흐름 뉴스에 더 민감하게 반영하여 모멘텀 현상이 더 강하게 나타날 수 있다.

하지만 Wu, Ohk, Ko(2021)는 로그-선형 벡터자기회귀 모형으로 현금흐름과 할인율 뉴스를 추정하면 뉴스 간 상관관계가 존재하여 정확하게 추정할 수 없다고 하였다. 또한 현금흐름 뉴스를 영구적인 현금흐름 뉴스와 일시적인 현금흐름 뉴스로 분해할 수 없는 한계점을 가진다고 하였다. 이에 본 논문에서는 로그-선형 모형에 구조형 벡터자기회귀 모형을 결합하여 예기치 못한 시장 초과수익률을 영구적인 현금흐름 뉴스, 일시적인 현금흐름 뉴스 및 할인율 뉴스로 분해시켰다. 그리고 각 뉴스 상황 하에서의 모멘텀 수익률을 측정하여 비교하고자 한다. 마지막으로, Cooper, Gutierrez, Hameed(2004)에 따라 모멘텀 수익률이 시장 상황에 기인하는 지를 분석하였다.

본 논문의 실증 결과는 다음과 같다. 첫째, 미국 주식시장에서는 모멘텀 효과가 강하게 나타난다. 둘째, 영구적인 현금흐름 뉴스와 일시적인 현금흐름 뉴스 상황 하에서 모두 모멘텀 전략에 의하여 유의적인 수익을 창출할 수 있다는 것을 보여 주었다. 셋째, 현금흐름 뉴스뿐만 아니라 할인율 뉴스 상황 하에서도 모멘텀 현상이 뚜렷하게 나타난다는 것을 알 수 있다. 마지막으로, 전체 시장이 상승기일 때 모멘텀 효과가 강하게 나타났다. 본 논문은 로그-선형 모형에 구조형 벡터자기회귀 모형을 결합하여 예기치 못한 시장 초과수익률을 영구적인 현금흐름 뉴스, 일시적인 현금흐름 뉴스 및 할인율 뉴스로 분해시켜 뉴스 상황 하에서 모멘텀 수익성을 측정하였다는 데에 있어 학술적 공헌을 가진다.

제2장은 본 논문에서 사용하는 데이터와 연구방법론이다. 제3장은 뉴스 상황 하에서의 모멘텀 수익률에 대한 실증 결과를 보여 주고 마지막 장은 본 논문의 결론이다.

2. 데이터 및 방법론

2.1. 데이터

본 논문은 NYSE, Nasdaq, Amex에 상장된 주식들을 연구 대상으로 하여 영구적인 현금흐름 뉴스, 일시적인 현금흐름 뉴스 및 할인율 뉴스 상황 하에서의 모멘텀 수익률을 구하였다. 논문에서 사용된 월별 시가총액과 현금배당이 반영된 수정주가는 CRSP(Center for Research in Security Prices) 데이터베이스에서 추출하였다. 영구적인 현금흐름 뉴스, 일시적인 현금흐름 뉴스 및 할인율 뉴스를 구하기 위하여 사용된 전체 시장의 월별 로그 이익 변화율, 로그 배당성향, 시장 초과수익률(ER) 등의 데이터는 Robert J. Shiller 홈페이지와 French 홈페이지에서 추출하였다. 이러한 내생변수들의 기초통계량은 Table 1에서 보여주고 있다.

Table 1. Descriptive statistics

	Mean	Median	Max.	Min.	Stdev.
△logE	0.0041	0.0063	0.7039	-0.5283	0.0436
logD/E	0.4349	0.4234	1.6042	0.2532	0.1391
ER	0.0045	0.0091	0.4075	-0.2741	0.0437

2.2. 구조형 벡터자기회귀 모형 및 뉴스의 분해

본 논문은 Lee(1998)의 이익 및 배당의 모형을 이론적 근거로 하여 로그 이익 변화율(△logEt), 로그 배당성향(△logD/Et) 및 시장 초과 수익률(ER)을 구조형 벡터자기회귀 모형의 내생변수로 선정하였다. 구조형 모형과 축약형 모형은 다음과 같은 관계를 가진다.

$$e_t = c^0 \epsilon_t \tag{1}$$

하지만 구조형 벡터 자기회귀 모형을 추정하기 위해서는 Lee(1998)의 이익 및 배당의 모형을 이론적 근거로 하여 다음과 같은 3개의 제약을 주었다.

$$\begin{bmatrix} e_t^{\triangle \log E} \\ e_t^{\log D/E} \\ e_t^{ER} \end{bmatrix} = \begin{bmatrix} c_{11}^0 & BQ & 0 \\ c_{21}^0 & c_{22}^0 & 0 \\ c_{31}^0 & c_{32}^0 & c_{33}^0 \end{bmatrix} \begin{bmatrix} \epsilon_t^{\triangle \log E} \\ \epsilon_t^{\triangle \log D/E} \\ \epsilon_t^{ER} \end{bmatrix} \tag{2}$$

여기에서 c_{ij}^0는 i번째 변수의 충격이 j번째 변수의 충격에 미치는 동시적 영향을 의미하고 BQ는 Blanchard and Quah(1988)의 일시적인 영향을 의미한다. 이러한 제약을 통하여 동시적 영향을 나타내는 계수들을 추정할 수 있다. 이에 수식 (2)의 세 번째 행을 다시 쓰면 다음과 같다.

$$\begin{aligned} r_t - E_{t-1} r_t = e_t^{ER} &= c_{31}^0 \epsilon_t^{\triangle \log E} + c_{32}^0 \epsilon_t^{\triangle \log D/E} + c_{33}^0 \epsilon_t^{ER} \\ &= N_{PCF,t} + N_{TCF,t} + N_{DR,t} \end{aligned} \tag{3}$$

여기서 e_t^{ER}는 예기치 못한 시장 초과수익률을 의미하고 $c_{31}^0 \epsilon_t^{\triangle \log E}$은 영구적인 현금흐름 뉴스($N_{PCF,t}$), $c_{32}^0 \epsilon_t^{\triangle \log D/E}$는 일시적인 현금흐름 뉴스($N_{TCF,t}$) 및 $c_{33}^0 \epsilon_t^{ER}$는 할인율 뉴스($N_{DR,t}$)를 의미한다.

3. 실증분석 결과

3.1. 모멘텀 수익률

Table 2는 현금흐름 뉴스와 할인율 뉴스 상황을 고려하지 않을 때 분석한 모멘텀 수익성이다. 괄호안에는 t-값이다. Table 2에서는 미국 주식시장에서는 모멘텀 효과가 강하게 나타난다는 것을 알 수 있다. 이러한 결과는 Chui, Titman, Wei(2010), Fama, French(2012)의 주장을 지지한다.

Table 2. Momentum returns

Loser	2	3	4	5	6	7	8	9	Winner	WML
1.581%	1.374%	1.242%	1.308%	1.262%	1.272%	1.325%	1.449%	1.590%	2.229%	0.648%
(5.270)	(5.520)	(5.934)	(6.678)	(6.971)	(7.119)	(7.969)	(8.920)	(9.322)	(11.239)	(2.658)

3.2. 뉴스 상황에 따른 모멘텀 수익성

Table 3은 영구적인 현금흐름 뉴스, 일시적인 현금흐름 뉴스 및 할인율 뉴스 상황을 고려할 때의 모멘텀 수익률을 보여 주고 있다. Panel A는 상승_영구적 현금흐름 상황과 하락_영구적 현금흐

름 상황일 때의 모멘텀 수익률의 평균값이고, Panel B는 상승_일시적 현금흐름과 하락_일시적 현금흐름 상황일 때의 모멘텀 수익률의 평균값이다. 상승_할인율 상황과 하락_할인율 상황일 때의 모멘텀 수익률의 평균값은 Panel C에서 보여 주고 있다. 그리고 괄호안의 값은 t-값을 나타낸다. Panel A에서는 상승_영구적 현금흐름 상황일 때 모멘텀 현상이 존재함을 보여 주고 Panel B에서는 상승_일시적 현금흐름과 하락_일시적 현금흐름 상황일 때 모두 모멘텀 수익률이 유의적임을 보여준다. Panel C에서는 상승_할인율 뉴스 상황일 때 모멘텀 전략의 수익성이 존재한다는 것을 확인할 수 있다.

4. 결론

본 논문에서는 로그-선형 구조형 벡터자기회귀 모형을 이용하여 예기치 못한 시장 초과수익률을 영구적인 현금흐름 뉴스, 일시적인 현금흐름 뉴스 및 할인율 뉴스로 분해시키고 각 뉴스 상황 하에서의 모멘텀 수익성을 측정하고자 한다. 실증 결과는 다음과 같다. 첫째, 미국 주식시장에서는 모멘텀 전략으로 투자수익을 창출할 수 있다. 둘째, 영구적인 현금흐름 뉴스와 일시적인 현금흐름 뉴스 상황 하에서 모두 모멘텀 수익률이 유의적이라는 것을 보여 주었다. 셋째, 현금흐름 뉴스뿐만 아니라 할인율 뉴스 상황 하에서도 모멘텀 현상이 뚜렷하게 나타난다는 것을 알 수 있다.

Table 3. Momentum returns conditional on permanent and temporary cash flow and discount rate news

Panel A. Permanent cash-flow states				
	Loser	5	Winner	WML
Up_CF states	1.469%	1.148%	2.140%	0.671%
	(12.231)	(14.756)	(19.333)	(7.235)
Down_CF states	1.996%	1.446%	2.191%	0.196%
	(9.058)	(10.686)	(14.073)	(1.177)
Up-down	-0.527%	-0.298%	-0.052%	0.475%
	(-2.100)	(-1.911)	(-0.270)	(2.497)
Panel B. Temporary cash-flow states				
	Loser	5	Winner	WML
Up_DR states	2.070%	1.574%	2.327%	0.257%
	(10.757)	(13.104)	(16.035)	(1.801)
Down_DR states	1.364%	1.007%	2.009%	0.645%
	(9.520)	(11.340)	(17.141)	(5.671)
Up-down	0.706%	0.568%	0.318%	-0.388%
	(2.944)	(3.802)	(1.703)	(-2.128)
Panel C. Discount-rate states				
	Loser	5	Winner	WML
Up_DR states	1.486%	1.340%	2.475%	0.989%
	(12.919)	(16.563)	(20.438)	(11.511)
Down_DR states	1.984%	1.209%	1.769%	-0.214%
	(8.728)	(9.005)	(12.461)	(-1.262)
Up-down	-0.497%	0.131%	0.705%	1.203%
	(-1.953)	(0.837)	(3.780)	(6.323)

Proceedings of the Korean Data Analysis Society
January 26-27, 2023, pp. 79-82

환경정책이 한국 제조업의 기업 성과에 미치는 영향 분석*

이진희[1], **송철종**[2]

요 약

이 연구는 환경정책의 불확실성과 엄격성이라는 두 가지 특성이 한국 제조업의 기업 성과에 미치는 영향을 분석하였다. 환경정책 변수로는 기후정책 불확실성 지수(Climate Policy Uncertainty Index)와 환경정책 엄격성 지수(Environmental Policy Stringency Index)를 사용하였다. 종속변수와 통제변수 자료는 기업활동조사에서 구하였다. 종속변수인 기업 성과 변수는 총비용, 부가가치, 총매출을 이용하였는데 이는 각각 환경정책이 기업 성과에 미치는 1차 효과, 2차 효과, 3차 효과를 의미한다. 통제변수로 종사자 수, 레버리지비율, R&D 투자, 수출집약도와 해외진출 여부, 그리고 실질실효환율을 사용하여 기업의 특성을 통제하였다. 2008년부터 2018년까지의 기간 동안 3년 이상 생존한 기업들을 대상으로 하였으며, 연간 5,082~5,805 개의 기업을 포함하고 있다. 패널고정효과 모형을 분석한 결과, 환경정책의 글로벌 불확실성은 기업의 총비용, 부가가치, 총매출을 증가시키고, 로컬 엄격성은 감소시키는 것으로 나타났다.

주요용어 : CPU 지수, EPS 지수, 환경정책, 기업 성과.

1. 서론

이 연구는 한국 제조업의 기업들을 대상으로 환경정책의 두 가지 특성인 불확실성과 엄격성이 한국 기업 성과에 어떠한 영향을 미치는지 분석하고자 한다. 이 연구에서 사용하는 환경정책의 불확실성은 Gavriilidis(2021)에서 구축한 미국의 기후정책 불확실성(Climate Policy Uncertainty: 이하 CPU로 표시) 지수이고, 환경정책의 엄격성은 OECD가 구축한 한국의 환경정책 불확실성(Environmental Policy Stringecy: 이하 EPS로 표시) 지수[1]이다. 미국의 CPU를 사용하는 가장 큰 이유는 기후정책 불확실성에 대한 지수는 현재 미국만 구축되어 있기 때문이다. 뉴스 기사를 통해 기후정책의 불확실성 지수를 구한 것인데 Berestycki et al.(2022)에서는 미국을 포함한 12개 국가에 대한 CPU를 구축하였으나 아직 구체적인 자료를 공개하지 않고 있다. 그런데 Ren et al.(2022)에서는 경제규모 1위인 미국의 CPU를 글로벌 기후 위기에 대한 대리 변수로 이용하여 기후정책의 불확실성이 중국 제조업 기업의 생산성에 미치는 영향을 분석한 바 있다. 이 연구에서도 Ren et al.(2022)

*이 논문은 2022년 교육부와 한국연구재단의 지원을 받아 수행된 연구(NRF)이며, 이 원고는 아이디어 차원의 초안이므로 인용을 허락하지 않음.

[1]07233 서울시 영등포구 의사당대로 1, 국회예산정책처 경제분석국 거시경제분석과 경제분석관. E-mail: dllete@naver.com

[2](교신저자) 31460 충남 아산시 탕정면 선문로 221번길 70, 선문대학교 국제경제통상학과 조교수. E-mail: cjsong5525@sunmoon.ac.kr

[1] 환경정책 엄격성 지수에 관한 보다 자세한 내용은 Kruse et al.(2022)을 참조하시오.

를 따라 미국의 CPU를 글로벌 측면에서의 환경정책의 불확실성에 대한 대리 변수로 이용하고자 한다. 반면에 EPS는 한국의 자료를 이용하여 기업의 성과에 대한 영향을 분석하고자 한다. 이는 환경정책의 글로벌 특성과 로컬 특성을 모두 고려하려는 시도이며, 본 연구의 의의이다. Dechezleprêtre, Sato(2018)에 따르면 기후변화 정책은 기업의 활동에 단계적으로 영향을 미친다. 정책변화는 먼저 기업의 비용 측면에서 변화를 주는데, 기업은 이에 대응하여 생산, 가격, 투자 등을 조정한다. 그 결과 기업의 매출이 변화하게 된다. 본 연구에서는 기후변화 정책이 기업의 활동에 미치는 영향을 단계적으로 살펴보며 기존 연구들과 차별화하였다.

2. 분석 자료

본 연구는 통계청의 「기업활동조사」를 이용하였다. 분석 대상은 2008년부터 2018년까지 제조업 기업 중 3년 이상 생존한 기업으로 한정하였다. 분석에 사용된 연간 기업 수는 최소 5,082 ~ 최대 5,805개이다. 산업은 KSIC 산업 중분류 수준에서 11개로 분류[2]하였으며, 총비용, 부가가치, 총매출은 산업별 생산자 물가지수를 이용하여 실질화하였다.

종속변수와 설명변수에 대한 설명은 Table 1에 제시하였다.

Table 1. Variable Description

Variables	Description
Dependent variables	
log of real total cost($\ln Cost$)	실질 총비용 = 비용/제조업ppi*100
log of real VA($\ln VA$)	실질 부가가치 = (법인세 차감 전 당기순이익 + 영업비용 + 영업외 이자비용 - 재료비)/제조업ppi*100
log of real total sales($\ln Sales$)	실질 총매출 = 매출액/제조업ppi*100)
Characteristic variables of Environmental policy	
lag of USA CPU index(CPU_{t-1}^{US})	월별 CPU를 연간지수로 전환한 후 0~1로 정규화
log of Korea EPS index(EPS_{t-1})	환경정책과 제도에 기초해 1~6(가장 엄격)으로 산출
Control variables	
log of employment($\ln Emp$)	종사자 수 = 상용종사자+임시일용종사자
log of leverage ratio($\ln Leverage$)	레버리지비 = 부채/자산*100
log of R&D share($\ln RnD$)	R&D 투자 집중도 = 연구개발비/ 매출액*100
log of export share($\ln ExpShare$)	수출집약도 = 직접수출거래액/매출액*100
dummy for GVC(GVC)	0 = GVC 불참, 1 = GVC 참여
log of REER by industry ($\ln REER$)	일본 경제산업연구소의 산업별 실질실효환율

분석대상 변수는 실질 총비용(1차 효과), 실질 부가가치(2차 효과), 실질 총매출(3차 효과)이며, 관심 변수는 기후정책 불확실성(Climate Policy Uncertainty: CPU) 지수와 환경정책 엄격성(environmental policy stringency: EPS) 지수이다. CPU 지수는 Gavriilidis(2021)가 미국의 8개 신문에

[2] 11개 산업은 다음과 같이 분류하였다. (1) 음식료품 (2) 섬유 및 가죽제품 (3) 목제 및 목제품 (4) 석탄 및 석유제품, 화학제품 (5) 고무제품 (6) 비금속광물제품 (7) 금속제품 (8) 컴퓨터, 전자 및 광학기기 (9) 전기장비 (10) 기계 및 장미 (11) 운송장비

서 실시간으로 생성되는 뉴스기사의 텍스트를 분석해 기후정책의 불확실성을 파악하기 위해 개발한 지수이다. CPU는 글로벌 기후정책의 불확실성에 대한 대리(proxy) 변수로 활용되었다. 본 연구에서는 분석 기간 중 불확실성이 가장 높은 연도를 1로 설정하고 다른 연도에서는 이 연도에 대한 비율로 표시하여, 0~1점의 지수로 변환하였다.[3] OECD는 미국, 영국, 일본, 한국 등 국가별 환경 정책의 엄격성을 비교하기 위해 EPS 지수를 발표하고 있다. EPS 지수는 1~6점으로 지수를 산출하며, 점수가 높을수록 환경 정책이 엄격하다는 것을 의미한다. 마지막으로 기업 및 산업의 특성을 통제하기 위해 기업의 규모를 나타내기 위한 종사자 수, 재무적 특성을 보여주는 레버리지비율, 기업의 혁신과 기후변화 적응에 대한 대리변수인 R&D 투자, 기업의 거래적 특성을 보여주는 수출집약도와 해외진출 여부, 대외 여건을 대표하는 산업별 실질실효환율을 포함하였다.

3. 모형 및 분석 결과

본 연구는 고정효과 패널데이터 분석을 실시하였으며, 그 회귀분석 식은 다음과 같다. y_{it}는 기업 성과를 나타내는 변수로 로그 실질 총비용, 로그 실질 부가가치, 로그 실질 총매출을 차례로 대입하였다. $index_{t-1}$는 환경정책의 특성을 나타내는 CPU와 EPS이다.

$$y_{it} = \beta_0 + \beta_1 index_{t-1} + \beta_2 \ln(employee_{it}) + \beta_3 \ln(Leverage_{it}) + \beta_4 \ln(RnD_{it}) + \beta_5 \ln(Export share_{it}) + \beta_6 \ln(REER_{it}) + \beta_7 GVC_{it} + \beta_8 Ind_j + \epsilon_{it}$$

분석 결과는 Table 2에 정리하였다. (1)열은 EPS만 포함한 경우, (2)열은 CPU만 포함한 경우, (3)열은 EPS와 CPU를 모두 포함한 경우의 분석 결과이다. 첫째, (1)열의 결과를 보면 한국의 환경정책이 엄격해질수록 제조 기업의 총비용과 부가가치, 총매출이 줄어드는 것으로 나타났다. 그런데 총매출 감소에 미치는 영향보다 총비용 하락에 미치는 영향이 더 큰 것으로 나타났다. 둘째, (2)열의 결과를 보면 글로벌 환경정책의 불확실성이 증가할수록 한국 제조업 기업의 총비용과 부가가치, 총매출이 늘어나는 것으로 나타났다. 그런데 CPU의 총매출 증가 효과보다는 총비용 상승 효과가 더 큰 것으로 나타났다. (1)열과 (2)열의 결과는, EPS와 CPU를 동시에 포함한 (3)열의 결과에서도 동일하게 나타났다. EPS는 기업의 총비용, 부가가치, 총매출을 하락시키며 CPU는 상승시키는 것으로 나타났다. 본 연구에서는 기업의 이윤에 대한 변수를 고려하지 않았지만 이윤을 총매출에서 총비용을 뺀 것으로 본다면, 위 결과는 EPS는 기업의 이윤을 증가시키지만 CPU는 기업의 이윤을 감소시킬 수 있음을 시사한다. 셋째, EPS가 부가가치에 미치는 영향이 총비용과 총매출에 비해 크게 나타났으나, CPU가 부가가치에 미치는 영향은 총비용과 총매출보다 작은 것으로 나타났다.

통제변수에 대한 결과 중 눈에 띄는 것은 다음과 같다. 넷째, 기업의 재무적 특성인 레버리지 비율은 총비용에 영향을 주지 않는 것으로 보인다. 반면에 작은 수준이긴 하지만 부가가치나 총매출에는 부정적 영향을 미치는 것으로 나타났다. 다섯째, R&D 투자 비율은 총비용, 부가가치, 총매출에 부정적 영향을 미치는 것으로 나타났다. R&D 투자가 총비용을 하락시키기는 하지만 총매출을 감소시키는 효과가 근소하게 더 큰 것으로 나타났다. 여섯째, 기업의 대외여건인 실질실효환율의 절상은 기업의 총비용, 부가가치, 총매출을 모두 증가시키며, GVC에 참여하는 경우에도 동일한 효과를 가져오는 것을 알 수 있다.

---

[3] 구글 트렌드에서 지수화하는 방식과 같다.

Table 2. Results of regression

	Cost			VA			Sales		
	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
EPS_KOR	-0.122*** (0.009)		-0.161*** (0.010)	-0.217*** (0.013)		-0.246*** (0.014)	-0.095*** (0.010)		-0.126*** (0.011)
CPU_USA		0.256*** (0.014)	0.318*** (0.016)		0.140*** (0.021)	0.236*** (0.023)		0.209*** (0.016)	0.257*** (0.017)
Employment	0.705*** (0.016)	0.711*** (0.016)	0.704*** (0.016)	0.717*** (0.019)	0.728*** (0.019)	0.716*** (0.019)	0.735*** (0.016)	0.740*** (0.016)	0.734*** (0.016)
Leverage	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	-0.003*** (0.001)	-0.003*** (0.001)	-0.003*** (0.001)	-0.001*** (0.000)	-0.001*** (0.000)	-0.001*** (0.000)
RnD share	-0.050*** (0.003)	-0.049*** (0.003)	-0.051*** (0.003)	-0.045*** (0.004)	-0.043*** (0.004)	-0.045*** (0.003)	-0.064*** (0.004)	-0.064*** (0.004)	-0.065*** (0.004)
Export share	-0.001 (0.002)	-0.002 (0.002)	0.000 (0.002)	0.002 (0.003)	0.000 (0.003)	0.003 (0.003)	-0.002 (0.002)	-0.002 (0.002)	-0.001 (0.002)
REER	0.408*** (0.034)	0.512*** (0.035)	0.360*** (0.034)	0.369*** (0.045)	0.566*** (0.047)	0.333*** (0.045)	0.357*** (0.036)	0.438*** (0.037)	0.318*** (0.036)
GVC dummy	0.091*** (0.007)	0.087*** (0.007)	0.066*** (0.006)	0.067*** (0.008)	0.081*** (0.008)	0.049*** (0.008)	0.075*** (0.007)	0.072*** (0.007)	0.055*** (0.007)
Constant	6.082*** (0.164)	5.133*** (0.160)	6.265*** (0.162)	5.858*** (0.221)	4.269*** (0.223)	5.996*** (0.220)	6.166*** (0.173)	5.427*** (0.168)	6.314*** (0.172)
Industry Dummy	yes	yes	yes	yes	yes	yes	yes	yes	yes
No. of Obs.	32,022	32,022	32,022	31,748	31,748	31,748	32,022	32,022	32,022
Adj. R-sq.	0.37	0.37	0.38	0.26	0.25	0.26	0.35	0.36	0.36

References

Berestycki, C., Carattimi, S., Dechezlepretre, A., Kruse, T. (2022). Measuring and assessing the effects of climate policy uncertainty, *OECD Economics Department Working Paper* No. 1724.

Dechezleprêtre, A., Sato, M. (2018). OECD Green Policies and Firms' Competitiveness. Issue Paper: Inclusive Solutions for the Green Transition. Green Growth: Knowledge Platform, 31.

Gavriilidis, K. (2021). Measuring Climate Policy Uncertainty. Available at SSRN: https://ssrn.com/abstract=3847388 or http://dx.doi.org/10.2139/ssrn.3847388

Kruse, T., Dechezlepretre, A., Saffar, R., Robert, L. (2022). Measuring environmental policy stringency in OECD countries: An update of the OECD composite EPS indicator, *OECD Economics Department Working Paper* No. 1703.

Ren, X., Zhang, X., Yan, C., Gozgor, G. (2022). Climate policy uncertainty and firm-level total factor productivity: Evidence from China, *Energy Economics* 113, 106209.

Proceedings of the Korean Data Analysis Society
January 26-27, 2023, pp. 83-87

한국 가계부채의 형성요인에 대한 실증분석

고동우[1], 윤성민[2]

요 약

한국의 가계부채는 2000년대 이후 규모가 지속적으로 증가함에 따라 꾸준히 연구되고 있다. 이 논문은 가계부채의 형성요인을 경제 환경 및 정책요인 등에 비추어 다각적으로 분석하고, 그 시사점을 도출해 보고자 한다. 경제 환경 요인으로서 경제성장 요인과 자산가격의 변동 요인을 살펴본다. 그리고 재정 및 통화정책 관점에서 재정지출과 유동성 요인을 함께 검토한다. 나아가 정부의 대출 억제대책 및 부양대책의 가계부채 관리 효과 역시 분석한다.

실증분석 결과는 다음과 같다. 우선, 경제 환경 요인으로서 실물부문의 경제성장 보다는 주택가격 상승이라는 자산 가격 요인이 가계부채 증가에 기여한 바가 큰 것으로 판단된다. 나아가, 재정지출 변동이 부채 증감에 미친 효과는 유효한 것으로 관측되나, 통화량 요인의 유효성은 불확실하다. 마지막으로 가계부채에 대한 정부의 제도적 개입은 주택담보대출 억제를 위한 정책 및 경기부양을 위한 대출요건 완화정책 모두 일정한 효과가 있었던 것으로 판단된다.

위 분석 결과를 종합한 정책적 시사점은 다음과 같다. 우선 가계부채 관리를 위하여 주택정책 등을 통한 안정적 자산 가격 관리가 수반될 필요가 있다. 또한 정부의 적절한 재정지출 및 가계대출 제도운용을 통하여 가계부채 규모에 대한 정책적 관리가 일정 수준 가능할 것이다.

주요용어 : 가계부채, 경제성장률, 주택가격, 재정지출, 유동성, 대출억제정책.

1. 표본자료와 분석방법

1.1. 표본자료

이 연구의 실증분석 대상은 2007년 4분기부터 2022년 3분기에 이르는 15년간 60분기에 해당하는 가계부채 규모의 시계열 자료를 활용한다. 해당 기간은 글로벌 금융위기를 전후한 시점을 포괄하여 최근 새 정부 출범 직후에 이르는 기간이다. 이는 과거 국제적 경기 침체와 최근 자산가격의 급등시점을 모두 포괄하여 가계부채의 변동 이력에 대한 종합적 분석이 가능할 것으로 기대된다.

Table 1과 같이 총가계신용(HTC)을 주요 종속변수로 하고, 대출성격에 따라 주택담보대출(HML) 및 보증부대출과 판매신용 등 기타대출(HOC)을 구분한다. 또한 설명변수로서 GDP성장률(GDP), 1인당소득(ICM), 가계대출금리(HIR), 주택매매가격지수(HPX), 수도권 주민등록세대수(POM), 재정수지 경상지출(GFC), 통화량(M2) 등을 활용한다.

[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

Table 1. Descriptive statistics

Categories	Variables	Definnition	Mean	St.d.	Max	Mean
Dependent variables	HTC	Household total credit	1,205,305.0	380,620.5	1,870,613.8	665,394.2
	HML	Household mortgage loans	630,721.2	208,608.6	1,007,871.0	343,798.9
	HOC	Household others credit	574,583.8	172,341.7	878,374.2	321,595.3
Independent variables	GDP	Gross domestic product (quarterly growth rate)	0.7	1.0	3.0	-3.3
	ICM	Personal Income	4,899,436.3	774,498.9	6,642,273.0	3,704,019.0
	HIR	Household loan interest rate	4.3	1.3	7.5	2.6
	HPX	Housing price index	71.3	11.3	100.8	56.9
	POM	Population of Metropolitan	10,315,016.2	691,967.9	11,676,865.0	9,225,724.0
	GFC	Central Government Fiscal Balance(Current Expenditure)	80,195.0	31,479.4	178,250.0	36,239.0
	M2	Broad money	2,262,472.3	680,286.8	3,734,164.2	1,250,018.2

1.2. 분석방법

이 연구는 대출규모 결정요인에 대한 시계열 회귀분석을 실증분석 도구로 활용한다. 우선 시계열 지표의 단위근 검정 및 차분, 변수 간 상관관계가 높은 설명변수의 배제 등을 통하여 자료의 안정성을 확보하고, 회귀분석을 통하여 가계부채 형성 원인에 대한 실증분석 결과를 도출한다.

우선, 경제 환경 및 정책효과를 이론적으로 고려하여 종속변수인 가계대출에 미치는 효과를 분석한다. 따라서 전체 가계대출(HTC)과 보증부 대출 및 판매신용 등 기타 대출(HOC)로 종속변수를 구분하고, 각 요인에 대하여 경제 환경 변수인 경제성장률(GDP)과 주택가격(dHPX), 그리고 정책변수인 재정수지(dGFC) 및 통화량(M2) 효과를 적용한 회귀분석을 식(1) 및 (2)와 같이 실시한다.

나아가 설명변수에 대한 가설검정 및 다중공선성 문제 등 모형의 타당성을 종합적으로 고려하여 가장 타당한 회귀분석 모형을 식(3)과 같이 도출한다. 나아가 시계열변수 활용에 따른 오차의 자기상관성(Auto-Correlation) 문제를 해소하기 위하여 (4)와 같이 코크래인-오컷 추정량(Cochrane-Orcutt estimation)을 활용한다.

$$dHTC = \beta_0 + \beta_1 GDP + \beta_2 dHPX + \beta_3 dGFC + \beta_4 dM2 + \epsilon \quad (1)$$

$$dHOC = \beta_0 + \beta_1 GDP + \beta_2 dHPX + \beta_3 dGFC + \beta_4 dM2 + \epsilon \quad (2)$$

$$Y = \delta_0 + \delta_i X_i + \epsilon \quad (3)$$

$$Y_t - \rho Y_{t-1} = \delta_0(1-\rho) + \delta_1(X - \rho X_{t-1}) + \nu_t \quad (4)$$

가계부채의 형성에 영향을 미치는 주요 요인을 분해하여 분석함으로써 적절한 회귀모형을 구축한 뒤 가계대출 관련 주요 정책효과에 대한 더미변수를 활용해 질적 분석을 더한다. 2008년 글로벌 금융위기 이후 금융시장 불안 및 자산 가격 하락 등이 이어지며 정부는 경기회복을 위한 다양

한 제도적 대안을 제시했다. 가계대출 부문에 있어서도 주택담보대출에 대한 LTV 및 DTI관련 규제를 완화하는 방향으로 대응했다. 반면, 2010년 중반 이후 경기회복이 완연해지고 최근 주택가격 급등 등에 대응하여 주택담보대출에 대한 LTV 및 DTI 등 대출규제를 강화해 왔다.

따라서 LTV 및 DTI 등 대출규제를 중심으로 대출규제가 완화된 시기와 강화된 시기를 구분하여 더미변수를 부여한다. 경제 및 정책 환경을 반영한 기본 모형에 대하여 위 대출규제에 대한 Table 4의 더비변수를 식(5)와 같이 반영하여 가계대출 관리를 위한 제도적 개입의 효과를 합리적으로 추출해 내고자 한다.

$$Y = \delta_0 + \delta_t X_t + \beta_j Dummy_j + \epsilon \tag{5}$$

$$Dummy_1 = Loan\ stimulus\ policy(1, \text{or}\, 0)$$
$$Dummy_2 = Loan\ restraint\ policy(1, \text{or}\, 0)$$

Table 4. Policy defining dummy variables

Dummy variables	Policy content	Application period
Loan stimulus policy	Relaxation of LTV and DTI application requirements	2010y 3Q~2017y 1Q
Loan restraint policy	Strengthening LTV and DTI application requirements	2017y 3Q~2022y 2Q

2. 분석 결과

2.1. 회귀분석 결과

가계부채에 대한 기본모형 회귀분석 결과는 다음 Table 5.와 같다. 경제 환경 요인 중 경제성장률(GDP)보다는 주택가격지수(dHPX) 요인이 총 가계신용(dHTC) 증가에 미치는 효과가 명확한 것으로 나타난다. 그리고 정책요인의 경우 통화량 증가(dM2)가 가계신용 증가에 미치는 영향은 다소 불확실한 반면, 재정지출(dGFC)은 가계신용 감소에 미치는 효과가 일부 있는 것으로 판단된다.

다만, 주택담보대출을 제외한 기타 대출(dHOC)에 대하여는 주택가격(dHPX)과 재정지출(dGFC)의 효과가 더욱 큰 것으로 나타나고, 통화팽창(dM2) 요인의 유의성은 상대적으로 더욱 낮은 것으로 확인된다. 그리고 총 가계신용(dHTC) 및 기타 대출(dHOC) 양측의 절편값(Intercept)은 모두 통계적으로 유의한 것으로 추정된다.

기본 모형(Basic)은 통계적 관점에서 대체로 타당하지만 일부 독립변수의 설명력이 낮고, 잔차의 자기상관 우려가 제기된다. 전체 가계대출(dHTC) 및 주택담보대출 제외 기타 대출(dHOC)에 대한 주택가격(dHPX)와 재정지출(dGFC)의 통계적 유의성은 각각 5% 및 1% 수준에서 유의하다. 다만, Ad.R2값이 0.318 및 0.331로 다소 낮고, D-W값이 1.18~1.20수준으로 자기상관 우려가 있다. 한편, 설명변수간 최대 VIF값은 1.19수준으로 다중공선성 우려는 낮고, 잔차에 대한 J-B검정 확률값은 0.27~0.72 수준으로 정규분포를 따르는 것으로 관측된다. 따라서 일부 설명력이 낮은 변수를 제외하고, 자기상관 위험에 대한 문제를 치유할 필요가 있다.

띠리서 기본 모형으로부터 설명력이 낮은 변수를 제외하고, 자기상관 우려를 치유하기 위하여 코크래인-오컷 추정량(Cochrane-Orcutt estimation)을 반영한 최적화 모형(Optimal)을 추가적으로 분석해 보았다. 해당 모형은 모두 Ad.R2값이 0.451~0.456 수준에서 상대적으로 양호한 설명력으로 갖는 것으로 확인된다. 또한 코크래인-오컷 추정량을 활용함에 따라 2.14~2.27수준의 D-W값이 추출되어

잔차의 자기상관이 상당부문 개선된 것으로 관측된다. 그리고 설명변수간 최대 VIF값은 1.00 수준으로 변수 간 다중공선성 문제 역시 낮다. 나아가 잔차에 대한 J-B검정 확률값 역시 0.46~0.43수준임에 따라서 대체로 정규분포를 따르는 것으로 관측된다.

앞서 모형 분석 결과를 종합하여 보자면, 가계대출의 증감은 주택가격 및 재정지출에 따른 효과가 큰 것으로 추정된다. 그리고 전체 가계대출 보다는 주택담보대출을 제외한 기타대출 모형에 있어 주택가격과 재정지출의 모수가 보다 높게 관측되어 보증부 대출이나 신용대출에 대하여 주택가격의 변동이나 재정지출의 증감이 미치는 효과가 상대적으로 더욱 큰 것으로 추정된다.

Table 5. Regression analysis result of basic model

	Dependent variable = dHTC				Dependent variable = dHOC			
	Basic		Optimal		Basic		Optimal	
	Coef.	t-stat	Coef.	t-stat	Coef.	t-stat	Coef.	t-stat
Intercept	0.0125	4.274***	0.0152	7.876***	0.0118	2.963***	0.0139	5.075***
GDP	0.0014	1.231			0.0018	1.185		
dHPX	0.1652	2.007**	0.3031	3.231***	0.2720	2.426**	0.4053	3.053***
dGFC	-0.0133	-4.287***	-0.0149	-6.515***	-0.0191	-4.515***	-0.0215	-6.664***
dM2	0.1968	1.406			0.1301	0.683		
Ad.R2	0.318		0.451		0.331		0.456	
D-W	1.18		2.27		1.20		2.14	
C-O's σ			0.454				0.461	
VIF	1.19		1.00		1.19		1.00	
J-B	0.72		0.46		0.27		0.53	

Note. The asterisk ***, **, and * mean statistical significances at 1%, 5%, and 10% level

4.2. 더미변수를 통한 정책효과 분석

가계부채에 대하여 주택가격지수, 재정정책 요인을 반영한 모델을 정책효과에 대한 더미변수를 추가하여 분석했다. 회귀분석 결과는 다음 Table 6과 같다. 대출 부양정책과 억제 정책 모두 대체로 유의한 결과가 도출되었으나, 부양정책 보다는 억제정책의 효과가 다소 더욱 뚜렷한 것으로 확인된다. 다만, 가계대출 전체에 대하여는 대책의 효과가 10% 유의수준 이내에서 유효한 것으로 확인 되지만, 주택담보대출을 제외한 경우 대책의 유효성이 유의수준 이내로 확인되지 않는다.

더미변수를 반영한 모델은 통계적 관점에서 전체적으로 타당하다. 해당 모형은 모두 Ad.R2값이 0.45~0.47 수준에서 양호한 설명력으로 갖는 것으로 확인된다. 또한 코크래인-오컷 추정량을 모두 활용하여 2.07~2.25수준의 D-W값이 추출되어 자기상관이 상당부문 해소된 것으로 관측된다. 그리고 설명변수간 최대 VIF값은 1.02~1.13수준으로 변수 간 다중공선성 문제 역시 낮다. 나아가 잔차에 대한 J-B검정 확률값 역시 0.25~0.47수준이며 대체로 정규분포를 따르는 것으로 관측된다.

앞서 정책 더미변수 반영 모형 분석 결과를 종합하여 보자면, 가계대출의 증감은 대출관련 제도 변화에 따른 효과가 일부 영향을 미치는 것으로 추정된다. 다만, 해당 정책이 주택담보대출에 대한 LTV나 DTI에 대한 규제로 적용되기 때문에 주택담보대출을 제외한 기타 보증부 및 신용대출 보다는 전체 대출에 대한 영향이 제한적으로 유의한 수준(10%)임이 관측된다. 이는 역설적으로 주택담보대출에 대한 LTV 및 DTI 규제의 완급이 가계대출 규모에 대체로 유의한 영향력을 행사한다는 점을 확인해 준다.

Table 6. Regression analysis result of policy dummy model

	Dependent variable = dHTC				Dependent variable = dHOC			
	PLUP Dummy		PLDN Dummy		PLUP Dummy		PLDN Dummy	
	Coef.	t-stat	Coef.	t-stat	Coef.	t-stat	Coef.	t-stat
Intercept	0.0129	5.856***	0.0170	9.834***	0.0119	3.525***	0.0154	5.563***
dHPX	0.3086	3.486***	0.3303	3.741***	0.4149	3.153***	0.4357	3.265***
dGFC	-0.0149	-4.468***	-0.0144	-6.028***	-0.0215	-6.591***	-0.0212	-6.403***
PLUP	0.0051	1.816*			0.0042	0.989		
PLDN			-0.0057	-1.993*			-0.0050	-1.126
Ad.R2	0.469		0.469		0.456		0.457	
D-W	2.25		2.07		2.14		2.07	
C-O's σ	0.385		0.325		0.436		0.406	
VIF	1.02		1.13		1.02		1.13	
J-B	0.27		0.47		0.36		0.25	

Note. The asterisk ***, **, and * mean statistical significances at 1%, 5%, and 10% level

Proceedings of the Korean Data Analysis Society
January 26-27, 2023, pp. 89-89

항만 개발과 운영이 미치는 사회적 영향에 관한 연구*

박호[1], 김상열[2], 김태훈[3], 장현미[4]

요 약

항만은 수출 위주의 경제성장 정책에서 중요한 역할을 담당하고 있으며, 우리나라는 국가 주도의 항만개발을 통해 경제 성장을 지원하고 있다. 항만의 개발의 영향과 효과에 대해서 다수의 선행연구들이 진행되었으나, 대부분의 선행연구는 경제적 효과를 중심으로 이루어졌다. 과거 산업화 시대는 공급 관점, 산업화 이후에는 수요와 효율성 관점으로 변화되었고, 최근에는 지속가능한 발전에 대한 국제적인 관심과 국민의식이 증대되고 있다. 특히, 항만은 도시와 다양한 단계를 거쳐 상호 영향을 주고받으며 진화하고 있으며, 항만을 포함한 타 SOC 산업이 지역사회 및 해당 지역 주민에 기여하는 사회적 가치가 크다는 점에 주목한 국내외 관련 연구가 증가하는 추세임을 고려하면 지역사회에 끼치는 경제적 영향뿐만 아니라 사회적 영향을 평가할 필요가 존재한다. 따라서, 본 연구에서는 항만의 사회적 영향과 효과에 대해 긍정과 부정적 영향을 포함한 평가모형을 구축하여 분석을 실시하고 시사점을 도출하였다. 이와 함께 항만 개발시 고려되었던 요인의 중요도 및 실제 이용자와 관리자가 생각하는 성과(만족도)를 분석하여 항만 개발의 효과를 높이기 위한 방안을 모색하였다. 연구의 분석 결과로서, 첫째, 항만의 개발의 사회적 영향에 대한 인식에서 부정적 영향이 높은 중요도를 보였으나, 지속적인 항만의 개발은 안정적인 공급망 확보를 위한 선결조건으로서 항만과 개발의 사회적 부정적인 영향을 최소화하고 긍정적인 영향을 극대화할 수 있는 방안 마련이 모색되어야 할 것이다. 둘째, 항만 개발의 사회적 영향에 대해서 공공/관리기관, 학계, 업계(이용자)와 기타로 구분하여 응답결과를 비교분석한 결과, 공공/관리기간과 학계보다는 업계(이용자)와 기타 집단이 부정적 영향효과에 상대적으로 높은 중요도를 나타내었다. 셋째, 항만 개발의 효과에 대해서 연관산업의 발달·기업의 집적화 및 기업 유치, 일자리/고용창출(고용의 질 향상), 물류/상업/도시/주거 기능의 확대(확보)로서 지속적인 유지 및 관리 전략 필요하며, 교통인프라의 확충, 지역 개발의 촉진(지역균형 발전)에 대해서 중요도에 비해 성과가 떨어져 현재 수준보다 높은 성과를 얻기 위한 집중적인 노력 요구되는 것으로 분석되었다.

주요용어 : 항만 개발, 항만 운영, IPA, AHP, 대응표본 t-검정.

---

*이 과제는 부산대학교 기본연구지원사업(2년)에 의하여 연구되었음.
[1](주저자) 전라북도 군산시 대학로 558, 군산대학교 국제물류학과 부교수. E-mail: hpark0321@kunsan.ac.kr
[2](교신저자) 부산시 금정구 부산대학로길 63번길 2, 부산대학교 국제전문대학원 교수. E-mail: ksy@pusan.ac.kr
[3]부산시 남구 수영로 309 경성대학교 경제금융물류학부 교수. E-mail: kdbdc@ks.ac.kr
[4]부산시 금정구 부산대학로길 63번길 2, 부산대학교 국제전문대학원 부교수. E-mail: jangh01@pusan.ac.kr

Proceedings of the Korean Data Analysis Society
January 26-27, 2023, pp. 91-94

비대칭 마진 SVM을 이용한 기업부실 예측모형의 범주 불균형 문제 해결*

조성임[1], 김명종[2]

요 약

본 연구는 범주 불균형 문제의 대표적인 예인 기업부실 예측모형의 성능향상을 위하여 대칭 마진 SVM(EMSVM)을 개선한 비대칭 마진 SVM(UMSVM)을 적용하고자 한다. 주요 실험 결과는 범주 불균형이 미미한 표본에서 UMSVM은 EMSVM의 성과 개선 효과가 미약한 반면, 범주 불균형이 심화된 표본에서 UMSVM은 EMSVM의 성과개선에 크게 공헌하고 있다.

주요용어 : 비대칭 마진 SVM, SVM, 범주 불균형, 부실예측, UMSVM.

1. 서론

기업부실은 국가경제에 막대한 손실을 초래하는데 이러한 기업의 부실 가능성을 사전에 예측하기 위해 정교한 예측모형의 개발은 매우 중요하다. 전통적인 통계모형 뿐만 아니라 (Support Vector Machine, SVM) 또는 의사결정트리 등의 다양한 머신러닝 기법을 적용하여 예측모형의 정확도를 개선시킬 수 있다. 특히, SVM은 구조화된 정규화 모수를 적용하여 일반화 능력이 탁월하여 부실예측 문제에서 우수한 성과를 보여주었다(Shin et al., 2015).

하지만 특정 범주의 데이터의 수가 다른 범주의 데이터의 수에 비하여 매우 많을 경우 발생하는 범주 불균형 문제는 예측모형의 성과 개선에 부정적인 영향을 미치는 것으로 알려져 있다. 이 경우, SVM은 다수 범주와 소수 범주 간 데이터 불균형 비율(Imbalance Ratio, IR)이 높아질수록 다수 범주의 경계영역이 확장되어 소수 범주 경계영역이 축소되고 소수 범주의 정확성(민감도)는 급격하게 감소한다(Li and Shawe-Taylor, 2003).

범주 불균형 문제를 해결하기 위하여 SVM의 일부 모듈을 수정한 비대칭 마진 SVM(Uneven Margin SVM, UMSVM)이 활용되고 있다. UMSVM은 전통적인 대칭 마진 SVM(Even Margin SVM, EMSVM)이 범주 간 동일 마진을 유지함으로써 범주 불균형 문제에 효과적으로 대처할 수 없다는 한계점을 개선하기 위하여, 다수 범주 마진에 비하여 소수 범주 마진을 크게 적용한다. 그 결과, 소수 범주의 경계영역을 확대하여 민감도를 개선하고, 과적합 문제에 강건하다는 장점이 있다(Li and Shawe-Taylor, 2003). 이러한 장점으로 문서분류, 이미지 인식 등 많은 분야에서 적용되고 있으나 비즈니스 분야에서 실증 적용된 사례는 극히 드물다. 본 연구의 공헌점은 비즈니스 분야에서

*본 연구는 과학기술정보통신부 및 정보통신기획평가원의 대학 ICT연구센터 육성지원사업의 연구결과로 수행되었음(IITP-2022-2020-0-01797).
[1]46241 부산광역시 금정구 부산대학로 63번길 경영관 부산대학교 경영학과 박사과정.
E-mail: reina337@pusan.ac.kr
[2]46241 부산광역시 금정구 부산대학로 63번길 경영관 부산대학교 경영학과 교수. E-mail: mjongkim@pusan.ac.kr

UMSVM의 도입 타당성에 대한 실증적 자료와 이론적 근거를 제공하는 것을 목적으로 한다.

본 연구의 주요 분석 결과로 범주 불균형이 미약한 표본에서는 EMSVM에 대한 UMSVM의 성과 개선 효과가 크게 나타나지 않은 반면, 범주 불균형이 심화된 표본에서 UMSVM은 EMSVM의 성과 개선에 크게 공헌하고 있음을 확인하였다.

본 연구는 다음과 같이 구성된다. 제 2 장은 EMSVM과 UMSVM의 알고리즘이다. 제 3 장은 연구 설계에 대하여 설명한다. 제 4 장에서 연구결과를 제시하고 제 5 장은 결론이다.

2. UMSVM의 최적화 알고리즘

EMSVM은 일반적으로 알려진 soft margin SVM과 목적식과 제한식이 동일하며 식(1)과 같이 정의된다(Cortes et al., 1995).

$$
\begin{aligned}
&function && \min < w,w > + C\sum_{i=1}^{n}\xi && ① \\
&constraints && < w,x^{+} > + b \geqq +1-\xi \quad \text{if } y_i = +1 && ② \\
& && < w,x^{-} > + b \geqq -1-\xi \quad \text{if } y_i = -1 && ③ \\
&\xi \geqq 0 && i = 1,...,n && \text{식(1)}
\end{aligned}
$$

UMSVM은 소수 범주의 마진을 다수 범주의 마진에 비하여 상대적으로 확장하여 민감도를 높이는 것을 목적으로 한다. 소수 범주의 마진(m^{+})과 다수 범주의 마진(m^{-})의 상대적인 비율인 $\frac{m^{-}}{m^{+}}$를 마진 파라미터(τ)라 할 때 UMSVM의 목적식과 제한식은 식(2)와 정의된다(Li and Shawe-Taylor, 2003).

$$
\begin{aligned}
&function && \min < w,w > + C\sum_{i=1}^{n}\xi && ① \\
&constraints && < w,x^{+} > + b \geqq +1-\xi \quad \text{if } y_i = +1 && ② \\
& && < w,x^{-} > + b \geqq -\tau-\xi \quad \text{if } y_i = -1 && ③ \\
&\xi \geqq 0 && i = 1,...,n && \text{식(2)}
\end{aligned}
$$

3. 연구설계

3.1. 표본수집과 변수선정

본 연구의 표본은 2015부터 2018년까지 국내 비금융업 외부 회계감사 법인을 대상으로 수집하였다. 국내 시중은행으로부터 수령한 데이터를 기준으로 500개 부실 기업을 선정하였으며, 7,500개 정상 기업으로 선정하여 총 8,000개의 표본을 구성하였다. 또한 불균형 비율에 따라 UMSVM이 EMSVM의 성과 개선에 미치는 효과를 확인하기 위하여 부실 기업을 기준으로 불균형 비율이 A(1:1), B(1:2), C(1:4), D(1:10) 그리고 E(1:15)로 상이한 5개의 하위 표본군을 구성하였다.

본 연구에 사용되는 변수를 선정하기 위하여 선행연구에서 사용된 비율을 기준으로 수집된 재무비율을 7개 재무비율 군(수익성, 부채상환능력, 레버지리, 자본구조, 유동성, 활동성, 규모)으로 분류하였으며, 최종 입력변수는 각 재무비율 군별로 AUC가 가장 높은 7개의 재무비율을 최종적으로 선정하였다(Kim et al, 2015).

3.2. 성능지표

현재 가장 보편적으로 사용되는 성과 측정치는 정확도이지만 범주 불균형 문제에서 다수 범주의 특이도에 크게 의존하며, 소수 범주의 민감도를 고려하지 못하는 단점이 있다. 이러한 문제점을 해결하기 위하여 기하평균 정확도와 같은 대체적인 성과지표들이 제안되어 왔다(Kubat et al., 1997). 본 연구는 부실 기업의 정보의 중요성을 감안하여 소수 범주의 민감도와 기하평균 정확도를 주요 성능지표로 활용하고자 한다. Table. 1은 이진 분류 모형의 정오분류표와 모형의 성과 측정에 사용되는 특이도, 민감도, 정확도 및 기하평균 정확도의 산출방법에 대하여 설명하고 있다.

Table 1. The concept of the confusion matrix and various performance measurement

		Prediction	
		Positive	Negative
Real	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

특이도(Specificity, SPE) : TN / (FP + TN)
민감도(Sensitivity, SEN) : TP / (TP + FN)
정확도(Arithmetic Accuracy or accuracy, ACC): (TP + TN) / (TP + FN + FP + TN)
기하평균정확도(Geometric Mean, GM) = $\sqrt{\text{민감도} \times \text{특이도}}$

4. 연구결과

본 연구에서는 입력 변수를 고차원의 입력 벡터로 전환하기 위하여 RBF 커널을 활용하였으며 RBF 커널의 분산과 손실계수는 그리드 탐색 기법을 적용하였다.

Table. 2는 데이터의 불균형 비율을 기준으로 조정한 하위 표본을 대상으로 UMSVM의 τ를 다양하게 조정하여 분류한 결과를 제시하고 있다. UMSVM은 마진의 비대칭성이 증가할수록 소수 범주의 경계영역이 확대되어 소수 범주의 민감도가 개선되므로 범주 불균형 문제를 완화할 수 있다. 마진의 비대칭성을 나타내는 마진 파라미터 τ는 0과 1 사이의 실수 값에서 결정된다.

Table. 2의 분석 결과를 요약하면 첫째, 모든 데이터 표본에서 τ가 낮을수록 특이도는 감소하는 경향을 보이고 있다. 이는 τ가 낮을수록 UMSVM의 분류 경계선의 이동 거리가 커져 다수 범주의 경계영역이 축소되고 이에 따라 다수 범주의 특이도가 감소함을 의미한다. 둘째, 모든 데이터 표본에서 민감도는 τ가 낮을수록 크게 개선되고 있으며, 특히 범주 불균형 상황이 심각할수록 그 효과가 크게 나타나는 것으로 분석되었다. 셋째, 정확도는 τ가 작아질수록 지속적으로 완만하게 감소하는 경향을 보인다. 이는 정확도가 다수 범주의 특이도에 크게 의존하고 소수 범주의 민감도에 대해서는 민감하게 반응하지 않는 특성을 가지고 있기 때문이다. 마지막으로, 기하평균 정확도는 데이터 표본별로 일관되지 않은 결과를 보이는데 이는 의 증가에 따라 특이도는 감소하는 반면, 민감도는 증가하는 상충효과가 범주 불균형에 따라 상이함을 의미한다.

Table 2. Comparison of classification performance by data sample according to τ

	특이도					민감도				
Data Sets/τ	0.001	0.25	0.5	0.75	1	0.001	0.25	0.5	0.75	1
A (1:1)	0.88	0.93	0.96	0.97	0.98	0.99	0.97	0.96	0.96	0.95
B (1:2)	0.89	0.94	0.96	0.97	0.98	0.96	0.94	0.94	0.92	0.91
C (1:4)	0.90	0.95	0.96	0.97	0.98	0.85	0.82	0.81	0.78	0.77
D (1:10)	0.90	0.96	0.98	0.98	0.99	0.68	0.44	0.34	0.30	0.27
E (1:15)	0.93	1	1	1	1	0.34	0.03	0.01	0.01	0.01
	정확도					기하평균 정확도				
Data Sets/τ	0.001	0.25	0.5	0.75	1	0.001	0.25	0.5	0.75	1
A (1:1)	0.93	0.95	0.96	0.96	0.96	0.93	0.95	0.96	0.96	0.96
B (1:2)	0.91	0.94	0.95	0.96	0.95	0.92	0.94	0.95	0.95	0.94
C (1:4)	0.89	0.91	0.92	0.92	0.92	0.87	0.87	0.87	0.87	0.86
D (1:10)	0.88	0.92	0.92	0.92	0.92	0.78	0.65	0.58	0.55	0.51
E (1:15)	0.90	0.94	0.94	0.94	0.94	0.53	0.16	0.09	0.08	0.07

5. 결론

본 연구는 범주 불균형 환경에서 SVM의 분류 성과 개선을 위하여 소수 범주에 편향된 분류 경계선을 다수 범주로 재이동시켜 소수 범주의 민감도를 개선하기 위하여 EMSVM의 알고리즘을 수정한 UMSVM을 기업부실 예측모형에 적용하였다. 분석 결과 UMSVM은 범주별 데이터 분포가 유사한 균형 데이터에서는 EMSVM의 성과 개선 효과에 유의적이지 못하지만, 범주 불균형이 심화된 데이터에서 EMSVM의 성과를 크게 개선함을 확인하였다. 본 연구는 SVM의 일반화 능력을 최적화하기 위한 이론적 배경과 실증 자료를 제공함으로써 범주 불균형 문제에 대한 SVM의 강건성을 제고한다는 공헌점이 있다.

References

Cortes, C., Vapnik, V., Saitta, L. (1995). “Support-Vector Networks Editor”, Machine Learning, *Kluwer Academic Publishers*, 273-297.

Kim, M. J., Kang, D. K., Kim, H. B. (2015). “Geometric mean based boosting algorithm with over-sampling to resolve data imbalance problem for bankruptcy prediction”, *Expert Systems with Applications*, 42(3), 1074 - 1082.

Kubat, M., Holte, R., Matwin, S. (1997). “Learning when negative examples abound”, ECML 1997: Machine Learning: *ECML-97*, 146-153.

Li, Y., Shawe-Taylor, J. (2003). “The SVM With Uneven Margins and Chinese Document Categorisation”, Language, Information and Computation: Proceedings of the 17th Pacific Asia Conference, Sentosa, *Singapore*, 1-3.

Shin, K., S., Lee, T., S., Kim, H., J., (2015). “An application of support vector machines in bankruptcy prediction model”, *Expert Systems with Applications*, 28(1), 127-135.

Proceedings of the Korean Data Analysis Society
January 26-27, 2023, pp. 95-95

통계용어 표준화 및 체계적 구축방안에 대한 연구

한상태[1], 장원철[2], 여인권[3], 임종호[4]

요 약

우리나라는 분산형 통계 제도에 따라 국가통계를 생산하고 있으며, 통계작성기관은 중앙행정기관, 지방자치단체 및 통계작성 지정기관(민간기관) 등으로 구분되어 있다. 분산형 통계 제도에 따라 국가통계를 생산하고 있는 경우 통계자료 수집, 활용, 해석과정에서 정확하고 표준화된 통계용어를 사용할 수 있도록 국가통계 작성기준(표준용어)을 제공하는 것이 통계 생산자와 이용자의 통계 이용상의 혼란을 막고 이해도를 높여서 국가통계의 활용도를 높일 수 있다. 본 연구는 통계용어의 표준화된 개념 및 정의 등을 적용함으로써 국가통계의 일관성을 제고하고 국제간 원활한 비교를 위해 통계용어의 체계적 구축(온라인사전 시스템, e-book)을 위한 것이다. 구체적으로 국가승인통계를 작성하고 관리하는 기관별로 통계용어집, 통계간행물·보도자료·설명자료를 종합 검토하여 관련 통계용어를 확정하고, 국제간 비교를 위한 용어 매칭을 위해 Eurostat, OECD와 같은 국제기구와 미국, 캐나다와 같은 통계선진국에서 제공하는 glossary를 검토하고 참고해서 확정된 통계용어에 대해 우리나라 통계용어사전 내용을 작성하고자 하는 것이다. 용어사전에서의 용어 설명은 (1) 정의, (2) 해설(가능한 경우 인포그래픽 포함), (3) 관련 용어, (4) 관련 통계조사명, (5) 관련 영문용어, (6) 분류의 여섯 가지 항목으로 구성하였다. 관련 영문용어는 적용 가능한 통계용어에 대해서 국제기구에 수록된 용어 설명이나 국가별 통계자료 수록 웹페이지로 연결하였다. 2019년부터 2022년까지 진행된 연구를 바탕으로 2023년에는 주제별 온라인 사전과 검색 온라인 사전으로 구성된 시스템 개발이 이루어질 예정인데, 사용자가 원하는 용어를 편리하게 검색할 수 있도록 시스템이 개발될것으로 기대된다.

주요용어: 국가통계, 승인통계, 통계용어, 온라인사전 시스템, 통계용어사전.

[1]31499 충청남도 아산시 배방읍 호서로 79번길 20, 호서대학교 빅데이터AI학과 교수.
[2]08826 서울시 관악구 관악로1, 서울대학교 통계학과 교수.
[3]04310 서울시 용산구 청파로47길 100(청파동2가), 숙명여자대학교 통계학과 교수.
[4]03722 서울시 서대문구 연세로 50, 연세대학교 응용통계학과 부교수.

Proceedings of the Korean Data Analysis Society
January 26-27, 2023, pp. 97-100

안전에 대한 객관적 지표와 주관적 인식 차이: 사회안전지수를 중심으로

박현수[1], **장안식**[2]

1. 들어가며

안전에 대한 평가는 특정 지역의 환경이 가지고 있는 위험에 따른 평가가 있을 수 있고, 이 환경의 위험에 대한 평가를 통해 인식되는 개인의 평가가 있을 수 있다. 지역의 안전을 위해 각 지자체에서는 많은 노력을 기울이고 있지만, 그 노력만큼 지역 주민의 안전에 대한 인식으로 이어지는 지에 대해서는 의문이다.

안전에 대한 객관적인 평가도 중요하지만 지역 주민이 가지고 있는 안전에 대한 주관적 인식은 삶의 질과 직결되므로 매우 중요한 부분이다. 이 연구에서는 안전에 대한 객관적 평가와 주관적 인식 간의 차이가 있는지 살펴보고자 한다. 그리고 이들 간의 차이가 있다면 그 특성은 무엇인지 설명하고자 한다.

2. 연구방법

2.1. 연구자료

이 연구의 연구자료로는 「2023 사회안전지수」의 자료를 사용하도록 한다. 이 자료는 시민들의 안전이나 불안감에 영향을 주는 4개 차원(경제활동, 생활안전, 건강보건, 주거환경)으로 구성하였고, 각 차원에 대한 사회안전도를 3~4개의 세부영역(경제활동 차원에서는 소득, 고용, 복지, 노후, 생활안전 차원에서는 치안, 소방, 교통안전 등)으로 구분하였다. 각 세부영역을 구성하는 지표는 정부나 지자체 통계자료를 이용한 정량지표와 해당 시군구 주민들의 체감도를 측정한 정성지표로 구성하였다.

사회안전지수는 4개 차원을 구성하는 세부영역별로 지표항목을 정량지표와 정성지표로 구분하고, 각 지표별로 표준점수를 산출하고, 각 차원별 정량지표와 정성지표를 합산하여 차원별 점수를 계산하였다. 이렇게 계산된 4개 차원의 점수를 합산가중치(정량지표 60%, 정성지표40%)에 따라 합산하여 사회안전지수를 산출하였다. 그리고 이 지수는 인구수가 적어 정성지표를 산출하기 위해 충분한 조사표본을 확보하기 어려운 시군구를 제외하고, 전국 184개 시군구를 대상으로 조사되었다.

2.2. 분석방법

이 연구는 사회안전지수 자료를 이용하여 안전의 객관적 수준과 주관적 수준의 차이를 살펴보고자 한다. 즉 정량지표 점수와 정성지표 점수를 이용하여 각 지표 간 점수의 차이를 보이는 지역의 특성을 파악하고자 한다.

---

[1]충북대학교 국가위기관리연구소 시민치안연구센터장.
[2]케이스탯컨설팅 공공사회정책연구소장.

정량지표 점수와 정성지표 점수를 모두 표준점수로 변환하여 각 지표 점수에 대한 지역별 차이를 살펴보도록 한다. 그리고 표준화된 정량지표 점수와 정성지표 점수의 차이 점수를 통해 어떤 지역에서 정량지표 점수가 정성지표 점수에 비해 더욱 높은지, 혹은 반대의 경우인지 살펴보고자 한다. 이를 위해 표준화된 정량지표 점수와 정성지표 점수에 대한 지역별 특성을 파악하기 위해 집단으로 구분하였다. 우선 정량지표 점수와 정성지표 점수의 지역별 분포를 살펴보기 위해 표준화된 각 지표 점수의 수준에 따라 집단으로 구분하였다. 이들 집단으로 구분하는 기준은 각 지표의 표준점수가 0.5점 초과면 '높음' 집단, 0.5점 이하에서 -0.5점 이상이면 '중간' 집단, 그리고 -0.5점 미만이면 '낮음' 집단으로 구분하였다. 그리고 정량지표 점수와 정성지표 점수의 차이에 대해서도 정량지표의 표준점수에서 정성지표의 표준점수를 뺀 점수를 기준으로 0.5점 초과면 '정량지표 우위' 집단, 0.5점 이하에서 -0.5점 이상이면 '상대적으로 차이 없음' 집단, 그리고 -0.5점 미만이면 '정성지표 우위' 집단으로 구분하였다.

3. 연구결과

사회안전지수에서는 점수에 따라 지역의 안전 등급을 A 등급에서 E 등급으로 구분하고 있다. 이 등급에 따른 정량지표와 정성지표 점수에 대한 각 지역의 산포도는 Figure 1과 같다.

정량지표와 정성지표 점수의 평균값을 기준으로 1사분면와 3사분면의 경우에는 대체로 이 두 지표의 점수 차이가 그리 크지 않다는 것을 알 수 있다. 그러나 2사분면과 4사분면에 위치한 지역의 경우에는 두 지표의 점수 차이가 상대적으로 클 것이고, 적지 않은 지역이 이들 사분면에 해당되고 있는 것으로 나타났다.

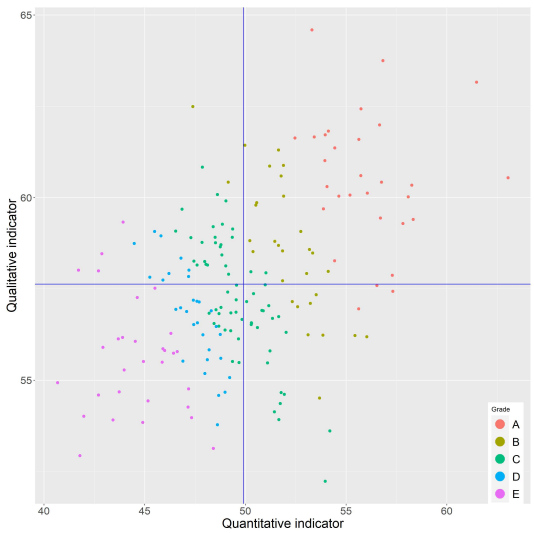
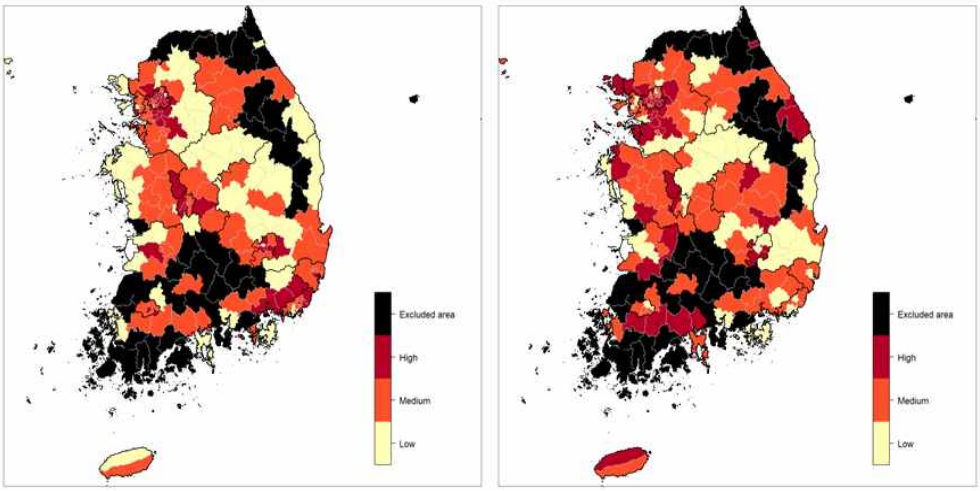


Figure 1. Distribution of qualitative and quantitative indices

정량지표와 정성지표 점수에 따른 지역별 분포를 살펴보면 Figure 2와 같다. 정성지표 점수에 비해 정량지표 점수의 경우에 '중간' 집단의 분포가 더욱 넓게 나타나고 있다. 특히, 경기도 북동부, 충청남도 북서부, 그리고 경상북도 북부의 경우에는 정성지표에서는 '낮음' 집단이었으나 정량지표

에서는 '중간' 집단이었다. 그리고 전라남도 남부에서는 정성지표에서는 '중간' 집단이었으나 정량지표에서는 '높음' 집단으로 나타났다. 이와 반대로 경상남도 남부에서는 정성지표에서는 '높음' 집단이었으나 정량지표에서는 '중간' 혹은 '낮음' 집단으로 나타나고 있다.



(a) Qualitative indicator (b) Quantitative indicator

Figure 2. Distribution of qualitative and quantitative indices by region

정량지표와 정성지표 간 점수 차이를 좀 더 명확히 살펴보기 위해 이들 지표의 지역별 표준점수 차이를 보면 Figure 3과 같다. 정량지표와 정성지표 간 점수의 차이를 보인 지역을 중심으로 살펴보면, 우선 정량지표 우위인 지역으로는 경기도 동북부, 강원도 동부, 충청남도 서부와 남부, 경상북도 북부, 전라남도 남부, 그리고 제주도 북부로 나타나고 있다. 다음으로 정성지표 우위인 지역으로는 서울, 인천, 부산 등 광역시 지역의 일부와 강원도 북부와 남부, 충청북도 북부인 것으로 나타나고 있다.

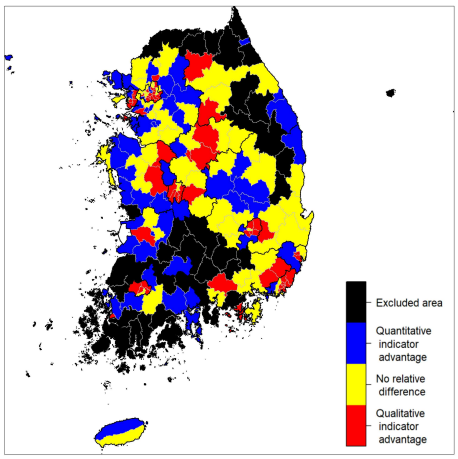


Figure 3. Regional distribution of differences in qualitative and quantitative indices

정성지표와 정량지표 간의 차이를 보인 지역의 특성을 살펴보기 위해 광역시 내 지역, 광역도 내 도시, 그리고 광역도 내 군 지역으로 구분하여 교차분석을 시행한 결과는 Table 1과 같다. 이 결과에서는 광역시 내 지역의 60.9%는 정성지표 우위인 지역인 것으로 나타나고 있다. 광역도 내 시 지역에서는 상대적으로 차이가 없는 지역이 42.9%로 가장 많은 것으로 나타났다. 그리고 광역도 내 군 지역에서는 정량지표 우위인 지역이 47.4%로 가장 많았다. 이러한 결과를 통해 광역시의 경우에는 객관적 안전 수준에 비해 주관적인 인식이 더욱 높은 것으로 나타난 반면, 군 지역에서는 객관적 안전 수준에 비해 주관적인 인식이 더 낮은 것으로 나타났다.

Table 1. Differences in qualitative and quantitative indices by regional types

	Counties within provinces		Cities within provinces		District within metropolitan city		Total	
Qualitative indicator advantage	4	(10.5%)	14	(18.2%)	42	(60.9%)	60	(32.6%)
No relative difference	16	(42.1%)	33	(42.9%)	21	(30.4%)	70	(38.0%)
Quantitative indicator advantage	18	(47.4%)	30	(39.0%)	6	(8.7%)	54	(29.3%)
Total	38	(100.0%)	77	(100.0%)	69	(100.0%)	184	(100.0%)

4. 결론

이 연구에서는 안전에 대한 객관적 평가와 주관적 인식 간의 차이를 살펴보고자 하였다. 그 결과 적지 않은 지역에서 이들 간의 차이가 있는 것으로 나타나고 있다. 특히 지역의 규모에 따라 광역시와 같은 대도시에서는 정량지표에 비해 정성지표의 점수가 높았고, 이에 비해 광역도의 군 지역에서는 반대의 경향이 나타났다. 안전에 대한 객관적 환경을 조성하는 것은 분명 중요한 부분이지만, 주관적 인식에 대한 향상도 주민의 삶의 질에 중요한 부분이므로 이를 위한 방안도 함께 모색하는 것이 필요할 것이다.

Proceedings of the Korean Data Analysis Society
January 26-27, 2023, pp. 101-105

가우시안 필터를 활용한 주가예측의 최적화*

김명종[1], 권성훈[2]

요 약

머신러닝과 딥러닝을 활용한 연구가 활발히 이루어지며 KOSPI 등 주가지수를 기계학습을 통해 예측하는 연구가 활발히 진행되고 있다. 그러나, 랜덤워크(Random-Walk) 데이터의 비선형성으로 인해 노이즈가 모델의 예측에 부정적인 영향을 미쳐 예측이 전날의 주가를 따라가는 후행성 문제가 발생한다. 따라서, 본 연구는 LSTM(Long Short-Term Memory)을 이용한 주가 예측에 디노이징 필터 중 하나인 가우시안 필터(Gaussian Filter)를 활용하여 스무딩 효과를 통해 후행성 문제를 해결하려 한다. 가우시안 필터는 값이 위치한 거리에 따라 가중치를 부여하여 효과적으로 노이즈를 제거한다는 장점이 있다. 예측을 위한 변수로 뉴스 키워드의 추출과 감성분석을 통해 감성지수(Sentiment Index)를 산출하고 코스피 지수의 기술적 분석을 통해 기술지수(Technical Index)를 산출하여 모델에 추가하여 주가 예측을 진행한 결과, 후행성 문제가 크게 개선됨에 동시에 모델의 예측 성능이 향상된 결과가 도출되었다.

주요용어 : 순환신경망, LSTM, 감성지수, 기술지수, 디노이징 필터.

1. 서론

KOSPI 주가지수의 변동성에 관한 연구가 활발하게 이루어지고 있으며, 주가 예측 연구는 크게 ARIMA 등의 시계열 분석 기법, 머신러닝(machine learning) ·딥러닝(deep learning) 기반 예측 연구와 텍스트마이닝(text mining) 기반 예측 연구로 구분할 수 있다. 현재 LSTM을 활용한 주가 예측에 대한 연구는 데이터의 노이즈로 인해 모델의 예측이 전날의 주가를 따라가는 후행성 문제가 있다. 이러한 후행성 문제의 원인은 랜덤워크(Random Walk) 데이터의 잡음으로 인해 인공신경망의 비선형적인 패턴의 추출이 불가하여 발생한다. 본 연구에서는 딥러닝의 순환신경망을 사용한 모델중 하나인 LSTM(Long Short-Term Memory)기법을 사용하여 KOSPI 지수의 종가를 예측하고자 한다. KOSPI 지수는 FnGuide API를 통해 데이터를 로딩하는 것으로 한다. KOSPI 종가 예측의 과정으로 디노이징 필터 중 하나인 가우시안 필터를 적용하여 노이즈를 제거하여 후행성 문제를 해결한다. 여기에, 뉴스빅데이터 분석시스템인 빅카인즈를 통해 뉴스키워드를 수집하고 감성분석(Sentiment Analysis)을 진행하여 감성지수를 산출한다. KOSPI 지수에 대한 기술적 분석을 통해 기술 지수를 산출한 후, 감성지수와 기술지수를 LSTM 모델에 적용하여 예측 성능을 향상시키고자 한다.

---

*본 연구는 과학기술정보통신부 및 정보통신기획평가원의 대학ICT연구센터육성지원사업의 연구결과로 수행되었음(IITP-2022-2020-0-01797).
[1]46241부산광역시 금정구 부산대학로 63번길 2, 부산대학교 경영학과 교수. E-mail: mjongkim@pusan.ac.kr
[2]46241부산광역시 금정구 부산대학로 63번길 2, 부산대학교 경영대학 학사과정. E-mail: rnjstjdgns216@pusan.ac.kr

2. 학습알고리즘

LSTM은 기존의 RNN이 출력과 먼 위치에 있는 정보를 기억할 수 없다는 단점을 보완하여 장/단기 기억을 가능하게 설계한 신경망의 구조를 의미한다. LSTM 네트워크는 Hochreiter and Schmidhuber(1997)에 의해 소개되었으며 기존의 바닐라 RNN의 장기 메모리를 보완하면서 메모리 셀과 게이팅 메커니즘을 확장한 것이다. 즉, RNN에 입력, 출력, 망각게이트가 추가되어 메모리 셀로부터 정보가 얼마나 유지될지 결정한는 기능을 한다. 시그모이드 층(Sigmoid Layer)에서 이전의 값을 입력받아 셀 상태(Cell State)에 업로드할지 결정된 이후 하이퍼볼릭 탄젠트 층(tanh layer)에서 새로운 후보 값들을 벡터로 만들어 셀 상태에 추가하게 된다. 즉, LSTM은 장기상태에 저장할 것, 읽어들일 것, 버릴 것을 학습하여 네트워크상 더 긴 기간 동안의 오류를 줄이는 작업을 거치게 된다.

가우시안 필터는 Karl Friedrich Gaussian이 제안한 가우시안 분포를 근사하여 생성한 필터 마스크를 사용하는 필터링 기법이다. 가우시안 분포는 평균을 중심으로 좌우 대칭의 종 모양을 갖는 확률분포로, 정규 분포(normal distribution) 이라고도 한다. 가우시안 필터는 주로 영상이나 이미지의 노이즈(Noise)를 처리하는 데 사용되며, 블러링(blurring) 기법을 통해 입력 영상이나 이미지를 부드럽게 만드는 스무딩(Smoothing) 기능을 한다. 가우시안 필터는 선형 필터로 평균값 필터(Mean Filter)와 다르게 가까운 픽셀은 큰 가중치를, 멀리 있는 픽셀은 작은 가중치를 두어 평균을 계산한다는 차이점이 있다. 가우시안 필터 마스크 행렬은 중앙부에서 큰 값을 가지고, 측면으로 갈수록 행렬의 원소 값이 0에 가까워지는 특성이 있다. 필터링 대상 픽셀과의 거리에 따라 가중치를 두어 가우시안 필터는 X,Y 방향의 시그마(Sigma)값에따라 스무딩 정도가 변화되는 특징이 있다. 2차원 가우시안 함수는 다음과 같다.

$$G(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}$$

3. 연구 방법론

본 연구에 활용된 주가 데이터 수집 기간은 2017년 1월 1일부터 2021년 12월 31일까지로, Finance Data Reader로부터 데이터를 수집하였다. 시계열 데이터의 효율적인 활용을 위해 일일 주기로 변수를 수집하였다. 가격 데이터는 KOSPI 지수의 저가/고가/시가/종가로 수집하였으며, 종속변수는 다음날의 KOSPI 지수의 종가로 두었다.

본 연구에 활용된 감성지수는 한국언론진흥재단에서 운영하는 뉴스 빅데이터 분석 서비스인 빅카인즈(Big kinds)에서 뉴스 분류(정치, 국제, 일반경제, 금융, 외환, 국제경제, 증권, 무역)에 따라 기사를 수집한 후, 감성분석을 실시하여 언론 감성지수를 계산하였다.

데이터 정제 이후 SO-PMI를 활용하여 감성분석의 기반이 되는 감성 사전을 구축하였다. 2017년 1월 1일에서 2021년 12월 31일 사이에 발행된 기사 중 30%를 무작위로 추출하여 감성 사전 학습에 이용하였다. 또한 SO-PMI를 활용하여 감성 사전을 구축하기 위해서는 긍정단어 집합과 부정단어 집합이 사전에 정의되어야 하는데, 이는 한국어 감성분석에서 흔히 이용되는 KNU 한국어 감성 사전을 이용하였다. 따라서 맞춤형 감성 사전을 일차적으로 구축한 후, 빈도수 상위 500개를 추

출하고 감성과 관련 없는 단어를 직접 제거하는 작업을 수행하였다. 이후 SO-PMI 점수를 -2에서 2의 범위로 정규화(normalization)한 후 단어의 극성을 강한 긍정(+2), 약한 긍정(+1), 약한 부정(-1), 강한 부정(-2)으로 재분류하였다(이상훈 외, 2016).

개별 뉴스의 감성지수는 단어별 감성 극성 점수를 활용하여 키워드에서 등장한 단어의 감성 극성 합으로 계산하였다. 단어의 감성 점수 합에 문장에서 등장한 단어의 개수를 나누어 문서의 길이가 감성지수에 영향을 미치지 않게 하였다. 이에 따라 경제/ 정치/ 국제별 뉴스에 따라 감성지수를 나누어 일자별 뉴스 감성지수의 합으로 시계열 데이터를 도출하였다.

본 연구에서는 13개의 기술적 지표가 기술적 분석을 통해 산출되었으며, 이 중 분산팽창지수(VIF)를 산출하여 다중공선성 문제가 있는 지표 4개를 제외한 7개의 지표를 기술지수로 산출하였다.

본 연구의 과정은 다음과 같다. 데이터 수집을 통해 불러온 각 변수들을 MinMaxScaler를 이용하여 값을 0~1 사이의 값으로 변환하였고, 2017년 1월 부터 2020년 12월까지의 코스피 종가를 train data, 2021년 1월부터 2021년 12월까지의 코스피 종가를 test data로 분할하였다. 또한, LSTM 학습을 위한 windowed data를 생성하여 예측을 수행하였다.

본 연구는 Tensorflow 버전 2.6.0을 사용하여 진행되었다. 그 중 인공신경망의 주요한 특성은 모수(Hyper-parameter)에 따라 성과가 다르게 나온다는 것이다. 따라서 본 연구에서는 반복적인 실험과 시행착오를 통해 최적의 epoch, 학습률, 은닉층의 노드, 정지조건 등의 모수를 선정하였다. 인공신경망은 입력변수와 상응한 수의 입력 노드, 32개의 은닉 노드, 1개의 출력 노드로 구성하였으며 모든 계층에서 활성화 함수는 relu, 손실함수로 MSE 함수를 활용하였다. 학습 알고리즘의 학습률은 0.1로 설정하였으며 학습의 종료 조건으로 epoch를 1,500으로 설정하였다.

4. 연구 결과

본 연구에서는 가우시안 필터를 통해 노이즈를 제거함으로써 LSTM의 후행성 문제를 다루고자 하였으며, 감성분석의 세분화를 통해 뉴스 카테고리별 예측 성과의 증분적 효과를 다루고자 하였다. 또한, KOSPI 지수의 기술분석을 통해 파생변수를 생성하여 감성지수와의 시너지 효과를 다루고자 하였다. 실증분석을 위해 2017년 1월 1일부터 2021년 12월 31일 까지의 KOSPI 지수의 시가, 고가, 저가, 그리고 종가를 사용하였다. 2017년부터 2020년 까지의 데이터를 학습데이터로 이용하였으며, 2021년의 데이터로 검증을 진행하였다.

연구 결과, 모든 입력변수의 경우에 가우시안 필터를 적용한 LSTM 모델의 예측성능이 필터를 적용하지 않은 LSTM모델보다 예측 정확도와 후행성 문제가 개선된 성과를 보여주었다. 감성지수의 세부 카테고리별 지수는 기술지수가 입력변수로 추가된 경우에 모든 지표에서 성과가 개선된 모습을 보여주었다. KOSPI 종가의 경우 예측의 성능에 지대한 영향을 미치는 변수로 판별되었으나, 기술지수와 감성지수, 그리고 KOSPI 종가의 모든 변수를 입력한 경우 예측의 성과가 감퇴하는 모습을 보여주었다.

Table 1. The RMSE of unfiltered model

	RMSE	RMSLE
KOSPI 종가	0.03246119	0.0269979
기술지수	0.08216006	0.0806925
감성지수	0.0881142	0.08590172

Table 2. The RMSE of Gaussian filtered model

	RMSE	RMSLE
KOSPI 종가	0.02027084	0.0233516
기술지수	0.03121306	0.03435473
감성지수	0.01855792	0.0214348

Table 3. Incremental Effects of Technical Index and Sentiment Index

	RMSE	RMSLE
경제지수	0.47959508	0.48132107
경제·기술지수	0.01923769	0.0229274
국제지수	0.44337733	0.44446292
국제·기술지수	0.02139174	0.02382935
정치지수	0.50275399	0.50385581
정치·기술지수	0.01922877	0.02468902
감성·기술지수	0.02946056	0.03248406
KOSPI 종가 · 기술 · 감성지수	0.011646	0.015759

References

이낙영, 오경주.(2019).디노이징 필터와 LSTM을 활용한 KOSPI200 선물지수 예측. *한국데이터정보과학회지*, 30(3),645-654.

전재헌. (2018). 딥러닝을 활용한 개인 투자자 감정지표 기반의 주가예측에 관한 연구

Yan Xu and Guosheng Zhang, "Application of Kalman Filter in the prediction of Stock Price", International Symposium on Knowledge Acquisition and Modeling, 2015

최홍식.(2021).LSTM 주가 예측의 후행성 문제 해결에 관한 연구

이상훈, 최정, 김종우.(2016). 영역별 맞춤형 감성사전 구축을 통한 영화리뷰 감성분석. *지능정보연구*, 22(2),97-113.

The optimization of Stock Price Prediction using Gaussian Filter*

Myong-Jong Kim[1], *and Seong-Hun Kwon*[2]

Abstract

Research using machine learning and deep learning is being actively conducted, and research on predicting stock indexes such as KOSPI through machine learning is being actively conducted. However, the nonlinearity of random-walk data negatively affects the prediction of the model, resulting in a lagging problem in which the prediction follows the stock price of the previous day. Therefore, this study attempts to solve the trailing problem through a smoothing effect by utilizing a Gaussian filter, one of the denoising filters, for stock price prediction using Long Short-Term Memory (LSTM). The Gaussian filter has the advantage of effectively removing noise by weighting according to the distance at which the value is located. As a variable for prediction, the Sentiment Index was calculated through the extraction and emotional analysis of news keywords, the technical index was calculated through the technical analysis of the KOSPI index, and stock price prediction was carried out.

Keywords : Recurrent Neural Network, sentiment index, technical index, gaussian filter, KOSPI index.

*This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program(IITP-2022-2020-0-01797) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation).

[1]Professor, Department of Business Administration, Pusan National University, 2, Busandaehak-ro 63 beon-gil, Geumjeong-gu, Busan, Republic of Korea. Email: mjongkim@pusan.ac.kr

[2]Student, Department of Business Administration, Pusan National University, 2, Busandaehak-ro 63 beon-gil, Geumjeong-gu, Busan, Republic of Korea. Email: rnjstjdgns216@pusan.ac.kr

Proceedings of the Korean Data Analysis Society
January 26-27, 2023, pp. 107-112

기하평균 정확도 최적화 기반의 부스팅 모델의 범주 불균형 데이터 분류/예측 문제 적용*

김명종[1], 안재현[2]

요 약

본 연구에서는 기하평균 정확도 최적화 기반의 부스팅 알고리즘, GMOPTBoost을 다양한 범주 불균형 문제 해결에 적용하고 기존 부스팅 알고리즘과의 성과를 비교하고자 한다. GMOPTBoost는 경사하강법 적용을 위하여 기하평균 정확도에 대한 다변량 정규분포의 가정을 통해 convex 함수로 전환하였으며, 경사하강법을 적용하여 기하평균 정확도를 최적화하였다. 다양한 불균형 문제에 대해 적용한 결과 유의적인 성과개선 효과를 확인하였다.

Keywords : class imbalance, multivariate normal distribution, Gaussian gradient descent method.

1. 서론

앙상블 학습의 일종인 부스팅 학습은 약분류자의 결론을 결합하여 최종 결론을 도출하는 학습 기법으로 많은 분류/예측 문제에 적용되어 훌륭한 학습성과를 보이고 있다. 그러나, 부스팅 모델은 분류 모형의 학습오류를 최소화함으로써 예측 정확도(prediction accuracy)를 극대화하는 것을 학습 목표로 한다. 이러한 학습 방법들은 예측 범주가 균형을 이루고 있는 데이터를 가정하고 있다. 하지만 데이터가 특정 범주에 편중되어 있는 불균형 데이터에서는 큰 한계점을 보인다. 다수 범주의 정확도(특이도)는 높게 나타나지만, 소수 범주의 정확도(민감도)는 매우 낮게 나타난다. 이는 산술평균 정확도의 극대화만을 목표로 하는 학습 알고리즘에 기인한다. 따라서 이러한 학습 알고리즘의 한계를 극복하기 위해 학습 목표를 기하평균 정확도, AUC와 같은 불균형 데이터에 적합한 학습목표를 사용해야 한다. 기하평균 정확도를 목표 함수로 두고 극대화함에 있어서 난제는 기하평균 정확도 함수는 non-covex 함수로 미분이 거의 불가능하며 미분이 가능하더라도 경사가 0으로 측정되기 때문에 최적화 작업이 불가능하다는 점이다. 본 연구에서는 다변량 정규분포의 가정을 통해 기하평균 정확도를 미분 가능한 convex 함수로 치환하고 경사 하강법을 이용해 최적화 작업을 수행하는 GMOPTBoost를 제안하며, 이를 3가지 불균형 데이터 샘플에 적용하여 기존 알고리즘과의 성과를 비교하고자 한다.

*본 연구는 과학기술정보통신부 및 정보통신기획평가원의 대학ICT연구센터육성지원사업의 연구결과로 수행되었음(IITP-2022-2020-0-01797).
[1]부산대학교 경영학과 교수. E-mail: mjongkim@pusan.ac.kr
[2](부산대학교 경영학과 학사과정. E-mail: bonjour8084@pusan.ac.kr

2. GMOPTBoost 알고리즘

부스팅 학습의 기저 분류기의 예측 결과를 $U=\{(C_k(x_i),y_i): i=1,...,n\}$라 할 때 $C_k(x_i)\in R^K$은 최종 분류기의 입력 벡터가 된다. 편의상 $C_k(x_i)$를 x_i로 간략하게 표기하고 확률변수 (X, Y)의 결합확률을 $P_{X,Y}(x,y)$라 하면 (x_i, y_i)는 $P_{X,Y}(x,y)$의 i.i.d(independently identically distributed) 관측치가 된다. x_i를 다수 범주 확률변수(X^-)의 관측치 x^-와 소수 범주 확률변수(X^+)의 관측치 x^+로 구분하면, $GM(x,\omega)$은 특이도($SPE(x^-,\omega)$)와 민감도($SEN(x^+,\omega)$)의 함수로 식 (1)로 정의된다.

$$GM(x,\omega)=\sqrt{SPE(x^-,\omega)\times SEN(x^+,\omega)} \tag{1}$$

$GM(x,\omega)$의 기대값 $E(GM(x,\omega))$는 식(2)의 누적분포함수로 측정할 수 있지만 $P_{X,Y}(x,y)$의 분포가 정의되지 않았기 때문에 직접적인 $E(GM(x,\omega))$의 계산이 불가능하게 된다.

$$\begin{aligned} E(GM(x,\omega)) = (&\int SPE(x^-,\omega)P_{X,Y}(x,y)dP(x,y) \\ &\times \int SEN(x^+,\omega)P_{X,Y}(x,y)dP(x,y))^{1/2} \end{aligned} \tag{2}$$

n^-와 n^+는 각각 다수 범주와 소수 범주에 속한 사례의 수이며, 임계값을 T라 하면, 위험 최소화 원칙에 따라 측정된 기하평균오류 $GE(x,\omega)$의 근사치는 식(3)과 같은 이산확률의 CDF로 계산될 수 있다. 그러나 식 (3)에서 $SPE(x^-,\omega)$와 $SEN(x^+,\omega)$는 0-1 손실함수로 정의된 비평활함수(nonsmooth function)로서 미분계산이 불가능하고 경사하강법의 적용이 불가능하게 된다.

$$\begin{aligned} GE(x,\omega) &= 1-GM(x,\omega)=1-(\frac{1}{n^-}\sum_{i=1}^{n^-}SPE(x_i^-,\omega) \\ &\times \frac{1}{n^+}\sum_{i=1}^{n^+}SEN(x_i^+,\omega))^{1/2} \end{aligned} \tag{3}$$

$$SPE(x^-,\omega)=\begin{cases}1 \text{ if } \omega^T x^- \le T\\ 0 \ otherwise\end{cases},\quad SEN(x^+,\omega)=\begin{cases}1 \text{ if } \omega^T x^{+>} T\\ 0 \ otherwise\end{cases}$$

본 연구에서는 $P_{X,Y}(x,y)$가 정규분포를 따른다는 가정을 도입하였다. 정규분포의 가정하에서 다수 범주의 확률변수 X^-와 소수 범주와 확률 변수 X^+와 기저 분류기의 선형 결합인 강분류기의 최종 예측값(W^TX)도 다음과 같이 정규 분포를 따르게 된다.

$$X^-\sim N(\mu^-,\Sigma^-)\rightarrow W^TX^-\sim N(\omega^T\mu^-,\omega^T\Sigma^-\omega)\sim N(\mu_Z^-,\sigma_Z^{2-})$$
$$X^+\sim N(\mu^+,\Sigma^+)\rightarrow W^TX^+\sim N(\omega^T\mu^+,\omega^T\Sigma^+\omega)\sim N(\mu_Z^+,\sigma_Z^{2+})$$

여기에서, μ^-와 μ^+는 다수 및 소수 범주 입력벡터의 평균벡터이며, Σ^-와 Σ^+는 다수 및 소수 범주 입력벡터의 공분산 행렬이다.

$\nabla GM(x,\omega)$의 계산을 위하여 식(2)에 자연로그 함수를 취한 후 이에 대한 1차 미분함수는 식(4)와 같이 계산된다.

$$GM(x,\omega) = \frac{1}{2}\left(LN(SPE(x^-,\omega)) + LN(SEN(x^+,\omega))\right) \tag{4}$$
$$\nabla GM(x,\omega) = \frac{1}{2}\left(\frac{\nabla SPE(x^-,\omega)}{SPE(x^-,\omega)} + \frac{\nabla SEN(x^+,\omega)}{SEN(x^+,\omega)}\right)$$

$SPE(x^-,\omega) = W^T X^- = T$ 이고, $SEN(x^+,\omega) = W^T X^+ = T$이며 T를 임계점이라 할 때, $SPE(x^-,\omega)$와 $SEN(x^+,\omega)$ 의 CDF는 각각 식(5)와 식(6)과 같이 정의된다.

$$\Phi(SPE(x^-,\omega) = \int_{\infty}^{SPE(x^-,\omega)} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{T-\omega^T\mu^-}{\sqrt{\omega^T\Sigma^-\omega}}\right)^2\right) d\omega \tag{5}$$

$$\Phi(SEN(x^+,\omega) = \int_{\infty}^{SEN(x^+,\omega)} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{T-\omega^T\mu^+}{\sqrt{\omega^T\Sigma^+\omega}}\right)^2\right) d\omega \tag{6}$$

$\Phi(SPE(x^-,\omega))$의 1차 미분 함수는 식(7)의 $\nabla SPE(x^-,\omega)$로 정의되며, $\Phi(SPE(x^+,\omega))$의 1차 미분 함수는 식(8)의 $\nabla SEN(x^+,\omega)$로 정의된다.

$$\begin{aligned}\nabla SPE(x^-,\omega) &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{T-\omega^T\mu^-}{\sqrt{\omega^T\Sigma^-\omega}}\right)^2\right) \times \left(\frac{\sqrt{\omega^T\Sigma^-\omega}\,(\mu^- - T) - \left(\frac{\omega^T\mu^- - T}{\sqrt{\omega^T\Sigma^-\omega}}\right)\Sigma^-\omega}{\omega^T\Sigma^-\omega}\right) \\ &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{T-\mu_Z^-}{\sigma_Z^-}\right)^2\right)\left(\frac{\sigma_Z^-(\mu^- - T) - \left(\frac{\mu_Z^- - T}{\sigma_Z^-}\right)\Sigma^-\omega}{\sigma_Z^{2-}}\right)\end{aligned} \tag{7}$$

$$\begin{aligned}\nabla SEN(x^+,\omega) &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{T-\omega^T\mu^+}{\sqrt{\omega^T\Sigma^+\omega}}\right)^2\right) \times \left(\frac{\sqrt{\omega^T\Sigma^+\omega}\,(\mu^+ - T) - \left(\frac{\omega^T\mu^+ - T}{\sqrt{\omega^T\Sigma^+\omega}}\right)\Sigma^+\omega}{\omega^T\Sigma^+\omega}\right) \\ &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{T-\mu_Z^+}{\sigma_Z^+}\right)^2\right)\left(\frac{\sigma_Z^+(\mu^+ - T) - \left(\frac{\mu_Z^+ - T}{\sigma_Z^+}\right)\Sigma^+\omega}{\sigma_Z^{2+}}\right)\end{aligned} \tag{8}$$

$\nabla SPE(x^-,\omega)$와 $\nabla SEN(x^+,\omega)$을 식(4)에 대입하면, $\nabla GM(x,\omega)$는 식(9)로 대체된다.

$$\nabla GM(x,\omega) = \frac{1}{2}\left(\frac{1}{SPE(\omega)}\right)\frac{1}{2\pi}\exp\left(-\frac{1}{2}\left(\frac{T-\mu_Z^-}{\sigma_Z^-}\right)^2\right)\times\left(\frac{\sigma_Z^-(\mu^- - T)-\left(\frac{\mu_Z^- - T}{\sigma_Z^-}\right)\Sigma^-\omega}{\sigma_Z^{2-}}\right) \quad (9)$$
$$-\frac{1}{2}\left(\frac{1}{SEN(\omega)}\right)\frac{1}{2\pi}\exp\left(-\frac{1}{2}\left(\frac{T-\mu_Z^+}{\sigma_Z^+}\right)^2\right)\times\left(\frac{\sigma_Z^+(\mu^+ - T)-\left(\frac{\mu_Z^+ - T}{\sigma_Z^+}\right)\Sigma^+\omega}{\sigma_Z^{2+}}\right)$$

3. 연구결과

연구에 사용하기 위해 데이터는 비즈니스 분야에서 예측 문제로 사용되는 대표적인 데이터 3가지를 수집하였다. 첫째, 기업 부도 데이터는 2015년 ~ 2018년의 국내 실제 제조업 기업 데이터로 총자산경상이익률, EBITDA/이자비용, 자기자본비율, 이익잉여금/총자산, 현금비율, 재고자산회전율, 총자산 총 7개 재무비율을 사용하여 부도 기업을 예측했다.(Kim et al, 2022). 둘째, 카드 대금 지급 불이행 데이터는 2005년 대만에서 일어난 실제 카드 거래 데이터로 신용수준, 성별, 나이, 혼인여부, 교육수준, 과거 카드결제 내역(정상지불, 연체 등), 과거 카드 사용금액 등 총 23개 변수를 사용하여 카드 대금 지급 불이행을 탐지했다.(Yeh, I-Cheng; Lien, Che-Hui, 2009). 셋째, 카드 사기 데이터는 2013년 9월 유럽에서 발생한 실제 카드 거래 데이터로 거래 정보에 대한 28개의 변수를 주성분분석(Principal Component Analysis)를 통해 차원축소(Dimensionality Reduction)을 진행한 9개 변수를 사용하여 카드 사기를 탐지했다.

Table 1. 실험 데이터 세트

데이터명	데이터수	소수범주	다수범주	불균형율
기업 부도	10,500	500	10,000	1:20
카드 대금 지급 불이행	30,000	6,636	23,364	1:3.52
카드 사기	284,806	492	284,314	1:577.8

성과비교 알고리즘으로 AdaBoost, GBM, XGB로 총 3개 알고리즘을 활용하였다. 모든 모델의 기저분류자는 동일하게 Max_Depth=2인 의사결정나무를 사용했으며, 교차 검증 타당성을 위해 데이터별로 10번의 교차 검증(Cross Validation)을 3번씩 진행하여 총 30회 교차 타당성 분석의 평균값을 통해 모델 성과를 비교했다.

주요 분석 결과는 다음과 같다. GMOPTBoost는 다른 부스팅 알고리즘과 비교하여 소수 범주 정확도가 월등히 개선되는 것으로 분석되었다. 세가지 데이터 모두에서 GMOPTBoost은 비교대상 알고리즘(AdaBoost, GBM, XGBoost)과 비교하여 단순평균정확도(ACC)는 낮았지만, 범주 불균형 성과지표인 기하평균 정확도(GM)와 AUC는 높은 성과를 보였다. 이러한 결과는 기존에 널리 사용되고 있는 알고리즘에 비해 본 연구에서 제안한 GMOPTBoost가 데이터의 종류와 무관하게 뛰어난 분류 정확성을 제공하고 있으며, 불균형 데이터에서 소수 범주의 예측 정확도를 개선하는 효과를 보여주는 강건한 알고리즘이라는 것을 의미한다.

Table 2. 실험 결과 지표

데이터	평가지표	Ada Boost	GBM	XGB	GMBoost
기업 부도 예측	SEN	0.11	0.05	0.05	0.73
	SPE	0.99	0.99	0.99	0.79
	ACC	0.95	0.95	0.95	0.79
	AUC	0.55	0.53	0.53	0.76
	GM	0.33	0.21	0.22	0.76
카드 대금 지급 불이행 탐지	SEN	0.36	0.32	0.35	0.70
	SPE	0.94	0.96	0.95	0.70
	ACC	0.81	0.82	0.82	0.70
	AUC	0.59	0.56	0.58	0.70
	GM	0.66	0.64	0.65	0.70
카드 사기 탐지	SEN	0.75	0.68	0.75	0.87
	SPE	0.99	0.99	0.99	0.95
	ACC	0.99	0.99	0.99	0.95
	AUC	0.87	0.82	0.87	0.91
	GM	0.88	0.84	0.87	0.91

References

Freund, Y. and R. E. Schapire, “A Decision Theoretic Generalization of Online learning and An Application to Boosting”, *Journal of Computer and System Science*, Vol. 55, No. 1, 1997, pp. 119-139.

Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, 29(5), 1189–1232.

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

Myoung-Jong Kim, Woo Seob Yun. (2022). The Optimization of Ensembles for Bankruptcy Prediction. *Information Systems Review*, 24(1), 39-57.

I-Cheng Yeh, Che-hui Lien, The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients, *Expert Systems with Applications*, Volume 36, Issue 2, Part 1, 2009, Pages 2473-2480, ISSN 0957-4174

Application of boosting model based on geometric mean accuracy optimization in categorical imbalance data classification/prediction problem

Myoung-Jong Kim, and Jae-Hyun Ahn

Abstract

GMOPTBoost, a boosting method based on geometric mean accuracy optimization, is used to solve several category imbalance problems in this study and its performance is compared to that of other boosting techniques. The geometric mean accuracy was maximized by converting GMOPTBost to a convex function under the assumption of a multivariate normal distribution, and the geometric mean accuracy was optimized by employing the gradient descent method. The experimental results with 3 business applications show significant performance improvement.

Keywords : class imbalance, multivariate normal distribution, Gaussian gradient descent method.

Proceedings of the Korean Data Analysis Society
January 26-27, 2023, pp. 113-116

서울서베이 표본추출틀 구축을 위한 방법론 연구

강현철[1], 조수정[2], 이동수[3], 이은혜[4], 최성용[5]

요 약

서울시에서는 빠른 사회변화의 흐름을 수용하는 동시에 지역별 특성을 반영하여 시정운영에 적절히 활용하기 위해, 행정자료 및 가구조사를 통하여 도시발전 지향을 지속적으로 모니터링 할 수 있는 정책지표의 개발 및 구축을 목적으로 하는 '서울서베이' 사업을 2003년부터 매년 시행하고 있다. 서울서베이의 초기에는 통계청에서 제공하는 인구주택총조사 기반의 조사구 추출틀을 이용하였으나, 5년마다 갱신되는 인구주택총조사 기반의 추출틀을 사용하는 경우 시의성에 대한 문제가 제기되었고, 주택유형 등 조사결과에 영향을 주는 세부 정보들을 사용할 수 없다는 점이 지적되었다. 따라서 재개발 및 재건축 등으로 인한 가구(주택)의 신축과 소멸을 반영하기 위하여, 2005년부터 사용가능한 최신의 서울시 '주민등록 DB'의 세대주 주소를 표본추출틀로 사용하여 왔고, '과세대장 DB' 또는 '건축물대장 DB'를 연계하여 각 세대주 주소에 해당하는 주택유형을 파악하고 있다. 본 연구에서는 2022년 기준 서울서베이 가구조사의 표본설계에 대한 내용을 요약하여 설명하고, 최신의 '주민등록 DB'와 '건축물대장 DB'를 연계하여 서울서베이 가구조사를 위한 표본추출틀을 구축하는 방안에 대하여 논의하고자 한다.

주요용어 : 서울서베이, 서울도시정책지표조사, 표본추출틀, 표본설계, 표본배분.

1. 서론

세계적으로 사회구조의 변화가 빠르게 나타나고 지역단위의 경쟁력이 주요 이슈로 등장하고 있는 시기를 맞아, 서울시에서는 빠른 사회변화의 흐름을 수용하는 동시에 지역별 특성을 반영하여 시정운영에 적절히 활용하기 위해 적절한 통계작성의 필요성을 인식하게 되었다. 따라서 서울시는 행정자료 및 가구조사를 통하여 도시발전 지향을 지속적으로 모니터링 할 수 있는 정책지표의 개발 및 구축을 목적으로 하는 '서울서베이' 사업을 2003년부터 매년 시행하고 있다.

서울도시정책지표조사로 명명되며 서울서베이의 가장 큰 부분을 차지하는 가구조사는 서울시 거주 가구(또는 세대)와 15세 이상의 모든 가구원(또는 세대원)을 목표 모집단으로 정의하고 있으나, 인구와 가구 유동성이 높은 서울시의 특성상 목표모집단과 정확히 부합하는 적절한 표본추출틀을 구축하는 것이 용이하지 않다.

[1]31499 충남 아산시 배방읍 호서로79번길 20, 호서대학교 빅데이터AI학부 교수. E-mail: hychkang@hoseo.edu
[2]04515 서울특별시 중구 덕수궁길15 서소문청사 1동, 서울특별시 디지털정책관 빅데이터담당관, 팀장.
[3]04515 서울특별시 중구 덕수궁길15 서소문청사 1동, 서울특별시 디지털정책관 빅데이터담당관, 전문경력관.
[4]04515 서울특별시 중구 덕수궁길15 서소문청사 1동, 서울특별시 디지털정책관 빅데이터담당관, 주무관.
[5]04515 서울특별시 중구 덕수궁길15 서소문청사 1동, 서울특별시 디지털정책관 빅데이터담당관, 주무관.

서울서베이의 초기에는 통계청에서 제공하는 인구주택총조사 기반의 조사구 추출틀을 이용하였으나, 5년마다 갱신되는 인구주택총조사 기반의 추출틀을 사용하는 경우 시의성에 대한 문제가 제기되었고, 주택유형 등 조사결과에 영향을 주는 세부 정보들을 사용할 수 없다는 점이 지적되었다.

따라서 재개발 및 재건축 등으로 인한 가구(주택)의 신축과 소멸을 반영하기 위하여, 2005년부터 사용가능한 최신의 서울시 '주민등록 DB'의 세대주 주소를 표본추출틀로 사용하여 왔고, '과세대장 DB' 또는 '건축물대장 DB'를 연계하여 각 세대주 주소에 해당하는 주택유형을 파악하고 있다.

본 연구에서는 최신의 '주민등록 DB'와 '건축물대장 DB'를 연계하여 2022년 기준의 서울서베이 가구조사를 위한 표본추출틀을 구축하는 방안에 대하여 논의하고자 한다.

2. 표본설계 요약

2021년 통계청 인구총조사 기준으로 서울시의 구별, 주택유형별 모집단 크기를 Table 1에 제시하였다. 모집단 전체 가구수는 4,046,799가구이며, 전체 가구원수(인구수)는 9,125,047명이고 그 중 만15세 이상 가구원수는 8,210,920명으로 파악되었다.

Table 1. 서울시 구별 모집단 크기

	가구					가구원	
구분	전체	단독주택 다가구주택	아파트	다세대주택	연립주택 기타	전체	15세 이상
전체	4,046,799	1,018,255	1,758,041	764,951	505,552	9,125,047	8,210,920
종로구	63,984	23,445	14,940	11,474	14,125	137,913	126,788
중구	56,116	13,521	24,781	6,637	11,177	114,942	106,157
용산구	96,989	32,493	35,260	17,137	12,099	205,288	187,139
성동구	122,756	34,791	66,360	9,399	12,206	276,273	249,292
광진구	153,962	65,544	31,981	37,182	19,255	328,395	299,876
동대문구	153,246	50,545	64,924	15,839	21,938	329,092	299,689
중랑구	167,260	64,272	55,402	33,256	14,330	373,443	340,742
성북구	181,827	56,517	79,321	28,761	17,228	422,428	379,095
강북구	130,329	46,859	33,958	38,559	10,953	287,693	264,449
도봉구	128,439	26,432	63,534	27,806	10,667	306,168	278,468
노원구	204,282	21,212	159,378	14,503	9,189	498,257	445,437
은평구	193,626	43,058	63,059	67,506	20,003	451,661	407,684
서대문구	134,544	38,523	53,951	26,358	15,712	299,912	271,594
마포구	164,080	36,319	71,087	33,314	23,360	352,832	317,139
양천구	168,355	26,573	89,869	39,080	12,833	428,082	375,573
강서구	249,984	34,172	112,459	62,546	40,807	551,470	495,196
구로구	166,197	31,284	79,992	27,214	27,707	383,334	344,114
금천구	107,430	30,426	32,179	23,423	21,402	222,191	204,534
영등포구	168,518	40,328	75,605	14,879	37,706	359,642	324,755
동작구	170,333	57,799	64,802	32,593	15,139	373,671	339,275
관악구	252,619	107,766	55,018	42,251	47,584	473,263	442,081
서초구	155,936	23,607	89,447	25,010	17,872	385,150	334,072
강남구	211,168	34,949	122,480	27,546	26,193	494,171	434,495
송파구	259,378	35,735	127,454	66,750	29,439	624,977	551,205
강동구	185,441	42,085	90,800	35,928	16,628	444,799	392,067

Table 2. 구별, 주택유형별 표본배분

구분	전체	단독주택 다가구주택	아파트	다세대주택	연립주택 기타
전체	20,000	5,110	8,600	3,755	2,535
종로구	510	185	120	90	115
중구	480	115	210	60	95
용산구	630	210	230	110	80
성동구	705	200	380	55	70
광진구	790	335	165	190	100
동대문구	790	260	335	80	115
중랑구	825	315	275	165	70
성북구	860	265	375	135	85
강북구	730	260	190	215	65
도봉구	725	150	360	155	60
노원구	910	95	710	65	40
은평구	885	195	290	310	90
서대문구	740	210	295	145	90
마포구	815	180	355	165	115
양천구	825	130	440	190	65
강서구	1,010	140	450	255	165
구로구	820	155	395	135	135
금천구	660	185	200	145	130
영등포구	830	200	370	75	185
동작구	830	280	315	160	75
관악구	1,015	435	220	170	190
서초구	795	120	455	130	90
강남구	925	155	535	120	115
송파구	1,025	140	505	265	115
강동구	870	195	425	170	80

2022년 서울서베이 가구조사의 전체 목표 표본크기는 조사의 비용 및 시간을 고려하여 20,000가구로 결정되었다. 또한 조사내용의 특성과 모집단의 대표성을 높이기 위해 지역구분과 주택유형을 층화변수로 고려하였다 - 구: 25개, 주택유형: 4개(단독주택, 아파트, 다세대주택, 연립주택/기타). 구별 표본배분에서는 다음과 같은 사항들을 고려하였다. (1) 구별 통계의 생산 및 비교가 가능하도록 구별로 최소 400표본이 할당될 수 있도록 하였다. (2) 구별 가구수의 배분을 위해 주거환경, 교육환경, 월소득의 표준편차를 고려한 네이만배분과 단순 비례배분 및 제곱근 비례배분을 검토하였다. (3) 제곱근 비례배분이 구별로 최소 400표본을 확보할 수 있고, 구별 표본크기의 차이가 크지 않으므로 제곱근 비례배분에 의해 구별 가구수를 배분하였다. 각 구내에서 주택유형별 표본배분은 단순비례배분을 사용하였다(Table 2 참조).

3. 표본추출틀 구축

이 절에서는 '주민등록 DB'와 '건축물대장 DB'를 연계하여 표본추출틀을 구축하는 방안을 설명한다. 먼저 2022년 8월 기준으로 서울시 '주민등록 DB'에 존재하는 세대주 주소는 4,384,199개로 파악되었으며, 각 세대주 주소는 '도로명+건물번호+동/호'와 같은 도로명 주소로 구분되어 있다.

Table 3. 구별, 주택유형별 세대주 주소의 개수

구분	전체	단독주택 다가구주택	아파트	다세대주택	연립주택 기타
전체	4,384,199	1,026,785	1,873,113	807,548	676,753
종로구	71,223	22,689	16,568	11,793	20,173
중구	61,726	12,477	25,335	6,702	17,212
용산구	107,040	31,737	40,971	17,950	16,382
성동구	131,670	32,367	69,115	10,260	19,928
광진구	167,184	67,746	34,688	39,979	24,771
동대문구	167,206	49,106	68,156	16,282	33,662
중랑구	185,518	68,971	61,021	35,704	19,822
성북구	194,803	53,589	83,608	30,521	27,085
강북구	142,531	49,837	35,728	40,606	16,360
도봉구	137,142	26,929	67,035	29,436	13,742
노원구	215,774	19,518	166,752	15,102	14,402
은평구	212,576	45,728	68,798	73,358	24,692
서대문구	142,447	35,274	57,926	27,875	21,372
마포구	177,300	35,650	74,887	36,959	29,804
양천구	179,284	28,217	94,917	40,900	15,250
강서구	271,035	35,859	120,908	67,430	46,838
구로구	179,980	32,145	88,022	28,638	31,175
금천구	118,002	33,259	33,017	23,568	28,158
영등포구	185,454	41,251	81,175	14,217	48,811
동작구	184,532	60,465	68,563	34,662	20,842
관악구	278,181	104,498	56,121	39,638	77,924
서초구	164,720	23,253	92,870	26,034	22,563
강남구	227,851	36,094	130,013	29,920	31,824
송파구	281,374	35,192	140,610	72,208	33,364
강동구	199,646	44,934	96,309	37,806	20,597

서울서베이를 위한 표본설계에서는 주택유형을 ‘단독주택/다가구주택, 아파트, 다세대주택, 연립주택/기타’로 구분하고 있는데, 이를 위해 ‘건축물대장 DB’의 건물용도 정보를 활용하여 각 세대주 주소의 주택유형을 파악하였다.

‘건축물대장 DB’의 건물용도 정보를 활용함에 있어서 몇 가지 문제점이 존재한다. 첫째, 동일한 ‘도로명+건물번호’ 내에서 건물용도가 다르거나 없는 경우가 있다. 일반적으로 동일한 ‘도로명+건물번호’ 내에 있는 주소(동/호)들은 동일한 건물용도를 가져야 한다. 둘째, 신축건물의 경우 과거의 건물용도로 구분되어 있는 경우가 있다. 예를 들어, 신축 아파트 단지 내의 일부 주소들이 근린생활시설 등으로 구분되어 있는 경우가 있다. 셋째, ‘주민등록 DB’에는 존재하나 ‘건축물대장 DB’에는 주소가 없는 경우가 있다.

이와 같은 문제점을 해결하기 위하여 다음과 같은 절차를 수행하였다. (1) ‘도로명+건물번호’ 단위로 가장 많은 건수를 가지는 건물용도를 기준으로 대표 주택유형(단독주택, 아파트, 다세대주택, 연립주택/기타)을 설정하여 데이터셋을 생성하고, ‘세대주 주소’ 데이터셋과 ‘n 대 1’ 매칭을 수행하였다. (2) ‘연립주택/기타’로 매칭되어 있는 신축 아파트의 경우 포털사이트에서 검색한 후 주택유형을 변경하였다. (3) 주택유형이 매칭되지 않은 세대주 주소(동/호)의 경우 근접지역의 주택유형 또는 ‘연립주택/기타’로 구분하였다.

Proceedings of the Korean Data Analysis Society
January 26-27, 2023, pp. 117-121

결정계수를 이용한 선형혼합모형의 선택*

이용희[1]

요 약

결정계수는 선형회귀모형에서 모형에 포함된 효과들이 자료의 총변동을 설명하는 상대적 비율을 나타내는 측도로서 모형을 선택하는데 가장 널리 사용되는 유용한 모형 선택의 측도이다. 선형혼합모형 또는 계층모형은 고정효과와 임의효과를 동시에 고려하는 모형으로서 독립이 아닌 복잡한 자료의 구조와 다양한 변동을 반영할 수 있는 모형이다. 선형혼합모형에서는 계층적인 자료의 구조에 의한 다양한 예측과 임의효과의 분산성분으로 정의할 수 있는 여러 가지 변동들이 존재하기 때문에 선형모형에서 정의된 결정계수를 확장하는 방법이 유일하지 않다. 본 논문에서는 선형혼합모형에서 모형선택을 위하여 제안된 다양한 결정계수의 개념과 정의를 고찰하였다.

주요용어 : 선형혼합모형, 결정계수, 모형의 선택.

1. 서론

선형혼합모형(linear mixed effects model)은 독립이 아닌 관측값으로 구성된 자료를 분석하는데 사용되는 모형이다. 선형혼합모형은 일반 선형회귀모형에서 사용되는 고정효과(설명변수)뿐만 아니라 독립이 아닌 관측값들의 상관성 또는 계층적 구조를 반영하기 위한 임의효과를 사용한다. 선형혼합모형을 적용할 수 있는 자료의 종류와 구조는 다양하지만 가장 흔하게 나타나는 자료 구조는 독립된 군집으로 이루어진 자료이다. 독립된 군집 자료는 전체 자료가 여러 개의 서로 독립적인 군집(개체, 그룹 등)들로 구성되어 있으며 군집 내에 속한 관측값들은 독립이 아닌 구조이다. 또한 독립 군집 자료는 의학연구에서 널리 사용되는 반복측정 자료를 포함하며 여러 개의 내포되는 층으로 구성된 계층적인 자료 구조로 확장할 수 있다.

독립 군집에 대한 자료를 설명하는 선형혼합모형에서 모형을 선택하는 경우 일반 선형회귀모형에서 사용되는 방법의 개념과 정의를 확장한 방법들을 사용한다. AIC와 같은 정보기준을 이용한 모형선택기준을 선형혼합모형에 확장한 조건부 AIC 또는 BIC 방법이 대표적인 사례이다 (Vaida et al., 2005; Muller et al, 2013; Lee, 2014, 2015). 또한 일반 선형회귀모형에서 자주 사용되는 결정계수(coefficient of determination; R^2)도 선형혼합모형에 적용할 수 있도록 확장된 개념과 방법들이 제안되었다.

선형혼합모형에서 사용되는 정보기준에 근거한 모형선택 측도들은 기존의 일반 선형회귀모형에서 사용된 정의를 쉽게 확장하여 유사한 방법으로 사용할 수 있어 현재까지 많은 연구자들이 사용하고 있다. 하지만 선형혼합모형에 확장하여 적용된 결정계수는 일반 선형회귀모형의 결정계수의 개념을 확장하는 방향이 매우 다양하고 통계적 특성도 상이하기 때문에 사용에 어려움이 있었다. 최근 선형

*이 논문은 2020년도 서울시립대학교 연구년교수 연구비에 의하여 연구되었음.
[1]02504 서울특별시 동대문구 서울시립대로 163, 서울시립대학교 통계학과. E-mail: @naver.com

혼합모형에 확장하여 적용된 결정계수를 사용하는 연구의 빈도가 늘어나고 계산을 할 수 있는 패키지들이 발간되면서 연구자들의 관심이 높아지고 있다. 본 논문은 이러한 추세에 따라서 선형혼합모형에 적용할 수 있는 결정계수의 다양한 개념과 정의, 사용법을 고찰하려고 한다. 2절에서는 선형혼합모형에서 변동을 설명하는 확장된 결정계수의 개념과 정의를 고찰하고 3절에서는 주변부와 조건부 결정계수의 개면과 정의를 알아본다. 결론에서는 계산을 위한 여러 가지 통계 패키지를 소개하였다.

2. 변동을 설명하는 결정계수

일반 선형회귀모형은 독립인 관측값들 y_i와 설명변수 벡터 $x_{i\cdot}=(1,x_{i1},x_{i2},\cdots,x_{ip})^t$ 의 관계를 다음과 같은 선형식으로 나타낸다. 아래 회귀식에서 오차 e_i는 서로 독립이고 평균이 0 이고 분산이 σ^2 을 가진다.

$$y_i = x_i^t\beta + e_i, \quad i=1,2,\cdots,n$$

일반 선형회귀모형에서 사용하는 결정계수는 관측값들이 가지는 총변동 중에서 주어진 회귀모형의 설명변수들이 설명할 수 있는 비율을 제곱합을 이용하여 나타낸 것이다.

$$R^2 = 1 - \frac{SSE}{SST}$$

일반 선형회귀모형에 대한 결정계수 R^2의 정의는 관측값들이 가지는 총변동을 나타내는 총제곱합 $SST=\sum(y_i-\bar{y})^2$에서 주어진 모형이 설명하지 못하는 변동을 나타내는 잔차제곱합 $SSE=\sum(y_i-\hat{y_i})^2$의 비율을 제외한 상대적 비율로 나타난다. 즉, 결정계수 R^2은 관측값이 가진 고유의 변동 중에서 주어진 모형의 독립변수들에 의하여 설명되는 변동의 비율이다.

이제 결정계수의 개념을 선형혼합모형에 확장하기 위하여 하나의 임의효과를 가지는 다음과 같은 독립군집 모형을 고려하자. 반응변수 y_{ij}는 i번째 군집에 속하는 j번째 개체의 관측값이다. i번째 군집에 대한 임의효과 b_i 는 독립적으로 $N(0,\tau^2)$ 을 따르며 오차항 e_{ij}는 $N(0,\sigma^2)$를 따른다. 또한 $x_{ij}=(1,x_{ij1},x_{ij2},\ldots,x_{ijp})^t$ 이고 $\beta=(\beta_0,\beta_1,\ldots,\beta_p)^t$ 이다.

$$y_{ij} = (\beta_0 + \sum_{k=1}^{p} x_{ijk}\beta_k) + b_i + e_{ij} = x_{ij}^t\beta + b_i + e_{ij}, \quad i=1,2,\cdots,I,\ j=1,2,\cdots,J \tag{1}$$

Snijder and Bosker(1994)은 선형혼합모형에 대하여 결정계수를 “주어진 모형으로 예측을 하는 경우 예측오차가 감소하는 비율”로 정의하는 방향을 제시하였다. 이렇게 예측에 대한 오차를 고려하는 경우 일반 선형회귀모형과 다르게 선형혼합모형에서는 예측의 대상이 단순하지 않다. 일반적으로 선형혼합모형이 적용되는 자료는 계층적인 구조를 가진다. 관측값이 수집되는 가장 낮은 1차 계층(예를 들어 학교 안의 학생들)에서 예측과 한 단계 상위인 2차 계층(그룹 또는 군집으로 예를 들어 학교)에 대한 예측을 따로 고려해야 한다. Snijder and Bosker(1994)는 위에서 고려한 독립 군집 모형에서 1차 계층과 2차 계층의 예측을 분리하여 다음과 같은 두 가지 결정계수의 개념과 정의를 고려하고 통계적 성질을 제시하였다.

$$R_1^2 = 1 - \frac{Var(y_{ij} - x_{ij}^t \beta)}{Var(y_{ij})}, \qquad R_2^2 = 1 - \frac{Var(\overline{y_i} - \overline{x}_i^t \beta)}{Var(\overline{y_i})} \tag{2}$$

먼저 위의 식에서 정의된 결정계수 R_1^2은 최종적으로 관측값이 얻어지는 1차 계층(예를 들어 학생의 성적)에 대한 예측을 고려한 모형선택의 측도이다. R_1^2은 "설명변수가 없는 최소모형"에서의 예측의 오차($Var(y_{ij})$)와 "설명변수가 포함된 모형"에 의한 예측 오차($Var(y_{ij} - x_i^t \beta)$)의 상대적인 비율로 나타내는 정의이다. 또한, 결정계수 R_2^2는 군집이나 그룹 단위인 2차 계층(예를 들어 학교)에 대한 예측을 고려하는 경우 예측의 오차가 감소하는 비율이다. 2차 단위를 예측하는 경우 군집의 평균값을 사용하므로 이 경우는 기본적인 예측의 오차가 $Var(\overline{y_i})$ 이다. 만약 선형혼합모형이 적용되는 자료의 계층이 k개이면 최소한 k개의 결정 계수를 정의할 수 있다. 이렇게 각 층에 대한 예측을 평가하는 결정계수가 따로 존재하므로 설명변수가 추가되는 경우 예측력에 대한 기여를 각 층마다 따로 평가해야 한다.

결정계수 R_1^2과 R_2^2 은 다음과 같은 특성을 가진다. (1) 위의 식에서 정의된 두 가지 결정계수 R_1^2과 R_2^2는 표본자료로 계산할 수 있는 통계량이 아니라 추정해야할 모수(parameters)이다. 즉 R_1^2과 R_2^2의 추정량이 필요하며 분산성분 τ^2과 σ^2의 추정법에 따라서 결정계수의 추정값이 다를 수 있다. 설명변수가 없는 최소모형과 설명변수가 포함된 후보모형에서의 분산성분 추정량을 각각 $\hat{\tau}_0^2$, $\hat{\sigma}_0^2$ 와 $\hat{\tau}^2$, $\hat{\sigma}^2$ 라고 하면 결정계수 R_1^2과 R_2^2 의 추정량은 다음과 같이 주어질 수 있다. 아래 식에서 J는 군집내의 개체 수를 의미한다.

$$\widehat{R_1^2} = 1 - \frac{\hat{\sigma}^2 + \hat{\tau}^2}{\hat{\sigma}_0^2 + \hat{\tau}_0^2}, \qquad \widehat{R_2^2} = 1 - \frac{\hat{\sigma}^2 + \hat{\tau}^2/J}{\hat{\sigma}_0^2 + \hat{\tau}_0^2/J}$$

(2) 모수로서의 결정계수 R_1^2과 R_2^2은 0보다 크지만 표본에 의한 추정량은 0보다 작을 수 있다. (3) 설명변수의 개수가 증가해도 결정계수의 추정값은 감소할 수 있다. (4) 결정계수 R_1^2과 R_2^2을 추정하기 위해서는 설명변수가 없는 최소모형과 설명변수가 포함된 후보모형에 대하여 별도로 모수를 추정해야 한다.

Xu(2003)은 Snijder and Bosker(1994)가 제안한 식 (2) 의 결정계수를 추정할 때 분선성분 추정, 제곱합 추정, KL-정보기준(Kullback-Leibler information criteria)을 이용한 추정 등을 다양하게 고려하여 그 효율을 비교하였다. Edward wt al.(2008)은 일반회귀모형에서는 F-검정에 사용되는 통계량과 결정계수의 관계를 선형혼합모형에 확장하여 주어진 모형의 설명변수에 대한 새로운 결정계수를 제안하였다.

3. 조건부 결정계수

Vonesh et al.(1996)은 비선형혼합모형에서 모형의 적합성(goodness of fit) 측도를 논의하면서 유사상관계수(concordance correlation coefficient)의 개념을 적용하여 조건부 결정계수(conditional R^2)을 제안하였다. 일반적으로 선형혼합모형에서 조건부 결정계수는 고정효과(또는 설명변수)와 임의효과에 의하여 설명되는 변동의 상대적인 비율을 나타내며 주변부 결정계수(marginal R^2)은 고정효과만으로

설명되는 변동의 비율을 나타낸다. 조건부 결정계수는 예측에 관심이 임의효과가 주어진 경우를 가정하고(예를 들어 학생이 특정 학교에 속한다면) 반응변수를 예측하는데 효율을 나타내는 측도이다. 조건부 결정계수는 Vaida(2005)가 제시한 조건부 AIC의 개념과 유사하다. 반대로 주변부 결정계수는 전체 집단(population)에서 반응변수를 예측하는 경우 모형의 효율을 나타내는 측도이다.

주변부 결정계수 R_m^2 과 조건부 결정계수 R_c^2 은 다음과 같이 정의할 수 있다.

$$R_m^2 = 1 - \frac{Var\,(y_{ij} \mid x_{ij}^t \beta)}{Var\,(y_{ij})}, \qquad R_c^2 = 1 - \frac{Var\,(y_{ij} \mid x_{ij}^t \beta, b_i)}{Var\,(y_{ij} \mid b_i)} \tag{3}$$

후보모형에서의 분산성분 추정량을 $\hat{\tau^2}$, $\hat{\sigma^2}$ 라고 하고 설명변수에 의하여 설명되는 변동을 $\sigma_f^2 = Var(\sum_{k=1}^{p} \beta_k x_{ijk})$라고 하면 주변부 결정계수 R_m^2과 조건부 결정계수 R_c^2의 추정량은 다음과 같이 주어질 수 있다.

$$\widehat{R_m^2} = \frac{\hat{\sigma_f^2}}{\hat{\sigma_f^2} + \hat{\tau^2} + \hat{\sigma^2}}, \qquad \widehat{R_c^2} = \frac{\hat{\sigma_f^2} + \hat{\tau^2}}{\hat{\sigma_f^2} + \hat{\tau^2} + \hat{\sigma^2}} \tag{4}$$

참고로 식 (2)에서 정의된 두 개의 결정계수 R_1^2과 R_2^2 은 모두 주변부 결정계수이다. 또한 조건부와 주변부 결정계수는 설명변수가 포함된 후보모형만을 적합하여 구할 수 있는 장점이 있다. 따라서 조건부 및 주변부 결정계수는 식 (2)에 주어진 변동을 설명하는 결정계수처럼 추정값이 음수로 나타나는 경우는 없다.

Orelien and Edward(2008)은 식 (2)에서 제시된 변동을 설명하는 결정계수와 식 (3)에 제시된 조건부와 주변부 결정계수의 효율을 모의실험으로 비교하고 식 (2)에 제시된 결정계수의 효율이 낮다는 연구를 제시하였다. Nakagawa and Schielzeth(2013)은 선형혼합모형에 대한 다양한 결정계수들의 정의와 장단점을 논의하고 계산하기 쉬운 조건부와 주변부 결정계수를 제시하였으며 더 나아가 일반화 선형혼합모형에 대한 결정계수도 제안하였다(Nakagawa et al., 2017). Jaeger et al.(2018)은 분산성분에 대한 선택을 할 수 있는 결정계수를 제안하였다.

4. 결론

혼합모형에서 결정계수를 계산할 수 있는 통계 프로그램의 명칭과 함수은 다음 표 1에 있다. 참고로 R 패키지 sjplot 의 tab_model 함수도 선형혼합모형의 요약를 출력하는 경우 주변부와 조건부 결정계수를 나타내어 준다.

선형혼합모형과 일반화선형혼합모형은 독립이 아닌 계층적 자료를 분석하는데 최근 많은 분야에서 사용되고 있는 모형이다. 일반 선형모형에서 사용되는 모형 선택 방법을 혼합모형에 확장하여 적용할 때 예측의 범위와 변동의 정의에 따라서 상이한 측도가 얻어진다. 실제 자료에 선형혼합모형을 적용하는 경우 만약 다수의 후보모형이 있는 경우가 흔하다. 이러한 경우 모형을 적합하는 목적에 따라서 모형선택을 위한 예측의 범위를 결정하거나(조건부 예측 또는 주변부 예측) 관심있는 변동을 목표로 정하여 적합한 결정계수를 사용해야 합리적인 분석이 가능하다. 일반적으로

개체(1차 계층)에 대한 예측과 그룹(2차 계층)에 대한 예측 중에서 더 중요한 예측을 목표로 정하고 조건부 결정계수와 주변부 결정계수 중 하나의 측도를 선택하여 모형을 선택하는 것이 바람직한 방향이다.

Table 1. Packages implementing R^2 for linear mixed effects model

package	language	function	R^2	참고문헌
performance	R	r2_nakagawa	conditional, marginal	Nakagawa and Schielzeth(2013, 2017)
r2glmm	R	r2beta	conditional, marginal variance components	Nakagawa and Schielzeth(2013,2017) Jaeger et al.(2018) Edward wt al.(2008)
partR2	R	partR2	conditional, marginal, recursive	Nakagawa and Schielzeth(2013,2017)
WLS	5.7%	1.7%	0.5%	0.1%

References

Edwards, L. J., Muller, K. E., Wolfinger, R. D., Qaqish, B. F., Schabenberger, O. (2008). An R2 statistic for fixed effects in the linear mixed model. *Statistics in Medicine*, 27(29), 6137 – 6157.

Jaeger, B. C., Edwards, L. J., Gurka, M. J. (2018). An R 2 statistic for covariance model selection in the linear mixed model. *Journal of Applied Statistics*, 46(1), 164 – 184.

Johnson, P. C. D. (2014). Extension of Nakagawa & Schielzeth's R2GLMM to random slopes models. *Methods in Ecology and Evolution,* 5(9), 944 – 946.

Lee, Y. (2014) Comparison of Model Selection Methods based on AIC for Linear Mixed Models under Crossover Design, *Journal of the Korean Data Analysis Society*, 16(1), 155-163. (in Korean).

Lee, Y. (2015) Simulation Study on Bayesian Model Selection Methods for Variance Components Models, *Journal of the Korean Data Analysis Society*, 17(5), 2383-2390. (in Korean).

Lee Y. (2015) A Note on Performance of Conditional Akaike Information Criteria in Linear Mixed Models. *Communications for Statistical Applications and Methods*, 22, 507-518.

Muller, S., Scealy, J. L., Welsh, A. H. (2013). Model Selection in Linear Mixed Models. *Statistical Science*, 28(2), 135 – 167.

Nakagawa, S. & Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed effects models. *Methods in Ecology and Evolution*, 4(2), 133 – 142.

Nakagawa, S., Johnson, P. C. D. & Schielzeth, H. (2017). The coefficient of determination R2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *Journal of The Royal Society Interface*, 14(134), 20170213.

Orelien, J. G., Edwards, L. J. (2008). Fixed-effect variable selection in linear mixed models using R2 statistics. *Computational Statistics & Data Analysis*, 52(4), 1896 – 1907.

Snijders, T., Bosker, R. (1994). Modeled Variance in Two-Level Models. *Sociological Methods & Research*, 22(3), 342 – 363.

Xu, R. (2003). Measuring explained variation in linear mixed effects models. *Statistics in Medicine*, 22(22), 3527 – 3541.

Vaida, F. & Blanchard, S. (2005). Conditional Akaike information for mixed-effects models, *Biometrika*, 92(2), 351 – 370.

Vonesh, E. F., Chinchilli, V. M., Pu, K. (1996). Goodness-of-fit in generalized nonlinear mixed-effects models. *Biometrics*, 52(2), 572 – 587.

Proceedings of the Korean Data Analysis Society
January 26-27, 2023, pp. 123-123

Quantile Estimation for Encrypted Data

Minjae Park[1], *Jaeseon Kim*[2], *Sungchul Shin*[3], *Cheolwoo Park*[4], *Jongjun Jeon*[5], *Hosik Choi*[6]

Abstract

Third-party analyzer which uses personal information may expose sensitive information. As data-based researches increase, such privacy issues are being raised seriously. Homomorphic encryption (HE) has been proposed as a way to avoid it. However, a challenge in ciphertext used in HE is a calculation time of elementary operations, which has significantly much higher complexity than that on plaintexts, resulting in that subsequent data analysis limiting various statistical analytics compared to plaintext data. In this paper, we consider an estimation method to get quantiles for encrypted data, where quantiles are core statistics to figure out data distribution in statistical data analysis. We propose a HE-friendly algorithm for large homomorphic encrypted data using an approximate quantile loss function. Numerical studies show that our proposed method improves calculation time over an estimation based on a sorting way for randomly generated and real data's homomorphically encrypted data. Furthermore, we suggest a modified boxplot for homomorphically encrypted data, where a boxplot is a primary statistical analysis method necessary for data analysis.

keyword : median; fully homomorphic encryption; optimization.

---

*본 연구는 인공지능혁신공유사업단 학술연구비를 자원을 지원 받아 수행되었습니다.

[1]08826 Research Engineer, Crypto Lab, 404, Gwank-ro, Gwank-gu, Seoul Korea. E-mail: minje@cryptolab.co.kr
[2]08826 Research Engineer, Crypto Lab, 404, Gwank-ro, Gwank-gu, Seoul Korea. E-mail: kjs@cryptolab.co.kr
[3]08826 Head of Data Business Development Division, Crypto Lab, 404, Gwank-ro, Gwank-gu, Seoul Korea. E-mail: scshin73@cryptolab.co.kr
[4]34141 Professor, Department of Mathematical Sciences, KAIST, 291 Daehak-ro Yuseong-gu Daejeon Korea. E-mail: parkcw2021@gmail.com
[5]02504 Associate Professor, Department of Statistics, University of Seoul 163 Seoulsiripdaero, Dongdaemun-gu, Seoul Korea. E-mail: jj.jeon@gmail.com
[6](Corresponding Author) Associate Professor, Department of Urban Big Data Convergence, University of Seoul 163 Seoulsiripdaero, Dongdaemun-gu, Seoul Korea. E-mail: choi.hosik@uos.ac.kr

Proceedings of the Korean Data Analysis Society
January 26-27, 2023, pp. 125-125

A study on multiple imputation modeling for Korean EAPS

Minjeong-Park[1], *Yoonjong-Bae*[2], *Joungyoun-Kim*[3]

Abstract

The Korean Economically Active Population Survey (KEAPS) is a national survey that produces employment-related statistics. The main purpose of the survey is to find out the economic activity status (employed/~unemployed/~non-employed) of the people. KEAPS has a unique characteristics caused by the survey method. In this study, through understanding of structural non-response and utilization of past data, we would like to present an improved imputation model. The performance of the proposed model is compared with the existing model through simulation. The performance of the imputation models is evaluated based on the degree of mathing/non-matching rates. For this, we employ the KEAPS data in November 2019. For the randomly selected ones among the total 59,996 respondents, the six explanatory variables, which are critical in determining the economic activity states, are treated as non-response. The proposed model includes industry variable and job status variable in addition to the explanatory variables used in the precedent research. This is based on the linkage and utilization of past data. The simulation results confirm that the proposed model with additional variables outperforms the existing model in the precedent research. In addition, we consider various scenarios for the number of non-responders by the economic activity status.

Keywords : Economically active population survey, Imputation, Multivariate imputation by chained equations (MICE), Structural non-response, Data linkage.

[1]Ph.D. Statistics Korea, 713 Hanbatdae-ro, Seo-gu, Daejeon, 35220, Korea. E-mail: mjstat@korea.kr
[2]M.S., Hanmi Pharm.Co.,Ltd, wilyeseongdae-ro 14, songpa-gu, Seoul, 05545, Korea.
E-mail: gripdash@naver.com
[3]Associate Professor, Department of Artificial Intelligence, University of Seoul, Seoulsiripdae-ro, Dongdaemun-gu, Seoul, 02504, Korea. E-mail: joungyoun@uos.ac.kr

Proceedings of the Korean Data Analysis Society
January 26-27, 2023, pp. 127-127

Development of Demand Response Transit mode choice models using Quantum Computing
-Optimization of Spatial Autocorrelation focusing on bus users in Seoul

Cheolgyu-Park[1], *Hosik-Choi*[2]

Abstract

Recently, various models have been used as a way to reflect the real of bigdata in models. The logic model is wildly used as a measure which transit mode choice. Since the traditional computing theory is based on the maximum random utility and probability, however, it is difficult to model a dynamic individual behavior causality. In this paper, we consider a regression Quantum Computing framework for modeling such a Spatial autocorrelation and bounded error of polynomial time. For this, we consider two models which are constructed by Ordinary Least Squares (OLS) Model and Quantum Geographically Weighted Regression (QGWR) Model based on Quantum Density Matrix method. In this study, we consider a develop an Quantum algorithm that efficiently calculates the predictor for demand response transit mode choice. As a practical example of the proposed method, we estimate effects of those of the users on transit mode choice from the Seoul. we apply floating population data, trip chain, land use attributes (metro data were aggregated in to 500m×500m buffer boundary of Seoul) data. We identify local spatial pattern of a relation with Quantum regression method. From two simulated data sets, as true model has more nonlinear structure in Hyper-computation, QGWR Model performs better than OLS Model in mean squared error rate and Local Moran's I Also, the result of the analysis about floating population rates shows that a Spatial autocorrelation effects of spatial heterogeneity and spatial similarities trip chain rates grows as floating population rate is higher. Therefore when selecting the best of mode choice, it must be based on the real time of polynomial space considering floating population due to Personalized optimization with user's.

Keywords : Demand Respons Transit mode choice, Quantum Computing, Polynomial.

---

[1]Member. Ph.D. Candidate, Department of Urban Big Data Convergence, University of Seoul, Seoulsiripdae-ro, Dongdaemun-gu, Seoul, 134-2, Korea. E-mail: carl86@nate.com
[2]Professor, Department of Urban Big Data Convergence, University of Seoul, Seoulsiripdae-ro, Dongdaemun-gu, Seoul, 134-2, Korea. E-mail: choi.hosik@uos.ac.kr

Proceedings of the Korean Data Analysis Society
January 26-27, 2023, pp. 129-132

슬링을 이용한 열린사슬운동과 닫힌사슬운동의 턱당기기 동작 시 목굽힘근 두께 비교*

이주훈[1], 한진태[2]

요 약

목적: 전방머리자세를 가진 사람을 대상으로 슬링을 이용하여 열린-닫힌사슬 운동 시 목굽힘근 두께 변화를 알아본다.

방법: 슬링을 이용하여 31명의 대상자에게 열린사슬과 닫힌사슬의 턱당기기운동을 실시하였다. 모든대상자는 슬링을 이용하여 턱당기기운동 시 초음파를 이용하여 목굽힘근 두께를 측정하였다. 열린사슬운동과 닫힌사슬운동 사이의 목굽힘근 두께 비교를 위해서 대응표본 t-검정을 사용하였으며, 총목굽힘근, 얕은목굽힘근 그리고 깊은목굽힘근의 두께를 비교하였다.

결과: 얕은목 굽힘근두께는 열린사슬운동 보다 닫힌사슬운동에서 더 감소하였다, 하지만 통계학적으로 유의하지는 않았다($p>0.05$). 반면, 총목굽힘근과 깊은목굽힘근의 두께는 열리사슬운동보다 닫힌사슬운동에서 증가하였으며, 통계학적으로 유의하였다($p<0.05$).

결론: 본 연구에서는 깊은 목굽힘근의 활성화를 위해서는 턱당기기 운동시 열린사슬운동보다 닫힌사슬운동이 효과적이며, 따라서, 슬링을 이용한 닫힌사슬의 운동이 전방머리자세를 가진 사람에게 열린사슬운동보다 효과적인 중재방법으로 제안한다.

주요용어 : 열린-닫힌사슬운동, 목굽힘근두께, 전방머리자세.

1. 서론

전방머리자세는 해부학적 중심선에 대해 머리가 앞으로 돌출되어 몸통보다 앞에 위치한 상태이며(Lee & Bae, 2015), 책상에 앉아서 장시간 일을 하게 되면 윗몸통이 경직된 상태에서 동일한 근육에 반복적으로 부하를 많이 받게 되고 그 결과 특정 근육과 인대의 과도한 긴장으로 해부학적 정렬의 변화가 일어나게 된다(Khayatzadeh et al., 2017). 특히 목빗근과 목갈비근은 과사용되고 긴목근과 긴머리근은 근력저하 및 기능 부재가 나타난다고 하였다(Falla et al., 2003).

열린사슬운동은 팔다리의 먼쪽에서 자유롭게 움직이고 몸쪽에서는 고정된 상태로 운동하는 방법이며, 닫힌사슬운동은 팔다리의 먼쪽이 고정되어 몸쪽과 먼쪽에서 저항을 동시에 적용하는 운동이다(Prentice et al., 2005). 관절의 안정화를 위해 실시하는 운동 중 닫힌사슬운동이 열린사슬운동보

---

*이 논문은 제1저자 이주훈의 석사학위논문의 축약본입니다.

[1]48434, 대한민국 부산광역시 남구 수영로 309, 경성대학교 물리치료학과 박사과정.
E-mail: dlwngns2722@naver.com
[2](교신저자) 대한민국 부산광역시 남구 수영로 309, 경성대학교 물리치료학과 교수.
E-mail: jthan2001@hanmail.net

다 더 많은 근육이 동원되었으며(Kwon & Lee, 2005) 근력강화에도 더 효과적이었다(Kwon et al., 2012). 또한 닫힌사슬운동이 열린사슬운동보다 고유수용성감각과 동적, 정적 균형능력을 더 향상시켰다(Park & Lee, 2018).

본 연구에서는 전방머리자세를 대상으로 슬링을 이용한 열린사슬운동과 닫힌사슬운동에서의 턱당기기 운동 시 목굽힘근의 두께에 미치는 영향을 알아보고, 임상에서 전방머리자세를 개선하기 위한 효과적인 운동방법을 제시하고자 한다.

2. 연구 방법

2.1. 연구대상자

본 연구의 대상자는 전방머리자세가 있는 20-30대 사무직 종사자 31명을 대상으로 하였다. 본 연구의 대상자 선정기준은 머리-목뼈각(craniovertebral angle, CVA)이 50도 이하인 자(Kim et al., 2011), 신체질량지수 범위가 18.5~24.9㎏/㎡인 자로 하였다. 제외기준은 목이나 어깨수술 이력이 있는 자, 고혈압 및 당뇨와 같은 만성질환이 있는 자, 누운자세에서 턱당기기 동작 시 통증을 호소하는 자로 하였다, 실험에 대한 충분한 설명과 방법을 인지한 후 본 실험에 참여하였다.

연구에 참여한 대상자의 일반적 특성은 다음과 같다.

Table 1. 일반적 특성

	대상자의 일반적 특성(n=31)
Age	29.25±8.86
BMI	22.54±1.26

Each value represents the mean±SD, BMI: Body Mass Index

2.2. 실험절차

운동방법은 바로 누운자세로 어깨가 치료테이블끝에 위치하도록 하고 슬링이 뒤통수 부위에 오도록한다. 검사자는 연구대상자에게 바깥귀길과 어깨봉우리를 일치시켜서 목의 중립자세를 만든 후 연구대상자가 목뼈 중앙부를 뒤쪽으로 천천히 누르게 하여 목뼈 앞굽음을 감소시키도록 하였다. 이때 이 자세를 최소한 10초 유지하라고 지시하였다(Yun & Kim, 2013). 열린사슬운동과 닫힌사슬운동에서의 턱당기기 운동을 5회 연습하여 충분히 동작을 익히도록 하였고 열린사슬운동과 닫힌사슬운동 시 목굽힘근의 두께를 총 3번 측정하여 그 평균값을 사용하였다. 열린사슬운동과 닫힌사슬운동은 뒤통수 부위에 위치하게 되는 줄의 종류로 구분하였으며, 턱당기기 운동 시 탄성이 있는 고무줄을 사용하여 머리가 움직인다면 열리사슬운동, 탄성이 없는 줄을 사용하여 몸통이 움직인다면 닫힌사슬운동이다. 목굽힘근 두께는 턱 당기기 동안 초음파 측정기(Prosound 2, Hitachi Aloka Medical, Japan)를 이용하여 얕은목굽힘근인 목빗근과 깊은목굽힘근인 긴목근과 긴머리근 두께를 측정하였다.

3. 자료 분석

턱 당기기 동안 슬링을 이용한 열린사슬운동과 닫힌사슬운동 시 목굽힘근의 차이를 비교하기 위하여 대응비교(paired t-test)를 이용하였고 목굽힘근 두께를 기술통계량으로 제시하였다. 자료 분석을 위해 SPSS version 25.0을 사용하였으며, 통계적 유의수준은 0.05로 하였다

4. 연구 결과

3.1. 총 목굽힘근 두께 비교

턱 당기기 동안 슬링을 이용한 열린사슬운동과 닫힌사슬운동 시 총 목굽힘근의 평균 두께는 열린사슬운동에서 26.49±6.10㎜이었고 닫힌사슬운동에서 28.11±6.23㎜로 나타났다 t값은 -4.55, p값은 0.00으로 통계학적으로 유의한 차이가 있었다.

Table 2. 총목굽힘근 두께 비교

	Open kinetic chain(mm)	Closed kinetic chain(mm)	t	p
Thickness of TNF	26.49±6.10	28.11±6.23	-4.55	0.00*

Each value represents the mean±SD, *p<0..05
TNF: Total Neck Flexor

3.2. 얕은목굽힘근 두께 비교

턱 당기기 동안 슬링을 이용한 열린사슬운동과 닫힌사슬운동 시 얕은목굽힘근인 목빗근의 평균 두께는 열린사슬운동에서 10.63±3.12㎜이었고 닫힌사슬운동에서 10.77±3.12㎜ t값은 -0.55, p값은 0.00로 통계학적으로 유의한 차이가 없었다.

Table 3. 얕은목굽힘근 두께 비교

	Open kinetic chain(mm)	Closed kinetic chain(mm)	t	p
Thickness of SNF	10.63±3.12	10.77±3.12	-0.55	0.58

Each value represents the mean±SD
SNF: Superficial Neck Flexor

3.3. 깊은목굽힘근 두께 비교

턱 당기기 동안 슬링을 이용한 열린사슬운동과 닫힌사슬운동 시 깊은목굽힘근인 긴목근과 긴머리근의 평균 두께는 15.85±3.81㎜이었고 닫힌사슬운동에서 17.33±4.15㎜ t값은 -5.57, p값은 0.00으로 통계학적으로 유의한 차이가 있었다.

Table 4. 발등 굽힘 시 하지 근육의 근활성도 비교

	Open kinetic chain(mm)	Closed kinetic chain(mm)	t	p
Thickness of DNF	15.85±3.81	17.33±4.15	-5.57	0.00*

Each value represents the mean±SD, *p<0..05
DNF: Deep Neck Flexor

5. 고찰

본 연구는 턱 당기기 동안 슬링을 이용한 열린사슬운동과 닫힌사슬운동이 목 굽힘근 두께 변화에 어떤 영향을 미치는지를 알아보기 위해서 실시하였다. 전방머리자세를 완화해서는 닫힌사슬위치에서 턱당기기운동이 목의 자세와 안정성에 기여하는 깊은목굽힘근의 활성화하기에 효과적일 것으로 사료된다.

References

Lee GC, Bae WS. (2015). The effect of shoulder exercise program for improving forward head posture. *Journal of The Korean Society of Integrative Medicine*, 3(3), 1-8.

Khayatzadeh S, Kalmanson OA, Schuit D, et al. (2017). Cervical spine muscle-tendon unit length difference between neutral and forward head posture: biomechanical study using human cadaveric specimens. *Physical Therapy*, 97(7), 756-766

Falla D, Jull G, Paul D. (2003). An electromyographic analysis of the deep cervical flexor muscles in performance of craniocervical flexor. *Physical Therapy*, 83(10), 899–906.

Prentice WE, Voight ML. (2005). Techniques in musculoskeletal rehabilitation. McGraw-Hill.

Kwon YJ, Park SJ, Kim K. (2012). The effect of open and closed chain exercise on lower extremity muscle activity in adults. *Journal of the Korean Society of Physical Medicine*, 7(2), 173-182.

Kwon SB, Lee HO. (2005). Effect of closed and open kinetic chain exercise after cruciate ligament reconstruction. *The Journal of Korean Society of Physical Therapy*, 17(3), 297-310

Park JM, Lee SH. (2018) The effect of knee joint position sense through close and open kinetic chain exercise on the proprioception and balance ability of elderly people. *Journal of Korean Society for Neurotherapy*, 22(2), 27-32.

Proceedings of the Korean Data Analysis Society
January 26-27, 2023, pp. 133-136

전방머리자세 환자의 목기능, 통증, 목빗근 두께가 삶의 질에 미치는 영향*

강민지[1], 한진태[2]

요 약

전방머리자세환자의 목기능 불능 지수(r= -0.87**), 머리-척추각(r= 0.87**), 목 폄 근력(r= 0.86**), 통증(r= -0.91**), 목빗근 두께(r= -0.85**)가 상관관계가 있음을 나타내므로 삶의 질과 매우 높은 상관성이 있었고 목 굽힘(r= 0.67**), 목 폄(r= 0.74**), 목 돌림(r= 0.63**), 목 굽힘 근력(r= 0.73**)은 삶의 질과 상관성이 있음을 확인하였다. 다중회귀분석 결과 회귀방정식은 다음과 같았다. 삶의 질 = 89.67 - 0.30*목기능 불능 지수 + 0.32*머리 척추각 + 0.02*목 굽힘 + 0.02*목 폄 + 0.02*목 돌림 + 0.01*목 굽힘 근력 + 0.04*목 폄 근력 – 2.63*통증 - 16.95*목빗근 두께로 나타났다. 그중 전방머리자세 환자의 삶의 질에 영향을 미치는 주요 항목은 목기능 불능 지수, 통증, 목빗근 두께로 나타났다($p<0.05$). 본 연구는 전방머리자세 환자들의 삶의 질을 개선시키기 위한 치료 중재 시 통증과 목기능 뿐만아니라 목빗근 두께가 중요한 요인임을 확인하였다. 따라서 향후 본 연구는 전방머리자세 환자의 삶의 질 개선을 위한 치료 중재의 근거자료로 사용할 수 있을 것이다.

주요용어 :전방머리자세, 삶의 질, 목빗근 두께, 목기능.

1. 서론

어깨와 목의 통증을 호소하는 환자 중 전방머리자세 환자가 60%라고 보고되었으며 전방머리자세가 있는 사람들은 목 주위 통증을 호소하고, 목의 불안정성과 불균형을 나타낸다(Chiu, 2002; Falla, 2004; Yip, 2008). 선행 연구에서 전방머리자세 그룹의 목빗근 두께가 더 두껍게 나타났다(Chow et al., 2007; Li & Chow, 2016 Bokaee et al., 2017).

최근 목통증은 삶의 질(quality of life)에 심각한 문제를 발생시키며 전체 인구의 67%가량이 목통증을 경험하고, 목통증에 대한 치료를 받지 않은 사람 중 57%가 시간이 지나도 통증이 완화되지 않아 이로 인해 심리적 불안으로 이어져 삶의 질이 저하되었다는 내용의 선행연구가 보고되었다(Wang et al., 2003). 따라서 본 연구는 전방머리자세 환자들을 대상으로 목기능 불능 지수, 머리-척추각, 목의 운동 범위, 목의 근력, 통증, 목빗근 두께 등이 삶의 질에 어떻게 영향을 미치는지 분석하여 전방머리자세 환자들을 위한 치료 중재 시 과학적 근거를 제시하고자 한다.

---

*이 논문은 제1저자 강민지의 석사학위논문의 축약본입니다.

[1]48434, 대한민국 부산광역시 남구 수영로 309, 경성대학교 물리치료학과 박사과정.
E-mail: alswl1633@naver.com
[2](교신저자) 대한민국 부산광역시 남구 수영로 309, 경성대학교 물리치료학과 교수.
E-mail: jthan2001@hanmail.net

2. 연구 방법

연구대상자는 B광역시 소재의 T병원 전방머리자세환자를 대상으로 선정기준으로는 머리-척추각(CVA)이 50°이하, 목기능 불능 지수(NDI) 5 이상, 목에 신경학적인 문제가 없는 대상자, 제외기준으로는 해당 부위에 외과적 수술, 정신과적인 약물을 복용 중, 심각한 외상, 전신에 퇴행성 질환, 섬유근육통이 있는 대상자는 제외하였다. 대상자들에게 실험 참여 동의를 받았으며, G-power 프로그램을 이용하여 측정횟수 1번, 효과 크기(effect size)는 0.35, 유의수준(α)은 0.05, 검정력은 80%로 설정하여 총 49명으로 산출되었다. 이후 탈락률 10%를 반영하여 총 54명의 대상자가 연구에 참여하였다.

2.1. 실험 절차

1) 삶의 질 평가

삶의 질 평가는 한국판 세계 보건기구 삶의 질 설문지로 평가하였다(WHOQoL-BREF)(M.J. Jeon, 2014).

2) 목기능 불능 지수((Neck Disability Index) 평가

한국판 목기능 불능 지수(NDI)를 사용하였다(이은우 등, 2007).

3) 머리-척추각 측정

C7의 가시돌기와 귀구슬(tragus)을 직선으로 연결하여 측정하였다(Nemmers et al., 2009).

4) 목 관절운동범위 측정

디지털 경사계(digital inclinometer, JTech Medical, Salt Lake City, Utah, USA)를 사용하여 굽힘, 폄, 돌림을 각각 측정하였다(Ghorbani et al., 2020).

5) 목 근력 측정

목의 근력은 디지털 근력측정기(Commander Power Track, Pro Healthcare Products, USA, 2011)를 이용하여 굽힘, 폄 근력을 검사하였다(박진식, 2015).

6) 시각 아날로그 척도 측정

눈금의 0부터 10까지 대상자가 느끼는 통증을 표시한 위치를 확인해 점수로 나타내었다(Jensen et al., 2003).

7) 목빗근 두께 측정

목빗근 두께는 초음파 영상진단 장치(ultra sound, Versana Premier, GE Medical systems, China)를 이용하여 측정하였다(Jesus et al., 2011)

3. 자료 분석

본 연구에서 전방머리자세 환자의 목기능과 통증 그리고 목빗근 두께와 삶의 질과의 상관성을 알아보기 위해 상관관계분석을 하였고 삶의 질에 가장 많은 영향을 미치는 요인을 확인하기 위해 다중 회귀분석을 하였다. 그리고 수집된 모든 자료를 컴퓨터를 이용해 SPSS version 26.0(IBM SPSS Inc., USA)을 사용하여 분석하였으며, 통계적 유의수준은 0.05로 하였다.

4. 연구 결과

본 연구에서 전방머리자세 환자의 목기능과 통증, 목빗근 두께를 평가한 결과는 목기능 불능 지수 24.35±3.58 점, 머리-척추각은 44.50±1.31 °, 목 굽힘 29.80±3.48 °, 목 폄 31.11±3.79 °, 목 돌림 26.89±2.36 °, 목 굽힘 근력 28.98±4.39 N, 목 폄 근력 29.20±5.33 N, 통증 7.68±0.73 점, 목빗근 두께는 0.95±0.05 mm으로 나타났다. 삶의 질 정도는 평균 63.93±4.80 점으로 나타났으며, 목기능 불능 지수(r= -0.87**), 머리-척추각(r= 0.87**), 목 굽힘(r= 0.67**), 목 폄(r= 0.74**), 목 돌림(r= 0.63**), 목 폄 근력(r= 0.86**), 목 굽힘 근력(r= 0.73**), 통증(r= -0.91**), 목빗근 두께(r= -0.85**)가 0.8이상은 매우 상관관계가 높음, 0.6이상은 상관관계가 높음, 0.4이상은 상관관계가 있음을 나타내므로 통계학적으로 삶의 질과 유의한 상관관계를 나타냈다($p<0.01$).

본 연구에서는 전방머리자세 환자의 삶의 질에 영향을 미치는 요인을 찾기 위해 목기능 불능 지수, 머리-척추각, 목 굽힘, 목 폄, 목 돌림 각도와 목 굽힘 근력, 목폄 근력, 통증, 목빗근 두께를 독립변수로 하고 동시입력으로 하여 다중 회귀분석을 시행하였다.

그 결과 유의한 예측모형이 확인되었으며 종속변수에 대해 87%의 설명력으로 높은 설명력을 보이고 잔차의 독립성을 나타내는 Durbin-Watson값이 2.251으로 자기상관이 없음을 나타냈다. 예측모형의 예측자는 목기능 불능 지수, 머리-척추각, 목 굽힘, 목 폄, 목 돌림 각도와 목 굽힘 근력, 목폄 근력, 통증, 목빗근 두께였으며, 예측 모형의 분산분석 결과 유의수준이 0.05 미만으로 나타나 적합한 회귀식이 있음을 알 수 있었다.

회귀방정식 도출을 위한 비표준화 계수 B값은 머리-척추각 0.32, 목빗근 두께 -16.95, 목 굽힘 0.02, 목 폄 0.02, 목 돌림 각도 .026와 목 굽힘 근력 0.01, 목 폄 근력 0.04, 통증 -2.63, 목기능 불능 지수 -0.30으로 이 변수들이 전방머리자세 환자의 삶의 질에 영향을 미치는 중요한 예측 인자임을 나타냈다.

본 연구의 회귀분석 결과에서 분산팽창요인(Variance Inflation Factor)은 10 이하로 다중공선성이 없다고 판단되며 이는 모형이 적절한 회귀모형임을 알 수 있다. 따라서 회귀방정식은 전방머리자세 환자의 삶의 질 = 89.67 - 0.30×목기능 불능 지수 + 0.32×머리 척추각 + 0.02×목 굽힘 + 0.02×목 폄 + 0.02×목 돌림 + 0.01×목 굽힘 근력 + 0.04×목 폄 근력 – 2.63×통증 - 16.95×목빗근 두께로 나타났다.

5. 고찰

본 연구에서 전방머리자세 환자의 목빗근 두께와 삶의 질 사이의 연관성을 확인한 것은 의미

있는 성과라고 생각한다. 따라서, 전방머리자세 환자들의 삶의 질을 증진시키기 위하여 목기능 불능 지수, 머리-척추각, 통증, 목빗근 두께 개선을 모두 고려하여 중재를 하여야 할 것이라 사료된다.

References

Chiu, T. T., Ku, W. Y., Lee, M. H. (2002). A study on the prevalence of and risk factors for neck pain among university academic staff in Hong Kong. *Journal of Occupational Rehabilitation*, 12(2), 77-91.

Falla, D. L., Jull, G. A., Hodges, P. W. (2004). Patients with neck pain demonstrate reduced electromyographic activity of the deep cervical flexor muscles during performance of the craniocervical flexion test. *Spine*, 29(19), 2108-2114.

Yip, C. H., Chiu, T. T., Poon, A. T. (2008). The relationship between head posture and severity and disability of patients with neck pain. *Manual Therapy*, 13(2), 148–154.

Li, S. S., & Chow, D. H. (2016). Multi-objective analysis for assessing simultaneous changes in regional spinal curvatures under backpack carriage in young adults. *Ergonomics*, 59(11), 1494-1504.

Chow, D. H,, Leung, K. T., Holmes, A. D. (2007). Changes in spinal curvature and proprioception of schoolboys carrying different weights of backpack. *Ergonomics*, 50(12), 2148-2156.

Bokaee, F., Rezasoltani, A., Farideh, D. et al. (2017). Comparison of cervical muscle thickness between asymptomatic women with and without forward head posture. *Brazilian Journal of Physical Therapy*, 21(3), 206-211.

Wang, W. T. J., Olson, S. L., Campbell, A. H. et al. (2003). Effectiveness ofphysical therapy for patients with neck pain: An individualized approach using a clinical decision-making algorithm, *American Journal of Physical Medicine and Rehabilitation*, 82(3), 203-218.

Jeon M-J., Jeon H-S., Yi C-H. (2014). Comparison of Elderly Fallers and Elderly Non-Fallers: Balancing Ability, Depression, and Quality of Life, *Physical Therapy Korea*, 21(3), 45-54.

Eun-woo Lee., Won-seob Shin., Kyoung-sim Jung., Yi-jung Chung. (2007). Reliability and Validity of the Neck Disability Index in Neck Pain Patient. *Physical Therapy Korea*, 14(3), 97-103.(in Korean)

Ghorbani, F., Kamyab, M., Azadinia, F. (2020). Smartphone Applications as a Suitable Alternative to CROM Device and Inclinometers in Assessing the Cervical Range of Motion in Patients With Nonspecific Neck Pain. *Journal of Chiropractic Medicine*, 19(1), 38-48.

Nemmers, T. M., Miller, J. W., Hartman, M. D. (2009). Variability of the Forward Head Posture in Healthy Community dwelling Older Women. *Journal of geriatric physical therapy*, 32(1), 10-14.

Park,Jin-Sik. (2015). Effects of closed and open kinetic chain exercise on thicknesses, strength, endurance of neck muscle in normal adults. *Departmentof Rehabilitation Science Graduate School, Daegu University Gyeongbuk, Korea.* (in Korean)

Jesus, M. F. R., Ferreira, P. H., Pereira, L. S. M. et al. (2011). Ultrasonographic analysis of the neck flexormuscles in patients with chronic neck pain andchanges after cervical spine mobilization. Journal of Manipulative and Physiolloqical Therpeutics, 34(8), 514-524.

Jensen, M. P., Chen, C., Brugger, A. M. (2003). Interpretation of Visual Analog Scale Ratings and Change Scores: A Reanalysis of Two Clinical Trials of Postoperative Pain. Journal of Pain, 4(7), 407–414.

Proceedings of the Korean Data Analysis Society
January 26-27, 2023, pp. 137-139

발가락 사용유무가 능동적 발등굽힘 가동범위 및 하지근활성도에 미치는 영향*

김인균[1], 최수홍[2], 이상열[3]

요 약

본 연구의 목적은 능동적 발등 굽힘 시 발가락의 움직임의 유무가 발목움직임에 미치는 영향과 하지근육 활성도 에 미치는 영향을 알아보고자 하였다. 연구에 참여한 인원은 25명으로 발목관절과 무릎관절에 이상이 없는 건강한 대상자가 참여 하였다. 연구방법은 대상자는 편안하게 누운 상태에서 능동적 발등 굽힘 시 발가락 움직임 유무에 따라 발목관절 가동범위 및 하지 근육 의 근활성도를 측정 하였다. 통계처리는 SPSS 25.0을 사용하였으며, 유의수준은 0.05로 하였다. 결과는 능동적 발등 굽힘 시 발가락 움직임이 있을 때 보다 발가락의 움직임 없을 때 발목관절 가동범위가 더 크게 나타났으며 통계적으로 유의하였다. 하지만 발가락 움직임이 없을 때 보다 발가락의 움직임이 있을 때 긴엄지 발가락 폄근과 앞정강이근, 넙다리곧은근의 근활성도가 더 증가하였으며 통계적으로 유의하였다. 따라서 본 연구에서는 능동적 발등 굽힘 시 발목의 가동범위 증진을 위해서는 발가락의 움직임을 사용하지 않아야 되지만, 하지 근육의 근활성도 증진을 위해서는 발가락의 움직임을 먼저 사용 후 발목 발등 굽힘을 시켜야 된다는 것을 알 수 있다.

주요용어 : 발등 굽힘, 가동범위, 근활성도.

1. 서론

정상적인 보행을 위해서는 발목 관절의 정상적인 발등 굽힘 가동범위가 중요하다(Jordan, 1979). 정상 발등 굽힘 가동 범위는 보행 중 유각기에 많은 영향을 미친다(Kim et al, 2001). 특히 정상 보행 시 최소 10도 이상 발등 굽힘 가동범위가 요구된다. 하지만 발등 굽힘에 문제가 생기면 보행뿐만 아니라 더 많은 발등 굽힘 가동 범위가 요구되는 계단 내려오기, 쪼그려 앉기와 같은 일상생활 동작이 어려울 수도 있다(Collins et al, 2004).

발등 굽힘 제한의 원인은 아킬레스 힘줄 염증(wilderd and sethi, 2004), 앞정강근 스트레스 증후군(Lillerdt et al, 1979), 무릎넙다리 증후군(Messier and Pittala, 1988), 불편한 신발 착용으로 생긴 과도한 근육 피로(Ko et al, 2008)와 같은 근육 및 힘줄 등의 문제로 발생된다.

하지만 대부분의 선행 연구들은 발등 굽힘 측정 시 발가락의 사용 유무에 대한 내용이 빠져 있다. 특히 임상에서 대부분 환자들은 발목 관절의 발등 굽힘 시 발가락의 움직임을 사용하며, 발가

---

*이 논문은 제1저자 김인균의 논문의 축약본입니다.

[1]48434, 대한민국 부산광역시 남구 수영로 309, 경성대학교 물리치료학과 박사과정.
E-mail: kinvirus2552@naver.com
[3](교신저자) 대한민국 부산광역시 남구 수영로 309, 경성대학교 물리치료학과 교수. E-mail: sjslh486@daum.net

락 움직임은 발목 관절의 발등 굽힘 가동 범위와 하지 근력에도 충분히 영향을 줄 수 있을 것이라고 생각된다.

그렇기 때문에 본 연구에서 발가락 움직임 유무가 능동적 발등 굽힘 가동범위와 하지 근육 활성도에 미치는 영향을 알아보고자 한다.

2. 연구 방법

2.1. 연구대상자

본 연구의 대상자는 발목관절과 무릎관절에 이상이 없는 20-30대의 정상인 25명을 대상으로 실험하였다. 실험 대상자는 자발적인 동의하에 참가하였으며, 실험에 대한 충분한 설명과 방법을 인지한 후 본 실험에 참여하였다.

2.2. 실험절차

대상자는 편안하게 누운상태에서 무릎관절 안정성을 위해 무릎관절 아래에 받침을 놓은 상태에서 시작하였다. 모든 측정은 우세발을 기준으로 측정하였다. 발목관절은 자연스러운 발목 굽힘 0도 상태에서 능동적으로 완전한 발등굽힘시에 발가락의 폄 사용 유무에 따른 발목관절 가동 범위 측정과 발가락의 폄 사용 유무에 따른 발등 굽힘 시 하지 근육 긴엄지발가락 폄근,앞정강이근과 넙다리곧은근의 근활성도를 측정하였다. 근육의 근전도 신호의 정규화하기 위해 최대 자발적 근수축(maximal voluntary contraction: MVC) 측정하였다. 하지 근육의 근활성도는 3번 측정 후 평균값을 측정하였다.

2.3. 자료 분석

발가락 사용유무에 따른 발목관절의 가동범위 비교를 위해 대응표본 t-test를 사용하였고, 발가락 사용에 따른 발목관절의 발등굽힘 근육간의 개시 시간 비교를 위해 대응표본 t-test 사용하였다. 발등 굽힘시 발가락 사용유무에 따른 하지근육의 근활성도 비교를 위해서 대응표본 t-test 사용하였고, 발가락 사용에 따른 하지근육간의 상관관계 분석을 위해 피어슨 상관관계분석을 사용하였다. 통계 프로그램은 SPSS 25.0(IBM SPSS Inc, USA)을 사용하였고, 유의수준은 0.05로 설정하였다.

3. 연구 결과

3.1. 발등 굽힘 시 발목 관절의 가동범위 비교

발목관절 가동범위는 발가락의 움직임을 사용하지 않은 상태에서 발등 굽힘시에는 22.63±4.96°로 나타났으며, 발가락의 움직임을 사용 한 발등 굽힘시에는 15.17±5.21° 로 통계학적으로 유의한 차이가 있었다.

3.2. 발가락 사용에 따른 발목관절의 발등굽힘 근육간의 개시 시간 비교

발가락 사용 후 발목관절의 발등 굽힘근의 개시 시간은 엄지발가락 폄근은 0.85±0.29로 나타났으며, 앞정강근은 1.05±0.35로 통계학적으로 유의한 차이가 없었다.

3.3. 발등 굽힘 시 발가락 사용 유무에 따른 하지근육의 근활성도 비교

발등 굽힙 시 긴엄지발가락 폄근의 근활성도는 발가락의 움직임을 사용하였을 때 13.75±5.54 이며, 발가락을 사용하지 않았을 때 13.91±7.58으로 통계적으로 유의한 차이가 없었다.

발등 굽힙 시 앞정강이근의 근활성도는 발가락의 움직임을 사용하였을 때 59.56±26.67 이며, 발가락을 사용하지 않았을 때 41.47±19.48으로 통계적으로 유의한 차이가 있었다.

발등 굽힘 시 넙다리곧은근의 근활성도는 발가락의 움직임을 사용하였을 때 12.95±8.01이며, 발가락을 사용하지 않았을 때 9.70±7.98으로 통계적으로 유의한 차이가 있었다.

3.4. 발등 굽힘 시 발가락 움직임 사용에 따른 하지 근육간의 상관 관계

발등 굽힙 시 발가락 폄 사용에 따른 긴엄지발가락과 앞정강근, 넙다리직근 상관 관계를 보기 위하여 피어슨의 상관계수를 측정하였다. 긴엄지발가락과 앞정강근은 유의한 양의 상관관계를 나타내었다($r=0.45$, $p<0.05$).

4. 고찰

이 연구의 목적은 발가락의 움직임 유무가 능동적 발등 굽힘의 가동 범위에 미치는 영향과 하지 근육(긴엄지발가락 폄근, 넙다리곧은근과 앞정강이근)에 미치는 영향을 알아보는 것이다. 그 결과 능동적 발등 굽힘 시 발가락 폄이 있을 경우 능동적 발등 굽힘 가동 범위가 유의하게 감소하였다. 반면에 발가락의 폄이 있을 경우 넙다리곧은근과 앞정강이근의 근활성도는 더 높은 수치의 유의함을 보였다.

References

Kim, D. Y., Park, C. I., Jang, Y. W., Park, S. Y. (2001). Kinematic and Kinetic Comparison between Stair Climbing and Level Walking, *Journal of the Korean Academy of Rehabilitation Medicine*, 25(6), 1048-1058.

Collins, N., Teys, P., Vicenzino, B. (2004). The initial effects of a Mulligan's mobilization with movement technique on dorsiflexion and pain in subacute ankle sprains, *Manual therapy,* 9(2), 77-82.

Ko, E. H., Choi, H. S., Kim, T. H., Cynn, H. S., Kwon, O. Y., Choi, K. H. (2008). The Effect of High-Heeled Shoes With Total Contact Inserts in the Gait Characteristics of Young Female Adults During Lower Extremity Muscle Fatigue, *Korean Research Society of Physical Therapy,* 15(1), 38-45.

Jordan, R. P., Cooper, M., Schuster, R. O. (1979). Ankle dorsiflexion at the heel-off phase of gait: a photokinegraphic study, *Journal of the American Podiatric Medical Association*, 69(1), 40-46.

Wilder, R. P., Shikha, S. (2004). Overuse injuries: tendinopathies, stress fractures, compartment syndrome and shin splints, *Clinics in sports medicine*, 23(1), 55-81.

Lilletvedt, J., Kreighbaum, E., Phillips, R. L. (1979). Analysis of selected alignment of the lower extremity related to the shin splint syndrome, *Journal of the American Podiatric Medical Association*, 69(3), 211-217.

Messier, S. P., Pittala, K. A. (1988). Etiologic factors associated with selected running injuries, *Medicine and Science in Sports and Exercise*, 20(5), 501-505.

Proceedings of the Korean Data Analysis Society
January 26-27, 2023, pp. 141-141

키워드 네트워크 분석을 통한 심부 디지털 트윈 국내외공공 연구기관 논문 분석*

안은영[1], 김성용[2]

요 약

영국 National Digital Twin, 버추얼 싱가포르(Virtual Singapore) 사례와 같이 많은 국가에서 공공 디지털 트윈으로 국토의 건물 및 시설물 내부까지 디지털화하고 있으나, 시설물보다 깊은 심부의 지질환경재해 정보를 제공하는 수준까지 미치지 못하고 있다. 자연재해인 지진은 단순히 개별 시설물 및 거주 인명에 위험뿐만 아니라 단층에 따라 도시 단위를 넘는 특성이 있으며 화재, 산사태 등의 1, 2차 재난 및 인공/자연 복합재해도 발생시킬 수 있다. 또한 지진뿐만 아니라 지역적 지질 특성, 가뭄/홍수 등 자연적인 현상, 주변 건물 및 공공시설의 개발에 따라 위해 광물자원 부존, 지하수 수위 하강, 지하공동화, 지하 폐기물 및 해수 침투 등 다양한 지질환경 위험이 존재한다. 본 연구는 지구 표면 정보에 기반한 현재의 도시 및 국가 규모의 공공 디지털 트윈에서 지질, 지형, 지질구조, 선구조, 광맥분포 등의 심부 정보를 제공하는 공공연구기관의 연구 현황을 알아보기 위해 논문 게재 정보를 기반으로 키워드 네트워크 분석을 실시하였다. 키워드 네트워크 분석은 텍스트로부터 의미 있는 개념들을 추출하여 이들의 관계를 분석하는 연구 방법으로 빈도분석, N-gram 네트워크분석, 연결중심성 및 TF-ID 가중치 분석 등을 실시한다(Lee, 2022). 본 연구에서는 심부 디지털 트윈에 관한 연구 자료 수집을 위해 3차 산업혁명(디지털화), 4차 산업혁명(인더스트리 4.0)의 중요 키워드인 digital, three dimension, real time, Artificial Intelligence가 포함된 심부지질 공공연구기관의 논문을 대상으로 분석하였다. Web of Science(WoS)에서 제목/초록/키워드(저자 및 WoS) 검색을 통해 국내외 논문을 수집하였으며, 소속 기준으로 상위 15개 공공연구기관의 총 1,702개 논문을 분석하였다. 미국지질조사소(USGS), 영국지질조사소(BGS)와 더불어 이탈리아 국립지구물리화산연구소(INGV)가 상위기관에 나타났으며 프랑스 국립지구과학천문연구소(INSU), 독일 지구과학연구센터(GFZ) 등이 나타났다. 이탈리아의 경우 심부지질 연구기관으로 지구과학광물자원연구소(IGG)가 있으나 분석 대상인 상위 15개 기관은 아니었다. 미국지질조사소(USGS) 외에 미국항공우주국(NASA)도 상위 7위 기관으로 나왔으며 프랑스의 경우 프랑스 국립지구과학천문연구소(INSU)가 제일 상위 기관이며 프랑스지질광산연구소(BRGM), 프랑스석유연구소(IFP) 또한 심부 디지털 트윈 관련 각각 10위, 11위 기관이다. 심부 디지털 트윈 구축을 위한 심부정보를 제공하는 공공연구기관을 우주이용, 자연재해, 지질조사 기관으로 구분할 수 있었으며 해당 기관의 논문 초록에 대한 빈도 분석, 워드 크라우드, 의미 연결망을 도출하여 지형학, 지도 등 일반 연구주제뿐만 아니라 탄성파, 화산, 지진, CO2, 지하수, 지열, 행성/위성, 인공위성/궤도선 등의 연구주제를 확인하였다.

References

Lee, J. Y. (2022). Research Trends on Adolescent Self-injury using Keyword Network Analysis: Focusing on Domestic Academic Journals Published from 2010-2021. *Journal of The Korean Data Analysis Society*, 24(1), 457-472.

---

*본 논문은 한국지질자원연구원 기본사업 지질자원 표본·기초학술연구와 선도형 R&D 정책/성과확산 연구에 의해 지원되었음(GP2020-020).

[1](교신저자) 34132 대전광역시 유성구 과학로 124, 한국지질자원연구원 책임연구원. E-mail: eyahn@kigam.re.kr
[2]34132 대전광역시 유성구 과학로 124, 한국지질자원연구원 책임연구원. E-mail: ksy@kigam.re.kr

Proceedings of the Korean Data Analysis Society
January 26-27, 2023, pp. 143-146

비가 온 후의 토양수분 변화 고찰 및 예측 모델 개발*

강소라[1], 조완현[2], 나명환[3]

요 약

토양수분은 식물이 자라는 데에 필수적인 요소로 기상환경 및 토양환경에 따라 생화학적으로 영향을 받는다. 구비대기 기간동안 적절한 토양수분 환경은 수확량 증대로 이어지고 관수 관리는 중요하다. 토양수분을 적절하게 유지해야 한다는 사실은 알려져 있으나 농업인들은 관리하는 데에 어려움을 겪고 다습한 환경조건에 관심이 많다. 기존에 통용적으로 사용하는 토양수분 예측 모델은 존재하지만 많은 환경 정보를 필요로 한다. 따라서 제한된 환경정보만을 이용하여 비가 온 이후의 즉각적인 토양수분 변화를 보여줄 수 있는 모델을 제시하고자 한다. 먼저, 상관분석을 통해 토양수분과 기상환경 및 토양환경 간의 관계를 규명하고, 토양수분 변화에 대한 시간적인 특성을 고려한다. 다음으로 선형 회귀분석과 시그모이드 함수를 활용한 비선형 회귀분석을 통해 구비대기 기간동안 비가 온 이후 양파, 마늘 토양수분 변화에 대한 통계적 모형식을 개발하였다. 궁극적인 목적은 생산성 향상으로 적정 토양수분을 알아내고 토양수분 예측 모델을 구축하고자 한다. 끝으로 밭과 논의 재배형식에 따른 차이를 비교하여 모델의 유용성을 확인하고 토양수분 관리에 도움을 줄 것으로 기대된다.

주요용어 : 토양수분, 양파, 마늘, 시차선형회귀분석, 비선형회귀분석.

1. 서론

식물이 자라는 데 있어서 수분은 필수적인 요소로 토양으로부터 흡수한다. 수분함량은 강수량, 관개 등 유입요소와 배수, 증발산 등 유출요소의 영향력에 따라 변화한다. 특히, 증발산량은 생육 및 기상환경, 토양의 상태에 따라 좌우되며, 주로 Pemman-Monteith 또는 Priestley- Taylor 방식으로 추정한다. 하지만 이러한 방법은 유입과 유출의 복잡한 관계식이 필요하며 기상 및 토양과 관련하여 많은 정보가 필요하다.

다음으로, 양파, 마늘은 다습한 환경에 민감하므로 토양수분의 적정량을 유지하는 것이 중요하다. 생육기 기간에는 10% 이상의 토양수분만 보유하면 생장에 무리가 없지만 구비대기 기간에 토양수분 함량이 높으면 열구, 무름병 등 작물 피해가 발생하고 수확량이 줄어든다. 따라서 비가 온 이후에는 관수가 중요하며, 반대로 물이 마르지 않도록 구비대기 기간에 일주일 간격으로 충분한

*이 논문은 농촌진흥청 연구사업(과제번호 PJ017089012022)의 지원과 교육부 및 한국연구재단의 4단계 두뇌한국21 사업 (4단계 BK21 사업, NO.5120200913674)으로 지원된 연구임.
[1](제1저자) 61186, 광주광역시 북구 용봉동 용봉로 77, 전남대학교 수학/통계학과 박사과정.
Email: sc12love@gmail.com
[2](공동저자) 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

물을 줌으로써 관리한다. 특히, 구비대기 기간인 4월~5월에 토양수분을 최적 상태로 유지하면 수확량이 증대한다.

본 연구에서는 토양수분과 기상환경 및 토양 환경과의 관계를 살펴보고 토양수분센서를 통해 수집된 기상환경 및 토양환경 자료만을 사용하여 비가 오고 난 직후에 토양수분 함량이 올라간 시점에서 낮아지는 동안의 토양수분 변화를 함수식으로 표현하였다. 뿌리 분포에 따라 수분의 흡수가 다르기 때문에 토양의 깊이 5m, 10m, 15m, 20m에 따라 추정하고자 한다. 또한, 관수는 중요한 정보이지만 모른다는 가정하에 비가 온 이후의 토양수분 변화를 관찰하였다. 최종적으로 상관분석, 선형 회귀분석을 통해 토양수분과 환경요인들의 관계를 알아보고 비선형 회귀분석을 통해 구비대기 기간동안 비가 온 이후 토양수분 변화를 예측하는 통계적 모델을 제시하려고 한다. 궁극적인 목적은 생산성 향상으로 적정 토양수분을 알아내고 토양수분 예측 모델을 구축하고자 연구가 진행되었다.

2. 분석자료 및 분석방법

2.1. 분석자료

분석 자료는 전라남도 우수 4농가로부터 수집된 기상환경 및 토양환경 자료이며, 토양수분센서를 통해 5분마다 측정되었다. 우수농가는 재배환경별(밭/논), 작물별(양파/마늘)로 다양하게 선정하였으며, 양파는 무안군, 밭마늘은 해남군, 논마늘은 고흥군 지역이다. 본 연구 자료는 2022년 4월 중하순에서 5월까지의 구비대기 기간 동안의 농가별 자료로 반응변수는 5cm, 10cm, 15cm, 20cm 토양수분, 설명변수는 온도(Temp), 습도(Humi), 일사량(Par), 강수량(RG), 토양수분장력(PF), 지온(Stemp), 토양EC(SEC) 등 7가지 환경정보이며, 자세한 내용은 다음 Table 1과 같다.

Table 1. List of response variable and explanatory variables

Variable type	Factor	Variable name	Explanation	Unit
Response	Soil moisture1	Swat1	Soil moisture measured at 5cm	cm
	Soil moisture2	Swat2	Soil moisture measured at 10cm	cm
	Soil moisture3	Swat3	Soil moisture measured at 15cm	cm
	Soil moisture4	Swat4	Soil moisture measured at 20cm	cm
Explanatory	weather environmental	Temp	Temperature	℃
		Humi	Humidity	%
		Par	Photosynthetically active radiation	W/m2
		RG	rainfall amount.	mm
	Soil environmental	PF	soil water tension	kPa
		Stemp	Soil temperature	℃
		SEC	Soil electrical conductivity	ms/cm

최종적으로 각 농가별로 토양수분 예측 모델에 영향을 미친 강우는 양파 4월 25일, 마늘 4월 29일 강수량으로 일일 누적 강수량은 다음 Table 2와 같다.

Table 2. Information of rainfall amount for soil moisture model

Division	Field-onion	Non-onion	Field-garlic	Non-garlic
Cumulative precipitation	2.2mm	3.0mm	3.2mm	1.2mm

2.2. 분석방법

먼저, 각 측정 구간별 토양수분과 기상환경 및 토양환경과의 관련성을 파악하기 위하여 추세 그래프를 통해 대략적으로 파악하고 이러한 관계를 통계적으로 검증하기 위해 상관분석을 실시하였다. 또한, 밭과 논의 재배방식과 양파 및 마늘 등 작물별로 차이가 있는지 비교분석하였다.

두 번째, 선형 회귀분석에서는 반응변수와 설명변수 간의 선형적 인과관계를 확인하고 비선형 회귀분석의 초기값 설정을 위한 목적으로 실행하였다 반응변수는 토양수분 로그값의 증가량으로 회귀분석의 선형성 가정을 만족하기 위해 로그 변환을 수행하였다. 설명변수는 온도, 습도, 일사량, 강수량 등 지상환경 요인과 토양수분장력, 토양EC, 지온 등 토양 환경요인이 사용되었다.

측정구간별 5cm, 10cm, 15cm, 20cm 토양수분의 관계를 확인을 위한 상관분석과 시차회귀분석의 목적은 지상환경과 가장 큰 영향을 미치는 5cm 토양수분에 토양의 깊이가 깊어질수록 몇 시차 간격을 두고 영향을 크게 미치는지 알아보기 위함이다. 회귀분석에서는 다음 (2.1)의 중선형 회귀모형을 고려하였고, 모형이 적합하다면 설명변수가 유의하지 않더라도 모든 환경을 사용하였다.

$$y = \beta_0 + \beta_1 \times x_1 + \beta_2 \times x_2 + \cdots + \beta_n \times x_n + \varepsilon \tag{2.1}$$

세 번째, 비선형 회귀분석에서는 변화가 없는 구간을 제외한 토양수분 증가량이 시그모이드 형태를 띤다는 점을 확인하고 토양수분 증가량의 누적 합계 값을 이용하여 시그모이드 함수를 적용하였다. 비가 오고 나서 토양수분 함량이 크게 올라간 이후 낮아지는 동안의 토양수분 함수식은 다음 (2.2)로 표현하였다. 비선형 회귀계수들의 초깃값 β는 토양수분 로그값의 증가량 회귀모델의 회귀계수, 초깃값 M은 토양수분의 최댓값을 토대로 작성하였고, EXP의 항은 log(토양수분 초깃값)에 log(토양수분)의 증가량의 누적값을 더한 값으로 토양수분의 로그값을 의미한다.

$$y = \frac{M}{1 + e^{cum(\log C_1 + \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n)}} \tag{2.2}$$

3. 결론 및 향후 연구

본 논문은 재배환경과 재배작물이 모두 다른 전라남도의 우수 4농가로부터 수집된 양파, 마늘의 토양수분 및 환경자료를 이용하여 토양수분과 기상환경 및 토양환경 간의 관계를 고찰하였다. 토양수분과 환경 변수들간의 관계는 추세 그래프와 상관 분석을 통해 확인하였으며, 측정 구간별 토양수분들 간의 관계는 지연시차 개념을 이용한 상관분석과 회귀분석을 이용하여 확인하였다. 마지막으로 토양수분 로그값의 증가량을 반응변수로 환경정보를 설명변수로 하여 선형 회귀분석을 실시하였고, 최종적인 토양수분 예측 모델은 시그모이드 함수를 가진 비선형 회귀분석을 활용하였다.

먼저 재배환경과 상관없이 토양EC, 토양수분장력, 강수량은 토양수분과 양의 상관성을 보였으며, 지온과 기온은 마늘 토양에서 토양수분과 음의 관계를 갖는 환경이었다. 강수량의 영향력은 적었으나 일일 누적량이 가장 많은 밭마늘에서는 토양수분에 가장 큰 연관성을 보였다. 또한, 논의 토양수분이 밭보다 높았으며, 수분함량의 변동 폭이 작고 안정적으로 변화하였으며. 강우 직후에 모두 급격하게 감소하였으나 몇 분 후에 안정적으로 변화하였다. 측정구간별 토양수분간의 관계는 측정구간이 깊어질수록 5cm 토양수분과 낮은 상관성을 보였으며, 시차를 두지 않았을 때 가장 높

은 상관성을 보였다. 특히 밭토양에서는 시차 간격이 길어질수록 상관성이 낮아졌으나 논토양에서는 큰 변화가 없었다.

선형 회귀분석의 결과는 일부 설명변수가 유의미하지 않았으나 비선형 회귀분석의 초깃값으로 사용하기에 큰 문제는 없었다. 또한, 변화가 없는 구간을 제외하고 토양수분 증가량이 시그모이드 함수 형태를 띠었다. 비선형 회귀분석 결과는 모든 환경 변수가 유의하게 나타났으며 모형의 정확도가 높게 나타났다. 비가 오고나서 토양수분함량이 크게 올라간 이후 낮아지는 동안의 토양수분 함수식은 실제값을 잘 추정하는 것으로 나타나 재배환경과 재배작물이 다르더라도 추정 모델이 유의하게 적용됨을 알 수 있었다.

본 연구는 관수 조절과의 상관성을 고려하지 않고 구비대기 동안의 토양수분을 관찰하였으며, 토양수분 증가량과 강수량의 상관성이 낮아 관수의 영향을 감안하기에는 부족한 부분이 있다. 이는 비가 오는 시점마다 모수의 추정치가 다르다는 점을 설명하며, 다른 모수는 고정된 상태에서 토양수분 최대값만 변하도록 모델을 구축할 필요가 있다. 향후 과제는 관수가 중요한 역할을 하므로 관수 관리 정보를 반영하여 비가 오지 않는 시점의 토양수분 변화를 고려할 필요가 있다.

References

Cho S. K., Lee E. T., Oh Y. J., Choi I. H., Kim Y. J., Kim J. G. (2006). Effect of paddy and upland conditions on yield and storage of onion bulbs. *Journal of the Korean Crop Society*, 51(5), 401-407.

Choi. W. J., Kim. Y. S., Shim. G. M., Heo. J. N., Jo, S. R., Kang. M. G. (2021). Setting criteria of suitable site for southern-type garlic using non-linear regression model. *Journal of the Korean Society of Agriculture and Meteorology*, 23(4), 366-373.

Lee, K. K., Ko, K. K., Lee, J. W. (2012). Correlation analysis between meteorological factors and crop products, *Journal of the Environmental Sciences*, 21(4), 461-470.

Lim. J. T., Lee. B. W., Yoon. J. I., Sin. J. C., Sin. J. H., Lee. C. G., Moon. G. H., Lee. K. O. (2009). *Theory and practice of crop growth modeling*. (2009). Gyeongjin. 1-364.

Na, M. H., Park, Y. H., Cho, W. H. (2017). A study on optimal environmental factors of tomato using smart farm data, *Journal of the Korean Data & Information Science Society*, 28(6), 1216-1226

Oh, S., Kim, M. S. (2017). Predicting onion production by weather and spatial time series model, *Journal of the Korean Data Analysis Society*, 19(5), 2447-2456.

Park. J. E., Jeong. S. J., Park. S. H., Shin. S. E., Park. S. H., Moon. T. Y., Lee. J. W. (2020). *A comparison of growth curve models for forensically important flies using sigmoid functions. Korean J Leg Med*, 44(2), 84-91.

Proceedings of the Korean Data Analysis Society
January 26-27, 2023, pp. 147-147

한국 노인의 건강정보이해력 영향요인 분석

김은경[1]

요 약

본 연구는 2021년 전국적으로 수행된 지역사회건강조사(Community Health Survey, CHS) 자료를 활용하여 앤더슨 모형을 기반으로 노인의 건강정보 이해력 실태를 파악하고, 이에 영향을 미치는 요인을 파악하기 위한 이차자료 분석연구이다. 65세 이상 노인인구 74,492명을 대상으로 자료 분석은 IBM SPSS statistics SPSS 21.0 프로그램을 이용하여 복합표본설계(complex sampling design)를 고려하여 분석하였고, 모집단 추정을 위해 개인 가중치를 적용하였다. 연구대상자의 일반적 특성에 따른 건강정보 이해력의 차이는 t-test로 통계적 검정을 하였고 상관관계는 피어슨 상관계수를 이용하여 분석하였다. 건강정보 이해력에 영향을 미치는 요인은 다중회귀분석을 이용하여 분석하였다. 앤더슨 모델을 노인의 건강정보 이해력에 적용한 결과, 노인의 건강정보이해 대한 영향요인은 연령, 성별, 거주지역, 주관적 건강상태, 우울, 만성질환 진단 유무였다. 즉 노인의 건강정보이해에는 다양한 요인들이 서로 복잡하게 상호작용하고 있음을 확인하였다. 연구결과를 바탕으로 노인의 건강정보 이해력과 복지향상을 위한 공공의료프로그램 구축이 필요하다.

주요용어 : 건강정보이해력, 노인, 앤더슨 모형.

[1](교신저자) 54150 전북 군산시 대학로 558, 군산대학교 E-mail: egkim@kunsan.ac.kr

Proceedings of the Korean Data Analysis Society
January 26-27, 2023, pp. 149-152

지역사회 노인의 여가생활

박정혜[1]

요 약

노년기는 다른 연령층에 비해 여가시간이 늘어나고 일상생활의 상당한 부분을 차지하기 때문에 여가 활동과 이에 대한 만족은 노년기 행복감이나 성공적 노화와 관련된 중요한 요소로 밝혀지고 있다. 또한 노년기 1인 가구와 부부가구는 급격히 증가하고 있다. 그러므로 본 연구는 지역사회에 거주하는 65세 이상 노인 중 1인 가구와 부부가구의 여가 활동 유형과 이에 대한 만족도를 확인함으로써 지역사회 노인의 행복한 노년을 위한 사회 인프라 구축에 기여하고자 한다. 이를 위하여 2020년 한국보건사회연구원의 노인실태조사자료를 분석하였다. 본 연구대상자는 1인 가구와 부부가구 형태로 거주하는 65세 이상 노인 8188명이다. 자료분석은 빈도와 백분율, 평균, 표준편차, 교차검증, 독립표본 T 검증, Kruskal-Wallis Test를 이용하였다. 그 결과, 여가생활에 대한 만족은 1인 가구 노인보다 부부가구 노인이 통계적으로 유의하게 높았고($t=-6.99$, $p<0.001$), 가구형태별 여가 활동 유형은 통계적으로 유의미한 차이가 있었다($\chi2=103.04$, $p<0.001$). 1인 가구 노인은 동적 오락활동(28.8%)이 가장 많았고, 그다음이 여가활동을 하지 않음(21.9%), 정적 오락활동(21.7%)이었다. 부부가구 노인도 동적 오락활동(34.8%)이 가장 많았고, 그다음은 정적 오락활동(22.8%), 여가활동을 하지 않음(19.3%)이었다. TV/라디오를 시청하는 시간은 1인 가구 노인이 통계적으로 유의하게 길었다($t=10.31$, $p<0.001$). 1인 가구 노인에서 여가생활 만족도가 가장 높은 여가활동 유형은 사회활동이었고, 그다음은 정적 오락활동이었다. 반면, 부부가구 노인은 정적 오락활동이 가장 높은 순위였고, 그다음이 동적 오락활동, 사회활동 순이었다. 본 연구결과는 1인 가구노인과 부부가구 노인의 여가생활 만족도와 여가활동 유형의 차이를 보여준다. 두 가구 형태 모두에서 정적 오락활동이 여가생활 만족도를 높일 수 있으나 1인 가구 노인은 소외나 외로움으로 부부가구 노인보다 사회활동이 더 필요할 수 있다. 이러한 결과는 지역사회에 1인 가구 노인의 사회적 교류를 위한 인프라가 필요함을 보여준다.

주요용어 : 노인, 여가, 여가 활동, 여가 만족도.

1. 연구의 필요성과 목적

여가생활에 대한 만족은 다양한 연령층에서 삶의 질과 행복감을 높이는 중요한 요소로 밝혀지고 있다. 그러나 은퇴 전이나 자녀가 어릴 때는 일과 육아로 여가생활을 제대로 누리지 못하는 경우가 많다. 반면, 65세 이상 노인은 은퇴와 함께 사회적 지위와 역할에 변화가 생기고, 하루 평균 5~6시간의 여가시간이 생긴다. 이것은 하루의 20~25%를 차지하는 많은 시간이지만 노인들은 이러한 시간을 제대로 활용하지 못하는 경우가 많다. 또한, 평균 수명이 길어지고 가족형태가 변화함에 따라 노인 1인 가구나 부부가구도 증가하고 있다. 그러므로, 본 연구는 지역사회에 거주하는 65세

[1]52725 경남 진주시 동진로 33, 경상국립대학교 간호학과 부교수. E-mail: masternur@gnu.ac.kr

이상 노인의 가구형태별 여가활동 유형과 여가생활에 대한 만족도를 확인하고자 한다.
1) 가구형태에 따른 여가활동 유형과 여가생활 만족도의 차이를 확인한다.
2) 가구형태별 여가활동 유형에 따른 여가생활 만족도의 차이를 확인한다.

2. 연구방법

2.1 연구대상자

본 연구의 대상자는 2020년 한국보건사회연구원의 노인실태조사 참여자 중 설문에 직접 응답한 1인 가구와 부부가구 노인 8188명이다.

2.2 연구변수

2.2.1 여가활동 유형

여가활동 유형은 없음, 정적 취미활동, 동적 취미활동, 종교활동, 사회활동으로 분류하였다. 정적 취미활동은 앉아서 하거나 움직임이 적은 활동으로, 영화, 전시회, 공연 등을 관람하거나 악기연주, 노래, 미술, 독서, 바둑, 장기, 화초가꾸기, 화투, 교양강좌 듣기, 자격취득이나 어학 공부 등이다. 동적 취미활동은 외부 활동을 하는 유형으로 축구, 야구 등 스포츠에 참여하거나 등산, 걷기 등이다. 종교활동은 교회나 절 등에서 종교와 관련된 활동이고, 사회활동은 단체나 집단에서 다른 사람과 교류하는 활동이다.

2.2.2 여가생활에 대한 만족도

여가생활에 대한 만족도는 사회・여가・문화 활동에 대해 어느 정도 만족하십니까?라는 단일 문항에 1점(전혀 만족하지 않음)부터 5점(매우 만족)까지 응답한 점수이다.

2.3 자료수집방법

본 연구를 위한 자료는 한국보건사회연구원에서 실시한 노인실태조사를 통해 수집되었고, 연구자는 이 자료를 MDIS(MicroData Intergrated Service) 홈페이지에서 내려받아 분석하였다.

2.4 윤리적 고려

본 연구에서 사용한 자료는 MDIS 홈페이지에 공개된 국가통계자료로, 필요한 사람은 누구나 자료를 내려받을 수 있다. 자료는 응답자를 특정할 수 있는 개인정보가 없어 대상자 식별이 불가능하다.

2.5 자료분석방법

본 연구의 자료는 IBM SPSS Statistics Version 23.0 프로그램을 이용하여 분석하였다.

3. 연구결과

3.1 대상자의 일반적 특성

1인 가구 노인 3117명 중 60.1%(1873명)가 여성, 39.9%(1244명)는 남성이었다. 이들의 연령은 평균 75.4(±6.84)세이었고, 최종학력은 초등교육이하가 59.9%, 종교가 있는 경우가 61.7%, 독립적인 일상생활을 하는 경우는 87.8%, 집을 소유한 경우는 66.6%이었다. 거주하는 집의 형태는 아파트가 가장 많았고, 가구의 연평균 수입은 1451만원, 주관적 건강상태는 5점 만점에 3.1(±0.89)점이었다. 부부가구 노인 5071명 중 60.8%(3085명)가 여성이었고, 평균 연령 72.2(±5.72)세, 최종학력은 중등교육(58.2%)이 가장 많았다. 종교가 있는 경우가 57.1%였고, 일상생활 독립은 91.6%, 집을 소유하고 있는 대상자는 87.6%였고, 아파트에 거주하는 대상자가 가장 많았다(47.5%). 연평균 가구수입은 약 2783만원, 주관적 건강상태는 3.5(±0.84)점이었다.

Table 1. Participants' characteristics

Variables		One-person household (n=3117)	Living with spouse only (n=5071)
		n(%) or Mean±SD	
Sex	Male	1244(39.9)	1986(39.2)
	Female	1873(60.1)	3085(60.8)
Age(years)		75.4±6.84	72.2±5.72
Educational background	≤Elementary	1867(59.9)	1755(34.6)
	Secondary	1189(38.1)	2950(58.2)
	≥Higher	61(2.0)	366(7.2)
Religion	No	1193(38.3)	2177(42.9)
	Yes	1924(61.7)	2894(57.1)
ADL	Dependent	2736(87.8)	4644(91.6)
	Independent	381(12.2)	427(8.4)
Housing ownership	Own house	2077(66.6)	4441(87.6)
	Others	1040(33.4)	630(12.4)
Type of housing	Apartment	1342(43.1)	2410(47.5)
	Detached house	1326(42.5)	2046(40.3)
	Others	449(14.4)	516(12.2)
Household income (yearly, ten thousand won)		1451.8±2444.43	2783.3±3257.71
Perceived health state(Score: 1-5)		3.1±0.89	3.5±0.84

3.2 가구형태에 따른 여가 활동 유형과 여가생활 만족도의 차이

1인 가구와 부부가구 노인의 여가 활동 유형은 통계적으로 유의미한 차이가 있었으며(χ^2=103.04, p<0.001), 1인 가구 노인은 동적 오락활동(28.8%)이 가장 많았고, 그다음이 여가활동을 하지 않거나(21.9%) 정적 오락활동이었다(21.7%). 부부가구 노인도 동적 오락활동(34.8%)이 가장 많았고, 그다음은 정적 오락활동(22.8%), 여가활동을 하지 않음(19.3%)이었다. TV/라디오를 시청하는 시간은 1인 가구 노인이 통계적으로 유의하게 길었고(t=10.31, p<0.001), 여가생활에 대한 만족은 부부가구 노인이 통계적으로 유의하게 높았다(t=-6.99, p<0.001).

Table 2. Differences in leisure activity types according to household types

Variables		One-person household (n=3117)	Living with spouse only (n=5071)	χ2/t(p)
		n(%) or Mean±SD		
Type of leisure	Nothing	683(21.9)	979(19.3)	103.04(<0.001)
	Static entertainment	677(21.7)	1158(22.8)	
	Dynamic entertainment	899(28.8)	1767(34.8)	
	Religious activity	493(15.8)	517(10.2)	
	Social activity	365(11.7)	650(12.8)	
Watching television/radio(hours/day)		4.4±2.20	3.9±1.88	10.31(<0.001)
Leisure satisfaction(Score: 1-5)		3.3±0.92	3.4±0.84	-6.99(<0.001)

3.3 가구형태별 여가 활동 유형에 따른 여가생활 만족도의 차이

1인 가구와 부부가구 노인은 여가활동 유형에 따라 여가생활 만족도에 통계적으로 유의미한 차이를 나타냈다(H=96.7, p<0.001; H=91.0, p<0.001). 1인 가구 노인에서 여가생활 만족도가 가장 높은 여가활동 유형은 사회활동이었고, 그다음이 정적 오락활동이었다. 반면, 부부가구 노인은 정적 오락활동이 가장 높은 순위였고, 그다음이 동적 오락활동, 사회활동 순이었다.

Table 3. Differences in leisure activity types and satisfaction according to household types

Type of leisure	Leisure satisfaction					
	One-person household (n=3117)			Living with spouse only (n=5071)		
	Mean±SD	Average ranking	H†(p)	Mean±SD	Average ranking	H†(p)
Nothing	3.1±1.00	1381.9	96.7 (<0.001)	3.2±0.96	2217.4	91.0 (<0.001)
Static entertainment	3.4±0.90	1695.1		3.5±0.76	2672.1	
Dynamic entertainment	3.3±0.87	1618.1		3.5±0.79	2667.1	
Religious activity	3.1±0.90	1363.6		3.3±0.85	2347.2	
Social activity	3.5±0.83	1756.3		3.4±0.81	2567.2	

†Kruskal-Wallis test

Proceedings of the Korean Data Analysis Society
January 26-27, 2023, pp. 153-156

노인의 우울 위험

박정혜[1]

요 약

65세 이상 1인 가구 노인은 2020년 현재 35.1%에 이르며 더 늘어날 전망이다. 또한 1인 가구 노인은 배우자나 자녀와 함께 생활하는 노인에 비해 사회적으로 고립되어 여러 가지 어려움을 경험하고, 우울증 발생 위험도 증가하는 것으로 보고되고 있다. 그러므로 본 연구는 1인 가구 노인의 여가활동이 우울에 미치는 영향을 확인하고자 한다. 이를 위하여 2020년 한국보건사회연구원의 노인실태조사자료 중 1인 가구 노인 3117명의 자료를 분석하였다. 자료분석은 빈도와 백분율, 평균, 표준편차, 교차검증, 독립표본 T 검증, Pearson 상관관계분석, 이분형 로지스틱 회귀분석을 이용하였다. 그 결과, 여가활동을 하지 않는 대상자에 비해 정적 오락활동을 하는 대상자의 우울 위험이 통계적으로 유의하게 감소하는 나타났다(OR=0.679, p=0.026). 다른 여가활동 유형은 우울에 통계적으로 유의한 영향을 나타내지 않았다. 한편, 여가활동 만족은 우울 위험을 통계적으로 유의하게 감소시키는 것으로 나타났고(OR=0.765, p=0.000), TV/라디오 사용시간은 우울 위험에 통계적으로 유의한 영향을 나타내지 않았다(OR=1.061, p=0.132). 그 외, 초기노년층(OR=1.458, p=0.040), 일상생활 활동이 의존적일 때(OR=1.780, p<0.001), 집을 소유하지 않았을 때(OR=1.348, p=0.005), 만성질환 수가 많을수록(OR=2.232, p<0.001), 주관적으로 건강하지 않다고 인지할수록(OR=0.892, p=0.023) 우울 위험이 통계적으로 높은 것으로 나타났다. 이러한 결과는 1인 가구 노인의 우울 감소를 위한 사회적 관심과 노년기에 적절한 여가활동을 위한 인프라가 필요함을 보여준다.

주요용어 : 노인, 여가, 우울 위험.

1. 연구의 필요성과 목적

65세 이상 1인 가구 노인은 2020년 현재 35.1%에 이르며 더 늘어날 전망이다(Statistic Korea, 2021). 또한 1인 가구 노인의 경우 배우자나 자녀와 함께 생활하는 노인에 비해 사회적으로 고립되어 여러 가지 어려움을 경험하고, 우울증 발생 위험도 증가하는 것으로 보고되고 있다(Kang, Kim, Kim, 2017). 한편, 여가활동과 이에 대한 만족은 노년기 삶의 만족이나 행복감을 증가시키는 것으로 보고되고 있다(Park, Oh, Cho, 2021). 그러므로 본 연구는 1인 가구 노인의 여가활동이 우울에 미치는 영향을 확인하고자 한다.

1) 대상자의 일반적 특성, 여가활동에 따른 우울의 차이를 확인한다.
2) 여가활동과 우울의 관계를 확인한다.
3) 여가활동이 우울에 미치는 영향을 확인한다.

[1]52725 경남 진주시 동진로 33, 경상국립대학교 간호학과 부교수. E-mail: masternur@gnu.ac.kr

2. 연구방법

2.1 연구대상자

본 연구의 대상자는 2020년 한국보건사회연구원의 노인실태조사 참여자 중 설문에 직접 응답한 1인 가구 노인 3117명이다.

2.2 연구변수

2.2.1 우울 위험

우울감은 단축형 노인 우울척도(Korean version of the short form of Geriatric Depression Scale, SGDS-K)를 사용하여 측정되었다. 이 도구는 노인의 우울증 선별검사를 위해 전세계에서 널리 사용되고 있는 Geriatric Depression Scale(GDS)의 한국어판 축약형 척도로 15문항, 이원 척도(binary options)로 구성되어 있고, 최적 절단점(cut-off)은 8점(Cho et al., 1999; Jung et al., 2017)이다.

2.2.2 여가활동

여가활동 변수는 선행연구고찰을 통해 여가활동 유형, 하루 평균 TV/라디오 사용시간, 여가활동에 대한 만족을 포함하였다. 여가활동 유형은 문화예술, 스포츠, 동적 오락활동, 정적 오락활동, 종교활동, 사회활동으로 분류되었다.

2.3 자료수집방법

본 연구를 위한 자료는 한국보건사회연구원에서 실시한 노인실태조사를 통해 수집되었고, 연구자는 이 자료를 MDIS(MicroData Intergrated Service) 홈페이지에서 내려받아 분석하였다.

2.4 윤리적 고려

본 연구에서 사용한 자료는 MDIS 홈페이지에 공개된 국가통계자료로, 필요한 사람은 누구나 자료를 내려받을 수 있다. 자료는 응답자를 특정할 수 있는 개인정보가 없어 대상자 식별이 불가능하다.

2.5 자료분석방법

본 연구의 자료는 IBM SPSS Statistics Version 23.0 프로그램을 이용하여 분석하였다.

3. 연구결과

3.1 대상자의 일반적 특성

연구대상자의 성별은 남자가 35.9%, 여자는 64.1%였다. 연령은 60세 이상이 69.2%, 교육 수준은 중등교육까지 받은 대상자가 50.0%로 가장 많았다. 직업은 주부가 42.9%, 무직이 30.1%였고, 가구

소득은 1분위가 35.9%로 가장 많았다. 병원 입원 기간은 평균 17.4(±19.49)일이었다. 대상자의 일반적 특성에 따른 병원 입원환자 만족도는 통계적으로 유의한 차이를 나타내지 않았다.

Table 1. Participants' characteristics

Category		Depressed (N=563)	Non-depressed (N=2554)	t/χ2	p
		N (%) or Mean (SD)			
Participants' characteristics					
Age (years)	65~74	223 (39.6)	1264 (49.5)	20.235	0.000
	75~84	267 (47.4)	1056 (41.3)		
	≥85	73 (13.0)	234 (9.2)		
Sex	Male	227 (40.3)	1017 (39.8)	0.048	0.826
	Female	336 (59.7)	1537 (60.2)		
Educational background	Elementary education	403 (71.6)	1464 (57.3)	39.428	0.000
	Secondary education	154 (27.4)	1035 (40.5)		
	Higher education	6 (1.1)	55 (2.2)		
Religion	Yes	373 (66.3)	1551 (60.7)	5.958	0.015
	No	190 (33.7)	1003 (39.3)		
ADL	Dependent	156 (27.7)	225 (8.8)	153.569	0.000
	Independent	407 (72.3)	2329 (91.2)		
Perceived health status		2.51(0.879)	3.29(0.825)	-19.132	0.000
Number of chronic diseases		3.00(1.938)	1.90(1.366)	12.768	0.000
Personal annual income*		1011.43(1868.859)	1464.76(2489.511)	-4.880	0.000
Housing ownership	Own house	315 (56.0)	1762 (69.0)	35.279	0.000
	Others	248 (44.0)	792 (31.0)		
Leisure activity					
Type	Nothing	159 (28.2)	524 (20.5)	34.695	0.000
	Cultural activity	14 (2.5)	116 (4.5)		
	Sports activity	13 (2.3)	89 (3.5)		
	Dynamic entertainment	114 (20.2)	579 (22.7)		
	Static entertainment/rest	98 (17.4)	553 (21.7)		
	Religious activity	112 (19.9)	381 (14.9)		
	Social activity	53 (9.4)	312 (12.2)		
Watching TV/Listening to radio(hrs/day)		1.58 (1.351)	1.43 (1.296)	2.297	0.022
Leisure satisfaction		2.84 (1.034)	3.37 (0.865)	-11.370	0.000

*Korean won, tens of thousands, ADL=activities of daily life

3.2 여가활동이 우울에 미치는 영향

이분형 로지스틱 회귀분석결과, 여가활동을 하지 않는 대상자에 비해 정적 오락활동을 하는 대상자의 우울 위험이 통계적으로 유의하게 것으로 감소하는 나타났다(OR=0.679, p=0.026). 그러나

다른 여가활동 유형은 우울에 통계적으로 유의한 영향을 나타내지 않았다. 한편, 여가활동 만족은 우울 위험을 통계적으로 유의하게 감소시키는 것으로 나타났고(OR=0.765, p=0.000), TV/라디오 사용시간은 우울 위험에 통계적으로 유의한 영향을 나타내지 않았다(OR=1.061, p=0.132). 그 외, 초기 노년층(65~74세) (OR=1.458, p=0.040), 일상생활활동이 의존적일 때(OR=1.780, p=0.000), 집을 소유하지 않았을 때(OR=1.348, p=0.005), 만성질환수가 많을수록(OR=2.232, p=0.000), 주관적으로 건강하지 않다고 인지할수록(OR=0.892, p=0.023) 우울 위험이 통계적으로 높은 것으로 나타났다.

Table 2. Effects of leisure activities on depression

Variables		B	p	OR	95% CI
Age (ref.=≥85)	65~74	0.377	0.040	1.458	1.018-2.088
	75~84	0.097	0.565	1.102	0.792-1.531
Education (ref.=elemenatry)	Secondary	-0.208	0.093	0.812	0.637-1.035
	Higher	-0.135	0.779	0.874	0.341-2.238
Religion(ref.=no)		0.178	0.182	1.194	0.920-1.550
ADL(ref.=independent)		0.576	0.000	1.780	1.347-2.352
Housing ownership(ref.=owner)		0.299	0.005	1.348	1.094-1.661
Personal income		0.000	0.214	1.000	1.000-1.000
Perceived health status		-0.115	0.023	0.892	0.808-0.985
Number of chronic disease		0.803	0.000	2.232	1.951-2.553
Leisure type (ref.=nothing)	Cultural activity	-0.509	0.128	0.601	0.312-1.157
	Sports activity	-0.150	0.665	0.861	0.437-1.696
	Dynamic entertainment	-0.228	0.176	0.796	0.572-1.108
	Static entertainment/rest	-0.387	0.026	0.679	0.483-0.955
	Religious activity	-0.170	0.363	0.843	0.584-1.218
	Social activity	-0.206	0.301	0.814	0.551-1.202
Watching TV		0.059	0.132	1.061	0.982-1.145
Leisure satisfaction		-0.267	0.000	0.765	0.680-0.861

ADL=activity of daily life

References

Kang, E., Kim, H., Kim, Y. (2017). Leisure activity types and depressive symptoms among middle-aged people living alone, *Health and Social Welfare Review*, 37(2), 184-215. DOI: https://doi.org/10.15709/hswr.2017.37.2.184

Park, J. H., Oh, Y., Cho, H. C. (2021). Factors related to subjective happiness of women in later life, *Journal of the Korea Academia-Industrial cooperation Society*, 22(9), 190-199. DOI: https://doi.org/10.5762/KAIS.2021.22.9.190

Statistic Korea (2021). 2021 Statistic on the aged, https://kostat.go.kr

Proceedings of the Korean Data Analysis Society
January 26-27, 2023, pp. 157-160

여대생의 신체감시성이 섭식절제행동에 미치는 영향: 신체수치심의 매개효과

김문정[1], 오은정[2]

요 약

여대생은 성인기 초기에 해당하는 집단으로 남성이나 다른 연령대에 비해 객체화 신체의식(신체감시성과 신체수치심)이 높고 외모관리행동을 많이 한다. 외모관리행동에는 다양한 꾸밈행동과 변형행동이 포함되며 그 가운데 섭식절제행동은 여성 청년의 건강에 지장을 초래하게 된다. 본 연구는 신체의식이 섭식절제행동에 영향을 미치는 심리적 기전을 파악하기 위하여 수행되었다. 2개 대학교 간호학과에 재학 중인 1~4학년 여대생 195명을 대상으로 온라인 설문조사를 실시하였다. 연구변수는 한국판 객체화 신체의식 척도(K-OBCS)의 하위척도인 '신체감시성'과 '신체수치심', 섭식행동질문지의 하위척도인 '섭식절제행동'으로 측정하였다. SPSS ver. 27.0 Program에서 Process macro ver. 4.2의 4번 모델을 적용하여 신체감시성과 섭식절제행동의 관계에서 신체수치심의 매개효과를 검정하였다. 연구결과 신체수치심은 신체감시성과 섭식절제행동의 관계를 완전 매개하였으며, 부트스트랩을 통한 간접효과 검정 결과는 유의하였다. 여성 청소년이 건강한 섭식행동을 통해 건강을 유지·증진하도록 돕기 위해서는 신체의식의 긍정적인 변화를 꾀하는 접근법도 필요하다고 여겨진다.

주요용어 : 매개효과, 신체감시성, 신체수치심, 섭식절제행동.

1. 서론

일반적으로 원하는 것보다 더 적게 먹으려는 개인의 경향성을 의미하는 섭식절제행동은 중요한 발달과업을 성취해야 하는 청년기에 영양 섭취의 부족, 신체적·심리적 손상, 대인관계에 부정적인 영향을 미칠 수 있다. 신체를 전인격적 주체의 일부로 보지 않고 관찰하고 이용할 객체로 간주하는 것을 신체의 객체화라 하는데, 객체화 신체의식은 외모관리행동 의도를 예측하는 것으로 나타나(Gim et al., 2007) 객체화 신체의식이 높을수록 날씬한 몸매를 위한 섭식절제행동도 많이 하리라 예상된다. 객체화 신체의식의 하위요인으로 신체 감시성과 신체 수치심이 있다. 신체감시성은 자신과 자기 신체의 관계를 외부관찰자와 객체로 형성한 채, 자기 신체를 문화적 기준에 맞추고 타인의 부정적 평가를 피하기 위하여 마치 감시하듯 바라보는 것이다. 신체수치심은 사회가 요구하는 이상적인 신체를 기준으로 자신의 신체가 이에 미치지 못하면 수치심을 느끼는 것이다(Gim et al., 2007). 신체감시성과 신체수치심 모두 다이어트 행동과 연관이 있었으나(Gim et al., 2007; Grabe, Hyde, 2009) 신체감시 수준이 높을 때 신체수치심이 유발되므로(Kim, Seo, 2011) 신체감시성이 선행변수임을 예측할 수 있다. 즉 신체감시성이 섭식절제행동에 직접 영향을 미치거나 신체수치심을 매개로 간접적인 영향을 미치는 경로를 추정할 수 있다. 따라서 본 연구에서는 여대생을 대상으로

[1](교신저자) 48513 부산광역시 남구 용소로 45, 부경대학교 간호학과 교수. E-mail: mjy7955@pknu.ac.kr
[2]44965 울산광역시 울주군 웅촌면 대학길 9, 춘해보건대학교 간호학과 조교수. E-mail: eunjung70@ch.ac.kr

신체감시성과 신체수치심이 섭식절제행동에 미치는 영향을 확인하고, 신체감시성과 섭식절제행동의 관계에서 신체수치심의 매개효과를 확인하고자 한다.

2. 연구방법

본 연구를 위한 자료수집은 P대학의 연구윤리심의위원회의 승인(1041386-****)을 받은 후 2020년 4~7월까지 이루어졌다. 연구대상자는 B와 U 지역에 소재한 2개 대학교 간호학과에 재학 중인 여대생 가운데 연구참여에 자발적으로 동의한 195명이었다. 포탈 ID 당 한 부의 설문지만 작성할 수 있게 온라인조사를 설정하여 연구대상자들에게 연구 안내문과 함께 링크주소를 전달하였다.

연구변수는 타당도와 신뢰도가 확립된 척도를 사용하여 측정되었다. 신체감시의 정도를 측정하기 위하여 한국판 객체화 신체의식 척도(The Korean Version of Objectified Body Consciousness Scale, K-OBCS)(Gim et al., 2007)의 하위 척도인 '신체감시성(Body Surveillance)'을 사용하였고, 신체수치심의 정도를 측정하기 위하여 K-OBCS의 하위 척도인 '신체수치심(Body shame)'을 사용하였다. 섭식절제행동의 정도를 측정하기 위하여 한글로 번안한 섭식행동질문지(Dutch Eating Behavior Questionnaire; DEBQ)(Kim et al., 1996) 중에서 음식섭취를 얼마나 통제하는지에 관한 섭식절제행동(Restrained Eating) 하위 척도를 사용하였다.

수집된 자료는 SPSS ver. 27.0 Program을 이용하여 분석하였다. 독립변수 간의 다중공선성과 종속변수의 자기상관을 확인하였으며, 회귀모형의 적합성은 잔차분석을 이용한 정규성 분포(Kolmogorov-Smirnov's)와 등분산성을 확인하였다. 대상자의 신체감시성과 섭식절제행동의 관계에서 신체수치심의 매개효과는 Process macro ver. 4.2에서 4번 모델을 적용하여 검증하였고(Hayes, 2018) 매개효과의 통계적 유의성은 부트스트랩(bootstrap)을 실시하여 검증하였다. 부트스트랩 re-sampling 횟수는 5,000회로 설정하였고, 95% bias-corrected bootstrap confidence interval [95% BC bootstrap CI]를 추정하여 검정하였다. 본 연구에 사용된 매개모형은 Figure 1과 같다. a는 신체감시성과 신체수치심의 관계에서 비표준화 회귀계수이다. b는 신체감시성이 통제된 상태에서 신체수치심과 섭식절제행동의 관계에 대한 비표준화 회귀계수이고 c′는 신체수치심이 통제된 상태에서 신체감시성과 섭식절제행동의 관계에 대한 비표준화 회귀계수이다.

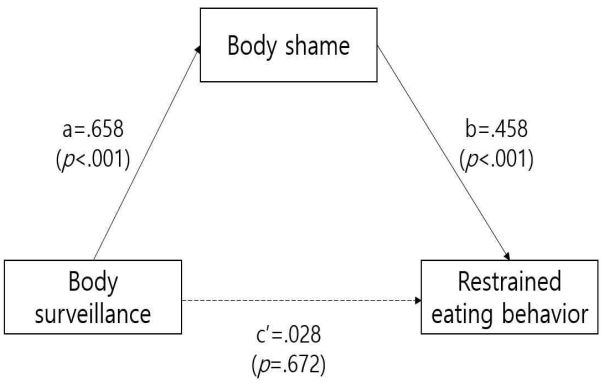


Figure 1. Body shame as mediator of body surveillance to restrained eating behavior

3. 연구결과

신체감시성과 섭식절제행동의 관계를 신체수치심이 매개하는지 알아보기 위하여 Process macro model 4를 사용하여 매개효과를 분석하였다(Table 1). 먼저 신체감시성이 매개변수인 신체수치심에 미치는 영향은 유의하였다(b=0.658, p<.001). 다음으로 신체수치심이 종속변수인 섭식절제행동에 미치는 영향은 유의하였으나(b=0.458, p<.001), 신체수치심이 통제된 상태에서 신체감시성이 섭식절제행동에 미치는 직접효과는 유의하지 않았다(b=0.028, p=.672).

Table 1. Simple mediation effect

Pathway			b	SE	t	p	95% CI	
							LL	UL
Body surveillance	→	Body shame	0.658	0.072	9.095	<.001	0.516	0.801
Body surveillance	→	Restrained eating behavior	0.028	0.065	0.424	.672	-0.101	0.156
Body shame	→	Restrained eating behavior	0.458	0.054	8.426	<.001	0.350	0.565

다음으로 부트스트랩을 실시하여 간접효과의 통계적 유의성을 검증한 결과(Table 2), 95% 신뢰구간 내 0이 포함되지 않아 신체감시성과 섭식절제행동의 관계에서 신체수치심의 간접효과가 통계적으로 유의한 것으로 나타났다.

Table 2. Results of bootstrap (n=5,000)

	Effect	SE	95% CI	
			LL	UL
Bootstrap	0.301	0.049	0.214	0.405

4. 논의 및 결론

본 연구는 간호여대생을 대상으로 객체화 신체의식의 하위요인인 신체감시성과 신체수치심이 섭식절제행동에 어떻게 영향을 미치는지 심리적 기전을 조사하였다. 그 결과 신체감시성은 섭식절제행동에 직접적인 영향을 미치지 않고, 신체수치심을 매개로 간접적인 영향을 미침을 확인하였다. 이는 단순회귀분석을 통해 신체감시성과 신체수치심 모두 이미지추구행동에 영향을 미치고, 신체감시성이 신체수치심을 유발한다고 보고한 Lee(2022)와 유사한 결과이다. 또한 신체감시성과 신체수치심 모두 식이조절다이어트를 포함한 전반적인 꾸밈행동 의도에 영향을 미치나, 다이어트 경험이나 의도만 두고 볼 때, 신체감시성은 차이가 없고, 신체수치심은 차이가 있다고 보고된 바(Gim et al., 2007)와 유사한 결과이다.

신체의식 변수와 외모관리행동의 관계에 대한 근거들이 많이 축적되었으나, 외모관리행동 가운데 건강과 직결된 섭식절제행동은 신체의식에 의해 어떤 영향을 받는지에 대한 과학적 근거가 부족하였는데, 본 연구를 통해 신체수치심이 신체감시성과 섭식절제행동의 관계를 완전매개하고 있음을 확인하였다. 본 연구의 결과는 여성 청년들의 건강을 유지·증진하기 위해서는 건강관리 능력

을 키워주는 노력 이외에 자기 신체를 스스로 의지를 발휘해서 다루는 전인격적 주체의 일부로 바라보며 긍정적인 신체경험을 할 수 있도록 이끌어주는 접근이 필요함을 시사한다. 이를 위해 섭식행동을 설명하는 신체 변인들에 대한 더 많은 탐색이 이루어질 필요가 있다.

References

Gim, W. S., Ryu, Y.J., Park, E. A. (2007). The Korean version of objectified body consciousness scales(K-OBCS): development and validation. The Korean Journal of Psychology, 26(2), 329-349. (in Korean).

Grabe, S., Hyde, J. S. (2009). Body objectification, MTV, and psychological outcomes among female adolescents. Journal of Applied Social Psychology, 39(12), 2840-2858. https://doi.org/10.1111/j.1559-1816.2009.00552.x

Hayes, A. F. (2017). Introduction to mediation, moderation, and conditional process analysis: a regression-based approach. New York, NY, USA: Guilford Publications.

Kim, H. J., Lee, I. S., Kim, J. H. (1996). The study of the reliability and validity of Korean version of the eating behavior questionnaire, Korean Journal of Clinical Psychology, 15(1), 141-150. (in Korean).

Kim, S. Y., Seo, Y. S. (2011). The relations of exposure to sexuall objectifying media to disordered eating, learning flow, and performance anxiety among female high school students, The Korean Journal of Counseling and Psychotherapy, 23(4), 1137-1160. (in Korean).

Lee, H. (2022). A study on image pursuit behavior according to body surveillance, body shame of women using image-based SNS, Journal of Fashion Business, 26(5), 22-35. https://doi.org/10.12940/jfb.2022.26.5.22 (in Korean).

Proceedings of the Korean Data Analysis Society
January 26-27, 2023, pp. 161-161

간호간병통합서비스 병동 간호사의 환자중심간호 영향요인에 대한 연구*

김미래[1], 안정아[2]

요 약

병원의 간호간병통합서비스 병동은 정부 차원에서 환자와 보호자의 간병부담을 해소하고 환자중심 의료서비스의 질 향상을 위해 추진된 제도로써, 현재 국내 병원별 확대 시행을 추진하고 있다. 본 연구는 간호간병통합서비스 병동 간호사의 환자중심간호, 임상적 의사결정 능력, 대상자 중심 의사소통 역량 및 간호근무환경의 수준을 파악하고, 이들의 환자중심간호(Patient-Centered Care)에 영향을 미치는 유의한 요인들을 파악하기 위해 시행된 횡단적 및 서술적 조사연구이다. 본 연구의 자료는 경기도 소재 2개 대학병원의 7개 간호간병통합서비스 병동 간호사 150명을 대상으로 2022년 4월부터 5월까지 구조화된 설문지를 이용하여 수집하였다. 수집된 연구 자료는 SPSS/WIN 23.0 프로그램을 이용하여 기술통계, independent t-test, one-way ANOVA, Pearson's correlation coefficient, multiple linear regression을 이용하여 분석하였다. 본 연구의 결과, 간호간병통합서비스 병동 간호사의 환자중심간호는 85점 만점에 평균 60.5점, 임상적 의사결정 능력은 200점 만점에 평균 132.1점, 대상자 중심 의사소통 역량은 114점 만점에 평균 81.9점, 간호근무환경은 116점 만점에 평균 76.8점이었다. 간호사의 환자중심간호는 임상적 의사결정 능력(r=.450, p<.001), 대상자 중심 의사소통 역량(r=.721, p<.001) 및 간호근무환경(r=.267, p=.002)과 모두 유의한 양의 상관관계를 나타냈다. 다중 회귀분석 결과, 간호사의 환자중심간호에 대해 임상적 의사결정 능력, 대상자 중심 의사소통 역량 및 간호근무환경을 포함한 회귀모형은 통계적으로 유의하였고(F=56.440, p<.001), 설명력은 56.0%였다. 간호사의 환자중심간호에 대한 유의한 영향요인은 대상자 중심 의사소통 역량(β=.632, p<.001), 임상적 의사결정 능력(β=.156, p=.016), 간호근무환경(β=.128, p=.031) 순으로 나타났다. 즉, 간호사의 대상자 중심 의사소통 역량이 높을수록, 임상적 의사결정 능력이 높을수록, 또한 간호사가 인식하는 간호근무환경이 좋을수록 환자중심간호 수행의 수준이 높음을 확인할 수 있었다.본 연구에서 도출된 영향요인들을 고려하여, 간호간병통합서비스 병동 간호사의 환자중심간호 수행 향상을 위해서는, 간호사 개인 차원의 대상자 중심 의사소통 역량과 임상적 의사결정 능력을 증진시키고, 병원 조직적 차원에서 간호근무환경을 개선하는 등의 전략이 필요할 것이다. 향후 간호간병통합서비스의 확대 시행 기대에 따라, 운영 취지에 적합한 효율적이고 질적인 전인간호 중재를 증진할 수 있도록 개인, 병원 및 국가적 차원의 관심과 적극적이고 구체적인 전략의 마련이 선행되어야 할 것이다. 이를 통해 궁극적으로 보다 양질의 환자중심간호 제공과 나아가 환자의 건강 결과 향상을 기대할 수 있을 것으로 사료된다.

주요용어 : 간호간병통합서비스, 환자중심간호, 임상적 의사결정, 의사소통, 간호근무환경.

---

*이 논문은 제 1저자 김미래의 석사학위논문의 축약본임.

[1]16499 경기도 수원시 영통구 월드컵로 164 아주대학교병원 간호사, 아주대학교 간호대학 석사과정생.
E-mail: krtt12345@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 26-27, 2023, pp. 163-163

심부전 환자와 의료진 간 질병양상 변화 및 완화 의료에 대한 의사소통 요구도의 차이*

김경화[1], 안정아[2]

요 약

심부전 질환관리 가이드라인에 따르면 주요 목표는 환자에게 최적의 증상 조절과 삶의 질 향상을 초점으로 하며, 이를 위해 의료진은 환자의 질병양상 변화에 따른 치료 방향에 대해 환자의 선호도에 대해 논의하고 의사결정에 적극 참여하도록 도와야 한다. 그러나 현재 임상에서 질병양상 변화, 특히 심부전 완화 의료에 대한 의사소통은 의료진과 환자 간 불편한 상황으로 여겨지며, 치료 방향의 결정에 대해서도 의료진과 환자 간 의사소통이 원활하지 못한 실정이다. 본 연구는 심부전 환자와 의료진 간 질병양상 변화 및 완화 의료에 대한 의사소통 요구도의 차이를 파악하기 위해 시행된 횡단적 및 서술적 조사연구이다. 연구의 자료는 경기도 소재 대학병원 순환기내과에 내원하는 심부전 환자 42명과 간호사 33명을 대상으로 2022년 11월부터 2개월간 구조화된 설문지를 이용하여 수집하였다. 수집된 연구 자료는 SPSS/WIN 23.0 프로그램을 이용하여 descriptive statistics와 independent t-test를 이용하여 분석하였다. 본 연구결과, 심부전 환자의 평균 연령은 66.4세였고 남자가 26명(61.9%)이었으며 심부전 유병기간은 평균 5.8년이었다. 또한 간호사의 심부전 환자 간호 경력은 평균 11.4년이었다. 심부전 환자와 간호사의 질병양상 변화 및 완화 의료에 대한 의사소통 요구도 점수(총점)는 평점 4점 만점에 심부전 환자는 평균 3.27점, 간호사는 평균 3.05점으로 나타났으며, 두 군간 통계적으로 유의한 차이를 보이지는 않았다(t=1.549, p=.126). 심부전 환자와 간호사의 질병양상 변화 및 완화의료에 대한 의사소통 요구도의 하위 영역별로 살펴보면, 특별히 생애 말기에 대한 의사소통 요구도의 경우 심부전 환자는 평균 3.25점, 간호사는 평균 2.81점으로, 심부전 환자의 해당 영역 의사소통 요구도가 간호사에 비해 유의하게 높음을 알 수 있었다(t=2.042, p=.045). 다음으로 심부전 환자와 간호사의 질병양상 변화 및 완화 의료에 대한 의사소통 요구도의 하위 영역별 순위를 살펴보면, 심부전 환자의 경우 1) 심박동기 또는 제세동기에 대한 의사소통, 2) 도움, 지원 및 치료에 대한 의사소통, 3) 심부전 질환이 일상에 미치는 영향에 대한 의사소통, 4) 생애 말기에 대한의사소통, 5) 가족구성원 관련 의사소통 순으로 나타났으며, 간호사의 경우 1) 심부전 질환이 일상에 미치는 영향에 대한 의사소통, 2) 도움, 지원 및 치료에 대한 의사소통, 3) 심박동기 또는제세동기에 대한 의사소통, 4) 가족구성원 관련 의사소통, 5) 생애 말기에 대한 의사소통 순으로 나타나, 두 군간 의사소통 요구도의 우선순위의 차이를 확인할 수 있었다. 심부전 환자와 의료진의 질병양상 변화 및 완화 의료에 대한 의사소통 요구도를 평가 및 비교한 본 연구결과는 향후 대상자의 요구도를 기반으로 한 의사소통 향상 중재 프로그램을 개발하는데 중요한 근거가 될 것으로 사료된다.

주요용어 : 심부전, 질병양상 변화, 완화 의료, 의사소통.

---

*(이 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2022R1F1A1075049).
[1]16499 경기도 수원시 영통구 월드컵로 164 아주대학교 간호대학 석사수료생. E-mail: summerlove@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 26-27, 2023, pp. 165-165

간호대학생을 위한 당뇨병 교육 프로그램: 체계적 문헌고찰*

김경아[1], 안정아[2]

요 약

당뇨병은 전 세계적으로 유병률이 높은 만성 대사질환으로 다양한 합병증을 유발한다. 간호교육은 전공지식과 더불어 임상수행능력을 갖추는 실무중심 교육을 지향하고 있다. 따라서 간호대학생은 학부기간 동안 간호술기를 비롯한 환자 및 의료진과 상호작용하는 의사소통의 기술, 임상 의사결정을 위한 추론 및 판단의 기술 등을 익혀야 한다. 환자들이 점점 더 개인의 사생활 보호와 안전 등의 이유로 실습 학생들에게 기회가 점점 줄어들고 있으며, 이에 대한 대응으로 간호대학에서는 실습의 기회로 디지털 교육을 많이 사용해 오고 있다. 또한 최근 예상하지 못한 Covid-19 팬더믹으로 인해 간호대학생의 실습 교육은 점점 더 온라인 디지털 교육으로 전환되고 있는 실정이다. 따라서 본 연구에서는 간호 학생을 위해 당뇨병 교육 프로그램이 어떻게 설계, 구현 및 평가되었는지에 대한 현재 상황을 알아보기 위한 체계적인 문헌고찰을 수행하였다. CINAHL, PubMed, EMBASE 및 Cochrane Library 데이터베이스를 사용하여 문헌을 검색하였다. 검색 전략은 (1) 간호 학생, (2) 당뇨병 및 (3) 교육 개념을 결합하였다. 2012년에서 2022년 사이에 원문이 영어인 논문, 저널에 발표된 무작위 통제 실험 또는 준실험 연구만을 대상으로 하였다. 체계적 문헌고찰 수행 결과 14편의 연구가 최종 분석대상으로 선정되었다. 14편의 논문은 미국(n=5, 35.7%), 터키(n=3, 21.4%), 캐나다(n=1), 영국(n=1), 인도(n=1), 한국(n=1), 태국(n=1), 브라질(n=1)에서 수행되었다. 14편 중 5편(35.7%)은 무작위대조시험, 나머지 6편(42.8%)은 1군 준실험연구, 나머지 3편(21.4%)은 2군 준실험연구이였다. 당뇨병 교육 프로그램은 독립적인 교육 과정으로 제공되거나 기존 과목에 통합되어 제공되고 있었으며, 당뇨병 교육에서 가장 선호되는 주제는 케톤산증에 대한 지식이었다. 대부분의 연구(n=13, 92.8%)가 기술에 기반한 교수법을 사용했으며, 전통적인 강의 기반 교수법을 사용한 연구는 1건에 불과했다. 구체적으로 4건(28.5%)은 고충실도 시뮬레이션을, 3건(21.4%)은 모바일 애플리케이션을, 1건은 가상현실 교육 방법을 사용하였다. 또한 다른 교육 방법으로 컴퓨터 보조 수업, 온라인 원격 교육, 가상 사례 시나리오 등을 활용하였다. 당뇨병 교육 프로그램에 대한 평가에서는 기술 기반의 양방향 교육 프로그램이 기존의 강의식과 비교하여 간호대학생의 지식뿐만 아니라 임상수행, 자신감, 자기효능감에서 유의미하게 향상시키는 것으로 나타났으며, 기술 기반 교육 프로그램을 경험한 간호대학생들은 학습 방법에 대한 만족도가 유의미하게 높아졌다고 보고하였다. 이러한 결과는 간호대학생의 지식과 임상 기술을 더 잘 향상시키기 위해 새로운 기술 기반 교육 프로그램의 중요성을 보여준다고 할 수 있다. 추후 간호대학생을 대상으로 한 기술 기반 교육 프로그램의 장기적 교육적 영향, 비용 효율성 등 다양한 교육적 측면을 고려한 연구가 필요하다고 하겠다.

주요용어 : 간호대학생, 당뇨병, 교육 프로그램.

---

*이 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단 생애 첫 연구사업의 지원을 받아 수행된 연구임 (NRF001664431G0003081).

[1]21936 인천광역시 연수구 함박뫼로 191 가천대학교 간호대학 간호학과 조교수. E-mail: kakim7213@gachon.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 26-27, 2023, pp. 167-167

임상간호사 전향적 코호트 연구동향 분석: 텍스트 네트워크 및 토픽모델링

이은교[1], 박성희[2], 김광현[3], 이영진[4]

요 약

본 연구의 목적은 임상간호사 대상의 전향적 코호트 연구에 대한 체계적 문헌고찰을 수행하고, 초록 기반의 텍스트 네트워크 및 토픽 모델링 분석을 통해 임상간호사 대상의 전향적 코호트 연구의 동향을 파악하고자 함이다. 본 연구는 체계적 문헌고찰을 수행하여 PubMed, Cochrane, Embase, CINAHL의 데이터베이스에서 임상간호사 대상 전향적 코호트 연구를 검색하였다. 검색된 연구 중 최종 분석대상으로 21개의 연구를 선정하여, 해당 논문의 영문 초록을 추출하고, R을 활용하여 텍스트 전처리를 실행하였다. 이후 해당 논문에 대한 핵심 주제어를 추출하고 텍스트 네트워크를 생성하였고, 네트워크 내 주제어의 중심성 분석 및 핵심 주제어를 선정하였다. 최종적으로 1994년부터 2022년까지 출판된 논문을 대상으로 토픽 모델링을 실시하여 전체적인 연구 토픽을 확인하였다. 추가적으로 논문의 출판 시기를 기준으로 2010년 이전과 이후의 논문을 비교하여 연구가 수행된 시기에 따른 연구주제의 변화를 파악해 보고자 하였다. 본 연구결과, 최종 분석대상 연구는 3개 연구주제로 분류되었다. 즉, 심리사회적 건강요인 관련 연구, 신체적 건강요인 관련 연구, 그리고 직무결과 관련 연구로 구분되었다. 연구가 수행된 시기에 따른 연구주제 동향은 2010년 이전은 환자안전, 직무결과, 신체적 건강 관련 연구가 다수 보고되었으며, 2010년 이후는 간호사의 직무결과, 심리사회적 건강요인, 전반적 삶의 질 관련 연구가 주요한 연구동향으로 분석되었다. 본 연구결과는 토대로 임상간호사 대상 전향적 코호트 연구는 2010년 이후 증가되고 있는 추세이며, 간호전문직의 직무에 영향을 미치는 건강 요인을 주제로 진행되고 있음을 파악해 볼 수 있었다. 특히, 임상간호사의 신체적 건강에 대한 주제가 점차 심리사회적 건강 문제와 전반적 삶의 질에 대한 관심으로 폭넓게 확대되고 있는 경향을 확인할 수 있었다. 이러한 연구결과는 추후, 임상간호사 대상의 전향적 코호트 연구에 대한 주제를 설정하고, 연구방법을 설계하기 위한 기초자료로 되며 궁극적으로는 임상간호사의 간호의 질 향상 및 직무결과 개선에 유용한 자료로 활용될 수 있다.

주요용어 : 임상간호사, 전향적 코호트 연구, 텍스트 네트워크, 토픽 모델링.

[1]아주대학교 간호과학연구소 특임연구원.
[2]아주대의료원 간호사, 아주대학교 간호대학 박사수료생.
[3]아주대의료원 간호사, 아주대학교 간호학과 석사과정생.
[4](교신저자) 아주대학교 간호대학·간호과학연구소 부교수.

Proceedings of the Korean Data Analysis Society
January 26-27, 2023, pp. 169-172

Testing for House Price Bubble in Seoul

Sung-Hoon Sim[1]

Abstract

Based on the simple theoretical framework of house pricing, which allows us to understand the background of the explosive behavior as bubbles, this paper examines whether bubbles exist in Seoul housing market. The results of sub ADF (SADF) and generalized ADF tests (GSADF) show that strong statistical evidence of explosive behavior in Seoul housing markets.

Keywords : House price bubble, sub ADF test, generalized ADF test.

1. Introduction

A sharp downfall of house prices during the global financial crisis in 2008 was followed by the drastic upward trend in many countries. Regarding this, many participants in Korean housing market think that the housing prices during the last decade go far from the economic fundamentals, and there has been growing concern about housing price overvaluation and thus a bubble during this period. However, the house prices in Seoul start to decrease along with the movement of interest rate that was record-low (0.50%) in May 2020. In fact, the interest rate has been continuously increasing up to 3.25% as of December 2022 from 0.75% in August in 2022. In this context, it is proper and timely to ask whether the house prices in Seoul contain a bubble. To detect housing bubbles, this paper adopts SADF test and GSADF test. The two ADF-type tests are used to calculate time-varying bubbles. Instead of stressing the fundamental values, these tests detect explosive behaviors in housing market and allow for changes of the starting and end points, as well as changes of the window widths. These features enable the test methods to be effective in detecting multiple bubbles and to perform better than other methods both with experimental data and real data.

2. Methodology

In this paper, we select the house price-to-rent ratio to calculate the degree of exuberance. Initially, we apply the right-tailed unit root test of Phillips, Shi, Yu (2015) to test evidence of explosive behavior in house price. The martingale null hypothesis with an asymptotic drift is specified as equation (1):

[1]Professor, Department of Economics, Pukyong National University, Youngso-ro 45, Nam-gu, Busan, 608-737, Korea. E-mail: shsim@pknu.ac.kr

$$H_0 : y_t = dT^{-\eta} + y_{t-1} + \epsilon_{t,} \quad \epsilon_t \sim iid(0.\sigma^2) \tag{1}$$

where y is the house price-to-rent ratio, d is a constant, T is the sample size, and localizing coefficient η is larger than 0.5. The alternative hypothesis is mildly explosive process as equation (2):

$$H_1 : y_t = \delta_T y_{t-1} + \epsilon_{t,} \tag{2}$$

where $\delta_T = 1 + c\,T^{-\theta}$ with $c > 0$ and $\theta \in (0,1)$.

We insert a transient dynamics to examine the degree of exuberance using rolling-window approach, in which the regression from the sample starts from the r_1^{th} fraction of the total sample (T) and ends the r_2^{th} fraction due to $r_2 = r_1 + r_w$, where r_w is the window size of T. Then, the following empirical model of (3) is estimated:

$$\triangle y_t = \alpha_{r_1, r_2} + \beta_{r_1, r_2}\, y_{t-1} + \sum_{i=1}^{k} \gamma^{i}_{r_1, r_2} \triangle y_{t-i} + \epsilon_t . \tag{3}$$

To investigate a behavior in housing market, we use right-tailed unit root test which is informative about mildly explosive behavior in the data and hence is useful as a form of market warning alert against mispricing. We introduce the concept of unit roots through the use of the null hypothesis: $\hat{\beta}_{r_1, r_2} = 0$, against the alternative hypothesis, namely mildly explosive behavior: $\hat{\beta}_{r_1, r_2} > 0$. Then We are interested in the following test statistics in (4):

$$ADF_{r_1}^{r_2} = \frac{\hat{\beta}_{r_1, r_2}}{se\,(\hat{\beta}_{r_1, r_2})} \tag{4}$$

This is a right-tailed statistic. Phillips et al. (2011) propose a recurse procedure on the estimation of $ADF_{r_2}^{r_1}$, using different subsamples of data since r_0 is the smallest sample window fraction, and this new test statistics is then supremum value of $ADF_0^{r_1}$. This proposed statistics is referred to as the sup ADF (SADF) test using a fixed start point as in (5):

$$SADF(r_0) = sup_{r_2 \in [r_0, 1]} ADF_0^{r_2} \tag{5}$$

The null hypothesis of unit root is rejected in favor of explosive behavior when its value is greater than the critical value. Otherwise, when its value is less than the critical value, the end data of this exuberance is detected.

One concern with the SADF is that it may not consistently detect the origination and termination when multiple episodes of exuberance are present. Phillips, Shi, Yu (2015) propose a generalized SADF (GSADF) to deal with multiple events of boom and bust in a single series as in (6):

$$GSADF(r_0) = sup_{\substack{r_1 \in [0, r_2 - r_0] \\ r_2 \in [r_0, 1]}} ADF_{r_1}^{r_2} \tag{6}$$

In equation (11), if the $GSADF(r_0)$ statistics is larger than the critical value, we reject the

null hypothesis in favor of the explosive alternative hypothesis. To obtain the GSADF test statistics Phillips, Shi, Yu (2015) propose a recursive backward SADF (BSADF) regression operation as in (7)

$$BSADF_{r_2}(r_0) = sup_{r_1 \in [0, r_2 - r_0]} ADF_{r_1}^{r_2} \tag{7}$$

Finally, we lay special emphasis on how to search for periods with explosive behavior, namely, data-stamping strategy. Phillips et al. (2015) define the initiation date of the bubble ($\hat{r_e}$) and termination ($\hat{r_f}$) of the bubble as

$$\hat{r_e} = inf_{r_2 \in [r_0, 1]} r_2 : BSADF_{r_2}(r_0) > scv_{r_2}^{\alpha}$$
$$\hat{r_f} = inf_{r_2 \in [\hat{r_e} + \frac{12}{T}, 1]} r_2 : BSADF_{r_2}(r_0) < scv_{r_2}^{\alpha}$$

where $scv_{r_2}^{\alpha}$ represents $(1-\alpha_T)$ critical value of the BSADF test statistics based on the $[Tr_2]$. We would like notice that the distribution of the SADF and the GSADF test statistics are non-standard, indicating that critical values can be obtained by case by case Monte Carlo simulation.

3. Results

Data from Koomin bank gathers the rent-to-price ratio during 2008:12~2022:10. Following the theoretical method, the data is transformed into price-to-rent ratio. Table 1 reports the empirical results for both the SADF and GSADF tests. We find that the hypothesis of no bubble is rejected for the SADF test for the house price-to-rent ratio at the 10%, 5%, and 1% level, respectively. However, with the limitation of the SADF test as argued in Phillips, Shi, Yu (2015), we also provide empirical results of GSADF test. Our GSADF test statistics are far greater than the critical values at all levels, implying that there are multiple episodes for explosiveness in the house price-to-rent ratio.

Four bubble periods, which are shaded in Fig. 1, appear to exist, based on the criteria that bubble are detected when the values of backward SADF sequence are larger than 95% critical value sequence and based on the criteria for a spurious bubble. The first group of bubble occurs during the period 2002:12~2005:10 and 2006:10~2009:12, which periods correspond to global financial crisis. The second group of bubbles is found during the periods 2010:11~2020:12 and 2021:04~2022:08. Over the periods, Seoul housing market has experienced sharp increases in the house prices mainly due to the huge amount of liquidity in the market stemming from historically low interest rate. Since the beginning of 2022, interest rate has started to rise in accordance with a sharp rise of US interest rate.

4. Conclusion

This paper presents new evidence casting doubt on the bubble hypothesis in Seoul housing market by using the GSADF test. We find evidence of bubble in the Seoul house

price during four separate periods. The first bubble periods are associated with global financial crisis, and the second periods correspond to the era of historically low interest rate in Korea. Regarding our empirical results, it is quite possible that the government might react to the sharp rises in house prices through various administrative means of intervention. These policies could include the tightening lending conditions (e.g., controlling bank mortgage policy and discriminative interest rate policy for investors), limiting housing purchases for multiple house-owner, increasing house supply, and etc., in order to dispel the momentum of rising house prices in Seoul and deter speculations by investors. Among others, the current downward movement of housing prices immediately results from the sharp increases in interest rate. Therefore, during prolonged bubble periods, policymakers should weigh options: first, the negative effect of increasing interest rate on domestic economy, relieving the house price bubbles; second, the negative influence of bubble itself on the economy.

Table 1. The SADF and the GSADF test

	SADF	GSADF
Test statistics	2.1663	6.3307
Finite sample critical values		
90%	1.2071	1.9540
95%	1.4165	2.1025
99%	1.9508	2.6963

Notes: The SADF and GSADF test statistics are estimated over the periods December 1998 to October 2022. The right-sided critical values are obtained from Monte Carlo simulations with 2,000 replications.

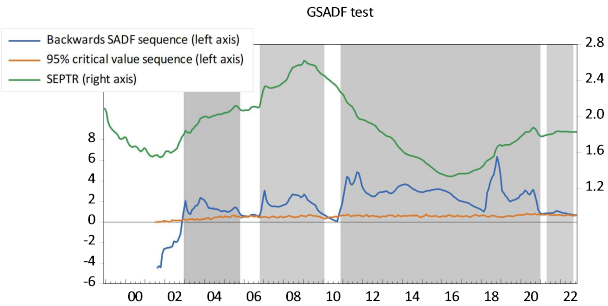


Fig. 1. GSADF for house price-to-rent ratio in Seoul

References

Flood, R.P., Hodrick, R.J. (1990). On testing for speculative bubbles. *The Journal of Economic Perspective*, 4(2), 85-101.

Harvey, D.I., Leybourne, S.J., Sollis, R., Taylor, A.R. (2016). Tests for explosive financial bubbles in the presence of non-stationary volatility. *Journal of Empirical Economics*, 38, 548-574.

Phillip, P.C.B., Shi, S.P., Yu, J. (2015). Testing for multiple bubbles: Historical episodes of exuberance and collapse in the S&P 500. *International Economic Review*, 56(4), 1043-1078.

Phillip, P.C.B., Wu, Y, Yu, J. (2011). Explosive behavior in the 1990s nasdaq: when did exuberance escalate asset values? *International Economic Review*, 52(1), 201-226.

Proceedings of the Korean Data Analysis Society
January 26-27, 2023, pp. 173-173

함수적 선형 모형을 활용한 아파트 매매 가격 지수 예측

김희진[1], 이은지[2]

요 약

우리나라 국민들 자산은 동산보다는 부동산이 더 많은 비중을 차지하고 있다. 이러한 상황 속에서 대표적인 부동산인 아파트 가격은 가구 부채 규모에도 영향을 주며 은행 이자에 따른 가구 잉여 소비력을 결정한다. 아파트 가격은 가구뿐만 아니라 국가 경제에 큰 영향을 미치고 있다. 이와 같이 국가 경제에 영향을 미치는 아파트 매매 가격 지수를 함수적 선형 모형을 활용해 예측해보고자 한다. 함수적자료분석 방법은 연속적으로 관측된 자료가 곡선의 함수적 형태로 나타나는 자료를 분석 대상으로 한다. 본 연구에서는 한국부동산 정보원에서 제공하는 2012년에서 2022년 9월까지의 85개 시 단위 아파트 매매지수를 함수적자료분석 대상으로 보고, 함수적 선형 모형으로 아파트 매매 가격 지수를 예측한다. 함수적 선형 모형을 활용한 아파트 매매 가격지수 예측 정확도를 확인하기 위해서 시계열 분석 방법인 자기회귀이동평균모형(ARIMA) 분석과 머신러닝 방법인 심층신경망을 동일한 조건에서 모형을 설계하여 모형 간 예측력을 RMSE로 비교한다. 연구결과는 전반적인 예측력과 함께, 시장이 안정적인 경우와 외부 요인에 의해서 급변하는 경우로 나누어 예측력을 검정한다.

주요용어 : 함수적선형모형, 아파트 매매 가격 지수, 시계열분석, 머신러닝.

References

Lee, M., Kwon S. (2020). Prediction of Stock Price Index Using Functional Linear Model, *Productivity Review*, 34(3)

Yoon, M., Kim, H., Eum, S. (2020) Comparative Study on the Prediction of the Apartment Sales Price Index, *Journal of the Korea Real Estate Academy*, (56), 5-24

Bae, S., Yu J. (2018) Predicting the Real Estate Price Index Using Machine Learning Methods and Time Series Analysis Model, *Housing Studies Review*, 26(1), 107-133

[1]34134 대전광역시 유성구 대학로 99, 충남대학교 통계학과 박사과정. E-mail: hjkim3036@gmail.com
[2](교신저자) 34134 대전광역시 유성구 대학로 99, 충남대학교 통계학과 교수. E-mail: eunjee.cnu@gmail.com

Proceedings of the Korean Data Analysis Society
January 26-27, 2023, pp. 175-179

기계학습을 이용한 대학출원 특허의 소멸예측

김용우[1], 우상미[2]

요 약

기술혁신의 핵심산출물 중 하나인 특허의 출원은 계속 증가하고 있지만, 대학이 출원한 특허의 사용률은 낮은 편이다. 본 연구에서는 대학에서 출원한 등록특허의 소멸여부를 예측하는 모형을 구축한다. 이를 위해 한국 및 미국의 대학이 출원한 등록특허를 수집하고, 기계학습 기법을 활용하여 특허의 소멸여부를 분류할 수 있는 예측모형을 구축하였다. 또한, 기계학습 모형이 예측하는 대학출원 특허의 소멸여부에 영향을 끼치는 변수들을 국가별로 제시하였다.

주요용어 : 기계학습, 특허, 분류.

1. 서론

특허는 기술혁신의 핵심적인 산출물 중 하나이다. 현대 기업의 경쟁력에 미치는 기술의 영향력이 커지면서(Jaffe, 2000), 특허 출원이 폭발적으로 증가하였다(Shapiro, 2000). 그러나, 모든 특허가 활용되지는 않는다. 특허청의 지식재산활동실태조사에 따르면 2019년 기준 보유 특허 중 기업의 특허 활용도는 77.7%이지만, 대학 및 공공연구기관의 특허 활용도는 22.1% 수준으로 낮게 나타났다. 이는 일반 기업이 아닌 대학의 미사용 특허 대부분은 기술이전, 사업화, 창업 등으로 연결되지 못하고 비교적 짧은 기간에 소멸되는 것을 의미한다. 이에 정부에서는 미활용 특허 문제를 중요한 정책과제로 설정하여 다양한 정책을 펴고 있으나 크게 개선되지 않는 상황이다. 이러한 점을 고려할 때 미활용 특허의 주요 특징을 살펴보고, 활용여부에 차이를 가져오는 요인을 알아보는 것은 의미가 있다. 본 연구에서는 기계학습을 활용하여 대학출원 특허의 미사용 여부를 예측하고 대학출원 특허의 활용 가능성을 높이기 위한 요인들을 식별하고자 한다. 예측에 사용되는 지표들이 다양하고 그 지표들 간 복잡한 관계가 존재하는 데이터의 특성을 고려할 때 기계학습 방식의 적용은 적합한 것으로 볼 수 있다. 기계학습을 통한 모델링은 복잡한 가정이 요구되지 않아 분석이 용이하며, 통제되지 않은 환경에서 얻은 데이터이거나 데이터에 많은 비선형 상호작용이 있는 경우 통계적 모델보다 더욱 효과적이기 때문이다(Bzdok et al., 2018). 본 연구에서는 대학출원 특허의 소멸여부에 영향을 미치는 요인들을 알아보기 위해 유럽특허청의 PATSTAT Database를 이용하여 특허정보를 수집하고 분석을 위한 서지정보를 생성한다. 이후 이진분류를 위한 다수의 기계학습 모델을 생성하고, 성능평가 지표를 통해 모델의 성능을 평가한다. 마지막으로, 최적의 분류모델을 이용하여 소멸 여부에 영향을 미치는 요인들을 도출한다.

[1](교신저자) 04763 서울시 성동구 왕십리로 222, 한양대학교 기술경영학과 박사과정 수료.
E-mail: ywkim@hanyang.ac.kr
[2]03722 서울시 서대문구 연세로 50, 연세대학교 행정학과 박사과정. E-mail: smwoo@yonsei.ac.kr

2. 실험 설계

분석을 위해 PATSTAT Database에서 1991~2014년 기간 동안 미국(US) 및 한국(KR)에 주소지를 둔 대학에서 해당 국가의 특허청에 단독 혹은 공동출원하여 등록된 특허들을 대상으로 연구를 진행하였다. 연차료 미납으로 인해 특허의 권리 만료가 가능한 시점을 고려하여 2015년 이후의 특허는 분석 대상에서 제외되었다. 또한, 한 특허의 출원인에 한국대학과 미국대학이 동시에 있는 경우도 제외하였다. 양국의 대학출원 특허를 비교하기 위해 한국의 대학이 한국 특허청에 출원하여 등록한 특허 76,434개, 미국의 대학이 미국 특허청에 출원하여 등록한 특허 104,722개 총 181,156개의 인스턴스(특허)가 분석에 사용되었다. 이 중에서 연차료를 납부하여 특허의 권리가 계속 유지된 특허와 연차료를 미납하여 최장존속기간 이전에 권리가 소멸된 특허를 구분하여 목적변수로 설정하였다. 소멸여부 예측에 사용된 변수들은 Table 2와 같으며, 특허 자체의 특성, 출원연도의 기술 트렌드 및 출원기관의 역량을 반영하고자 하였다.

Table 1. Patents by country and expiration status

Country	1 (Lapsed)	0 (Not Lapsed)	Total
KR	46,558	29,876	76,434
US	65,984	38,738	104,722
Total	112,542	68,614	181,156

Table 2. Variables

No.	Variable Type	Variable Name	Description
0	Control Variable	APPLN_FILING_YEAR	특허의 출원연도. 연차료 미납으로 인한 특허 권리의 소멸은 오래된 특허일수록 일어나기 쉬우므로 이를 조절할 변수가 필요.
1	Patent Characteristics	NB_INVENTORS	특허의 발명인 수
2	Patent Characteristics	NB_APPLICATNS	특허의 출원인 수
3	Patent Characteristics	DOC_DB_FAMILY_SIZE	특허의 DOCDB 패밀리 사이즈
4	Patent Characteristics	CNT_CTRY_CODE	특허의 출원인 국적 수
5	Patent Characteristics	CNT_CPC	특허의 CPC Symbol 개수
6	Patent Characteristics	CNT_CPC_4DIGIT	특허의 CPC Symbol (앞 4자리) 개수
7	Patent Characteristics	CNT_IPC	특허의 IPC Symbol 개수
8	Patent Characteristics	CNT_IPC_4DIGIT	특허의 IPC Symbol (앞 4자리) 개수
9	Patent Characteristics	CNT_BACK_CIT	특허의 Backward Citation
10	Patent Characteristics	CNT_FORWARD_CIT	특허의 Forward Citation
11	Patent Characteristics	LEN_ABSTRACT	특허의 초록 글자 수
12	Innovation Trend	IPC_AVG_CNT_APPLN_DIST	해당 특허가 속하는 IPC Symbol의 출원연도 출원 수 (여러 IPC에 속하는 경우 평균)

Table 2. Variables (continued)

No	Variable Type	Variable Name	Description
13	Innovation Trend	IPC_AVG_DOCDB_FSIZE	해당 특허가 속하는 IPC Symbol의 출원연도 DOCDB 패밀리 사이즈 (여러 IPC에 속하는 경우 평균)
14	Innovation Trend	IPC_AVG_AVG_NB_INVENTORS	해당 특허가 속하는 IPC Symbol의 출원연도 발명인 수 (여러 IPC에 속하는 경우 평균)
15	Applicant Competency	RE_AVG_CNT_BACK_SIT	해당 특허를 출원한 출원인의 출원연도 평균 Backward Citation (여러 출원인이 출원한 경우 평균)
16	Applicant Competency	RE_AVG_FOWARD_CIT	해당 특허를 출원한 출원인의 출원연도 평균 Forward Citation (여러 출원인이 출원한 경우 평균)
17	Applicant Competency	RE_AVG_AVG_DOCDB_FSIZE	해당 특허를 출원한 출원인의 출원연도 평균 DOCDB 패밀리 사이즈 (여러 출원인이 출원한 경우 평균)
18	Applicant Competency	RE_AVG_CNT_APPLN_DIST	해당 특허를 출원한 출원인의 출원연도 평균 출원 수 (여러 출원인이 출원한 경우 평균)
19	Applicant Competency	RE_AVG_CNT_IPC	해당 특허를 출원한 출원인의 출원연도 평균 IPC Symbol 수 (여러 출원인이 출원한 경우 평균)
20	Applicant Competency	RE_AVG_CNT_IPC_4DIGIT	해당 특허를 출원한 출원인의 출원연도 평균 IPC Symbol (앞 4자리) 수 (여러 출원인이 출원한 경우 평균)

Extra Trees(ET), Generalized Linear Model(GLM), Gradient Boosting Models(GBM), 그리고 Random Forest(RF)를 이용하여 특허의 소멸여부를 예측하는 모델을 구축한다. 학습 및 테스트 데이터는 8:2 비율로 분할하고, 모델 학습 시 모델의 초매개변수(hyperparameter) 조정은 5겹 교차검증(5-fold cross validation)을 통해 진행한다. 이후 생성된 여러 모델을 결합(stacking)하는 super learner(Van der Laan et al., 2007)를 생성한다. 국가 간 결과 비교를 위해 이러한 과정을 한국대학 그리고 미국대학이 출원한 특허를 대상으로 각각 실시한다.

3. 분석결과

한국 특허를 대상으로 한 개별 모델의 성능평가 지표를 살펴보면 AUC(Area Under the ROC Curve) 82.96%~88.54%, AUCPR(Area Under the Precision-Recall Curve) 74.21%~81.33% 수준이다. 개별 모델 중에서는 GLM의 성능이 가장 좋지 못했고 GBM의 성능이 가장 우수했다. 이러한 성능 차이는 복잡한 비선형 모델링이 특허의 권리소멸을 예측하는 데에 있어서 유리할 수 있음을 의미한다. 개별 모델 중에서 종류별로 가장 우수한 4개를 결합한 super learner의 경우 AUC 88.84%, AUCPR 81.85%, MCE(Mean per Class Error) 18.48%로 개별 모델의 성능보다 미세하게 우수한 것으로 나타났다. 그러나 제시된 예측변수를 이용한 개별 모델의 성능이 충분히 우수하므로 다수의 모델을 결합한 성능향상 방법 적용의 필요성은 적은 것으로 판단할 수 있다. 미국 특허를 대상으로 한 모델들 역시 한국특허를 대상으로 분석한 결과와 대체로 유사한 결과를 보였으나, 전체적인 모델의 성능은 한국특허를 대상으로 했을 때보다 낮은 것이 특징이었다.

복잡한 기계학습 모형을 사용할 때의 약점은 예측된 결과에 대한 설명이 부족하다는 점이다. 본 연구에서는 두 국가에 대해서 공통적으로 성능이 우수했던 GBM 모델을 대상으로 순열 중요도(Altmann et al., 2010) 및 TreeSHAP(Lundberg et al., 2018)을 적용하여 개별 변수가 특허 소멸여부에 미치는 영향을 알아본다. 한국 및 미국 특허의 소멸여부를 예측하는 GBM 모델에서 추출한 변수중요도 순위는 Table 4와 같다. 미국의 경우 전방인용, 특허 출원인의 평균 특허 패밀리 사이즈, 특허

가 속하는 IPC의 평균 발명인 수 등이 중요한 변수로 나타났다. 반면 한국에서는 모두 출원인의 연구역량에 관련된 변수들이 중요도 상위를 차지하였다.

Table 3. Classification results

Methods	US (USPTO)			KR (KIPO)		
	AUC	AUCPR	MCE	AUC	AUCPR	MCE
GLM	78.15(%)	63.08(%)	28.60(%)	82.96(%)	74.21(%)	24.76(%)
GBM	80.55(%)	67.39(%)	26.55(%)	88.54(%)	81.33(%)	19.04(%)
RF	80.34(%)	67.59(%)	26.56(%)	88.48(%)	81.27(%)	19.07(%)
ET	80.06(%)	66.96(%)	26.62(%)	87.95(%)	80.42(%)	19.34(%)
Super learner	80.97(%)	68.38(%)	26.35(%)	88.84(%)	81.85(%)	18.48(%)

Table 4. Variable importance

Variable Name	Permutation Importance		TreeSHAP		Variable Name	Permutation Importance		TreeSHAP	
	KR	US	KR	US		KR	US	KR	US
NB_INVENTORS	12	17	17	19	LEN_ABSTRACT	8	8	11	15
NB_APPLICATNS	13	14	3	2	IPC_AVG_CNT_APPLN_DIST	9	12	14	14
DOC_DB_FAMILY_SIZE	17	9	8	3	IPC_AVG_DOCDB_FSIZE	11	13	15	17
CNT_CTRY_CODE	20	20	20	20	IPC_AVG_AVG_NB_INVENTORS	7	3	9	13
CNT_CPC	14	16	16	6	RE_AVG_CNT_BACK_SIT	6	7	7	4
CNT_CPC_4DIGIT	18	18	19	16	RE_AVG_CNT_FOWARD_CIT	3	10	1	5
CNT_IPC	16	11	12	11	RE_AVG_AVG_DOCDB_FSIZE	4	2	6	7
CNT_IPC_4DIGIT	19	19	18	18	RE_AVG_CNT_APPLN_DIST	5	6	5	8
CNT_BACK_CIT	10	5	13	9	RE_AVG_CNT_IPC	1	15	2	12
CNT_FORWARD_CIT	15	1	10	1	RE_AVG_CNT_IPC_4DIGIT	2	4	4	10

3. 논의 및 한계점

본 연구에서는 특허의 다양한 서지정보를 수집하여 권리소멸 여부를 기계학습 모형을 통해 예측하였다. 이후 정확도 향상을 위한 stacking 기법을 사용하여 추가적인 성능 향상을 모색하였고, 중요 변수 도출을 위해 순열중요도 및 TreeSHAP을 이용하였다. 본 연구를 통해 대학으로부터 출원된 특허의 활용 가능성을 높이는 데 영향을 미치는 요인들을 확인할 수 있었다. 제시된 결과는 추후 특허의 소멸 여부를 예측하기 위한 연구에 기여할 수 있고, 대학의 미사용 특허를 줄이기 위한 정책적 방향 설정에 도움을 줄 수 있을 것으로 기대된다. 추후에는 더 다양한 변수들을 포함한 모델을 구축하여 정확도 향상을 꾀하고, 소멸여부에 영향을 미치는 변수들이 동시에 변화할 때 소멸여부와 어떤 관계를 보이는지 설명할 수 있도록 규명하는 작업이 진행되어야 할 것으로 보인다.

References

Altmann, A., Toloşi, L., Sander, O., & Lengauer, T. (2010). Permutation importance: a corrected feature importance measure. Bioinformatics, 26(10), 1340-1347.

Bzdok, D., Altman, N., & Krzywinski, M. (2018). Statistics versus Machine Learning. Nature Methods, 15, 233-234.

Criscuolo, P. (2006). The 'home advantage' effect and patent families. A comparison of OECD triadic patents, the USPTO and the EPO. Scientometrics, 66(1), 23-41.

Jaffe, A. B. (2000). The US patent system in transition: policy innovation and the innovation process. Research policy, 29(4-5), 531-557.

Lundberg, S. M., Erion, G. G., & Lee, S. I. (2018). Consistent individualized feature attribution for tree ensembles. arXiv preprint arXiv:1802.03888.

Van der Laan, M. J., Polley, E. C., & Hubbard, A. E. (2007). Super learner. Statistical applications in genetics and molecular biology, 6(1).

Proceedings of the Korean Data Analysis Society
January 26-27, 2023, pp. 181-184

층화추출법에 의한 조건부 Quatember 확률화응답모형*

이기성[1], 홍기학[2], 손창균[3], 박근화[4], 홍성준[5]

요 약

본 논문에서는 모집단이 층들로 구성되어 있을 때 층화추정이 가능한 층화 조건부 Quatember 모형을 제안하였고, 층화 조건부 Quatember 모형을 무관한 속성의 모비율이 미지일 때도 사용할 수 있는 층화 조건부 Quatember 이표본 모형으로 확장하였다.

주요용어 : 덜 민감한 속성, Quatember 모형, 층화추출, 조건부 확률화응답모형.

1. 서론

Warner(1965)는 민감한 속성을 효율적으로 추정하기 위한 확률화응답모형(randomized response model)을 처음으로 제안하였고, Loynes(1976)는 민감한 질문과 “예”라는 강요응답으로 구성된 강요질문모형을 제안하였다. 그리고 Carr, Marascuilo.(1982)는 덜 민감한 속성과 강요질문으로 만들어진 확률장치로부터 “예”라고 답한 응답자들에게만 Loynes의 모형을 이용하여 민감한 속성의 모비율을 추정할 수 있는 조건부 확률화응답모형을 제안하였다. 또한, Quatember(2009)는 5개의 질문 문항 즉, 민감한 질문, 민감한 질문과 배반되는 질문, 무관한 질문, 예, 아니오로 만들어진 새로운 형태의 확률화응답모형을 제안하였다.

본 논문에서는 모집단이 층들로 구성되어 있을 때 층화추정이 가능한 층화 조건부 Quatember 모형을 제안하였고, 층화 조건부 Quatember 모형을 무관한 속성의 모비율이 미지일 때도 사용할 수 있는 층화 조건부 Quatember 이표본 모형으로 확장하고자 한다.

2. 층화 조건부 Quatember 확률화응답모형

2.1. 층화 조건부 Quatember 모형

이 절에서는 조건부 Quatember 모형에 층화추출법을 적용하여 민감한 속성의 층화 추정이 가능한 층화 조건부 Quatember 모형을 제안하고자 한다. 모집단의 크기는 N이고, 각 층별 크기는

*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

$N_h\,(h=1,2,\cdots,L)$인 L개의 층들로 구성되어 있으며 상호 배반이고 각 층의 크기를 안다고 가정하기로 하자. 첫 번째 단계에서는 h층에서 단순임의추출된 n_h명의 응답자들에게 아래와 같은 Loynes의 모형의 확률장치로부터 선택된 질문에 대해 응답하게 한다.

질문 1 : 귀하께서는 덜 민감한 속성 B_h를 소유하고 계십니까?

질문 2 : "예"라고 응답하세요.

여기서, 질문 1의 선택 확률은 T_h이고, 질문 2의 선택 확률은 $1-T_h$이다. 이 때, 응답자들은 확률장치로부터 질문 1이 나오면 "예" 또는 "아니오"라고 응답하고, 질문 2가 나오면 "예"라고 응답하게 된다.

h층의 첫 번째 단계에서 응답자들이 "예"라고 응답할 확률 λ_{h1}는 아래와 같다.

$$\lambda_{h1} = T_h\pi_{h1} + (1-T_h)\,.$$

여기서, π_{h1}은 h층의 덜 민감한 속성 B_h의 모비율이다.

이 때, h층의 n_h명의 응답자들 중에서 "예"라고 답한 응답자들의 수를 n_{h1}명이라 두면, $\hat{\lambda}_{h1} = \dfrac{n_{h1}}{n_h}$이 되어 모비율 π_{h1}의 추정량 $\hat{\pi}_{h1}$는 아래와 같다.

$$\hat{\pi}_{h1} = \frac{1}{T_h}[\hat{\lambda}_{h1} - (1-T_h)] = \frac{1}{n_h T_h}[n_{h1} - (1-T_h)\,n_h]\,.$$

두 번째 단계에서는 h층의 전 단계에서 "예"라고 답한 n_{h1}명의 응답자들에게만 Quatember 모형의 확률장치에 의해서 선택된 질문에 대해 응답하게 한다. h층의 확률장치로부터 질문 1이 선택될 확률은 p_{h1}, 질문 2가 선택될 확률은 p_{h2}, 질문 3이 선택될 확률은 p_{h3}, 질문 4가 선택될 확률은 p_{h4}, 질문 5가 선택될 확률은 p_{h5}이다($\sum_{j=1}^{5} p_{hj} = 1$). 이 때, 응답자들은 확률장치로부터 선택된 질문에 대해 "예" 나 "아니오"라고 응답하게 된다. 따라서 h층의 응답자들이 첫 번째 단계에서 "예"라고 답했다는 조건하에서 두 번째 단계에서 "예"라고 응답할 확률 λ_{h2}는 아래와 같다.

$$\lambda_{h2} = (p_{h1} - p_{h2})\frac{\pi_{h2}}{\lambda_{h1}} + (p_{h2} + p_{h3}\pi_{hy} + p_{h4})\,.$$

여기서, π_{h2}는 h층의 민감한 속성 A의 모비율이고, π_{hy}는 h층의 무관한 속성 Y의 모비율이며 π_{hy}는 알고 있는 값이라고 가정한다.

h층의 n_{h1}명의 응답자들 중에서 "예"라고 답한 응답자들의 수를 n_{h2}라 두면 $\hat{\lambda}_{h2} = \dfrac{n_{h2}}{n_{h1}}$가 되므로 모비율 π_{h2}의 추정량 $\hat{\pi}_{h2}$는 아래와 같다.

$$\hat{\pi}_{h2} = \frac{1}{n_h(p_{h1}-p_{h2})}\left[n_{h2} - (p_{h2} + p_{h3}\pi_{hy} + p_{h4})n_{h1}\right], \quad (p_{h1} \neq p_{h2})\,.$$

따라서 민감한 속성 A의 모비율 π_2의 추정량 $\hat{\pi}_{2st}$는 아래와 같다.

$$\hat{\pi}_{2st}=\sum_{h=1}^{L}W_h\hat{\pi}_{h2}=\sum_{h=1}^{L}W_h\frac{1}{n_h(p_{h1}-p_{h2})}\left[n_{h2}-(p_{h2}+p_{h3}\pi_{hy}+p_{h4})n_{h1}\right],\ W_h=\frac{N_h}{N}$$

$\hat{\pi}_{2st}$는 모비율 π_2의 비편향추정량이며, 그 분산은 아래와 같다.

$$V(\hat{\pi}_{2st})=\sum_{h=1}^{L}\frac{W_h^2}{n_h}\left[\pi_{h2}(1-\pi_{h2})+\frac{\pi_{h2}(1-p_{h1}-p_{h2}-2p_{h3}\pi_{hy}-2p_{h4})}{p_{h1}-p_{h2}}\right.$$
$$\left.+\frac{\{T_h\pi_{h1}+(1-T_h)\}(p_{h2}+p_{h3}\pi_{hy}+p_{h4})(1-p_{h2}-p_{h3}\pi_{hy}-p_{h4})}{(p_{h1}-p_{h2})^2}\right].$$

2.2. 층화 조건부 Quatember 이표본 모형

이 절에서는 층화 조건부 Quatember 모형에서 무관한 속성을 알지 못하는 경우 두 독립표본을 이용하여 민감한 속성을 층화 추정할 수 층화 조건부 Quatember 이표본 모형으로 확장하고자 한다. 층화 조건부 Quatember 모형의 두 번째 단계에서 모비율 π_{hy}를 알지 못하는 경우에는 서로 다른 두 독립표본이 필요하며, 첫 번째 단계에서 "예"라고 답한 n_{h1}명의 응답자들로부터 단순임의복원추출된 서로 독립인 $n_{h1i}\,(i=1,2)$명의 응답자들은 층화 조건부 Quatember 모형에서 사용한 확률장치로부터 선택된 질문에 대하여 "예" 나 "아니오"라고 응답한다. 이 때, h층에서 질문 1이 선택될 확률은 p_{hi1}, 질문 2가 선택될 확률은 p_{hi2}, 질문 3이 선택될 확률은 p_{hi3}, 질문 4가 선택될 확률은 p_{hi4}, 질문 5가 선택될 확률은 p_{hi5}이다($\sum_{j=1}^{5}p_{hij}=1, i=1,2$).

따라서 h층의 응답자들이 첫 번째 단계에서 "예"라고 답했다는 조건하에서 두 번째 단계의 n_{h1i}명 응답자들이 "예"라고 응답할 확률 $\lambda_{h2i}\,(i=1,2)$는 아래와 같다.

$$\lambda_{h2i}=(p_{hi1}-p_{hi2})\frac{\pi_{h2}}{\lambda_{h1}}+(p_{hi2}+p_{hi3}\pi_{hy}+p_{hi4}).$$

h층의 n_{h1i}명의 응답자들 중에서 "예"라고 답한 응답자들의 수를 n_{h2i}라 두면 $\hat{\lambda}_{h2i}=\frac{n_{h2i}}{n_{h1i}}$가 되므로 모비율 π_{h2}의 추정량 $\hat{\pi}_{h2}$는 아래와 같다.

$$\hat{\pi}_{h2}=\frac{p_{h23}\frac{n_{h1}n_{h21}}{n_{h11}}-p_{h23}(p_{h12}+p_{h14})n_{h1}-p_{h13}\frac{n_{h1}n_{h22}}{n_{h12}}+p_{h13}(p_{h22}+p_{h24})n_{h1}}{n_h\{p_{h23}(p_{h11}-p_{h12})-p_{h13}(p_{h21}-p_{h22})\}},$$
$$p_{h23}(p_{h11}-p_{h12})\neq p_{h13}(p_{h21}-p_{h22}).$$

따라서 민감한 속성 A의 모비율 π_2의 추정량 $\hat{\pi}_{2st}$는 아래와 같다.

$$\hat{\pi}_{2st} = \sum_{h=1}^{L} \frac{W_h}{n_h} \left[\frac{p_{h23}\frac{n_{h1}n_{h21}}{n_{h11}} - p_{h23}(p_{h12}+p_{h14})n_{h1} - p_{h13}\frac{n_{h1}n_{h22}}{n_{h12}} + p_{h13}(p_{h22}+p_{h24})n_{h1}}{p_{h23}(p_{h11}-p_{h12}) - p_{h13}(p_{h21}-p_{h22})} \right].$$

$\hat{\pi}_{2st}$는 π_2의 비편향추정량이며, 그 분산은 아래와 같다.

$$V(\hat{\pi}_{2st}) = \sum_{h=1}^{L} \frac{W_h^2}{n_h} \frac{\frac{p_{23}^2\alpha_h}{n_{h11}} + \frac{p_{h13}^2\beta_h}{n_{h12}}}{\{p_{h23}(p_{h11}-p_{h12}) - p_{h13}(p_{h21}-p_{h22})\}^2},$$

여기서,

$$\alpha_h = \lambda_{h21}(1-\lambda_{h21})\{\lambda_{h1}(1-\lambda_{h1}) + n_h\lambda_{h1}^2\} + n_{h11}\lambda_{h1}(1-\lambda_{h1})[\lambda_{h21}^2 + (p_{h12}+p_{h14})\{(p_{h12}+p_{h14}) - 2\lambda_{h21}\}],$$

$$\beta_h = \lambda_{h22}(1-\lambda_{h22})\{\lambda_{h1}(1-\lambda_{h1}) + n_h\lambda_{h1}^2\} + n_{h12}\lambda_{h1}(1-\lambda_{h1})[\lambda_{h22}^2 + (p_{h22}+p_{h24})\{(p_{h22}+p_{h24}) - 2\lambda_{h22}\}]$$

이다.

3. 결론

본 논문에서는 모집단이 층들로 구성되어 있을 때 층화추정이 가능한 층화 조건부 Quatember 모형을 제안하였고, 층화 조건부 Quatember 모형을 무관한 속성의 모비율이 미지일 때도 사용할 수 있는 층화 조건부 Quatember 이표본 모형으로 확장하였다.

Reference

Carr, J. W., Marascuilo, L. A. (1982). Optimal randomized response models and methods for hypothesis testing, *Journal of Educational Statistics*, 7, 295-310.

Loynes, R. M. (1976). Asymptotically optimal randomized response procedures, *Journal of the American Statistical Association*, 71, 924-928.

Quatember, A. (2009). A standardization of randomized response strategies, *Survey Methodology*, 35(2), 143-152.

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 26-27, 2023, pp. 185-187

토픽모델링을 기반으로 한 디지털 신기술에 대한 트위터 데이터의 키워드 분석*

이윤진[1], 정호현[2]

1. 서론

현 정부가 2026년까지 '100만 디지털 인재양성'을 목표로 전 국민의 디지털 교육 기회 확대와 역량 강화를 지원하기 위해 '디지털 인재양성 종합방안'을 마련했다. 이에 따라 4차 산업혁명과 관련된 디지털 신기술에 대한 대중의 관심과 동향을 파악하기 위해 트위터 데이터를 분석하였다. 토픽모델링을 기반으로 디지털 신기술에 대한 키워드 분석을 통해 대중들의 관심 분야를 알아보았다. 키워드 동시출현네트워크를 이용한 네트워크 분석을 통하여 디지털 신기술 키워드의 관계도와 연계성 측면에서의 키워드 중요도를 알아보았다. 소셜네트워크 분석을 통해 국내외 대중들의 관점에서 4차 산업혁명을 해석하였다는 데 의의가 있고, 추후 유망 분야에 대한 정부 정책이나 투자를 위한 기초 자료로 활용될 가치가 있을 것으로 기대한다.

2. 연구방법

본 연구는 디지털 신기술에 대한 키워드로 한국직업능력연구원이 디지털 신기술이라고 언급한 기술들 중 인공지능, 빅데이터, 사물인터넷, 클라우드, 블록체인, 소프트웨어, 메타버스를 국문 키워드로 AI, BigData, Iot, Cloud, BlockChain, SoftWare, Metaverse를 영문 키워드로 선정하였다.

우선 국문 키워드를 중심으로 트위터에서 2020년 7월부터 2021년 12월까지 '인공지능, 빅데이터, 사물인터넷, 클라우드, 블록체인, 소프트웨어, 메타버스' 중 하나라도 포함된 트윗 500개를 하루 간격으로 추출하여 6개월을 하나의 구간으로 분류하여 데이터를 수집하였다. 수집한 데이터는 한글 이외의 문자, 한 글자 키워드, 불용어를 제거하는 전처리 과정을 수행하였다. 마찬가지로 같은 기간 동안 영문 키워드인 'AI, BigData, Iot, Cloud, BlockChain, SoftWare, Metaverse' 중 하나라도 포함된 트윗 500개를 하루 간격으로 추출하여 6개월을 하나의 구간으로 분류하여 데이터를 수집하였다 (Jeong et al, 2015). 수집한 데이터는 대소문자를 통합하는 정규화, 토큰화, 표제어 추출, 불용어 제거의 전처리 과정을 수행하였다.

본 연구에서 디지털 신기술의 수요를 파악하고 각광받는 분야를 살펴보기 위해 LDA(Latent Dirichlet Allocation) 토픽모델링을 수행하였다. 국문 키워드의 구간별 토픽 수는 10개, 중요 단어는

*이 논문은 한국연구재단(2021R1G1A1094103)의 지원을 받아 연구된 것임.
[1]02844 서울특별시 성북구 돈암동 보문로34다길 2, 성신여자대학교 통계학과 학석사연계과정.
E-mail: 20200778@sungshin.ac.kr
[2](교신저자) 02844 서울특별시 성북구 돈암동 보문로34다길 2, 성신여자대학교 수리통계데이터사이언스학부 조교수.
E-mai: hhjung@sungshin.ac.kr

10개로 설정하고, 영문 키워드의 경우도 구간별 토픽 수는 10개, 중요 단어는 10개로 설정하고 LDA모델에 적용하여 토픽모델링을 수행하였다.

본 연구는 하나의 트윗에 단어가 동시에 출현한 패턴을 분석하여 해당 트윗들이 가지고 있는 키워드 간 연관관계를 파악하는 방법인 동시출현단어 네트워크 분석(co-occurence network analysis)을 수행하였다. 토픽모델링을 위해서 수집한 트위터 데이터를 이용하여, 하나의 문서에 출현한 단어들을 노드로, 인접하여 출현했을 때를 엣지로 연결하여 동시출현단어 네트워크를 생성하였다. 생성된 네트워크를 바탕으로 키워드 연계 관점에서의 키워드들의 영향력을 파악하기 위해서 차수 중심성과 매개 중심성을 측정하였다.

3. 연구결과

Table 1. 기간별 국문, 영문 키워드 토픽

Topic number	Period 1 20.07~20.12		Period 2 21.01~21.06		Period 3 21.07~21.12	
	Korean	English	Korean	English	Korean	English
1	블록체인	software	인공지능	blockchain	클라우드	blockchain
2	클라우드	data science	플랫폼	platform	빅데이터	IoT
3	인공지능	software	블록체인	big data	메타버스	AI
4	빅데이터	cloud	클라우드	cryptocurrency	사물인터넷	metaverse
5	인공지능 로봇	blockchain	빅데이터	metaverse	게임 산업	nft
6	소프트웨어	AI robot	추천시스템	nft	블록체인	software
7	게임 산업	mobile industry	메타버스	bitcoin	주식 시장	data mining
8	모바일	fintech	사물인터넷	software	플랫폼	stock market
9	정보 처리	machine learning	인공지능 로봇	game industry	인공지능 로봇	cryptocurrency
10	플랫폼	cryptocurrency	소프트웨어	AI	정보 처리	robotics

Table 1은 기간별 국문 키워드와 영문 키워드의 중요 단어들에 대한 토픽이다. 모든 기간별로 블록체인, 클라우드, 인공지능에 관련된 단어들이 많이 언급되었다는 것을 알 수 있다. 또한 게임 산업에서 메타버스를 이용한 게임들을 많이 출시했다는 것도 알 수 있다(Chae et al, 2021).

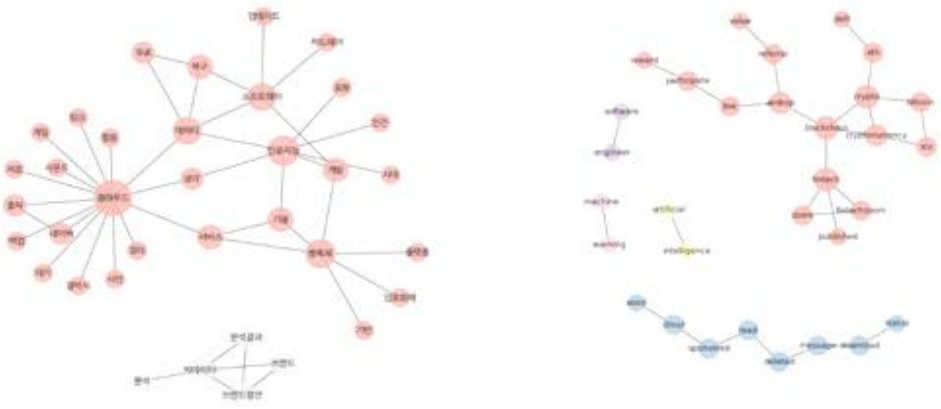


Figure 1. 기간 1 동시출현단어 네트워크

Figure 1은 기간 1의 국문과 영문 트위터 데이터의 동시출현단어 네트워크이다. 노드의 크기는 차수 중심성과 비례하게 나타내었다. 기간 1의 국문 동시출현단어 네트워크에서는 클라우드, 인공지능, 블록체인, 빅데이터의 차수 중심성이 높은 것을 확인할 수 있다. 기간 1의 영문 동시출현단어 네트워크에서는 blockchain이 차수 중심성이 높은 것을 확인할 수 있으며, 의미가 비슷한 단어들끼리 군집을 이루고 있는 것도 확인할 수 있다.

4. 결론

트위터 데이터를 토픽모델링과 동시출현단어를 이용하여 분석하여 2020년 7월부터 2021년 12월까지 디지털 신기술에 대한 대중의 일반적 인식을 파악하였다. 모든 기간에 블록체인, 인공지능, 클라우드가 핵심적인 토픽으로 추출된 것을 보아 앞으로도 해당 키워드에 대한 주제들은 디지털 신기술에 영향력있는 분야가 될 것이라고 판단된다. 이 결과를 바탕으로 실무적으로 정부 정책이나 투자를 위한 기초 자료로 활용될 가치가 있을 것으로 기대한다.

추후 연구 과제로 향후 대중의 관심도와 실제 정책이나 특허, 논문 등이 어떤 차이가 있는 지 알아보고 정책적으로 지원해야 할 공백기술 영역을 파악할 수 있다. 또한, 동시출현단어 네트워크에 Exponential Random Graph Models(ERGMs)와 같은 모델을 적용하여 4차 산업혁명 키워드 간의 관계를 심층적으로 분석한다면 디지털 사회의 특성을 파악하는 데 더 의미 있는 결과를 도출할 수 있을 것이다.

References

Chae H-G., Lee G-H., Lee J-Y. (2021). Analysis of Domestic and Foreign Financial Security Research Activities and Trends through Topic Modeling Analysis. *Journal of the Korea Industrial Information Systems Research*, 26(1), 83-95.

Jeong B., Kim T., Lee J., Kim J. (2015). Twitter Topic Extraction and Topic Category Decision using LDA Model. *Journal of the Korean Institute of Information Scientists And Engineers*, 787-788.

Proceedings of the Korean Data Analysis Society
January 26-27, 2023, pp. 189-191

토픽모델링과 시계열 분석을 활용한 국내외 독과점 산업 동향 분석*

이수지[1], 정호현[2]

1. 서론

국내는 1960년대 시작된 수출주도형 공업화 전략으로 특정 산업을 중심으로 기업의 대형화, 독과점화 현상이 발생하였다. 이에 1981년 공정거래법이 시행되었지만, 시간이 지남에 따라 신산업이 등장하며 법과 간극이 발생하였다(Lee, 2005). 국외 중 대표적으로 미국을 보면 미국산업의 75%에서 소수 기업이 독과점으로 이익을 누리고 있어 소비자, 중소기업 등에서 피해가 증가했고, 미국 내 대기업의 특정 산업군에 집중된 인수합병으로 경쟁력이 약화되고 있다(Jang & Kang, 2021).

RISS와 Web of Science 논문 데이터를 이용하여 텍스트 마이닝 기법으로 국내외 연구 동향을 구조화하고, 시계열 분석을 통하여 연구 동향의 추세 파악 및 미래 방향을 예측하였다. 본 연구의 목적은 국내외 독과점 산업 동향을 파악 및 예측하여 해당 독과점 산업의 경쟁 시장 촉진을 위해 어떠한 연구에 비중을 두어야 하는지 제안하고, 국가 정책 수립 및 R&D 예산 책정에 필요한 기초 자료로 활용함에 있다. 이를 바탕으로 현행되고 있는 국내외 반독점법의 사각지대를 줄일 수 있기를 기대한다.

2. 연구 방법

본 연구는 RISS와 Web of Science에서 1995년부터 2021년까지 독과점 관련 논문 초록 데이터를 분석하였다. 전처리 과정과 함께 데이터 추출 시 사용했던 키워드는 불용어로 간주하고 제거하였고, 토픽 수는 15개, 토픽별 단어 수는 5개로 설정하여 LDA 토픽모델링을 진행하였다.

토픽 중요도를 계산하기 위해 연도별로 문서를 분류하고, 연도별 단어 출현 횟수를 리스트 형태로 저장하였다. 토픽모델링 과정에서 계산된 각 토픽의 빈도에 연도별 단어 출현 횟수를 곱한 것의 합에서 전체 토픽 빈도에 년도 별 단어 출현 횟수를 곱한 것의 합으로 나누고 100을 곱하여 각 토픽의 연도별 중요도를 계산하였다. 이 중요도를 시계열 데이터로 하여 AIC를 기준으로 토픽마다 최적의 ARIMA(Autoregressive Integrated Moving Average) 모델을 구축하였다. 시계열 모형을 바탕으로 이후 3년의 토픽 동향을 예측하였다.

추가로 1995년부터 2021년 동안의 토픽 동향을 알아보기 위해 회귀분석을 진행하였다. 종속변수를 토픽의 중요도, 독립변수를 연도로 하여 분석하였고, 유의확률 0.05에서 회귀계수의 p-value가 유의한 토픽을 대상으로 hot topic과 cold topic을 구분하였다.

---

*이 논문은 한국연구재단의 지원을 받아 연구된 것임(2021R1G1A1094103).

[1]02844 서울특별시 성북구 보문로 34다길 2, 성신여자대학교 통계학과 학석사연계과정. E-mail: 20200457@sungshin.ac.kr

[2](교신저자) 02844 서울특별시 성북구 보문로 34다길 2, 성신여자대학교 수리통계데이터사이언스학부 조교수. E-mail: hhjung@sungshin.ac.kr

3. 연구 결과

Table 1은 토픽모델링 결과 추출된 토픽별 단어를 바탕으로 토픽을 정한 것이다. Table 2는 각 토픽 회귀분석 결과를 정리한 것이다. 토픽 중 유의하며 계수 값이 양수이면 hot topic, 음수이면 cold topic으로 분류하였다.

Table 1. 국내외 토픽

Topic number	Korean topic	Global topic
0	재판	price competition
1	지식 재산권	online market
2	계약	development of industry
3	에너지	government policy
4	방송	education
5	연구 개발	balanced development
6	온라인 거래	online competition
7	기업 발전 연구	game theory
8	제도 개혁	network management
9	통신	energy
10	기업 투자	merge
11	공정 거래	communication expense
12	기업 정책	patent
13	지위 남용	software project
14	신용	trade

Table 2. 국내외 토픽 동향 선형 회귀분석 결과

Topic number	Korean topic			Global topic		
	coefficient	p-value	AIC	coefficient	p-value	AIC
0	0.084	<.001	56.36	-0.0512	0.071	83.71
1	0.0704	0.03	89.94	0.0213	0.041	28.91
2	0.1072	<.001	64.74	0.0324	0.003	28.91
3	-0.0888	<.001	68.64	-0.0497	0.006	56.92
4	-0.437	0.354	112.5	0.0584	<.001	31.23
5	-0.1921	<.001	85.07	0.0191	0.243	54.87
6	0.1389	<.001	84.95	-0.0066	0.627	45.66
7	-0.0373	0.18	83.34	-0.0084	0.362	24.17
8	-0.1344	0.001	97.3	0.0243	0.192	61.75
9	-0.0733	0.012	83.59	-0.0802	0.018	92.01
10	0.0252	0.124	54.38	-0.012	0.235	28.75
11	0.1225	<.001	78.42	-0.0374	0.001	28.71
12	-0.0851	<.001	68.41	0.0624	0.004	66.71
13	0.0615	0.002	59.58	-0.0119	0.234	28.58
14	0.045	0.477	128.5	0.0397	0.001	33

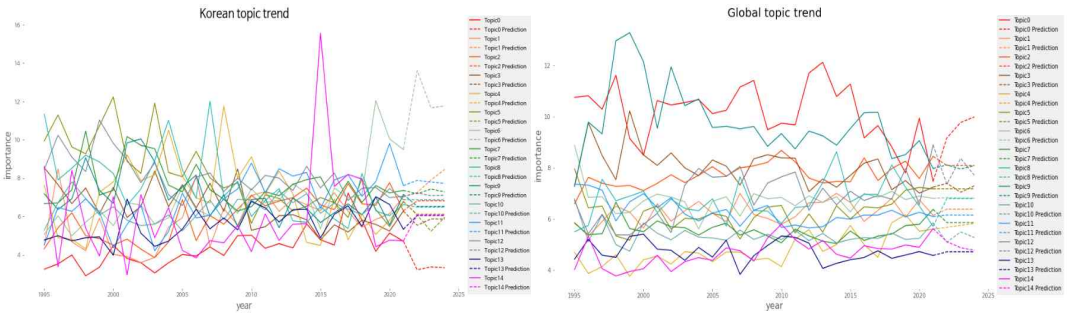


Figure 1 . 국내외 토픽 중요도 동향 및 예측

Figure 1은 국내외 토픽 중요도 동향과 예측값을 그래프로 나타낸 것이다. Table 2와 Figure 1을 종합적으로 보았을 때, Hot topic이고 미래에도 중요도가 올라가는 토픽은 국내에서는 topic 6(온라인 거래), 국외에서는 topic 4(education)으로 확인하였다. Cold topic이고 미래에도 중요도가 꾸준히 감소하는 토픽은 국내에서는 topic 3(에너지)와 topic 9(통신), 국외에서는 topic 3(government policy)으로 확인하였다.

4. 결론

본 연구는 국내외 독점 산업 동향을 토픽모델링과 시계열 모델을 통해 분석하였다. 국내에서는 '온라인 거래'의 중요도가 오르며 연구가 활발히 되고 있는데, 비대면 거래와 온라인 플랫폼이 활성화되며 일부 대형 기업이 온라인 시장을 장악함에 따라 이에 관한 규제 연구가 중요해지고 있다고 판단된다. 반면 '에너지', '통신'에 관한 연구 중요도는 낮아지고 있다. 그 이유로 신재생 에너지 보급이 확대됨에 따라 '한국 전력' 등 일부 기업의 에너지 독점 현상이 완화되었고, 구독 서비스 활성화에 따라 통신사 간 경계가 모호해지며 통신 3사 과점에서 경쟁 시장으로 변화되는 과정이 반영된 것으로 생각된다. 국외에서는 'education'에 관한 연구 중요도가 증가하는 반면 'government policy'에 관한 연구 중요도는 낮아지고 있다. 국내에서도 대학 서열화에 따른 특정 대학의 과점 현상이 발생하는 상황에서 국내 교육 시장 독과점 연구도 더욱 밀도 있게 진행되어야 한다고 생각한다. 마지막으로 본 연구 결과가 독과점 관련 연구를 바탕으로 한 정책 수립 및 R&D 연구과제 예산 책정의 기초 자료로 사용될 수 있을 것으로 기대한다.

References

Lee, S. B.(2005). Current Status and prospects of Industrial Organization Studies in Korea, *The Korean Journal of Industrial Organization*, 13(4), 1-41. (in Korean)

Jang, Y. S., Kang, G. S.(2021). Competition policy and platform antitrust legislative trends and implications in the US, *Korea institute for international economic policy*, 21-16. (in Korean)

Proceedings of the Korean Data Analysis Society
January 26-27, 2023, pp. 193-194

함수적 회귀모형을 활용한 알츠하이머 관련 유전자 변이 탐지 연구

이윤석[1], 이은지[2]

요 약

알츠하이머병(Alzheimer's disease)은 노인의 건강에 영향을 주며, 큰 사회적 영향을 미치는 신경퇴행성질환이다. 이는 치매의 가장 흔한 형태로 치매 환자의 약 60~80%를 차지하며, 신경계의 노화가 진행되는 상태에서 다양한 환경 및 유전적 요인에 의해 발병이 가능한 유전질환이다. 알츠하이머병 발병에는 연령과 같은 환경적 기저요인과 더불어 인간 게놈(Genome)에서 가장 빈번하게 발생하는 유전적 변이인 SNP(Single Nucleotide Polymorphism)이 발병의 위험요인으로 밝혀진 바 있다. 현재 GWAS(Genome-wide association Studies)를 통해 알츠하이머병을 포함한 여러 유전질환의 Bio-Marker로 900만 개가 넘는 SNP이 발견되었는데, 대표적으로 APOE-ϵ4(Apolipoprotein E(ϵ4 allele), rs429358)는 알츠하이머병(AD)에 가장 강력한 유전적 위험인자로 인지 기능 저하와 직접적인 관련이 있으며, 그 밖에 CR1(Complement Component receptor-1, rs3818361), BIN1 (Bridging Integrator-1, rs744373) 등 알츠하이머병과 관련된 다양한 SNP이 연구되었다. 이러한 SNP은 수집된 표본과 전처리 방법에 따라 대상마다 적게는 30만에서 많게는 100만개 가량 기록된다. 특히 인접한 SNP들은 서로 유사한 값을 갖는 특징이 있는데, 엄밀히 말해서 SNP은 이산화 된 공간(Discrete Space)에서 관측된 자료이지만 충분히 촘촘한 염색체 영역에서 발생하며 인접한 SNP사이에 밀접한 관계가 있으므로 우리가 관측한 SNP들을 연속(Continuous)인 염색체 영역을 정의역으로 하는 어떤 연속형 함수의 이산화 된 관측값으로 가정할 수 있다. 위 가정을 토대로 SNP에 함수적 자료분석(Functional Data Analysis) 방법론을 적용하여 알츠하이머병 발병과 관련된 유전학적 변수로 활용하고자 한다. 따라서 본 연구에서는 알츠하이머병 환자와 정상인지 능력을 갖춘 노인, 두 그룹을 판별하는 로지스틱 회귀모형에 SNP을 함수적 설명변수로 활용하여 알츠하이머병 발병에 유의미한 SNP의 탐지를 목표로 한다.

주요용어 : Alzheimer's disease, GWAS, SNP, Functional Data Analysis, Logistic Regression.

[1]34134 대전광역시 유성구 대학로 99, 충남대학교 정보통계학과 석사과정. E-mail: toatoal123@gmail.com
[2]34134 대전광역시 유성구 대학로 99, 충남대학교 정보통계학과 교수. E-mail: eunjee@cnu.ac.kr

References

DeTure, M. A., Dickson, D. W. (2019) The neuropathological diagnosis of Alzheimer's disease, *Molecular Neurodegeneration*, 14, 32. https://doi.org/10.1186/s13024-019-0333-5

Fan, R., Wang, Y., Mills, J. L., Wilson, A. F., Bailey-Wilson, J. E., Xiong, M. (2013). Functional Linear Models for Association Analysis of Quantitative Traits, *Genet Epidemiol*, 37(7), 726-742. doi:10.1002/gepi.21757

Jung, J. Hwan., Ji, S. Jin., 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(6), 603–624. https://doi.org/10.29220/CSAM.2020.27.6.603

Luo, L., Boerwinkle, E., Xiong, M. (2011). Association studies for next-generation sequencing, *Genome Research*, 21, 1099-1108. doi:10.1101/gr.115998.110

Nettiksimmons, J., Tranah, G., Evans, S. D., Yokoyama, J. S., Yaffe, K. (2016) Gene-based aggregate SNP associations between candidate AD genes and cognitive decline. *AGE*, 38, 41. doi:10.1007/s11357-016-9885-2

Ridge, P. G., Mukherjee, S., Crane, P. K., Kauwe, J. S. K. (2013) Alzheimer's Disease: Analyzing the Missing Heritability, *PLoS ONE*, 8(11), e79771. https://doi.org/10.1371/journal.pone.0079771

Proceedings of the Korean Data Analysis Society
January 26-27, 2023, pp. 195-198

아파트 주거만족도에 영향을 미치는 주거환경 요인에 대한 메타분석

김명인[1], 윤상복[2], 이현경[3]

요 약

최근 경제발전과 함께 국민의 생활수준이 향상되면서 주택의 양적 충족뿐만 아니라 삶의 질 향상에 대한 욕구가 급속하게 증가하게 되었고, 이에 주거의 질을 평가하기 위한 척도로 주거만족도와 이에 영향을 미치는 요인에 대한 연구가 활발하게 이루어지고 있는 실정이다. 이와 같은 주거만족도와 이를 평가하는 지표로서의 주거환경 요인은 국민의 주거안정과 주거만족 수치를 파악하기 위하여 다양하게 연구되고 있으나, 연구의 관점, 목적, 방법 등에 따라 의미와 범주가 축소 또는 확대 해석되어 정리되고 있기 때문에 이를 체계적이고 계량적으로 통합하여 분석할 필요성이 매우 높기 때문에, 본 연구에서는 주거만족도란 동일한 주제에 대하여 영향을 미치는 요인에 관한 다양한 연구결과를 메타분석을 통해 체계적으로 계량하고 분석하여 주거환경 요인에 대한 효과크기를 다음과 같이 도출하였다. 물리적 요인 중 가장 큰 효과크기를 보인 요인은 환경요인(0.59)으로 나타났으며, 다음으로 입지적 요인(0.47), 내부시설 요인(0.45), 단지 외부요인(0.43)순인 것으로 나타났다. 그리고 사회적 요인 중 가장 큰 효과크기를 보인 요인은 아파트 관리(0.45)로 나타났으며, 다음으로 지역 커뮤니티 요인(0.44), 경제적 요인(0.38)순인 것으로 나타났다.

주요용어 : 아파트, 주거환경, 영향요인, 메타분석, 효과크기.

1. 서론

우리나라는 1970년대 이후 급속한 경제성장으로 수도권 지역이 팽창하고 도시화 및 핵가족화가 빠르게 진행되면서 도시 지역의 심각한 주택 부족 문제를 야기하였고, 주택문제 부족을 해결하기 위하여 정부는 1980년 「택지개발촉진법」을 제정하고 대규모로 택지를 개발하고 양적인 아파트 공급을 시행하였으며, 국내의 대표적인 주거형태로 자리 잡게 되었다.

그러나 최근 경제발전과 함께 국민의 생활수준이 향상되면서 주택의 양적 충족뿐만 아니라 삶의 질 향상에 대한 욕구가 급속하게 증가하게 되었고, 이에 주거의 질을 평가하기 위한 척도로 주거만족도와 이에 영향을 미치는 요인에 대한 연구가 활발하게 이루어지고 있는 실정이다.

이와 같은 주거만족도와 이를 평가하는 지표로서의 주거환경 요인은 국민의 주거안정과 주거만족 수치를 파악하기 위하여 다양하게 연구되고 있으나, 거주지 대한 만족도는 물리적, 경제적, 소비재적 측면, 입지적 측면, 환경적인 측면 등 다양한 측면에서 연구되고 있으며, 연구의 관점, 목

[1]47340 부산광역시 부산진구 엄광로 176, 동의대학교 공과대학 도시공학과 박사과정.
E-mail: kmi930424@naver.com
[2](교신저자) 47340 부산광역시 부산진구 엄광로 176, 동의대학교 공과대학 도시공학과 교수.
E-mail: yoonsb@deu.ac.kr
[3]47340 부산광역시 부산진구 엄광로 176, 동의대학교 공과대학 도시공학과 교수. E-mail: lhk@deu.ac.kr

적, 방법 등에 따라 의미와 범주가 축소 또는 확대 해석되어 정리되고 있기 때문에 이를 체계적이고 계량적으로 통합하여 분석할 필요성이 매우 높다.

따라서 본 연구에서는 주거만족도란 동일한 주제에 대하여 영향을 미치는 요인에 관한 다양한 연구결과를 체계적으로 계량하고 분석하여 근거기반 연구의 실천을 달성하고 주거환경 요인에 대한 효과크기를 도출하여 통계적 유의성을 검증함으로써 설득력 있는 근거를 제공하고자 한다.

2. 연구 설계와 방법

2.1. 연구대상의 선정과 자료 수집

분석대상 논문은 아파트를 연구대상으로 설정하고 주거만족도를 종속변수로 하는 계량적 연구로 선정하였다. 분석 자료의 수집은 학술연구정보서비스(www.Riss.kr)와 한국학술정보(www.kiss.kstudy.com)를 활용하여 '주거만족도'를 키워드로 논문을 검색하였다.

먼저 주제 선정 키워드인 '주거만족도'로 논문을 검색하여 3,434편을 도출하였으며, 결과 내 재검색 기능을 사용하여 '아파트'를 대상으로 연구를 수행한 논문 973편을 선정하였다. 다음으로 1차 선정과정은 도출된 973편의 논문을 대상으로 질적 연구, 사례연구 등을 제외한 논문을 선별하였으며, 이 중 주거만족도와 관계없는 연구, 주거환경이 통제변수가 아닌 연구를 제외한 64편을 선정하였다.

2차 선정과정에서는 통일성 있는 분석을 위하여 주거만족도에 관련된 연구이더라도 아파트 이외의 주거형태가 혼재된 연구, 주거만족도 이외의 변수를 종속변수로 설정한 논문을 제외하여 최종적으로 19편의 논문을 선정하였으며, 세부적인 자료 수집과정은 Figure 1.과 같다.

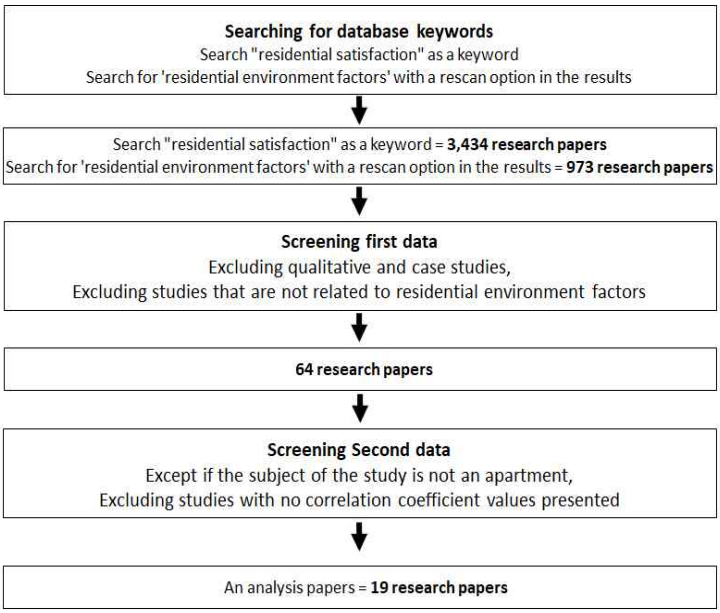


Figure 1. The process of selecting papers to be analyzed

2.2. 독립변수의 분류

메타분석의 목적을 달성하기 위해서는 먼저 기존의 다양한 연구에서 사용한 독립변수들을 체계적으로 분류하는 작업이 요구된다. 따라서 본 연구에서는 주거만족도에 영향을 미치는 요인을 독립변수로 설정하고 선정된 분석대상 연구들에서 사용된 독립변수들을 추출하여 정리하였다. 효과크기(상관계수) 비교를 위해서는 최소 2개 이상의 효과크기가 제시되어야하므로 효과크기가 1개인 변수들은 제외하였으며, 최종적으로 7개의 변수 62개의 효과크기를 도출하였다(Table 1).

Table 1. Independent variable to be analyzed

Sortation	Variable	Number of cases
Physical factors	Locational factor	10
	Outside the residential factor	14
	Interior Facilities factor	13
	Environmental factor	5
Social factors	Economic factor	8
	Local community factor	4
	apartment management	8

2.3. 메타분석 절차 및 방법

메타분석의 분석절차 및 방법은 다음과 같이 구분할 수 있다.

첫째, 메타분석은 상호 다른 연구물의 효과크기를 통합하는 분석기법이기 때문에 이를 타당화하기 위해서는 개별 효과크기에 대한 이질성 검증이 요구된다. 본 연구에서는 이질성 검정을 위하여 Q통계량 및 I^2 값을 사용하였다. 이질성 판단의 기준은 Hwang(2018)의 연구에 따르면, Q통계량의 카이제곱검증 결과가 통계적으로 유의하면 이질성이 높다고 판단하고, I^2이 25%이하면 이질성이 낮고, 25 ~ 75%면 중간 정도, 75%이상이면 이질성이 높다고 판단할 수 있다. 그리고 동질성 검증 결과 개별 효과크기의 이질성이 확인된 경우에는 각 연구로부터 추출된 효과 크기가 연구에 따라 달라진다고 가정하는 무선효과모형(Random effect model)을 사용하여 평균효과 크기를 추정하였다.

둘째, 메타분석은 기존 연구의 결과물을 효과크기(Effect size)로 산출하여 해석하게 되며, 본 연구에서는 개별 연구에서 추출한 상관계수를 사용하여 효과크기를 산출하였다. 또한, 산출된 효과크기의 해석은 일반적으로 Cohen(1998)의 해석기준에 따라 해석하게 되며, 상관계수(r) 값이 0.1이면 작은 효과크기, 0.3이면 중간 효과크기, 0.5이면 큰 효과크기로 해석하였다.

3. 분석 결과

메타분석 모형 결정을 위해 개별 연구에서 도출된 62개 효과크기(상관계수)에 대한 동질성 검증을 실시하였다. 그 결과 Q값이 1165.09(df=61, p<.000.1)로 나타났으며, Q통계량에 대한 p값이 0.05 이하이므로 분석 대상으로 선정된 연구의 결과들이 서로 이질성을 띈다고 할 수 있다(Table 2).

주거만족도에 영향을 미치는 주거환경 요인의 효과크기를 분석한 결과는 Table 3과 같다.

물리적 요인은 주로 아파트 주거환경 요인 중 주거만족도에 영향을 미치는 그 형태나 모양이 아파트의 위치, 형태, 모양, 주변 여건 등 유형의 형태를 가지고 있는 요인으로 판단할 수 있다. 본

연구에서는 입지적 요인, 단지외부 요인, 내부시설 요인, 환경 요인으로 분류되었다. 물리적 요인의 하위 변수가 주거 만족도에 미치는 효과크기는 환경 요인(0.59), 입지적 요인(0.47), 내부시설 요인(0.45), 단지 외부요인(0.43) 순인 것으로 나타났다.

사회적 요인은 주로 아파트 주거환경 요인 중 주거만족도에 영향을 미치는 그 형태나 모양이 경제적 여건, 이웃과의 관계, 공동체 의식, 아파트 관리 상태 등 무형의 형태를 가지고 있는 요인으로 판단할 수 있다. 본 연구에서는 경제적 요인, 지역 커뮤니티 요인, 아파트 관리로 분류되었다. 사회적 요인의 하위 변수가 주거 만족도에 미치는 효과크기는 아파트 관리(0.45), 지역 커뮤니티 요인(0.44), 경제적 요인(0.38) 순인 것으로 나타났다.

Table 2. Heterogeneity Verification Results

Number of cases(k)	Q	df	p	I^2
62	1165.09	61	0.0001	94.8%

Table 3. Derive effect size by factor

	Sortation	Number of cases(k)	ES(r)	95% CI		I^2
				lower confidence limit	upper confidence limit	
Physical factors	Locational factor	10	0.47	0.36	0.58	94%
	Outside the residential factor	14	0.43	0.35	0.50	88%
	Interior Facilities factor	13	0.45	0.32	0.56	96%
	Environmental factor	5	0.59	0.27	0.79	98%
Social factors	Economic factor	8	0.38	0.23	0.50	93%
	Local community factor	4	0.44	0.25	0.59	86%
	apartment management	8	0.45	0.25	0.61	97%

References

An, W. H., Lee, J. S., Choi, E. C., (2022). Analysis of the Effect of each National Park Characteristics on Willingness To Pay: Using Meta-analysis, *Journal of The Korean Data Analysis Society*, 24(6), 2289-2298. (in Korean). https://doi.org/10.37727/jkdas.2022.24.6.2289

Cohen, J., (1988) Statistical Power Analysis for Behavioral Sciences. Hillsdale, NY: Lawrence Erlbaum Associates.

Hwang, S. D., (2016). Publication Bias in Meta-Analysis: Its Meaning and Analysis, *The Korean Journal of the Human Development*, 23(1), 1-19. (in Korean). http://dx.doi.org/10.15284/kjhd.2016.23.1.1

Jang, J. S., (2020). A Meta Analysis of Influencing Factors of Police Performance, T*he Journal of Police Science*, 20(2), 61-91. (in Korean).

Park, H. O., Jung, T. S., Sung, M. Y., (2009). A Study on the Residential Environment Satisfaction of Apartment Complex, *Korea Real Estate Society*, 27(2), 223-240. (in Korean).

Yoo, S. M., (2019). Accuracy Comparison of Effect Size Calculation Methods in Meta-analysis, *Journal of The Korean Data Analysis Society*, 19(3), 1233-1246. (in Korean).

Proceedings of the Korean Data Analysis Society
January 26-27, 2023, pp. 199-202

인공신경망을 활용한 미세먼지 예측 모델 구축 기초 연구*

이재현[1], 이현경[2], 윤상복[3]

요 약

미세먼지에 대한 사회적인 이슈가 확산됨에 따라 기상청, 한국환경공단에서 실시간으로 제공하는 정보를 토대도 새로운 정보 제공 서비스와 미세먼지를 예측하는 방법 연구가 많아 지고 있다. 향후 미세먼지 수치 예측으로 정부의 정책, 도시설계의 변화 등 대비에도 필요하다. 본 연구에서는 2019~2021년 3년간의 데이터를 활용하였으며, 미래 예측에 인공신경망 모델을 활용하였다. 효과적인 인공신경망 모델의 학습값으로 훈련용 데이터의 비율80%, 테스트 데이터의 비율20%가 적절한 비율로 나타났으며, PM10에서 6,432번의 학습, 90.494842의 Error, PM2.5에서는 3,741번의 학습, 257.923995의 Error가 나타났다. 향후 오차범위를 줄이는 방법으로 데이터의 목록을 인식하도록 하여 데이터의 인식률을 올려 Error을 낮추고 적중률을 계산하여 정확성을 높여야 할 필요가 있다. 미세먼지를 예측하는 기초연구로서 연구모델을 검증하였다. 연구모델의 정확성을 높여 미세먼지와 도시설계에 연계 등 타 분야 까지 활용할 수 있는 기초 연구가 되기를 기대한다.

주요용어 : 인공신경망, R Studio, 미세먼지, 도시설계, 예측모델.

1. 서론

미세먼지에 대한 사회적인 이슈가 확산됨에 따라 한국환경공단에서는 미세먼지 정보를 제공하고 있으며 관련 새로운 정보 제공 서비스가 생겨나고 있다. 미세먼지를 수집하는 과정에서 데이터 센서의 오류, 고장, 기기 점검 등 정보제공 서비스가 원활히 이루어지지 않는다. 이러한 문제를 해결하기 위하여 측정되지 못한 미세먼지 데이터 예측이 필요하다. 대기오염의 속도 및 오염 정도를 낮추는 정책이 펼쳐지고 있으나 미비한 실정으로 미세먼지에 대한 장기적인 예측이 필요하며 이에 알맞은 정보 예측 데이터가 필요하다.

본 연구에서는 미세먼지를 예측하기 위하여 인공신경망을 활용하여 예측 모델을 설계한다. 이를 위하여 과거의 데이터를 1시간 간격으로 수집된 2년간 데이터를 수집하여 학습에 활용하고 다음 해 1년의 자료에 검증하여 모델의 정확성을 확인하고 이를 활용할 수 있도록 기초데이터를 형성한다. 연구의 범위는 2019~2021년 부산광역시 내부 28개소의 측정소에서 측정한 PM10, PM2.5의 미세먼지 수치를 활용한다. 미세먼지 수치를 예측하는 것에 있어 인공신경망 모델 검증과 도시계획 및 설계, 건축 등 다양한 분야와 결합할 수 있는 기초데이터를 형성하는 것을 연구 목적으로 둔다.

[1]47340 부산 부산진구 엄광로 176, 동의대학교 대학원 도시공학과 박사과정. E-mail: 15484@deu.ac.kr
[2]47340 부산 부산진구 엄광로 176, 동의대학교 건설공학부 도시공학전공 조교수. E-mail: lhk@deu.ac.kr
[3](교신저자)47340 부산 부산진구 엄광로 176, 동의대학교 건설공학부 도시공학전공 교수. E-mail: yoonsb@deu.ac.kr

2. 연구 방법 및 분석의 틀

본 연구에서는 데이터의 객관성을 확보하기 위하여 공공데이터포털(data.go.kr)에서 제공하는 Open API를 활용한다. 공공기관에서 제공하는 데이터는 장비의 정확성과 현재 민간에 제공되는 데이로 신뢰성을 확보할 수 있어 데이터의 활용이 자유롭다. 데이터를 정제하기 위하여 Excel - 데이터 가져오기를 하여 CSV로 저장한다.

예측모델은 R Studio 4.2.2버전에서는 neuralnet 패키지를 활용한다. neuralnet 패키지는 다른 인공신경망 패키지보다 시각화가 쉬우며 plot 함수에 결과 모델 객체의 정확한 값이 도출되기 때문에 사용이 편리하고 좀 더 정확한 결괏값이 도출된다. 공공데이터포털에서 제공하는 '한국환경공단_에어코리아_대기오염정보'에서 '측정소별 실시간 측정정보조회'를 활용하였다. 엑셀에서 가져오기를 하여 CSV로 변환·저장하였다. 전국 측정소·시간별로 제공되는 데이터를 2019~2021년 월별 데이터를 36개 파일로 변환하였으며 1년 단위로 통합하였다. 이후 공간적 범위에 맞도록 부산광역시에 위치한 지역, 측정소(28개소), PM10, PM2.5, 측정일시, 도시대기로 정리하였다.

	A	B	C	D	E	F	G	H	I	J	K	
1	지역	망	측정소코드	측정소명	측정일시	SO2	CO	O3	NO2	PM1	PM2	주소
2	부산 중구	도시대기	221112	광복동	2019010101	0.003	0.3	0.015	0.025	31	19	부산 중구 광복로 55번길 10
3	부산 중구	도시대기	221112	광복동	2019010102	0.003	0.3	0.018	0.021	33	21	부산 중구 광복로 55번길 10
4	부산 중구	도시대기	221112	광복동	2019010103	0.003	0.3	0.016	0.022	29	19	부산 중구 광복로 55번길 10
5	부산 중구	도시대기	221112	광복동	2019010104	0.003	0.4	0.015	0.023	36	20	부산 중구 광복로 55번길 10
6	부산 중구	도시대기	221112	광복동	2019010105	0.004	0.4	0.021	0.018	44	20	부산 중구 광복로 55번길 10
7	부산 중구	도시대기	221112	광복동	2019010106	0.004	0.3	0.025	0.015	37	23	부산 중구 광복로 55번길 10
8	부산 중구	도시대기	221112	광복동	2019010107	0.003	0.3	0.027	0.013	29	21	부산 중구 광복로 55번길 10
9	부산 중구	도시대기	221112	광복동	2019010108	0.003	0.3	0.028	0.011	36	22	부산 중구 광복로 55번길 10
10	부산 중구	도시대기	221112	광복동	2019010109	0.003	0.3	0.027	0.012	26	18	부산 중구 광복로 55번길 10
11	부산 중구	도시대기	221112	광복동	2019010110	0.003	0.3	0.027	0.014	31	19	부산 중구 광복로 55번길 10
12	부산 중구	도시대기	221112	광복동	2019010111	0.003	0.3	0.029	0.013	27	16	부산 중구 광복로 55번길 10
13	부산 중구	도시대기	221112	광복동	2019010112	0.003	0.3	0.03	0.011	27	15	부산 중구 광복로 55번길 10
14	부산 중구	도시대기	221112	광복동	2019010113	0.003	0.3	0.031	0.01	32	15	부산 중구 광복로 55번길 10

Figure 1. Data collection process

PM10, PM2.5 측정값에 미입력된 값(오류, 전송오류 등)등 결측값을 제거하는 과정을 진행하였다. 28개 측정소에서 245,280개의 측정값이 있으며 제거된 최종값은 165,176개의 값을 본 연구에서 표본으로 활용하였다. 연구 과정으로는 ① 가져온 데이터를 정리를 한다. ② 데이터를 패키지에 적용할 수 있도록 정규화, 결측값을 실시 및 제거한다. ③ 데이터를 학습용과 테스트용으로 분리를 한 후 ④ 입출력 변수를 지정한다. ⑤ 인공신경망에 학습용 데이터를 무작위로 학습을 실시하고 ⑥ 테스트 값에 적용한 후 오차를 검정한다.

3. 분석 및 결과

3.1. 데이터 분석

3.1.1 데이터 가져오기 및 정규화

R Studio 4.2.2버전을 활용하여 인공신경망 패키지와 결괏값을 엑셀파일로 저장할 수 있는 openxlsx 패키지를 설치한다. 데이터 정리가 완료된 엑셀파일을 R Studio에 가져온다. 가져온 데이

터가 결측값을 검사하여 무응답데이터를 확인한다. 각 열을 기준으로 데이터를 0~1사이로 정규화 한다. 정규화하는 식은 아래와 같다. 최댓값과 최솟값을 할당하여 0~1의 범위로 정규화 한다.

3.1.2 데이터 분리 및 입출력 변수 지정

전체 데이터 165,176개 표본 중 70~90%의 행을 학습용으로, 나머지를 테스트용으로 분리한다. 본 연구에서는 70%, 80%, 90%로 나누어 진행한다. 학습용 데이터를 무작위로 산출하기 위하여 난수를 사용하고 추후 난수를 활용하여 검증이 가능하다. train에서는 학습용 데이터, test는 테스트용 데이터로 할당한다. 입출력 변수들로 인공신경망 모형의 입출력 형식으로 지정한다. 형식은 종속변수(y)~독립변수(x)이며, 변수가 여러개인 경우 +로 연결한다. 본 연구에서 종속변수는 2021년의 자료이며 독립변수는 2019년, 2020년으로 두 개의 자료를 +로 연결하여 2021~2019+2020 형태로 지정한다.

3.1.3 학습 및 예측

5개의 입력 노드와 은닉층, 1개의 노드 출력으로 실행한 인공신경망은 은닉층의 노드 수는 입력에서 출력 방향의 순서로 5개씩으로 가중치가 설정되어 있으며 각 6개의 학습하고 예측하였다. 테스트용 데이터 세트에서 입력 노드의 데이터 세트에 대응하는 출력을 예측하고 실제 값 크기의 비정규화 데이터로 복원한다. 이후 예측값들의 오차를 측정한다. test와 학습된 모형 값을 비교하고 오차를 검증한다.

3.1.4 검정

모든 학습 데이터에서 66.459%의 MAPE(평균절대백분율오차, 절댓값을 사용하여 양수 및 음수 오차가 서로를 상쇄하지 않도록 하고 상대 오차를 사용하여 시계열 모델 간에 예측 정확도를 비교할 수 있게 하는 상대 오차 측정항목)가 나타났으며 입력 변수의 수를 변경할 경우 오차 값이 변경된다.

3.2. 분석 결과

학습용 데이터를 80%로 지정하였을 때 학습단계(Steps)이 가장 높았으며 70%, 90%는 낮은 값을 알 수 있었다. PM2.5에서 90%의 학습용 데이터를 지정하였을 경우 계산 불가(weights were not calculated)의 오류가 발생하였다. 66.459%의 MAPE가 나타났다. 오차 범위가 66%로 데이터 간의 오차가 높은 것으로 나타났다.

Table 1. Artificial Neural Network Analysis Results

Sortation	PM10 70%	PM10 80%	PM10 90%	PM2.5 70%	PM2.5 80%	PM2.5 90%
Error	228.115568	90.494842	100.687137	228.113116	257.923995	Weights were not calculated
Steps	1,531	6,432	1,637	761	3,741	

4. 결론 및 한계

본 연구에서는 2019~2021년 3년간의 데이터를 활용하였으며, 미래 예측에 인공신경망 모델을 활용하였다. 효과적인 인공신경망 모델의 학습 값으로 훈련용 데이터의 비율 80%, 테스트 데이터의 비율 20%가 적절한 비율로 나타났으며, PM10에서 6,432번의 학습, 90.494842의 Error, PM2.5에서는 3,741번의 학습, 257.923995의 Error가 나타났다. 데이터의 학습 값과 2021년의 값을 비교하여 오차 값 검증을 하였다. 검증은 오차를 비교하는 방법을 활용하였으며 오차율은 66%로 나타나 본 연구의 한계점으로 발생하였다. 학습용 데이터를 80%로 지정하였을 때 학습 단계(Steps)이 가장 높았으며 70%, 90%는 낮음을 알 수 있었다. PM2.5에서 90%의 학습용 데이터를 지정하였을 경우 계산 불가(weights were not calculated)의 오류가 발생하였다. 가중치를 계산할 수 없으므로 입력층과 은닉층의 비율을 감소하거나 표본을 정리하여 계산 불가와 Error, 오차를 낮추는 방법이 필요하다.

References

Byun, J, Y., Jun, C. H. (2021). Prediction of Air Quality Based on Spatio-Temporal Neural Networks from Seoul, Korea, *KSCE 2021 CONVENTION*, 279-280. (in Korean).

Cho, K. W., Jung, Y. J., Kang, C. G, Oh, C. H. (2019). Conformity Assessment of Machine Learning Algorithm for Particulate Matter Prediction, *Journal of the Korea Institute Of Information and Communication Engineering*, 23(1), 20-26. (in Korean). https://doi.org/10.6109/jkiice.2019.23.1.20

Sung, S. H., Kim, S. J., Ryu. M. H. (2020). A Comparative Study on the Performance of Machine Learning Models for the Prediction of Fine Dust: Focusing on Domestic and Overseas Factors, *Innovation studies*, 15(4), 339-357. (in Korean). https://doi.org/10.46251/INNOS.2020.11.15.4.339

Jung, Y. J., Cho, K. W., Kang, C. G, Oh, C. H. (2018). Particulate Matter Prediction Model using Artificial Neural Network, *sdsd*, 22(2), 0623-0625. (in Korean).

Proceedings of the Korean Data Analysis Society
January 26-27, 2023, pp. 203-206

한국 인공지능 학술논문 저자들 연결 네트워크 분석

양희건[1], 양희성[2], 양희창[3], 강은미[4]

요 약

인공지능분야 학술논문의 최근 연구동향 분석을 위해, 사회연결망 분석방법을 활용하여 국내 연구자들의 공저자 관계를 연구하였다. 연구범위는 인공지능 분야에서 게재된 2001년부터 2022년까지 총 6,398편의 한국학술지인용색인(KCI, Korea Citation Index) 학술논문이다. 사회연결망 분석을 위해 NodeXL Pro 프로그램을 사용하여 연구자간 관계를 분석하였다. 본 연구를 통해 2016년부터 인공지능 연구가 급속히 증가한 것을 알 수 있고, 2020년부터는 1천편이 넘는 논문이 게재되는 것으로 나타났다. 인공지능 분야 공저자 관계를 분석한 결과에서 노드 연결정도(Degree) 평균은 1.525, 네트워크 밀도(Density)는 0.0002, 노드간 평균 거리(Distance)는 3.216로 특정 연구자간의 공저자 관계를 확인할 수 있었다. 향후 인공지능 분야 공저자간의 상호관계 연결고리, 공저자간 상호인용에 대해서도 심층분석이 필요하다고 생각한다.

주요용어 : 사회연결망 분석, 네트워크모형, 빅데이터, 인공지능, 연구동향, NodeXL.

1. 서론

사회연결망 분석(social network analysis)는 사회관계 또는 네트워크 연구에서 행위자간의 관계를 수치로 표현하고, 상세 분석을 제시하며, 이해하기 쉽게 그래프로 관계를 보여주는 연구방법이다 (Burt, 2004; Hanson, et al., 2019; Wasserman & Faust, 1994). 사회과학을 포함해서 최근 사람을 매개로한 확산이 이루어지는 코로나로 인한 상황에서 감염정도를 파악하는데에 이르기까지 다양한 학문분야에서 사용되고 있다(Cho, 2020; Newman, 2001; Yan & Ding, 2009; Yu et al., 2013). 특히 최근 4차 산업혁명이 확산되면서 휴대폰의 일상화는 각 개인이 서로 네트워크로 연결될 수밖에 없는 상황에서 더욱 활발히 이용되고 있다.

본 연구에서는 ChatGPT처럼 인공지능의 발달로 다른 분야에 비해서 가장 관심이 뜨거운 인공지능 관련 연구자들의 관계를 살펴보기 위해 인공지능 분야 학술논문을 게재한 연구자들의 공저관계를 분석하였다. 최근까지 인공지능 분야가 어떻게 발전해왔는지, 어떤 방식으로 연구자들이 서로 연결되어 있는지를 살펴보는 것은 인공지능분야 연구동향을 파악하는데도 유용할 뿐만 아니라 향후 공동연구가 더욱 활발해질 수밖에 없는 인공지능 분야의 연구자간의 관계를 예측해 보는데도 기초자료로 활용될 수 있을 것이다.

[1]03722 서울시 서대문구 연세로 50, 연세대학교 기계공학과 석박과정. E-mail: vnfms5004@gmail.com
[2](교신저자)13574 경기도 성남시 분당구 새마을로 177번길 81, 국군수도병원 중위.
E-mail: brianyang955@naver.com
[3]04542 서울시 중구 청계천로 100, Open Future Labs 연구원. E-mail: heechangyang11@gmail.com
[4]07238 서울시 영등포구 국회대로 70길 15-1, SeeArt 대표. E-mail: irisesem@gmail.com

2. SNA 분석방법

본 연구에서 인공지능 분야 연구자간의 관계를 살펴보기 위해 인공지능 키워드를 사용해 검색한 한국학술지인용색인(KCI, Korea Citation Index)에 게재된 2001년부터 2022년까지 총 6,398편의 논문을 대상으로 분석하였다. 최종분석에서는 단독저자를 제외한 3,293편의 논문을 대상으로 SPSS와 NodeXL Pro 분석 프로그램을 사용해서 사회연결망 분석을 진행하였다.

3. SNA 공저자 네트워크 분석 결과

3.1 인공지능 학술논문 추이

인공지능 분야 학술논문은 2001년부터 2015년까지는 연간 50편 미만으로 매우 적게 게재된 것으로 나타났으나, 2016년부터는 매년 2배 정도 논문 편수가 급속히 증가하는 모습을 보이고 있다. 이런 경향은 GPU 성능향상과 최근 머신러닝과 딥러닝 학습모델의 본격적으로 응용되는 추세와도 관련이 있다.

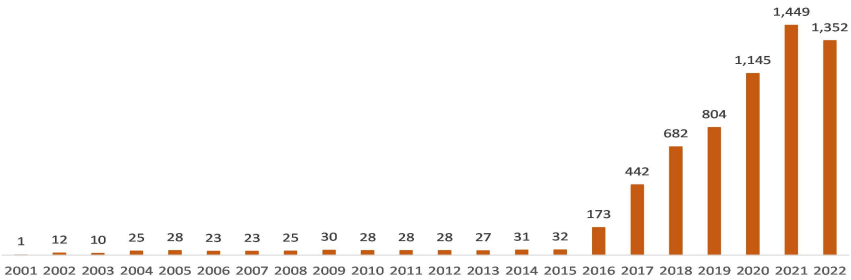


Figure 1. Trend of research paper related to artificial intelligence

또 본 연구에서 분석대상인 논문은 단순히 컴퓨터공학이나 소프트웨어공학에서만 인공지능 연구들이 제한되지 않고 전체 논문 중에서 사회과학과 복합학 분야에서도 인공지능 관련 학술논문이 많이 게재된 것을 알 수 있다. 이런 추세는 인공지능 분야가 최근 광범위하게 다학제적 성격을 띠고 있는 것과 관련이 있는 것으로 보인다.

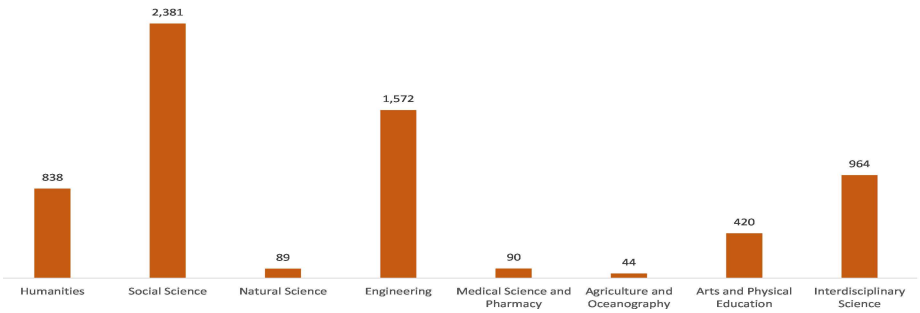


Figure 2. Distribution of research paper related to artificial intelligence by subfields

3.2 인공지능 학술논문 SNA 분석결과

인공지능 학술논문을 NodeXL 프로그램을 이용해서 분석해 보면, 공저자간 긴밀한 연결이 학술논문에서 파악되었다. 가장 많이 공저 관계를 보인 연구자는 김수환(충신대학교)였으며, 다음으로 이재호(경인교육대학교), 김기현(한국과학기술정보연구원) 등은 20편 이상 공저관계를 보여주고 있었다. 인공지능 학술논문에서 10편 이상의 공저 관계를 보이는 연구자는 23명으로 나타났다.

공저자 사회연결망 분석을 통해 공저자간 연결정도를 파악해 보면, 네트워크 노드 간 평균 거리는 3.216에 이르고 있다. 또 네트워크 밀도는 0.0002였고, 네트워크 긴밀도를 보여주는 연결정도는 1.525로 나타났다. 다양한 학문분야에서 각 연구자간의 공동저자 네트워크가 활발히 이루어지고 있는 것을 알 수 있다.

인공지능 분야에서 학술논문은 앞으로도 점점 많이 게재될 가능성이 높고, 실제 이런 논문의 저자들이 어떤 관계가 있는지 알아보는 것은 앞으로도 더욱 의미가 있을 것이다. 앞으로 국내 인공지능 분야 학술연구 공동체의 현재 모습을 파악하는 과정을 통해 미래의 연구경향을 예측할 수 있기 때문에, 공저자 관계에 대해 좀더 세부적으로 공저자간의 상호관계 연결고리 그리고 공저자간 상호인용에 대해서도 심층분석이 이루어질 필요가 있다.

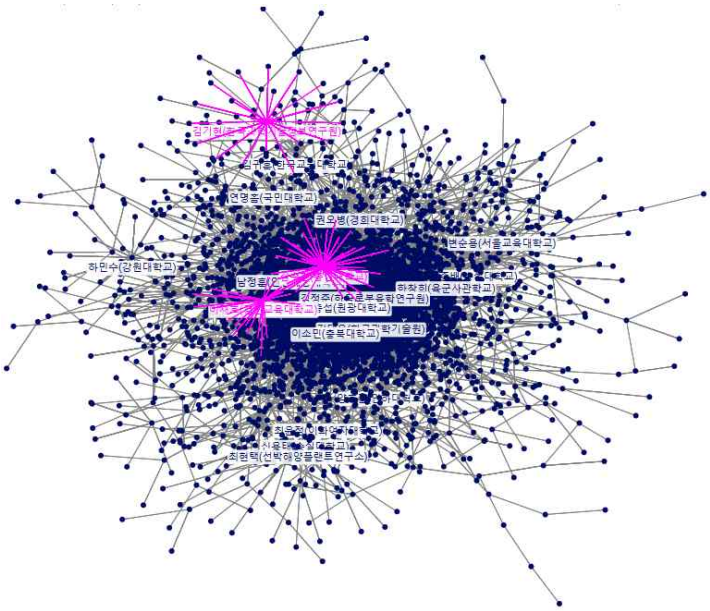


Figure 3. Coauthorship network on artificial intelligence

Table 1. Statisics of coauthorship network on artificial intelligence

Metric	Value
Number of nodes	7,833
Number of links	19,388
Density	0.0002
Average degree	1.525
Average distance	3.216

References

Burt, R. S. (2004). Structural holes and good ideas. *American Journal of Sociology, 110*(2), 349-399.

Cho, J. (2020). Analysis of open access status of domestic author's papers published in international journals: Based on highly cited papers. *Journal of the Korean Society for Library and Information Science, 54*(1), 325-341.

Hanson, D., Shneiderman, B., Smith, M. A., Himelboim, I. (2019). *Analyzing social media networks with NodeXL: Insights from a connected world(2nd ed),* Boston: Morgan Kaufmann.

Newman, M. E. J. (2001). Scientific collaboration networks. II: Shortest paths, weighted networks, and centrality. *Physical Review E, 64*(1), 016132.

Wasserman, S., Faust K. (1994). *Social network analysis: Methods and applications*, New York: Cambridge University Press.

Yan, E., & Ding, Y. (2009). Applying centrality measures to impact analysis: A coauthorship network analysis. *Journal of the American Society for Information Science and Technology, 60*(10), 2107-2118.

Yu, Q., Shao, H., He, P., & Duan, Z. (2013). World scientific collaboration in coronary heart disease research. *International Journal of Cardiology, 167*(3), 631-639.

Proceedings of the Korean Data Analysis Society
January 26-27, 2023, pp. 207-207

모집단이 부분층화 가능한 경우의 비추정치

(Ratio estimation for partially stratifiable population)

손경호[1]

요 약

본 연구에서는 비추정치의 효율을 높이기 위한 기본 조건들에 대해서 검토하고, 이러한 조건들이 개선되었을 때 실제로 비추정치의 신뢰도가 향상되는지를 검토하였다. 또한, 비추정치의 신뢰도 제고를 위한 방법으로 추가적인 보조변수를 활용하여 모집단의 일부만을 층화 했을 때의 비추정치 개선 정도도 확인해보았다. 연구에 활용한 데이터는 통계청 승인통계로 산림청에서 실시한 2020년 임산물생산비조사의 지역별 밤 생산량 조사 데이터를 사용하였고, 밤 재배 임가의 지역, 재배면적, 수령, 생산량 정보를 활용하여 재배면적 규모에 따른 층화를 실시하였다.

국가통계로 진행되는 임산물생산비조사와 동일한 층화 기준을 적용하여 각 세부 층별 재배면적과 생산량의 관계를 산점도와 상관계수를 통해 파악하고, 밤 생산량 추정치의 RSE를 검토하였다. 이를 통해 지역별로 4개 구분된 층화 기준을 2개로 줄이고, 각 층에서 비추정치의 오차를 증가시키는 극단치를 제거하였다. 이 결과 충남, 경남 및 기타 지역에서 비추정치가 개선되는 것을 확인하였다. 전남 지역에서는 층화 기준 변경, 극단치 제거에도 불구하고 비추정치가 개선되지 않았는데, 이는 재배면적 뿐 아니라 추가적인 요인이 생산량에 영향을 미치는 것으로 예상할 수 있었다. 따라서 추가적인 보조변수인 밤나무의 수령 정보를 활용하여 전남 지역에서만 추가적인 층화를 실시하였다. 수령에 따른 추가 층화는 밤나무의 수령이 최대 생산량을 내는 수령에 진입했는지 여부만을 구분할 수 있도록 20년 미만과 20년 이상으로 구분하였다. 이렇게 전남 지역에서만 일부 추가 층화를 실시하여 다시 비추정치를 구하고 기존의 비추정치와 비교해 본 결과 보조변수를 활용하여 추가 층화를 실시했을 때의 비추정치가 그렇지 않을 때에 비해 개선되는 것을 확인하였다.

주요용어 : 비추정치 개선, 층화기준 조정, 수령 변수 활용, 부분층화.

[1]서울 서초구 반포대로24길 76 월드빌딩 5층, 한신대학교 일반대학원 빅데이터융합학과 석사과정, ㈜케이스탯 리서치 연구2본부 2팀장. E-mail: khson@kstat.co.kr

Proceedings of the Korean Data Analysis Society
January 26-27, 2023, pp. 209-212

서울시 전기차 충전기 입지 선정 최적화*

박진아[1], 김혁구[2], 전종준[3]

요 약

본 연구는 서울시 공공데이터와 서울 데이터 펠로우십을 통해 제공받은 서울시 행정 데이터를 이용하여 전기차 수요와 전기차 충전 수요를 분석하였다. 급속 충전기와 완속 충전기를 나누어 설치 장소에 대한 충전기 설치 확률과 예상 충정량을 분류 모형과 회귀 모형으로 각각 학습한다. 다음으로 서울시의 총 전기차 충전기 설치 예산을 추정하여 예산 제약 하의 최대 충전량을 산출하고, 종류별 충전기 개수와 설치 장소를 선정한다. 해당 연구를 통해 서울시에서는 향후 전기차 충전기 선정에 있어 효율적 의사결정이 가능할 것이라 기대한다.

주요용어 : 서울시, 전기차, 충전량 예측, 최적화, 급속, 완속.

1. 서론

서울시는 탄소중립 정책으로 전기차 도입과 전기차 충전기 설치 정책을 추진하고 있다. 전기차 충전기는 매년 증가하지만 전기차의 증가 속도에 비해 더딘 상황이다. 전기차의 가파른 보급률에 맞춰 충전 수요를 분석하고, 부족한 충전 인프라를 보완하기 위해 충전 수요를 바탕으로 먼저 설치할 지역을 선정하여 충전기 보급 문제를 우선적으로 해결할 필요가 있다. 이에 본 연구는 서울시 공공 데이터와 서울 데이터 펠로우십을 통해 제공된 서울시 행정 데이터를 수집 및 가공하여 지리적 정보에 대한 분석을 기반으로 충전 종류별 전기차 충전기 이용량에 영향을 미치는 지역적인 특성을 파악하였다. 또한 분류 모형과 회귀 모형을 적합하여 충전기 설치 확률 기반의 충전량을 예측하였다. 또한 서울시의 전기차 충전기 예산을 추정하여 예산 하에서의 최대 예상 충전량을 산출하였다. 이를 활용해 전기차 충전기 설치가 시급한 지역에 우선적으로 설치하는 방안을 제시할 수 있다.

*이 논문은 제1저자 박진아의 석사학위논문의 축약본입니다. 이 논문은 2022년도 서울특별시의 재원으로 서울 데이터 펠로우십의 지원을 받아 작성된 논문입니다.
[1]02504 서울특별시 동대문구 서울시립대로 163, 서울시립대학교 통계학과 석사과정.
E-mail: jinah3316@naver.com
[2]02504 서울특별시 동대문구 서울시립대로 163, 서울시립대학교 도시빅데이터융합학과 석사과정.
E-mail: hyksun2015@gmail.com
[3](교신저자) 02504 서울특별시 동대문구 서울시립대로 163, 서울시립대학교 통계학과 부교수.
E-mail: jj.jeon@gmail.com

2. 변수 소개

본 연구에서 사용된 변수는 Table 1.과 같다. 건물연면적, 거주인구 등의 변수는 공공데이터를 통해 수집하였고, 자동차 등록현황, 충전기 충전이력 등은 서울시 등에서 제공받았다. 전기차 수요 변수의 경우 Bass 확산모형을 적용한 회귀분석을 이용해 신규 전기자동차 수요예측을 시행하고 전기차 수요 변수를 생성하였다. 그리고 교통량 변수의 경우 TOPIS 공공데이터를 활용해 인접한 격자의 교통량을 추정하였다.

Table 1. 변수 설명

구분	변수명	내용
자동차	대수(대)	용도별 연료별 차종별 자동차 등록현황
	전기차 수요(대)	전기차 수요 대수
건물	거리(m)	용도별 주요 시설물의 최단거리
	면적(㎡)	생활시설, 의료시설, 교육시설 등
인구	인구(명)	주거인구, 생활인구
교통량	교통량(대)	자동차 통행대수

3. 데이터 전처리

본 연구에서는 각 충전기별 지리적 정보를 부여하고자 100m 격자를 인덱스로 설정하고 다음과 같은 절차로 데이터를 가공하였다. 그리고 법정동/행정동으로 인덱싱된 데이터의 경우 동 전체 거주인구 대비 각 격자의 거주인구 데이터를 가중치로 추정하였다.

또한 지리적 인접성을 적용하고자 각 격자로부터 1km 반경 격자들의 거리별 차등 1km 반경 격자들의 거리별 차등 가중치 행렬을 생성하여 인접정보가 필요한 변수(전기차수요, 연령대별 주거인구, 용도별 건물연면적, 자동차 등록현황)에 해당 가중치 행렬을 적용하여 인접한 격자의 정보를 가중하였다.

4. 연구 방법

본 연구에서는 확률 기반의 충전량 예측 모형을 응용하여 추정된 예산 하에서 최대 충전량을 도출하는 최적화 식을 도출하여 입지를 선정하였다.

4.1. 확률 기반의 충전량 예측

충전기가 없는 지역의 전기차 충전량 예측을 위해 안승환(2016)이 제안한 모형을 적용하였다. 이 모형을 통해 충전기가 없는 지역에 대해 학습된 분류 모형으로 충전기 설치 확률을 부여하고, 회귀 모형으로 충전량을 예측하였다.

공공데이터포털(https://www.data.go.kr/index.do)에서 제공한 서울시내 전기차 충전기의 충전이력 정보를 보면 충전기별로 사용량에 차이가 있는 것으로 나타난다. 로그변환 충전량지표는 다봉분포 형태를 띄므로 다봉분포 가운데의 밀도가 작은 부분을 분류 기준으로 설정하여 분류 모형을 학습하였다. 임계값을 기준으로 사용량이 적은 충전기는 충전량이 0으로 가충전기, 그렇지 않으면 충전량이 0이 아니므로 실충전기라는 가정을 세웠다. 이 가정의 수식은 다음과 같다.

$$For\ any\ given\ X,\ E[Y|Z=0,X]=0,$$
$$where\ X\text{: 설명변수, } Y\text{: 충전량, } Z\text{: 충전소 유무}$$

가정을 통해 다음의 모형을 유도하였다.

$$E[Y|X] = E[Y|Z=1,X] \cdot P[Z=1|X]$$

LightGBM(Ke et al., 2017)을 통해 분류모형으로 실충전기일 확률을 계산하고 회귀모형으로 예측 충전량을 계산하여 곱한 결과, 실충전기일 확률을 기반으로 예측 충전량을 산출하였다.

4.2. 예산 제약 하의 최대 충전량 도출

확률 기반의 충전량 예측 모형에서 분류 모형 결과의 로짓에 가중치를 더하여 실충전기일 확률을 계산한다. 실충전기일 확률이 0.5 이상이면 충전기를 설치할 지역, 아니면 설치하지 않을 지역으로 구분하는 지시함수를 적용하여 확률을 이진으로 분류하고, 예상 충전량을 곱하였다. 이 모형을 통해 실충전기 여부에 따른 예상 충전량을 도출하였다. 수식은 다음과 같다.

$$\max_{t_1,t_2} \sum_{i=1}^{n} \left[I\left(\frac{1}{1+e^{-\hat{E}[y|Z=1,X_{1i}]+t_1}} \geq 0.5 \right) \cdot \hat{P}[Z=1|X_{1i}] + I\left(\frac{1}{1+e^{-\hat{E}[y|Z=1,X_{2i}]+t_2}} \geq 0.5 \right) \cdot \hat{P}[Z=1|X_{2i}] \right]$$
$$s.t. \sum_{i=1}^{n} \left[P_1 \cdot I\left(\frac{1}{1+e^{-\hat{E}[y|Z=1,X_{1i}]+t_1}} \geq 0.5 \right) + P_2 \cdot I\left(\frac{1}{1+e^{-\hat{E}[y|Z=1,X_{2i}]+t_2}} \geq 0.5 \right) \right] \leq C$$

이때 n은 데이터의 개수, x_{1i}는 급속충전기 격자의 데이터, x_{2i}는 완속충전기 격자의 데이터를 나타내며, $\hat{E}[y|Z=1,x_{1i}]$와 $\hat{E}[y|Z=1,x_{2i}]$는 급속 및 완속 실충전기 확률을 예측하는 모형 결과의 로짓이다. 또한 $\hat{P}[Z=1|x_{1i}]$와 $\hat{P}[Z=1|x_{2i}]$는 급속충전기와 완속충전기의 충전량을 각각 예측하는 모형이다. P_1과 P_2는 급속충전기와 완속충전기의 지원 가격, 그리고 C는 추정 예산이다. 예산 안에서 그리드 서치를 적용하여 t_1과 t_2를 계산하고, 이용량을 최대로 하는 t_1과 t_2를 산출하여 최적화 문제를 풀었다.

본 논문은 다양한 서울시 데이터를 활용하여 분석하고 비용 제약 하의 최대 충전량을 예측하였다. 이를 통해 충전 종류별 최적 설치 개수와 설치 장소를 제시하였다.

본 논문의 결과를 기반으로 제약된 예산 하에서 서울시가 전기차 충전량을 최대로 할 수 있는 입지선정이 가능할 것이라 기대된다.

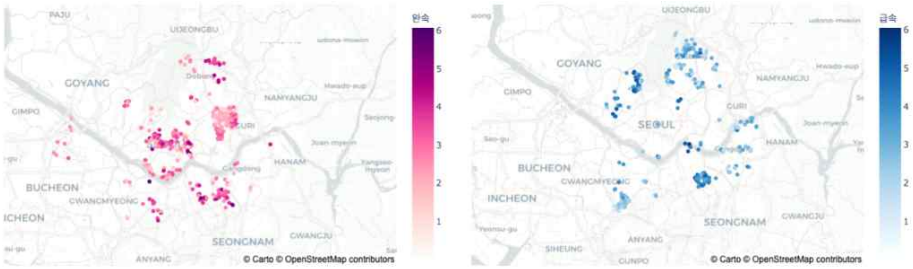


Figure 1. 급속 및 완속충전기 예상 충전량

References

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decisiontree. Advances in neural information processing systems, 30.

김경현 (2020). 교통량 데이터를 활용한 전기차 충전기 위치 최적화 방안연구. PhD thesis, 서울대학교 대학원.

김여진and허진(2019). 전기차 충전수요 분산을 위한 신규 전기차충전기 위치 선정에 관한 연구. *조명·전기설비학회논문지*, 33(5):28 – 34.

김예술, 박시온, and 박건웅 (2020). 서울 공공자전거 신규 대여소를 위한 수요량 예측 분석. *응용통계연구*, 33(6):739 – 751.

김장영 (2017). 서울시 최적의 전기자동차 충전기 위치 선정. *한국정보통신학회논문지*, 21(8):1575 – 1580.

김태경, 김장영, and 양윤기 (2016). 네비게이션 데이터를 바탕으로 한 서울시의 공공 전기차 충전기 위치. *Information Systems Review*, 18(4):1 – 15.

석인 (2019). 전기차 사용자의 충전기 이용에 영향을 주는 지리 환경적, 사회경제적 요인에 대한연구. *한국자동차공학회춘계학술대회*, pages 1197 – 1206.

심혜영 (2022). 서울시 전기차 보급을 위한 공공 급속충전기의 최적입지분석. *한국기후변화학회지*, 13(4):511 – 523.

안승환 (2016). 기상 데이터를 이용한 산불 피해규모의 예측방법 연구. *통계청논문공모전*, pages 67 – 98.

윤충식(2019).빅데이터 표준분석모델을 활용한 전기차 충전 인프라 우선설치지역 도출 사례연구. *한국통신학회학술대회논문집*, pages 23 – 26.

이순정, 손서은, and 박민혁 (2017). 국내 전기자동차 확산 전망과 영향. *한국혁신학회지*, 12(3):117 – 133.

이우범 and 황태성 (2019). 최대연결거리를 고려한 전기자동차 급속충전기 최적 위치선정 연구. *한국물류학회지*, 29(4):19 – 28.

홍진슬 and 홍성조 (2021). 전기차 충전기 입지의 공간적 형평성 분석. *대한건축학회논문집*, 37(12):211 – 221.

Proceedings of the Korean Data Analysis Society
January 26-27, 2023, pp. 213-213

초정밀 금형 프레스 공정 데이터 분석을 통한 이상 탐지 AI 모델 개발 및 성능 평가

조한별[1], 설현욱[2], 정현진[3], 백범수[4], 김영순[5]

요 약

IT 기술의 발전으로 모든 분야에서의 IT 융합이 보편화되면서 현대사회는 DNA(Data, Network, AI), ABCi(AI, Bigdata, Cloud, IoT) 등으로 일컬어지는 첨단 ICT 기술과의 융합이 가속화되고 있다. 이러한 흐름 속에서 제조업에서도 스마트 공장이라는 이름으로 ICT 기술과의 융합을 시도하고 있다. 이를 통해 단순히 제품 생산을 자동화하여 제품을 대량 생산하는 것뿐만 아니라 제품이 생산되는 과정으로부터 제조 데이터, 에너지 데이터, 물류 데이터 등을 자동으로 수집하고 분석하면서 서비스의 질을 높이고 있다. ICT 기술이 융합된 금형 프레스 공정은 금형의 내구도가 허락하는 한 많은 양의 제품을 빠르게 쏟아내는 양산에 특화된 공정 방식으로 공정과 동시에 공정 당시의 온·습도 데이터와 제품의 형상 이미지 데이터가 수집되고 있다. 본 연구는 머신러닝과 딥러닝 기법을 적용하여 금형 프레스 공정이 진행되면서 수집된 데이터를 분석하여 불량률을 감소시키는 모델을 구현하고자 하였다. 연구를 위해 온·습도 데이터는 1초 단위로 건물 내부, 건물 외부, 그리고 프레스 장비 내부에 대한 데이터를 35일 동안 약 230만 건의 데이터를 수집하였고, 이미지 데이터는 제품 하나의 이미지를 촬영하여 40일 동안 약 128만 건의 데이터를 수집하였다. 그리고 수집한 데이터를 적절히 전처리한 후 모델을 구현하기 위해 온·습도 데이터에 대해서는 LSTM을 적용하였으며, 이미지 데이터에 대해서는 이미지 처리를 위한 CNN을 적용하였다. 모델을 구현한 결과 온·습도 데이터에 대한 모델은 다소 아쉬운 성능을 보였으나 이미지 데이터에 대한 모델은 좋은 성능을 보였음을 확인하였다.

주요용어 : ICT, AI, 머신러닝, 딥러닝, 이상 탐지.

[1]52858 경남 진주시 진주대로 501, 경상국립대학교 정보통계학과 석사과정. E-mail: hanbyeoljo98@gmail.com
[2]52858 경남 진주시 진주대로 501, 경상국립대학교 정보통계학과 석사과정. E-mail: seolleim_@naver.com
[3]52858 경남 진주시 진주대로 501, 경상국립대학교 정보통계학과 석사과정. E-mail: hyunjin.jugn27@gmail.com
[4]52858 경남 진주시 진주대로 501, 경상국립대학교 정보통계학과 학석사통합과정.
E-mail: qorqjatn1388@gmail.com
[5]52858 경남 진주시 진주대로 501, 경상국립대학교 정보통계학과 조교수. E-mail: youngsoonkim@gnu.ac.kr

Proceedings of the Korean Data Analysis Society
January 26-27, 2023, pp. 215-217

장애인 거주시설 물리치료사들의 업무 만족도와 시간 외 근무시간 정도

노효련[1]

요 약

장애인 거주시설에 근무하는 물리치료사들의 시간외 근무시간과 업무 만족도에 대하여 알아보았다. 장애인 거주시설에 근무하는 전국의 물리치료사 117명을 대상으로 설문조사를 실시하였다. 월 16~20시간 정도의 시간외 근무를 하는 경우가 42.7%로 가장 많았고 업무에 대한 매우 만족과 만족이 65.8%로 나타났다. 물리치료사의 연령이 높을수록 업무 만족도는 높은 것으로 나타났다. 따라서 시간외 근무시간이 장애인 거주시설에서 근무하는 물리치료사들의 직무 만족도에 영향을 주지 않는 것으로 보인다.

주요용어 : 물리치료사, 시간외 근무시간, 업무 만족도, 장애인 거주시설.

1. 서론

장애인 거주시설은 장애인 복지법에 의해 설립된 장애인을 위한 생활시설이다. 장애인 복지법에서는 거주시설의 장애인들을 위하여 재활치료 서비스를 하도록 되어 있어서 물리치료사 또는 작업치료사들이 장애인 거주시설에서 근무하고 있다. 이들 물리치료사들은 소수를 차지하고 있고 근무환경 자체도 다수의 물리치료사들이 근무하고 있는 의료기관과 상당히 상이하다. 의료기관과 달리 가정처럼 24시간을 생활하는 시설이다 보니 공휴일이나 기념일이 되면 생활인들을 위한 행사나 프로그램 준비 및 시행에 참여하게 된다. 이와 같은 업무 등으로 물리치료사들은 물리치료 업무 외에도 다른 업ㅁ로 인하여 시간외 근무를 하는 경우가 있다.

이에, 본 연구에서는 장애인 거주시설에 근무하는 물리치료사들의 업무만족도와 시간외 근무시간 정도에 대하여 알아보고자 한다.

2. 연구 방법

본 연구는 장애인 거주시설에 근무하는 물리치료사들을 대상으로 하여 구조화된 설문지를 이용하여 자료를 수집하고자 한다. 일반적인 특성, 업무 만족도, 물리치료 외 업무 시간을 알아보았다. 직무 만족도는 5점 리커트(Likert) 척도를 사용하였으며, 1점은 “전혀 아니다”, 2점은 “아니다”, 3점은 “보통이다”, 4점은 “그렇다”, 5점은 “매우 그렇다”로 가장 높은 만족도는 5점으로 하였다.

성별과 업무 만족도와 시간외 근무시간은 교차분석을 실시하였고 업무 만족도에 영향을 주는 요인에 대하여서는 회귀분석을 실시하였다.

[1]25940 강원도 삼척시 도계읍 황조길 346 강원대학교 물리치료학과 교수. E-mail: withtry@kangwon.ac.kr

3. 연구 결과

1) 성별과 시간 외 근무시간

월 시간외 근무시간에서는 성별에서는 차이가 나타나지 않았으며, 월 16~20시간 정도의 시간외 근무를 하는 경우가 42.7%로 가장 높았고 월 21~25시간 외 근무는 18.8%로 나타났다(Table 1).

Table 1. Gender and working hour (N(%))

	Hour	Male	Female		χ2
Working hour(month)	0	0(0)	4(3.4)	4(3.4)	.094
	1~5	0(0)	1(0.9)	1(0.9)	
	6~10	6(5.1)	9(7.7)	15(12.8)	
	11~15	2(1.7)	2(1.7)	4(3.4)	
	16~20	18(15.4)	32(27.4)	50(42.7)	
	21~25	15(12.8)	7(6.0)	22(18.8)	
Total		49(41.9)	68(58.1)	117(100)	

2) 성별과 업무 만족도

성별에 따른 업무 만족도 정도는 통계학적으로 차이가 나타나지 않았다. 업무에 만족 45.3%, 보통이다가 29.9%, 매우 만족 20.5% 순이었으며 만족과 매우 만족이 65.8%로 과반수 이상이 만족한다로 나타났다(Table 2).

Table 2. Gender and job satisfaction (N(%))

	Male	Female	Total	χ2
Very dissatisfaction	1(0.9)	0(0)	1(0.9)	.586
Dissatisfaction	1(0.9)	3(2.6)	4(3.4)	
Moderate	13(11.1)	22(18.8)	35(29.9)	
Satisfaction	22(18.8)	31(26.5)	53(45.3)	
Very satisfaction	12(10.3)	12(10.3)	24(20.5)	
Total	49(41.9)	68(58.1)	117(100)	

3) 업무 만족도와 시간외 근무시간의 상관관계

업무 만족도와 월시간 외 업무시간, 연봉, 나이와 상관관계를 알아보았다. 업무 만족도는 시간외 업무시간, 연봉과 관련성이 나타나지 않았고 연령과는 양의 상관관계($r=.233$, $p<0.01$)를 보였다. 따라서 물리치료사의 연령이 높을수록 업무 만족도는 높은 것으로 나타났다(Table 3).

4. 결 론

장애인거주시설에서 일하는 물리치료사들은 월 16~20시간 정도의 시간외 근무를 하는 경우가 가장 높았고 과반수 이상이 업무에 만족하는 것으로 나타났으며, 물리치료사의 연령이 높을수록 업무 만족도는 높은 것으로 나타났다.

Table 3. Correlation among job satisfaction, working hour and age

	Working hour/month	Salary	Age
Job satisfaction	0.004	-0.066	.233**

** p<0.01

References

National Law Information Center. Elderly welfare act. 2021.(in Korean).

S. M. Lee. & M. K. Choi. (2003). Relationships between Job Characteristics and Turnover Intention of the Physical Therapists of Hospitals. *The journal of Korean academy of physical therapist.* 10(2), 145-152.(in Korean).

Proceedings of the Korean Data Analysis Society
January 26-27, 2023, pp. 219-221

만성 뇌졸중 환자의 마비측과 보행기능 및 속도

김찬우, 노효련

요 약

뇌졸중 환자의 마비측과 보행 속도와 보행기능 정도를 알아보고자 한다. 뇌졸중 환자의 보행 속도를 알아보기 위하여 10m 보행검사를 실시하였고 기능적 보행지수를 이용하여 보행기능 정도를 알아보았다. 연구대상자는 만성뇌졸중 환자 121명이었다. 뇌졸중 환자의 마비측과 보행속도 및 보행기능 정도는 차이가 나타나지 않았다. 따라서 뇌졸중 환자의 보행은 마비측과는 관련이 없는 것으로 보인다.

주요용어 : 기능적 보행지수, 뇌졸중, 마비측, 10m 보행검사.

1. 서론

뇌졸중 발병 후 발생하는 편마비는 일반적인 증후군으로 대뇌반구(hemispere)의 손상측에 따라 우측 편마비 또는 좌측 편마비가 된다. 좌측 대뇌반구의 손상이 있을 경우에는 우 측편마비가 발생하고 우측 대뇌반구의 손상이 있을 경우에는 좌측 편마비가 발생하는데, 좌측 편마비들은 우측 편마비에 비해 균형 상실이 나타나기 쉽고 기능 회복은 일반적으로 우측 편마비보다 좌측 편마비가 나쁘다. 뇌졸중으로 인한 편마비이라 할지라도 손상 대뇌반구에 따라 나타나는 증상과 균형, 보행 등의 특성이 달라지므로 이에 본 연구에서는 뇌졸중 환자의 마비측과 10m 보행검사와 기능적 보행지수를 이용하여 보행 속도와 보행기능 정도를 알아보고자 한다.

2. 연구 방법

본 연구는 뇌졸중 환자의 환측과 10m 보행검사와의 관계를 알아보고자 한다. 10m 보행 검사는 노인의 균형능력을 측정하기 위해 개발된 검사방법으로 뇌출혈이나 뇌경색으로 인한 뇌졸중, 파킨슨 환자들의 보행평가와 신경학적 손상을 받은 환자의 보행능력과 수행능력의 정도를 평가하는데 사용된다(Duncan et al., 2002). 이 검사는 평가 방법이 간단하며 특별한 훈련이나 장비가 필요로 하지 않아서 상요하기가 용이하다. 10m 보행검사는 14m의 직선거리에서 처음과 마지막의 각 2m를 가속과 감속을 위한 거리로 설정하고 초시계로 중간 10m의 거리에 대한 보행시간을 측정하는데, 환자가 가장 안전하다고 느끼면서 걷는 속도를 측정하였다. 측정자간, 측정자내 신뢰도는 r=0.95~0.96으로 높은 수준으로 보고하였다(Dobkin, 2006). 기능적 보행 지수(FAC)는 기능적인 이동능력의 측정하기 위해 개발된 평가 도구로 뇌졸중 환자의 보행을 필요한 도움의 정도와 자립성에

25940 강원도 삼척시 도계읍 황조길 346 강원대학교 물리치료학과 교수. E-mail: withtry@kangwon.ac.kr

따라서 쉽고 빠르게 평가할 수 있는 효율적인 검사방법으로 최소 0점에서 최대 5점으로 6점 척도로서 보행이 불가능한 경우는 0점, 1인의 간헐적인 도움이 필요한 경우는 2점, 신체적 접촉 없이 지시 또는 관찰이 필요한 경우는 3점, 독립적으로 평지는 걸을 수 있으나, 불안한 평지나 계단을 걸을 때에 도움이 필요는 한 경우는 4점, 독립적인 보행이 가능한 경우는 5점으로 구성되어 있다. 점수가 높을수록 이동능력 정도가 좋은 것으로 판단한다.

본 연구의 대상자는 총 121명이다. 전체 연구대상자 중 우측 편마비는 59명(48.8%) 좌측 편마비가 62명(51.2%) 이다. 전체의 평균연령은 57.62±8.93세이었다.

3. 연구 결과

1) 성별과 마비측에 따른 보행 속도의 차이

성별에 따른 보행 속도는 통계학적으로 유의하지 않게 나타났다. 따라서 성별에 따른 보행속도의 차이는 나타나지 않는 것으로 보인다.

마비측에 따른 보행 속도는 통계학적으로 유의하지 않아서 손상부위에 따른 보행속도의 차이는 나타나지 않는 것으로 보인다 (Table 1).

Table 1. Gender and 10m speed (unit:sec)

		Mean	SD	t	p
Gender	Male	18.56	17.40	-1.693	0.093
	Female	28.24	51.27		
Affected side	Rt	19.68	16.57	-.870	0.386
	Lt	25.21	46.07		

2) 마비측에 따른 보행기능 정도의 차이

기능적 보행지수를 이용하여 마비측에 따른 보행 속도를 알아보았다. 신체적 접촉 없이 지시 또는 관찰이 필요한 경우인 3점을 차지한 경우가 30.6%로 가장 많았고, 보행이 불가능한 경우는 0점인 경우가 6.6%로 가장 적었으며 마비측에 따른 보행기능 정도에 대하여서는 통계학적으로 유의하지 않아서 손상부위에 따른 보행기능 정도의 차이는 나타나지 않는 것으로 보인다(Table 2).

Table 2. Affected side and 10m speed (unit:sec)

Score	0	1	2	3	4	5	Total	χ2
Rt	3	9	8	17	15	7	59	2.870
	5.10%	15.30%	13.60%	28.80%	25.40%	11.90%	100.00%	
Lt	5	6	7	20	20	4	62	
	8.10%	9.70%	11.30%	32.30%	32.30%	6.50%	100.00%	
Total	8	15	15	37	35	11	121	
	6.60%	12.40%	12.40%	30.60%	28.90%	9.10%	100.00%	

4. 결 론

뇌졸중 환자의 마비측과 보행 속도와 보행기능 정도를 알아본 결과 뇌졸중 환자의 마비측과 보행속도 및 보행기능 정도는 차이가 나타나지 않았다. 따라서 뇌졸중 환자의 보행은 마비측과는 관련이 없는 것으로 보인다.

References

Dobkin, B., Apple, D., Barbeau, H., Basso, M., Behrman, A., Deforge, D., & Harkema, S. (2006). Weight-supported treadmill vs over-ground training for walking after acute incomplete SCI. *Neurology*, 66(4), 484-493.

Duncan, P. W., Horner, R. D., Reker, D. M., Samsa, G. P., Hoenig, H., Hamilton, B., & Dudley, T. K. (2002). Adherence to postacute rehabilitation guidelines is associated with functional recovery in stroke. *Stroke*, 33(1), 167-178.

Holden MK, Gill KM, Magliozzi MR, et al. Clinical gait assessment in the neurologically impaired: reliability and meaningfulness. *Physical Therapy.* 1984;64(1):35-40.

Proceedings of the Korean Data Analysis Society
January 26-27, 2023, pp. 223-223

농식품 소비자 패널을 활용한 포도 소비행태 분석 - 수도권 소비자를 중심으로

이정명[1], 박건환[2], 김혜형[3], 이영순[4]

요 약

사회·경제적 환경의 변화에 따라 구매력을 갖춘 소비계층이 다양해지고 삶의 질에 대한 관심이 증가하면서 과일 소비에서도 소비구매 행태가 바뀌는 양상을 보이고 있다. 그 중 포도의 경우, 한국농촌경제연구원 농업전망에 따르면 폐원과 작목 전환으로 지속적인 감소세를 가지다가 최근 샤인머스캣 신규식재가 늘면서 2020년 증가로 전환되었다. 이에 따라 포도 재배 농가에서는 실제 소비자들의 구매행태 변화를 파악함으로서 판매 마케팅 전략을 수립할 수 있을 것이다. 본 연구는 농촌진흥청이 수집한 2010년부터 2020년까지의 농식품 소비자 패널 자료를 활용하여 수도권 소비자의 포도 소비행태를 연도별로 비교 분석하고 과일 소비트렌드 변화를 나타내었다. 수도권 가구당 연간 평균 포도 구입액은 53,277원, 구입횟수는 2020년 기준 4.1회로 나타났다. 연령대별 가구당 포도 구입액은 50대> 60대 이상> 40대> 30대 이하 순이며, 가구 소득수준별 구입액은 5분위> 4> 3> 2> 1 순으로 소득수준이 높을수록 포도를 많이 구매하는 정상재 임을 알 수 있었다. 포도 구입처 비율은 슈퍼마켓에서의 구매가 가장 많았으며, 전통시장에서의 구매횟수 비율은 꾸준히 줄어들었다. 샤인머스캣의 구입액이 2018년도부터 큰 폭으로 증가하여 2020년에는 6,657천원으로 가장 높은 지출이 나타났으며, 2020년 기준 샤인머스캣> 거봉> 청포도> 캠벨얼리 순으로 구입액이 높았다.

주요용어 : 농식품소비자패널, 과일소비, 포도소비, 소비행태.

[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: ghpark21@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 26-27, 2023, pp. 225-228

Product spacing estimation of exponential logistic distribution under multiply type hybrid censoring

Kyeongjun Lee[1], *Chankeun Park*[2]

Abstract

Under classical estimation set up, the maximum product spacings method is quite effective and several authors advocated the use of this method as an alternative to maximum likelihood method, and found that this estimation method provides better estimates than maximum likelihood estimates in various situations. In this paper, we propose the estimators of the parameter of the exponential distribution under multiply hybrid censoring scheme. First, we derive the maximum likelihood estimator and maximum product spacings estimator for the parameter of exponential distribution. Also, we derive the approximate maximum product spacings estimators for the parameter of exponential distribution using Talyor series expansion. And we compare the proposed estimators in the sense of mean squared error and bias under multiply hybrid censoring scheme. Finally, the validity of the proposed methods are demonstrated by a real data.

Keywords : maximum likelihood estimator, maximum product spacings estimator, multiply hybrid censoring scheme, exponential distribuiton.

1. Introduction

Consider a exponential distribution with pdf

$$g(x;\theta)=\frac{1}{\theta}e^{-x/\theta}, \qquad x \geq 0, \quad \theta > 0,$$

and cdf

$$G(x;\theta)=1-e^{-x/\theta}.$$

The estimation of the parameters of exponential distribution has been researched by researchers. Abd El-Raheem (2018) considered optimal plans and estimation of constant-stress accelerated life tests for the extension of the exponential distribution under type-I censoring.

[1]Assistant Professor, Department of Mathematics and Big Data Science, Kumoh National Institute of Technology, Gumi 39177, Korea. E-mail: leekj@kumoh.ac.kr
[2](Corresponding Author) Professor, Department of Data Informatics, Korea Maritime and Ocean University, Busan 49112, Korea. E-mail: chapark@kmou.ac.kr

Alshenawy et al. (2020) considered progressive type-II censoring schemes of extended odd Weibull exponential distribution with applications in medicine and engineering. Cho and Lee (2020) derived exact inference for an exponential parameter under generalized adaptive progressive hybrid censored competing risks data. Cho and Lee (2021) derived exact likelihood inference for a competing risks model with generalized type II progressive hybrid censored exponential data. Seong and Lee (2022) derived exact likelihood inference for parameter of exponential distribution under combined generalized progressive hybrid censoring scheme.

Under the hybrid censoring scheme, some units can be failed between two points of observation with exact times of failure of these units unobserved. Therefore, Lee et al (2014) introduced a multiply hybrid censoring scheme. A schematic representation of the multiply hybrid censoring scheme is presented in Figure 1. Under the multiply hybrid censoring scheme, $X_{a_i:n}$ denote the a_ith failure. $a_s = r + \sum_{i=1}^{s-1} l_i$ if $r + \sum_{i=1}^{s-1} l_{i<n}$, and $a_s = n$ if $r + \sum_{i=1}^{s-1} l_i \geq n$. $l_i = a_{i+1} - a_{i+1}$, where $i = 1,2,\cdots,s-1$. In Case II, $X_{a_d:n} < T < X_{a_{d+1}:n}$ means that the a_dth failure took place before T, and no failure took place between $X_{a_d:n}$ and T.

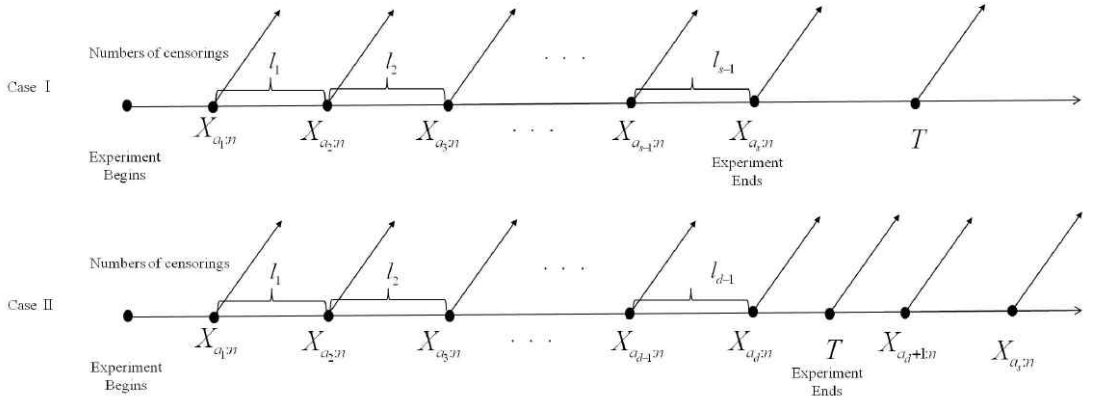


Figure 1. The multiply hybrid censoring scheme

Under classical estimation set up, the maximum product spacings method is quite effective and several authors advocated the use of this method as an alternative to maximum likelihood method, and found that this estimation method provides better estimates than maximum likelihood estimator in various situations (Cheng and Amin, 1983). Lee (2019) derived maximum product spacingsE and approximate maximum product spacings estimator for the parameter of a Rayleigh distribution under doubly hybrid censoring scheme. Cho and Lee (2019) derived maximum product spacings estimator and approximate maximum product spacingsE for the parameters of a Weibull distribution under progressive censoring scheme. Lee (2022a) derived maximum product spacingsE and approximate maximum product spacings estimator for the parameters of a Weibull distribution under progressive hybrid censoring

scheme. Lee (2022b) derived maximum likelihoodE, maximum product spacings estimator and approximate maximum product spacingsE for the parameter of half logistic distribution under progressive hybrid censoring scheme.

The main aim of this paper is to propose the estimators of the parameters of exponential distribution under multiply hybrid censoring scheme. First, we derive the maximum likelihood estimator and maximum product spacingsE for the parameter of exponential distribution. And we derive the approximate maximum product spacingsEs for the parameter of exponential distribution using Talyor series expansion. We also compare the proposed estimators in terms of the mean squared error and bias under multiply hybrid censoring scheme.

The rest of the paper is organized as follows. The maximum likelihood estimator and maximum product spacings estimator of the parameter of exponential distribution are presented in Section 2. Also, we derive approximate maximum product spacings estimator of the parameter of exponential distribution by using Taylor series expansion in Section 2. In Section 3, a Monte Carlo simulation of proposed estimators is carried out. Also, real data has been analyzed. And finally in Section 4, we conclude the paper.

2. Estimation

2.1 maximum likelihood estimator

Under multiply hybrid censoring scheme, the likelihood function is given by

$$\text{Case I. } L_I \propto \prod_{i=1}^{s} g\left(x_{a_{i:n}}\right)\left[1-G\left(x_{a_{i:n}}\right)\right]^{n-a_u} \prod_{i=2}^{s}\left[G\left(x_{a_{i:n}}\right)-G\left(x_{a_{i-1}:n}\right)\right]^{l_{i-1}},$$

$$\text{Case II. } L_{II} \propto \prod_{i=1}^{d} g\left(x_{a_{i:n}}\right)[1-G(T)]^{n-a_d} \prod_{i=2}^{d}\left[G\left(x_{a_{i:n}}\right)-G\left(x_{a_{i-1}:n}\right)\right]^{l_{i-1}}.$$

Therefore, Case I and Case II can be written as

$$L \propto \prod_{i=1}^{u} g\left(x_{a_{i:n}}\right)[1-G(W)]^{n-a_u} \prod_{i=2}^{u}\left[G\left(x_{a_{i:n}}\right)-G\left(x_{a_{i-1}:n}\right)\right]^{l_{i-1}}, \tag{1}$$

where $D = x_{a_{s:n}}$ and $u = s$ if Case I ($x_{a_{s:n}} < T$), and $D = T$ and $u = d$ if Case II ($x_{a_{s:n}} > T$).

Let $Z_{a_{i:n}} = X_{a_{i:n}}/\theta$ and then the variables $Z_{a_{i:n}}$ have a standard exponential distribution, we may rewrite the likelihood function as

$$L \propto \theta^{-u} \prod_{i=1}^{u} f\left(z_{a_{i:n}}\right)[1-F(W)]^{n-a_u} \prod_{i=2}^{u}\left[F\left(z_{a_{i:n}}\right)-F\left(z_{a_{i-1}:n}\right)\right]^{l_{i-1}}. \tag{2}$$

On differentiating the log likelihood function with respect to θ of Eq (2) and to zero, we obtain the estimating equation as

$$\frac{\partial \log L}{\partial \theta} = -\frac{1}{\theta}\left[u - \sum_{i=1}^{u} z_{a_{i:n}} - (n - a_{u)z_D} + \sum_{i=2}^{u} l_{i-1}\left\{\frac{f\left(z_{a_{i:n}}\right)z_{a_{i:n}} - f\left(z_{a_{i-1}:n}\right)z_{a_{i-1}:n}}{F\left(z_{a_{i:n}}\right) - F\left(z_{a_{i-1}:n}\right)}\right\}\right] = 0. \tag{3}$$

We can find the maximum likelihood estimator of θ that maximize the log likelihood function by solving Eq (3). We use the Newton-Raphson method to obtain the numerical solution of Eq (3).

References

Abd El-Raheem, A. M. (2018). Optimal plans and estimation of constant-stress accelerated life tests for the extension of the exponential distribution under type-I censoring, *Journal of Testing and Evaluation*, **47**, 3781-3821.

Alshenawy, R., Al-Alwan, A., Almetwally, E. M., Afify, A. Z. and Almongy, H. M. (2020). Progressive type-II censoring schemes of extended odd Weibull exponential distribution with applications in medicine and engineering, *Mathematics*, **8**, 1679.

Cheng, R. C. H. and Amin, N. A. K. (1983). Estimating parameters in continuous univariate distributions with a shifted origin, *Journal of the Royal Statistical Society B*, **45**, 394-403.

Cho, S. and Lee, K. (2019). Approximate maximum product spacing estimation for a Weibull distribution under progressive censoring, *Journal of the Korean Data Analysis Society*, **21**, 2203-2213.

Cho, S. and Lee, K. (2021). Exact likelihood inference for a competing risks model with generalized type II progressive hybrid censored exponential data, *Symmetry*, **13**, 887.

Cho, Y. and Lee, K. (2020). Exact inference for an exponential parameter under generalized adaptive progressive hybrid censored competing risks data, *Symmetry*, **12**, 2005.

Lee, K. (2019). Approximate maximum product spacings estimation for the parameter of a Rayleigh distribution under doubly hybrid censoring, *Journal of the Korean Data Analysis Society*, **21**, 1121-1129.

Lee, K. (2022a). Approximate maximum product spacing estimation of Weibull distribution under progressive hybrid censoring, *Journal of the Korean Data & Information Science Society*, **33**, 167-176.

Lee, K. (2022b). Product spacing estimation of half logistic distribution under progressive hybrid censoring, *Journal of the Korean Data & Information Science Society*, **33**, 337-346.

Lee, K., Sun, H. and Cho, Y. (2014). Estimation of the exponential distribution based on multiply Type I hybrid censored sample, *Journal of the Korean Data & Information Science Society*, **25**, 633-641.

Seong, Y. and Lee, K. (2022). Exact likelihood inference for parameter of exponential distribution under combined generalized progressive hybrid censoring scheme, *Symmetry*, **14**, 1764.

Proceedings of the Korean Data Analysis Society
January 26-27, 2023, pp. 229-233

엘라스틱 넷 벌점항을 이용한 비선형 지지벡터기계의 변수선택*

황호석[1], 최호식[2], 박창이[3]

요 약

지지벡터기계 (support vector machines)는 주어진 훈련 데이터를 이용하여 마진 (margin)을 최대화 하는 최적 초평면 (hyperplane)을 찾는 학습방법이다. 지지벡터기계는 다양한 문제에 적용이 가능하며 고차원데이터를 효과적으로 다룰수 있기 때문에 분류문제에서 폭넓게 사용된다. 그러나 지지벡터기계는 잡음변수 (noise variable)들이 많은 경우 예측력이 저하될 수 있고 선택된 최종 모형의 해석이 어렵다. 이러한 단점을 보완하고자 LASSO (least absolute shrinkage and selection operator) 벌점에 기반한 변수선택법을 고려할 수 있다. 회귀분석에서 변수들간에 강한 상관관계가 있는 경우에 그 변수들중에서 하나만 선택하는 LASSO 벌점의 문제점을 보완하기 위해 엘라스틱 넷 (elasticnet, EL) 벌점이 제안되었다. 본 논문에서는 비선형 분류문제의 변수선택법인 COSSO (component selection and smoothing operator)를 적용한 COSSO-SVM을 고려한다. 비선형 커널의 성분선택을 위해 LASSO 대신 엘라스틱 넷 벌점항을 적용한 EL-COSSO-SVM을 소개한다. 모의실험에서 제안한 방법이 COSSO-SVM보다 변수선택과 예측력 측면에서 성능이 향상됨을 보인다.

주요용어 : 커널, COSSO, LASSO.

1. 서론

지지벡터기계는 주어진 훈련 데이터를 이용하여 마진을 최대화 하는 최적의 초평면을 찾는 학습 방법으로, 모형의 변형이 쉬워 다양한 문제에 적용이 가능하며 고차원데이터를 잘 다룰수 있기 때문에 다양한 분류문제에서 폭넓게 사용된다. 그러나 Hastie 등 (2001)에 따르면 지지벡터기계는 잡음변수들이 많은 경우 예측력이 저하될 수 있고 선택된 최종 모형의 해석이 어렵다. 이러한 단점을 보완하고자 벌점항을 이용해 설명력이 높은 변수들을 선택하는 변수선택 방법들이 연구되어 왔다.

선형 문제에서는 LASSO와 같은 벌점을 계수에 직접 적용하면 입력변수들을 선택할 수 있는 반면, 비선형 문제에서는 이러한 벌점을 계수에 직접 적용하면 입력변수 대신 커널의 기저함수가 선택되는 난점이 있다 (Kim과 Park, 2013). 비선형 문제에서 지지벡터기계의 대표적인 변수선택법으로 COSSO를 적용한 Zhang (2006)의 COSSO-SVM이 있다. COSSO-SVM에서는 설명력 있는 함수성

*이 논문은 2022년도 서울시립대학교 교내학술연구비에 의하여 지원되었음.
[1]02504 서울 동대문구 서울시립대로 163, 서울시립대학교 통계학과 대학원생.
[2]02504 서울 동대문구 서울시립대로 163, 서울시립대학교 도시빅데이터융합학과 부교수.
[3](교신저자) 02504 서울 동대문구 서울시립대로 163, 서울시립대학교 통계학과 교수. E-mail: park463@uos.ac.kr

분을 선택하고자 대응되는 성분들의 계수에 LASSO 타입의 벌점항을 적용하는데, 성분들간 양의 상관관계가 있는 경우 특정 성분 하나만 선택될 수 있는 문제가 있다.

따라서 본 논문에서는 엘라스틱 넷 벌점항을 이용하여 COSSO-SVM 알고리즘을 개선하고자 한다. 본 논문의 구성은 다음과 같다. 2절에서는 엘라스틱 넷을 이용한 EL-COSSO-SVM 및 알고리즘에 대해 설명한다. 3절에서는 모의실험을 통해 COSSO-SVM과 EL-COSSO-SVM의 성능을 비교한다.

2. EL-COSSO-SVM 알고리즘

Wahba (1990)와 Gu(2002)에 따르면 SS-ANOVA 모형은 고차원 함수 추정의 틀을 제공하며, COSSO-SVM은 이러한 함수 추정의 틀에서 지지벡터기계의 분류함수 $f \in F$ 에 대하여 다음과 같은 함수 ANOVA 분해를 적용한다.

$$f(x) = b + \sum_{j=1}^{d} f_j(x_j) + \sum_{j<k} f_{jk}(x_j, x_k) + \ldots$$

여기서 b는 상수항이고 f_j, f_{jk}는 각각 변수 x_j의 주효과와 변수 x_j, x_k들의 이인자 교호작용을 나타내며 나머지는 고차항의 교호작용에 해당한다.

분류함수에 대응되는 함수공간 F 에대한 AVONA는

$$F = \{1\} \oplus \sum_{j=1}^{d} \overline{H_j} \oplus \sum_{j<k} (\overline{H_j} \otimes \overline{H_k}) \oplus \ldots$$

로 나타낼 수 있다. 여기서 $\overline{H_j}$는 상수공간 {1}에 직교하는 x_j에 의해 정의되는 RKHS의 부분공간이다. 고려 대상이 되는 부분공간들을 $F_v, v = 1,\ldots,q$로 나타내면 실제 함수공간은 다음과 같다.

$$F = 1 \oplus_{v=1}^{q} F_v$$

Zhang (2006)에서는 재정의된 F상에서 COSSO-SVM을 제안하였고 다음과 같은 최적화 문제를 풀게 된다.

$$\min_{f \in F, \theta} \ \frac{1}{n} \sum_{i=1}^{n} [1 - y_i f(x_i)]_+ + \frac{\lambda}{2} \sum_{v=2}^{q} \theta_v^{-1} \| P^v f \|^2 + \lambda_\theta \sum_{v=1}^{q} \theta_v \tag{2.1}$$

$$\text{subject to } \lambda > 0, \lambda_\theta > 0, \theta_v \geq 0, v = 1,\ldots,q.$$

여기서 λ, λ_θ는 벌점모수, $\| \cdot \|$ 은 F_v상의 커널 노음, P^v는 F_v상의 사영연산자 (projection operator), θ_v는 스케일 모수이다. 식 (2.1)은 스케일 모수에 LASSO 타입의 벌점을 적용함으로써 설명력있는 함수성분 (functional component)를 선택한다. 그러나 Zou 와 Hastie (2005)에 따르면 LASSO 벌점항은 변수간 상관관계가 높은 경우 상관된 변수중 하나의 변수만 선택하는 문제를 지적한 바 있다. 본 논문에서는 이러한 문제를 보완하기 위하여 식 (2.1)에 엘라스틱 넷 벌점항을 적용한 다음과 같은 최적화 문제를 다룬다.

$$\min_{f \in F, \theta} \ \frac{1}{n} \sum_{i=1}^{n} [1 - y_i f(x_i)]_+ + \frac{\lambda}{2} \sum_{v=2}^{q} \theta_v^{-1} \| P^v f \|^2 + \lambda_\theta (\gamma \sum_{v=1}^{q} \theta_v + (1-\gamma) \sum_{v=1}^{q} \theta_v^2) \tag{2.2}$$

여기서 추가적인 벌점모수인 $0 \le \gamma \le 1$ 는 LASSO 벌점과 능형회귀의 벌점간의 비중을 조절하며 $\gamma = 1$인 경우 COSSO-SVM 최적화 (2.1)와 일치하게 되며 식 (2.2)의 최적화 문제에 해당하는 분류 방법을 EL-COSSO-SVM이라 부르기로 한다.

3.모의실험 설계 및 분석

COSSO-SVM과 EL-COSSO-SVM의 예측력과 변수선택 측면에서의 성능을 비교하기 위하여 다음과 같은 세 가지 모의실험을 고려하였다.

1) 모의실험 A

$X_1,...,X_p$는 서로 독립적으로 N(0,1)에서 생성되며, $X = x$로 주어졌을 때 $Y = 1$의 조건부확률에 대하여 다음과 같은 로지스틱 모형을 가정한다.

$$\log\left(\frac{\mathbb{P}(Y=+1|X=x)}{\mathbb{P}(Y=-1|X=x)}\right) = \beta^T \Phi(x).$$

여기서 $\beta = (3,1.5,1,1,2,1,1,1)^T$이고 $\Phi(x) = (x_1,...,x_8)^T$이다. 이 모의실험의 경우 $X_1,...,X_8$은 설명 력 있는 변수들이고 나머지 $X_9,...,X_p$는 잡음 변수가 된다. 설명력 있는 변수들간의 쌍의 상관관계는 $corr(i,j) = 0.5^{|i-j|}, (i,j = 1,...,8)$이며 선형 문제로 선형 커널을 적용했다.

2) 모의실험 B

모의실험 A와 마찬가지로 $X_1,...,X_p$는 서로 독립적으로 N(0,1)를 따르며, $X = x$로 주어졌을때 $Y = 1$의 조건부확률은 (3.1)의 로지스틱 모형을 따른다고 하자. 단 $\beta = (\underbrace{2,...,2}_{5}, \underbrace{1,...,1}_{5})^T$이고 $\Phi(x) = (x_1,...,x_{10})^T$로 주어진다. $X_1,...,X_{10}$은 설명력 있는 변수이고 나머지 $X_{11},...,X_p$는 잡음 변수가 된다. 설명력 있는 변수들간의 쌍의 상관관계는 $corr(i,j) = 0.5, (i,j = 1,...,10)$이며 선형문제로 선형 커널을 적용했다.

3) 모의실험 C

$X_1,...,X_p$는 서로 독립적으로 N(0,1)에서 발생되며, $\boldsymbol{h} = (x_1,x_2,x_3,x_1x_2,x_1x_3,x_2x_3)^T$라 하자.비선형 성분들간의 상관관계를 주기 위하여 은닉층이 없고 입력과 출력층으로만 구성된 신경망 모형을 이용하여 데이터를 생성한다. 출력노드값은 다음과 같다.

$$o_i = (i + \exp(-w_i^T \boldsymbol{h}))^{-1}, \ i \in \{1,2\}.$$

여기서 가중치 벡터는 $w_1 = (2.1,1.4,-1.2,2.1,1.0,2.4)^T, w_2 = (-1.1,2.4,1.5,3.4,2.5,0.7)^T$이다. 출력 변수는 $y = \arg max_i o_i$이다. 이 모의실험에서는 출력변수 y와 관련된 변수들은 x_1,x_2,x_3이며 나 머지 변수는 잡음변수이다. y와 관련된 변수들간의 상관계수는$corr(i,j) = 0.3, (i,j = 1,...,3)$ 이며 비선형 문제로 가우스 커널을 적용했다.

각 모의실험에서 $p \in \{10,50,100,150\}$와 $n \in \{100,150\}$의 각 수준에서 n개의 훈련데이터와 10,000개의 시험데이터를 생성하였다. 조율모수인 $\lambda, \lambda_\theta, \gamma$ 값을 결정하기 위해서 검증오차에 대한 5-묶음 교차확인법 (5-fold crossvalidation)을 적용하였다. 또한 실험의 변동성을 측정하기 위하여 데이터 생성, 훈련, 시험데이터상의 평가의 전 과정을 100회 반복 실시하였다.

Table 1-3은 각 모의실험에서 COSSO-SVM과 EL-COSSO-SVM의 예측력 (평균 오분류율과 표준오차)과 변수선택 (평균 F_1 스코어와 표준오차) 측면의 성능을 비교한 결과이다. 모의실험 A 와 B는 선형문제로 선형 커널을 적용했고 모의실험 C는 비선형 문제로 가우스 커널을 사용했다.

세 모의실험에서 공통적으로 고정된 n에 대하여 p가 증가할수록 예측력이 감소하는 경향을 보이며, EL-COSSO-SVM이 예측력 뿐만 아니라 변수선택 측면에서 COSSO-SVM에 비하여 더 나은 성능을 보여주며 이는 엘라스틱 넷 벌점항의 그룹핑 효과(grouping effect)와 관련된 것으로 보인다. 모의실험 C의 경우 유의한 변수들로 이루어진 성분들과 노이즈 변수로 이루어진 성분들이 명확히 분리되는 반면에 모의실험 A와 B 에서는 커널에 의해 특성공간으로 매핑 (mapping)되는 과정에서 변수들이 뒤섞이며 유의미한 변수들이 명확히 구분되지 않는다. 따라서 모의실험 A와 B의 경우 고정된 p에 대해서 n이 커질수록 F_1스코어 값이 떨어지는 경우가 나타난다.

Table 1. Result for simulation A, B, C

-	n	p	Average missclassification error rate		Average F1-score	
			COSSO-SVM	EL-COSSO-SVM	COSSO-SVM	EL-COSSO-SVM
A	100	10	0.2513(0.0021)	0.2385(0.0016)	0.4140(0.0099)	0.4933(0.0371)
		50	0.2728(0.0037)	0.2425(0.0022)	0.3248(0.0193)	0.4681(0.0452)
		100	0.2995(0.0040)	0.2544(0.0029)	0.3103(0.0403)	0.4647(0.0328)
		150	0.3133(0.0046)	0.2574(0.0038)	0.2948(0.0217)	0.4357(0.0184)
	150	10	0.2412(0.0018)	0.2324(0.0012)	0.3876(0.0434)	0.5043(0.0232)
		50	0.2500(0.0027)	0.2349(0.0015)	0.3030(0.0235)	0.4339(0.0274)
		100	0.2703(0.0035)	0.2399(0.0018)	0.3089(0.0421)	0.4753(0.0342)
		150	0.2912(0.0033)	0.2428(0.0022)	0.2168(0.0326)	0.3829(0.0434)
B	100	10	0.2513(0.0021)	0.2385(0.0016)	0.4140(0.0099)	0.4933(0.0371)
		50	0.2728(0.0037)	0.2425(0.0022)	0.3248(0.0193)	0.4681(0.0452)
		100	0.2995(0.0040)	0.2544(0.0029)	0.3103(0.0403)	0.4647(0.0328)
		150	0.3133(0.0046)	0.2574(0.0038)	0.2948(0.0217)	0.4357(0.0184)
	150	10	0.2412(0.0018)	0.2324(0.0012)	0.3876(0.0434)	0.5043(0.0232)
		50	0.2500(0.0027)	0.2349(0.0015)	0.3030(0.0235)	0.4339(0.0274)
		100	0.2703(0.0035)	0.2399(0.0018)	0.3089(0.0421)	0.4753(0.0342)
		150	0.2912(0.0033)	0.2428(0.0022)	0.2168(0.0326)	0.3829(0.0434)
C	100	10	0.2513(0.0021)	0.2385(0.0016)	0.4140(0.0099)	0.4933(0.0371)
		50	0.2728(0.0037)	0.2425(0.0022)	0.3248(0.0193)	0.4681(0.0452)
		100	0.2995(0.0040)	0.2544(0.0029)	0.3103(0.0403)	0.4647(0.0328)
		150	0.3133(0.0046)	0.2574(0.0038)	0.2948(0.0217)	0.4357(0.0184)
	150	10	0.2412(0.0018)	0.2324(0.0012)	0.3876(0.0434)	0.5043(0.0232)
		50	0.2500(0.0027)	0.2349(0.0015)	0.3030(0.0235)	0.4339(0.0274)
		100	0.2703(0.0035)	0.2399(0.0018)	0.3089(0.0421)	0.4753(0.0342)
		150	0.2912(0.0033)	0.2428(0.0022)	0.2168(0.0326)	0.3829(0.0434)

References

Hastie, T., Tibshirani, R. and Friedman, J. H. (2001).The Elements of Statistical Learning, 2nd Ed,Springer,New York.

Kim, K. and Park, C. (2013). Comparison of feature selection methods in support vector machines. J*ournal of the Korean Data & Information Science Society*, 26, 131-139. (in Korean)

Zhang, H. (2006). Variable selection for SVM via smoothing spline ANOVA. *Statistica Sinica*, 16, 659-674.

Wahba, G. (1990). Spline methods for obsevational data, CBMS-NSF Regional Conference Series, *SIAM*, Philadelphia.

Gu, C. (2002). Smoothing spline anova models, Springer, New York.

Proceedings of the Korean Data Analysis Society
January 26-27, 2023, pp. 235-242

다단집락추출에서의 분산 추정량의 비교*

김유미[1]

요 약

본 연구에서 국가승인 통계 자료를 통해 임의추출방식을 활용한 다단집락추출의 분산추정량과 테일러급수 선형전개를 활용한 분산추정량에 실제 차이가 어느 정도 발생하는지를 확인하고, 다단집락추출을 적용한 설계의 분산추정량을 테일러급수 선형전개를 활용하여 산출하는 것이 적절한 것인지를 검증하였다. 자료는 조사구를 집락으로 활용하는 『국민여가활동조사(문화체육관광부)』의 로데이터를 활용하여 두 분산추정량의 결과를 비교하였다. 두 분산추정 방법간 결과를 비교한 결과 다단집락추출의 분산식으로 산출된 분산추정량과 테일러 급수 선형 전개를 활용한 분산추정량이 비슷한 수준으로 산출되었다. 이를통해 집락내 표본이 거의 유사한 속성을 가지고 있고, 표본의 크기가 충분히 크다면 집락내 분산이 거의 0에 수렴한다고 가정하는 테일러 급수 선형 전개를 활용해도 모평균에 대한 분산추정량에는 큰 차이가 없음을 확인할 수 있다.

주요용어 : 다단집락추출, 분산추정량, 테일러급수선형전개.

1. 서론

국가승인통계 중 개인 또는 가구를 대상으로 진행되는 사회조사에서는 통계청에서 제공하는 조사구 또는 집계구를 조사단위로 활용하는 다단집락추출(Multistage Cluster Sampling)을 많이 활용하고 있다. 다만, 실제 설문조사 과정에서는 원표본의 응답거부 또는 조사가 불가능한 상황으로 단위 무응답이 발생되어 이에 대해 대체표본으로 대체가 되거나 단위 무응답이 발생한 그대로 자료가 분석되게 된다. 특히 다단집락추출을 적용한 설계에서 사후층화, 단위 무응답에 대한 보정 등이 이루어지면 정확한 추정식을 만드는 것이 어렵다. 따라서 이러한 복잡한 조사에서는 점추정량(point estimator)에 대한 근사분산(approximate variance) 추정이 필요하며, 분산의 근사 추정 방법으로 테일러 급수 선형화 기법을 많이 활용하고 있다(Sung, 2000).

본 연구는 다단집락추출 방법을 활용하는 국가승인통계 데이터를 활용하여 다단집락추출의 분산추정량과 테일러 급수 선형전개를 활용한 분산추정량에 실제 차이가 어느 정도 발생하는지를 확인하여 테일러 급수 선형전개를 활용한 분산추정량 사용의 적절성을 검증하고자 한다.

---

*이 논문은 제1저자 김유미의 석사학위논문의 축약본임.

[1]18101 경기도 오산시 한신대길 137, 한신대학교 일반대학원 빅데이터융합전공 석사.
E-mail: corntea2119@naver.com

2. 다단집락추출에 대한 분산추정식 검토

2.1 다단집락추출에 따른 분산추정식

표본조사 결과로부터 모평균 μ의 추정량은 다음의 식을 적용하여 산출할 수 있다.

$$\hat{\mu}=\left(\frac{N}{M}\right)\frac{\sum_{i=1}^{n}M_i\bar{y}_i}{n}=\frac{1}{\overline{M}}\frac{\sum_{i=1}^{n}M_i\bar{y}_i}{n} \tag{2.1}$$

Table 1. 첨자에 대한 설명

N	모집단에 있는 집락 수(총 조사구)
n	단순임의추출로 뽑힌 집락 수(표본으로 추출된 조사구)
M_i	집락 i에 있는 조사단위의 수(추출된 조사구의 총 가구수)
m_i	집락 i로부터 단순임의추출로 뽑힌 조사단위의 수(추출된 조사구의 표본 가구수)
$M=\sum_{i=1}^{N}M_i$	모집단에 있는 총 조사단위의 수(총 가구수)
$\overline{M}=\frac{M}{N}$	모집단의 평균 집락크기(조사구내 평균 가구수)
y_{ij}	i번째 집락에 있는 j번째 관측값
$\bar{y}_i=\frac{1}{m_i}\sum_{j=1}^{m_i}y_{ij}$	i번째 집락의 표본평균

각 단계에서 단순임의추출을 가정하여 다단집락추출을 진행했을 때, 모평균 $\hat{\mu}$의 추정분산은 집락간 분석과, 집락내 분석의 각 단계별 분산을 구하고 이를 더하는 형태로 산출되어야한다. 즉, 총분산 = 집락간 분산 + 집락내 분산은 다음과 같은 형태로 산출된다.

$$\hat{V}(\hat{\mu})=\left(1-\frac{n}{N}\right)\left(\frac{1}{n\overline{M}^2}\right)s_b^2+\frac{1}{nN\overline{M}^2}\sum_{i=1}^{n}M_i^2\left(1-\frac{m_i}{M_i}\right)\left(\frac{s_i^2}{m_i}\right) \tag{2.2}$$

여기서 $s_b^2=\frac{\sum_{i=1}^{n}(M_i\bar{y}_i-\overline{M}\hat{\mu})^2}{n-1}$ 이고, $s_i^2=\frac{\sum_{j=1}^{m_i}(y_{ij}-\bar{y}_i)^2}{m_i-1}$ 이다.

여기서 s_b^2은 $M_i\bar{y}_i$의 표본에 대한 분산이고 s_i^2은 집락 i로부터 추출된 표본들의 표본분산이다.

2.2 테일러 급수 선형전개를 적용한 분산추정식

통계청에서 제공하는 인구총조사용 조사구를 집락으로 이용하여 통계를 생산하는 경우를 중심으로 설명하자면, 17개 지역(h)(서울, 부산, 대구, 인천, 대전, 광주, 울산, 세종, 경기, 강원, 충북, 충남, 전북, 전남, 경북, 경남, 제주), 각 지역에 조사구가 N_h개가 있으며, i번 조사구는 M_{hi}개 가구로 구성되어있다. 조사모집단은 L개의 층으로 구분되어있고, 각 층은 N_h(h=1,2,3....,L)개의 1차 추출단

위(PSU: primary sampling unit)으로 구성되어있다. 즉 조사구가 1차 추출단위가 된다. 각 조사구(PSU)는 2차 추출단위(SSU: second stage sampling unit)인 추출단위 즉 가구들을 포함하고 있으며, h 층의 i번 조사구(PSU)는 2차 추출단위인 가구 M_{hi}(h=1,2,3....,L, i=1,2,3,...N_h)개를 포함하고 있다. 조사구와 조사가구는 다음과 같다.

$$PSU\text{의 총 수} : N=\sum_{h=1}^{L} N_h \text{(총 조사구 수)}$$

$$SSU\text{의 총 수} : M=\sum_{h=1}^{L}\sum_{i=1}^{N_h} M_{hi} \text{(총 가구의 수)}$$

모집단의 총계 Y의 추정량은 다음과 같이 산출할 수 있다.

$$\text{총계 추정량} : \hat{Y}=\sum_{h=1}^{L}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w_{hij}y_{hij}=\sum_{h=1}^{L}\frac{n_h}{n_h}\sum_{i=1}^{n_h} y_{hi.}=\sum_{h=1}^{L} n_h\bar{y}_{h..} \tag{2.3}$$

총계 추정량의 분산은 다음과 같이 산출할 수 있다.

$$\begin{aligned}\hat{Var}(\hat{Y}) &= \hat{Var}(\sum_{h=1}^{L} n_h\bar{y}_{h..}) \\ &= \sum_{h=1}^{L}\frac{n_h(1-f_h)}{n_h-1}\sum_{i=1}^{n_h}(y_{hi.}-\bar{y}_{h..})^2\end{aligned} \tag{2.4}$$

모평균 추정량 $\bar{y}$는 다음과 같이 산출할 수 있다.

$$\bar{y}=\frac{\sum_{h=1}^{L}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w_{hij}y_{hij}}{w...}=\frac{\hat{Y}...}{\hat{M}} \tag{2.5}$$

테일러 급수 선형전개를 이용하면 $\hat{\bar{Y}}=\frac{\hat{Y}}{\hat{M}}\approx\frac{Y}{M}+\frac{1}{M}(\hat{Y}-\frac{Y}{M}\hat{M})$으로 정리될 수 있다. 따라서 이를 적용하면 모평균 추정량은 다음과 같이 정리할 수 있다.

$$\begin{aligned}&= \bar{Y}+\frac{1}{M}(\sum_h\sum_i\sum_j w_{hij}y_{hij}-\bar{Y}\sum_h\sum_i\sum_j w_{hij}) \\ &= \bar{Y}+\frac{1}{M}\sum_h\sum_i\sum_j w_{hij}(y_{hij}-\bar{Y})\end{aligned} \tag{2.6}$$

모평균 추정식에서 M과 $\bar{Y}$는 알 수 없으므로 $\hat{M}=w...$와 $\bar{y}$로 대체하여 다음과 같이 분산을 추정할 수 있다.

$$\hat{Var}(\bar{y}) = \hat{Var}\left(\frac{\sum_h\sum_i\sum_j w_{hij}(y_{hij}-\bar{y})}{w...}\right)\text{이다.}$$

여기에서 $\left(\frac{\sum_j w_{hij}(y_{hij}-\bar{y})}{w...}\right)=e_{hi.}$이라고 하면, 위 식은 아래와 같이 정리될 수 있다.

$$\hat{V}ar\left(\sum_h \frac{n_h}{n_h}\sum_i e_{hi}\right)=\hat{V}ar\left(\sum_{h=1}^{L} n_h \bar{e}_{h..}\right) \tag{2.7}$$

$$=\sum_h^L n_h^2 \frac{s_{h..}^2}{n_h} \;=\sum_{h=1}^{L}\frac{n_h^2(1-f_h)}{n_h}\frac{1}{n_h-1}\sum_{i=1}^{n_h}(e_{hi}-\bar{e}_{h..})^2 \;=\sum_{h=1}^{L}\frac{n_h(1-f_h)}{n_h-1}\sum_{i=1}^{n_h}(e_{hi}-\bar{e}_{h..})^2$$

여기서 모평균 추정량에 대한 상대표준오차는 $\frac{\sqrt{\hat{V}(\bar{y})}}{\bar{y}}\times 100(\%)$으로 산출된다.

모비율에 대한 추정은 층 h의 i번 조사구의 j번째 표본 가구의 응답은 0 혹은 1의 값을 갖고 있을 뿐 모평균 추정량과 동일한 식을 사용할 수 있다. 따라서 모비율 추정량은 다음의 식으로 산출할 수 있다.

$$\hat{p}=\frac{\sum_{h=1}^{L}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w_{hij}y_{hij}}{w...}$$

모비율에 추정량에 대한 분산 추정량은 아래의 식으로 산출된다.

$$\hat{V}ar(\hat{p})=\sum_{h=1}^{L}\frac{n_h(1-f_h)}{n_h-1}\sum_{i=1}^{n_h}(e_{hi}-\bar{e}_{h..})^2 \tag{2.8}$$

여기에서 $e_{hi}=\left(\frac{\sum_{j=1}^{m_{hi}} w_{hij}(y_{hij}-\hat{p})}{w...}\right)$ 이고, 모비율 추정량의 상대표준오차는 $\frac{\sqrt{\hat{V}(\hat{p})}}{\hat{p}}\times 100(\%)$ 이다.

단, 모비율 추정치에 대한 상대표준오차 사용시 질문하는 방식에 따라 응답의 내용이 크게 달라지는 경우(예를들어, “귀하는 원자력발전소 건설을 찬성하십니까?”와 “귀하는 원자력발전소 건설을 반대하십니까?”)에는 모비율 추정치는 큰 차이가 있지만 분산은 같기 때문에 상대표준오차는 큰 차이를 보인다. 따라서 이러한 경우에는 신뢰구간 또는 허용오차를 사용하는 것이 적절하다.

3. 다단집락추출에서 분산추정량 비교

3.1 통계청의 조사구를 활용한 조사에서의 분산추정량 비교

모평균 추정치에 대한 분산을 비교하기 위해 통계청의 MDIS(MicroData Integrated Service)에서 다운로드가 가능한 자료를 활용하였다. 문화체육관광부의 「국민여가활동조사」는 통계청 승인통계(통계승인번호 제 113014호)로 전국 만15세 이상 인구의 여가활동을 파악하기 위해 매년 실시되는 조사이다.

본 연구에서는 공표된 데이터만을 활용하여 분석이 가능하므로, 공표데이터 중 일부 데이터를 활용하여 가상의 조사구를 만들어 분산을 추정하였다. 연구에 활용한 지역은 서울 지역 정보를 활용하여 4개 조사구를 임의 부여하였다.

Table 2. 2020년 국민여가활동조사 층화 기준(출처 : 2020년 국민여가활동조사 보고서)

- 1차 층화는 전국 17개 시도(7개의 특별·광역시, 세종시, 9개의 도지역)를 도시지역과 농촌지역의 특성을 반영하여 읍/면부로 세부 층화 - 2차 층화는 위의 행정구역에 따라 구성된 1차 층을 조사구 특성(일반조사구, 아파트조사구)에 따라 2차 층화를 하여 구분 - 표본추출은 지역별 및 동 및 읍/면부별, 집락별로 층화한 후 1,000개 조사구를 계통추출하고, 각 조사구에서 10가구를 추출하여 총 10,000가구 조사하는 다단계층화집락추출(Stratified Multi-Stage Cluster Sampling)을 사용 - 가구는 표본조사구의 요도와 가구명부에서 10가구를 계통추출법으로 선정하고 순서대로 방문하여 적격 조사대상가구를 선정

본 조사 자료 자료를 활용하여 산출한 조사지역의 평일 여가시간은 평균 3.7시간($\hat{\mu}$)를 보낸 것으로 집계되었다.

서울 지역의 평일 평균 여가시간 추정치에 대한 분산은 다음과 같이 산출된다.

$$\hat{V}(\hat{\mu}_{hi}) = \left(1-\frac{n_h}{N_h}\right)\left(\frac{1}{n_h\overline{M}_h^2}\right)s_b^2 + \frac{1}{n_h N_h \overline{M}_{hi}^2}\sum_{i=1}^{n} M_{hi}^2\left(1-\frac{m_{hi}}{M_{hi}}\right)\left(\frac{s_{hi}^2}{m_{hi}}\right) \tag{3.1}$$

$$\text{여기서 } s_b^2 = \frac{\sum_{i=1}^{n}(M_{hi}\bar{y}_{hi} - \overline{M}_{hi}\hat{\mu}_{hi})^2}{n_h - 1} = \frac{150.8}{(4-1)} = 50.267 \text{이며,}$$

$$\sum_{i=1}^{n} M_i^2\left(1-\frac{m_i}{M_i}\right)\left(\frac{s_i^2}{m_i}\right) = 60^2\left(1-\frac{10}{60}\right)\left(\frac{50.267}{10}\right) = 3{,}600\times 0.833\times 5.0267 = 15{,}080.1 \text{이다.}$$

집락간 분산은

$$\left(1-\frac{n}{N}\right)\left(\frac{1}{n\overline{M}^2}\right)s_b^2 = \left(1-\frac{4}{68{,}832}\right)\left(\frac{1}{4\times 60^2}\right)\times 50.267 = 0.99994\times 0.00349 = 0.00349 \text{이 되며,}$$

집락내 분산은 아래와 같이 계산된다.

$$\frac{1}{nN\overline{M}^2}\sum_{i=1}^{n} M_i^2\left(1-\frac{m_i}{M_i}\right)\left(\frac{s_i^2}{m_i}\right) = \frac{1}{4\times 68{,}832\times 60^2}\times 15{,}080.1 = 0.000015$$

서울 지역의 평일 여가시간 모평균 추정량 $\hat{\mu}$에 대한 최종 분산은 집락간 분산과 집락내 분산의 합계이므로 다음과 같이 계산된다.

$$\hat{V}(\hat{\mu}) = 0.00349(\text{집락간분산}) + 0.000015(\text{집락내 분산}) = 0.003506$$

허용오차는 95% 신뢰수준에서 $1.96\times\sqrt{0.003506} = 0.12$시간로 계산되어 서울지역의 평일 여가시간은 3.7시간 $\pm$ 0.12시간로 추정된다.

위에서 확인한 것처럼, 표본의 크기가 적절하다면 2단 집락추출의 경우 집락내 분산값이 0.000015로 작아 총 분산추정량의 0.4% 수준으로 총 분산에 영향을 거의 미치지 않는 것이 확인된다.

3.2 테일러 급수 선형 전개를 활용한 분산추정량 산출

평일 평균 여가시간 3.7시간($\bar{y}$)에 대한 분산을 테일러 급수 선형 전개를 활용하여 분산을 추정하면 다음과 같이 계산된다.

$$\hat{V}(\hat{\mu}) = \sum_{h=1}^{L} \frac{n_h(1-f_h)}{n_h-1} \sum_{i=1}^{n_h} (e_{hi.} - \bar{e}_{h..})^2, \tag{3.2}$$

여기서 $f_h = \frac{n_h}{N_h}$ 이고, $e_{hi.} = \frac{\left(\sum_{j=1}^{m_{hi}} w_{hij} - \bar{y}\right)}{w...}$ 이고, $\bar{e}_{h..} = \frac{\left(\sum_{i=1}^{n_h} e_{hi.}\right)}{n_h}$ 이다.

$\sum_{h=1}^{L} \frac{n_h\left(1-\frac{n_n}{N_n}\right)}{n_h-1} = 1.33325$이므로, 따라서 분산은 아래와 같이 산출된다.

$$\hat{V}(\hat{\mu}) = \sum_{h=1}^{L} \frac{n_h(1-f_h)}{n_h-1} \sum_{i=1}^{n_h} (e_{hi.} - \bar{e}_{h..})^2 = 1.33325 \times 0.0023245 = 0.003099$$

따라서 허용오차는 95% 신뢰수준에서 $1.96 \times \sqrt{0.003099} = 0.109$시간로 산출된다.

결과적으로 서울지역 응답가구의 평일 여가시간에 대한 분산은 3.7시간 ± 0.11시간으로 추정되어 다단집락추출의 분산식으로 산출된 3.7시간 ± 0.12시간과 비슷한 수준으로 산출된다. 따라서 조사구 내 가구 또는 가구원의 속성이 거의 유사할 때 집락내 분산이 거의 0에 수렴한다고 가정하는 테일러 급수 선형 전개를 활용해도 모평균에 대한 분산추정량에는 큰 차이가 없음을 확인할 수 있다.

4. 결론

본 연구에서는 다단집락추출을 통한 모수 추정시 해당 추정치에 대한 분산이 원칙적으로는 집락간 분산과 집락내 분산의 합으로 이루어져야하나, 표본의 크기가 적절하고(혹은 충분히 크고), 집락내 표본의 유사성이 확보된다면 테일러의 1차 전개 만으로도 충분히 만족스러운 결과가 얻어진다는 것을 확인하였다.

많은 조사통계 보고서에서 다단집락추출로 표본을 추출했음에도 추정치에 대한 분산식이 다단집락추출에 대한 분산식이 아닌 테일러 급수 선형 전개를 활용한 분산추정식이 사용되고 있는 것은 바로 위와 같은 결과 때문인 것으로 보인다.

다만 자료 이용자의 입장에서 표본추출 방식과 다른 분산추정식이 기재되어 있는 경우 자료 활용에 혼란이 있을 수 있는데, 많은 조사통계 보고서에서 테일러 급수 선형 전개를 활용한 이유에 대해서는 언급이 생략되어 있는 경우가 많으므로 다단집락추출의 분산식과 테일러 급수 선형 전개를 활용한 분산추정량 간에 큰 차이가 없음을 언급하고 해당 분산추정식을 통한 추정량을 도출하는 것이 더 적절할 것으로 보인다.

Reference

Sung N. G., (2000). Software packages for survey data analysis, *Survey research* v.1 no.1 , pp.109 - 123

Han, G. S., (2005). Intracluster Correlation Adjusted Variance, *Journal of the Korean Data Analysis Society*, Vol. 7, No. 3, June 2005, pp. 8

Jung, G. H., Lee, J. H., Kim, N. K., (2008). Comparison of Variance Estimation for the Economically Active Population Survey, *Journal of Korean official statistics*, vol.13, no.2, pp. 204-218 (15 pages)

Song., J. H., Jea., H. E., Park. M. G., (2011). A Study of Sample Size for Two-Stage Cluster Sampling, *The Korean journal of applied statistics* v.24 no.2 , pp.393 - 400

Lee, G. S. (2017). An unrelated question model by PPS sampling, *Journal of the Korean Data Analysis Society*, Vol. 19, No. 4 (B), pp. 1897-1906

Jeong, H. S., Hong, S. J., Son, C. K. (2020). A Study on the Calibation Estimator for an Adaptive Cluster Samplin, *Journal of the Korean Data Analysis Society*, 22(5), 1767-1777.

Lee, G. S., Hong, S. J.,(2020). A Conditional Multiplicative Unrelated Quantitative Randomized Response Model by TwoStage Cluster Samplin, *Journal of the Korean Data Analysis Society*, 22(4), 1361-1370.

Comparison of Variance Estimates in Multistage Cluster Sampling*

Yumi Kim[1]

Abstract

In this study, we checked the actual difference between the variance estimation of multi-stage cluster extraction using random sampling and the variance estimation using Taylor series linear development, and verified whether it is appropriate to calculate the variance estimation of the design using Taylor series linear development. For the data, the results of the two variance estimates were compared using raw data from the 『National Leisure Activity Survey (Ministry of Culture, Sports and Tourism)』, which uses the survey district as a colony. As a result of comparing the results between the two variance estimation methods, the variance estimation calculated by the variance equation of multi-stage cluster extraction and the variance estimation using Taylor series linear development were calculated at a similar level. From this, it can be confirmed that if the sample in a colony has almost similar properties and the size of the sample is large enough, there is no significant difference in the variance estimate for the population mean using Taylor series linear deployment, which assumes that the variance in the colony converges to almost zero.

Keywords : Multi-stage cluster extraction, variance estimation, Taylor series linear development.

[1]Master's degree, Big data Convergence, Hanshin University, 137 Hansindae-gil, Osan-si, Gyeonggi-do, 18101, Korea. E-mail: corntea2119@naver.com

Proceedings of the Korean Data Analysis Society
January 26-27, 2023, pp. 243-243

Text To Table을 적용한 Robotic Process Automation 사례 연구*

이준서[1], 염승민[2], 이다운[3], 백송희[4], 최호식[5]

요 약

120다산콜센터는 교통 정보, 수도 요금, 지방세, 민원 신고, 정책 문의 등 서울특별시와 25개 자치구의 행정에 대한 전문적인 상담을 제공하는 서울특별시 행정상담 전문 콜센터로 전화, 문자, 수어, 챗봇, SNS, 외국인의 경우 외국어 상담까지 시민들이 일상에서 궁금한 점이나 느끼는 불편 사항 등 다양한 질의에 대한 상담 창구역할을 수행하고 있다. 본 연구에서는 2020년 1월부터 2022년 6월까지 접수된 총 14,786,174건의 민원 데이터로부터 빈발한 민원을 발굴하고, 이를 활용하여 민원 중 큰 비중을 차지하는 전화상담 및 전화번호 안내 등을 자동으로 처리하는 방법에 대해 살펴본다. 우선, 지역명 및 유관기관, 부서명 등을 고려하여 데이터를 표 형태의 데이터로 변환(Text to Table)하여 각 기관 및 자치구의 업무 내용으로 구성된 데이터베이스(Database)를 구축하고 이와 민원 데이터에서 민원인이 제기한 질문 쌍을 Sentence-BERT 언어모형에 적합하여 구문 유사도 학습을 통해 임베딩 공간을 구성한다. 다음으로 전화상담 및 전화번호 요청 민원에 대해 민원 텍스트 문장 임베딩 코사인 유사도가 가장 높은 업무 내용을 선정 및 추천하는 프로세스(robotic process automation)를 구축하였다. 임의 추출된 2,000개의 민원 표본에 대하여 업무 추천 프로세스를 진행한 결과, 평균 63.28%의 구문 유사성이 있음을 확인할 수 있었으며, 향후 상담원들의 업무 부담을 완화하는 방안이 될 것으로 기대한다.

주요용어 : 임베딩, 텍스트, RPA, NLP, 추천시스템.

---

*본 연구는 서울시립대학교 도시과학빅데이터AI연구원의 슈퍼컴퓨팅 자원을 지원 받아 수행되었습니다.
[1]02504 서울특별시 동대문구 서울시립대로 163, 서울시립대 일반대학원 도시빅데이터융합학과 석사과정.
E-mail: jsyi@uos.ac.kr
[2]02504 서울특별시 동대문구 서울시립대로 163, 서울시립대 일반대학원 도시빅데이터융합학과 석사과정.
E-mail: yeom0913@uos.ac.kr
[3]02504 서울특별시 동대문구 서울시립대로 163, 서울시립대 융합전공학부 학부재학생.
E-mail: dlekdns07@uos.ac.kr
[4]02504 서울특별시 동대문구 서울시립대로 163, 서울시립대 일반대학원 도시공학과 박사과정.
E-mail: thdgml1445@uos.ac.kr
[5](교신저자) 02504 서울특별시 동대문구 서울시립대로 163, 서울시립대 도시빅데이터융합학과 부교수.
E-mail: choi.hosik@uos.ac.kr

Proceedings of the Korean Data Analysis Society
January 26-27, 2023, pp. 245-249

정규화 특이값 분해 기반 추천 시스템 모델 제안 및 비교 연구

김재훈[1], 이도현[2], 이상헌[3], 김도윤[4], 임광묵[5], 육현섭[6], 서문민건[7], 홍석수[8], 허태영[9]

요 약

추천 시스템은 사용자의 과거 행동과 선호도를 바탕으로 관심을 가질 만한 아이템을 제공하는 것이다. 추천 시스템은 영화, 음악, 도서 그리고 생필품 추천과 같은 다양한 분야에서 활용되고 있다. 협업 필터링은 추천 시스템 분야에서 좋은 성능을 보이는 기법이며 그중에서도 행렬 분해를 통한 접근 방법은 추천 시스템 데이터와 같이 행렬의 높은 희소도 문제를 해결하는 데에 뛰어난 성능을 보인다. 본 논문에서는 특이값 분해 기반의 영화 추천 모델을 설계하고자 하였다. 우리는 정규화 기법을 이용한 Normalized SVD 방법을 제안하고 LLORMA(Local Low-Rank Matrix Approximatio; 지역 저차원 행렬 접근법) 아이디어를 제안된 모형에 추가함으로써 고성능의 추천 시스템 모델을 생성하고자 하였다. Movielens 벤치마크 데이터 세트에 대하여 제안된 모형은 기존 SVD 및 SVD++ 알고리즘 기반 추천 시스템 모형과 비교되었다. 제안된 추천 모델의 성능은 RMSE(Root Mean Square Error; 평균제곱 오차)로 설명되었으며 실험 결과 LLORMA 기반의 정규화 SVD 모델이 기존의 SVD 및 SVD++ 모델보다 더 좋은 성능을 보이는 것을 확인할 수 있었다.

주요용어 : 추천 시스템, 행렬 분해, 특이값 분해, LLORMA, RMSE.

1. 서론

추천 시스템(recommendation system)은 최근 많은 기업이 실제 경영에 활용하고 있는 연구 분야이다(Isinkaye et al, 2015). 추천 시스템은 사용자의 과거 행동과 선호도를 바탕으로 사용자들이 좋아할 만한 상품을 추천해주는 시스템을 의미하여 영화, 음악, 생필품 등 다양한 분야에서 활용되고 있다(Bobadilla et al. 2010; Zhou, Y. 2008; Sivapalan, 2014). 기존의 추천 시스템 방법론은 크게 콘텐츠 기반 필터링과 협업 필터링으로 나뉘게 되며 최근 연구에서는 기존 추천 시스템 모델에 다양한 아이디어를 적용한 모델들이 좋은 성능을 보이고 있다(Vančura et al, 2021; Shenbin et al, 2020).

[1]충북 청주시 서원구 충대로1, 충북대학교 통계학과 박사과정. Email: kjh664425@naver.com
[2]충북 청주시 서원구 충대로1, 충북대학교 통계학과 석사과정. Email: dhguslee@naver.com
[3]충북 청주시 서원구 충대로1, 충북대학교 통계학과 석사과정. Email: dltkdgjs0313@gmail.com
[4]충북 청주시 서원구 충대로1, 충북대학교 통계학과 박사과정. Email: doyun2516@gmail.com
[5]충북 청주시 서원구 충대로1, 충북대학교 통계학과 석사과정. Email: imgm3452@naver.com
[6]충북 청주시 서원구 충대로1, 충북대학교 통계학과 석사과정. Email: youk2222@naver.com
[7]충북 청주시 서원구 충대로1, 충북대학교 통계학과 석사과정. Email: seomun97@naver.com
[8]충북 청주시 서원구 충대로1, 충북대학교 통계학과 박사과정. Email: blindsage10@gmail.com
[9](교신저자) 충북 청주시 서원구 충대로1, 충북대학교 통계학과 교수. Email : theo@cbnu.ac.kr

본 논문에서는 협업 필터링에서 SVD(Singular Value Decomposition; 특이값 분해)를 활용한 추천 시스템 모델을 제안한다. SVD는 행과 열의 길이가 다른 행렬의 분해에 사용할 수 있으므로 추천 시스템에서 널리 사용되는 행렬 분해 기법의 하나다 (Ma et al, 2008; Zhou et al, 2015). 우리는 기존의 SVD에 정규화 기법과 LLORMA(Local Low-Rank Matrix Approximation) 아이디어를 추가함으로써 보다 나은 성능의 추천 시스템 모델을 만들고자 하였다(Lee et al, 2016).

본 논문의 세부 구성은 다음과 같다. 2절에서는 본 연구의 모델 개발에 사용된 아이디어의 내용을 기술하였다. 3절에서는 우리가 제안하는 모델과 기존의 SVD 기반의 모델들 간의 성능을 비교하였다. 4절에서는 향후 연구의 발전 과정에 관하여 기술하였다.

2. 방법론

본 연구에서의 모델에 사용된 SVD의 형태를 도식화하면 Figure 1과 같으며, 다음의 도식은 Albatayneh(2014)에서 다루어진 바 있다. Figure 1에서 SVD 행렬 분해(Matrix Factorization)를 적용한 알고리즘은 다음과 같은 도식으로 표현할 수 있다.

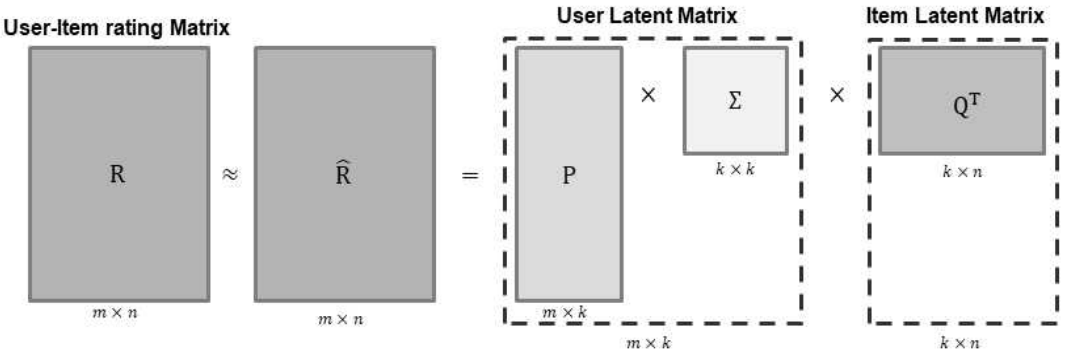


Figure 1. SVD algorithm applied matrix factorization

본 연구에서는 기존 SVD 알고리즘과 다르게 SVD로 분해된 행렬의 성질에 집중한 Normalized SVD 알고리즘을 제안하고 다음과 같은 식을 이용하여 행렬 R을 예측하고자 한다.

$$R = P\Sigma Q^T$$

행렬 P와 Q는 모두 SVD의 조건인 정규성과 직교성을 만족하는 행렬이다. 따라서 두 행렬 모두 열벡터가 정규 직교 벡터이며, 행렬 Σ의 경우 원소들이 k개의 대각 원소 $\sigma_1, \sigma_2, \cdots, \sigma_k$로 구성되어 있으며 $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_k > 0$와 같은 성질을 가진다.

SVD로 인해 분해 된 잠재 행렬 P와 Q의 열벡터들은 정규직교 벡터이다. 본 논문에서는 정규 벡터 성질을 만족시키기 위해 SVD로 분해된 사용자 잠재 행렬(User Latent Matrix)와 아이템 잠재 행렬(Item Latent Matrix)의 열 벡터(Column vector)들을 다음과 같은 수식으로 표현하였다.

$$\tilde{p}=(p_1,p_2,\cdots,p_k)\Rightarrow\left(\frac{p_1}{\sqrt{p_1^2+p_2^2+\cdots+p_k^2}},\frac{p_2}{\sqrt{p_1^2+p_2^2+\cdots+p_k^2}},\cdots,\frac{p_k}{\sqrt{p_1^2+p_2^2+\cdots+p_k^2}}\right)$$
$$\Rightarrow(p_1^*,\ p_2^*,\ \cdots,\ p_k^*)$$

$$\tilde{q}^T=(q_1,q_2,\cdots,q_k)\Rightarrow\left(\frac{q_1}{\sqrt{q_1^2+q_2^2+\cdots+q_k^2}},\frac{q_2}{\sqrt{q_1^2+q_2^2+\cdots+q_k^2}},\cdots,\frac{q_k}{\sqrt{q_1^2+q_2^2+\cdots+q_k^2}}\right)$$
$$\Rightarrow(q_1^*,\ q_2^*,\ \cdots,\ q_k^*)$$

또한 잠재 행렬 Σ은 대각 행렬이고 각 대각 원소는 양수인 조건을 만족해야하기 때문에, 지수 함수를 이용해 각 원소 $\sigma_1,\cdots,\sigma_k$를 다음과 같이 정의하였다.

$$\tilde{\sigma}=(\sigma_1,\cdots,\sigma_k)=(\exp(\beta_1),\cdots,\exp(\beta_k)),\ \tilde{\beta}\in\mathbb{R}^k$$

기존 SVD 방법과 동일하게 해당 알고리즘 또한 잠재 행렬 및 각 원소들에 대해 RMSE(Root Mean Square Error)를 최소화하도록 최적화를 진행하였다.

LLORMA(Local Low-Rank Matrix Approximation)는 SVD로 분해된 사용자 잠재 행렬과 아이템 잠재 행렬을 입력값으로 하는 커널 함수(Kernel Function)을 이용하여 사용자들과 아이템들을 그룹화하며, 그 후 각 그룹에 대해 SVD를 실시하는 방법이다. 그룹화를 위한 커널 함수는 다음과 같이 정의된다.

$$K_h(s_1,s_2)\propto I[d(s_1,s_2)<h]$$

이 때, s_1과 s_2는 각 잠재 행렬들의 원소 위치이고 h는 그룹화를 위한 임계치이다. 또한, 함수 $d(\cdot)$은 거리 함수이며 각 원소 위치에 해당하는 잠재 벡터 $(p_{s_1}^*,p_{s_2}^*)$ 또는 $(q_{s_1}^*,q_{s_2}^*)$의 함수이다. LLORMA 알고리즘에서, 각 잠재 행렬의 원소 위치를 임의로 q개 선택하고 선택된 각 원소의 순서쌍 $(a_t,b_t),\ t=1,\cdots,q$은 그룹화에 사용되는 고정된 위치의 의미를 갖는다. 커널 함수를 이용하면 각 잠재 행렬 벡터 $(i,j),\ i=1,\cdots,m,j=1,\cdots,n$은 각 그룹에 속하는지 다음과 같이 판단할 수 있다.

$$K_{(a_t,b_t)}(i,j)=K_h(a_t,i)K_h(b_t,j)$$

이때, $K_{(a_t,b_t)}(i,j)$는 0 또는 1의 값을 가지며 1일 경우 순서 쌍 (i,j)는 (a_t,b_t) 그룹에 속한다. LLORMA는 각 그룹에 SVD 적용하기 위해 다음과 같은 손실 함수를 정의한다.

$$L_{LLROMA}=\sum_{(i,j)\in\Omega}\sum_{t=1}^{q}K_{(a_t,b_t)}(i,j)([P^{(t)}\Sigma^{(t)}Q^{(t)T}]_{i,j}-R_{i,j})^2$$

해당 손실 함수는 각 그룹에 대한 잠재 행렬 $P^{(t)},Q^{(t)}$과 행렬 $\Sigma^{(t)}$를 RMSE를 최소로 하여 추정하는 문제로 인식할 수 있다. 벌점 항을 추가하여 각 잠재 행렬에 대한 정규화(Regularization)를 실시할 수 있다. 최종적으로, 각 위치에 대한 예측값은 다음과 같이 나타낼 수 있다.

$$\hat{R}_{i,j}=\sum_{t=1}^{q}\frac{K_{(a_t,b_t)}(i,j)}{\sum_{s=1}^{q}K_{(a_s,b_s)}(i,j)}[\hat{P}^{(t)}\hat{\Sigma}^{(t)}\hat{Q}^{(t)T}]_{i,j}$$

3. 모델간 성능 비교

기존의 SVD 기반 추천 시스템 모델들과 우리가 개발한 LLORMA 기반 Normalized SVD 모델의 성능 비교를 벤치마크(Benchmark) 데이터인 MovieLens 데이터셋을 기준으로 하여 비교하였다. 성능 비교는 기존 연구들과 마찬가지로 90%의 훈련 자료와 10%의 검증 자료를 이용하여 비교하였다. 성능 비교를 위한 평가 측도로는 RMSE를 사용하였으며 수식은 다음과 같다.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \tilde{y}_i)^2}$$

여기서 N은 예측된 전체 평점의 수를 의미하며 y_i와 $\hat{y}_i$는 각각 실제 평점과 예측 평점을 의미한다. RMSE가 낮을수록 추천 시스템이 사용자 평가를 더 정확하게 예측한다는 것을 의미한다. Table 1은 RMSE를 평가 측도로 하여 기존 SVD 기반 모델들과 우리가 제안하는 모델의 성능을 비교한 것으로 모든 데이터셋에서 우리가 제안하는 모델의 성능이 높은 것을 확인할 수 있다.

Dataset	RMSE			
	Bias SVD	SVD++	Normalized SVD	LLORMA based Normalized SVD
ML-100k	0.942	0.938	0.930	0.918
ML-1M	0.863	0.844	0.842	0.829

4. 향후 연구

본 연구에서는 기존의 SVD 기반의 추천 시스템 알고리즘의 성능 향상을 높이기 위하여 SVD 알고리즘의 행렬 분해 과정에서 각 열 벡터들의 정규화와 LLORMA 아이디어를 추가하였다. 향후 연구에서는 분해 행렬의 열 벡터에 대한 가정인 직교성(Orthogonality) 또한 만족시켜 SVD의 모든 조건 충족시키고자 하며 LLORMA 과정에서 거리 함수로 K-평균 군집화(K-means clustering)와 GMM(Gaussian Mixture Model)를 사용하여 보다 나은 성능의 모델을 개발하고자 한다.

References

Isinkaye, F. O., Folajimi, Y. O., & Ojokoh, B. A. (2015). “Recommendation systems: Principles, methods and evaluation.”, *Egyptian informatics journal*, 16(3), 261-273.

Bobadilla, J., Serradilla, F., & Bernal, J. (2010). “A new collaborative filtering metric that improves the behavior of recommender systems.” *Knowledge-Based Systems*, 23(6), 520-528.

Zhou, Y., Wilkinson, D., Schreiber, R., & Pan, R. (2008, June). “Large-scale parallel collaborative filtering for the netflix prize.” *In International conference on algorithmic applications in management* (pp. 337-348). Springer, Berlin, Heidelberg.

Sivapalan, S., Sadeghian, A., Rahnama, H., & Madni, A. M. (2014, August). “Recommender systems in e-commerce.” *In 2014 World Automation Congress (WAC)* (pp. 179-184). IEEE.

Vančura, V., & Kordík, P. (2021, September). “Deep Variational Autoencoder with Shallow Parallel Path for Top-N Recommendation (VASP).” *In International Conference on Artificial Neural Networks* (pp. 138-149). Springer, Cham.

Shenbin, I., Alekseev, A., Tutubalina, E., Malykh, V., & Nikolenko, S. I. (2020, January). Recvae: “A new variational autoencoder for top-n recommendations with implicit feedback.” *In Proceedings of the 13th International Conference on Web Search and Data Mining* (pp. 528-536).

Ma, C. C. (2008). “A guide to singular value decomposition for collaborative filtering.” *Computer (Long Beach, CA)*, 2008, 1-14.

Zhou, X., He, J., Huang, G., & Zhang, Y. (2015). “SVD-based incremental approaches for recommender systems.” *Journal of Computer and System Sciences*, 81(4), 717-733.

Lee, J., Kim, S., Lebanon, G., Singer, Y., &Bengio, S. (2016). “LLORMA: Local low-rank matrix approximation.” Proceedings of the 30th International Conference on Machine Learning, *PMLR* 28(2):82-90, 2013.

Albatayneh, N. A., Ghauth, K. I., & Chua, F. F. (2014). “A Semantic content-based forum recommender system architecture based on content-based filtering and latent semantic analysis.” *In Recent Advances on Soft Computing and Data Mining* (pp. 369-378). Springer, Cham.

Proceedings of the Korean Data Analysis Society
January 26-27, 2023, pp. 251-254

방한 외래관광객의 관광지 연관분석을 통한 한국관광 개선방안

오유진[1], 윤상복[2], 이현경[3]

요 약

본 연구는 코로나19 이전 관광객들의 한국 관광 동향을 파악하고 코로나19 종식 이후 방문할 외래관광객을 위한 한국관광 개선방안을 제시하였다. 연도별 관광지 간 이동 경로를 파악하기 위하여 연관분석을 실시하였으며, 다음과 같은 결과를 도출할 수 있었다. 첫 번째, 방한 외래관광객 관광경로의 주요 HUB는 '고궁', '동대문시장', '명동'으로 도출되었다. 두 번째, 2011년을 기점으로 관광경로가 증가하였으나, 시간이 지날수록 관광경로 수가 감소하고 단순화되었다. 세 번째, 관광순위가 감소하였던 '박물관(기념관)', '에버랜드/캐리비안베이', '롯데월드'의 주요 연관 관광지는 '고궁', '동대문시장', '명동'이며, 타 관광지와의 연계성이 강하지 않은 것으로 도출되었다. 네 번째, 비수도권역 관광지의 높아지는 방문 순위에 비하여 비수도권역 간의 관광지 연계성이 강하지 않은 것으로 도출되었다. 분석 결과를 바탕으로, 본 연구에서 제시한 한국 관광 품질 향상 방안은 다음과 같다. 첫째, 주요 관광지와 타 관광지와의 접근성을 강화할 필요성이 있다. 둘째, 주요 관광지에 방문하는 관광객들의 욕구를 파악하고 연계할 수 있는 관광 자원을 개발할 필요성이 있다. 셋째, 방문 순위가 하락하는 관광지는 기존 욕구를 강화하거나 새로운 매력 창출을 통해 관광객을 유입할 수 있도록 정비할 필요성이 있다. 마지막으로, 비수도권역의 관광지 간 연계 강화 방안 마련의 필요성이 있다.

주요용어 : 외래관광객, 한국관광, 관광지연계, 연관분석, 데이터마이닝.

1. 서론

K-POP, 한국 문화에 대한 관심 증대 등 한류의 영향으로 한국 관광은 꾸준한 상승세를 이어왔다. 하지만, 2019년 11월 중국 우한시에서 첫 코로나 환자가 발생하면서, 전 세계적으로 국내외의 이동을 각국의 정부가 주도하여 통제하게 되었다(Kim, Jin, 2021). 그 결과, 2020년 3월을 기준으로 국제선 여객이 전년 동기 대비 약 93.5%가 감소하고, 인천국제공항의 일평균 여객이 91.1% 감소하는 등 항공을 비롯한 관광산업은 큰 타격을 받게 되었다(Kim et al., 2020). 하지만, 팬데믹 기간동안 넷플릭스와 같은 글로벌 디지털콘텐츠 플랫폼의 인기로 외국인들의 한국 문화에 대한 관심이 증대되었고, 국가 간 이동 규제가 완화되면서 방한 외래관광객의 규모가 다시 증가하는 추세를 보이고 있다(Lee, 2022). 따라서, 본 연구에서는 코로나19 종식 이후 증가할 것으로 예상되는 방한 외

[1]47340 부산광역시 부산진구 엄광로 176, 동의대학교 대학원 도시공학과 박사과정.
E-mail: mid1207@naver.com
[2](교신저자) 47340 부산광역시 부산진구 엄광로 176, 동의대학교 공과대학 도시공학과 교수.
E-mail: yoonsb@deu.ac.kr
[3]47340 부산광역시 부산진구 엄광로 176, 동의대학교 공과대학 도시공학과 조교수. E-mail: lhk@deu.ac.kr

래관광객의 관광만족도를 높이기 위하여, 코로나19 이전의 관광 데이터를 기반으로 연관분석을 진행하고, 분석결과를 바탕으로 추후 방한 외래관광객을 위한 한국 관광의 개선방안을 제시하고자 한다.

2. 연구자료 및 방법론

본 연구의 시간적 범위는 코로나 발생 이전 10개년(2010년 ~ 2019년)으로 설정하였으며, 공간적 범위는 대한민국 전국구를 대상으로 하였다. 연구 표본은 문화체육관광부에서 제공하는 '외래관광객 실태조사 원자료'를 활용하였으며, 설문문항의 '한국여행 방문지' 약 30개를 추출하여 사용하였다.

본 연구는 크게 2가지의 분석 과정으로 이루어진다. 먼저, 연관분석을 위한 희소행렬(Sparse Matrix)를 형성하는 과정에서 1차적으로 연도별 관광지 순위를 추출하여 대략적인 관광지 방문 행태를 파악하였다. 이후, 외래관광객의 관광지 간 연관관계를 분석하기 위하여 연관분석(Association Analysis)을 진행하였다. 연관분석이란 '데이터 안에 존재하는 항목 간의 연관규칙(Association Rules)을 발견하는 과정으로, 두 가지 항목의 연관성을 통하여 한 항목의 값을 바탕으로 다른 항목의 값을 예측해내는 방법(Ryu et al., 2006)'이다. 연관분석은 전체 트랜잭션(transaction)의 비율을 의미하는 지지도(Support), 조건의 구매가 발생한 후에 결과의 구매가 일어난 비율인 신뢰도(Confidence), 한 품목의 구매확률이 다른 품목의 구매에 따라 얼마나 증가하였는지를 나타내는 향상도(Lift)의 3가지 값을 모두 고려하여 연관성을 판단한다. 연관분석에는 FP-Growth, FPV와 같이 다양한 알고리즘이 존재하지만, 본 연구에서는 Apriori 알고리즘을 선택하여 분석을 진행하였다. Apriori 알고리즘은 최소 지지도값 이하의 경우의 수들을 분석에서 제외할 수 있으므로, 분석의 효율을 높여주는 장점이 있다.

3. 연구결과

3.1. 빈도분석 결과

방한 외래관광객의 관광지 방문 현황을 파악하기 위하여 연도별 빈도분석 데이터를 추출하였으며 그 결과는 Figure 1, Figure 2와 같다. 2010년부터 2019년까지 인기 관광지 빈도분석 결과를 통하여 도출할 수 있는 결론은 첫 번째, 인기 관광지의 대부분이 수도권에 편중되어 있으며, 그 중에서도 '명동', '동대문시장', '고궁'은 연도마다 조금씩 순위의 변동은 있었으나, 10년간 꾸준히 높은 순위를 보였다. 두 번째, 박물관(기념관)과 에버랜드/캐리비안베이, 롯데월드의 인기 순위가 하락세를 보임을 알 수 있다. 세 번째, 2014년을 기점으로 부산, 제주도, 대구, 경주 등 지방권역의 관광지 방문이 높아지고 있음을 알 수 있다.

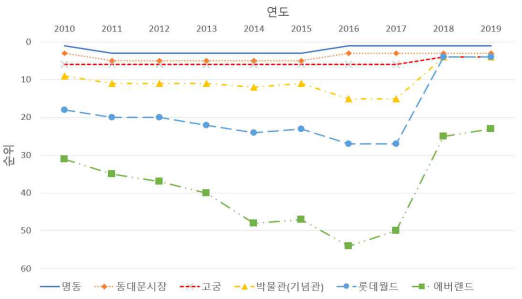


Figure 1. Ranking of Major Tourist Attraction

Figure 2. Ranking Tourist Attractions by Region

3.2. 연관분석 결과

연관분석을 위한 희소행렬(Sparse Matrix)를 생성하고, Apriori 알고리즘을 활용하여 최소 Support 0.2 & Confidence 0.7 & Minlen 2를 조건으로 연관 규칙을 생성하여 연관분석을 진행하였다. 분석 결과, 2010년 주 관광경로가 3개에서 시작하여, Figure 3와 같이 2011년을 기점으로 관광경로가 증가하였으나, 2016년부터 점차 그 수가 감소하기 시작하더니 2019년도에는 Figure 4와 같이 5개로 관광경로 수가 감소하고 단순화되는 현황을 알 수 있다. 또한, 관광지와의 연계성에 있어 가장 중심이 되는 관광 HUB는 '고궁'인 것으로 도출되었으며, '명동', '동대문시장', '남산/N서울타워'가 연도에 상관없이 꾸준히 각광받는 관광지로 도출되었다.

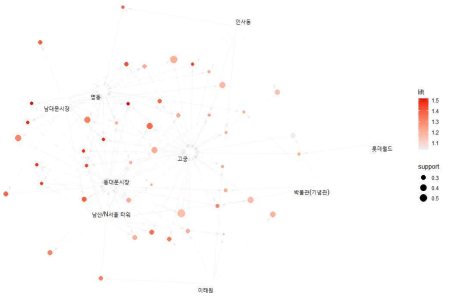


Figure 3. Results of Association Analysis in 2011

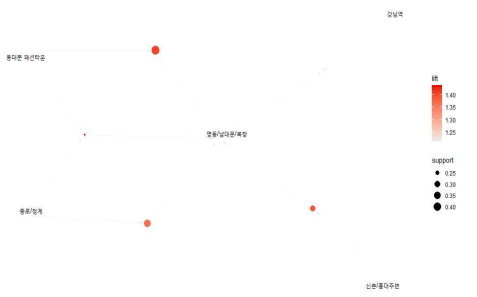


Figure 4. Results of Association Analysis in 2019

3.3. 분석결과를 통한 시사점

본 연구에서 실시한 빈도분석과 연관분석의 결과를 종합한 시사점은 다음과 같다. 첫 번째, 방한 외래관광객 대부분의 관광지 연계가 '고궁', '동대문시장', '명동'과 함께 이루어졌으므로, 방한 외래관광객의 주요 관광 HUB는 '고궁', '동대문시장', '명동'으로 결론내릴 수 있다. 두 번째, 점차 방문순위가 감소하고 있는 박물관(기념관), 에버랜드/캐리비안베이, 롯데월드의 관광 이동 경로가 해마다 점차 감소하고 단순화되는 현황을 알 수 있다. 세 번째, 시간이 지날수록 주요 관광지 수가 감소하고, 관광지 간의 이동 경로도 단순화되고 있음을 알 수 있다. 마지막으로, 2014년을 기점으로 부산과 제주도의 관광지 방문 순위가 점차 높아지고, 2017년부터는 대구, 경주 등 지방 관광지의 방문 순위도 높아졌으나, 수도권을 제외한 비수도권에서는 지역 내 관광지 간의 연계성이 강하지 않은 것으로 도출되었다.

3.4. 한국 관광 품질 개선 방안 제시

연관분석을 통하여 도출된 결론을 바탕으로, 코로나19 종식 이후 방한 외래관광객의 관광 품질 향상을 위한 방안을 제시하고자 한다. 첫 번째, 주요 관광지(고궁, 동대문시장, 명동)에 방문하는 관광객들이 다른 관광지로 이동하기 용이하도록 교통, 정보 접근성 등 관광지 간의 접근성을 강화하여 관광지 간의 연계가 강화되도록 유도한다. 두 번째, 박물관(기념관)과 에버랜드/캐리비안베이, 롯데월드의 방문순위가 감소하는 현황에는 관광지의 매력요소 부족, 떨어지는 접근성, 관광 욕구 변화 등 다양한 변수에 의하여 발생될 수 있다. 따라서, 점차 방문순위가 하락하는 관광지를 다각적인 측면에서 면밀히 분석하고 기존 관광 욕구를 강화 및 새로운 관광 욕구 창출, 관광 인프라

강화 등의 대책을 수립할 필요성이 있다. 세 번째, 관광지 간의 이동 경로가 감소하고 단순화되는 것은 관광객들의 관광 욕구 감소와 관광 욕구를 충족시켜 줄 수 있는 대체 관광자원의 부족이라 판단할 수 있다. 따라서, 주요 관광지에 방문하는 관광객들의 관광 욕구를 파악하고 이를 다양한 관광 활동으로 체험할 수 있는 관광자원의 개발이 필요하다. 마지막으로, 본 연구의 공간적 범위가 대한민국 전역이었음에도 불구하고, 유효한 관광 연관 경로는 수도권에 편중되어 있는 현황을 알 수 있다. 따라서, 비수도권역 내의 관광지 간 연계 강화 방안 마련이 시급하다.

4. 결론

본 연구에서는 데이터 마이닝 기법 중, 연관분석을 사용하여 방한 외래관광객의 관광 행태를 분석하고 도출된 데이터를 기반으로 코로나 19 종식 이후 한국 관광 품질 향상을 위한 방안을 제시하고자 하였다. 방한 외래관광객의 방문 관광지 데이터를 구축하기 위하여, 문화체육관광부에서 제공하는 외래관광객 실태조사 원자료를 정제 및 가공하였고, 정제 및 가공 과정에서 빈도분석으로 연도별 주요 방문 관광지를 추출하여 대략적인 관광 행태를 파악하였다. 빈도분석을 통하여 '인기 관광지의 수도권 편중', '각 관광지의 방문 순위 변화', '지역별 관광 현황 변화'와 같은 시사점을 도출하였으며, 연관분석 결과 데이터에서 도출한 시사점과 조합하여 추후 방한 외래 관광객이 한국 관광 품질 향상을 위한 개선방안을 제시하였다. 개선방안은 '주요 관광지와 타 관광지와의 접근성 강화로 관광지 간의 연계성 강화 유도', '관광객들의 관광 욕구 파악을 기반으로 다양한 관광자원 개발', '방문객이 감소하는 관광지의 현황 분석 및 대책 마련', '지방권역의 관광지 간 연계 강화 방안 마련'으로 총 4가지의 개선방안을 제시하였다.

본 연구에서는 데이터마이닝 기법 중 하나인 연관분석으로 방한 외래관광객의 관광 행태 동향을 분석하여 한국관광의 개선방안을 제안하고자 하였다. 분석 결과를 바탕으로, 본 연구 결과는 추후 방한 외래방문객의 만족도 향상을 위한 관광 정책, 관광 자원 개발 등의 부문에서 기초 자료로 활용될 수 있을 것으로 기대된다.

References

Kim, D. J., Choi, H. J., Cho, H. K., Kim, K. Y., (2020), A Study on the Development of the Tourism Industry of Post-COVID-19 Pandemic, *Korean Journal of Hospitality & Tourism*, 29, 4, 129-144. (in Korean) https://doi.org/10.24992/KJHT.2020.06.29.04.129.

Kim, Y. S., Jin, H, J., (2021), Analysis for Economic Impacts of the Tourism/Culture and Whole Industry due to Reduced Inbound Visitors after the COVID-19 Outbreak, *Journal of Tourism Sciences*, 45, 3, 33-53. (in Korean) https://doi.org/10.17086/JTS.2021.45.3.33.53

Lee, Y. S. (2022), Changes in Interests in the Korean Traveling Industry of International Visitor at the Post-Covid 19 by Analysis of Frequency in searching keyword. *Journal of the Korea Institute of Information and Communication Engineering*, 26, 9, 1399-1404. (in Korean)

Ryu, G. Y., Moon, Y. S., Choi, S. D., (2006). A Case Study on Data Mining Techniques, *Journal of The Korean Data Analysis Society* 8, 3, 1021-1033. (in Korean). https://doi.org/10.6109/jkiice.2022.26.9.1399

케이스탯리서치(Kstat)

sas viya
A faster, more productive AI and analytics platform
What if you could make decisions more decisively? And your team could work more productively?
You can with SAS® Viya®. SAS is committed to creating technology that is not only collaborative, powerful and intuitive, but also ethical, equitable and sustainable. Together, we can build a better, more productive future for all.
sas.com/viya
SAS ®Ondemand for Academics
· SAS 클라우드 기반 소프트웨어를 무료로 체험해보십시오.
· 먼저 SAS Profile을 생성하고 Sign Up 해주십시오.
SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries.
® indicates USA registration. Other brand and product names are trademarks of their respective companies. © 2023 SAS Institute Inc. All rights reserved. G252637.0123

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

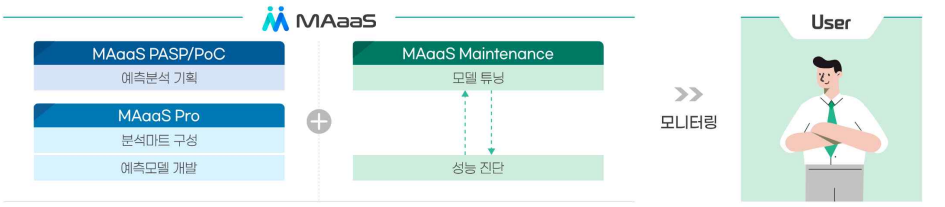
수요 예측	제조 | 수요예측 분석 시스템 구축 유통 | 판매량 예측 모델 개발 물류 | 물동량 예측 시스템 구축 공공 | 열 수요 예측 모델 개발	텍스트 분석	통신 | VoC 분석 금융 | 텍스트 분석을 활용한 평판 분석 유통 | 텍스트 분석을 통한 마케팅 증권 | 챗봇 메시지 분석
상품 추천	유통 | AI 상품 추천 프로그램 구축 유통 | 온라인 쇼핑몰 개인화 상품 추천 모델 개발 유통 | 개인화 추천 서비스 모델 구축 금융 | 고객관리 시스템을 통한 상품 추천	품질 분석 및 예지보전	제조 | 취약 부위 예측 모델 개발 제조 | 특정 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