The Journal of the Convergence on Culture Technology (JCCT)
Vol. 10, No. 3, pp.145-154, May 31, 2024. pISSN 2384-0358, eISSN 2384-0366

http://dx.doi.org/10.17703/JCCT.2024.10.3.145

JCCT 2024-5-18

행정 빅데이터 환경에서 컷오프-투표 분류기를 활용한 빅데이터 예측모형의 실험

Operation Plan of Big Data Prediction Model using Cut-off-Voting Classifier in Administrative Big Data Environment

이우식*

Woosik Lee

요 약 행정 빅데이터를 활용하는 예측 모형을 운영하기 위해서는 정책의 변화 및 변동성 심한 데이터의 특성이 고려가 되어야만 한다. 이런 상황을 고려하여 본 연구에서는 Cut-off Voting Classifier(CVC) 알고리즘을 제안한다. 제안하는 알고리즘은 여러개의 약 분류기를 활용하여 적중률이 급격하게 하락하는 것을 방지하는 알고리즘이다. 본 연구에서는 제안하는 알고리즘을 실험을 통해 성능을 검증한다. 성능검증 결과 급격하게 예측모형 적중률이 하락하는 상황에서도 안정적으로 예측률을 유지한다는 것을 입증할 수 있었다.

주요어 : 빅데이터, 머신러닝, 인공지능, Voting Classifier

Abstract In order to operate predictive models utilizing administrative big data, it is crucial to consider policy changes and the characteristics of highly volatile data. Considering this scenario, this study proposes the Cut-off Voting Classifier (CVC) algorithm. This proposed algorithm prevents a sharp decline in accuracy by utilizing multiple weak classifiers. The study validates the proposed algorithm's performance through experiments. The performance evaluation demonstrates the ability to maintain stable prediction rates even in situations with a sharp decline in predictive model accuracy.

Key words : Big data, Machine Learning, Artificial Intelligence, Voting Classifier

Ⅰ. 서 론

한국사회보장정보원에서는 '15년 12월 이후 복지사각지대 발굴관리시스템 오픈한 이후 지금까지 6년 이상 AI 기반 예측모델을 운영해오고 있다. 사회보장분야에서 복지 분야와 AI 분야가 접목되어 사회 취약계층을 위해 찾아가는 서비스라는 측면에서 매우 혁신적으로 알려져 있다.

복지사각지대 발굴관리시스템은 1년에 6회 운영이 되고 있으며, 2달에 한 번씩 적재되고 있는 데이터를 기계학습과 분석에 활용하고 있다. 하지만 2달에 한 번씩 적재되고 있는 데이터의 수집 주기는 상의하며, 지

*정회원, 한국사회보장정보원 사회보장정보연구소 연구센터 부연구위원 (제1저자)
접수일: 2024년 3월 4일, 수정완료일: 2024년 4월 10일
게재확정일: 2024년 4월 20일

Received: March 4, 2024 / Revised: April 10, 2024
Accepted: April 20, 2024
*Corresponding Author: wslee@ssis.or.kr
Research Center, Korea Social Security Information Service, Korea

자체로부터의 조치결과 또한 상이하므로 시점에 대한 고려가 매우 중요한 시스템이다.

이러한 이유로 한국사회보장정보원과 같은 중앙 기관에서 서로 다른 기관에서 수집되는 행정 데이터를 안정적으로 수집하여 예측에 활용하기 위해서는 최적의 모형을 개발을 하는 것도 중요하지만 안정적인 모형의 운영을 하는 것 또한 매우 중요한 요소라고 할 수 있다.

예측모델이 안정적으로 운영되어야 하는 또 다른 이유는 한국사회보장정보원에 수집되는 데이터는 2달에 한 번씩 입수될 때 사회적 이슈 또는 정책적 상황에 따라 달라질 수 있으며, 수집되는 데이터의 종수가 늘어날 때 과거 비슷한 데이터가 있으면 다중공선성으로 인해 과거 개발된 예측모델이 매우 흔들리는 경우가 발생할 수 있다. 이러한 상황은 행정 데이터를 많이 다루고 있는 타 기관에서도 비슷한 상황이라고 생각이 든다. 예를 들어 건강보험공단의 부정적 수급자를 예측하는 등이 있을 수 있다.

본 논문에서는 행정 데이터 또는 불완성이 높은 데이터를 다루는 기관에서 예측모델을 운영할 때 안정적으로 운영하기 위해 고려해야 할 사항 등을 정리해본다. 그뿐만 아니라 실제 데이터 기반으로 예측모델을 안정적으로 운영하기 위한 대상자 선정방법인 컷오프-선출 분류기 알고리즘을 제안한다. 또한, 본 논문에서는 제안하는 알고리즘을 시뮬레이션을 통해 검증하고자 한다.

II. 이론적 배경

1. 행정 빅데이터와 관련기관 및 활용연구

행정 빅데이터는 업무 또는 사업운영 과정에 생성되는 행정 절차상의 데이터 집합이다 [1]. 대부분 공공분야에서 생성되며[2], 자료수집대상이 전국민 또는 대규모 인구집단이므로 양적 측면에서 이미 빅데이터라 볼 수 있다. 때에 따라 특정 서비스를 제공하기 위해 각 행정 기관이 보유한 데이터를 연계·통합한 것을 행정 빅데이터라 하기도 한다 [3]. 행정 빅데이터는 별도의 자료수집, 저장 과정 없이 기존에 축적된 데이터를 활용할 수 있다는 장점이 있다. 그러나 분석이나 사업운영 외 활용을 위해 수집된 데이터가 아니므로 빅데이터의 활용 시 데이터 현황 파악, 분석데이터로의 가공(cleaning) 과정 등이 중요하다. 이와 관련하여 「데이터기반행정 활성화에 관한 법률(약칭: 데이터기반행정법)」이 '20년 12월 10일 첫 시행 되었다. 이후 행정안전부를 중심으로 공공에서 생산하는 빅데이터의 활용을 촉진하고 있다 [4]. 국가에서 제공하는 공공 빅데이터의 현황과 내용은 '공공데이터 포털(Data.go.kr)에서 일괄 확인이 가능하다. 데이터의 현황을 제공하는 것 외에도 단일 기관에서 보유한 빅데이터를 일부 가공하여 연구자 또는 분석자에게 제공하는 경우(국민건강보험공단, 건강보험심사평가원), 다기관 보유 데이터를 연계·통합하여 AI 기반 예측에 활용하는 경우(한국사회보장정보원, 한국교육학술정보원) 등이 있다.

다음으로 행정 빅데이터 예측모델 기반으로 서비스 모델인 국민연금의 장애연금 및 유족연금 부정수급자 예측모형 [5], 클린센터의 부정수급 모형 [6], 건강보험공단의 체납자의 체납 징수 가능성 사례에 대해 살펴보고자 한다 [7].

차경엽 [5]은 국민연금 부정수급 유형의 전체자료 중 제3자 가해로 인한 장애연금 및 유족연금 수급자(손해배상금 불성실 신고 대상)를 대상으로, 로지스틱 회귀모형, 인공신경망모형, 의사결정나무 모형에 적용하여 분석하였다. 분석 결과 국민연금 부정수급자 예측모델로는 의사결정나무 모형의 예측력이 가장 우수하게 나타났다. 이 예측모델을 이용하여 수급권자가 연금 청구 시 위험이 크게 나타나면 심층심사를 통해 부정수급을 예방할 수 있도록 하였다.

김영선 외 3명 [6]은 클린센터에 접수된 부정수급 제보 건에 대한 현장점검 자료로 이상 결제 모니터링 유형에 대한 전체자료, 이상 결제 자료, 정상/비정상 자료, 장애인활동지원서비스 자료를 수집하여, 로지스틱 회귀모형, 신경망모형, 의사결정나무 모형과 위 3개의 결과를 종합하여 앙상블모형으로 부정수급 가능성이 큰 결제 건을 판별하고, 오차를 감소시키는 결과를 제시하였다.

나영균 외 4명 [7]은 건강보험공단 자체 데이터베이스 체납자의 인구사회 경제적 특성 정보인 나이, 총소득, 체납빈도, 미납빈도, 성별, 생계가입자 여부와 체납처분 여부, 보험료 경감 여부 등의 데이터를 기반으로 로지스틱회귀 모형을 적용하여 체납 징수 가능성 예측모델을 개발하였다. 이 예측모델을 활용하여 체납자 특성 맞춤형 개선방안인 징수전략을 수립하고 최저보험료 정부 지원 및 지역가입자 보험료부과체계 보완책을

제시하였다.

그 밖의 중앙정부와 지방정부의 빅데이터를 활용하여 부정수급 및 범죄 위험 등 예측사례를 살펴보도록 하자. 본 논문에서는 기획재정부 [8], 행정안전부 [9], 경찰청 [10]에서 활용하는 사례를 조사하였다[11].

기획재정부 [8]는 보조금을 지금 받은 사업자가 거래를 취소하는 사례를 축적하여, 빅데이터 통계 모델을 구축하였다. 이를 활용하여 부정수급 행동과 연관성이 큰 의심점수를 형성하여, 이를 기준점으로 지속 적용해 유사한 부정수급 의심 사례를 찾아내는 방식이다.

행정안전부 [9]와 고용노동부는 공공 빅데이터를 통해 실업급여 부정수급 방지 방안을 마련하였다. 부정수급 조사관을 인터뷰하여 부정수급 적발 노하우와 개선사항을 파악하고 분석하여 새로운 유형의 부정수급 패턴을 발굴하였다. 이를 통해 실업급여 신청자 및 사업장의 위험 점수를 측정하여 부정수급 우선순위 리스트를 제공하여 부정수급 적중률을 향상시켰고, 적발 모델을 발전시켜 실업급여 부정수급 예방률을 높일 계획이다.

경찰청 [10]은 치안데이터와 행정 데이터 기반으로 범죄 위험도 예측분석시스템을 활용하여 시간대별 위험등급 예측치를 도출하였다. 과거 범죄 발생 및 112 신고 건수, 유흥시설 수, 교통사고 수, 경찰관 수 등 치안데이터와 인구, 기상, 요일, 면적, 실업·고용률, 건물 유형과 노후도, 공시지가 등의 공공데이터를 종합적으로 활용하였다. 범죄예측 건수와 실제 발생 건수 비교 결과 정확도가 83.1%로 높게 나타났다.

서울시는 성별, 나이, 이동유형, 이동시간, 이동인원 등의 행정 빅데이터, 한국교통연구원의 교통 다각형(출발지, 도착지 행정동 데이터) 데이터, kt의 휴대전화 LTE+5G 신호 데이터를 융합하여 서울시 생활 이동 데이터를 개발하고, 향후 교통정책, 주택정책의 기초자료로 활용할 예정이다 [18][19].

위의 선행연구와 정부 사례에서처럼 행정 데이터를 활용하여 효율적인 재정 절감, 국민의 삶의 질 향상, 정책 의사결정 지원 등의 효과가 있었다. 향후 데이터를 활용한 맞춤형 서비스 제공, 데이터 간 융·복합도 지속해서 필요하며, 민간데이터 활용도 검토할 필요가 있다.

III. 행정 빅데이터 활용 시 고려해야할 사항

행정 빅데이터를 활용하여 기계학습 알고리즘을 활용하기 위해서는 행정 빅데이터의 각종 특성을 이해한 이후 활용을 해야만 운영단계에 가서 문제가 발생하지 않는다. 본 장에서는 비주기적으로 신규 변수가 추가되거나, 다양한 기관으로부터 자료가 수집될 때 발생할 수 있는 각종 문제점 분석 및 고려해야 할 사항을 자세히 보고자 한다. 이를 위해 본 논문에서는 실제 행정빅데이터 기반의 서비스를 운영하고 있는 운영담당자 밀첩 인터뷰 및 업무프로세스 등을 참고하여 고려사항을 선정하였다.

이를 위해 본 논문에서는 다양한 기관으로부터 데이터가 특정 조건으로 A 기관에 수집되는 상황을 바탕으로 고려해야 할 사항 등을 정리하고자 한다. 그림 1은 외부 기관으로부터 중앙 수집기관인 A 기관의 자료가 수집되는 예시를 보여준다.

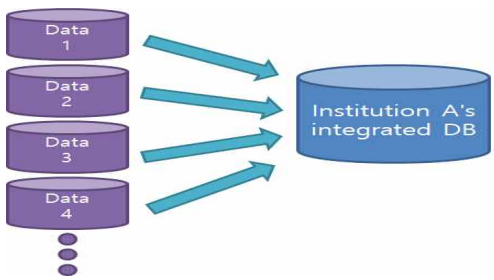


그림 1. A 기관의 자료가 수집되는 예시
Figure 1. Example of data collected from organization A

이러한 상황에서 본 논문에서는 아래와 같이 7가지 행정 빅데이터 활용 시 고려해야 할 사항들을 제시하고자 한다.

1) 신규 변수 투입 시 기존 변수와의 데이터 불균형
2) 특정 변수의 데이터 크기의 갑작스러운 증가
3) 원천 기관의 자료 품질
4) 피드백 결과에 대한 품질
5) 시계열 특성의 변동성
6) 누적되는 데이터의 고착 패턴과 이외 패턴
7) 데이터의 입수조건과 정책 변화

우선 1) 신규 변수 투입 시 기존 변수와의 데이터 불균형에 관해 설명을 하고자 한다. 현재 A 기관에서 4개

의 자료를 수집하고 있는 상황에서 새로운 정보가 입수되어 5개의 자료가 수집되는 상황에서 통합 DB에 누적된 데이터는 타 정보보다 현저하게 부족한 정보를 가지고 있으므로 해당 데이터가 누적되어 활용되기까지 정보 수집이 필요하다.

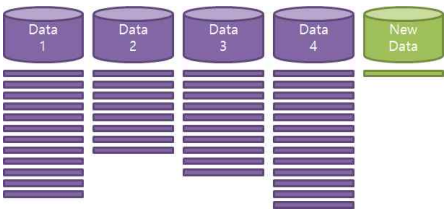


그림 2. 불균형 데이터의 예시
Figure 2. Example of imbalanced data

2) 특정 변수의 데이터 크기의 갑작스러운 증가의 경우는 그림 3과 같이 볼 수 있다. 시간 1과 시간 2에서는 비슷한 수준의 데이터가 입수되어 기존에 구동되고 있는 분류기가 잘 작동이 되었다. 하지만 시간 3에 데이터 2의 갑작스러운 데이터 증가로 인해 기존 분류기가 잘 작동이 안 될 수도 있다.

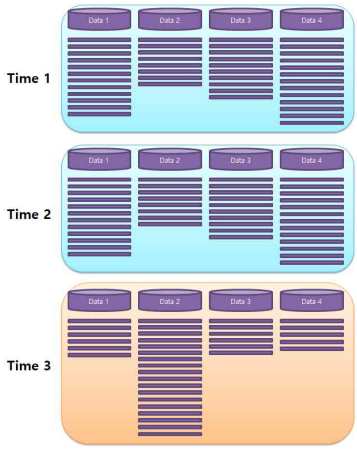


그림 3. 갑자기 증가한 데이터의 예시
Figure 3. Example of suddenly increased data

3) 원천 기관의 자료 품질의 경우 데이터를 송신하고 있는 외부 기관의 데이터 품질을 의미한다. 이러면 데이터 수신을 하는 A 기관에서는 데이터의 품질을 수정하기 위해서는 전처리 과정에 문제가 있는 자료를 수정하거나 원천 기관에 자료의 품질 향상을 위한 조치를 해주라는 요청과정을 수행해야 한다.

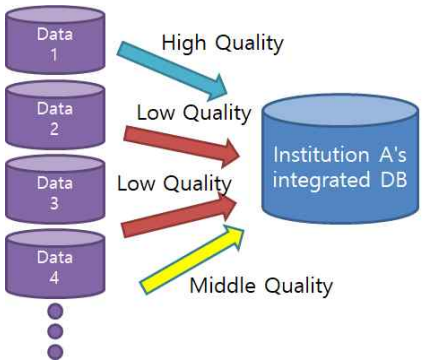


그림 4. 서로 다른 품질을 가지는 데이터 수집의 예시
Figure 4. Examples of data collection with different qualities

4) 피드백 결과에 대한 품질의 경우 데이터의 정답을 입력할 때 라벨링의 품질에 대한 것을 의미한다. 즉, 기계학습 알고리즘이 감독학습을 할 때 정답을 입력할 때 정답에 대한 품질을 말한다. 이때 원천 기관 자료의 품질과 마찬가지로 행정 데이터의 경우 사람이 정답을 입력하기 때문에 사람이 잘못 입력을 한 오류에 대한 품질 관리를 수행해야 한다.

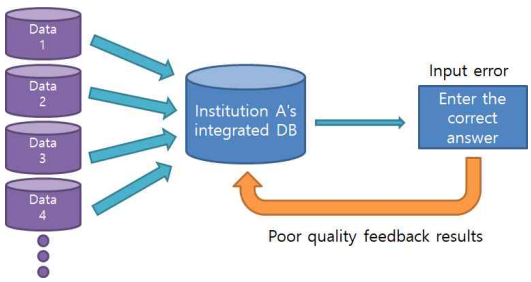


그림 5. 피드백 데이터에 대한 질 낮은 품질 예시
Figure 5. Examples of poor quality feedback data

5) 시계열 특성의 변동성의 경우 2번째 행정 데이터의 특징과 비슷하게 시간이 지남에 따라서 데이터의 값의 변화가 많이 발생하는 경우를 의미한다. 즉, 사람의 소득 관련 데이터 또는 건강 상태와 같이 변동성이 큰 값의 경우에 시간에 따라 변동이 크게 발생할 수 있는 값이다. 이러한 이유로 변동성이 높은 데이터를 분류기에 활용하기 위해서는 과적합이 많이 되지 않고 데이터 변동성이 고려된 분류기가 더욱더 적합하다.

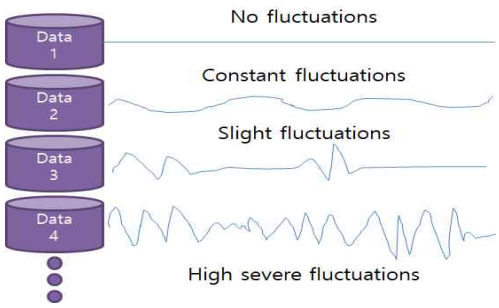


그림 6. 피드백 데이터에 대한 질 낮은 품질 예시
Figure 6. Examples of various types of data based on time series

6) 누적되는 데이터의 고착 패턴과 이외 패턴의 경우 데이터마다 특징이 있고 각 데이터를 보유하고 있는 칼럼의 경우 특정 형태를 보일 수 있는 이야기다. 즉, 데이터 1과 데이터 2의 값이 특정한 값을 가지는 경우 그리고 특정한 값이 반복적으로 발생하는 경우 하나의 패턴이 되어 앞으로도 같은 패턴이 나오면 같은 결과가 나올 수 있을 가능성이 커진다는 이야기다. 그러므로 시계열 패턴을 활용하는 방식의 분석은 단순 시점을 분석하는 것에 비해 많은 정보를 획득할 수 있다.

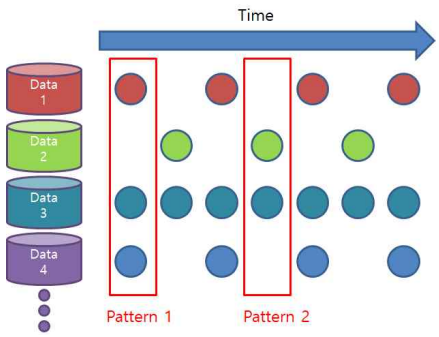


그림 7. 데이터 특성에 따라 반복적으로 발생하는 다양한 패턴 예시
Figure 7. Examples of various patterns that occur repeatedly depending on data characteristics

7) 데이터의 입수주기와 정책 변화는 행정 데이터가 다른 데이터와 큰 차이가 있는 점이라고 할 수 있다. 행정 데이터의 경우 업무 또는 사업의 운영 과정에서 생성되는 데이터로 정책의 변화에 따라 업무와 사업이 변경되기 때문에 정책의 영향을 매우 많이 받는다. 따라서 행정 데이터를 활용하여 기계학습에 적용하기 위해서는 정책의 변화를 살핀 후 정책의 변화 전후의 데이터 변화와 기계학습 학습 결과 비교 분석을 수행해야 한다. 예를 들어 입수기준이 거주금액이 1억 원이 되어야만 들어오는 데이터가 있다고 가정하자, 지난 차수까지 잘 입수가 되었다. 하지만 정책의 변화로 인해 거주금액의 가격대가 4천만 원으로 줄었다고 하면 기존에 잘 입수되었던 대상자가 줄어들고 전체 금액의 최소 최댓값 또한 변동이 있을 것이며, 이런 결과로 기계학습 알고리즘은 잘 작동이 안 될 것이다.

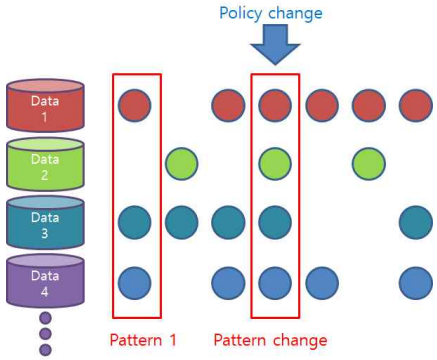


그림 8. 정책 변화에 따라 특정 패턴이 변경되는 예시
Figure 8. Example of a specific pattern changing due to policy changes

이상 7가지 행정빅데이터의 특성에 대하여 살펴보았다. 표 1은 7가지 행정빅데이터의 특성에 대하여 요약한 표를 보여주고 있다. 특성은 크게 불균형, 품질, 데이터로 구분할 수 있으며, 각 내용에 따라 데이터의 모수의 차이, 품질, 시계열 및 패턴의 생성 등이 존재할 수 있다. 이것에 따라 다양한 해결방안이 존재하겠지만 기본적으로 많이 활용되고 있는 샘플링 기법, 알고리즘, 시계열 알고리즘 적용 등이 대표적인 해결 방안으로 고려해볼 수 있다. 본 연구에서 제안하는 방식은 행정 빅데이터 활용시 모든 고려사항을 해결해 줄 수는 없지만, 제시하는 방법은 행정 빅데이터로 모형을 운영할 때 매우 안정적인 운영을 수행하면서 낮은 적중률 하락을 발생하고, 많은 특성을 해소할 수 있는 특징을 가질 수 있다는 장점이 있다.

표 1. 행정빅데이터 활용시 고려사항
Table 1. Considerations when using administrative big data

특성	설명	해결방안
불균형	특정DB의 모수 부족함	오버샘플링
불균형	특정DB의 모수 과도함	언더 샘플링
품질	불균형 데이터 품질	데이터 정제
품질	피드백결과 품질 낮음	데이터 정제 및 지자체 교육
데이터	다양한 시계열의 데이터 유형	시계열데이터 전용 분석기
데이터	반복적 발생 패턴	패턴 기반 알고리즘
데이터	정책에 따른 데이터 변화	정책 반영을 위한 별도 알고리즘

IV. 다수 분류기를 활용한 예측모델의 효율적 운영을 위한 컷오프-다수결 알고리즘(CVC)

본 장에서는 행정 데이터와 같이 불안전성이 매우 높은 데이터를 활용하는 빅데이터 기계학습 알고리즘이 효율적으로 운영이 되려는 방안을 고민하고 제시하고자 한다. 이를 위해 본 연구진은 다수 분류기를 활용하려는 방안 제시를 한다.

제시하는 운영방안은 다수의 분류기를 활용하여 변동성이 심한 행정 데이터 환경에서 빅데이터 예측모형을 안정적으로 운영하는 방법이다.

특히 정책적 지자체 공무원의 수 및 최대 대상자의 한계로 인한 특정 컷-오프 지점인 x%를 활용할 때의 특정 상황을 본 연구에서는 가정하고자 한다. 이런 상황은 위험 점수 분포가 존재할 때 상위 점수 대상자 일부만 내려보내는 방식이다. 즉, 새로운 데이터 입수에 따라 극소수 대상자만 선출할때의 해결 선정 방식을 의미한다.

예를 들어 전체 100만 명의 대상자 중에서 상위 5% 대상자는 5만 명의 대상자가 된다. 이때 기계학습 알고리즘의 위험 상위 5만 명이 대상자가 되고, 컷-오프 선은 5만 명이 된다.

본 논문에서는 상위 위험 대상자의 컷-오프와 다수결 분류기(Voting Classifier) [20]의 앙상블 알고리즘이 융합된 컷-오프 다수결 분류기(CVC: Cut-Voting Classifier) 알고리즘을 제안한다. 컷-오프 다수결 분류기 알고리즘은 고위험 대상자를 선정하기 위해 상위 위험 대상자를 컷-오프만큼 추출한 이후, 하드 보팅(Hard Voting) 선정자로 설정한다. 하드 보팅 선정자로 설정하게 되면 다수의 분류기에서 동시에 발생하는 대상자와 그렇지 않은 대상자로 분류된다. 이때 다수의 분류기 모두 선정되는 대상자 타입-1(type 1: 전체중복)'과 일부 분류기에서만 선정되는 대상자 타입-2(type 2: 일부 중복)이 존재할 수 있으며, 단일 분류기에서만 선정되는 대상자 타입-3(type 3: 미중복)가 존재할 수 있다.

예를 들어 3개의 분류기를 활용한다면 그림 9와 같이 볼 수 있다. 3개의 분류기가 모두 겹치는 곳은 전체적으로 겹치는 곳으로 하드 보팅 대상자로 선정될 수 있다. 그리고 2개의 분류기만 겹치는 곳은 일부 대상자가 겹치는 곳으로 소프트 보팅(Soft Voting) 방식을 활용하여 대상자를 선정하게 된다. 만약 최종 대상자의 수가 부족할 경우 단독 분류기 기반으로 대상자를 선정하게 된다.

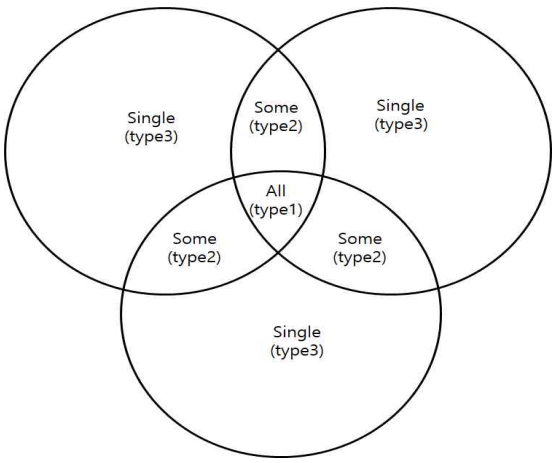


그림 9. 다수 분류기를 활용하여 Voting을 수행하는 방식
Figure 9. Voting method using multiple classifiers

컷오프 투표 알고리즘인 컷-오프 다수결 분류기가 순서도는 그림 10과 같이 볼 수 있다. 우선 분류기에 입력될 수 있는 입력데이터를 호출한 이후 분류기의 개수를 설정한다. 다음으로 분류기에서 선정하게 될 데이터의 개수 상위 컷오프를 설정한다. 모든 설정이 완료되면, 다수의 분류기를 구동한 이후 각 분류기의 대상자를 컷오프에 맞춰서 선정하게 된다. 이때 모든 분류기의 중복되는 대상의 타입-1이 컷오프 수를 넘기면 모든 대상자를 선정하는 하드 보팅 방식으로 선정한다. 그렇지 않고 일부 대상자를 포함해야 컷오프를 넘기면 타입-1 하드 보팅 방식으로 대상자를 선정하며 타입-2의 경우 소프

트 보팅 방식으로 대상자를 선정한다. 만약 중복되는 대상자를 모두 선정하고도 남는 대상자가 존재한다면 타입-1과 타입-2를 모두 선정한 이후의 타입-3은 상자를 단일 분류기 비율에 맞게 선정하게 된다.

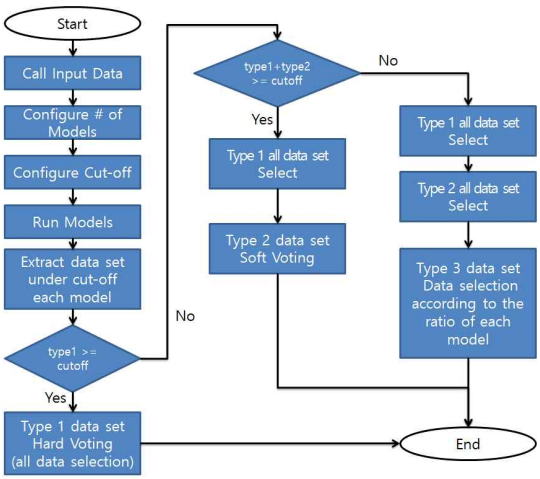


그림 10. 컷오프-다수결-알고리즘 순서도
Figure 10. Cutoff-Majority-Algorithm Flowchart

타입-3의 경우 단일 분류기의 대상자를 비율로 선정하는 방식으로 분류의의 개수에 따라 비율이 달라질 수 있으며, 가장 최적의 비율은 서비스의 목적, 데이터, 기계학습 알고리즘 등에 따라 변경될 수 있다.

본 논문에서는 3개의 분류기를 활용하는 환경에서 기계학습 알고리즘을 효율적으로 운영하기 위한 비율로 5:3:2 비율을 활용한다. 50% 비율을 차지하는 분류기는 가장 성능이 좋은 비율이고, 30% 비율을 차지하는 분류기는 성능의 개선이 필요한 비율이며, 20% 비율을 차지하는 분류기는 신규로 도입하는 분류기 또는 성능이 저조하여 폐기하는 단계에 있는 분류기다. 그림 11은 타입-3의 단일 분류기의 대상자 분류기의 비율을 결정하는 예시를 보여주고 있다.

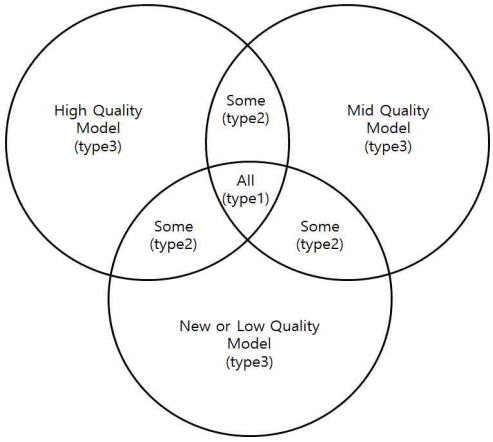


그림 11. 단일 분류기의 대상자 비율을 가지는 구조
Figure 11. Structure with a proportion of subjects for a single classifier

V. 컷오프-다수결 알고리즘의 실험 및 분석

컷오프-다수결 알고리즘의 성능을 파악하기 위해서 본 논문에서는 단일 모형에서 성능이 좀 떨어지는 모형으로 변경하였을 때 행정 데이터의 특징을 고려하지 않고 100% 변경하였을 때의 상황과 행정 데이터의 특징을 고려하여 컷-오프 다수결 분류기 알고리즘을 적용하였을 때의 성능을 비교하고자 한다.

본 논문에서 성능 비교를 하기 위해 분류기의 수를 3개로 한정하여 시뮬레이션을 진행한다. 이때 데이터의 전체 수는 41,738개로 다수의 칼럼을 가지고 있는 데이터로 가정하여 시뮬레이션을 진행하고자 한다. 그리고 분류기의 적중률은 80%, 50%, 20%로 설정하여 실험을 진행한다. 적중률이라고 하는 것은 실제 대상자를 대상자로 예측이 성공적으로 이루어졌을 때의 비율을 말한다. 즉, 100명 중에 100명 모두 정답이면 100%를 말하고, 80명만 정답이면 80%라고 하는 것을 의미한다.

표 2. 실험 환경
Table 2. Experimental Environment

Category	Contents
Models	Three
Data set	417,38
Hit Ratio each Model	80%, 50%, 20%
Cut off	Top 10%
Extract data	41,738

본 연구에서는 80% 예측모델로 잘 구동되고 있는 상황에서 예측력이 떨어지는 20% 예측모델로 변경해야 하는 상황에서의 표 5와 같이 시나리오 3개를 고려하였다. 시나리오는 그림 7 또는 8과 같이 AS-IS 환경에서 정책 또는 데이터 특성의 변형으로 적중률이 변동되어 TO-BE로 변경되는 상황을 의미한다.

표 3. 3가지 시나리오
Table 3. Three scenarios

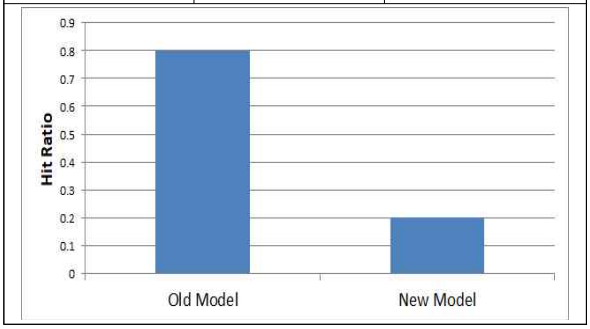
특성	AS-IS	TO-BE
시나리오1 (1개분류기)	80% 적중률 분류기 활용	20% 적중률 분류기 활용
시나리오2 (2개분류기)	80% 적중률 분류기 활용	20% 적중률 분류기 활용시 1:1 비율적용
시나리오3 (3개분류기)	80% 적중률 분류기 활용	80%, 50%, 20% 분류기 모두 활용

첫 번째 시나리오의 경우는 표 4와 같이 볼 수 있다. 우선 상위 80% 적중률을 보이는 분류기가 20% 적중률을 보이는 분류기로 변경하였을 때의 실제 운영 환경에서 60% 적중률 하락이 발생하는 최악의 경우가 생긴다. 이럴 때 백업이나 실제 대상자를 바탕으로 수행되는 기계학습의 경우 곤란한 상황이 발생할 것이다.

첫 번째 시나리오 결과 모든 대상자 중복인 6,237명에 대해 대상자가 고려되지 않은 상황에서 변경이 이뤄진다는 것을 확인할 수 있었으며, 그림 12와 같이 적중률의 큰 폭의 하락으로 인한 곤란한 상황이 발생한다는 것을 확인할 수 있었다.

표 4. 시나리오 1의 실험결과
Table 4. Experimental result of scenario 1

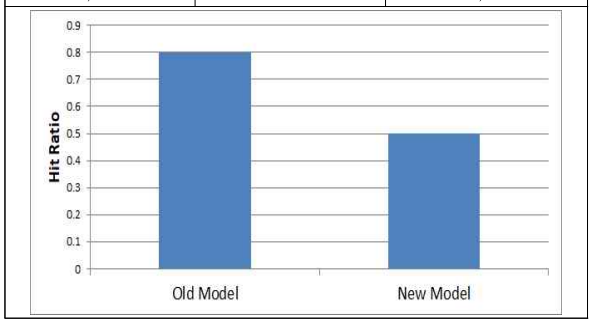
타입-1	타입-2	타입-3
(6,237)	0	41,738



다음으로 시나리오 2 분류기 2개를 활용하여 1:1 비율로 추출하는 경우를 고려해보자. 이 경우 시나리오 1과 거의 비슷하지만 타입-1이 고려되어 적중률에 많은 대상자가 들어갈 수 있다는 것을 보여주고 있다. 이러한 이유로 전체 적중률이 20%에서 49.8%까지 적중률이 약 30% 상승시킬 수 있는 효과를 볼 수 있었다.

표 5. 시나리오 2의 실험결과
Table 5. Experimental result of scenario 2

타입-1	타입-2	타입-3
6,237	0	35,501



마지막 3번째 시나리오의 경우 타입-1, 타입-2, 타입-3이 모두 활용되는 경우로, 타입-3의 경우 5:3:2 비율로 추출하는 경우를 보여준다. 모든 분류기가 겹치는 경우인 타입-1의 경우 2,331명이, 2개의 분류기를 활용하여 대상자가 겹치는 경우 타입-1을 제외하고 19,088명 그리고 나머지 개별 분류기에서 대상자를 20,319명 추출하는 경우이다. 변경 후 적중률의 향상이 매우 큰 것을 확인하였다. 시나리오 1의 적중률 20%, 시나리오 2의 적중률 49.8%에서 시나리오 3의 경우 적중률이 69.1%까지 올라간 것을 확인할 수 있다.

표 6. 시나리오 3의 실험결과
Table 6. Experimental result of scenario 3

타입-1	타입-2	타입-3
2,331	19,088	20,319

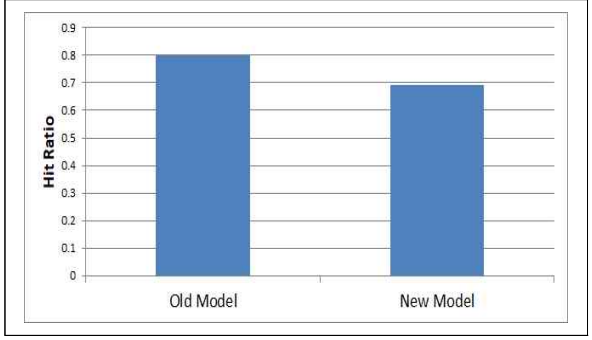


그림 12의 경우 3개의 시나리오의 종합적인 결과 그래프를 보여주고 있다. 결과적으로 모형을 변경할 때 한 번에 변경하는 것에 비해 본 논문에서 제안한 컷-오프 다수결 분류기를 적용한 방식을 활용하는 것이 적중률

하락을 저하하면서 안정적으로 운영할 수 있다는 것을 보여주었다.

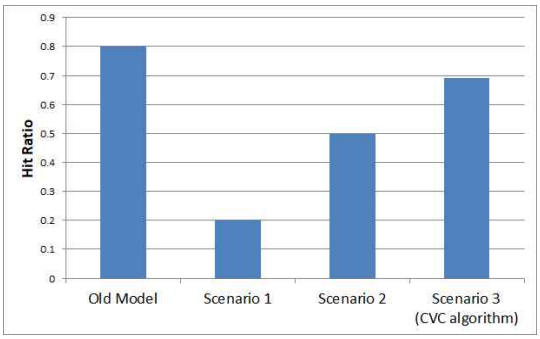


그림 12. 기존 것과 시나리오 1,2,3의 실험결과
Figure 12. Experimental results of existing and scenario 1, 2, and 3

Ⅵ. 결론

본 논문에서는 행정 빅데이터를 활용하여 예측모델을 개발할 때 고려해야 할 사항을 7개의 범주로 분류하여 제시하였다. 제시하는 범주는 주기적으로 행정 빅데이터가 수집되는 환경에서 변동성이 발생할 수 있는 요인으로 행정 데이터를 활용하는 대다수 기관이 비슷한 상황일 것이라 생각이 든다. 즉, 행정 빅데이터 보유 기관에서는 최소 본 논문에서 제시한 7개 이상은 고려해서 예측모형 설계 및 개발을 수행해야지, 그렇지 않으면 향후 발생하는 변동성 높은 상황에 대한 대처가 힘들어질 수가 있다.

본 논문에서는 행정 빅데이터와 같은 변동성이 높은 데이터에 대응하기 위해 컷오프-투표 분류기 알고리즘을 제안하였다. 컷오프-투표 분류기의 경우 기존 컷-오프와 다수결 분류기 선출 방식의 결합 버전으로 다수결 분류기의 하드 보팅과 소프트 보팅 모두를 활용하는 알고리즘이다. 제안하는 알고리즘은 활용하는 데이터가 증가하거나 정책이 변화하는 행정 빅데이터 활용 시 예측력이 급격하게 떨어지는 것을 방지하면서 지속해서 개선할 수 있는 운영을 보장해준다.

본 논문에서는 실제 실험을 통해 20% 적중률로 떨어지는 상황에서 제안하는 컷오프-투표 분류기 알고리즘을 활용하게 되면 적중률 하락을 크게 방지할 수 있음을 보여주었다.

본 논문에서 제시한 알고리즘은 초기 버전으로 좀 더 깊이 있게 개선되어 시계열, 딥러닝 등을 동시에 고려될 수 있는 알고리즘으로 개선할 계획하고 있다.

References

[1] Yoo Jong-seong, Jeon Byung-yu, Shin Kwang-young, Lee Do-hoon, Choi Seong-su. Utilization of administrative data for evidence-based policy research. Korea Social Policy Review. 2020;27(1):5-37
[2] Elias, P. Administrative data: Facing the future: european research infrastructures for the humanities and social sciences, SCIVERO. 2014:47-48.
[3] Woollard, M. Administrative data: problems and benefits. A perspective from the United Kingdom. Facing the future: european research infrastructures for the humanities and social sciences, SCIVERO. 2014.
[4] Big Data Analysis and Utilization Division, Ministry of Public Administration and Security. New design of public service based on public experience with data - Confirmation and announcement of 「First Basic Plan for Data-Based Administration Revitalization (2021~2023)」 1st meeting held-. Ministry of Public Administration and Security, 2021.2.20.
[5] Cha Kyung-yeop, "A study on the development of a model for predicting illegal receipt of the National Pension using data mining - for infidelity reporting of damages -. Proceedings of the Korean Statistical Society," 2010:17(1):1-8
[6] Kim Young-sun, Park Seon-mi, Choi Ki-jung, Park Eun-ang, "Measures to improve social service abnormal payment and illegal supply and demand monitoring system," Korea Social Security Information Service. 2017.12.
[7] Na Young-gyun, Cha Ye-rin, Choi Dae-gyu, Im Seung-ji, Kim Na-young, "A study on improvement of health insurance arrears collection. Health Insurance Corporation," 2020.12.
[8] 6494 cases of illegal subsidies caught by AI of the Ministry of Strategy and Finance... Electronic newspaper with list confirmed by May, 2021.03
[9] Prevention of Unauthorized Receiving of Unemployment Benefit-Public Big Data Part 5-. Ministry of Public Administration and Security, 2017.03.
[10] Expansion of crime prevention activities using police big data and artificial intelligence (AI)

nationwide. Police Agency Press Release, 2021.04
[11] Yujin Gil, Yoon Chung and Sangsoo Park, “Diabetes Prevalence and Diagnosis Rates, and Risk Factor Effect Analysis,” JCCT, 2024.1.31.
[18] Seoul Life Movement Data Manual. Seoul City. Korea Transportation Research Institute, KT, 2021.09
[19] Hye-Sun Lee, “Study of mental disorder Schizophrenia, based on Big Data,” IJACT, 2023.11
[20] F. Mutaher and Ba-Alwi, Comparative Study for Analysis the Prognostic in Hepatitis Data: Data Mining Approach, International Journal of Scientific and Engineering Research, 2013.8.