한국빅데이터학회지
제2권 제2호, 2017, pp. 95-104

RandomForest와 XGBoost를 활용한 한국어 텍스트 분류: 서울특별시 응답소 민원 데이터를 중심으로

Korean Text Classification Using Randomforest and XGBoost Focusing on Seoul Metropolitan Civil Complaint Data

하지은 · 신현철 · 이준기†

연세대학교 정보대학원

요 약

2014년 서울시는 시민의 목소리에 신속한 응대를 목표로 '서울특별시 응답소' 서비스를 시작하였다. 접수된 민원은 내용을 바탕으로 카테고리 확인 및 담당부서로 분류 되는데, 이 부분을 자동화시킬 수 있다면 시간 및 인력 비용이 감소될 것이다. 본 연구는 2010년 6월 1일부터 2017년 5월 31일까지 7년치 민원 사례 17,700건의 데이터를 수집하여, 최근 화두가 되고 있는 XGBoost 모델을 기존 RandomForest 모델과 비교하여 한국어 텍스트 분류의 적합성을 확인하였다. 그 결과 RandomForest에 대비 XGBoost의 정확도가 전반적으로 높게 나타났다. 동일한 표본을 활용하여 업 샘플링과 다운 샘플링 시행 후에는 RandomForest의 정확도가 불안정하게 나타난 반면, XGBoost는 전반적으로 안정적인 정확도를 보였다.

■ 중심어 : 머신러닝, 텍스트마이닝, 민원 분류, RandomForest, XGBoost

Abstract

In 2014, Seoul Metropolitan Government launched a response service aimed at responding promptly to civil complaints. The complaints received are categorized based on their content and sent to the department in charge. If this part can be automated, the time and labor costs will be reduced. In this study, we collected 17,700 cases of complaints for 7 years from June 1, 2010 to May 31, 2017. We compared the XGBoost with RandomForest and confirmed the suitability of Korean text classification. As a result, the accuracy of XGBoost compared to RandomForest is generally high. The accuracy of RandomForest was unstable after upsampling and downsampling using the same sample, while XGBoost showed stable overall accuracy.

■ Keyword : Machine Learning, Text Mining, Civil Complaint Classification, RandomForest, XGBoost

Ⅰ. 서 론

다양한 사람들이 모여 사는 현대 사회에서 분쟁이나 불편 사항을 해결할 수 있는 수단으로 민원 제도를 떠올릴 수 있다. 민원은 '사회 구성원이 행정 기관에 대하여 원하고 요구하는 바를 신청하는 일'로 정의된다. 과거 행정 기관 방문이나 유선으로 이루어지던 민원 접수는 최

2017년 11월 23일 접수; 2017년 12월 25일 수정본 접수; 2017년 12월 26일 게재 확정
† 교신저자 zlee@yonsei.ac.kr

근 정보통신기술의 발달에 힘입어 인터넷 환경에서 언제 어디서든 손쉽게 접근 가능해졌고, 그에 따라 발생되는 민원 양도 크게 증가하고 있다[8]. 이러한 변화에 발맞춰 대부분의 행정기관들은 처리 효율을 높이고자 경찰 민원포털, 자동차민원 대국민포털, 국민건강보험 사이버민원센터, 식품의약품안전처 의약품민원 등 각자의 사이버 민원 접수 페이지를 운영하고 있다.

2014년 3월 5일 서울시는 시민의 목소리에 신속한 응대를 하는 것을 목표로 '서울특별시 응답소'라는 통합 민원관리시스템을 개소하였다. 개설 이전 서울시의 분리된 민원제안시스템으로는 시민들이 필요한 채널을 찾는데 상당한 불편함이 있었고, 시정 관련 지식이 부족한 시민들은 민원 신청 부서를 찾는 것에 어려움을 겪어왔다. 이는 시민의 입장에서도, 서울시의 민원처리 프로세스 측면에서도 비효율적인 형태였다. 따라서 '서울특별시 응답소'의 목표는 서울시정에 대한 응답성과 접근성 증진, 서울시의 공공생산성 향성, 서울시정 운영의 효과성 제고에 초점이 있다. 오픈 이후 민원 처리 기간이 1년 사이에 평균 3.8일에서 2.9일까지 단축되는 효과를 보였으며 시민의 만족도 및 운영 효율성 또한 향상되었다[20].

민원처리 프로세스에는 사람이 직접 담당부서를 할당하고 분배하는 과정이 포함되어 있는데, 이를 자동화시킬 수 있다면 인력 비용을 감소시킬 수 있고 처리 시간 감소에 따른 시민 만족도를 높일 수 있을 것이다. 따라서 본 연구는 이러한 민원처리 프로세스 효율 증진을 위한 텍스트 분류를 목적으로 하며, 분류 알고리즘으로 최근 화두가 되고 있는 XGBoost 모델을 기존 RandomForest 모델과 비교하여 한국어 텍스트 분류의 적합성을 확인하고자 한다.

Ⅱ. 이론적 배경

2.1 문서 분류

문서 분류에 있어 일반적으로 사용되는 분류 모델로는 Naive Bayes, SVM, Decision Tree, Random Forest, Deep Learning 알고리즘 등이다. Tang et al.[29]는 20,000건의 뉴스 기사를 20개의 클래스로 분류하는 과정에서 각 클래스별로 강조된 컨텐츠들로부터 설정된 순서에 기반한 베이지안 접근법을 제시하였다. 김수아, 조희선, 이현아[3]는 8,000건의 블로그 포스트를 수집하여 나이브 베이즈를 통해 클래스를 분류를 시도하는 한편, 어휘빈도, 문서빈도, 분류별 빈도 등의 가중치를 조합하여 TF-CTF-IECDF라는 블로그 포스트에 적합한 가중치 방식을 제시하였다. 손남례, 김서영[4]는 강원도 민원데이터 21,474건을 수집하였고, 나이브베이즈 분류기와 지역별 민원 통계분석, 민원 주제어 분류 및 검출, 민원 주제어간 상관관계 분석을 통해 민원 데이터 자동 분류 시스템을 제안하였다. Amor et al.[10]는 KDD CUP 1999년 출제 데이터를 활용하여 나이브 베이즈 모델과 의사결정나무 모델의 성능을 비교하여 해당 데이터에서 나이브 베이즈 모델보다 의사결정 나무 모델의 성능이 더 높은 것을 확인하였다. Hebert[19]는 대용량 제조산업 데이터 100,000건으로부터 랜덤포레스트와 XGBoost 성능을 비교하였고, 랜덤포레스트가 미세하게 더 높은 성능을 보이는 것을 확인하였다. Fragos et al.[17]는 나이브 베이즈와 Maximum Entropy 분류기법을 결합하여 Reuters newswire에 실린 기사 데이터인 Reuters-21578 문서를 분류하였고, 각각의 개별 분류기보다 높은 성능을 확인하였다. Zhang et al.[30]는 Character-level convolutional networks를 활용하여 AG's News, Sogou News, DBPedia 등 다양한 텍스트 데이터 분류를 시도하였고 n-grams TFIDF, LSTM 등의 다른 기법들과 성능을 비교하였다.

2.2 국내 민원 데이터 관련 문헌 연구

2013년 '공공데이터 제공 및 이용활성화에 관한 법률(이하 공공데이터법)'이 시행되면서 정부적 차원에서 공공정보를 적극적으로 개방하기 시작하였고 공공데이터 포털 운영에(http://www.data.go.kr) 따른 데이터 개방이 대폭 확대되었다[5]. 이러한 개방화에 따라 공공 데이터를 활용한 비즈니스나 연구들이 활발해지고 있고, 본 연구에서 사용하고자 하는 민원 데이터를 대상으로 한 연구도 증가하고 있는 추세이다[4].

김성표 등의 연구와 최해옥의 연구는 민원 주제 별 빈도 분석을 통하여 정책 우선순위를 결정하는 방법을 제시하였다[2, 9]. 원태홍, 유환희는 ARIMA 분석을 통해 전주시 전자민원의 추이를 분석하고 지속적으로 민원이 접수되는 분야를 발견하고자 하였다[6]. 손남례, 김서영[4]은 민원을 통계 분석하여 시각화 하는 방법을 제안하고, K-means 알고리즘을 이용하여 민원 내용에서 추출한 주제어들 간의 거리를 분석하여 민원의 카테고리를 새롭게 재정의 한 뒤 Naive Bayes 분류기를 이용하여 연구 담당부서를 자동 분류하였다[4].

이처럼 기존 문서 분류 연구들은 다양한 영역에서 진행되었으나 보고서, 뉴스기사 등과 같은 띄어쓰기 등 맞춤법이 잘 맞는 비교적 정형화된 텍스트 데이터를 활용한 연구가 대부분이다. 반면 민원데이터는 특성상 약어, 은어 등의 요소들이 들어가 있으며 띄어쓰기가 잘 되어있지 않은 등 맞춤법에 어긋나는 경우가 나타난다는 특징을 가지고 있다. 특히나 한국어는 교착어의 특성을 가지며 입력단위와 형태소 분석 사전과 직접적으로 일치하지 않는 비율이 높아 더욱 복잡한 언어라고 볼 수 있으며 의미 분석이 어렵다[1]. 따라서 본 연구에서는 최근 분류 모델로써 각광받고 있는 XGBoost와 RandomForest 기법을 활용하고 비교하여 민원 내용을 분류함으로써 한국어 비정형 텍스트 분류의 적합성을 확인하고자 한다.

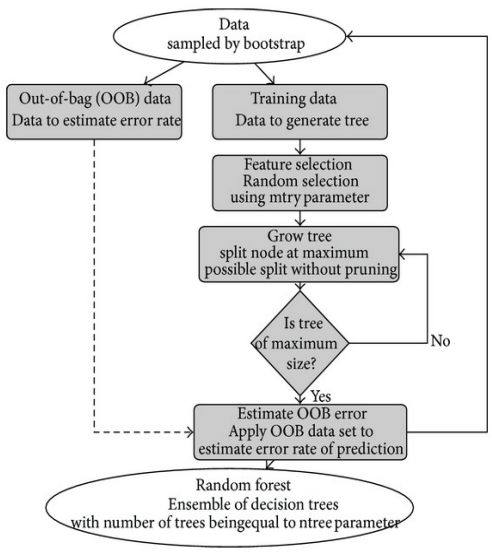
2.3 RandomForest

랜덤포레스트는 의사결정나무와 배깅 모델을 소개했던 Breiman[12]이 2001년 소개한 방법으로 다수의 의사결정 나무를 결합하여 하나의 모형을 생성하는 방법이다. 즉 붓스트랩 표본을 다수 생성하여 의사결정나무 모형을 적용하고 그 결과를 종합하는 방법이다. 다른 앙상블 모형과의 가장 큰 차이점은 붓스트랩 표본을 생성하는 부분에서 임의성(random)이 도입될 뿐만 아니라 의사결정나무 모형 적합 시 각 마디에서 설명변수를 선택함에 있어서도 임의성(random)이 도입된다는 것이다[7]. 이에 따라 랜덤 포레스트는 임의성을 최대로 가지게 되며 이는 결정트리 간 상관관계를 낮추어 예측오차를 줄어들게 한다. 또한 의사결정나무의 수가 증가할수록 예측오차가 줄어들며, 의사결정나무의 수가 많아도 과적합 하지 않는다는 장점이 있는 모델이다[11].

일반적으로 모형 검증 방법으로 데이터를 훈련 데이터와 검증 데이터로 나누어 검증을 하거나, cross-validation 등을 하여 정분류율을 구하는 방법을 사용한다. 반면 랜덤포레스트에서는 데이터를 훈련, 검증 데이터로 나눌 필요 없이 붓스트랩 표본을 생성 할 때, 붓스트랩 표본으로 뽑히지 않은 자료들을 OOB(Out-Of-Bag) 자료라 부르며 이를 검증 데이터 대신 사용하여 모형 검증을 수행할 수 있다[12].

랜덤포레스트는 기존 의사결정나무가 분기 시 한 변수만을 사용하기 때문에 모형의 설명력은 높으나 예측력이 떨어지며 모형 안정성이 떨어지는 단점을 임의성을 최대로 하여 붓스트랩하는 방법으로 해결하였으며, 특히 설명변수가 다수일 때 높은 예측력을 보이며 매우 안정적인 모형을 제공한다[28]. 또한 랜덤포레스트는 최

대우도법 등과 같은 기존의 매개변수를 이용한 기법과 단일의사결정나무, 신경망 등과 같은 기계학습 알고리즘에 비해 보다 정확하고 좋은 결과를 낼 수 있는 대안으로 떠오르고 있다[16, 27]. 많은 연구들이 랜덤 포레스트를 이용하여 분류를 수행하였고, 다른 분류 기법들과 비교하여 우수성을 보여주고 있다[16, 25]. <그림 1>은 랜덤포레스트 알고리즘을 나타내고 있다.



〈그림 1〉 RandomForest 알고리즘[22]

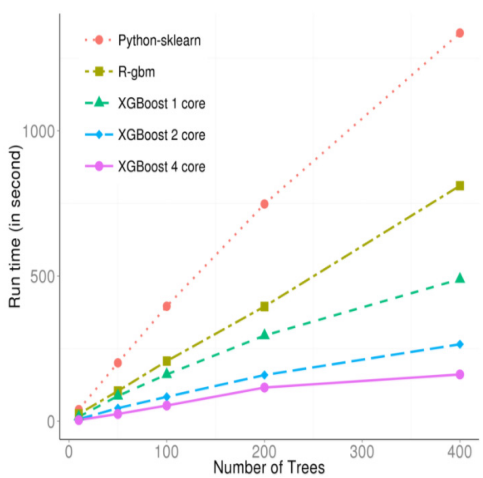
2.4 XGBoost

XGBoost는 선형 모델이나 tree 기반 모델에서의 과적합 문제를 해결하고, 규모가 큰 데이터셋의 안정성과 훈련 속도를 향상시키기 위한 목적으로 Tianqi Chen과 Carlos Guestrin이 소개한 방법이다. eXtreme Gradient Boosting의 약자로 boosting algorithm 기반 모델이며, 회귀와 분류, 순위 및 사용자 정의 objective을 지원하는 유연한 모델이다[14].

Boosting은 sequential 프로세스로 이전 tree로부터 얻은 정보를 다음 tree를 생성하는데 활용하여 약한 개체들을 강한 개채들로 변환시키면서 오차를 교정한다. 일반적인 gradient boosting에서는 tree pruning과정에서 negative loss가 발생하면 그 과정을 멈추는 반면 XGBoost는 모델 수행 시 파라미터로 지정한 max_depth까지 진행한 후 loss function에서 개선이 일정 수준에 못 미칠 경우 역방향으로 pruning과정을 진행한다[14, 18].

XGBoost는 결측치를 내부적으로 자동 처리하며 나무를 생성할 때 병렬적으로 생성하며 새로운 데이터에 대해 이전의 학습 모델 결과를 반영하여 그 성능을 더 향상시키는 방법으로 훈련하는 연속적인 훈련 알고리즘적인 특징을 가지고 있다. 또한 학습하는 동안 모든 CPU 코어들을 사용하는 parallel computing, 메인 메모리 공간에 맞지 않는 데이터를 처리하기 위한 out-of-core computing, 최적의 하드웨어 사용을 위해 데이터 구조 및 알고리즘에 대한 cache optimization과 같은 시스템적 특징을 가지고 있다[14]. 검증 방법에 있어서도 cross validation function을 내장하고 있어 검증이 용이하다[21].

<그림 2>는 기존 다른 패키지와의 속도를 비교한 것이다. Pytnon의 scikit-learn과 R의 Generalized Boosted Regression Models 에 비해 tree의 수가 많으면 많을수록 처리 속도가 훨씬 뛰어나다는 것을 확인할 수 있다.



〈그림 2〉 Tree 수에 따른 처리 속도 비교[15]

즉 처리 속도가 빠르며, 모델의 이전 결과를 활용하여 모델을 계속적으로 개선하고 훈련하는 등의 방법으로 성능이 뛰어나 Google, MS Azure, Alibaba 등 실무에서도 많이 활용되고 있는 모델이며, 데이터 과학자들이 특정 문제를 해결하기 위해 경쟁을 하는 온라인 플랫폼인 kaggle에서 2015년 한 해 동안 Machine Learning Challenge에서 승리한 팀 중 17팀 이상이 XGBoost를 사용하여 성과를 거둔 기록이 있다[14].

Ⅲ. 연구 방법

3.1 데이터 수집

응답소 홈페이지(https://eungdapso.seoul.go.kr)에서 2010년 6월 1일부터 2017년 5월 31일까지 7년치 민원 사례 데이터를 수집하였으며, 총 13개 카테고리 17,700건의 데이터를 수집하였다. 이 중 7년 사이에 일정 기간 동안에만 민원 신청이 있었던 카테고리 5개를 제외한 8개 카테고리의 데이터 11,457건을 대상으로 연구를 진행하고자 하며, 데이터 예시는 <표 1>과 같다.

〈표 1〉 민원 데이터 예시

민원내용	카테고리 (처리부서)
오늘 처음으로 상암동 롯데몰 부지에서 주민 80여 명이 모여 집회를 하였습니다. 방송국에서 취재도 다녀갔고 점차 주민들이 참는데 한계를 …	경제/일자리 (경제진흥본부)
시장님 잘 아시다시피 날로 심각해져 가는 환경오염과 미세먼지로 인해 호흡기질환이 급증하고 국민들의 삶의 질이 급격하게 악화되어가고 …	환경/공원/ 상수도 (기후환경본부)
성수it종합센터에 있는 회사에 다니고 있습니다. 출근 퇴근시에 근처 도로에 인도가 없어서 너무나 괴롭습니다. 인도는 없고 한쪽에는 지정주차 …	교통 (안전총괄본부)
…	… …

민원 카테고리 별 데이터 수는 <표 2>와 같으며 민원 처리부서 별 데이터 수는 <표 3>와 같다.

〈표 2〉 민원 카테고리 별 데이터 수

카테고리	데이터 수
경제/일자리	1,381
교통	1,983
기획/감사/교육	1,364
문화/관광/체육	1,477
복지/어르신/장애인	1,310
안전/소방/민방위	1,223
주택/도시계획/부동산	1,394
환경/공원/상수도	1,325
합 계	11,457

〈표 3〉 민원 처리부서 별 데이터 수

처리부서	데이터 수	처리부서	데이터 수
경제 진흥본부	709	소방재난 본부	301
기획 조정실	439	시민 건강국	212
기후환경 본부	522	시민소통 기획관	1,131
도시 계획국	618	안전총괄 본부	752
도시교통 본부	1,837	여성가족 정책실	564
도시재생 본부	318	주택 건축국	924
문화본부	947	평생교육 정책관	270
복지본부	563	푸른 도시국	634
서울혁신 기획단	249	행정국	467
합 계			11,457

3.2 데이터 전처리

3.2.1 처리부서 통합 작업

본 연구에서 수집한 데이터는 7년 동안의 서울시 민원 사례 데이터이다. 7년 동안 서울시 행정부서 체계 변동으로 처리부서의 명칭이 변경된 경우 및 부서가 통합 된 경우가 존재하였다. 이에 서울시 정보소통광장의 결재문서를 참고하여 통합작업을 진행하였다.

3.2.2 NIADic을 사용한 형태소 분석 및 명사 추출

NIADic은 국립국어원 우리말샘 사전과 브랜드, 유명인, 장소, 신조어 등의 명사 위주로 구성된 SNS 분석 전문 기업인 인사이터에서 구축한 사전을 기반으로 최신 단어로 구성된 형태소 사전이다. 2016년 12월 14일 배포된 NIADic은 약 93만 단어로 구성되어 있으며, 기존 한국어 형태소 사전인 꼬꼬마, 한나눔, 세종사전, 시스템 사전 등 에 비해 단어의 폭이 넓다[23]. 민원 데이터의 특성 상 신조어, 유명인, 장소 등의 내용이 많이 포함되어 있어 NIADic이 적합한 형태소 사전이라고 판단되어 이를 사용하여 형태소 분석 및 명사 추출을 진행하였다.

3.2.3 업 샘플링

업 샘플링은 해당분류에 속하는 데이터가 적은 쪽을 표본으로 더 많이 추출하는 방법으로 카테고리와 담당부서에 해당하는 민원 수의 불균형으로 인해 모델 정확성이 떨어짐에 따라, 데이터 중복 추출 방식의 업샘플링을 통해 데이터 균형 맞춘 다음 모델에 적용시켰다.

3.2.4 다운 샘플링

다운 샘플링은 해당분류에 속하는 데이터가 많은 쪽을 표본으로 더 적게 추출하는 방법으로 카테고리와 담당부서에 해당하는 민원 수의 불균형으로 인해 모델 정확성이 떨어짐에 따라, 다운샘플링을 통해 데이터 균형 맞춘 다음 모델에 적용시켰다.

Ⅳ. 데이터 분석 및 결과

4.1 검증 방법

4.1.1 OOB(Out-Of-Bag) 검증

랜덤포레스트에서는 붓스트랩 표본을 생성할 시, 붓스트랩 표본으로 뽑히지 않은 자료들을 OOB(Out-Of-Bag) 자료라 부르며 이를 검증 데이터 대신 사용하여 모형 검증을 수행할 수 있다. OOB 오차는 K-fold cross validation 오차와 거의 동일하며, 계산상 교차검증을 수행하기 힘든 규모가 큰 데이터셋에 대해 배깅을 수행할 때 편리하다[13].

4.1.2 K-fold cross validation 검증

K-fold cross validation 검증은 데이터를 동일한 크기를 가진 k개의 그룹으로 분할하고, 첫 번째 fold는 검증 데이터로 사용하고 나머지 k-1개의 fold에 대해 훈련하여 검증 데이터 셋의 오차를 계산한다. 이 과정은 순차적으로 k번 반복되며 그 결과 얻어지는 오차 값들을 평균하여 계산한다. 이 검증 방법은 지나치게 높은 편향과 높은 분산으로 인한 문제없이 검증을 진행할 수 있다는 장점을 가지고 있다[24, 26]. XGBoost는 자체적으로 cross validation 검증을 내장하고 있으며 이를 활용하여 모델을 검증하였다.

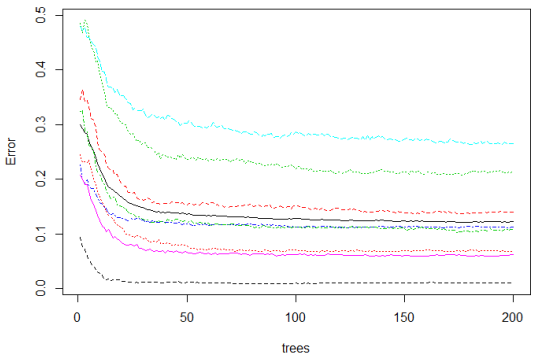
4.2 결과

민원 카테고리 분류를 검증하는데 있어, 업/다운 샘플링을 하여 데이터 균형을 맞춘 다음 RandomForest 모델에 적용시키는 과정을 50회 실시한 뒤 OOB 오차를 산술 평균하였다. 민원처리부서 또한 위의 과정과 동일하게 진행한 뒤 OOB 오차를 산술 평균한 값을 사용하였다.

XGBoost 또한 업/다운 샘플링을 하여 데이터

균형을 맞춘 뒤 민원 카테고리와 처리부서 각각에 대한 분류과정을 진행하였고, 5 fold cross validation 오차 값을 산술 평균하여 사용하였다.

<그림 3>은 트리 수에 따라 랜덤포레스트의 오차율이 어떻게 변하는지 OOB 자료를 이용하여 나타낸 그래프이다. 트리의 수가 많아질수록 오차율이 점점 낮아지며, 트리의 수가 100개일 때 오차율이 안정적으로 수렴하는 것을 확인할 수 있다. 이에 따라 랜덤포레스트 모델 수행 시 트리 생성 수를 100개로 설정한 뒤 모델을 수행하였다.



〈그림 3〉 Tree 수에 따른 오차율 변화

민원 카테고리 분류 결과는 <표 4>와 같으며, 민원 처리부서 분류 결과는 <표 5>와 같다.

〈표 4〉 민원 카테고리 분류 결과

	카테고리 분류 정확도	
	Random Forest	XGBoost
Upsampling	87.41%	93.59%
Downsampling	74.27%	93.43%

〈표 5〉 민원 담당부서 분류 결과

	담당부서 분류 정확도	
	RandomForest	XGBoost
Upsampling	89.63%	91.39%
Downsampling	69.26%	92.68%

RandomForest를 사용하여 카테고리 분류한 결과 예시는 <표 6>, <표 7>과 같으며 아래 표와 같은 과정을 50회 실시한 뒤 그 오차값을 산술평균하여 정확도를 계산한 것이 위의 <표 4>, <표 5>이다. 담당부서 분류의 경우에도 이와 과정이 동일하므로 이는 생략하였다.

〈표 6〉 카테고리 분류 결과 예시(RandomForest, upsampling)

	경제 일자리	교통	기획 감사 교육	문화 관광 체육	복지 어르신 장애인	안전 소방 민방위	주택 도시계획 부동산	환경 공원 상수도	OOB class error
경제 일자리	1704	46	35	109	36	10	30	13	0.14069592
교통	63	1558	48	91	22	45	68	88	0.21432173
기획 감사 교육	93	16	1760	61	20	7	19	7	0.11245587
문화 관광 체육	144	116	54	1456	42	18	62	92	0.26575895
복지 어르신 장애인	18	23	16	32	1860	3	15	16	0.06202723
안전 소방 민방위	2	6	3	5	1	1962	2	2	0.01059002
주택 도시계획 부동산	25	19	11	42	8	12	1848	18	0.06807867
환경 공원 상수도	28	57	12	45	22	7	43	1769	0.10791730
					251.375				

〈표 7〉 카테고리 분류 결과 예시(RandomForest, Downsampling)

	경제 일자리	교통	기획 감사 교육	문화 관광 체육	복지 어르신 장애인	안전 소방 민방위	주택 도시계획 부동산	환경 공원 상수도	OOB error
경제 일자리	789	66	41	149	62	24	63	29	0.3548651
교통	66	914	25	38	22	45	51	62	0.2526574
기획 감사 교육	113	22	917	82	34	16	30	9	0.2502044
문화 관광 체육	22	94	57	554	68	33	93	122	0.5470155
복지 어르신 장애인	48	23	24	44	1030	6	24	24	0.1578087
안전 소방 민방위	3	6	11	4	2	1184	8	5	0.0318888
주택 도시계획 부동산	35	20	24	43	4316	21	1038	26	0.1512674
환경 공원 상수도	46	74	18	79	24	31	72	879	0.2812756
					694.75				

Ⅴ. 결론 및 시사점

본 연구에서는 서울시 민원 사례 데이터를 RandomForest와 XGBoost 두 모델을 활용하여 분류 과정을 수행하였다. 분류 결과 RandomForest 대비 XGBoost의 정확도가 전반적으로 높게 나타났으며, 비정형 한국어 텍스트 분류의 적합성을 확인할 수 있었다.

동일한 표본을 활용하여 업 샘플링과 다운 샘플링 후 모델을 수행하였을 시 RandomForest는 샘플링 방법에 따라 분류 정확도가 불안정하게 나타난 반면, XGBoost는 전반적으로 안정적인 정확도를 보였다. 즉 분류하고자 하는 데이터의 클래스의 수가 많거나 클래스 간의 데이터 불균형이 높을 때 RandomForest대비 XGBoost의 성능이 특히 부각된다. 이를 통해 민원분야 외에도 클래스 간 불균형이 발생하는 데이터에 XGBoost를 적용함으로써 정확도와 안정성을 확보할 수 있음을 알 수 있다.

실무적인 측면으로는 서울특별시 응답소에 올라오는 민원 내용을 토대로 담당부서를 할당하고 분배하는 과정을 높은 정확도로 자동화할 수 있음을 확인하였고, 응답소 개설의 목적인 신속한 민원 응대 측면에서 민원 처리업무 프로세스 효율성 향상과 처리 시간을 단축에 기여할 수 있을 것이다.

한편 현재 서울특별시 응답소에 공개된 데이터는 과거 데이터가 섞여 있어 행정체계 변경에 따른 부서 및 민원 카테고리 변경이 업데이트 되지 않은 상태이다. 현재 사용되고 있는 카테고리 및 부서로 통합작업을 진행하여 데이터 품질을 높일 수 있다면 보다 높은 분류 정확도를 확보할 수 있을 것으로 보인다.

참 고 문 헌

[1] 강승식, “형태소 분석 결과의 인코딩 기법과 어절 사전 구축”, 한국정보과학회 언어공학연구회 학술발표논문집, 제16권, 제1호, pp.112-117, 2004.

[2] 김성표, 박훈진, 나영우, 최병길, “공간 빅데이터

를 활용한 환경민원 분석에 관한 연구", 한국측량학회, 제15권, 제2호, pp.333-334, 2015.

[3] 김수아, 조희선, 이현아, "다양한 어휘 가중치를 이용한 블로그 포스트의 자동 분류", *Journal of the Korean Society of Marine Engineering*, 제39권, 제1호, pp.58-62, 2015.

[4] 손남례, 김서영, "공공민원 빅데이터 분석을 통한 민원통계 및 담당부서 자동분류 시스템, 한국차세대컴퓨팅학회, 제13권, 제1호, pp.22-35, 2017.

[5] 송효진, 황성수, "정부 3.0 추진에 따른 공공데이터 개방과 지방정부의 방향성 모색: 공공데이터법에 관한 이해와 개방 사례를 중심으로", 한국지역정보학회지, 제17권, 제2호, pp.1-28, 2014.

[6] 원태홍, 유환희, "진주시 전자민원 추이분석", 한국지형공간정보학회 춘계학술대회논문집, pp.106-109, 2016.

[7] 유진은, "랜덤 포레스트: 의사결정나무의 대안으로서의 데이터 마이닝 기법", *Journal of Educational Evaluation*, 제28권, 제2호, pp.427-448, 2015.

[8] 조응래, 지우석, 홍명기, 최서윤, "경기도 교통불편 민원의 효율적인 처리방안", 정책연구, pp. 1-76, 2016.

[9] 최해옥, "환경·위생분야 민원분석을 통한 정책우선순위 결정에 관한 연구", 환경정책, 제24권, 제2호. pp.45-57, 2016.

[10] Amor, N.B., S. Benferhat, and Z. Elouedi, "Naive Bayes vs Decision Trees in Intrusion Detection Systems", *ACM Symposium on Applied Computing*, pp.420-424, 2004.

[11] Breiman, L. and A. Cutler, "Random Forests", https://www.stat.berkeley.edu/~breiman/RandomForests, 2014.

[12] Breiman, L., "Random Forests", *Machine Learning*, Vol.45, pp.5-32, 2001.

[13] Bylander, T., "Estimation Generalization Error on Two-Class Datasets Using Out-of-Bag Estimates", Machine Learning, Vol.48, pp.287-297, 2002.

[14] Chen, T. and C. Guestrin, "XGBoost: A Scalable Tree Boosting System", *KDD'16*, pp.785-794, 2016.

[15] Chen, T. and C. Guestrin, "XGBoost: Reliable Large-sclae Tree Boosting System", 2017.

[16] Chen, W., X. Li, Y. Wang, G. Chen, and S. Liu, "Forested Landslide Detection using LiDAR data and the Random Forest Algorithm: A Case Study of the Three Gorges, China", *Remote Sensing of Environment*, Vol.152, pp.291-301, 2014.

[17] Fragos, K., P. Belsis, and C. Skourlas, "Combining Probabilistic Classifiers for Text Classification", Procedia-Social and Behavioral Sciences, Vol.147, pp.307-312, *3rd International Conference on Integrated Information (IC-ININFO)*, doi:10.1016/j.sbspro.2014.07.098, 2014.

[18] Friedman, J.H., "Greedy Function Approximation: A Gradient Boosting Machine", *The Annals of Statistics*, Vol.29, pp.1189-1232, 2001.

[19] Hebert, J., "Predicting Rare Failure Event using Classification Trees on Large Scale Manufacturing Data with Complex Interactions", *IEEE International Conference on Big Data*, pp.2024-2028, 2016.

[20] Holzer, M. and A. P. Manoharan et al., "서울시 전자정부 우수사례 연구보고서", Rutgers University School of Public Affairs and Administration, 2016.

[21] https://xgboost.readthedocs.io "XGBoost 공식 홈페이지".

[22] Kibinge, N., S, Ikeda, N. Ono, M. Altaf-Ul-Amin, and S. Kanaya, "Integration of Residue Attributes for Sequence Diversity Characterization of Terpe-

noid Enzymes", *BioMed Research International*, Vol.2014, pp.1-10, 2014.

[23] K-ICT 빅데이터센터, NIADic tutorial, 2017.

[24] Kohavi, R., "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection", *Appears in the Internationa l Joint Conference on Artificial Intelligence*, 1995.

[25] Pal, M., "Random Forest Classifier for Remote Sensing Classification", *International Journal of Remote Sensing*, Vol.26, No.1, p.217, 2005.

[26] Refaeilzadeh, P., L. Tang, et al., "Cross-Validation", Encyclopedia of Database Systems, pp.532-538, 2009.

[27] Rodriguez-Galiano, V.F., M. Chica-Olmo, F. Abarca-Hernandez, P.M., Atkinson, and C. Jeganathan, "Random Forest Classification of Mediterranean Land Cover using Multi-seasonal Imagery and Multi-seasonal Texture", *Remote Sensing of Environment*, Vol.121, pp.93-107, 2012.

[28] Siroky, D.S., "Navigating Random Forests and related advances in algorithmic modeling", *Statistics Surveys*, Vol.3, pp.147-163, 2009.

[29] Tang, B., H. He, P.M. Baggenstoss, and S. Kay, "A Bayesian Classification Approach Using Class-Specific Features for Text Categorization", *IEEE Transactions on Knowledge and Data Engineering*, Vol.28, No.6, pp.1602-1606, 2016.

[30] Zhang, X., J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification", in *Advances in neural information processing systems*, pp.649-657, 2015.

저 자 소 개

하 지 은(Ji-Eun Ha)

- 2015년 : 동아대학교 경영정보학 (학사)
- 2017년 : 연세대학교 정보대학원 디지털 경영 (석사)
- 관심분야 : Big Data Analytics, Open collaboration, Digital Marketing, Digital Transformation

신 현 철(Hyun-Chul Shin)

- 2016년 : 경희대학교 경영학, 철학 (학사)
- 2016년~현재 : 연세대학교 정보대학원 비즈니스 빅데이터 분석 (석사과정)
- 관심분야 : Big Data Analytics, Data Mining, Deep Learning, Text Mining

이 준 기(Zoon-Ky Lee)

- 1985년 : 서울대학교 계산통계학 (학사)
- 1991년 : 카네기멜론대학 사회심리학 (석사)
- 1999년 : 남가주대학교 경영정보학 (박사)
- 2004년~현재 : 연세대학교 정보대학원 교수
- 관심분야 : Big Data Analytics, Digital Transformation, Open Collaboration