Journal of The Institute of Electronics and Information Engineers Vol.58, NO.4, April 2021 https://doi.org/10.5573/ieie.2021.58.4.43
ISSN 2287-5026(Print) / ISSN 2288-159X(Online)

논문 2021-58-4-6

클래스 불균형 문제가 있는 특허분류 데이터의 자동분류 성능 개선을 위한 모델 재귀적 오버샘플링 방법

(Recursive Oversampling Method for Improving Classification Performance of Class Unbalanced Data in Patent Document Automatic Classification)

김 성 훈*, 김 승 천**

(Sunghoon Kim and Seungcheon Kim©)

요 약

클래스 불균형이란 정의된 클래스 간 샘플 개수의 차이가 매우 커서 샘플의 대부분을 샘플의 수가 적은 소수 클래스(minority class) 보다 샘플의 수가 많은 다수 클래스(majority class)로 예측하게 되는 경우를 말한다. 본 논문에서는 클래스 불균형 상태의 특허데이터로 부터 생성된 분류기의 문제점을 해결하기 위하여 모델 재귀적 오버샘플링이라는 기법을 제안한다. 클래스 불균형인 데이터로 훈련시킨 분류기 생성 후, 특허문서를 기반으로 랜덤 데이터를 생성하고 생성된 랜덤데이터를 앞서 정의된 분류기로 분류한다. 분류기에 의해 예측된 랜덤 데이터 중 소수 클래스로 예측된 데이터를 샘플링 하는 방식이 모델 재귀적 오버샘플링 기법이다. 모델 재귀적 오버샘플링을 통해 만들어진 분류기를 최초 분류기와 비교했을 때 소수 클래스의 precision, recall, f-score가 증가하였다. 특히 SMOTE 오버샘플링 기법을 이용한 분류기와 비교했을 때도 소수 클래스의 정확도가 증가함을 확인하였다.

Abstract

Class imbalance refers to a case in which the difference in the number of samples between the defined classes is very large, so that most of the samples are predicted as a majority class with a larger number of samples than a minority class with a small number of samples. In this study, we propose a technique called recursive oversampling to solve the problem of classifiers generated from patent data in class imbalance. After generating a classifier trained with class imbalanced data, random data is generated based on the patent document, and the generated random data is classified with the previously defined classifier. The recursive oversampling is a method of sampling data predicted by a minority class among random data predicted by a classifier. When comparing the classifier made through recursive oversampling with the original classifier, the precision, recall, and f-score of the minority class were increased. In particular, it was confirmed that the accuracy of the minority class increased even when compared to the classifier using the SMOTE oversampling technique.

Keywords : 특허분류, 클래스 불균형, 재귀적 오버샘플링, MLP

* 학생회원, 한성대학교 스마트융합컨설팅학과
(Department of Smart Convergence Consulting Hansung University)
** 평생회원, 한성대학교 스마트융합컨설팅학과
(Department of Smart Convergence Consulting Hansung University)
© Corresponding Author(E-mail: kimsc@hansung.ac.kr)
※ 본 연구는 한성대학교 교내학술연구비 지원과제임.
Received ; January 5, 2021 Revised ; March 7, 2021
Accepted ; March 13, 2021

Ⅰ. 서 론

특허를 심사하는 과정은 여러 단계로 나뉘는데, 그 첫 번째 과정은 접수된 특허들을 기술의 영역에 맞춰서 분류하는 과정이 된다. 이런 분류는 발명가에 의해서 스스로 분류되거나 혹은 접수된 키워드에 의해서 기술 영역이 분류되도록 하는 것이 일반적이나 이는 자칫 시간이 많이 소요되거나 혹은 잘못된 분류로 인해서 심사

가 늦어지는 요인이 되기도 한다.

특허문서 분류는 특허문서를 효율적으로 심사, 검색, 분석, 구조화하기 위해 특허문서를 미리 정의된 클래스로 할당하는 작업이다. 특허문서는 IPC(International Patent Classification), CPC(Cooperative Patent Classification) 등 공식적인 클래스와 사용자의 필요에 따른 사용자 정의 클래스로 나뉠 수 있다.

이러한 사용자 정의 클래스와 클래스에 할당된 특허문서를 바탕으로 경쟁사 특허분석, 산업동향 분석 등을 수행한다. 대부분의 분류 작업은 특허분류 전문가에 의해 수작업으로 진행되고 있으므로 많은 시간과 비용이 소요된다. 이러한 비용과 시간을 절약하기 위해, 전문가에 의해 분류된 특허 데이터를 학습 데이터로 사용하여 생성된 특허문서 자동분류기는 좋은 해결책이 될 수 있다. 그러나 전문가에 의해 분류된 데이터가 클래스별로 할당된 특허의 개수가 차이가 많을 경우가 생길 수 있는데 이러한 클래스 불균형 데이터를 학습 데이터로 사용한다면 분류기의 분류 성능 저하가 발생한다. 실제 산업계에서 생산되는 특허문서 분류 작업의 결과물은 클래스 불균형 데이터일 가능성이 크다.

클래스 불균형이란 각 클래스에 할당된 학습 데이터 개수의 차이로 인해 클래스에 할당되지 않은 데이터의 클래스를 예측할 때 학습 데이터의 개수가 많은 클래스로 주로 예측하게 되는 경우를 말한다. 텍스트 데이터에서 클래스 불균형 문제가 많이 발생하며, 이러한 데이터가 학습데이터로 사용되는 알고리즘은 분류 성능이 떨어진다고 보고되었다[1~2].

이러한 문제를 극복하기 위해 소수 클래스의 데이터를 추가하기 위한 오버샘플링 방법이 있다. 대표적인 오버샘플링 방법은 SMOTE이다. 이는 소수 클래스의 보간을 통해서 합성 데이터를 생성해서 소수 클래스의 데이터를 증가시키는 방법이다[3]. SMOTE는 소수 클래스의 샘플을 취하여 소수 클래스의 가장 가까운 이웃을 연결하는 선 세그먼트를 따라 합성 데이터를 생성하여 오버샘플링 하는 방법이다[3].

SMOTE를 기반으로 한 Borderline-SMOTE는 경계선 근처의 소수 클래스의 샘플만 이용하여 합성 데이터를 생성하는 오버샘플링 하는 방법이다. 이러한 접근 방식은 무작위 오버샘플링 및 SMOTE 보다 높은 f-score를 달성하였다[4].

SMOTE 알고리즘은 소수 클래스의 데이터를 합성하는 기술이다. 다수 클래스의 데이터를 샘플링하여 기존 소수 데이터의 샘플을 보간하여 새로운 소수 클래스의 데이터를 합성한다. 그러나 이렇게 합성된 데이터는 키워드로 이루어진 문서로 변환이 불가하다.

키워드로 구성되어진 문서로 변환될 수 없는 합성데이터를 모델의 훈련 데이터로 사용된다면 정확도 및 향후 확장성에 문제가 있을 것으로 판단된다.

본 논문에서는 훈련데이터로 사용된 특허문서에 포함된 키워드를 랜덤하게 추출하여 대량의 문서를 만들고 이를 불균형한 데이터로 학습된 분류기로 예측하여 소수 클래스로 분류되는 문서만을 샘플링 하는 재귀적 오버샘플링 기법을 제안한다. 특히 SMOTE와 재귀적 오버샘플링을 이용한 분류기의 성능을 비교하여 모델 재귀적 오버샘플링의 유효성을 검증하고자 한다.

Ⅱ. 본 론

1. 모델 재귀적 오버샘플링

본 연구에서 제안하는 모델 재귀적 오버샘플링 기법은 클래스 불균형 데이터로 학습되어진 불균형 분류모델을 사용한다.

불균형 분류모델의 입력 형식에 맞게 랜덤 데이터를 생성하고 분류모델에 의해 소수의 클래스로 예측될 경우 랜덤하게 생성된 데이터를 저장한다. 이 과정을 클래스 불균형 데이터가 클래스 균형데이터가 될 때 까지 반복한다. 이 과정을 통해 만들어진 클래스 균형 데이터를 학습 데이터로 사용하여 알고리즘을 재학습하여 모델을 생성한다. 이 과정이 1차 모델 재귀적 오버샘플링 방법이다.

2차 모델 재귀적 오버 샘플링 방법은 1차에서 생성된 분류모델에 다시 랜덤 데이터를 생성하고 2차에서 생성된 모델에 의해 소수의 클래스로 예측될 경우 생성된 데이터를 저장한다. 이 과정을 1차때와 마찬가지로 클래스 불균형 데이터가 클래스 균형데이타가 될 때 까지 반복한다. 즉 최초의 불균형 데이터와 분류모델에 의해 생성된 소수클래스로 예측된 데이터를 함께 데이터 셋으로 만들고 이를 학습데이터로 사용하여 알고리즘을 다시 학습시켜 모델을 생성한다. 이 과정이 2차 모델 재귀적 오버 샘플링 방법이다.

이러한 과정을 계속 반복하여 모델의 성능을 높이고자 하는 것이 모델 재귀적 오버 샘플링 기법이다.

2. 랜덤 데이터 생성

본 실험을 위해 사용하는 알고리즘은 Multi Layer Perceptron(MLP) 이다. MLP의 입력값은 N차원 벡터

표 1. 재귀적 오버 샘플링 기법의 pseudo code
Table1. Pseudo code of recursive oversampling method.

	pseudo code
1	mlp_model := TRAIN(imblanced_dataset)
2	REPEAT n times
3	train_set := imblanced_dataset
4	DO
5	make random_data
6	if PREDICT(mlp_model, random_data) is minor class:
7	train_set := train_set U random_data
8	WHILE(train_set is balanced)
9	mlp_model := TRAIN(train_set)
10	

이다. N차원의 N 값은 bag of words 전처리를 통해 만들어지는 전체 words의 개수이다. 본 실험에서 전처리를 거친 총 word의 수는 15,985개이다. 그러므로 랜덤 데이터는 15,985 차원의 벡터로 생성된다. 본 실험에 사용된 특허 문서 하나당 100 ~ 150개 정도의 word를 가지는 것으로 파악되었기 때문에 랜덤데이터도 100 ~ 150개의 차원에 값을 가지며 그 값들은 정규분포를 가지도록 구성하였다. 즉 15,985 차원의 벡터를 생성하며 임의의 100~150개의 차원에 값을 가지며 그 값들은 정규분포를 따르며 나머지 차원은 0으로 채워지는 sparse 벡터로 생성하였다. 랜덤 데이터의 형식은 실제 특허문서가 벡터로 변환될 때의 형식과 같다.

이렇게 생성된 랜덤데이터를 모델 재귀적 오버샘플링 기법의 입력 데이터로 사용하고 분류 모델의 예측 결과에 따라 폐기되거나 학습데이터에 추가된다.

Ⅲ. 실 험

1. 데이터

가. 학습데이터 및 테스트 데이터 구성

본 실험을 위해 사용된 데이터는 미국 등록특허 1,200건이다. 총 4개의 CPC 분류(H01, H02, H03, H04)에 포함된 특허로서 각 분류당 300건의 특허를 최신 출원일 순으로 선택하였다.

실험의 정확도 측정을 위한 테스트 데이터는 분류당 각 100건씩 총 400건을 사용하였다.

분류 균형 조건에서 사용한 학습데이터는 분류당 각 200건씩 총 800건을 사용하였다. 분류 불균형 조건에서 사용한 학습데이터는 H03 분류에 포함된 특허를 50건, 그 외 분류에 대해서는 200건씩 사용하여 총 650건의 학습데이터를 사용하였다. 본 실험에서 분류 불균형 조건은 H03 분류가 다른 분류(H01, H02, H04)에 비해 포함된 특허건수가 적을 때를 가정하였다.

표 2. 실험에 사용된 학습데이터와 테스트 데이터
Table 2. Training and test data used in the experiment.

	H01	H02	H03	H04	Sum.
Balanced training set	200	200	200	200	800
Imbalanced training set	200	200	50	200	650
Test set	100	100	100	100	400
Total set	300	300	300	300	1,200

나. 데이터 전처리

본 실험에서 사용하는 특허는 영어로 기술되어 있다. 특허문서에서 분류 판단에 사용될 구성요소는 특허의 제목, 요약 그리고 대표청구항이다. 이들 구성요소를 합친 전체 텍스트가 사용되었으며 이 텍스트에 대해 소문자 변환, 특수문자 와 숫자의 제거, 불용어 제거 전처리를 수행하였다. 이러한 전처리는 단어의 출현 빈도를 특징으로 사용하는 본 실험 모델인 MLP에 적합하기 때문이다.

2. MLP(Multi Layer Perceptron)모델

가. 모델 및 실험조건

본 실험에서는 총 5가지의 경우에 대해서 실험하였으며 5번의 경우에 대해서 동일한 구조와 동일한 하이퍼파라메타를 사용하였다.

실험의 기본적인 모델은 MLP 이다. 64개의 노드를 가지는 2개의 은닉층을 가지며 출력층은 4개의 노드를 가지는 구조로 구성되었다. dropout rate는 0.2, 배치사이즈는 32, learning rate는 1e-4 그리고 옵티마이저는 adam을 사용하였다. 최대 150 에폭까지 학습하며 특정 조건을 만족하면 학습을 중지시키는 early stopping을 적용하였다. early stopping의 조건은 테스트 데이터를 validation 데이터로 사용하여 validation loss가 30 에폭이 지나도 감소하지 않는 것이다. validation loss가 최소값일 때의 모델을 선택하였다. 선택된 모델의 성능은 동일한 테스트 데이터를 이용하여 측정한다.

3. 실험 방법

가. 5가지 실험 조건

총 5가지의 경우에 대해서 실험을 실시한다. 2가지 경우는 오버샘플링 없이 분류 균형 학습데이터와 분류 불

균형 학습데이터에 대해 MLP로 훈련시키고 테스트 데이터를 통해 정확도와 f1-score를 측정한다.

3가지의 경우에 대해서는 분류 불균형 학습데이터 상태에서 3가지 오버샘플링 방법을 적용하여 H03 분류를 150건 증가시킨다. 오버샘플링을 통해 균형 학습데이터를 만든 후 동일한 방법으로 MLP를 훈련시키고 테스트 데이터를 통해 정확도와 f1-score를 측정한다.

표 3. 5가지 실험조건의 데이터 구성
Table3. Data composition of 5 experimental conditions.

Condition	H01/H02/H04	H03	Sum.
Class Balanced	600(each 200)	200	800
Class Imbalanced	600(each 200)	50	650
SMOTE OverSampling	600(each 200)	50 150(o.s.)	800
Borderline SMOTE OverSampling	600(each 200)	50 150(o.s.)	800
Model Recursive OverSampling	600(each 200)	50 150(o.s)	800

나. 조건별 오버샘플링

본 실험에서 분류 불균형은 H03 분류에서 발생한다. H01, H02, H04 분류에는 각 200건씩의 특허를 포함하고 있지만 H03 분류에는 50건의 특허를 포함하고 있다. H03 분류의 불균형을 해결하기 위해 150건의 특허를 오버샘플링 하여 H03분류의 특허 건수를 다른 분류와 마찬가지로 200건의 데이터를 포함하게 하였다.

실험에서 사용한 오버샘플링 기법은 SMOTE, Borderline SMOTE 그리고 본 연구에서 제안하는 모델 재귀적 오버샘플링 이다.

4. 실험 결과

가. 분류 균형에서 테스트 데이터 추론 결과

그림 1은 분류 균형 상태에서 테스트 데이터의 추론 결과이다. 4가지 분류의 f1-score은 최소 0.786(H02), 최대 0.899(H04)이며 평균은 0.844 이다. 최대값과 최소값의 차는 0.113 이다.

recall은 최소 0.77(H02), 최대 0.93(H04)이며 평균은 0.845 이다. 최대값과 최소값의 차는 0.16 이다.

나. 분류 불균형에서 테스트 데이터 추론 결과

그림 2는 분류 불균형 상태에서 테스트 데이터 추론 결과이다. 4가지 분류의 f1-score은 최소 0.615(H03), 최대 0.879(H04)이며 평균은 0.763 이다. 최대값과 최소값의 차는 0.264 이다.

recall은 최소 0.48(H03), 최대 0.879(H04)이며 평균은 0.763 이다. 최대값과 최소값의 차는 0.399 이다.

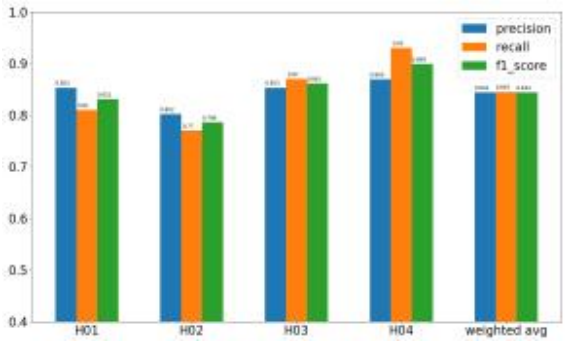

그림 1. 분류 균형 조건에서 테스트 데이터 추론 결과
Fig. 1. Test data inference results under classification balanced condition.

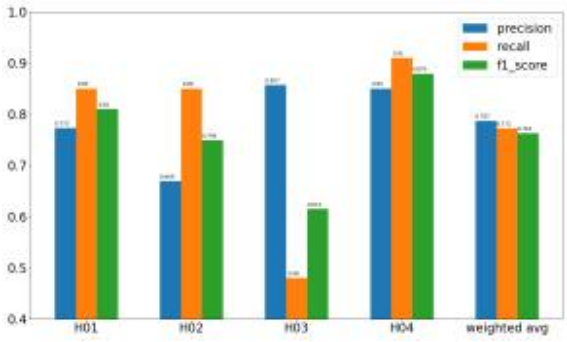

그림 2. 분류 불균형 조건에서 테스트 데이터 추론 결과
Fig. 2. Test data inference results under classification imbalanced condition.

다. SMOTE 오버샘플링에서 테스트 데이터 추론 결과

그림 3은 SMOTE 오버샘플링 상태에서 테스트 데이터 추론 결과이다. 4가지 분류의 f1-score은 최소 0.679(H03), 최대 0.894(H04)이며 평균은 0.792 이다. 최

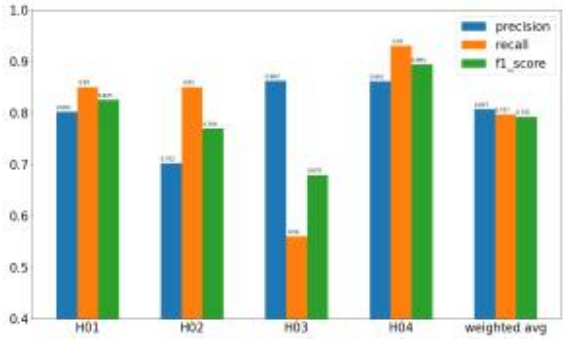

그림 3. SMOTE 오버샘플링의 테스트 데이터 추론 결과
Fig. 3. Test data inference results of SMOTE oversampling.

대값과 최소값의 차는 0.215 이다.

recall은 최소 0.56(H03), 최대 0.93(H04)이며 평균은 0.797 이다. 최대값과 최소값의 차는 0.37 이다.

라. Borderline SMOTE 오버샘플링에서 테스트 데이터 추론 결과

그림 4는 Borderline SMOTE 상태에서 테스트 데이터 추론 결과이다. 4가지 분류의 f1-score은 최소 0.617(H03), 최대 0.852(H04)이며 평균은 0.749 이다. 최대값과 최소값의 차는 0.235 이다.

recall은 최소 0.5(H03), 최대 0.92(H04)이며 평균은 0.757 이다. 최대값과 최소값의 차는 0.163 이다.

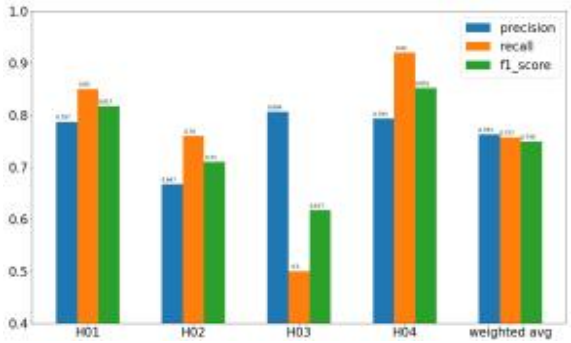

그림 4. Borderline SMOTE 오버샘플링에서 테스트 데이터 추론 결과

Fig. 4. Test data inference results of Borderline SMOTE oversampling.

마. 모델 재귀적 오버샘플링에서 테스트 데이터 추론 결과

모델 재귀적 방법은 재귀 횟수에 따라 테스트 데이터 추론 결과는 다음과 같다. 1회 적용한 재귀적 방법에서 f1-score은 최소 0.667(H03), 최대 0.896(H04)이며 평균은 0.782이다. 최대값과 최소값의 차는 0.229이다.

recall은 최소 0.54(H03), 0.9(H04) 이며 평균은 0.787 이다. 최대값과 최소값의 차는 0.36 이다.

2회 적용했을 경우 f1-score은 최소 0.701(H03), 최대 0.883(H04)이며 평균은 0.793이다. 최대값과 최소값의 차는 0.09이다.

recall은 최소 0.62(H03), 최대 0.87(H04) 이며 평균은 0.795이다. 최대값과 최소값의 차는 0.25 이다.

3회 적용했을 경우 f1-score은 최소 0.652(H03), 최대 0.865(H04)이며 평균은 0.771이다. 최대값과 최소값의 차는 0.213이다.

recall은 최소 0.58(H03), 0.85(H02) 이며 평균은 0.772 이다. 최대값과 최소값의 차는 0.27 이다.

1회 적용시에 비해 2회 적용시에 성능이 증가함을 보였으나 3회 적용시에는 다시 성능이 감소함을 보였다.

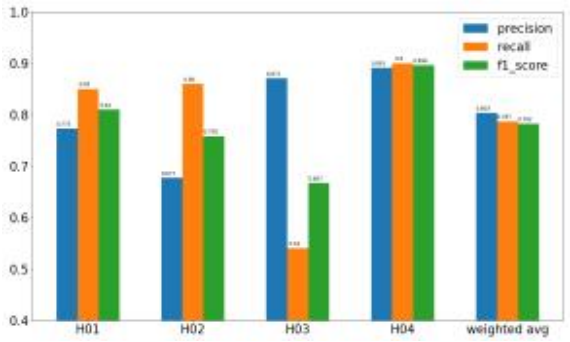

그림 5. 1회 모델 재귀적 오버샘플링에서 테스트 데이터 추론 결과

Fig. 5. Test data inference results from one-time model recursive oversampling.

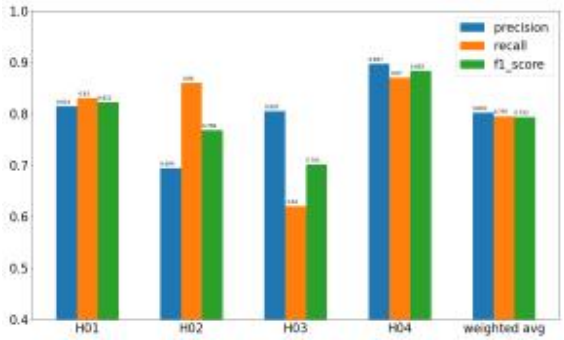

그림 6. 2회 모델 재귀적 오버샘플링에서 테스트 데이터 추론 결과

Fig. 6. Test data inference results from two-time model recursive oversampling.

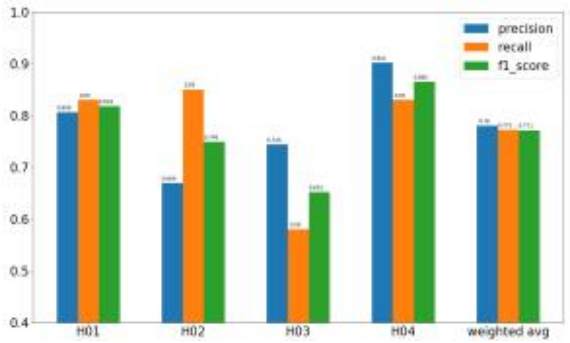

그림 7. 3회 모델 재귀적 오버샘플링에서 테스트 데이터 추론 결과

Fig. 7. Test data inference results from three-time model recursive oversampling.

바. 불균형 분류인 H03 관점에서의 비교

불균형 분류를 초래한 H03 분류에 대해서 7가지 경

우에 대해서 테스트 데이터의 추론 결과는 아래 그림과 같다. 균형 분류에 비해서 불균형 분류의 recall과 f1-score은 현저하게 줄어들었다.

H03에 대해 오버샘플링을 기존의 방법(SMOTE, Borderline SMOTE)과 본 연구에서 제안한 방법(모델 재귀적 오버샘플링)으로 수행하였다. 모델 재귀적 오버샘플링은 1회, 2회, 3회를 수행한 결과를 비교하였다.

불균형 조건에서의 추론결과에 비해 오버샘플링 기법을 적용하면 추론결과가 개선되었다. 오버샘플링 기법중 2회 적용된 모델 재귀적 오버 샘플링 기법이 가장 높은 성능을 보였으며 SMOTE, Borderline SMOTE, 1회 적용 모델 재귀적 기법, 3회 적용 모델 재귀적 기법 순으로 성능 개선을 보였다.

모델 재귀적 기법에서는 1회에 비해 2회 적용했을 때 높은 성능을 보였으나 3회 적용했을 때는 다시 성능이 하락함을 보였다.

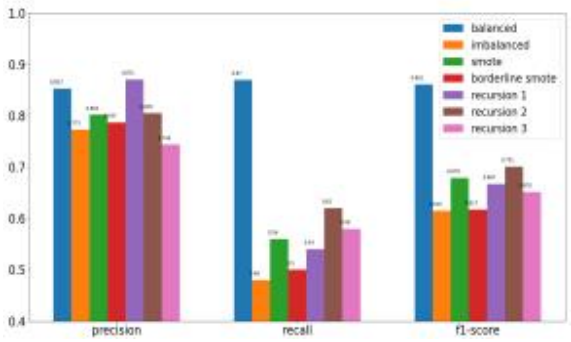

그림 8. 불균형 분류인 H03의 조건별 추론 결과 비교
Fig. 8. Comparison of inference results by condition of H03, which is an imbalanced class.

Ⅳ. 결 론

MLP를 이용한 특허문서 분류에서 분류 균형일 경우와 분류 불균형일 경우에 데이터가 적은 분류에서는 다른 분류에 비해 f1-score와 recall이 상대적으로 낮은 것을 확인 할 수 있었다. 이러한 불균형을 해소하기 위해 다양한 오버샘플링 기법이 존재하는데 이를 특허 문서 분류에 적용하였다. 본 실험에서는 SMOTE와 Borderline SMOTE를 특허 문서 분류에서 분류 불균형에 대해서 적용하였고 이를 통해 성능이 개선됨을 확인하였다. 또한 모델 재귀적 오버샘플링 기법을 제안하여 동일한 조건에서 실험하여 결과를 비교하였다. 모델 재귀적 기법에서 횟수를 2회 적용했을 때 기존의 오버샘플링에 비해 높은 성능을 보임을 확인하였다. 그러나 모델 재귀적 오버샘플링 기법에서는 재귀 횟수에 따라 성능이 상승 후 다시 하강하는 것을 확인 할 수 있었다.

특허 자동 분류에 있어 학습데이터가 분류 불균형 상태일 경우 오버 샘플링은 성능 개선에 좋은 해결책이 될 수 있다. 또한 본 연구에서 제안한 모델 재귀적 오버 샘플링 기법은 재귀 횟수에 대한 조절을 통해서 기존의 오버샘플링 기법의 대안이 될 수 있으리라 기대된다.

REFERENCES

[1] H. He and E. A. Garcia, Learning from Imbalanced Data, IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 9, pp. 1263-1284, Sept. 2009, doi: 10.1109/TKDE.2008.239.

[2] Liu, Y. et al., Imbalanced text classification: A term weighting approach, Expert Systems with Applications, vol. 36, no. 1, 690-701, Jan. 2009, doi:10.1016/j.eswa.2007.10.042

[3] Chawla, N. V. et al., SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research 16, 321-357, 2006, doi:10.1613/jair.953

[4] Hui Han. et al., Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning, Advances in Intelligent Computing. ICIC 2005. Lecture Notes in Computer Science, vol 3644. Springer, Berlin, Heidelberg. doi:10.1007/11538059_91

저 자 소 개

김 성 훈(학생회원)
1997년 경희대학교 화학공학과 학사 졸업(공학사)
2000년 경희대학교 화학공학과 석사 졸업(공학석사)
2018년～현재 한성대학교 스마트 융합컨설팅학과 박사 과정
2016년～현재 (주)웹스 서비스개발부 부서장
<주관심분야 : 인공지능, 딥러닝, NLP, 특허분석>

김 승 천(평생회원)
1994년 2월 연세대학교 전자공학과 학사 졸업(공학사).
1996년 2월 연세대학교 전자공학과 석사 졸업(공학석사)
1999년 8월 연세대학교 전기컴퓨터공학과(공학박사)
2000년 1월～2001년 1월 Univ. of Sydney Research Fellow
2001년 2월～2003년 8월 LG전자 DTV/DA 연구소 선임 연구원
2009년 7월～2010년 7월 Univ of Oregon 방문교수
2003년 3월～현재 한성대학교 IT 융합공학부 교수
<주관심분야 : 네트워크 보안, 블록체인 서비스, 사물인터넷 보안, 5G 이동통신망 서비스>