The Korean Journal of Applied Statistics (2014) DOI: http://dx.doi.org/10.5351/KJAS.2014.27.3.357
27(3), 357–371

A Comparison of Ensemble Methods Combining Resampling Techniques for Class Imbalanced Data

Hee-Jae Lee[a] · Sungim Lee[a,1]

[a]Department of Applied Statistics, Dankook University

(Received August 26, 2013; Revised October 7, 2013; Accepted December 16, 2013)

Abstract

There are many studies related to imbalanced data in which the class distribution is highly skewed. To address the problem of imbalanced data, previous studies deal with resampling techniques which correct the skewness of the class distribution in each sampled subset by using under-sampling, over-sampling or hybrid-sampling such as SMOTE. Ensemble methods have also alleviated the problem of class imbalanced data. In this paper, we compare around a dozen algorithms that combine the ensemble methods and resampling techniques based on simulated data sets generated by the Backbone model, which can handle the imbalance rate. The results on various real imbalanced data sets are also presented to compare the effectiveness of algorithms. As a result, we highly recommend the resampling technique combining ensemble methods for imbalanced data in which the proportion of the minority class is less than 10%. We also find that each ensemble method has a well-matched sampling technique. The algorithms which combine bagging or random forest ensembles with random undersampling tend to perform well; however, the boosting ensemble appears to perform better with over-sampling. All ensemble methods combined with SMOTE outperform in most situations.

Keywords: Imbalance data, bagging, boosting, random forest, under-sampling, over-sampling, SMOTE.

1. 서론

데이터마이닝의 기본적인 분석주제 중 하나는 분류(classification)이다. 분류에서 어려운 문제 중 하나는 목표변수의 클래스들 중 한 클래스의 크기가 다른 경우보다 압도적으로 크거나 혹은 매우 작은 경우에 발생하는 불균형 문제이다. 예를 들어, 신용카드 회사에서 범주형 목표변수를 카드사용의 사기여부라고 할 때, 일정기간 동안 신용카드 전체 사용건수들 중 비정상적으로 사용된 경우는 매우 희박하고, 정상적으로 사용된 경우가 대부분이다. 또 제조업에서 목표변수를 제품의 불량여부라고 한다면, 회사에서 생산된 제품들은 대부분 양품으로 불량품은 그 비율이 매우 작게 된다. 희귀질병에 대한 진단을 위해 분류모형을 적합할 경우에도 목표변수의 불균형 문제는 발생하게 된다. 일반적으로 분류모형은 단순히 다수의 집단으로 예측하는 단순규칙(naive rule)보다 더 나은 예측결과를 제공해야 하는데, 이들 예제와 같은 불균형 데이터에서는 단순규칙도 정확도가 매우 높게 나타나 단순규칙보다 성능이 좋은

This research was supported by the research fund of Dankook University.
[1]Corresponding author: Department of Applied Statistics, Dankook University, 152, Jukjeon-ro, Suji-gu, Yongin-si, Gyeonggi-do 448-701, Korea. E-mail: silee@dankook.ac.kr

분류모형을 찾는 것이 쉬운 문제가 아니다. 이처럼 많은 응용분야에서 접할 수 있는 데이터의 불균형 문제는 최근에 학자들의 많은 관심을 받아왔고 이에 관한 많은 연구가 있어왔다 (Yangrhk Wu, 2006; Zhu와 Song, 2010; Khreich 등, 2010; Mazurowski 등, 2008; Kubat 등, 1998; Hur와 Kim, 2007; Kim과 Park, 2012). Galar 등 (2012)은 목표변수의 불균형을 어떻게 다루었는지에 따라 기존의 연구결과들을 크게 세 가지로 분류하였다. 첫째는 기존의 알고리즘을 수정하거나 새롭게 제안하는 방식이고 (Zadrozny와 Elkan, 2001; Wu와 Chang, 2005), 둘째로는 목표변수의 불균형을 균형을 맞출 수 있도록 데이터의 전처리과정을 고려하는 방법 (Batista 등, 2004; Chawla 등, 2002, 2004; Oh와 Zhang, 2001)을 소개하였다. 마지막으로 목표변수의 오분류비용을 계산하여 알고리즘과 데이터의 전처리과정을 함께 고려하는 방법 (Chawla 등, 2008; Kim과 Jeong, 2004; Freitas 등, 2007)이 있다. 이러한 방법들 외에도 분류기법들의 앙상블을 사용하는 방법들이 사용되어 왔는데, 앙상블이란 주어진 자료로부터 여러 개의 분류모형을 만든 후 이러한 분류모형들을 결합하여 최종 분류모형을 만드는 것으로 분류모형의 정확성을 높이도록 고안된 방법이다. 그런데 불균형 데이터에 적용된 경우는 목표변수의 균형을 맞추도록 데이터의 전처리 과정과 함께 앙상블 기법을 사용한 경우에서 분류모형이 효과적으로 적합된다는 사실이 알려져 있다 (Chawla 등, 2003; Seiffert 등, 2010; Liu 등, 2009; Kang과 Cho, 2006). 일반적으로 알고리즘과 오분류 비용관점에서 불균형 데이터를 다룬 경우에는 분류모형이 데이터의 특수성이나 문제의 특성에 의존하는 반면에, 목표변수의 균형을 맞추기 위해 데이터를 전처리하거나 앙상블 기법을 적용하는 방식은 분류모형과는 독립적으로 사용될 수 있기 때문에 이들 접근법이 불균형 문제에 훨씬 더 다양하게 적용될 수 있다. 이에 본 논문에서는 불균형데이터를 맞추기 위해 데이터 전처리과정을 크게 과소표집(under sampling), 과대표집(over sampling), 그리고 과소표집과 과대표집을 모두 고려한 하이브리드 표집(hybrid sampling)으로 나누고, 각각에 러프리 밸런스드 배깅(roughly balanced bagging) (Hido 등, 2009), 배깅(bagging), 부스팅(boosting) 그리고 랜덤 포리스트(random forest) 등의 대표적인 앙상블 기법을 모두 적용하여 그 성능을 비교검토하고자 한다. Kim과 Jeong (2004)은 과소표집과 과대표집을 이용한 데이터의 전처리와 부스팅을 결합하여 그 성능을 비교평가하였는데, 본 논문에서는 과소표집과 과대표집을 결합한 하이브리드 표집방법을 추가하고, 대표적인 앙상블 방법을 모두 활용한 모형들을 사용하여 불균형 자료에 대한 분류성능을 비교평가하고자 한다. 앙상블 기법의 예측모형으로는 CART (Breiman, 1984)를 사용한다. 또한 시뮬레이션을 통해 이들 모형의 성능을 비교평가하기 위해 척추모형(Backbone model, Nathalie와 Shaju (2002))을 사용하였다. 이 모형은 불균형 정도를 조절하여 데이터를 발생할 수 있어, 불균형 정도에 따라 모형을 비교평가 하기에 효율적이다. 시뮬레이션을 통하여 데이터 전처리의 효과가 어떤 불균형 정도에서 효과적인지도 함께 살펴보고자 한다. 마지막으로, UCI repository (Newman 등, 1998)와 NASA Metrics Data Program (Sayyad와 Menzies, 2005)에서 구한 실제 불균형 데이터에 여러 모형들을 적용해 봄으로써 불균형 자료에 대하여 분류성능을 높을 수 있는 모형이 무엇인지 알아보기로 한다.

2. 불균형 데이터에서의 성능평가

이 절에서는 불균형 데이터에 대한 여러 모형들을 비교평가하기 위해 가장 대표적인 평가 기준 4가지를 간단히 소개하기로 한다. 우선 Table 2.1은 분류모형에 대한 정오분류표를 나타낸 것이다.

Table 2.1에 나타난 결과로부터 다음과 같은 4가지의 분류기준을 정의할 수 있다.

- AUC: ROC(Receiver Operating Characteristic) 곡선은 분류기준의 변화에 따라 x축에는 1-특이도(가짜 양성률) y축에는 민감도(진짜 양성률)을 나타낸 것으로 이 곡선의 아래 영역을 AUC(Area Under Curve)로 정의한다. 이 영역이 넓을수록 분류결과가 좋다는 것을 나타낸다 (Huang과 Ling,

Table 2.1. The confusion matrix of a classification model

Actual Class	Predicted Class	
	Minority(Positive)	Majority(Negative)
Minority(Positive)	a (= True Positive)	b (= False Negative)
Majority(negaitive)	c (= False Positive)	d (= True Negative)

2005). Table 2.1의 정오분류표로부터 다음과 같이 정의한다.

$$\text{AUC} = \frac{1}{2}\left(1 + \frac{a}{a+b} - \frac{c}{c+d}\right).$$

- ACC: 정확도(Accuracy)는 분류모형의 성능을 평가할 때 가장 대표적으로 사용되는 평가 기준이다. 이 기준은 분류모형의 정확성과 부정확성을 동시에 살펴 볼 수가 있다. 정확도에 대한 정의는 다음과 같다.

$$\text{ACC} = \frac{a+d}{a+b+c+d}.$$

- F-measure: 이 값은 소수집단을 분류하는 데에 있어서 실제 소수집단을 모형에 의해 소수로 분류할 확률, 즉 민감도(sensitivity) $(= a/(a+b))$와 소수로 분류한 사람 중 실제로 소수집단에 속할 확률, 즉 양성예측도(positivie predictive value) $(= a/a+c)$의 조화평균으로 정의된다. 따라서 이 값이 클수록 소수집단에 대한 분류모형의 성능이 높다는 것을 나타낸다. 이 둘의 조화평균을 간단히 정리하면 다음과 같다.

$$F\text{-measure} = 2 \cdot \frac{\text{민감도} \times \text{양성예측도}}{\text{민감도} + \text{양성예측도}} = \frac{2a}{2a+b+c}.$$

- G-mean: Kubat와 Matwin (1997)이 제안한 이 값은 민감도와 특이도의 기하평균(geometric mean)으로 계산된다. 따라서 다수집단이 정확하게 분류되었지만 소수집단에 대한 예측력이 낮다면 이들의 기하평균값인 G-mean 값은 낮게 된다. 즉, 소수집단의 정분류율이 좋을수록 G-mean 값도 높아지는 경향이 있다. 이 값의 정의는 다음과 같다.

$$G\text{-mean} = \sqrt{\frac{a}{a+b} \times \frac{d}{c+d}}.$$

본 논문에서는 시뮬레이션 데이터와 실제 데이터 모두 훈련용 데이터와 평가용 데이터의 비를 7 : 3으로 분할 후, 이들 지표값을 계산하기 위해 평가용 데이터를 사용하며 이를 100번 반복 수행하여 결과를 분석하기로 한다.

3. 불균형 데이터의 전처리 과정 및 앙상블 기법

불균형 데이터에 대한 재표집(resampling) 방법은 여러 가지 방법이 있는데 다수집단에서 단순임의추출방법을 사용하여 소수집단과의 비율을 조정하는 과소표집 방법 (Ling과 Li, 1998), 소수집단을 반복추출하여 다수집단과의 비율을 조정하는 과대표집 방법 (Kubat와 Matwin, 1997), 그리고 이들 두 개의 표집 방법을 결합(hybrid-sampling)하여 분류성능을 높이려는 방법 등이 있다. 대표적인 예로 Chawla 등 (2002)에 의하여 연구된 SMOTE(Synthetic Minority Oversampling Technique)가 있다. SMOTE는 소수집단을 단순임의추출하여 복제를 하는 것이 아니라, 소수집단 개체의 최근접 이웃(k-

Nearest Neighbors) k개 중에서 임의로 추출하는 방법이다. 본 논문에서는 데이터 재표집 방법을 사용할 때 과소표집법은 다수집단을 소수집단의 수에 맞게 추출하여 그 비율을 1:1로, 과대표집법은 소수집단을 추출하여 다수집단의 수에 맞게 그 비율을 1:1로, SMOTE 방법은 소수집단의 수를 2배로 과대표집 후 다수집단을 과대표집된 소수집단의 수에 맞게 추출하여 그 비율을 1:1로 조정하였다.

앙상블(ensemble)이란 주어진 자료로부터 붓스트랩 표본을 얻고 이를 통해 여러 개의 예측모형을 생성한 후, 이들 예측모형들을 결합하여 하나의 최종 예측모형을 만드는 방법이다. 앙상블은 각 단일 예측모형의 오분류율을 낮추게 되어 분류성능을 높일 수 있다. 본 연구에서는 앙상블 기법 중에서 배깅(baggin, Breiman (1996)), 러프리 밸런스드 배깅(roughly balanced bagging, Hido 등 (2009)), 부스팅(boosting, Freund와 Schapire (1997)), 랜덤 포리스트(random forests, Breiman (2001)) 방법을 활용하여 불균형데이터에 대한 분류성능을 비교평가 하도록 한다. 예측모형으로는 CART 모형을 사용하였다. 본 논문에서는 시뮬레이션을 위해 R 프로그램을 사용하였으며, 러프리 밸런스드 배깅은 알고리즘을 생성하였고, 배깅의 경우에는 R 프로그램의 ipred 패키지 속에 포함된 bagging 함수(bagging, Breiman (1996)), 부스팅의 경우에는 ada 패키지 속에 포함된 ada함수 (Culp 등, 2006), 랜덤포리스트의 경우에는 randomForest 패키지 속에 포함된 randomForest함수(random forests, Breiman (2001))를 사용하였다.

- 배깅: 배깅은 훈련용 데이터 집합으로부터 크기가 같은 붓스트랩 표본을 생성하고, 각 표본에 맞는 단일예측모형을 생성한 후 평균예측모형을 토대로 오분류율을 낮추는 방법이다.
- 러프리 밸런스드 배깅: 이 방법은 배깅 전 데이터를 재표집할 때, 소수집단과 다수 집단의 균형을 맞추기 위해 과소표집을 사용한다. 다만 다수집단에서 추출되는 표본의 수가 음이항 분포에 따라 확률적으로 결정된다. 음이항 분포의 모수는 소수집단의 수로 결정된다. 따라서 각 표본에서 소수집단과 다수집단의 비율이 정확히 1:1은 아니지만, 데이터에 대한 전처리 없이 배깅만 사용하는 경우에 비해 데이터의 불균형을 조절할 수 있는 장점이 있고, 분류성능도 배깅보다 우수하다고 알려져 있다.
- 부스팅: 배깅에서는 데이터를 복원추출할 때 동일한 확률로 반복 추출하였지만, 부스팅은 데이터를 동일한 확률로 복원추출하지 않고, 오분류된 관측치에 가중치를 높여 표본으로 선택될 확률을 높여주게 된다. 본 연구에서 사용한 아다부스트(AdaBoost) 알고리즘은 매 반복마다 오분류된 관측치의 가중치는 증가시키고, 정분류된 가중치는 감소시키면서 예측모형을 적합시킨다. 일반적으로 불균형이 심할 경우 소수계급이 잘못분류될 가능성이 크고, 따라서 그런 데이터가 다음 반복에서 높은 가중치를 갖게되므로 좀 더 효과적인 방법이 될 개연성이 있다.
- 랜덤 포리스트: 이 방법은 이름처럼 나무모형의 결합을 나타내는데, 각 나무모형은 전체 예측변수들 중에서 독립적으로 선택된 변수들의 부분집합을 이용하게 된다. 이러한 방법은 모형예측에 있어 로버스트할 뿐 아니라 정확도가 높은 알고리즘으로 알려져 있다.

본 논문에서는 앞에서 살펴본 것처럼 불균형 데이터의 사전처리와 여러 앙상블 방법들을 고려하여 Figure 3.1에서 보는 바와 같이 모두 13가지 모형을 선정하고, 척추모형으로부터 불균형 정도를 조절하여 발생시킨 데이터와 실제 데이터에 이들 모형을 적합을 한 후, 이를 통해 데이터의 사전처리 효과와 앙상블 기법들의 성능을 비교 검토하고자 한다.

4. 시뮬레이션 데이터 분석

불균형 데이터의 비율에 따라서 앞서 소개한 13가지 모형들의 성능을 비교하기 위해 척추모형 (Nathalie와 Shaju, 2002)을 사용하기로 한다. 척추모형은 불균형의 비(i), 복잡도(c), 데이터의 크기(s)를

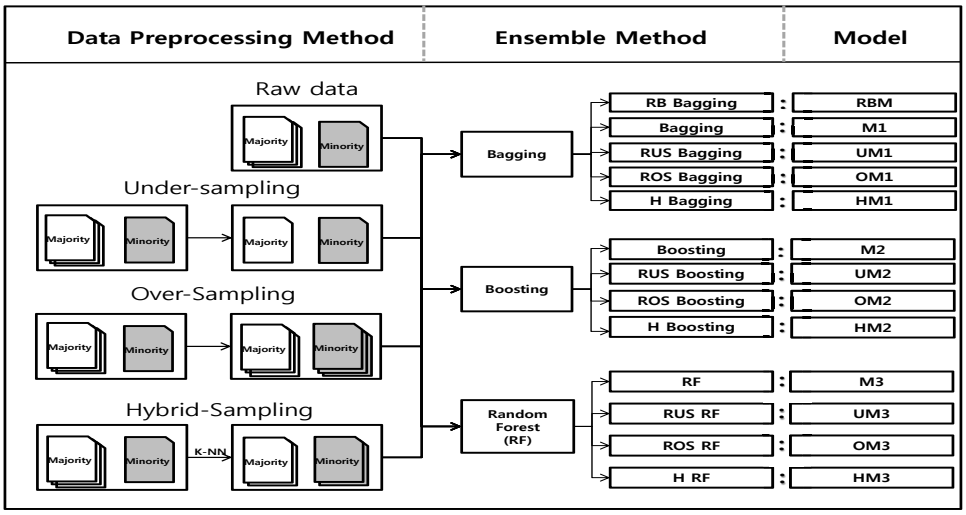

Figure 3.1. Thirteen models according to data preprocessing technique with ensemble methods for imbalanced data

Table 4.1. Data size of majority class and minority class generated by the Backbone model with $c = 3$ and $s = 1$

Imbalance Ratio (i)	Majority	Minority	Proportion of Minority
1	294	18	5.8%
3	250	16	19.9%
5	156	156	50.0%

모두 고려하여 데이터를 생성한다. 먼저 범위 $[0, 1]$의 균등분포 하에서 예측변수를 추출한 후 복잡도 c가 1인 경우 $2^c = 2$만큼 범위를 2등분 하여 $[0, 0.5)$의 범위에서는 목표변수값을 1(소수집단)로, $[0.5, 1]$범위는 0(다수집단)으로 목표변수를 생성한다. 만약 복잡도인 $c = 3$인 경우에는 $[0, 1]$범위를 8등분하여 $[0, 0.125), (0.25, 0.375), (0.5, 0.625), (0.75, 0.875)$의 범위에서는 목표변수를 1(소수집단)로 생성하고, 나머지 범위는 0으로 한다. 목표변수를 생성할 때 불균형의 비율(i)과 데이터의 크기(s)도 고려해서 목표변수가 0(다수집단)인 경우 데이터의 크기는 식 (4.1)을 따르고, 목표변수가 1인 경우 데이터의 크기는 식 (4.2)와 같다.

$$S_0 = \left(\frac{5000}{32} \times 2^s\right) / \left(1 + \frac{32}{2^i}\right) \times \frac{32}{2^i},$$
$$S_1 = \left(\frac{5000}{32} \times 2^s\right) / \left(1 + \frac{32}{2^i}\right).$$

이처럼 예측변수의 수를 고정하고, 불균형 비(i), 복잡도(c), 데이터 크기(s)에 따라 척추모형으로부터 데이터를 발생하면 Figure 3.1과 같다.

Table 4.1은 척추모형에서 복잡도 $c = 3$과 데이터의 크기 $s = 1$로 설정 후 불균형의 비(i)에 따라 생성된 데이터 크기를 나타낸다. $i = 1$일때 소수집단의 수가 5.8%로 불균형이 심하고, $i = 3$ 또는 $i = 5$일 경우에는 약 20% 또는 50%로 불균형의 정도가 심하지 않은 경우를 나타낸다. 이들 데이터로부터 Figure 4.1의 13가지 모형을 적용한 결과는 다음과 같다.

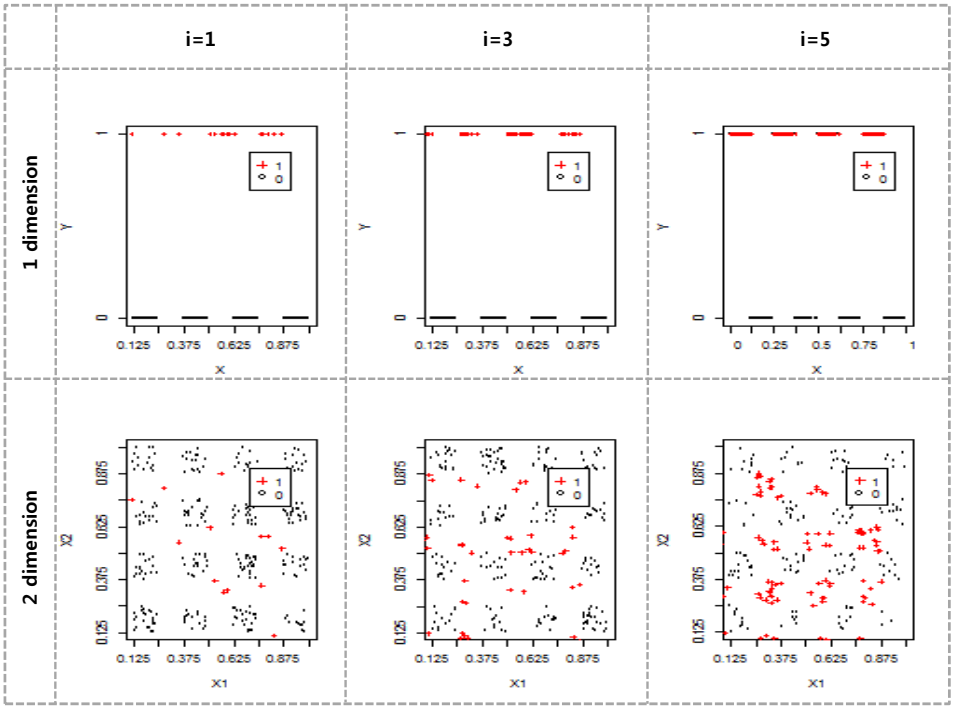

Figure 4.1. Data description generated by Backbone model according to the imbalance ratio(i) for $c = 3$ and $s = 1$): First-row figures display the data with one predictor variable and second-row figures the data with two predictor variables

Table 4.2. Performance for all the models applied to the data generated by Backbone model with one predictor variable with $i = 1$, $c = 3$, $s = 1$

Imblance ratio ($i = 1$)	AUC	ACC	F-measure	G-mean
RBM	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
M1	0.989 ± 0.074	0.979 ± 0.071	0.875 ± 0.108	0.988 ± 0.078
M2	0.857 ± 0.078	0.979 ± 0.073	0.833 ± 0.071	0.845 ± 0.078
M3	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
UM1	0.839 ± 0.077	0.725 ± 0.076	0.483 ± 0.113	0.795 ± 0.082
UM2	0.680 ± 0.080	0.549 ± 0.076	0.250 ± 0.109	0.593 ± 0.076
UM3	0.994 ± 0.073	0.978 ± 0.073	0.925 ± 0.101	0.984 ± 0.074
OM1	0.989 ± 0.074	0.982 ± 0.073	0.867 ± 0.099	0.978 ± 0.075
OM2	0.989 ± 0.085	0.979 ± 0.076	0.867 ± 0.089	0.984 ± 0.075
OM3	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
HM1	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
HM2	0.890 ± 0.074	0.893 ± 0.073	0.520 ± 0.101	0.805 ± 0.073
HM3	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000

예측변수가 1개인 Table 4.2에서는 모든 기준치에서 러프리 밸런스드 배깅이 배깅보다 성능이 좋게 나타났지만, 예측변수가 두 개인 Table 4.4와 Table 4.5에서는 그렇지 않았다. 또, Table 4.2에서 살펴보

Table 4.3. Performance for all the models applied to the data generated by Backbone model with one predictor variable with $i = 3$, $c = 3$, $s = 1$.

Imblance ratio ($i = 3$)	AUC	ACC	F-measure	G-mean
RBM	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
M1	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
M2	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
M3	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
UM1	0.994 ± 0.082	0.986 ± 0.076	0.972 ± 0.113	0.993 ± 0.077
UM2	0.954 ± 0.076	0.980 ± 0.076	0.942 ± 0.109	0.945 ± 0.080
UM3	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
OM1	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
OM2	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
OM3	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
HM1	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
HM2	0.992 ± 0.064	0.981 ± 0.072	0.974 ± 0.101	0.985 ± 0.063
HM3	0.995 ± 0.073	0.991 ± 0.075	0.974 ± 0.057	0.994 ± 0.072

Table 4.4. Performance for all the models applied to the data generated by Backbone model with two predictor variables with $i = 1$, $c = 3$, $s = 1$

Imblance ratio ($i = 1$)	AUC	ACC	F-measure	G-mean
RBM	0.682 ± 0.052	0.984 ± 0.074	0.384 ± 0.074	0.680 ± 0.080
M1	0.989 ± 0.074	0.979 ± 0.071	0.745 ± 0.108	0.988 ± 0.078
M2	0.857 ± 0.078	0.979 ± 0.073	0.362 ± 0.071	0.845 ± 0.078
M3	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
UM1	0.827 ± 0.077	0.725 ± 0.076	0.483 ± 0.113	0.795 ± 0.082
UM2	0.680 ± 0.080	0.949 ± 0.076	0.200 ± 0.109	0.598 ± 0.076
UM3	0.998 ± 0.073	0.978 ± 0.068	0.500 ± 0.101	0.984 ± 0.074
OM1	0.984 ± 0.074	0.982 ± 0.072	0.867 ± 0.099	0.978 ± 0.075
OM2	0.985 ± 0.085	0.979 ± 0.076	0.865 ± 0.089	0.984 ± 0.075
OM3	0.980 ± 0.085	0.980 ± 0.069	0.825 ± 0.087	0.875 ± 0.075
HM1	0.895 ± 0.053	0.765 ± 0.085	0.325 ± 0.102	0.768 ± 0.052
HM2	0.890 ± 0.074	0.893 ± 0.073	0.253 ± 0.098	0.862 ± 0.073
HM3	0.951 ± 0.053	0.875 ± 0.052	0.520 ± 0.101	0.924 ± 0.025

면 부스팅의 경우 AUC 기준으로 과대표집의 경우에 성능이 향상 되었으나, 과소표집이나 하이브리드 표집에서는 오히려 분류 성능이 크게 떨어짐을 알 수 있다. 특히 F-measure를 통해서 볼 때 소수계급에 대한 예측이나 양성예측도를 살펴보면 과소표집이나 하이브리드 표집의 방법이 오히려 앙상블 성능을 저하시키고 있음을 알 수 있다. 같은 조건에서 데이터의 불균형 비율을 $i = 3$으로 하는 경우 Table 4.3에서 알 수 있듯이 데이터의 전처리 효과는 거의 없다는 것을 알 수 있다. 오히려 과소표집이나 하이브리드 표집의 경우 성능이 약간 떨어짐을 알 수 있다.

Table 4.4는 Figure 4.1에서 두 개의 예측변수를 갖고, $i = 1$인 불균형 정도가 심한 데이터에 대한 적합결과이다. 부스팅 기법의 경우 Table 4.2에서의 결과와 마찬가지로 과소표집한 경우 전처리가 없는 경우나 과대표집한 경우에 비해 덜 효과적임을 알 수 있었다. 그러나 Table 4.5의 결과에서 알 수 있듯이 불균형 정도가 심하지 않고 소수집단의 비율이 약 20%정도만 되어도 불균형을 맞추려는 데이터의 전처

Table 4.5. Performance for all the models applied to the data generated by Backbone model with two predictor variables with $i = 3$, $c = 3$, $s = 1$

Imblance ratio ($i = 3$)	AUC	ACC	F-measure	G-mean
RBM	0.984 ± 0.074	0.982 ± 0.073	0.867 ± 0.099	0.978 ± 0.075
M1	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
M2	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
M3	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
UM1	0.998 ± 0.073	0.978 ± 0.073	0.925 ± 0.101	0.984 ± 0.074
UM2	0.984 ± 0.072	0.982 ± 0.068	0.867 ± 0.099	0.978 ± 0.068
UM3	0.998 ± 0.068	0.978 ± 0.054	0.925 ± 0.048	0.984 ± 0.064
OM1	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
OM2	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
OM3	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
HM1	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
HM2	0.989 ± 0.074	0.979 ± 0.071	0.875 ± 0.108	0.988 ± 0.078
HM3	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000

Table 5.1. Summary description of the imbalanced data sets used in this study

Dataset	No. of Obs.	No. of predictors	Minority(%)
Ecoli	336	7	10.4
Page-block	5473	10	10.2
Abalone	731	8	9.9
Flag	194	26	9.8
CM1	496	21	9.7
Vowel	988	13	9.2
Cleveland	173	13	7.5
Hyper-Thyroid	2278	23	5.9
Letter	20000	16	4.0
PC1	1107	21	0.7

리 과정은 불필요해 보인다.

Figure 4.1에서 살펴보면 척추모형에 의한 데이터발생은 예측변수들의 영역을 반복적으로 분할하여 목표집단을 정했기 때문에, 예측모형으로 CART를 사용할 경우 데이터의 불균형이 심하지 않은 경우에는 AUC나 ACC 등의 값이 대부분 1로 완벽히 예측되는 것을 알 수 있다. 그러나 불균형이 심한 Table 4.4나 Table 4.5 경우에는 불균형이 심하지 않은 경우에 비해 예측성능이 떨어지며, 데이터의 전처리로부터 예측성능을 높일 수 있음을 알 수 있다. 또한 앙상블 방법들 중 랜덤 포리스트의 성능이 높게 나타났음을 알 수 있다.

5. 실제 데이터 분석

이 절에서는 실제의 불균형 데이터에 예측모형을 적합할 때 어떤 방법이 가장 효과적이었는지 알아보기 위해, 앞서 소개한 13가지의 모형을 실제 자료에 적합해보도록 한다. 시뮬레이션 데이터와 마찬가지로 훈련용 데이터와 평가용 데이터를 7 : 3으로 분할하여, 100번 반복 수행하기로 한다. 또한, 앙상블을 사용할 때 예측모형의 수를 100개로 똑같이 고정하기로 한다.

Table 5.2. AUC results for 10 data set using thirteen models

	Ecoli	Page-Block	Abalone	Flag	CM1	Vowel	Cleveland	Hyper-thyroid	Letter	PC1
RBM	0.816±0.002	0.942±0.002	0.806±0.002	0.784±0.004	0.709±0.003	0.985±0.002	0.806±0.003	0.981±0.004	0.967±0.003	0.813±0.006
M1	0.735±0.012	0.924±0.020	0.653±0.022	0.544±0.021	0.694±0.035	0.998±0.034	0.595±0.033	0.979±0.035	0.975±0.034	0.675±0.025
M2	0.731±0.034	0.921±0.026	0.722±0.022	0.567±0.035	0.694±0.025	0.999±0.033	0.721±0.037	0.975±0.038	0.989±0.035	0.606±0.036
M3	0.784±0.003	0.938±0.007	0.638±0.025	0.586±0.004	0.739±0.026	0.995±0.003	0.615±0.005	0.972±0.004	0.974±0.003	0.653±0.028
UM1	0.869±0.048	0.957±0.045	0.763±0.044	0.787±0.038	0.739±0.035	0.985±0.042	0.675±0.043	0.978±0.041	0.996±0.045	0.843±0.038
UM2	0.907±0.035	0.964±0.034	0.797±0.037	0.696±0.036	0.718±0.032	0.976±0.037	0.582±0.038	0.984±0.036	0.978±0.036	0.812±0.037
UM3	0.812±0.045	0.968±0.048	0.800±0.042	0.678±0.041	0.745±0.028	0.989±0.043	0.812±0.040	0.980±0.042	0.981±0.042	0.824±0.041
OM1	0.806±0.038	0.925±0.035	0.718±0.024	0.725±0.030	0.649±0.015	0.945±0.031	0.712±0.033	0.991±0.035	0.997±0.032	0.715±0.036
OM2	0.773±0.011	0.961±0.026	0.815±0.017	0.717±0.015	0.687±0.010	0.998±0.023	0.832±0.027	0.989±0.025	0.987±0.017	0.831±0.029
OM3	0.685±0.067	0.956±0.069	0.670±0.036	0.865±0.068	0.723±0.028	0.962±0.065	0.745±0.061	0.981±0.059	0.987±0.063	0.725±0.067
HM1	0.831±0.045	0.985±0.048	0.781±0.041	0.856±0.041	0.731±0.043	0.990±0.044	0.795±0.045	0.985±0.040	0.988±0.048	0.842±0.043
HM2	0.884±0.028	0.989±0.030	0.784±0.026	0.892±0.027	0.745±0.029	0.991±0.024	0.815±0.023	0.991±0.021	0.980±0.020	0.815±0.026
HM3	0.893±0.037	0.990±0.040	0.825±0.038	0.918±0.036	0.755±0.032	0.992±0.037	0.843±0.033	0.995±0.035	0.987±0.038	0.832±0.034

Table 5.3. ACC results for 10 data set using thirteen models

	Ecoli	Page-Block	Abalone	Flag	CM1	Vowel	Cleveland	Hyper-thyroid	Letter	PC1
RBM	0.814±0.005	0.942±0.003	0.699±0.003	0.593±0.004	0.718±0.002	0.939±0.003	0.808±0.002	0.961±0.002	0.954±0.003	0.694±0.005
M1	0.849±0.001	0.948±0.002	0.945±0.005	0.834±0.007	0.906±0.003	0.983±0.005	0.904±0.008	0.993±0.009	0.996±0.011	0.945±0.012
M2	0.858±0.003	0.967±0.006	0.960±0.003	0.842±0.008	0.886±0.004	0.997±0.007	0.942±0.001	0.989±0.002	0.999±0.005	0.926±0.008
M3	0.913±0.004	0.945±0.005	0.938±0.007	0.864±0.010	0.926±0.012	0.990±0.013	0.923±0.015	0.990±0.012	0.997±0.011	0.923±0.014
UM1	0.842±0.034	0.966±0.031	0.698±0.058	0.610±0.042	0.664±0.027	0.956±0.028	0.698±0.034	0.968±0.029	0.984±0.025	0.756±0.022
UM2	0.832±0.037	0.968±0.034	0.743±0.044	0.525±0.037	0.652±0.034	0.933±0.031	0.581±0.029	0.959±0.024	0.997±0.028	0.742±0.025
UM3	0.822±0.026	0.965±0.020	0.756±0.047	0.576±0.039	0.681±0.034	0.960±0.035	0.895±0.031	0.948±0.038	0.990±0.035	0.698±0.031
OM1	0.861±0.021	0.921±0.028	0.932±0.008	0.797±0.012	0.829±0.026	0.970±0.019	0.904±0.020	0.978±0.015	0.977±0.014	0.925±0.012
OM2	0.879±0.016	0.949±0.014	0.925±0.005	0.814±0.015	0.805±0.014	0.997±0.013	0.942±0.016	0.982±0.011	0.999±0.018	0.937±0.017
OM3	0.859±0.042	0.935±0.038	0.954±0.005	0.864±0.013	0.893±0.016	0.980±0.019	0.938±0.015	0.979±0.019	0.996±0.013	0.838±0.011
HM1	0.851±0.034	0.954±0.031	0.788±0.032	0.881±0.024	0.748±0.025	0.973±0.022	0.769±0.028	0.968±0.025	0.989±0.021	0.824±0.026
HM2	0.871±0.042	0.969±0.029	0.782±0.026	0.797±0.028	0.698±0.025	0.980±0.027	0.846±0.024	0.970±0.022	0.995±0.029	0.812±0.025
HM3	0.881±0.034	0.970±0.024	0.792±0.031	0.847±0.024	0.753±0.035	0.976±0.030	0.923±0.031	0.979±0.031	0.999±0.008	0.825±0.028

Table 5.4. *F*-measure results for 10 data set using thirteen models

	Ecoli	Page-Block	Abalone	Flag	CM1	Vowel	Cleveland	Hyper-thyroid	Letter	PC1
RBM	0.457±0.002	0.735±0.005	0.282±0.002	0.294±0.005	0.253±0.007	0.782±0.008	0.483±0.006	0.816±0.005	0.795±0.004	0.485±0.007
M1	0.526±0.029	0.875±0.028	0.425±0.031	0.185±0.035	0.232±0.036	0.926±0.027	0.395±0.031	0.962±0.033	0.961±0.029	0.685±0.024
M2	0.516±0.048	0.866±0.042	0.585±0.045	0.165±0.048	0.216±0.042	0.975±0.041	0.495±0.046	0.950±0.050	0.984±0.051	0.596±0.049
M3	0.632±0.068	0.892±0.051	0.375±0.071	0.226±0.058	0.352±0.060	0.965±0.062	0.462±0.068	0.953±0.066	0.968±0.061	0.671±0.067
UM1	0.529±0.035	0.779±0.034	0.265±0.037	0.316±0.038	0.262±0.035	0.816±0.036	0.325±0.039	0.816±0.032	0.862±0.037	0.513±0.038
UM2	0.541±0.038	0.824±0.031	0.301±0.033	0.152±0.035	0.235±0.037	0.794±0.039	0.235±0.034	0.807±0.038	0.971±0.036	0.425±0.034
UM3	0.471±0.027	0.775±0.028	0.264±0.033	0.205±0.030	0.245±0.029	0.842±0.027	0.685±0.026	0.793±0.025	0.931±0.025	0.419±0.022
OM1	0.560±0.035	0.850±0.034	0.471±0.044	0.261±0.038	0.185±0.037	0.868±0.036	0.516±0.035	0.935±0.039	0.816±0.038	0.485±0.037
OM2	0.571±0.025	0.896±0.026	0.539±0.024	0.361±0.028	0.216±0.025	0.986±0.024	0.735±0.029	0.967±0.022	0.978±0.025	0.675±0.029
OM3	0.560±0.068	0.900±0.057	0.483±0.079	0.218±0.062	0.234±0.068	0.915±0.063	0.621±0.069	0.959±0.060	0.965±0.067	0.681±0.071
HM1	0.513±0.030	0.839±0.034	0.319±0.038	0.543±0.032	0.315±0.033	0.906±0.036	0.485±0.037	0.865±0.031	0.901±0.038	0.591±0.034
HM2	0.581±0.024	0.836±0.021	0.316±0.024	0.465±0.020	0.296±0.025	0.926±0.026	0.516±0.024	0.895±0.028	0.975±0.027	0.576±0.029
HM3	0.574±0.038	0.848±0.034	0.334±0.037	0.539±0.033	0.265±0.035	0.968±0.037	0.875±0.039	0.906±0.031	0.969±0.040	0.587±0.037

Table 5.5. *G*-mean results for 10 data set using thirteen models

	Ecoli	Page-Block	Abalone	Flag	CM1	Vowel	Cleveland	Hyper-thyroid	Letter	PC1
RBM	0.807±0.003	0.923±0.002	0.796±0.002	0.778±0.004	0.709±0.003	0.967±0.005	0.804±0.002	0.979±0.003	0.959±0.004	0.812±0.002
M1	0.728±0.035	0.922±0.031	0.558±0.038	0.534±0.031	0.562±0.032	0.991±0.019	0.598±0.011	0.974±0.018	0.969±0.020	0.621±0.033
M2	0.729±0.031	0.915±0.038	0.667±0.037	0.548±0.019	0.554±0.036	0.998±0.020	0.700±0.019	0.972±0.012	0.980±0.019	0.606±0.020
M3	0.784±0.026	0.933±0.024	0.534±0.006	0.563±0.017	0.542±0.002	0.994±0.022	0.600±0.025	0.961±0.027	0.963±0.022	0.615±0.006
UM1	0.868±0.035	0.955±0.043	0.757±0.043	0.702±0.041	0.721±0.039	0.976±0.034	0.651±0.030	0.971±0.035	0.983±0.030	0.841±0.031
UM2	0.907±0.034	0.958±0.039	0.793±0.038	0.469±0.035	0.675±0.030	0.963±0.027	0.534±0.025	0.981±0.029	0.971±0.027	0.776±0.024
UM3	0.812±0.027	0.950±0.043	0.799±0.042	0.587±0.038	0.732±0.047	0.978±0.025	0.757±0.029	0.976±0.034	0.974±0.030	0.761±0.038
OM1	0.806±0.025	0.920±0.021	0.675±0.033	0.617±0.019	0.562±0.034	0.931±0.012	0.679±0.015	0.960±0.014	0.984±0.011	0.637±0.013
OM2	0.768±0.025	0.951±0.019	0.806±0.020	0.695±0.029	0.591±0.028	0.998±0.015	0.789±0.019	0.984±0.020	0.986±0.028	0.827±0.016
OM3	0.685±0.067	0.947±0.040	0.583±0.062	0.563±0.030	0.571±0.064	0.954±0.027	0.700±0.025	0.972±0.036	0.987±0.031	0.718±0.028
HM1	0.829±0.026	0.961±0.038	0.779±0.042	0.844±0.035	0.704±0.034	0.985±0.032	0.694±0.031	0.975±0.037	0.988±0.035	0.812±0.026
HM2	0.884±0.045	0.963±0.035	0.783±0.025	0.889±0.030	0.708±0.022	0.989±0.025	0.736±0.026	0.988±0.021	0.980±0.022	0.796±0.014
HM3	0.885±0.039	0.963±0.024	0.790±0.038	0.917±0.029	0.698±0.035	0.987±0.030	0.779±0.024	0.976±0.025	0.981±0.022	0.816±0.019

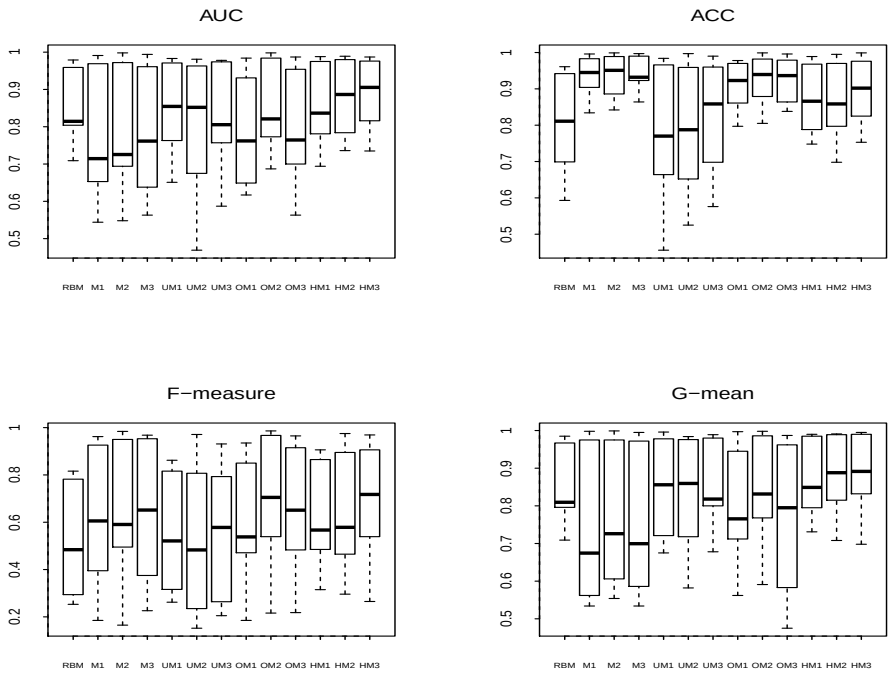

Figure 5.1. Performances of thirteen models averaged across all data sets

5.1. 실제 데이터 소개

본 논문에서는 UCI repository (Newman 등, 1998)와 NASA Metrics Data Program (Sayyad와 Menzies, 2005)로부터 소수집단의 비율이 10% 이하인 실제 데이터 10개를 선정하여 분석에 사용하였다. 데이터에 관한 자세한 설명은 다음의 Table 5.1과 같다.

5.2. 모형 비교 결과

Table 5.2부터 Table 5.5까지는 13가지 예측모형을 100번의 반복실험을 통해 평가용 데이터에 대입하여 구한 AUC, ACC, F-measure, 그리고 G-mean에 대한 평균값과 표준오차를 나타낸다. 대체적으로 데이터의 전처리 없이 앙상블만 적용한 모형들에 비해 데이터의 전처리를 사용한 모형에서 예측 성능이 향상된 것을 알 수 있다.

Figure 5.1은 모형별로 10개 데이터로부터 계산된 지표값들의 분포를 나타낸 그림이다. Figure 5.1에서 살펴보면 모형 HM3는 ACC를 제외한 나머지 분류기준인 AUC, F-measure, 그리고 G-mean 기준에서 평균적으로 성능이 우수한 것으로 나타났다. 이에 HM3을 기준으로 나머지 12개 모형들과의 성능차이에 대해 각각의 분류기준별로 윌콕슨 순위 검정을 실시하여, Table 5.6에 그 결과를 표로 작성하였다. 그 결과 AUC 측면에서 봤을 때 하이브리드 표집법을 사용 후 랜덤포리스트 모형을 수행한 HM3 모형과 다른 모형들의 성능 비교 결과, HM3모형의 예측성능이 가장 좋은 것으로 나타났다. 그리고 Figure 5.1을 보면 ACC 측면에서 살펴보면, 데이터 전처리를 수행하지 않은 모형들이 데이터 전처리를 사용한 모형보다 예측성능이 높은 것을 볼 수 있는데, 이것은 소수집단보다 다수집단을 더 정확하게 분류한 것에 기인한 것으로 보인다.

Table 5.6. Results for model comparison based on Wilcoxon test

comparison	p-value			
	AUC	ACC	F-measure	G-mean
HM3 vs. RBM	0.005**	0.005**	0.005**	0.203
HM3 vs. M1	0.007**	0.575	0.646	0.009**
HM3 vs. M2	0.013*	0.314	0.646	0.013*
HM3 vs. M3	0.007**	0.051	0.575	0.007**
HM3 vs. UM1	0.022*	0.005**	0.005**	0.169
HM3 vs. UM2	0.013*	0.005**	0.007**	0.053
HM3 vs. UM3	0.005**	0.005**	0.005**	0.097
HM3 vs. OM1	0.009**	0.959	0.059	0.007**
HM3 vs. OM2	0.018*	0.192	0.646	0.575
HM3 vs. OM3	0.008**	0.314	0.760	0.009**
HM3 vs. HM1	0.025*	0.059	0.059	0.202
HM3 vs. HM2	0.005**	0.011*	0.074	0.314

$*$: p-value < 0.05; $**$: p-value < 0.01

6. 결론

본 논문에서는 분류문제에 있어 소수집단의 경우가 10% 이하인 불균형 데이터에 대하여 데이터의 균형을 맞추기 위한 데이터의 전처리와 앙상블 기법을 결합한 여러 알고리즘들에 대한 성능비교를 위해, 시뮬레이션과 실제의 불균형 데이터를 이용하여 비교 검토해 보았다. 그 결과 목표변수의 불균형이 심한 데이터의 경우에는 데이터의 전처리를 통하여 균형을 맞춘 데이터를 기초로 앙상블을 사용할 때 데이터의 전처리 없이 앙상블을 사용할 때보다 예측성능이 좋아짐을 확인하였다. 예측변수의 수가 1개인 시뮬레이션 데이터에서 데이터 균형에 대한 전처리 없이 부스팅한 결과가 과소표집으로 균형을 맞춘 후 부스팅한 예측성능만 나빠졌을 뿐, 예측변수가 여러 개인 실제 데이터에서 살펴보아도 균형을 맞춘 데이터의 분류성능이 높아짐을 알 수 있다. 다만 불균형 비율이 20% 정도일때 모의실험 된 데이터에서 살펴보면 데이터의 계급간 균형을 맞추려는 사전처리 후 앙상블 기법을 사용한 것이나 앙상블만 사용한 알고리즘이나 그 성능에 큰 차이가 없어 데이터의 사전처리는 불균형 정도가 심한 경우에만 효과적이라는 것을 짐작할 수 있다. 또한 시뮬레이션과 실제 자료의 분석결과 앙상블 기법마다 결합하여 사용할 때 효과적인 재표집방법이 있는 것으로 나타났는데, AUC나 G-mean 기준에서 살펴볼 때, 배깅이나 랜덤포리스트의 경우에는 과소표집, 부스팅의 경우에는 과대표집법과 함께 쓰일 때 분류 성능이 많이 향상되는 경향이 있는 것으로 나타났다. 배깅의 경우 이미 배깅만 사용한 경우보다 러프리 밸런스드 배깅을 사용한 경우에 예측성능이 좋아진다는 사전결과를 보아도, 배깅의 경우 과소표집법과 함께 쓰일 때 예측 성능이 좋아짐을 알 수 있다. 본 논문에서 사용한 SMOTE와 같이 과대표집법과 과소표집법을 함께 사용하는 재표집방법의 경우 AUC나 G-mean 기준에서 살펴볼 때 모든 앙상블 기법에서 그 성능이 향상되는 것으로 나타났다. AUC 기준으로 데이터의 사전처리를 SMOTE를 사용하고 랜덤포리스트를 사용한 경우에 성능이 가장 좋은 것으로 나타났다. 마지막으로 본 논문에서 수행한 시뮬레이션 데이터의 경우 분류를 하기 쉬운 형태로 생성되기 때문에, 다음 연구에서는 다양한 시뮬레이션 데이터와 좀 더 많은 실제 데이터를 사용하여 연구할 필요가 있다고 하겠다.

References

Batista, G. E. A. P. A., Prati, R. C. and Monard, M. C. (2004). A study of the behavior of several methods

for balancing machine learning training data, *Special Interest Groups Knowledge Discovery in Data Explorations Newsletter*, **6**, 20–29.

Breiman, L. (1984). Algorithm CART, *California Wadsworth International Group, Belmont, CA.*, **6**, 20–29.

Breiman, L. (1996). Bagging predictors, *Machine Learning.*, **24**, 123–140.

Breiman, L. (2001). Random forests, *Machine Learning*, **45**, 5–32.

Chawla, N. V., Bowyer, W. K., Hall, L. O. and Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique, *Journal Of Artificial Intelligence Research* , **16**, 321–357.

Chawla, N. V., Cieslak, D., Hall, L. and Joshi, A. (2008). Automatically countering imbalance and its empirical relationship to cost, *Data Mining and Knowledge Discovery*, **17**, 225–252.

Chawla, N. V., Japkowicz, N. and Kolcz, A. (2004). Special issue learning imbalanced datasets, *Special Interest Groups Knowledge Discovery in Data Explorations Newsletter*, **6**, 1–6.

Chawla, N. V., Lazarevic, A., Hall, L. O. and Bowyer, W. K. (2003). Smoteboost: Improving prediction of the minority class in boosting, *Seventh European Conference on Principles and Practice of Knowledge Discovery in Databases*, 107–119.

Culp, M., Johnson, K. and Michailidis, G. (2006). Ada : An r package for stochastic boosting, *Journal of Statistical Software*, **16**, 321–357.

Freitas, A., Costa- Pereira, A. and Brazdil, P. (2007). Cost-sensitive decision trees applied to medical data, *Data Warehousing and Knowledge Discovery*, **4654**, 302–312.

Freund, Y. and Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting, *Journal of Computer and System Sciences*, **55**, 119–139.

Galar, M., Fernandez, A., Barrenechea, E., Bustince, H. and Herrera, F. (2012). A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches, *Institute of Electrical and Electronics Engineers*, **42**, 463–484.

Hido, S., Kashima, H. and Takahashi, Y. (2009). Roughly balanced bagging for imbalanced data, *Statistical Analysis and Data Mining*, **2**, 412–426.

Huang, J. and Ling, C. (2005). Using AUC and accuracy in evaluating learning algorithms, *Knowledge and Data Engineering, Institute of Electrical and Electronics Engineers*, **17**, 299–310.

Hur, J. and Kim, J. (2007). Decision tree induction with imbalanced data set: A case of health insurance bill audit in a general hospital, *Information systems review*, **9**, 45–65.

Kang, P. and Cho, S. (2006). EUS SVMS: Ensemble of under-sampled SVMS for data imbalance problems, *Lecture Notes in Computer Science*, **4232**, 837–846.

Khreich, W., Granger, E., Miri, A. and Sabourin, R. (2010). Iterative boolean combination of classifiers in the roc space: An application to anomaly detection with HMMs, *Pattern Recognition*, **43**, 2732–2752.

Kim, J. and Jeong, J. (2004). Classification of class-imbalanced data: Effect of over-sampling and under-sampling of training data, *The Korean Journal of Applied Statistics*, **17**, 445–457.

Kim, J. and Park, H. (2012). Imbalanced data analysis using sampling methods, Inha University.

Kubat, M., Holte, R. C. and Matwin, S. (1998). Machine learning for the detection of oil spills in satellite radar images, *Machine Learning*, **30**, 195–215.

Kubat, M. and Matwin, S. (1997). Addressing the curse of imbalanced data sets: One-sided sampling, *Proceedings of the Fourteenth International Conference on Machine Learning*, 179–186.

Ling, C. X. and Li, C. (1998). Data mining for direct marketing: Problems and solutions., *Knowledge Discovery in Data-98*.

Liu, X., Wu, J. and Zhou, Z. (2009). Machine learning for the detection of oil spills in satellite radar images, *Institute of Electrical and Electronics Engineers*, **39**, 539–550.

Mazurowski, M. A., Habas, P. A., Zurada, J. M., Lo, J. Y., Baker, J. A. and Tourassi, G. D. (2008). Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance, *Neural Networks*, **21**, 427–436.

Nathalie, J. and Shaju, S. (2002). The class imbalance problem: A systematic study, *Intelligent Data Analysis*, **6**, 429–449.

Newman, D., Hettich, S., Blake, C. and Merz, C. (1998). UCI repository of machine learning databases, *Department of Information and Computer Science*.

Oh, J. and Zhang, B. (2001). Kernel perceptron boosting for effective learning of imbalanced data, *Proceed-*

ings of the Korean Information Science Society Conference, 304–306.

Sayyad, S. J. and Menzies, T. J. (2005). The promise repository of software engineering databases, *http://promise.site.uottawa.ca/SERepository*.

Seiffert, C., Khoshgoftaar, T., Van Hulse, J. and Napolitano, A. (2010). Rusboost: A hybrid approach to alleviating class imbalance, *Institute of Electrical and Electronics Engineers*, **40**, 185–197.

Wu, G. and Chang, E. (2005). KBA: Kernel boundary alignment considering imbalanced data distribution, *Institute of Electrical and Electronics Engineers*, **17**, 786–795.

Yang, Q. and Wu, X. (2006). 10 challenging problems in data mining research, *International Journal of Information Technology and Decision Making*, **5**, 597–604.

Zadrozny, B. and Elkan, C. (2001). Learning and making decisions when costs and probabilities are both unknown, *Knowledge Discovery in Data '01 Proceedings of the seventh Association for Computing Machinery Special Interest Groups Knowledge Discovery in Data international conference on Knowledge discovery and data mining*, 204–213.

Zhu, Z. and Song, Z. (2010). Fault diagnosis based on imbalance modified kernel fisher discriminant analysis, *Chemical Engineering Research and Design*, **88**, 936–951.

데이터 전처리와 앙상블 기법을 통한 불균형 데이터의 분류모형 비교 연구

이희재[a] · 이성임[a,1]

[a]단국대학교 응용통계학과

(2013년 8월 26일 접수, 2013년 10월 7일 수정, 2013년 12월 16일 채택)

요 약

최근 들어 데이터 마이닝의 분류문제에 있어 목표변수의 불균형 문제가 많은 관심을 받고 있다. 이러한 문제를 해결하기 위해, 이전 연구들은 원 자료에 대하여 데이터 전처리 과정을 실시했는데, 전처리 과정에는 목표변수의 다수계급을 소수계급의 비율에 맞게 조정하는 과소표집법, 소수계급을 복원추출하여 다수계급의 비율에 맞게 조정하는 과대표집법, 소수계급에 K-최근접 이웃 방법 등을 활용하여 과대표집법을 적용 후 다수계급에는 과소표집법을 적용한 하이브리드 기법 등이 있다. 또한 앙상블 기법도 이러한 불균형 데이터의 분류 성능을 높일 수 있다고 알려져 있어, 본 논문에서는 데이터의 전처리 과정과 앙상블 기법을 함께 고려한 여러 모형들을 사용하여, 불균형 자료에 대한 이들모형의 분류성능을 비교평가한다.

주요용어: 불균형 데이터, 배깅, 부스팅, 랜덤 포리스트, 과소표집, 과대표집, SMOTE.

이 연구는 2014년도 단국대학교 대학연구비의 지원으로 연구되었음.

[1]교신저자: (448-701) 경기도 용인시 수지구 죽전로 152, 단국대학교 응용통계학과.
E-mail: silee@dankook.ac.kr