The Korean Journal of Applied Statistics (2015) DOI: http://dx.doi.org/10.5351/KJAS.2015.28.3.495
28(3), 495–509

Classification Analysis for Unbalanced Data

Dongah Kim[a] · Suyeon Kang[a] · Jongwoo Song[a,1]

[a]Department of Statistics, Ewha Womans University

(Received March 20, 2015; Revised April 27, 2015; Accepted April 28, 2015)

Abstract

We study a classification problem of significant differences in the proportion of two groups known as the unbalanced classification problem. It is usually more difficult to classify classes accurately in unbalanced data than balanced data. Most observations are likely to be classified to the bigger group if we apply classification methods to the unbalanced data because it can minimize the misclassification loss. However, this smaller group is misclassified as the larger group problem that can cause a bigger loss in most real applications.

We compare several classification methods for the unbalanced data using sampling techniques (up and down sampling). We also check the total loss of different classification methods when the asymmetric loss is applied to simulated and real data. We use the misclassification rate, G-mean, ROC and AUC (area under the curve) for the performance comparison.

Keywords: up-sampling, down-sampling, asymmetric loss, misclassification rate, G-mean, ROC, AUC, logisitic regression, SVM, random forest

1. 서론

다양한 종류의 분류 문제들을 보면 균형이 맞는 데이터 보다는 균형이 맞지 않는 데이터의 경우를 종종 볼 수 있다. 예를 들어 스팸메일여부, 은행에서 대출해주는 기업의 파산여부, 공항입국자의 테러리스트 여부, 보험회사에서 고객의 가입여부 등등 여러 종류의 불균형 데이터들이 있다. 특히, 두 집단의 비율 차이가 아주 큰 경우에는 분류방법론을 이용하여 두 집단을 정확하게 분류하기가 쉽지 않다. 많은 경우에 작은 집단은 큰 집단으로 오분류 되지만 전체적인 분류 성능을 본다면, 오분류율이 작으므로 좋게 나타나는 경우가 대부분이다. 하지만, 이런 불균형데이터의 분류의 경우에는 작은 집단을 큰 집단으로 잘못 분류하면 그 반대상황 보다 훨씬 더 큰 손실(loss)을 가져오는 경우가 많다. 극단적인 예로, 공항 입국자의 경우에 거의 대부분 테러리스트가 아니겠지만, 만약 테러리스트를 일반 입국자라고 분류하면 그 반대 상황보다 훨씬 큰 대가를 치룰 것이다. 이런 경우에 접근할 수 있는 간단한 방법론은 sampling을 통하여 두 집단의 비율을 비슷하게 맞춘 후에 분류 방법론을 적용하는 것이다. 우리는 이런 경우에 원데

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the ministry of Education, Science and Technology (No. NRF-2013R1A1A2012817).
[1]Corresponding author: Department of Statistics, Ewha Womans University, Seoul 120-750, Korea.
E-mail: josong@ewha.ac.kr

이터를 사용해서 분류하는 것보다 어느 정도 성능을 개선할 수 있는 지를 simulation data와 실제 데이터를 이용하여 비교 분석하고자 한다. 이 경우에 사용할 수 있는 sampling 방법은 큰 집단에서 임의의 관측치 표본을 이용한 down-sampling이나 작은 집단에서 bootstrap sample이나 아주 작은 noise를 더한 sample을 이용한 up-sampling 방법을 사용할 수 있다. 이런 sampling 기법을 사용한 방법론으로는 SHRINK (Kubat 등, 1997)와 SMOTE (Chawla 등, 2002)가 있다. 물론 up-sampling을 통해서 기존의 정보량보다 더 많은 것을 얻을 수는 없지만 두 클래스를 비슷한 규모로 만듦으로써 비슷한 weight을 줄 수 있는 장점이 있다. 그리고 Chen등이 2004년에 weighted Random Forest를 이용하여 unbalanced data의 분류 문제를 분석하고 다른 방법론과 성능을 비교하였다 (Chen 등, 2004).

우리는 또한 asymmetric loss를 가정한 경우에 각각의 분류 방법론들의 분류 정확성에 대해서도 분석하고자 한다. 분석에 사용되는 분류 방법론들은 logistic regression, support vector machines (Vapnik, 1998), 그리고 Random Forest (Breiman, 2001)이다. 본 논문에서 모든 계산은 R (R Development Core Team, 2010)을 이용하여 이루어 질 것이며, 우리는 R에서 제공하는 다양한 분류 함수에서 어떤 옵션이 asymmetric loss를 줄 수 있는 지를 상세한 R코드를 제공함으로써 알리고자 한다. 본 논문은 다음과 같은 순서로 되어있다. 2장에서는 sampling 방법과 R의 다양한 분류 함수에서 어떻게 asymmetric loss를 주는 가를 설명한다. 그리고 성능 비교를 위하여 사용되는 오분류율, G-mean (Kubat 등, 1997), ROC, 그리고 AUC (Park 등, 2011)에 대해서 간략히 설명한다. 3장에서는 simulation data와 실제 데이터에서 각각의 분류 방법론의 성능을 원데이터와 sampling방법론을 이용하여 비교하고 또한 어떤 방법론에서 total loss가 최소화 되는 지를 알아본다. 4장에서는 본 연구의 결과를 요약한다.

2. 방법론

이번 장에서는 불균형 데이터의 분류에 쉽게 사용될 수 있는 sampling 방법과 오분류가 일어났을 때 비대칭 손실을 R의 분류함수에서 적용시키는 방법을 설명한다. 그리고 분류 방법론간의 성능비교에 사용될 수 있는 오분류율과 G-mean, ROC, AUC 등을 설명한다.

2.1. Sampling 방법

Sampling 방법은 두 가지로 나누어진다. 첫 번째 방법은 큰 집단을 작은 집단의 크기에 맞추어 무작위로 뽑는 방법(Down-sampling)이고, 두 번째 방법은 작은 집단을 큰 집단의 크기에 맞추어 반복추출하는 방법(Up-sampling)이다. Figure 2.1은 이변량 정규분포를 따르는 두 개의 그룹이며 규모가 900 : 100로 unbalanced data이다. Original 데이터를 가지고 down-sampling, up-sampling을 하였을 때 어떻게 데이터가 변하는지를 그림으로 보여주었다. Up-sampling의 경우 규모가 작은 집단을 규모가 큰 집단에 맞추기 때문에 100개의 데이터 $(X_1, Y_1), (X_2, Y_2), (X_3, Y_3), \ldots, (X_{100}, Y_{100})$를 9번 반복해서 추출하게 된다. 이 경우에 bootstrap 샘플과 같이 with replacement를 허용하는 방법과 아주 작은 노이즈를 더해주는 두 가지 방법을 생각해 볼 수 있는데, 본 연구에서는 bootstrapping 기법을 이용하였다.

2.2. 비대칭 손실(Asymmetric Loss) 가정

비대칭 손실의 적용은 작은 집단을 큰 집단으로 잘못 추정하는 것에 더 큰 페널티를 주는 방법으로, 예를 들어 큰 집단과 작은 집단의 비율이 9 : 1인 데이터의 경우 큰 집단을 작은 집단으로 잘못 추정할 경우 1이라는 페널티를 주고 작은 집단을 큰 집단으로 잘못 추정하는 경우 9라는 페널티를 주어서 작은 집단을 큰 집단으로 잘못 추정하는 것을 피하고자 노력하게 하는 방법이다. 우리는 R에서 어떤 옵션을 이용

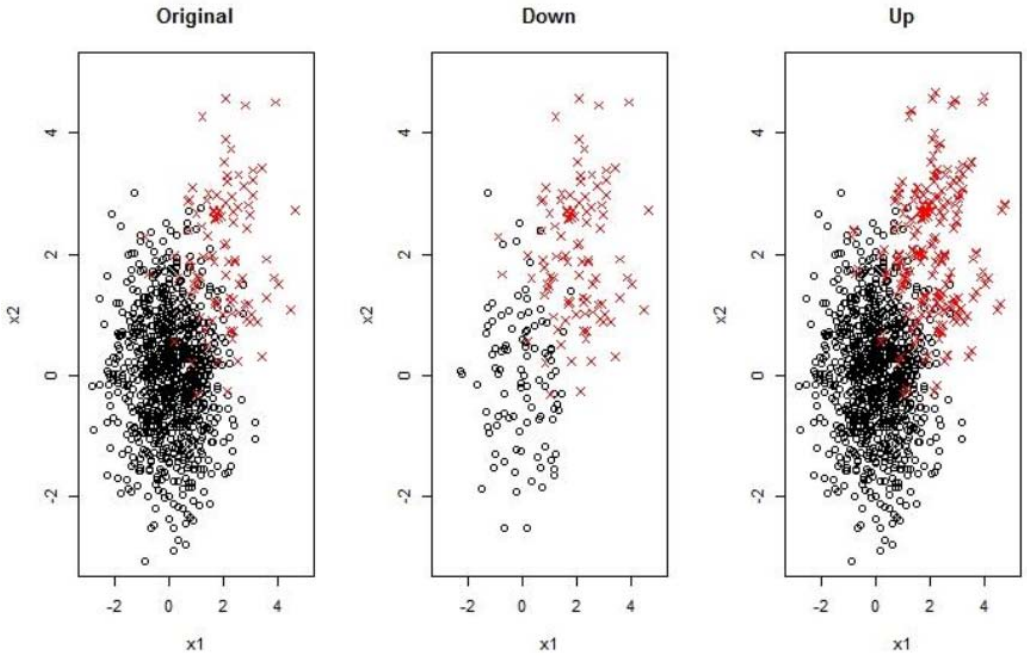

Figure 2.1. Scatter plots of Original, Down-sampling, and Up-sampling data.

해서 비대칭 손실을 줄 수 있는지를 설명하고자 한다. 일반적으로 weight 옵션을 이용해서 다른 손실을 줄 수 있는데, 대표적인 분류 방법론인 Logistic Regression, SVM, 그리고 Random Forest의 경우에는 다음과 같다.

2.2.1. Logistic Regression(LOGREG) Logistic regression은 분류방법론 중에 가장 많이 쓰이는 방법론 중에 하나로 각각의 클래스의 사후확률 odds비가 설명변수 X와 선형관계를 가진다고 가정한다.

$$\log \frac{P(G=k|X=x)}{P(G=K|X=x)} = \beta_{k0} + \beta_k^t x, \quad k = 1, \ldots, K-1.$$

이 모형 하에서 회귀계수에 대한 추정은 IRLS(Iterative Reweighted Least Squared)를 이용하여 최대우도추정량(MLE)를 일반적으로 사용한다. Logistic regression의 가장 큰 장점중에 하나는 해석의 용이함이다. 예를 들어, 특정 설명변수가 1 증가했을 때 특정 클래스의 사후확률의 증가/감소 계산이 가능하므로 설명변수의 변화가 사후확률에 어떤 영향을 미치는 지를 파악할 수 있다.

R에서 Logistc Regression의 경우에는 `glm()` 함수 또는 nnet 패키지의 `multinom()` 함수를 이용해서 모형 적합이 가능하다. `glm()` 함수는 반응변수가 2가지 값을 갖는 2-class 분류 문제에 적용할 수 있고, `multinom()` 함수는 반응변수가 3가지 이상의 값을 갖는 multi-class 분류 문제에도 적용이 가능하다. 2-class 문제에는 두 함수 모두 적용될 수 있는데, 본 연구에서는 multi-class 분류 문제로도 자연스럽게 확장 가능한 `multinom()` 함수를 이용하였다.

아래 코드는 $y = 0$인 관측치들이 큰 집단이고, $y = 1$인 관측치들이 작은 집단일 경우의 예이다.

multinom() 함수의 weights 옵션에는 총 관측치 개수만큼의 길이를 갖는 벡터가 들어갈 수 있는데, 큰 집단과 작은 집단의 비율이 9:1인 데이터의 경우에는 작은 집단을 큰 집단으로 잘못 추정하는 것에는 9의 가중치를 주었고, 큰 집단을 작은 집단으로 잘못 추정하는 것에는 1의 가중치를 주었다.

```
> library("nnet")
> wts1 <- ifelse(traindata$y==0, 1, 9)
> logis.res <- multinom(y~., data= traindata, weights=wts1)
```

2.2.2. Support Vector Machine(SVM) SVM은 Vapnik이 제안한 분류방법론으로 기본적인 아이디어는 관측치에서 decision boundary까지의 최단거리를(이를 마진이라 한다) 최대화하는 decision boundary를 찾는 것이다. 이렇게 관측치와 decision boundary사이에 가능한 큰 버퍼를 둠으로서 오분류가 일어날 확률을 줄일 수 있고 따라서 많은 실제 자료에서 좋은 성능을 보인다. SVM에서 가장 많이 사용되는 tuning parameter는 gamma와 cost이다. SVM은 여러 가지 다양한 커널을 사용할 수 있지만 가장 많이 사용되는 것은 Gaussian radial basis function이고 gamma는 이 커널에서 사용된다. Cost는 오분류된 관측치의 페널티를 결정하는데 사용된다. R에서 SVM에 관련된 패키지로는 kernlab, e1071, klaR, svmpath 등이 있는데 (Karatzoglou 등, 2006), 우리는 e1071 패키지의 svm() 함수를 이용하여 분석할 것이다. e1071 패키지는 livsvm 라이브러리를 R에서 사용할 수 있게 하며 다양한 커널과 parameter tuning, visualization 이 가능하다는 장점이 있다.

먼저 tune.svm() 함수를 이용해 CV error를 최소화하는 최적의 모수 값을 찾은 후, svm() 함수를 통해 모형 적합을 할 수 있다. svm() 함수의 class.weights 옵션이 바로 asymmetric loss와 관련된 옵션인데, class 개수만큼의 길이를 갖는 벡터가 들어가야 하며 작은 집단을 큰 집단으로 추정했을 때와 큰 집단을 작은 집단으로 추정했을 때의 가중치를 다르게 줄 수 있다.

```
> library("e1071")
 ## parameter tuning
> tobj <- tune.svm(y~., data=traindata, gamma=2^(-8:3), cost=2^(-8:3))
> bestGamma <- tobj$best.parameters[[1]]
> bestC <- tobj$best.parameters[[2]]
> wts <- table(traindata$y)
> wts[1] <- 1; wts[2] <- 9
> svm.res <- svm(y~., data=traindata, type="C-classification",
+      kernel="radial", gamma=bestGamma, cost=bestC,
+      class.weights = wts)
```

2.2.3. Random Forest(RF) Random Forest는 Brieman이 2001년에 제안한 방법론으로 상관계수가 낮은 tree들을 이용한 bagging 방법론이다. Tree를 만들어 나갈 때 모든 설명변수를 고려하지 않고 임의로 작은 개수의 설명변수만 고려해서 tree를 만들어 나가면 각각의 bootstrap 샘플에서 만들어진 tree들의 상관계수가 낮아지게 되고 따라서 전체 bagged된 estimates의 분산을 줄일 수 있다. RF는 이런 임의의 설명변수의 개수(m)과 bootstrap의 샘플 수와 같은 tuning parameter가 있지만 별도의 tuning없이도 상당히 좋은 결과를 보여준다. R에서는 randomForest 패키지가 있지만 class별로 다른 loss를 주기 위해서 우리는 party 패키지에 있는 cforest() 함수를 이용하였다.

cforest() 함수에서는 weights 옵션을 통해 비대칭 손실을 적용할 수 있으며, LOGREG 모형 적합에

Table 2.1. Confusion matrix

	Predicted 0	Predicted 1
True 0	TN (True Negative)	FP (False Positive)
True 1	FN (False Negative)	TP (True Positive)

서와 마찬가지로 총 관측치 개수만큼의 길이를 갖는 벡터가 들어갈 수 있다. `control` 옵션에서의 `mtry` 값은 RF 모형을 적합할 때 각 노드에서 랜덤으로 선택되는 설명변수의 개수를 의미하며, 분류분석에서는 일반적으로 전체 설명변수 개수의 루트값을 내림하여 `mtry` 값으로 사용한다.

```
> library("party")
> wts1 <- ifelse(traindata$y==0, 1, 9)
> m <- floor(sqrt(ncol(traindata)-1))
> rf.res <- cforest(y~., data=traindata, weights=wts1,
+      control=cforest_unbiased(mtry=m))
```

2.3. 모형 평가

2.3.1. 오분류에 대한 다양한 추정치 분류방법론에서는 적합된 모형이 분류를 얼마나 정확하게 했는지를 보기 위해서 여러 가지 추정치를 사용한다. 우선 추정치들을 설명하기에 앞서 꼭 살펴보아야 할 것이 있는데, 바로 오분류표이다. 본 연구에서 사용하는 데이터는 이집단 분류(two-class classification)이므로 오분류표는 Table 2.1과 같은 2×2 matrix가 된다. 오분류표는 적합된 모형의 분류 결과를 시각화한 것으로서, 모형 평가 시 사용되는 추정치들을 정의하는 데 기본이 된다. 이 표를 이용하여 본 연구에서 사용할 추정치들을 정의하면 다음과 같다.

$$\begin{aligned}
\text{오분류율(misclassification rate)} &= \frac{\text{FP}+\text{FN}}{\text{TN}+\text{FP}+\text{FN}+\text{TP}},\\
\text{정분류율(accuracy)} &= \frac{\text{TN}+\text{TP}}{\text{TN}+\text{FP}+\text{FN}+\text{TP}},\\
\text{민감도(sensitiyity)} &= \frac{\text{TP}}{\text{FN}+\text{TP}},\\
\text{특이도(specificity)} &= \frac{\text{TN}}{\text{TN}+\text{FP}}.
\end{aligned}$$

즉, 오분류율이란 전체 데이터에서 잘못 분류한 자료의 비율이며 값이 작을수록 보다 정확한 예측이 가능하다고 할 수 있다. 하지만 오분류율은 기본적으로 대칭 손실을 가정하기 때문에 대부분의 불균형 자료에 바로 적용하기에는 한계가 있다. 왜냐하면, 일반적으로 작은 그룹에 속한 관측치를 큰 그룹으로 잘못 분류하는 것이 그 반대보다 손실이 큰데, 불균형 자료에서는 대부분의 관측치가 큰 그룹으로 추정되는 경향을 보이므로 단순히 오분류율 값을 기준으로 모형을 선택할 경우 매우 큰 손실이 발생할 수 있기 때문이다. 정분류율(accuracy)은 (1 − 오분류율)로 정의되며 값이 클수록 좋다.

그리고 민감도는 $y = 1$ 클래스에 속한 자료 중 정분류된 자료의 비율(True Positive Rate; TPR)이며, 특이도는 $y = 0$인 클래스에 속한 자료 중 정분류된 자료의 비율(True Negative Rate; TNR)이다. 이 값들은 절단값을 변경함으로써 바꿀 수 있는데, 극단적으로 절단값을 0으로 놓으면 모든 자료를 $y = 1$ 클래스에 할당하므로 민감도는 1이고 특이도는 0이 된다. 반대로 절단값을 1로 놓으면 민감도는 0, 특이도는 1이 된다. 일반적으로 민감도와 특이도를 동시에 크게 하는 것은 아주 어려운 일이고 두 가지를

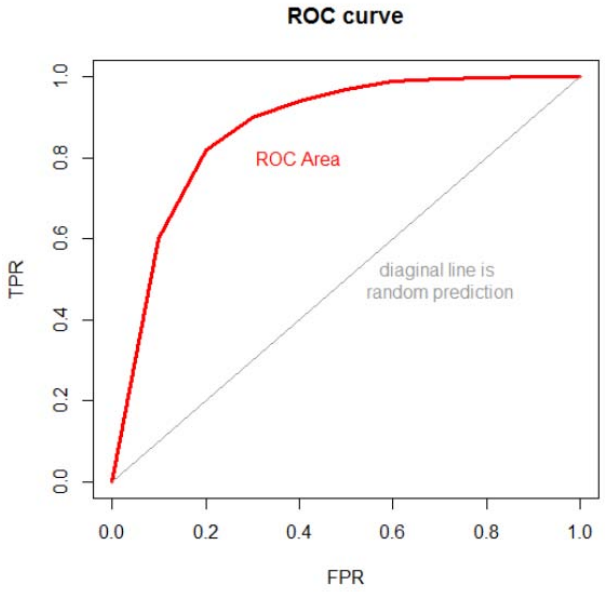

Figure 2.2. ROC curve

동시에 크게 할 수 있는 방법론이 좋은 분류 방법론이라고 할 수 있다.

$$\text{G-mean} = \sqrt{\text{sensitivity} \times \text{specificity}} = \sqrt{\text{TPR} \times \text{TNR}}.$$

그런데 불균형 자료의 경우 분류기가 작은 그룹($y = 1$ 클래스)에 속한 관측치를 큰 그룹($y = 0$ 클래스)의 관측치로 추정하는 경향을 보이므로 특이도 값은 크지만 민감도 값은 매우 작게 된다. 이런 이유로 민감도와 특이도의 기하평균인 G-mean이 새로운 평가 지표로 제안되었고 (Kubat 등, 1997), 이후로도 G-mean은 불균형 자료에서 모형의 성능을 평가하는 데 많이 이용되었다 (Kubat 등, 1997; Wu와 Chang, 2003). 이렇듯 불균형 자료의 경우에는 단순히 오분류율을 비교하는 것보다 G-mean으로 비교하는 것이 의미가 있기 때문에 본 연구에서는 각 방법론에서의 G-mean 값도 계산하여 비교하였다.

2.3.2. ROC(Receiver Operating Characteristic) 곡선

ROC 곡선은 여러 절단값에서의 민감도(sensitivity; TPR)와 특이도(specificity; TNR)의 관계를 보여준다. 이 곡선을 통하여 절단값의 변화에 따른 분류기의 성능을 눈으로 확인할 수 있다. ROC 곡선의 x축은 1-specificity이고 y축은 sensitivity로, 절단값을 변화시키며 구한 점들을 연결하면 ROC 곡선이 된다.

$$\text{sensitiyity} = \frac{\text{TP}}{\text{FN} + \text{TP}} = \text{TPR},$$
$$1 - \text{specificity} = 1 - \frac{\text{TN}}{\text{TN} + \text{FP}} = \frac{\text{FP}}{\text{TN} + \text{FP}} = \text{FPR}.$$

위에서도 언급하였듯 민감도와 특이도의 관계는 한 쪽을 크게 하면 다른 한 쪽이 작아지는 관계이므로, ROC 곡선의 x축과 y축을 나타내는 민감도와 (1 − 특이도)는 반대로 양의 상관관계를 갖는다. 따라서 ROC 곡선은 Figure 2.2와 같이 증가하는 함수의 형태를 보인다. 또한, ROC 곡선 아래의 면적인

AUC(Area Under the Curve; ROC Area)는 분류방법론에서 성능평가를 위해 자주 사용되는 통계량으로 그 값이 클수록 예측력이 우수한 분류기라고 평가할 수 있다. 완전 랜덤하게 자료를 분류한 경우의 ROC 곡선은 $y = x$ 직선이 되고 AUC 값은 0.5이므로, AUC가 이 값보다 작으면 랜덤으로 분류하는 것보다 좋지 않기 때문에 잘못된 예측이라 한다.

3. 모의실험 및 실제 데이터 분석 결과

3.1. 모의실험 결과

본 절에서는 모의실험으로 다양한 비율을 가진 불균형 데이터를 여러 가지 분류방법론을 이용하여 모형을 적합한 후 각 모형의 분류정확도를 비교할 것이다. 우리는 정규분포를 이용하여 simulation data를 생성하였고 단계별 모의실험 과정은 다음과 같다.

(단계 1) 다음과 같은 이변량 정규분포를 따르는 X_1을 900개, X_2를 100개 생성한다.

$$X_1 \sim N(\mu_1, \Sigma_1), \quad \text{where } \mu_1 = (0,0),\ \Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

$$X_2 \sim N(\mu_2, \Sigma_2), \quad \text{where } \mu_2 = (2,2),\ \Sigma_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

(단계 2) 900개의 데이터를 가지는 집단에 y값 0을, 100개의 데이터를 가지는 집단에 y값 1을 주어 unbalanced training data를 만든다.

(단계 3) Up/Down sampling 기법을 사용하여 데이터를 5 : 5의 비율로 만든다.

(단계 4) LOGREG, SVM, RF의 3가지 분류방법론을 사용하여 모형을 적합한다. 이 때, Original 데이터와 Up/Down sampling 기법을 사용한 데이터 각각에 대해 3가지 방법론을 적용하여 총 9개의 모형을 만든다.

(단계 5) 위에서 만들어진 training 데이터와 동일한 분포와 비율을 갖는 데이터 1000개를 생성하여 test 데이터로 사용한다.

(단계 6) 적합된 모형을 이용하여 test 데이터에서 오분류율, G-mean, AUC 값을 구하고 ROC 곡선을 그린다. 이 때, 정확성을 위해 1000회 반복하여 평균을 구한다.

(단계 7) Original 데이터에 비대칭 손실을 적용하여 (단계 6)을 시행한다.

(단계 8) 단계 1–7을 다른 비율의 데이터를 가지고 동일하게 시행한다.

3.1.1. Sampling 기법을 사용한 경우의 모의실험 결과 1000회 반복하여 나온 오분류율의 평균과 표준편차 결과는 Table 3.1과 같다. 오분류율 값은 작을수록 좋지만 unbalanced data의 경우는 이 값만으로 판단하는 것은 한계가 있다. Table 3.1의 오분류율 값만 보면 현재 unbalanced data를 보정해주는 sampling 방법을 사용하기 전의 Original data가 가장 좋은 것으로 확인된다. 어떻게 분류되었는지 오분류표를 살펴보면 다음과 같다.

Table 3.2–3.4에서 보는 것과 같이 Original 데이터의 경우 오분류한 부분을 살펴보면 실제로 작은 집단(1)을 큰 집단(0)이라 분류하는 쪽이 반대의 경우보다 큰 것으로 확인된다. 하지만 sampling 방법을 사용한 결과를 보면 오분류율은 높지만 Original 데이터와는 반대의 결과를 보인다. 예를 들어, 공항입국자의 테러리스트 여부를 판단하는 경우 보통 사람(0)을 테러리스트(1)로 잘못 분류하는 것보다 테러

Table 3.1. Misclassification rate for 9 : 1 dataset (Means (Std. deviations))

	Original	Down	Up
LOGREG	0.0393 (0.0059)	0.0808 (0.0142)	0.0786 (0.0119)
SVM	0.0425 (0.0058)	0.0809 (0.0124)	0.0772 (0.0121)
RF	0.0417 (0.0057)	0.0969 (0.0209)	0.0611 (0.0105)

Table 3.2. Confusion matrix of LOGREG model (average value)

		Predict					
		Original		Down		Up	
		0	1	0	1	0	1
True	0	886.87	13.13	327.08	72.92	829.45	70.56
	1	26.19	73.81	7.86	92.15	8.02	91.97

Table 3.3. Confusion matrix of SVM model (average value)

		Predict					
		Original		Down		Up	
		0	1	0	1	0	1
True	0	893.56	6.44	827.14	72.86	830.77	69.23
	1	36.08	63.92	8.08	91.92	7.92	92.08

Table 3.4. Confusion matrix of RF model (average value)

		Predict					
		Original		Down		Up	
		0	1	0	1	0	1
True	0	887.15	12.85	811.80	88.20	853.16	46.84
	1	28.86	71.14	8.71	91.29	14.23	85.77

리스트(1)를 보통사람(0)으로 잘못 분류하는 것이 훨씬 더 심각한 문제를 초래하기 때문에 sampling 방법을 사용한 결과가 더 좋은 결과라 볼 수 있다. 추가적으로 각 방법론에서 적합된 모형을 각종 통계량을 이용하여 비교해보기 위해 2.3절에 제시하였던 모형 평가 지표들을 살펴보았다. 단, ROC 곡선은 1회 적합한 결과를 사용하여 그렸으며 ROC 곡선의 성능을 나타내는 AUC 값은 1000회 반복한 평균값을 사용하였다.

Figure 3.1은 sampling 기법을 적용했을 때와 하지 않았을 때의 ROC 곡선을 겹쳐 그린 그림이다. 각 방법론에서 도출된 ROC 곡선이 거의 유사한 것으로 보인다.

Table 3.5는 앞에서 적용했던 모든 방법론들의 결과값을 정리한 표이다. Accuracy와 G-mean, 그리고 AUC는 0부터 1까지의 값을 가지며 1에 가까울수록 성능이 좋다고 할 수 있다. Accuracy는 LOGREG, SVM, RF 3가지 모형 모두 original data에서 가장 높으나 G-mean은 데이터의 균형을 맞추는 sampling 기법을 적용했을 때 더 높은 값을 보였다. 이는 불균형 데이터를 그대로 쓰는 것보다 균형을 맞추고 분석을 진행했을 때 예측이 더 잘 된다는 의미이다. AUC의 경우에는 모든 방법론에서 비슷한 값을 보였다. G-mean과 AUC에서의 결과가 다른 것은 특정 절단값에서의 성능을 보는 G-mean과는 달리 AUC는 모든 절단값에서의 전체적인 성능을 보는 평가 지표이기 때문인 것으로 보인다.

3.1.2. 비대칭 손실을 적용한 경우의 모의실험 결과 본 절에서는 어떤 방법론에서 total loss가 최소화되는지를 살펴볼 것이다. Total loss란 오분류된 자료에 대한 손실(loss)을 모두 합한 값으로, 불균

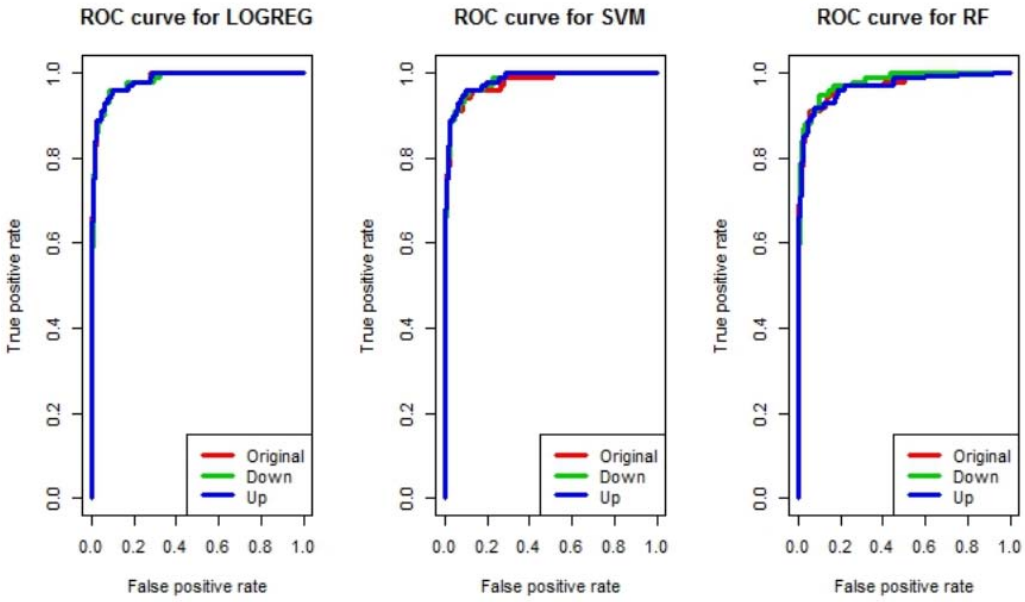

Figure 3.1. ROC curves comparing classification performance of three machine learning models.

형 자료에 대한 total loss는 큰 집단을 작은 집단으로 오분류한 경우와 작은 집단을 큰 집단으로 오분류한 경우에 다른 loss를 적용하여 계산한다. 예를 들어, 앞에서 생성한 모의실험 데이터와 같이 큰 집단이 0, 작은 집단이 1의 값을 갖고 그 비율이 9 : 1인 데이터가 주어졌다고 하자. 이 경우 total loss는 큰 집단(0)을 작은 집단(1)으로 오분류한 FP에는 loss 1을 주고 작은 집단(1)을 큰 집단(0)으로 오분류한 FN에는 더 큰 값인 loss 9를 주어 구할 수 있다. 즉, 이처럼 positive sample이 작은 집단에 해당할 경우 total loss는 다음과 같은 식으로 표현될 수 있다.

$$\begin{aligned}\text{total loss} &= \text{FP} \times \frac{\text{FN}+\text{TP}}{\text{TN}+\text{FP}+\text{FN}+\text{TP}} \times 10 + \text{FN} \times \frac{\text{TN}+\text{FP}}{\text{TN}+\text{FP}+\text{FN}+\text{TP}} \times 10 \\ &= \text{FP} \times \text{작은집단비율} \times 10 + \text{FN} \times \text{큰집단비율} \times 10.\end{aligned}$$

Table 3.6은 각 방법론에서 도출된 total loss 값을 정리한 표이다. 표의 3, 4번째 열은 각각 모형을 적합할 때 대칭/비대칭 손실을 적용한 경우의 total loss 값이며 마지막 열은 비대칭 손실을 적용함으로써 total loss가 얼마나 감소되었는지를 나타낸다. 표를 보면 total loss의 절대적 수치와 감소율 모두 SVM 모형에서 가장 좋으며, 세 모형에서 공통적으로 작은 집단의 비율이 작아질수록 total loss의 감소율이 커지는 것을 관찰할 수 있다. 이는 작은 집단의 비율이 매우 작을 경우, 즉 자료의 불균형이 매우 심할 경우 비대칭 손실을 적용했을 때의 보정 효과가 더 크기 때문인 것으로 보인다. 특히 작은 집단의 비율이 1%일 때 1000개의 관측치 모두를 큰 집단으로 예측하였던 SVM과 RF 모형 중 SVM 모형은 비대칭 손실을 적용함으로써 80%가 넘는 total loss 감소율을 보였다.

3.2. 실증분석

3.2.1. 실증 데이터 본 장에서는 총 3가지의 실제 데이터를 이용하여 모의실험에서의 분석을 실시하였다. 분석에 사용된 데이터는 Table 3.7과 같다.

Table 3.5. Performance metrics for simulation data (Means (Std. deviations))

Model	Data	Accuracy	G-mean	AUC
LOGREG	original	0.9608 (0.0059)	0.8524 (0.0278)	0.9769 (0.0064)
	down	0.9193 (0.0142)	0.9200 (0.0145)	0.9767 (0.0065)
	up	0.9215 (0.0119)	0.9205 (0.0143)	0.9769 (0.0063)
SVM	original	0.9575 (0.0058)	0.7958 (0.0366)	0.9721 (0.0089)
	down	0.9192 (0.0124)	0.9190 (0.0141)	0.9781 (0.0065)
	up	0.9229 (0.0121)	0.9217 (0.0135)	0.9769 (0.0063)
RF	original	0.9584 (0.0057)	0.8368 (0.0304)	0.9704 (0.0086)
	down	0.9032 (0.0209)	0.9070 (0.0166)	0.9694 (0.0077)
	up	0.9389 (0.0105)	0.9014 (0.0204)	0.9684 (0.0091)

Table 3.6. Total loss for simulation data (Means (Std. deviations))

Model	%작은집단	Symmetric Loss	Asymmetric Loss	%Decrease
LOGREG	10%	248.82 (43.01)	144.75 (26.27)	41.83%
	5%	175.94 (35.32)	77.96 (17.44)	55.69%
	1%	58.48 (14.61)	19.38 (10.22)	66.86%
SVM	10%	326.68 (48.69)	144.23 (26.12)	55.85%
	5%	312.84 (56.49)	78.15 (20.73)	75.02%
	1%	99.00 (0.00)	17.10 (9.30)	82.73%
RF	10%	272.59 (45.98)	193.90 (38.20)	28.87%
	5%	200.85 (34.46)	100.20 (23.00)	50.11%
	1%	99.00 (0.00)	35.45 (17.57)	64.20%

첫 번째 데이터는 전화를 사용하여 은행에서 직접 광고를 하였을 때 정기예금에 가입하는지 여부에 대한 데이터이다. 전체데이터는 4,521개이며 그중 가입을 한 사람 수는 521명으로 전체의 11.5%가 정기예금에 가입한 unbalanced data이다. 사용된 변수는 나이, 직업, 결혼상태(결혼, 싱글, 이혼), 교육, 부채유무, 년 평균잔고, 주택담보, 전화기 종류(모름, 집전화, 휴대폰), 마지막으로 연락한 날, 마지막으로 연락한 달, 마지막 연락 후 기간, 연락횟수, 이전 연락 후 다음연락 까지 걸린 기간, 이전 광고에서의 효과이다.

두 번째 데이터는 UCI data sets에서 adult.test라는 데이터로, 미국 경제지표자료이며 성인 근로자의 수입이 5만 달러 이상인지 아닌지에 대해 이 분류 되어있는 자료이다. 총 16,281개의 자료 가운데 1,221개의 자료에 결측값이 존재하는데, 이는 전체의 약 7.5%로 적은 수이므로 정확한 분석을 위하여 삭제하기로 한다. 이제 남은 15,060개의 관측치 중 수입이 5만 달러 이상인 사람은 3,700명으로 전체의 24.57%를 차지한다. 이 때 데이터의 불균형을 더 심하게 하기 위해 5만 달러 이상인 사람 중 30%을 랜덤으로 추출하였다. 그래서 전체 데이터 수는 12,470개이며, 5만 달러 이상인 사람의 수가 1,110명으로 작은 집단이 전체의 8.90%를 차지하는 데이터가 만들어진다. 사용된 변수로는 나이, 직업분류(자영업, 회사원, 등등) 교육, 교육받은 기간, 결혼여부, 직종, 인종, 성별, 출신국가, 노동시간, 부양가족 수 등이 있다.

세 번째 데이터는 Pima Indian 21세 이상 여성의 당뇨병 발병여부에 관한 자료이다. 사용한 변수는 임신한 횟수, 공복혈당 농도, 혈압, 삼두근 피부 주름 두께, 2시간 혈청 인슐린, BMI, 당뇨 집안 내력, 나이이다. 전체 데이터 수는 768개이며, 당뇨병을 앓고 있는 사람의 수는 268명인데 그 중 20%를 임의로 추출하여 전체 500개의 데이터 중 당뇨병 환자가 54명으로 작은 집단이 9.7%인 데이터를 만들어 주었다.

Table 3.7. Summary of datasets

Dataset	사용한 변수 개수	총 데이터 수	%작은집단
Bank	16	4,521	11.5
Adult	14	12,470	8.9
Diabetes	8	554	9.7

Table 3.8. Misclassification rate for real data (Means (Std. deviations))

Dataset	Model	Original	Down	Up
Bank	LOGREG	0.0297 (0.0017)	0.0537 (0.0037)	0.0497 (0.0028)
	SVM	0.0320 (0.0015)	0.0600 (0.0033)	0.0402 (0.0024)
	RF	0.0316 (0.0011)	0.0644 (0.0035)	0.0400 (0.0027)
Adult	LOGREG	0.0223 (0.0008)	0.0646 (0.0029)	0.0598 (0.0020)
	SVM	0.0217 (0.0005)	0.0688 (0.0057)	0.0505 (0.0027)
	RF	0.0210 (0.0003)	0.0720 (0.0057)	0.0601 (0.0041)
Diabetes	LOGREG	0.0317 (0.0040)	0.0759 (0.0105)	0.0691 (0.0094)
	SVM	0.0319 (0.0031)	0.0856 (0.0128)	0.0682 (0.0095)
	RF	0.0307 (0.0007)	0.0891 (0.0125)	0.0473 (0.0071)

먼저 데이터를 7 : 3으로 나누어 전체의 70%에서 Logistic Regression, SVM, Random Forest를 가지고 모형을 만든 후 나머지 30%의 데이터에서 위에서 만든 모형이 얼마나 잘 맞는지 확인하였다.

3.2.2. Sampling 기법을 사용한 결과

오분류율만을 확인하였을 때는 Original 데이터가 가장 적은 값을 보이며 분류를 잘 한 것처럼 보이지만, 단순히 작은 집단에 속한 대부분의 관측치들을 큰 집단으로 예측함으로써 전체적인 오분류율이 낮아졌을 가능성이 크기 때문에 실제로 어떻게 분류되었는지 살펴보기로 하였다.

Table 3.9는 은행 데이터에서의 결과값이므로 이와 관련하여 결과를 해석해보자. 반응변수의 값이 0인 것이 해당 고객이 정기예금 가입을 하지 않은 경우이고 1인 것이 가입을 한 경우이다. Original 데이터는 가입을 했는데(1) 안했다고(0) 예측하는 값이 많고, 가입을 안했는데 했다고 예측하는 값이 적다. 하지만 sampling 기법을 적용한 경우에는 그 반대의 결과를 보여준다. 두 경우 중 은행의 입장에서 더 필요한 모형은 무엇일까?

Original 데이터를 이용하여 적합한 모형을 A, sampling 데이터를 이용한 모형을 B라고 하자. 은행은 새로운 고객 정보가 들어오면 A 또는 B 모형을 이용하여 새로운 고객들의 정기예금 가입여부를 예측하고자 할 것이다. 고객들에게 가입을 권유하는 광고를 할 때에 가입할 것으로 예측되는 고객들을 광고 대상으로 선정하는 것이 좋기 때문이다. 이 때 A 모형은 가입할 가능성이 있는 고객들의 대부분을 가입하지 않을 것이라고 예측하므로, 은행의 목적에 부합하지 않게 된다. 따라서 B 모형이 A 모형보다 은행에게 더 필요한 모형이라고 할 수 있다. 즉 모형을 적합할 때 데이터의 불균형을 보정하는 sampling 방법을 적용한 데이터를 사용하는 것이 좋다. 다른 방법론과 다른 데이터에서도 이와 비슷한 패턴을 보이기에 이곳에서는 생략하기로 한다. 추가적으로 각 방법론에서 적합된 모형을 비교해보기 위해 모형 평가 지표들과 ROC 곡선을 함께 살펴보았다. 단, ROC 곡선은 1회 적합한 결과를 사용하여 그렸으며 AUC를 비롯한 평가 지표들은 1000회 반복한 평균값을 사용하였다.

Figure 3.2와 Table 3.10의 마지막 열을 보면, 각 방법론에서 도출된 ROC 곡선과 그 성능을 나타내는 통계량인 AUC가 거의 동일하였다. 이는 모의실험에서의 결과와 마찬가지로 ROC 곡선과 AUC는

Table 3.9. Confusion matrix of LOGREG model for Bank dataset (average value)

		Predict					
		Original		Down		Up	
		0	1	0	1	0	1
True	0	1170.78	29.22	990.83	109.17	1011.05	188.95
	1	105.22	51.78	33.70	123.30	35.75	121.25

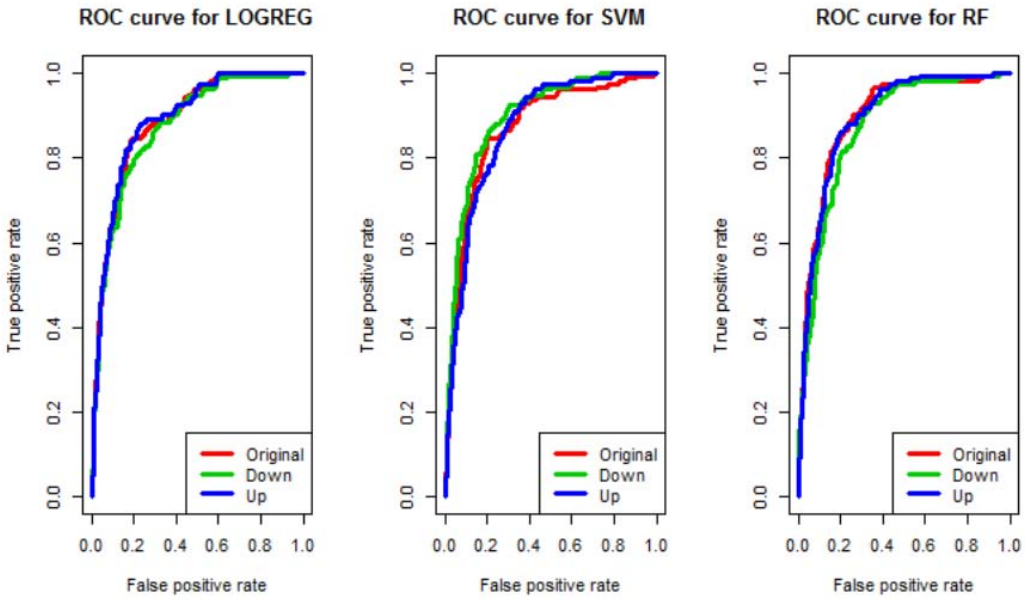

Figure 3.2. ROC curves of three machine learning models for Bank dataset.

모든 절단값에서의 전체적인 성능을 보는 지표이기 때문인 것으로 보인다. 그리고 Table 3.10의 Accuracy와 G-mean 열을 보면, Accuracy의 경우에는 sampling 기법을 적용했을 때의 값이 original data를 사용했을 때의 값보다 작거나 비슷했는데 G-mean 값은 sampling 기법을 적용한 결과값이 그렇지 않은 경우보다 월등히 높았다. 이는 데이터의 불균형을 보정해주는 기법을 사용했을 때의 예측력이 더 높다는 의미이다.

여기서의 결과값들은 Bank 데이터 셋에서 구한 값이지만, Adult와 Diabetes 데이터 셋에서의 결론도 위와 동일하였다. Down 또는 up-sampling 기법을 사용한 모형의 G-mean 값이 그렇지 않은 경우보다 약 2배가량 높게 나타나 3가지 데이터에서 공통적으로 sampling 기법을 적용한 모형의 예측력이 더 좋다는 결론을 얻을 수 있었다.

3.2.3. 비대칭 손실을 적용한 결과 본 절에서는 모의실험에서와 마찬가지로 어떤 방법론에서 total loss가 최소화되는지 살펴볼 것이다. 이때 total loss의 계산은 모의실험의 계산식을 이용하였다. 3가지 데이터 셋에서 공통적으로 관찰되는 것은, 비대칭 손실을 적용했을 때의 total loss 값과 그 감소율이 LOGREG 모형에서 가장 좋다는 사실이다. total loss의 감소율이 가장 적은 모형은 Bank와 Diabetes 데이터에서는 SVM으로, Adult 데이터에서는 RF 모형으로 나타났다. 하지만 그 차이는 크지 않았고,

Table 3.10. Performance metrics for Bank dataset (Means (Std. deviations))

Model	Data	Accuracy	G-mean	AUC
LOGREG	original	0.9726 (0.0017)	0.5663 (0.0319)	0.8878 (0.0112)
	down	0.9463 (0.0037)	0.8050 (0.0155)	0.8840 (0.0109)
	up	0.9504 (0.0028)	0.8064 (0.0172)	0.8910 (0.0102)
SVM	original	0.9680 (0.0015)	0.5100 (0.0319)	0.8694 (0.0125)
	down	0.9400 (0.0033)	0.8098 (0.0143)	0.8821 (0.0113)
	up	0.9599 (0.0024)	0.7075 (0.0219)	0.8749 (0.0108)
RF	original	0.9685 (0.0011)	0.4005 (0.0342)	0.9081 (0.0080)
	down	0.9356 (0.0035)	0.8074 (0.0064)	0.8911 (0.0069)
	up	0.9600 (0.0027)	0.7933 (0.0175)	0.9055 (0.0085)

Table 3.11. Total loss for real data (Means (Std. deviations))

Dataset	Model	Symmetric Loss	Asymmetric Loss	%Decrease
Bank	LOGREG	964.62 (50.26)	529.25 (45.19)	45.13%
	SVM	999.12 (37.62)	682.39 (45.95)	31.70%
	RF	1175.94 (38.23)	697.70 (58.76)	40.67%
Adult	LOGREG	2169.66 (65.42)	1193.38 (64.87)	45.00%
	SVM	2277.63 (41.29)	1293.63 (52.75)	43.20%
	RF	2258.38 (19.57)	1432.89 (93.55)	36.55%
Diabetes	LOGREG	133.20 (11.62)	80.54 (16.86)	39.53%
	SVM	146.59 (6.92)	105.43 (15.73)	28.08%
	RF	153.12 (2.21)	104.78 (18.85)	31.57%

대체로 total loss의 감소율이 3∼40% 이상이었다. 결론적으로 데이터와 방법론의 종류에 관계없이 비대칭 손실을 적용함으로써 total loss를 효과적으로 줄일 수 있었다.

4. 결론

본 논문에서는 불균형 데이터의 분류분석에 많이 이용되는 sampling 기법들의 성능을 각종 평가 지표들을 이용하여 비교하였다. 분석에 사용된 데이터는 이변량 정규분포를 따르는 난수를 생성하여 만든 모의실험 데이터와 은행의 정기예금 가입 여부, 미국 성인 근로자의 수입, 그리고 당뇨병 발병 여부에 관한 실제 데이터 3가지로, 모두 반응변수 가 2개의 범주를 가지므로 2집단 분류문제에 해당한다. 각 데이터에 대해 logistic regression (LOGREG), support vector machine(SVM), 그리고 random forest(RF) 방법론을 이용하여 모형을 적합했으며 모형 평가 지표로는 오분류율, 정분류율, G-mean, ROC 곡선, AUC, 그리고 total loss를 이용하였다.

모의실험 데이터와 실제 데이터에서 공통적으로 관찰되는 사실은 다음과 같다. 첫째, down 또는 up-sampling을 적용한 경우의 모형 성능이 더 좋았다. 이 때 모형의 성능이 좋다는 것은 여러 평가지표로 설명될 수 있는데, 대표적으로 본 연구에서 사용한 평가지표 중 민감도와 특이도의 기하평균인 G-mean 값을 예로 들 수 있다. Sampling 기법을 적용한 모형의 G-mean 값이 그렇지 않은 모형에 비해 월등히 높은 값을 보였는데, 특히 실제 데이터에서 그 차이가 컸다. 또 다른 평가 지표인 오분류율은 실제로 어떻게 분류되었는지와 관계없이 단순히 오분류된 자료의 비율을 보는 지표이기 때문에 불균형 자료에서의 모형 평가 지표로 활용하기에 한계가 있다. ROC 곡선이나 AUC의 경우 각 모형에서 큰 차이를 보이지는 않았는데, 이는 이 평가 지표들이 특정 절단값에서의 성능을 평가하는 것이 아니라 decision

boundary를 변화시키면서 전체적인 성능을 평가하기 때문인 것으로 보인다. 둘째, 비대칭 손실을 가정한 경우의 total loss 값이 그렇지 않은 경우보다 훨씬 작았고, 그 감소율은 자료의 불균형이 심해질수록 컸으며 모의실험 데이터에서는 SVM 모형에서, 실제 데이터에서는 LOGREG 모형에서 감소율이 가장 컸다. 셋째, SVM과 RF 모형이 LOGREG 모형에 비해 자료의 불균형에 더 민감하였다. 즉 sampling 기법이나 비대칭 손실을 적용하지 않고 원 자료를 일반적인 방법론으로 분류한 경우 대부분의 자료가 큰 집단으로 예측되는 쏠림 현상이 더 심했다. 그렇기 때문에 자료의 불균형이 심할수록 그 불균형을 보정해주는 방법을 사용했을 때 G-mean의 증가율이나 total loss의 감소율이 비교적 더 컸다.

우리는 본 논문을 통해서 불균형 자료의 분류분석에 있어 sampling 기법이나 비대칭 손실을 적용하는 것이 모형의 성능을 효과적으로 개선시킬 수 있음을 보였다. 많은 실제 분류문제에서 이런 unbalanced data가 나오는 경우가 많은데 sampling 기법이나 비대칭 손실을 이용하면 그렇지 않았을 때보다 total loss를 줄일 수 있고 따라서 더 좋은 결과를 가져올 것이라고 예상한다.

References

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

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

Chen, C., Liaw, A. and Breiman, L. (2004). Using random forest to learn imbalanced data, Technical Report 666.

Karatzoglou, A., Meyer, D. and Hornik, K. (2006). Support vector machines in R, *Journal of Statistical Software*, 15.

Kubat, M., Holte, R. and Matwin, S. (1997). Learning when negative examples abound. In *Proceedings of ECML-97, 9th European Conference on Machine Learning*, 146–153.

Kubat, M. and Matwin, S. (1997). Addressing the curse of imbalanced training sets: One-sided selection, *Proceedings of the 14th International Conference on Machine Learning*, 179–186.

Park, C., Kim, Y., Kim, J., Song, J. and Choi, H. (2011). *Datamining using R*, Kyowoo, Seoul.

R Development Core Team (2010). R: A Language and Environment for Statistical Computing, *R Foundation for Statistical Computing*, Vienna, Austria, ISBN 3-900051-07-0. http://www.R-project.org

Vapnik, V. (1998). *Statistical Learning Theory*, Wiley, New York.

Wu, G. and Chang, E. (2003). Class-boundary alignment for imbalanced dataset learning, *In ICML 2003 Workshop on Learning from Imbalanced Data Sets II*, Washington, DC.

불균형 자료에 대한 분류분석

김동아[a] · 강수연[a] · 송종우[a,1]

[a]이화여자대학교 통계학과

(2015년 3월 20일 접수, 2015년 4월 27일 수정, 2015년 4월 28일 채택)

요 약

일반적인 2집단 분류(2-class classification)의 경우, 두 집단의 비율이 크게 차이나지 않는 경우가 많다. 본 논문에서는 두 집단의 비율이 크게 차이나는 불균형 데이터(unbalanced data)의 분류 문제에 대해서 다루고자 한다. 불균형 데이터의 분류방법은 균형이 맞는 데이터(balanced data)의 경우보다 분류하기 어려운 경우가 많다. 이런 자료에서 보통의 분류모형을 적용하게 되면 많은 경우에 대부분의 관측치가 큰 집단으로 분류 되는 경우가 많은데 실질적인 어플리케이션에서는 이런 오분류가 손해가 더 큰 경우가 대부분이다. 우리는 sampling 기법을 이용하여 다양한 분류 방법론의 성능을 비교 분석 하였다. 또한 비대칭 손실(asymmetric loss)을 가정한 경우에 어떤 방법론이 가장 작은 loss를 생성하는 지를 비교하였다. 성능 비교를 위해서는 오분류율(misclassification rate), G-mean, ROC, 그리고 AUC(Area under the curve) 등을 이용하였다.

주요용어: up-sampling, down-sampling, 비대칭 손실, 오분류율, G-mean, ROC, AUC

이 논문은 2013년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. NRF-2013R1A1A2012817).

[1]교신저자: (120-750) 서울특별시 서대문구 이화여대길 52, 이화여자대학교 통계학과.
E-mail: josong@ewha.ac.kr