석 사 학 위 논 문

불균형 자료에서 랜덤 포레스트에 기반한 분류 방법의 성능 비교

Performance Comparison of Classification methods based on the Random Forest in Class Imbalanced Data

고려대학교 대학원

의학통계학 협동과정

문 선 영

2018년 2월 일

안 형 진 교수지도
석 사 학 위 논 문

불균형 자료에서 랜덤 포레스트에 기반한 분류 방법의 성능 비교

Performance Comparison of Classification methods based on the Random Forest in Class Imbalanced Data

이 논문을 의학통계학 석사학위 논문으로 제출함.

2018년 2월 일

고려대학교 대학원
의학통계학 협동과정
문 선 영

문선영의 의학통계학 석사학위논문 심사를 완료함.

2018 년 2 월 일

위 원 장 안 형 진 ㊞

위 원 이 준 영 ㊞

위 원 지 희 정 ㊞

Abstract

Performance Comparison of Classification methods based on the Random Forest in Class Imbalanced Data

Sun Young Moon

Department of Biostatistics

Graduate School of Korea University

(Supervising Professor : Hyonggin An, Ph.D.)

Objectives : Because of the very rare occurrence of events of interest in many fields, the class of response variables shows a highly imbalanced distribution (minority class, majority class). However, general classification algorithms assume a balanced distribution of classes and aim at minimizing the overall misclassification rate. Thus, the use of this general classification algorithms for class imbalance data leads to a very poor classification accuracy for minority class that we

are interested. Several methods have been suggested to overcome this class imbalanced problem. The purpose of this study is to compare the classification performance of various methods that can correct the class imbalanced problem according to the degree of class imbalance(IR).

Methods : In this study, random forest is used as the basic classification algorithm and random forest based algorithms are used among the various methods to solve the class imbalanced problem. A random forest was applied to modified sample through ROS, RUS, and SMOTE, which are data-level approaches that adjust the distribution of the sample. Also the algorithm-level approaches such as Weighted Random Forest, Balanced Random Forest, Isolation Forest were used. Additionally, classification of class imbalance data can not use 'accuracy', which is a general model performance evaluation index, for comparing performance of models because it is interested in more accurate classification of minority class than majority class. Therefore, in this study, F1 measure and AU-ROC were used as performance evaluation index.

Results : Simulation results show that the Isolation Forest method, which is an anomaly detection algorithm based on spatial division, shows the best performance for both the F1 measure and the AU-ROC index in all IRs. In addition, as the number of samples increases, both indicators show a slight increase, indicating that the number of samples in the classification of class imbalance also affects the size of samples.

Also, overall trends show that there was a difference in the decrease of each method, but as the degree of imbalance increased, the values of both indicators decreased in all methods.

Conclusion : If the distribution of response variables with a value of 0 or 1 is highly imbalanced, proportion of correct classification of minority classes in the general classification algorithm is somewhat lower. However, there are many factors that affect the class imbalanced problem such as within imbalance or sample size, so it is important to select the appropriate method for the structure of the data.

Key words : Imbalanced Data, IR, Random Forest, ROS, RUS, SMOTE, Weighted Random Forest, Balanced Random Forest, Isolation Forest

목 차

Abstract ····· i

Ⅰ. 서론 ····· 1

Ⅱ. 본론 ····· 5

1. 분류 알고리즘 ····· 5

2. 계급 불균형 자료의 분류 성능 향상을 위한 방법 ····· 9

3. 모형 성능 평가 기준 ····· 20

Ⅲ. 모의실험 ····· 25

1. 모의실험 설계 ····· 25

2. 모의실험 결과 ····· 27

Ⅳ. 실제자료 분석 ····· 34

1. 분석 자료 ····· 34

2. 분석 결과 ····· 35

Ⅴ. 고찰 ····· 39

Ⅵ. 결론 ····· 42

참고문헌 ····· 43

국문요약

List of Tables

Table 1. Weight by Class in WRF method ····· 14

Table 2. Confusion Matrix ····· 20

Table 3. Results of simulation study for various methods according to IR when N=1,000 & 2,000 (F1 measure) ·· 32

Table 4. Results of simulation study for various methods according to IR when N=1,000 & 2,000 (AU-ROC) ·· 33

Table 5. Summary of Actual data ····· 35

Table 6. Results of actual data analysis (F1-measure) ····· 37

Table 7. Results of actual data analysis (AU-ROC) ····· 38

List of Figures

Figure 1. Random Forest ····· 6

Figure 2. Illustration of ROS & RUS Algorithm ····· 10

Figure 3. Illustration of SMOTE Algorithm ····· 11

Figure 4. The random partitioning of a normal point versus an anomaly ····· 16

Figure 5. Example of ROC Curve ····· 24

Figure 6. Results of simulation study N=1,000 (F1 measure) ···· 28

Figure 7. Results of simulation study N=2,000 (F1 measure) ···· 28

Figure 8. Results of simulation study N=1,000 (AU-ROC) ····· 30

Figure 9. Results of simulation study N=2,000 (AU-ROC) ····· 30

Ⅰ. 서론

다양한 분야의 자료 분석에서 주어진 정보를 이용하여 개체를 분류한다. 분류는 반응 변수가 C개의 계급(class) 중 하나의 값을 가질 때 주어진 설명 변수에 대해 반응 변수가 어느 계급에 속할지를 예측하는 것이다. 이를 위해 의사결정 나무(Decision Tree), 베이지안 분류기(naive Bayesian Classifier), 서포트 벡터 머신(Support Vector Machine) 등 다양한 분류 알고리즘이 개발되어 사용되고 있다. 이러한 일반적인 분류 알고리즘은 계급의 균형 분포(balanced class distributions)를 가정하고 전체적인 오분류율을 최소화하는 것을 목적으로 한다. 따라서 계급의 비율이 비슷할 때 이러한 분류 알고리즘이 잘 작동한다고 알려져 있다(He, H., & Garcia, E. A., 2009).

하지만 많은 경우 계급의 균형 분포를 가정할 수 없다. 의학 분야에서 희귀 질환 진단 자료의 경우, 병원에 방문하는 다수의 사람 중에서 희귀 질환을 가진 환자는 매우 드물게 발생한다. 금융 분야에서 다수의 카드 사용 건 중 소수의 부정 사용, 품질 관리 분야에서 다수의 상품 중 소수의 불량, 그리고 네트워크 분석 분야에서 다수의 네트워크 트래픽 중 소수의 해킹 등 다양한 분야에서도 마찬가지이다. 이처럼 자료를 구성하는 하나 이상의 계급이 다른 계급과 비교하여 매우 낮은 빈도를 갖는 경우를 계급 불균형(class imbalance)이라고 한다. 본 논문에서는 반응 변수가 0과 1로 이루어진 이진 분류 문제에서의 계급 불균형을 다룰 것이다. 즉, 반응 변수 1(소수계급)의 빈도가 반응 변수 0(다수계급)에 비해 매우 낮은 경우이다. 이와 같은 계급 불균형 자료에서는 소수계급을 잘못 분류할 때 큰 비용이

발생하기 때문에 주로 소수계급을 잘 분류하는 것에 관심이 있다. 예를 들어 의학 분야에서 암 환자를 암이 아니라고 진단하는 것은 암이 아닌 환자를 암으로 진단하는 것보다 훨씬 큰 비용을 가져온다. 만약 이러한 계급 불균형 자료에 일반적인 분류 알고리즘을 적용하게 되면 소수계급을 제대로 분류하지 못하는 문제가 발생한다. 이를 계급 불균형 문제(class imbalanced problem)라고 한다.

계급 불균형 자료 문제는 계급의 불균형 정도에 영향을 받는다. 계급의 불균형 정도는 다음과 같이 정의된 IR(imbalance ratio)로 나타낸다.

$$IR = \frac{n^+}{n^-}$$

n^+: the number of instances in the majority class

n^-: the number of instances in the minority class

이는 단순히 표본에서 다수계급의 빈도가 소수계급의 빈도에 비해 얼마나 큰지를 나타낸다. 따라서 IR 값은 1보다 큰 값을 가지며, 커질수록 계급 불균형 정도가 심해진다. IR이 9 이상이면 소수계급이 전체 자료의 10% 이하인 매우 불균형한 자료를 의미한다. 하지만 분류 성능을 저해하는 불균형 정도는 아직 명확하게 알려져 있지 않다(Sun et al., 2009).

계급의 불균형 정도 외에도 다른 일반적인 분류 문제와 마찬가지로 적은 표본의 수(small sample size)와 계급 간 겹침(class overlapping) 역시 계급 불균형 자료 문제에 영향을 미친다. 표본의 수가 적어지면 특히나 소수계급의 특징에 대한 정보가 부족해지게 된다. 따라서 모형이 소수계급과

다수계급을 분류할 적절한 규칙을 발견하기 힘들어진다. 만약 훈련 자료의 표본 수가 증가하면 계급 불균형 자료의 분류에서 오분류율이 감소한다 (Japkowicz & Stephen, 2002). 또한 계급 간 겹침 정도가 커질수록 소수계급과 다수계급을 분리할 경계(boundary)를 찾기 힘들어진다. 이때 일반적인 분류 알고리즘은 전반적인 정확성을 최대화하기 위해 겹친 구간(overlapping region)을 다수계급으로 분류하고, 소수계급을 잡음(noise)으로 인식해버린다. 추가적으로 한 계급이 여러 개의 작은 클러스터들로 구성되는 경우와 같은 단일 계급 내의 불균형(within class imbalance) 역시 계급 불균형 문제에 영향을 준다.

이와 같은 계급 불균형 자료에서 소수계급의 분류 정도를 높이기 위해 다양한 방법이 연구되고 있다. 방법은 크게 자료 수준의 접근 방법과 알고리즘 수준의 접근방법으로 나눌 수 있다. 첫 번째로 자료 수준의 접근방법(data-level approaches)은 표본추출(sampling) 방법을 통해 두 계급의 불균형한 분포를 수정하는 방법이다. 일종의 전처리(pre-process) 방법으로 분류 알고리즘과 독립적으로 사용할 수 있고 적용이 쉽다. 두 번째로 알고리즘 수준의 접근방법(algorithm-level approaches)은 소수계급과 관련한 학습을 강화하도록 분류 알고리즘을 수정하는 방법이다.

본 논문에서는 계급 불균형 자료의 분류에서 랜덤 포레스트에 기반한 분류 성능을 향상시키기 위한 방법을 고찰하고, 모의실험을 통해 불균형 정도에 따른 각 방법의 분류 성능을 비교하고자 한다. 본 논문의 구성은 다음과 같다. 본론의 1장에서는 기본 분류 알고리즘으로 사용한 랜덤 포레스트에 대해 설명한다. 2장에서는 랜덤 포레스트에 기반한 계급 불균형 자료의 분류 성능 향상을 위한 방법론을 소개하며, 3장에서는 계급 불균형

자료에 맞는 모형 평가 측도에 대해 소개한다. 그 후 모의실험과 실제 자료 분석을 통해 각 방법의 성능을 비교한다. 이를 통해 결론에서 각 방법의 특징을 알아보고, 상황에 따른 적절한 방법을 제시한다.

Ⅱ. 본론

1. 분류 알고리즘

기계학습에서 분류란 관측된 값의 설명 변수와 반응 변수를 사용하여 모형을 훈련시키고, 이 훈련된 모형을 이용하여 새로운 관측 값이 주어졌을 때 어느 쪽으로 분류되는지 예측하는 것을 말한다. 본 논문에서는 다수의 의사결정나무를 임의적으로 학습하는 랜덤 포레스트 알고리즘을 기본 분류 알고리즘으로 사용하였다.

Random Forest(2001)

랜덤 포레스트는 Leo Breiman(2001)이 제안한 분류 알고리즘의 하나로, 여러 개의 의사결정나무를 임의적으로 학습하는 방식의 앙상블(Ensemble) 학습 방법을 사용한 모형이다. 앙상블 학습 방법은 주어진 자료로부터 여러 개의 모형을 학습한 후 여러 모형의 예측 결과들을 종합하여, 예측 시 정확도를 높이는 방법이다. 일반적으로 하나의 의사결정나무를 이용하면, 훈련 자료에 따라 생성되는 의사결정나무가 매우 달라져 분류 성능의 변동 폭이 크다는 단점이 있다. 랜덤 포레스트는 이러한 의사결정나무의 특징을 고려한 방법이다.

랜덤 포레스트 알고리즘은 크게 학습(training) 단계와 검정(test) 단계로 이루어진다. 학습 단계에서는 먼저 기존의 훈련 자료에서 부트스트랩

표본을 구성한다. 그 후 p개의 설명 변수 중 m개를 무작위로 선택하여 가지치기 없이 CART(Classification and Regression Trees)(Brieman et al., 1984) 알고리즘을 이용해서 의사결정나무를 생성한다. 이러한 과정을 B번 반복하게 된다. 검정 단계에서는 학습 단계에서 사용되지 않은 새로운 표본을 B개의 의사결정나무에 적용한다. 각 의사결정나무의 끝마디(terminal node)는 입력된 새로운 표본에 대해 출력 값을 갖는 분류기(classifier)를 갖게 된다. 최종적으로 랜덤 포레스트는 이러한 각 의사결정나무 분류기를 단순 다수결을 통해 하나의 분류기로 만들어 낸다. 이상의 과정은 아래의 <Figure 1>과 같이 표현할 수 있고 알고리즘을 정리하면 다음과 같다.

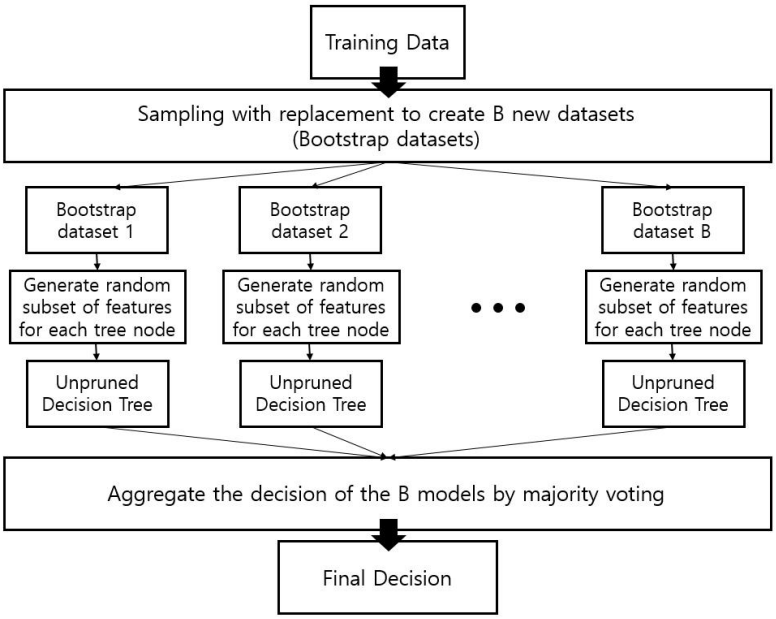

Figure 1. Random Forest (Dittman et al., 2015)

Random Forest Algorithm

step 1) b = 1, ... , B에 대해:

A. 훈련 자료에서 크기가 N인 부트스트랩 표본 L^*를 생성한다.

B. 부트스트랩 표본 L^*에서 의사결정나무의 각 끝마디(terminal node)가 최소 마디 크기인 $n_{\min}$에 도달할 때까지 아래의 i, ii, iii 단계를 반복하여 랜덤 포레스트 나무 $C_b(x)$를 생성한다.

i) p개의 설명 변수 중 m개를 무작위로 선택한다.
(m= $\sqrt{p}$ 또는 $p/3$)

ii) m개의 설명 변수 중 가장 좋은 변수/분할점(split-point)을 선택한다.

iii) 마디를 두 개의 자식 마디(daughter nodes)로 분할한다.

step 2) B개의 랜덤 포레스트 나무를 단순 다수결을 통해 결합하여 최종 학습기 $C^*(x)$를 만든다.

$$C^*(x) = \text{majority vote } \{C_b(x)\}_1^B$$

랜덤 포레스트의 가장 큰 특징은 무작위성(randomness)에 의해 서로 조금씩 다른 특성을 갖는 의사결정나무들로 구성된다는 점이다. 무작위성을 최대로 주기 위해 의사결정나무의 생성과정에서 부트스트랩 표본과 더불어 설명 변수들에 대한 무작위 추출을 결합한다(박창이, 김진석, 2008). 결과적으로 각 의사결정나무들이 서로 비상관화(decorrelation) 되게 하여 일반화 성능을 향상시킨다. 또한 부트스트랩 과정은 의사결정나무의 편향

은 그대로 유지하면서, 분산은 감소시키기 때문에 랜덤 포레스트의 성능을 향상시킨다.

이와 같이 랜덤 포레스트는 여러 개의 의사결정나무를 사용해 한 개의 의사결정나무를 사용하는 경우보다 정확성과 안정성이 더 좋다. 이 때 랜덤 포레스트를 구성하는 의사결정나무의 수는 랜덤 포레스트의 분류 성능을 결정하는 주요 모수이다. 이 크기가 작으면 의사결정나무를 구성하고 검정하는데 걸리는 시간이 짧지만, 일반화 능력이 떨어질 수 있다. 또한 하나의 의사결정나무에서 뿌리마디(root node)부터 끝마디까지의 깊이(최소 마디 크기) 역시 랜덤 포레스트의 주요 모수이다. 이 깊이가 너무 크면 과대적합이 발생하고 너무 작으면 과소적합이 발생하기 때문에 적절한 값을 설정하는 것이 중요하다.

랜덤 포레스트의 단점으로는 분류 속도가 느리고 중간 과정을 알 수 없어 의사결정나무가 갖는 장점 중 하나인 설명력을 잃게 된다는 점이 있다. 또한 랜덤 포레스트 역시 다른 분류 알고리즘처럼 두 계급의 동일한 분포를 가정하고 오분류율을 최소화하기 때문에 계급 불균형 자료에서 분류 성능에 문제를 갖는다. 특히나 극도로 불균형한 자료에서 부트스트랩 표본을 구성할 때, 부트스트랩 표본에 포함된 소수계급 표본이 매우 적거나 아예 추출되지 않을 수도 있다.

2. 계급 불균형 자료의 분류 성능 향상을 위한 방법

계급 불균형 자료의 분류 문제를 해결하기 위한 접근법으로 크게 자료 수준의 접근 방법과 알고리즘 수준의 접근 방법이 있다.

2.1 자료 수준의 접근 방법

자료 수준의 접근 방법(data-level approach)이란 모형을 만드는 훈련 자료의 표본을 조절하는 방법이다. 즉, 표본 추출(sampling)을 통해 직접 훈련 자료 내 표본의 분포의 균형을 맞춰 소수계급의 분류 정도를 높이는 방법이다. 자료의 분류 전에 표본을 조절하는 방법이기 때문에 표본의 분포를 조정한 후 일반적으로 사용하는 분류 알고리즘을 적용하기 쉬워 현실적으로 많이 쓰이는 방법이다. 여기에는 크게 소수계급의 표본 수를 늘리는 오버샘플링과 다수계급의 표본 수를 줄이는 다운샘플링(혹은 언더샘플링) 방법이 있다. 최적의 분류 성능을 보이는 불균형 정도는 알려져 있지 않기 때문에 효과적인 오버샘플링 혹은 다운샘플링 비율은 데이터마다 다르다(Sun et al., 2009). 조정된 두 계급의 비율이 50 : 50인 경우가 항상 최적의 분류 성능을 보이지는 않지만, 대체적으로 50 : 50의 균형 잡힌 분포에서 분류 성능이 좋았다(Ali et al., 2015).

본 논문에서는 자료 수준의 접근 방법으로 간단한 알고리즘을 가진 ROS와 RUS 및 SMOTE 방법을 살펴본다.

2.1.1 ROS(Random Over Sampling) & RUS(Random Under Sampling)

랜덤오버샘플링 방법은 <Figure 2. (b)>와 같이 소수계급의 표본 수를 늘리기 위해 소수계급의 표본을 무작위로 선택하여 반복 추출하는 방법이다. 소수계급의 표본 수가 증가하는 만큼 자료의 크기는 커지지만, 단순히 소수계급의 표본을 반복하는 것이기 때문에 정보의 양이 늘어나는 것은 아니다. 따라서 정보의 손실은 없으나 같은 표본의 중복으로 인해 과대적합(over fitting) 문제가 발생할 수 있고, 자료의 크기가 커지므로 분류 알고리즘을 적용할 때 학습시간이 증가한다.

랜덤언더샘플링은 <Figure 2. (c)>와 같이 소수계급의 표본은 모두 사용하지만, 다수계급의 표본은 무작위로 선택된 일부만 추출하는 방법이다. 따라서 다수계급의 표본 수가 감소하는 만큼 자료의 크기는 작아지고, 그만큼의 정보 손실이 발생한다. 하지만 랜덤오버샘플링 방법에 비해 자료의 크기가 작은 만큼 크기가 큰 데이터를 다룰 때 학습 과정에 소요되는 시간을 단축할 수 있다.

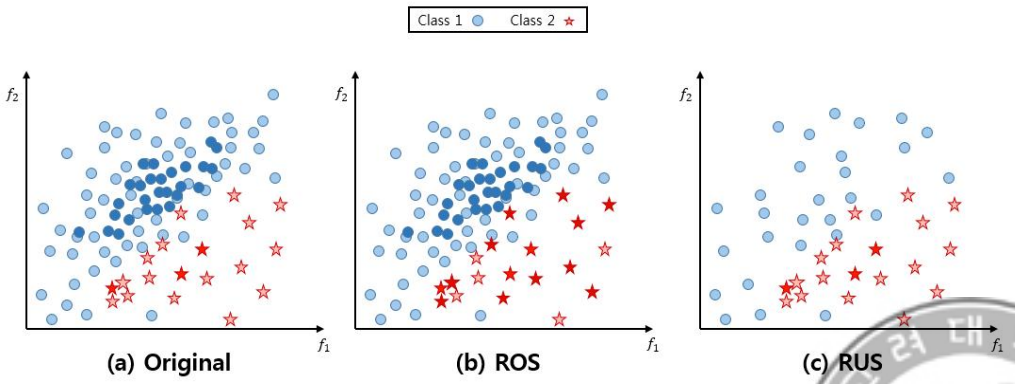

Figure 2. Illustration of ROS & RUS Algorithm

2.1.2 SMOTE(Synthetic Minority Oversampling Technique)(2002)

SMOTE는 Chawla et el.(2002)이 제안한 방법으로 <Figure 3>와 같이 소수계급에 속하는 새로운 표본을 만들어 두 계급의 균형을 맞춰주는 방법이다. <Figure 3>를 보면 소수계급에 속하는 표본 x_i에 대해 5개의 최근접 이웃(nearest neighbor)을 구한다. 그 후 임의의 표본 $\hat{x}_i$을 선택하여 두 표본 사이에 새로운 표본 x_{new}를 생성한다. 즉, 랜덤오버샘플링에서 같은 표본의 중복으로 인해 생기는 과대적합 문제를 보완하기 위해 기존의 표본을 약간씩 이동시킨 점들을 훈련 자료에 추가하는 방법이다. SMOTE 방법의 알고리즘은 다음과 같다(He, H., & Garcia, E. A., 2009).

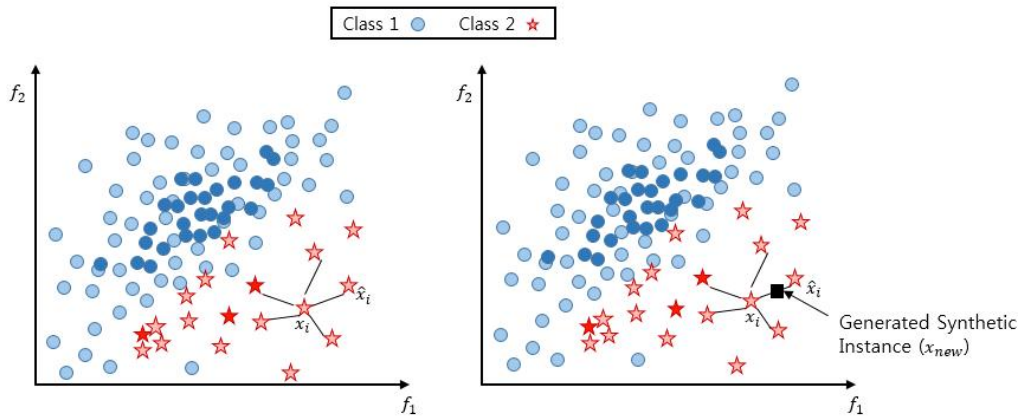

Figure 3. Illustration of SMOTE Algorithm (He, H., & Garcia, E. A., 2009)

SMOTE Algorithm	
step 1)	소수계급에 속하는 표본 x_i에 대해 K-최근접 이웃(K-nearest neighbors)을 계산한다.
step 2)	step 1에서 구한 K-최근접 이웃 중 무작위로 한 이웃($\hat{x_i}$)을 선택하여 표본 x_i와의 유클리디안 거리(Euclidian distance)를 계산한다.
step 3)	step 2에서 구한 유클리디안 거리에 [0, 1] 사이의 임의의 값을 곱하여 원래의 표본에 더하여 새로운 표본 x_{new}를 만든다. $$x_{new} = x_i + (\hat{x_i} - x_i) \times \delta \ , \quad \delta \in [0,1]$$
step 4)	step 3에서 새로 만든 표본 x_{new}를 훈련 자료에 추가한다.

2.2 알고리즘 수준의 접근 방법

알고리즘 수준의 접근 방법(algorithm-level approach)은 소수계급의 학습을 개선하기 위해 기존 알고리즘을 조정하는 방법이다. 알고리즘 수준의 접근 방법 내에서도 One-class learning, Cost-sensitive learning, Improved algorithm, Hybrid approach, ...등으로 다시 범주를 나눌 수 있다. One-class learning은 두 계급의 차이를 학습하는 것이 아닌 소수계급 혹은 다수계급의 단일계급 내에서 학습하는 방법으로 특정 알고리즘에서만 가능하기 때문에 많이 쓰이지 않는 방법이다. Cost-sensitive learning은 반응 변수가 잘못 분류 되었을 때의 비용 크기를 계급별로 비대칭적으로 부

여하는 방법이다. 불균형자료에서의 주 관심사항은 소수계급을 잘 분류하는 것이므로, 소수계급을 다수계급으로 잘못 분류했을 때의 비용을 다수계급을 소수계급으로 잘못 분류했을 때보다 크게 부여한다. Improved algorithm은 기존 알고리즘을 수정하여 소수계급의 학습을 강화한 방법이다. 마지막으로 Hybrid approach는 다양한 방법을 융합시킨 방법이다.

본 논문에서는 Cost-sensitive learning 방법인 WRF 방법과 Improved algorithm 방법인 BRF 방법 및 Isolation Forest 방법을 살펴볼 것이다.

2.2.1 WRF(Weighted Random Forest)(2004)

랜덤 포레스트 분류 알고리즘 역시 두 계급의 동일한 분포를 가정하고 오분류율을 최소화하기 때문에 계급 불균형 자료에서 분류 성능에 문제를 갖는다는 점을 보완하기 위해 Chen et al.(2004)은 Weighted Random Forest(WRF) 방법을 제안하였다. WRF 방법은 각 계급에 가중치를 부여하는데, 소수계급에 더 큰 가중치를 부여함으로써 소수계급을 잘못 분류하는 것에 더 큰 비용을 준다. 즉, WRF 방법은 소수계급을 잘못 분류하는 것에 큰 비용을 부과하고, 전반적인 비용을 감소시키는 방법이다.

이렇게 부여한 각 계급 가중치는 랜덤 포레스트 알고리즘의 두 부분에 영향을 미친다.

(1) 지니 계수(Gini criterion): 의사결정나무를 생성하는 단계에서 마디 선택 기준인 지니 계수(Gini criterion)에 영향을 준다.

(2) 가중 다수결(weighted majority vote): 각 의사결정나무의 끝마디가 나타내는 결과를 통합(aggregate)할 때 단순 다수결 방법이 아닌 해당 계급의 가중치를 곱한 값을 비교하는 가중 다수결 방법을 통해 최종 모형이 결정된다.

WRF 방법에서 분류 성능 향상을 위해 적절한 계급의 가중치를 정하는 것이 중요하다. 하지만 일반적으로 각 계급의 오분류 비용은 알려져 있지 않기 때문에 본 논문에서는 <Table 1>과 같이 상대 집단의 크기에 비례하는 값으로 가중치를 부여하였다.

Table 1. Weight by Class in WRF method

	계급의 표본 크기 (Sample size)	가중치 (Weight)
소수계급 (Minority Class)	n1	n2/(n1+n2)
다수계급 (Majority Class)	n2	n1/(n1+n2)

2.2.2 BRF(Balanced Random Forest)(2004)

본론의 1절에서 살펴보았듯이 랜덤 포레스트는 훈련 자료의 부트스트랩 표본에서 다수의 의사결정나무를 만든다. 하지만 IR이 매우 큰 계급 불균형 자료에서 구성된 부트스트랩 표본 내에는 소수계급이 거의 없거나 또는 전혀 없을 확률이 높다. 따라서 이렇게 학습된 랜덤 포레스트 분류기에서는 소수계급의 예측력이 떨어지게 된다. 이러한 문제를 해결하기 위해

Chen et al.(2004)은 WRF 방법과 함께 Balanced Random Forest(BRF) 방법을 제안하였다.

BRF 방법은 훈련자료의 부트스트랩 표본의 계급 구성비를 50 : 50으로 만들어 주는 방법이다. 따라서 다수계급의 표본을 제거하는 것이 아니기 때문에 본론의 2.1.1절에서 설명한 랜덤언더샘플링의 정보 손실 단점을 보완해준다. BRF 방법의 알고리즘은 부트스트랩 표본의 구성 단계에서 소수계급과 다수계급을 나누어 표본을 추출한다는 약간의 차이가 있다는 점 외에는 랜덤 포레스트와 동일한 알고리즘을 갖는다.

BRF Algorithm	
step 1)	소수계급에서 부트스트랩 표본을 추출한다. 그 후 소수계급의 부트스트랩 표본 크기와 같은 크기로 다수계급에서 부트스트랩 표본을 구성한다.
step 2)	이후 과정은 랜덤 포레스트의 알고리즘과 동일

2.2.3 Isolation Forest(2008)

Isolation Forest는 Liu et al.(2008)이 제안한 공간분할 기반의 이상 탐지(anomaly detection) 방법이다. 본 논문에서는 소수계급을 이상치(anomaly)로 다수계급을 정상치(normal)로 가정하여 불균형 자료의 분류 문제에 적용하였다.

Isolation Forest는 기본적으로 랜덤 포레스트의 구조를 이용한 공간

분할 방법으로 이상치는 전체 자료에서 차지하는 수가 적을 것이고, 대부분의 표본들과는 다른 패턴을 가질 것(few and different)이라는 전제에서 출발한다. 자료 내의 한 표본을 선을 그어 다른 표본들과 독립된 공간에 위치시키게 되는데 <Figure 4>와 같이 군집 내부에 있는 정상치 x_i의 경우 공간 내에 한 점만 남기고 완전히 고립시키려면 많은 횟수의 공간 분할을 수행해야 하지만, 군집에서 멀리 떨어진 이상치 x_o는 적은 횟수의 공간 분할만으로도 해당 표본을 고립시킬 수 있다.

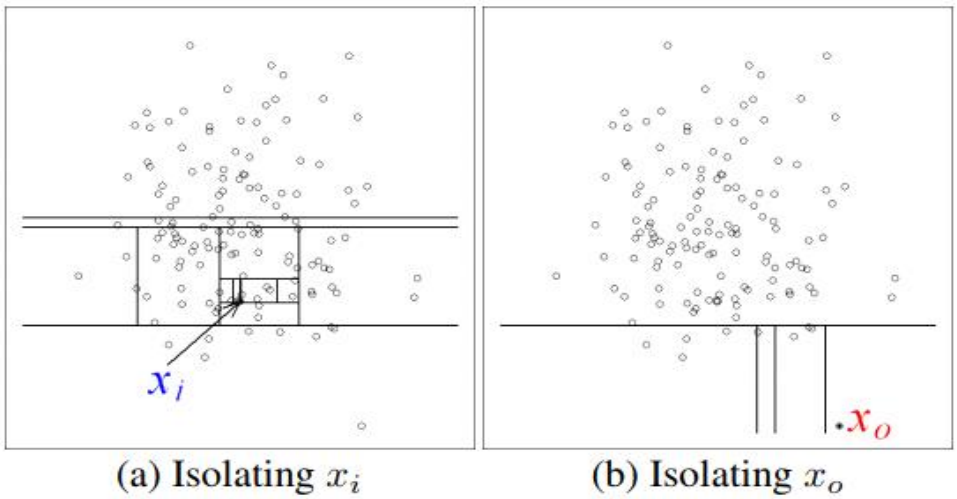

Figure 4. The random partitioning of a normal point versus an anomaly (Liu et al., 2008)

위의 그림에서 공간을 분할 할 때, 무작위로 선택된 설명 변수에 대해 해당 설명 변수의 최솟값과 최댓값 사이의 임의의 값을 이용하여 해당 표본을 다른 표본들로부터 분리(partitioning)하게 된다. 이와 같이 공간 분할을 설명 변수와 분할점(split-point)으로 표현할 수 있으므로, 여러 번의 공

간 분할은 의사결정나무 형태로 표현할 수 있다. 또한 표본의 분리에 요구되는 분리 횟수는 의사결정나무의 뿌리마디부터 끝마디까지의 깊이(마디 크기)와 동일하게 볼 수 있다. 이러한 관점에서 보면 정상치일수록 그 표본을 완전히 고립시키기 위해서는 의사결정나무를 깊숙하게 타고 내려가야 한다. 반대로 이상치의 경우, 의사결정나무의 상단부만 타더라도 해당 표본은 고립될 가능성이 높다. 이런 특성을 이용하면 의사결정나무를 몇 회 타고 내려가야 해당 표본이 고립되는가를 기준으로 정상치와 이상치를 분류할 수 있다.

이런 의사결정나무를 분리나무라고 칭하는데, 분리나무를 여러 개 모아서 앙상블 모델을 만들면 안정적인 이상 점수(anomaly score)를 산출 할 수 있다. Liu et al,(2008)에 따르면 약 50개에서 100개 정도의 의사결정나무를 이용하면 이상점수가 안정화된다.

이상의 Isolation Forest의 알고리즘을 정리하면 아래와 같다.

Isolation Forest Algorithm

step 1) t = 1, ... ,T에 대해:

A. 훈련 자료에서 비복원 표본 추출을 통해 크기가 ψ인 표본을 생성한다.

B. 생성된 자료에 대해 표본이 완전히 고립되거나 분리나무(iTree)의 높이가 $l = ceiling(\log_2 \psi)$에 도달 할 때까지 표본을 반복적으로 분리하여 분리나무(iTree)를 생성한다.

i) Q개의 설명 변수 중 q개를 무작위로 선택한다.

ii) q개의 설명 변수 중 무작위로 한 변수/분할점(설명 변수의 최솟값과 최댓값 사이의 임의의 값)을 선택한다.

iii) 마디를 두 개의 자식 마디로 분할한다.

step 2) T개의 분리나무를 통합하여 최종 학습기를 만들어 이상 점수 $s(x, \psi)$를 출력한다.

$$s(x, \psi) = 2^{-\frac{E(h(x))}{c(\psi)}}$$

where h(x) = a single path length

$$c(\psi) = \begin{cases} 2H(\psi - 1) - 2(\psi - 1)/n & \text{for } \psi > 2, \\ 1 & \text{for } \psi = 2, \\ 0 & \text{otherwise.} \end{cases}$$

where H(i) = the harmonic number,

$\approx \ln(i) + 0.5772156649$(Euler's constant)

step 3) 이상 점수 $s(x, \psi)$를 내림차순으로 나열하여 상위 m개를 이상치로 분류한다.

Liu et al,(2008)에 따르면 이상 점수 $s(x,\psi)$가 0.5보다 작으면 정상치로 봐도 안전하며 1에 가까우면 확실히 이상치라고 본다. 따라서 본 논문에서는 이상 점수 $s(x,\psi)$를 내림차순으로 나열하여 상위 10%에 속하면서 0.55보다 큰 표본을 이상치 즉, 소수계급으로 분류하였다.

3. 모형 성능 평가 기준

구축된 분류 모형의 성능을 평가하는 방법에서 가장 중요한 요소는 예측력이다. 이진 분류에서 분류 모형의 예측 성능을 평가하는 평가 메트릭(metric)은 혼동 행렬(confusion matrix)로부터 계산한다. 혼동행렬은 모형을 통해 구한 예측 값의 빈도와 자료의 실제 값의 발생 빈도를 정리하여 나타낸 분류표이다.

Table 2. Confusion Matrix

		Actual value	
		Positive	Negative
Predicted value	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

위의 혼동행렬에서 positive로 표현된 계급은 계급 불균형 자료에서 관심 계급인 소수계급을 나타낸다. TP에 해당하는 칸은 실제 값이 Positive(class=1)이고 예측 값도 Positive(class=1)인 경우의 빈도를, FP에 해당하는 칸은 실제 값이 Negative(class=0)인데 예측 값은 Positive(class=1)인 경우의 빈도를 의미한다. 이처럼 계급의 앞에 붙은 True(T)와 False(F)는 예측이 정확했는지 틀렸는지를 나타낸다. 또한 앞으로의 수식에서 Positive(P)는 소수계급을, Negative(N)는 다수계급을 대체하여 사용한다.

3.1 정확도(Accuracy)

위의 혼동행렬을 통해 계산할 수 있는 가장 대표적인 메트릭으로 정확도가 있다. 정확도는 식(1)과 같이 구할 수 있다.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (1)$$

정확도는 모든 계급의 정분류 정도를 고려한 측도로 분류 알고리즘에서 가장 보편적으로 사용되는 기준이다. 하지만 계급 불균형자료에서는 소수계급이 정확도에 미치는 영향력이 상대적으로 작기 때문에 모형의 평가 측도로 사용하기에 부적합하다. 예를 들어 희귀질환의 유병률이 전체 모집단의 1%일 경우, 모든 개체를 희귀질환이 없다고 분류하면 99%의 정확도를 얻을 수 있지만 이는 의미 있는 수치가 아니다. 일반적으로 계급불균형 자료의 분류 문제에서는 소수계급의 정확도를 높이는 것이 목적이므로 정확도가 아닌 다른 측도를 이용하여 모형을 평가해야 한다.

3.2 F1 measure

위의 혼동행렬을 이용하여 계산할 수 있는 또 다른 메트릭으로 정밀도(precision)와 재현율(recall)이 있다. 정밀도는 식(2)를 통해 구할 수 있다.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

정밀도는 Positive class로 예측된 것 중 실제로도 Positive class인 경우의 비율을 나타낸다. 따라서 이 값이 클수록 모형이 실제 다수계급을 소수계급으로 예측하지 않았다는 것을 의미한다.

재현율은 식(3)을 통해 구할 수 있다.

$$\text{Recall} = \frac{TP}{TP+FN} \tag{3}$$

재현율은 실제로 Positive class에 속하는 자료 중 Positive class로 예측이 된 경우의 비율을 나타낸다. 즉, 이 값이 클수록 모형이 소수계급을 잘 훈련했다는 것을 의미한다.

따라서 이 두 값이 큰 경우 소수계급에 대한 분류 성능이 좋다고 할 수 있다(Buckland, M., & Gey, F., 1994). 하지만 정밀도와 재현율은 서로 상충(trade-off) 관계가 있기 때문에 정밀도가 커질수록 재현율은 작아지게 되고, 정밀도가 작을수록 재현율은 커지게 된다. 따라서 이 두 측도를 모두 반영하여 식 (4)와 같이 F measure를 계산 할 수 있다.

$$F\,measure = \frac{1}{\alpha\frac{1}{Recall}+(1-\alpha\frac{1}{Precision})} \tag{4}$$

이 때 정밀도와 재현율의 중요도에 따라 가중치 α를 조정할 수 있는데, 둘의 가중치를 동일하게 0.5로 주어 계산한 것이 F1 measure이다.

$$F1\ measure = \frac{2Recall \times Precision}{Recall + Precision}$$

이는 정밀도와 재현율의 조화평균으로 0과 1 사이의 값을 가지며 두 값 중 더 작은 값에 가까워지는 특성이 있다. 따라서 F1 measure가 높다는 것은 정밀도와 재현율이 높다는 것이므로, F1 measure가 더 클수록 소수계급의 분류 성능이 좋은 모형이라 할 수 있다(Sun et al., 2009).

3.3 AU-ROC

혼동행렬에서 대표적으로 계산되는 메트릭에 민감도(sensitivity)와 특이도(specificity)가 있다. 민감도는 실제로 Positive class에 속하는 자료 중 Positive class로 예측이 된 경우로 3.2절에서 언급한 재현율과 같다. 특이도는 실제로 Negative class에 속하는 자료 중 Negative class로 예측이 된 경우의 비율을 나타낸다. 민감도와 특이도의 식은 (5), (6)과 같다.

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{5}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{6}$$

ROC 곡선(Receiver Operating Characteristics Curve)은 모형의 분류결과를 민감도(Sensitivity)와 특이도(Specificity)를 이용해 나타낸 것으로, 두 지표의 교환 정도를 비교할 수 있다. ROC 곡선의 형태는 <Figure 5>와 같이 나타난다.

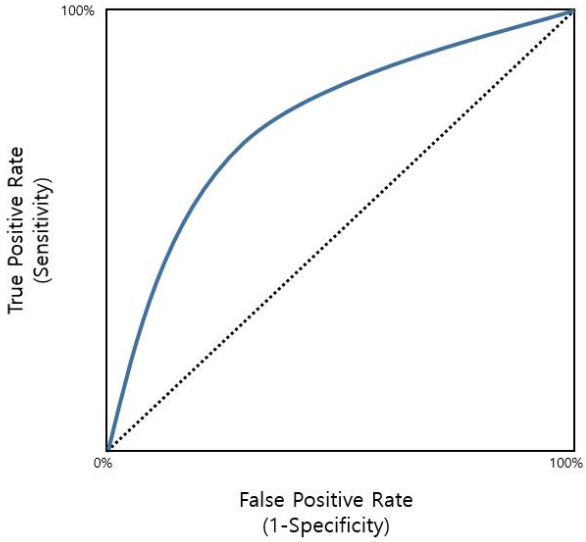

Figure 5. Example of ROC Curve

AU-ROC는 ROC 곡선의 아래쪽 면적을 나타내는 값으로, ROC 곡선을 단일 통계량으로 나타내어 여러 모형의 ROC 곡선을 비교할 수 있다. AU-ROC 값은 False Positive Rate를 조금 증가시키면서 True Positive Rate를 많이 증가시킬 때 넓어진다. 즉, 다수계급을 소수계급으로 잘못 분류하는 것에 비해 소수계급을 소수계급으로 잘 분류한다면 AU-ROC 값이 커지므로, AU-ROC 값이 1에 가까울수록 모형의 분류 성능이 좋다고 평가할 수 있다.

본 논문에서는 F1 measure와 AU-ROC를 분류 성능의 평가 기준으로 사용하였다.

Ⅲ. 모의실험

랜덤 포레스트를 이용하여 계급 불균형 자료를 분류 할 때, 불균형 정도에 따른 각 방법들의 분류 성능을 모의실험을 통해 비교해보고자 한다. 본 논문에서 비교를 위해 사용된 방법은 앞에서 언급한 6가지로, 비교에 사용한 방법은 아래와 같이 표현할 것이다.

- RF: 원자료(original data)에 랜덤 포레스트 이용
- ROS: Random Over Sampling을 적용한 자료에 랜덤 포레스트 이용
- RUS: Random Under Sampling을 적용한 자료에 랜덤 포레스트 이용
- SMOTE: SMOTE를 적용한 자료에 랜덤 포레스트 이용
- BRF: 원자료(original data)에 Balanced Random Forest 이용
- WRF: 원자료(original data)에 Weighted Random Forest 이용
- IF: 원자료(original data)에 Isolation Forest 이용

위의 방법들을 비교하기 위한 성능 평가 기준으로는 본론의 3절에서 제시한 2가지 통계량인 F1-measure와 AU-ROC를 사용하였다.

1. 모의실험 설계

먼저 불균형 자료를 만들기 위해 표준정규분포 $N(0,1)$, 지수분포 $\mathrm{Exp}(1)$, 균등분포 $Unif(0,1)$에서 서로 독립인 설명 변수를 각각 3개씩 생성한다. 설명 변수 $X_1, \ldots, X_9$와 표준정규분포를 따르는 오차항을 이용하여

만든 임의의 변수 η를 통해 η가 $(1-\frac{IR}{IR+1})/2$분위수 보다 작거나 $(1+\frac{IR}{IR+1})/2$ 분위수 보다 큰 경우 반응 변수 Y를 소수계급으로 하고 그 이외의 경우는 반응 변수 Y를 다수계급으로 하여 반응 변수가 두 계급을 갖도록 생성하였다. 표본 크기는 1,000개와 2,000개로 하였다.

$$\begin{aligned} X_1, X_2, X_3 &\sim N(0,1) \\ X_4, X_5, X_6 &\sim \mathrm{Exp}(1) \\ X_7, X_8, X_9 &\sim Unif(0,1) \end{aligned}$$

$$\eta = (\exp(X_1 - X_2) + X_4^3)/X_7 + \epsilon, \quad \epsilon \sim N(0,1)$$

$$Y = \begin{cases} Class\,0 & \text{if} \quad \eta_{(\frac{n}{a1})} < \eta < \eta_{(\frac{n}{a2})} \\ Class\,1 & \text{if} \quad o.w \end{cases}$$

$$where \;\; a1 = (1-\frac{IR}{IR+1})/2 \,*100 \;\; \text{and} \;\; a2 = (1+\frac{IR}{IR+1})/2 \,*100$$

설명 변수 X_3, X_5, X_6, X_8, X_9의 경우 반응 변수 Y의 생성에 무관하지만, 정보량을 늘리기 위해 생성하였다. 또한 실제로는 설명 변수의 분포를 알 수 없기 때문에, 다양한 분포를 가정하고 비선형 함수를 이용하여 반응 변수 Y를 생성하였다. 불균형 정도는 IR을 통해서 조정하였고 IR의 범위는 {9, 19, 29, 49, 99, 199}로 9 이상의 매우 불균형한 자료를 설정하였다. 각 IR 값에 대응되는 전체 자료에서 소수계급이 차지하는 비율은 {10%, 5%, 3%, 2%, 1%, 0.5%}이다.

모형 적합에는 통계 소프트웨어 R의 'Caret', 'DMwR', 'randomForest', 'isofor' 패키지를 사용하였다. 각 방법에서 공통적으로 추정해야하는 모수는 의사결정나무의 개수(ntree, n)로 본 모의실험에서는 100으로 두었다. 또한 RF, ROS, RUS, SMOTE, BRF, WRF 방법에서 추정해야하는 모수인 노드 생성에 고려할 변수의 개수(mtry)는 3으로 설정하였다. WRF 방법에서 가중치는 <Table 1>에 따라 상대 집단의 크기에 비례하는 값으로 부여하였다. IF 방법에서 sub-sample 크기(ψ)는 128로 두었다.

생성된 데이터 중 모형의 학습에 사용되는 훈련 자료(train dataset)와 평가에 사용되는 평가 자료(test dataset)의 비율은 8 : 2로 하였다. 각 실험은 500번 반복하였고 분류 성능 평가 기준 값은 500번 반복의 평균값을 이용하였다.

2. 모의실험 결과

모의실험의 결과 F1 measure의 값은 <Table 3>에 나타나있고 **굵은 숫자**는 제일 높은 값을, 밑줄 친 숫자는 제일 작은 값을 나타낸다. 이 결과를 이용하여 그래프를 그리면 <Figure 6, 7>과 같다.

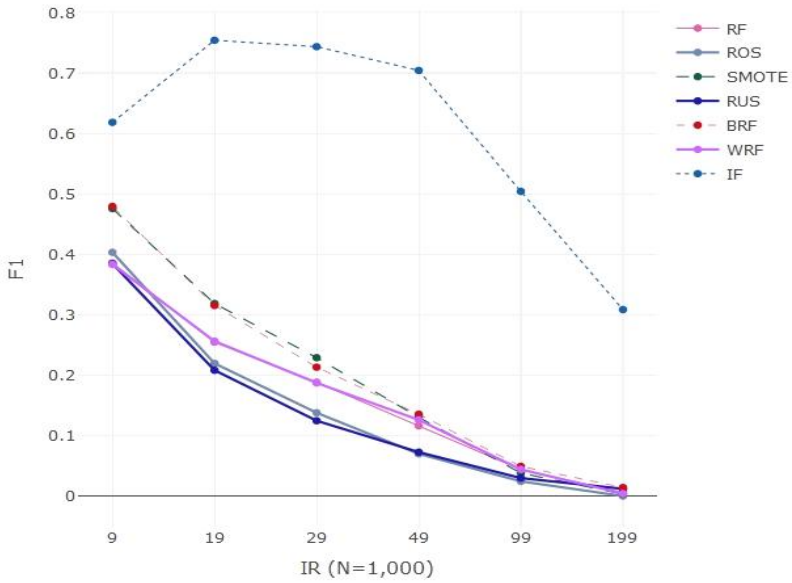

Figure 6. Results of simulation study N=1,000 (F1 measure)

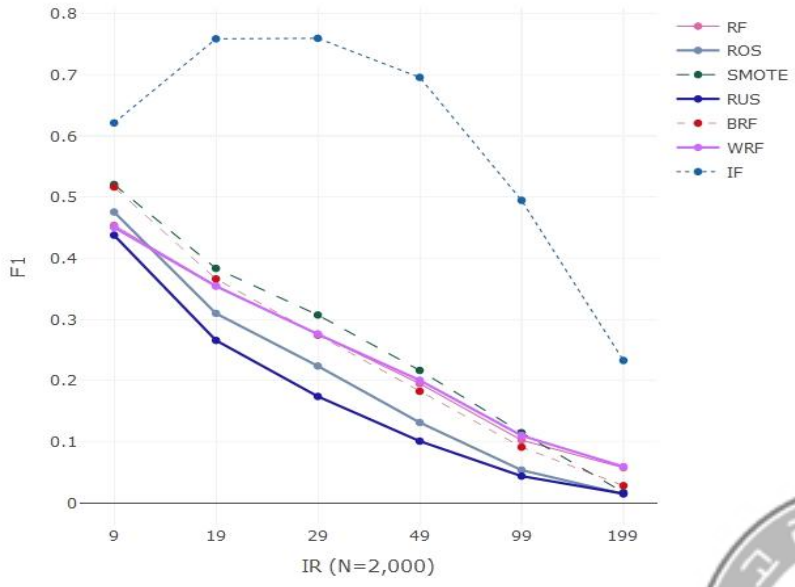

Figure 7. Results of simulation study N=2,000 (F1 measure)

<Table 3>와 <Figure 6, 7>에서 한눈에 확인 할 수 있듯이 모든 IR에서 IF 방법이 가장 좋은 성능을 보였다. 모든 방법에서 불균형 정도가 심해질수록 F1 measure는 점차 감소하였다. 다만 IF 방법의 경우, IR이 9에서 19로 증가할 때는 F1 measure가 증가하는 경향을 보였다. IR이 50보다 낮을 경우 IF 방법의 F1 measure는 0.6보다 높은 값을 유지하였다. 표본의 크기가 1,000개일 때와 2,000개일 때 IF 방법을 제외한 나머지 방법의 F1 measure는 값이 조금씩 증가하였다. 하지만 모든 방법들이 표본크기의 증가에 따라 분류성능이 크게 좋아지지는 않았다. 거의 모든 IR값에서 ROS 방법과 RUS 방법은 RF 방법을 적용한 것 보다 분류 성능이 조금씩 떨어졌다. 하지만 RF 방법이 뛰어나다고 볼 수는 없었다.

다음으로 <Table 4>와 <Figure 8, 9>은 모의실험 결과의 AU-ROC 값을 나타낸다. AU-ROC에서도 마찬가지로 IF 방법이 가장 분류 성능이 좋았다. F1 measure와 마찬가지로 모든 방법에서 불균형 정도가 심해질수록 AU-ROC 값은 점차 감소하였다. 표본의 크기가 증가하면 전체적으로 AU-ROC 값이 조금씩 증가하였다. 모든 IR값에서 RUS는 RF보다 분류 성능이 떨어졌다. 또한 AU-ROC 값만 봤을 때 RF 방법만을 적용한 결과 값이 나쁘지는 않았다.

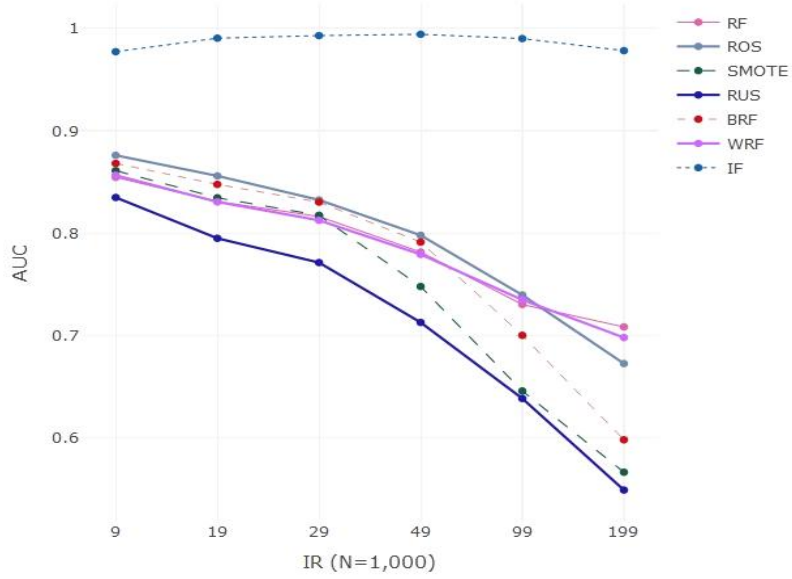

Figure 8. Results of simulation study N=1,000 (AU-ROC)

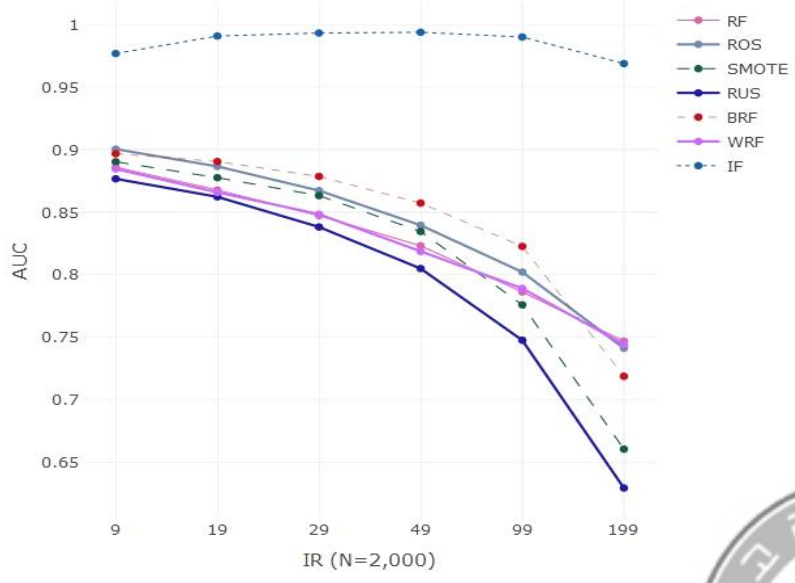

Figure 9. Results of simulation study N=2,000 (AU-ROC)

모의실험 결과, 전체적인 추세를 보면 방법 간의 감소폭은 차이가 있지만 불균형 정도가 증가함에 따라 모든 방법의 F1 measure와 AU-ROC 값은 감소하였다. 표본 수 역시 계급 불균형 자료의 분류 성능에 영향을 미침을 알 수 있었다.

Table 3. Results of simulation study for various methods according to IR when N=1,000 & 2,000 (F1 measure)

N	IR	RF	ROS	RUS	SMOTE	BRF	WRF	IF
1,000	9	0.3856	0.4032	0.3846	0.4756	0.4792	0.3832	**0.6182**
	19	0.2546	0.2195	0.2081	0.3185	0.3153	0.2562	**0.7540**
	29	0.1884	0.1378	0.1246	0.2289	0.2132	0.1872	**0.7434**
	49	0.1162	0.0698	0.0726	0.1303	0.1353	0.1260	**0.7041**
	99	0.0437	0.0243	0.0297	0.0380	0.0491	0.0440	**0.5040**
	199	0.0060	0	0.0119	0.0032	0.0141	0.0040	**0.3084**
2,000	9	0.4538	0.4755	0.4376	0.5205	0.5163	0.4502	**0.6212**
	19	0.3550	0.3097	0.2656	0.3833	0.3661	0.3542	**0.7587**
	29	0.2759	0.2236	0.1739	0.3071	0.2744	0.2757	**0.7596**
	49	0.1949	0.1313	0.1008	0.2165	0.1823	0.2004	**0.6957**
	99	0.1023	0.0534	0.0436	0.1146	0.0910	0.1099	**0.4946**
	199	0.0573	0.0143	0.0149	0.0170	0.0280	0.0591	**0.2327**

Table 4. Results of simulation study for various methods according to IR when N=1,000 & 2,000 (AU-ROC)

N	IR	RF	ROS	RUS	SMOTE	BRF	WRF	IF
1,000	9	0.8545	0.8760	0.8348	0.8610	0.8681	0.8565	**0.9773**
	19	0.8309	0.8559	0.7949	0.8345	0.8476	0.8305	**0.9905**
	29	0.8162	0.8324	0.7712	0.8174	0.8303	0.8125	**0.9930**
	49	0.7815	0.7979	0.7127	0.7478	0.7912	0.7792	**0.9943**
	99	0.7302	0.7395	0.6384	0.6457	0.7001	0.7348	**0.9901**
	199	0.7083	0.6725	0.5489	0.5663	0.5980	0.6979	**0.9783**
2,000	9	0.8860	0.9004	0.8767	0.8903	0.8968	0.8846	**0.9768**
	19	0.8678	0.8866	0.8623	0.8776	0.8905	0.8659	**0.9908**
	29	0.8473	0.8673	0.8382	0.8633	0.8786	0.8484	**0.9932**
	49	0.8232	0.8396	0.8050	0.8346	0.8573	0.8186	**0.9938**
	99	0.7863	0.8022	0.7477	0.7760	0.8227	0.7891	**0.9900**
	199	0.7470	0.7413	0.6296	0.6607	0.7189	0.7444	**0.9688**

Ⅳ. 실제자료 분석

다양한 분야의 실제 자료에서도 계급 불균형 문제가 발생한다. 실제 자료의 경우 각 자료의 구조에 따라 좋은 성능을 보이는 방법이 달라질 수도 있다. 따라서 본 논문에서는 실제 자료에 앞서 언급한 방법들을 적용하여 성능을 비교하였다. 각 방법들의 분류 성능 비료를 위해 5-fold Cross Validation(CV)을 사용하였다. 그 외에 랜덤 포레스트의 나무 수(ntree)와 WRF 방법의 오분류 비용, IF 방법의 sub-sample 크기(ψ) 등은 앞의 모의실험과 동일한 세팅을 사용하였다.

1. 분석 자료

분석에 사용된 자료는 총 6개로 반응 변수가 두 개의 계급으로 이루어진 범주형 변수이며, IR이 9 이상인 계급 불균형이 심한 자료이다. 'Mammography' 자료는 Woods et al.(1993)에서 사용된 자료이며, 나머지 자료는 KEEL(http://www.keel.es/data/php)에서 제공하는 자료이다. <Table 5>는 분석에 사용된 각 자료의 이름과, 관측치 수(Obs.), 실수인 설명 변수의 수(Attri.), IR 및 소수계급과 다수계급의 빈도와 비율을 정리한 내용이다.

Table 5. Summary of Actual data

Data set	Obs.	Attri.	IR	Minority(%)	Majority(%)
Clevland-0vs4	173	13	12.3	13 (7.5)	160 (92.5)
Winequality-red-4	1,599	11	29.2	53 (3.3)	1,546 (96.7)
Yeast6	1,484	8	41.4	35 (2.4)	1,449 (97.6)
Mammography	11,183	6	42.0	260 (2.3)	10,923 (97.7)
Poker-89vs6	1,485	10	58.4	25 (1.7)	1,460 (98.3)
Abalone19	4,174	7	129.4	32 (0.8)	4,142 (99.2)

2. 분석 결과

<Table 6>는 각 자료별로 앞에서 언급한 방법들을 적용한 분류 결과의 F1 measure 값을 나타낸다. **굵은 값**이 가장 높은 F1 measure 값, 밑줄친 값이 가장 작은 F1 measure 값을 보인 방법을 나타낸다. 분석 결과 모든 방법에서 높은 값을 보인 방법은 없었다. 다만 RF의 경우 대체적으로 가장 낮은 F1 measure 값을 보이는 것을 확인할 수 있었다. 앞의 모의실험에서 대체적으로 안정적인 값을 보였던 IF 방법 역시 IR이 커질수록 값이 매우 떨어졌다. 'Mammography' 자료에서는 다른 자료에서는 상대적으로 성능이 떨어졌던 샘플링 방법이나 아무것도 하지 않은 RF 방법이 좋은 값을 보였다. 이는 앞에서도 언급했듯, 불균형 자료 문제에 영향을 미치는 것이 단순히 불균형 정도(imbalance ratio) 뿐만이 아니라는 것을 보여준다.

<Table 7>은 AU-ROC 값을 나타낸다. 마찬가지로 **굵은 값**이 방법이 가장 높은 AU-ROC 값, 밑줄 친 값이 가장 작은 AU-ROC 값을 보인 방법을 나타낸다. 여기서는 모두 IF 방법이 0.9보다 큰 값을 가지며 안정적으로 제일 높은 값을 보인다. 자료 'Clevland-0vs4'와 'Abalone19'의 경우 F1 measure 값과 마찬가지로 RF 방법이 가장 낮은 AU-ROC 값을 보였다. 따라서 불균형 자료에서 기존의 분류 알고리즘을 사용하는 것에는 다소 한계가 있다는 것을 다시 한 번 확인할 수 있었다.

F1 measure로 본 각 방법의 분류성능과 AU-ROC로 본 각 방법의 분류성능이 약간의 차이가 났다. 이는 두 측도가 보여주는 의미가 다르기 때문인 것으로 판단된다. 이러한 점을 고려했을 때 전반적으로 IF 방법이 실제 자료의 적용에도 나쁘지 않은 성능을 보인다고 볼 수 있다. 하지만 데이터 구조에 따라 분류 성능을 향상시키는 방법이 달라질 수 있으니 데이터에 맞는 방법을 선택하는 것이 중요하다.

Table 6. Results of actual data analysis (F1-measure)

Data set	IR	RF	ROS	RUS	SMOTE	BRF	WRF	IF
Clevland-0vs4	12.3	0.3000	0.4048	0.3883	0.7000	0.6124	0.3333	**0.8095**
Winequality-red-4	29.2	0	0	0.1146	0.1901	0.1842	0	**0.2522**
Yeast6	41.4	0.4962	0.5282	0.2613	0.4143	0.3547	0.5161	**0.7153**
Mammography	42.0	0.6803	**0.6952**	0.3460	0.6141	0.4712	0.6763	0.4238
Poker-89vs6	58.4	0	0.3310	0.0811	**0.5952**	0.4230	0	0.2400
Abalone19	129.4	0	0	0.0377	**0.0772**	0.0420	0	0.0669

Table 7. Results of actual data analysis (AU-ROC)

Data set	IR	RF	ROS	RUS	SMOTE	BRF	WRF	IF
Clevland-0vs4	12.3	0.9089	0.9510	0.9122	0.9853	0.9645	0.9698	**0.9844**
Winequality-red-4	29.2	0.7974	0.7784	0.7254	0.7621	0.7744	0.7933	**0.9371**
Yeast6	41.4	0.9219	0.9193	0.9405	0.9304	0.9501	0.9062	**0.9942**
Mammography	42.0	0.9484	0.9553	0.9568	0.9555	0.9602	0.9527	**0.9855**
Poker-89vs6	58.4	0.9395	**0.9832**	0.8735	0.9801	0.9310	0.9756	0.9579
Abalone19	129.4	0.6674	0.7101	0.7830	0.7745	0.7941	0.6727	**0.9148**

Ⅴ. 고찰

본 논문에서는 계급 불균형 자료의 분류 문제에서 분류 성능을 개선하기 위한 다양한 방법 중 랜덤 포레스트에 기반한 방법을 고려하였다. 표본의 분포를 직접 조정하는 자료 수준의 접근 방법으로 ROS, RUS, SMOTE 방법으로 훈련 자료를 조정 후 랜덤 포레스트를 적용하였고, 알고리즘 수준의 접근 방법으로 랜덤 포레스트 알고리즘에 변형을 준 Weighted Random Forest, Balanced Random Forest, Isolation Forest를 고려하였다. 모의실험과 실제 자료의 분석을 통해, IR이 9 이상인 매우 불균형한 분포를 보이는 계급 불균형 자료에서 IR의 변화에 따른 각 방법의 분류 성능을 비교해 보았다. 모의실험 결과 고려된 방법 중 자료 수준의 접근법은 분류 성능이 알고리즘 수준의 접근법에 비해 상대적으로 좋지 않았다. 이는 앞에서 ROS와 RUS의 단점으로 언급했던 과대적합과 정보 손실 때문인 것으로 보인다. 이러한 단점을 보완하기 위한 방법들이 제안되었다. Evolutionary Under Sampling(EUS)(Garcia, S., & Herrera, F., 2009), Cluster-based under sampling(Yen, S. J., & Lee, Y. S., 2009) 등은 RUS의 정보 손실을 최소화하기 위해 중요 정보를 담고 있는 대표 표본을 선택하는 방법이다. Borderline-SMOTE(Han et al., 2005)는 다수계급과 소수계급의 경계(borderline)에 있는 표본을 이용하여 SMOTE를 적용한 방법으로 SMOTE 보다 소수계급에 유사한 표본을 만들어서 소수계급의 정확도를 높이는 방법이다.

또한 모의실험에서 IR이 커질수록 F1 measure와 AU-ROC가 거의 모든 방법에서 떨어짐을 확인 할 수 있었으나, 공간분할 기반의 이상탐지 알

고리즘인 Isolation Forest의 경우 다른 방법들에 비해 비교적 안정적인 성능을 보였다. 오히려 Isolation Forest는 IR이 작은 자료 즉, 균형분포에 가까운 자료일수록 성능이 다소 떨어졌다. 이러한 현상은 Isolation Forest의 기본 전제인 'few and different'에서 그 이유를 찾을 수 있다. Isolation Forest 알고리즘 입장에서 보면 IR이 작아진다는 것은 이상치인 표본의 수가 정상치인 표본의 수와 비슷해진다는 것이다. 그렇기 때문에 소수계급을 분리시킬 때의 의사결정나무의 깊이와 다수계급을 분리시킬 때의 의사결정나무의 깊이가 비슷해지게 되므로 분류 성능이 떨어지게 되는 것이다. 따라서 Isolation Forest 방법은 심한 계급 불균형 자료에 사용하는 것이 좋다고 판단된다.

본 논문에서 살펴본 6가지 방법론 이외에도 다양한 방법론들이 많이 소개되고 있다. 예를 들어 계급 불균형 자료에 영향을 미치는 다양한 요소를 보완하기 위한 방법론이 많이 소개되고 있다. Jo et al.(2004)은 계급 간 불균형(between-class imbalance)과 단일 계급 내 불균형(within-class imbalance)을 동시에 고려하기 위해 각 계급별로 클러스터(Cluster)를 구성하고, 그 클러스터에 근거하여 오버샘플링을 하는 Cluster-based over sampling 방법을 제안하였다. Wasikowski(2010)는 표본 크기가 작은 고차원 자료에서 변수선택(feature selection)이 계급 불균형 문제의 완화에 도움이 될 수 있다는 것을 확인하였다. 따라서 추후 계급 불균형 자료의 분류에 영향을 미치는 다양한 요소를 반영한 모의실험 연구가 필요하다.

또한 가장 좋은 성능을 보인 Isolation Forest 알고리즘의 경우 랜덤 포레스트 뿐만 아니라 다른 알고리즘과도 여러 측면에서 성능을 비교하는 연구가 필요하다. 특히 자료의 크기가 큰 빅데이터의 경우 모형의 훈련에

많은 시간이 소요된다. 본 논문에서 실제 자료를 분석하였을 때, 표본의 크기가 2,000개 보다 작은 자료의 분석에 비해 표본의 크기가 4,174개인 'Abalone19' 자료와 11,183개인 'Mammography' 자료의 분석에 소요되는 시간이 더 길어졌다. 따라서 동일한 성능을 유지하며 학습시간을 줄일 수 있는 방법에 대한 연구가 필요하다.

추가적으로 본 논문에서는 반응 변수가 두 개의 계급을 갖는 이진 분류 문제에 대해 살펴보았는데, 여러 개의 계급을 갖는 다중 계급 불균형 자료(multi-class imbalanced data)로 연구를 확장할 필요가 있다.

Ⅵ. 결론

반응 변수가 0과 1로 이루어진 자료에서 하나의 계급이 다른 계급에 비해 매우 낮은 빈도를 갖는 계급 불균형 자료를 분류할 때에는 이러한 불균형 분포를 고려한 방법이 필요하다. 특히 계급 불균형 자료에서는 일반적으로 소수계급을 정확히 분류하는 것이 중요한데, 두 계급의 균형 분포를 가정하고 전체적인 오분류율을 최소화하는 일반적인 분류 알고리즘을 적용 시 다수계급에 편향된(biased) 결과를 얻게 된다. 또한 모형의 분류 성능을 비교할 때에도 두 계급의 균형 분포를 가정한 측도인 정확도(accuracy) 보다는 소수계급의 분류 정도를 반영할 수 있는 F1 measure나 AU-ROC와 같은 측도를 사용해야 한다. 따라서 본 논문에서는 이러한 문제를 해결하기 위한 자료 수준의 접근 방법과 알고리즘 수준의 접근 방법에 속하는 6가지 방법을 F1 measure와 AU-ROC를 통해 비교해 보았다.

모의실험 결과 불균형 정도가 심해질수록 F1 measure와 AU-ROC가 거의 모든 방법에서 떨어졌지만 Isolation Forest의 경우는 좋은 성능을 보였다. 그리고 표본의 수를 늘리면 F1 measure와 AU-ROC 값은 조금씩 증가하는 것으로 보아 계급 불균형 자료의 분류에 표본 수 역시 영향을 미침을 확인하였다. 모의실험 결과와 달리 실제 자료 분석에서 랜덤 포레스트는 더욱 안 좋은 결과를 보였고 반면에 ROS 방법은 모의실험 결과와 달리 실제 자료 분석에서 좋은 결과를 보였다. 실제 자료의 분석에서는 거의 모든 방법에서 모의실험 결과와 달리 일정한 경향을 보이지 않았으므로 자료 특성에 맞는 방법을 선택하는 것이 중요하다.

참고문헌

박창이, 김진석. (2008). R을 이용한 데이터마이닝. 교우사

Ali, A., Shamsuddin, S. M., & Ralescu, A. L. (2015). Classification with class imbalance problem: a review. Int. J. Advance Soft Compu. Appl, 7(3).

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). Classification and regression trees. CRC press.

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

Buckland, M., & Gey, F. (1994). The relationship between recall and precision. Journal of the American society for information science, 45(1), 12.

Chen, C., Liaw, A., & Breiman, L. (2004). Using random forest to learn imbalanced data. University of California, Berkeley, 110.

Dittman, D. J., Khoshgoftaar, T. M., & Napolitano, A. (2015, August). The effect of data sampling when using random forest on imbalanced bioinformatics data. In Information Reuse and Integration (IRI), 2015 IEEE International Conference on (pp. 457-463). IEEE.

García, S., & Herrera, F. (2009). Evolutionary undersampling for

classification with imbalanced datasets: Proposals and taxonomy. Evolutionary computation, 17(3), 275-306.

Han, H., Wang, W. Y., & Mao, B. H. (2005). Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. Advances in intelligent computing, 878-887.

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on knowledge and data engineering, 21(9), 1263-1284.

Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. Intelligent data analysis, 6(5), 429-449.

Jo, T., & Japkowicz, N. (2004). Class imbalances versus small disjuncts. ACM Sigkdd Explorations Newsletter, 6(1), 40-49.

Liu, F. T., Ting, K. M., & Zhou, Z. H. (2008, December). Isolation forest. In Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on (pp. 413-422). IEEE.

Sun, Y., Wong, A. K., & Kamel, M. S. (2009). Classification of imbalanced data: A review. International Journal of Pattern Recognition and Artificial Intelligence, 23(04), 687-719.

Yen, S. J., & Lee, Y. S. (2009). Cluster-based under-sampling approaches for imbalanced data distributions. Expert Systems with Applications, 36(3), 5718-5727.

Wasikowski, M., & Chen, X. W. (2010). Combating the small sample class imbalance problem using feature selection. IEEE Transactions on knowledge and data engineering, 22(10), 1388-1400.

국문 요약

계급 불균형 자료에서 랜덤 포레스트에 기반한 분류 방법의 성능 비교

문 선 영

고려대학교 대학원 의학통계학협동과정

(지도교수: 안 형 진, Ph.D.)

목적 : 많은 분야의 자료에서 관심 사건이 매우 드물게 발생하기 때문에 반응 변수의 계급이 매우 불균형한 분포(소수계급, 다수계급)를 보인다. 하지만 일반적인 분류 알고리즘은 계급의 균형 분포를 가정하고 전체적인 오분류율을 최소화하는 것을 목적으로 한다. 따라서 소수계급에 관심이 있는 계급 불균형 자료에 이러한 일반적인 분류 알고리즘을 사용할 경우 소수계급의 분류 정확도가 매우 떨어지게 된다. 이러한 계급 불균형 문제를 보완하기 위한 여러 가지 방법이 제시되었다. 본 연구의 목적은 계급 불균형 문제를 보정할 수 있는 여러 방법들의 분류 성능을 계급 불균형 정도(IR)에 따라 비교하는 것이다.

방법 : 본 연구에서는 랜덤 포레스트를 기본 분류 알고리즘으로 사용하고, 계급 불균형 문제를 보완할 수 있는 여러 방법들 중에서 랜덤 포레스트에 기반한 알고리즘을 사용하였다. 표본의 분포를 조정하는 자료 수준의 접근 방법인 ROS, RUS 및 SMOTE 방법을 통해 수정한 표본에 랜덤 포레스트를 적용하였고, 알고리즘을 수정하는 알고리즘 수준의 접근 방법으로는 Weighted Random Forest, Balanced Random Forest, Isolation Forest 방법을 사용하였다. 또한 계급 불균형 자료의 분류에서는 소수계급의 보다 정확한 분류에 관심이 있기 때문에 일반적인 모형 성능 평가 지표인 정확도를 사용할 수 없다. 따라서 본 연구에서는 F1 measure와 AU-ROC를 성능 평가 지표로 사용하였다.

결과 : 시뮬레이션 결과 모든 IR에서 F1 measure와 AU-ROC 지표 모두 공간분할을 통한 이상 탐지 알고리즘인 Isolation Forest 방법이 좋은 성능을 보였다. 추가적으로 표본의 수를 늘리면 두 지표 모두 약간의 증가를 보여 계급 불균형 자료의 분류에 표본 수 역시 영향을 미침을 확인할 수 있었다. 또한, 전체적인 추세를 보면 각 방법 간의 감소폭에는 차이가 있었지만, 불균형 정도가 심해질수록 모든 방법에서 두 지표의 값은 감소하였다.

결론 : 0 또는 1의 값을 갖는 반응 변수의 분포가 매우 불균형한 경우 일반적인 분류 알고리즘의 소수계급 분류 정도는 다소 떨어진다. 하지만 단일 계급 내 불균형이나 표본의 크기 등 계급 불균형 자료에 영향을 미치는 여러 가지 요인들이 있으므로 자료의 구조에 맞는 알맞은 방법을 선택하여 사용하는 것이 중요하다.

주제어 : 계급 불균형 자료, 불균형 정도, 랜덤 포레스트, 랜덤오버샘플링, 랜덤언더샘플링, SMOTE, 가중 랜덤 포레스트, 균형 랜덤 포레스트, 아이솔레이션 포레스트