creative commons
COMMONS DEED

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

碩 士 學 位 論 文

불균형 자료의 분류 문제에서 분류 향상을 위한 방법 비교 연구

高麗大學校 大學院

統計學科

李 恩 景

2016年 12月 日

LIBERTAS JUSTITIA VERITAS

李 宰 遠 敎授指導
碩 士 學 位 論 文

불균형 자료의 분류 문제에서 분류 향상을 위한 방법 비교 연구

이 論文을 統計學碩士 學位論文으로 提出함.

2016年 12月 日

高麗大學校 大學院
統計學科
李 恩 景

李恩景의 統計學碩士 學位論文

審査를 完了함.

2016年 12月 日

委員長 李寧達 (印)

委 員 조형준 (印)

委 員 신승준 (印)

개 요

불균형 자료는 이항 반응변수의 분류 문제에서 쉽게 접할 수 있는 자료이다. 일반적인 분류 알고리즘은 두 집단의 균형 분포와 동일한 오분류 비용(cost)을 가정한다. 따라서 로짓 회귀, 의사결정나무와 같은 알고리즘을 이용하여 분류할 경우, 소수집단의 분류가 제대로 이루어지지 않는 문제가 발생하게 된다. 앞선 연구를 연장하여 본 논문에서는 기존의 불균형 자료의 분류 문제 해결을 위한 Cost-Sensitive learning, Tree Ensemble 방법과 Resampling 방법을 비교하려고 한다. 다양한 상황에서 방법 간의 비교를 통해 각 방법의 특징을 알아보고 상황별 방법을 제안하고자 한다.

고려대학교 KOREA UNIVERSITY 1905

목 차

고려대학교 KOREA UNIVERSITY 1905

표 목차

고려대학교 KOREA UNIVERSITY 1905

그림 목차

제 1 장

연구의 배경과 목적

다양한 통계적 방법에 대한 연구 중 한 가지는 두 그룹에 대한 분류 문제이다. 분류 문제에서 다루는 자료 중 한 그룹의 비율이 다른 그룹에 비해 현저하게 높은 불균형 자료는 현실의 다양한 상황에서 발생하고 있다. 의학 분야의 희귀병 진단, 사회과학 분야의 전쟁 발발 예측, 경제 분야의 거래 이상 감지, 과학 분야의 인공위성 이미지 판별 등의 다양한 상황에서 불균형 자료가 이에 해당하는 예이다. 대부분의 불균형 자료에서는 소수 집단의 예측이 주요 관심사이다. 예를 들어, 희귀병 진단 자료의 경우, 희귀병을 가진 환자는 환자가 아닌 사람보다 현저히 적게 발생한다. 하지만 희귀병 진단 자료의 분류 목표는 희귀병을 가진 사람을 제대로 진단해 내는 것이다.

분류를 위한 대다수의 분류 알고리즘은 두 그룹의 균형 분포와 동일한 오분류 비용(cost)를 가정하고 있으며 전체적인 오분류율을 최소화하는 것을 분류 목적으로 한다. 따라서 일반적인 분류 알고리즘에서는 소수 집단의 개체가 잡음(noise)로 인식되어 불균형 자료의 분류 문제가 발생하게 된다. 이에 대한 결과로 전체적인 오분류율에 비해 소수 집단은 상당히 높은 오분류율을 갖게 된다. 즉, 소수 집단이 다수 집단에 비해 제대로 분류되지 않는 문제가 발생하게 된다. 따라서, 불균형 자료를 분류하는 경우에는 소수 집단에 대한 예측력 향상을 위한 방법이 필요하다.

이산형 반응 변수를 갖는 불균형 자료에서 불균형 문제 해소를 위해 다양한 분야에서 연구가 진행되고 있다. 첫 번째는 의사결정나무모형을 발전시킨 Tree Ensemble 방법, 두번째는 자료의 각 집단의 오분류 비용(cost)

고려대학교 KOREA UNIVERSITY 1905

을 다르게 부여하는 Cost-sensitive learning 방법, 마지막으로 Resampling 방법인 'Data-level approaches'가 있다. Cost-sensitive방법과 Data-level approaches의 경우, 적용이 간단하고 알고리즘과 독립적으로 사용할 수 있다는 장점이 있다.

Resampling 방법은 자료의 분포를 수정하는 방법으로 기본적인 방법으로는 Random under-sampling(RUS)와 Random over-sampling(ROS)이 있다. RUS의 경우 기존의 자료가 지닌 정보를 손실한다는 단점을 가지고 있으며 ROS의 경우 동일한 개체의 중복으로 과대 적합(over-fitting)의 단점을 지니고 있다. 따라서 이를 보완한 Tomek Link, SMOTE 방법 등이 있으며, SMOTE방법을 확장시킨 Borderline-SMOTE, Safe-level SMOTE 방법이 연구되고 있다. (Bunkhumpornpat C, Sinapiromsaran K & Lursinsap C)

불균형 자료에서 예측력을 판단할 때, 일반적으로 사용하는 오분류율이 적절하지 않다는 문제를 가지고 있다. 오분류율이 작더라도 소수 집단에 대한 높은 오분류율을 반영되어 있지 못한다. 따라서 불균형 자료에 대한 연구에서는 소수 집단의 분류를 고려하는 적절한 평가 기준을 사용해야 한다.

이에 본 논문에서는 불균형 자료에서 의사 결정 나무 알고리즘이 보일 수 있는 문제점을 알아보고자 한다. 그리고 이에 대한 해결책으로 자료 수준의 접근 방법과 기존에 불균형 자료를 다루기 위한 방법들과 비교해 보려고 한다. 먼저, 모의실험을 통하여 불균형 비율, 두 class 사이의 overlapping 정도의 다양한 상황에서 비교를 통해 각 상황에서 방법들의 특징을 살펴보고자 한다. 이어서 실제 불균형 자료에서 적용을 통해 방법 간 비교를 하려고 한다. 이를 통해 다양한 상황에서 방법 간 비교를 통해 방법의 특징을 알아보고 각 상황에 따른 알맞은 방법을 제시하고자 한다.

고려대학교 KOREA UNIVERSITY 1905

제 2 장
방법론 검토

2.1. 분류 알고리즘

두 그룹의 분류를 위한 다양한 알고리즘이 존재한다. 본 연구에서는 대표적인 분류 알고리즘으로써 적용이 간단하고 해석이 용이한 모형인 의사결정나무 알고리즘을 사용하였다.

2.1.1. 의사결정나무(Decision Tree) - CART(Classification And Regression Tree)

의사결정나무는 의사결정규칙을 나무구조로 도표화하여 분류와 예측을 수행하는 분석방법이다. 본 연구에서는 의사결정나무 알고리즘 중 가장 널리 쓰이는 CART 알고리즘을 사용하였다. CART는 분류규칙을 형성할 때 이진분리를 하며, 지니계수를 가장 많이 감소시키는 설명변수를 찾고, 해당 변수의 분리 기준점을 정한다. 이전 단계에서 선택된 설명변수와 분리기준에 의해 생성된 노드들에 동일한 분류규칙을 반복함으로써 분류 규칙을 형성하게 된다. CART 알고리즘은 해석이 쉽다는 장점을 가지고 있지만 자료에 따라 생성된 분류규칙이 다양하여 불안정하다는 단점을 가지고 있다.

의사결정나무에서 모형을 구축하는 과정은 결정나무 구축(tree building)과 가지치기(prunning) 두 과정으로 나눌 수 있다. 불균형 자료를 분류할 때, 몇몇 결정나무 구축 과정에서 소수 집단을 판별하기 전에 종료가 될 수 있다. 또한, 분류에 대한 일반성을 잃지 않기 위해 소수 집단을 판별하

고려대학교 KOREA UNIVERSITY 1905

는 가지를 제거할 수 있다. 이러한 문제는 소수집단을 분류하는 과정이 모형의 오류를 줄이는 데 큰 기여를 못하기 때문이다. 따라서 이러한 과정으로 불균형 자료에서 의사결정나무를 사용하여 분류를 하는 데 문제가 발생하게 된다.

2.2. 불균형 문제 해결을 위한 방법

2.2.1. Tree Ensemble 방법

Tree Ensemble 방법은 기존의 original 자료로부터 다수의 분류규칙을 생성하여 이를 종합함으로써 더 나은 분류 예측력을 얻는다는 장점을 가지고 있다.

1. AdaBoost(1995)

부스팅 알고리즘 1995년에 Freund와 Schaprie이 제안한 방법론이다. 부스팅 알고리즘은 잘못 분류된 개체에 더 관심을 가지고 이를 더 잘 분류하도록 제안된 재표본 방법이다. 부스팅 알고리즘은 분류규칙을 순차적으로 생성하며, 이전의 분류규칙에 의해 분류된 자료를 분석용 자료로 이용하여 분류규칙을 적용해 나간다. 이 과정에서 가중치를 부여하게 되는데 가장 처음의 분석용 자료의 각 관측치에는 동일한 가중치를 부여하며, 분류규칙을 적용한 후 잘못 분류된 관측치에 높은 가중치를 부여한다. 분류규칙을 적용해 나가면서 가중치를 재조정하는 과정을 반복하게 된다. 이렇게 잘못 분류된 관측치에 높은 가중치를 부여하는 과정을 거치면서 부스팅 알고리즘은 분류하기 힘든 관측값을 더 잘 분류하도록 한다.

고려대학교 KOREA UNIVERSITY 1905

2. Random Forest(2001)

Random Forest 알고리즘은 2001년에 Brieman이 제안한 방법이다. 단일 의사결정나무를 사용하는 경우보다 정확성, 안정성 측면에서 더 나은 결과를 보이는 것으로 알려져 있는 방법이다. Random Forest 분류 규칙은 원 자료로부터 생성된 동일한 크기의 Bootstrap 표본들에 의해 만들어진 의사결정나무로 구성된다. 먼저, 원자료에 Bootstrap 표본을 추출하여 훈련용 자료로 뽑히지 않은 개체들을 테스트용 자료로 한다. 훈련용 자료에서 임의로 선택된 설명변수를 이용하여 나무를 최대한 확장시킨다. 이렇게 여러번 반복하여 만들어진 의사결정나무에 의한 결과를 결합하여 최종 분류 규칙을 생성하게 된다.

2.2.2. 자료 수준 접근 방법 - Resampling 방법

Resampling 방법은 자료 수준의 접근 방법으로 크게 소수 집단의 개체를 일정한 알고리즘을 이용하여 다수 집단과의 비율을 맞춰주는 Over-sampling 방법과 다수 집단의 개체를 일정한 알고리즘을 이용하여 소수 집단과의 비율을 맞춰주는 Under-sampling 방법이 있다. 본 연구에서는 대표적인 Over-sampling 방법인 SMOTE(2002) 방법을 보완한 Hybrid Resampling 방법을 사용하였다. Hybrid Resampling 방법 중 SMOTE+ENN, Borderline1-SMOTE 방법을 사용하였다.

1. SMOTE(Synthetic Minority Oversampling Technique) (2002)

SMOTE 방법은 ROS(Random OverSampling)을 응용한 Oversampling

방법으로 Chawla 외 3명이 2002년에 제안한 방법이다. 이 방법의 주요 아이디어는 소수 집단 개체 사이의 보간법(interpolating)을 이용하여 과대적합(over-fitting)을 피하고자 하는 것이다. SMOTE 방법의 구체적인 알고리즘은 다음과 같다.

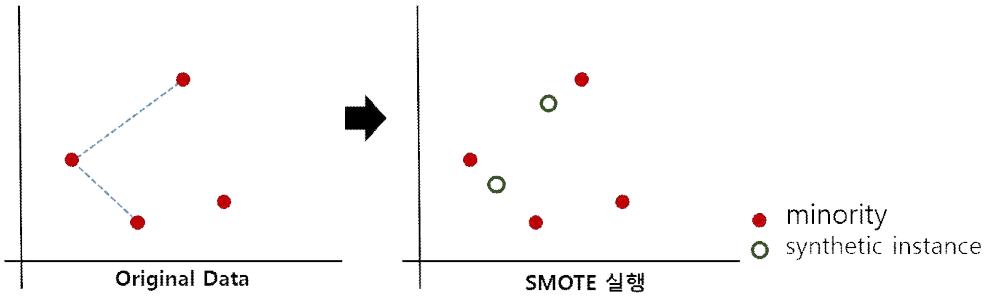
(단계 1) A는 소수 집단(Y=1), B는 다수 집단(Y=0)의 집합이라고 정의한다.

(단계 2) 집합 A에 속하는 개체 x에 대하여 k-NN(k nearest neighbor)을 계산한다.

(단계 3) 2에서 구한 x의 이웃(neighbors) 중 무작위로 r(≤k)개를 선택한다.

(단계 4) 3에서 선택된 개체와 x 사이의 차이(diff)를 계산한다.

(단계 5) 새롭게 생성된 synthetic point = x + u*diff (u는 0~1 사이의 임의의 수)가 된다.



[그림 2.1 SMOTE 방법]

2. Borderline1-SMOTE (2005)

Borderline1-SMOTE 방법은 SMOTE 방법으로 생성된 개체가 소수 집단의 개체와 유사하지 않을 수 있다는 점을 보완한 방법으로 Han 외 2명이 2005년에 제안하였다. 다수집단과 소수집단의 경계에 위치한 개체를 이

고려대학교 KOREA UNIVERSITY 1905

용하여 SMOTE를 적용한 방법으로 구체적인 알고리즘은 다음과 같다.

(단계 1) 소수 집합의 개체들의 집합을 P라고 정의한다.

(단계 2) P의 모든 p_i에 대해서 m-nearest neighbors 계산, m_i는 m개 중 다수 집단에 속하는 개체의 개수를 나타낸다.

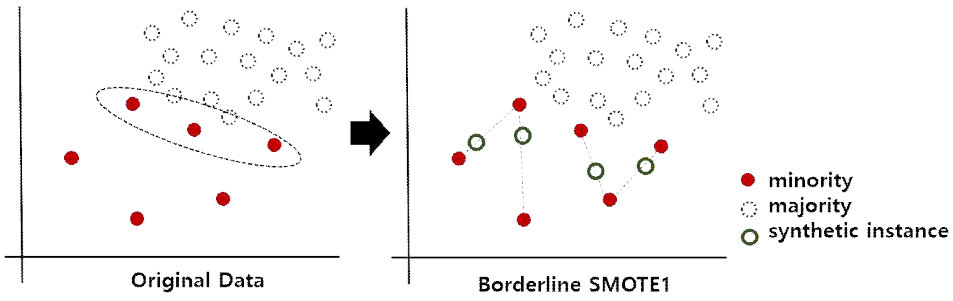
(단계 3-1) m_i=m 일 때, p_i는 잡음(noise)로 생각하고 이후의 과정 진행하지 않는다.

(단계 3-2) $\frac{m}{2} \leq m_i < m$ 일 때, p_i를 'DANGER'라는 집합에 포함시킨다.

(단계 3-3) $0 \leq m_i < \frac{m}{2}$ 일 때, p_i는 적절한 소수집단의 개체라고 생각하고 이후의 과정 진행하지 않는다.

(단계 4) DANGER에 속한 개체들과 P 집합에 속한 개체 사이의 SMOTE 진행한다.

synthetic point = x + u*diff (u는 0~1 사이의 임의의 수)

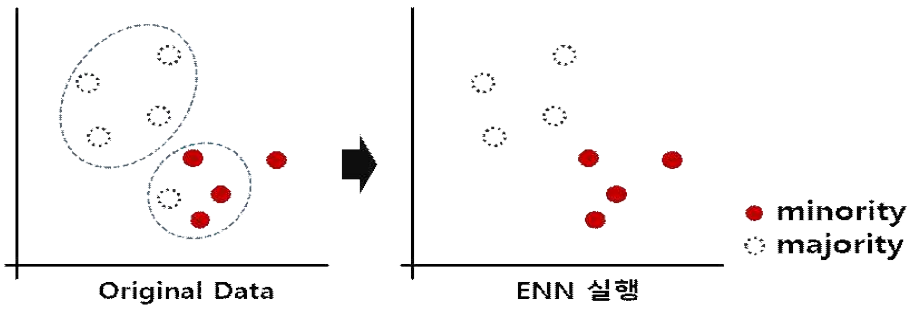


[그림 2.2] Borderline1 SMOTE 방법

3. SMOTE + ENN

먼저 자료에 SMOTE 방법을 적용한 후, ENN(Wilson, 1972)를 이용하여 두 집단의 경계에 있는 개체를 제거하는 방법이다. 기존 ENN방법은 다수 집단의 개체만을 제거했지만 SMOTE 방법을 적용한 후에는 두 집단의 개체 모두 제거한다.

ENN(Wilson's Edited Nearest Neighbor Rule)방법은 다수 집단의 개체를 제거하는 방법이다. 개체의 k-NN을 계산한 후, k개의 Nearest Neighbors 중에 해당 개체와 다른 집단에 속하는 개체가 반 이상인 경우에 해당 개체를 제거한다. 예를 들어, 3개의 Nearest Neighbor 중 2개 이상의 개체가 해당 개체와 다른 집단에 속하는 경우 해당 개체를 분석 자료에서 제거한다.



[그림 2.3] ENN 방법

2.2.3. Cost-sensitive learning - 가중치(Weight) 부여

각 집단에서 잘못 분류 되었을 때의 손실의 크기를 비대칭적으로 부여하는 방법이다. 불균형 자료의 연구에서는 소수집단을 제대로 분류하는 것이 주 관심사이므로 소수 집단을 잘못 분류 했을 경우의 손실을 다수집단의 손실보다 크게 부여하였다. 본 연구에서는 두 집단의 크기를 고려하여 가

고려대학교 KOREA UNIVERSITY 1905

중치를 상대 집단의 크기에 비례하는 값으로 부여하였다.

<표 2.1> 가중치

	그룹 개체 수	가중치(Weight)
그룹 1(다수집단)	n1	n2/(n1+n2)
그룹 2(소수집단)	n2	n1/(n1+n2)

2.3. 성능 평가 기준

분류 문제에서 성능 평가 기준은 적합된 모형의 분류력을 평가하거나 적절한 분류 모형을 세우는 데에서 중요한 역할을 한다. 일반적으로 분류 문제에서 이러한 목적을 만족하기 위해 오분류율(Misclassification rate)을 사용한다. 하지만, 오분류율은 불균형 자료의 분류에서 소수집단의 분류를 적절하게 고려하지 못한다. 예를 들어, 전체 자료 중 소수 집단이 차지하는 비율이 1%일 경우, 모든 관측치를 다수집단으로 분류하더라도 오분류율은 1%가 된다. 따라서 이를 보완한 다른 평가 기준이 필요하다.

<표 2.2> Confusion Matrix

실제 \ 예측	Y=1	Y=0
Y=1	True Positive(TP)	False Negative(FN)
Y=0	False Positive(FP)	True Negative(TN)

- Sensitivity(민감도) = $\frac{TP}{TP+FN}$
- Specificity(특이도) = $\frac{TN}{FP+TN}$

고려대학교 KOREA UNIVERSITY 1905

- Precision = $\frac{TP}{TP+FP}$

2.3.1. Misclassification Rate (오분류율)

오분류율 = $\frac{FP+FN}{Total}$, $Total = TP+FP+FN+TN$

2.3.2. F-measure

소수 집단에 대한 분류력에 대해 평가할 경우, Sensitivity(민감도)와 Precision이 중요하게 고려되어야 한다. 따라서 이 두 가지 값을 적절하게 고려한 통계량이 F-measure이다. Sensitivity와 Precision이 모두 적절하게 높은 값을 갖는다면, F-measure 값 또한 높은 값을 갖게 된다.

F-measure = $\frac{2*Sensitivity \times Precision}{Sensitivity + Precision}$

2.3.3. AUC

AUC는 ROC(Receiver Operating Characteristics) 곡선의 아래쪽 면적을 나타내는 값으로 ROC 곡선을 단일한 통계량으로 나타낸 것이다. AUC를 이용하여 모형의 전반적인 분류력을 평가할 수 있다.

본 연구에서는 F-measure와 AUC를 평가 기준으로 사용하였다.

제 3 장
모의실험

앞에서 언급한 Tree Ensemble 방법, 가중치 방법, Resampling 방법을 모의실험을 통해 비교하고자 한다. 먼저, 분류를 위해 의사결정나무(Decision Tree)를 이용하였다. 본 논문에서 불균형 자료에서의 분류 성능 비교를 위해 사용된 방법은 기존의 자료 이용, 자료에 가중치(Weights) 부여, Boosting, Random Forest, SMOTE+ENN, Borderline1의 6가지 방법이다.

비교에 사용된 방법은 각각 다음과 같이 표현하고자 한다.

1. Tree : 기존의 자료에 의사결정나무 이용
2. WTree : 가중치(Weights)를 부여한 의사결정나무 이용
3. Boost : AdaBoost 이용
4. RF : Random Forest 이용
5. ENN : SMOTE+ENN를 적용한 자료에 의사결정나무 이용
6. BL1 : Borderline1 SMOTE를 적용한 자료에 의사결정나무 이용

방법 비교를 위해 분류력 평가 기준으로는 앞에서 제시한 2가지 통계량을 사용하였다.

1) AUC
2) F-measure

불균형 자료에서의 분류향상을 위한 방법 비교를 위해서 다음과 같은 두 가지 자료를 설정하였다.

Ⅰ. 이변수 정규 분포(Bivariate Normal Distribution) 자료

이변수 정규 분포를 이용하여 모의실험 자료를 생성하였다.

$$Group1 : \begin{pmatrix} X_{11} \\ X_{12} \end{pmatrix} \sim BN\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{11} & 0 \\ 0 & \sigma_{12} \end{pmatrix}\right]$$

Ⅱ. 원(Circle) 자료

균일 분포[0,1]를 이용하여 생성한 자료를 이용하여 소수집단이 다수집단에 둘러싸인 원형 형태의 모의 자료를 설정하였다. 자료 생성에 관한 구체적인 단계는 다음과 같다.

(단계 1) 두 집단의 비율을 맞추기 위해 균일 분포[0,1]에서 1000개의 난수 생성

(단계 2-1) 1에서 생성한 난수 중 소수집단의 비율보다 작은 수를 갖는 경우, 중심이 (0.2, 1.2)이고 반지름이 0.35인 원 내부에 속하는 임의의 개체를 생성하고 소수 집단으로 분류

(단계 2-2) 1에서 생성한 난수 중 소수집단의 비율보다 큰 수를 갖는 경우, 균일 분포[0,1]에서 개체를 생성

(단계 3-1) 2-2에서 생성한 개체와 원의 중심 사이의 거리가 (반지름+overlap) 이상이면 다수 집단으로 분류

(단계 3-2) 2-2에서 생성한 개체와 원의 중심 사이의 거리가 (반지름+overlap) 미만이면 균일 분포[0,1]에서 다시 개체를 생성하여 원의 중심 사이의 거리가 (반지름+overlap) 이상을 만족하는 개체 생성

(단계 4) 3에서 생성한 다수집단 개체 중에 out 비율에 맞춰 소수집단의 이상치로 배정

위와 같은 과정에서 두 집단의 경계에서 두 집단의 겹치는 정도를 조정

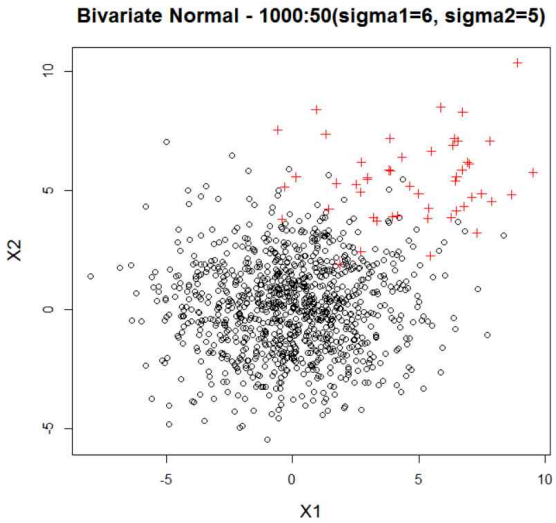
고려대학교 KOREA UNIVERSITY 1905

하기 위해 겹치는 구간(overlap)과 이상치 비율(out)을 설정하였다. 예를 들어, 겹치는 구간이 -0.1인 경우 원의 중심에서 0.25(=0.35-0.1) 떨어진 곳에 다수집단의 개체가 소수집단의 개체와 overlap 되도록 하는 자료이다.

3.1. 모의실험 계획

비교를 위한 자료는 불균형 자료에서 분류에 영향을 주는 두 집단의 비율, overlap 정도를 변화시켜 자료의 특성을 다르게 하였다.

먼저, 각 방법의 분류력 비교를 위해 이변수정규분포(Bivariate Normal Distribution)에서 모의실험 자료를 생성하였다. 다수 집단 대비 소수 집단의 비율을 1%, 5%, 10%로 조정하였고 각 집단의 분산은 아래와 같이 조정하였다. 두 집단의 평균은 다수 집단은 (0, 0), 소수 집단은 (5, 5)로 고정하였다. 모의 자료의 상황은 다음과 같다.



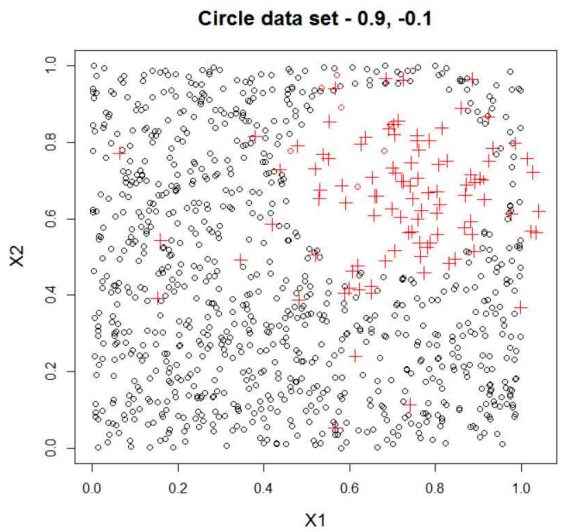
[그림 3.1] 정규 분포 자료

고려대학교 KOREA UNIVERSITY 1905

<표 3.1> 정규분포 모의실험 자료 상황

시나리오		
두 집단의 분산-공분산 행렬 집단1(다수 집단)	집단2(소수 집단)	두 집단의 비 (소수집단 비율)
$\begin{pmatrix}2&0\\0&2\end{pmatrix}$	$\begin{pmatrix}1&0\\0&1\end{pmatrix}$	10 : 990 (1%) 50 : 950 (5%) 100 : 900 (10%)
$\begin{pmatrix}6&0\\0&6\end{pmatrix}$	$\begin{pmatrix}5&0\\0&5\end{pmatrix}$	10 : 990 (1%) 50 : 950 (5%) 100 : 900 (10%)
$\begin{pmatrix}10&0\\0&10\end{pmatrix}$	$\begin{pmatrix}9&0\\0&9\end{pmatrix}$	10 : 990 (1%) 50 : 950 (5%) 100 : 900 (10%)

두 번째 모의실험 자료는 다수집단(Y=0)이 원형을 이루는 소수집단(Y=1) 주변에 분포하는 모의실험 자료이다. 각 집단의 비율, 소수 집단과 다수 집단 사이의 경계의 겹침 정도, 이상치의 비율을 고려한 상황은 아래와 같다.



[그림 3.2] 원(circle) 자료

고려대학교 KOREA UNIVERSITY 1905

<표 3.2> 원(circle) 자료 모의실험 상황

시나리오	
소수 집단의 비율	1 % 5 % 10 %
두 집단의 경계 겹침 구간 (overlap)	-0.1 -0.2 -0.3
이상치 비율(out)	0.01 0.05

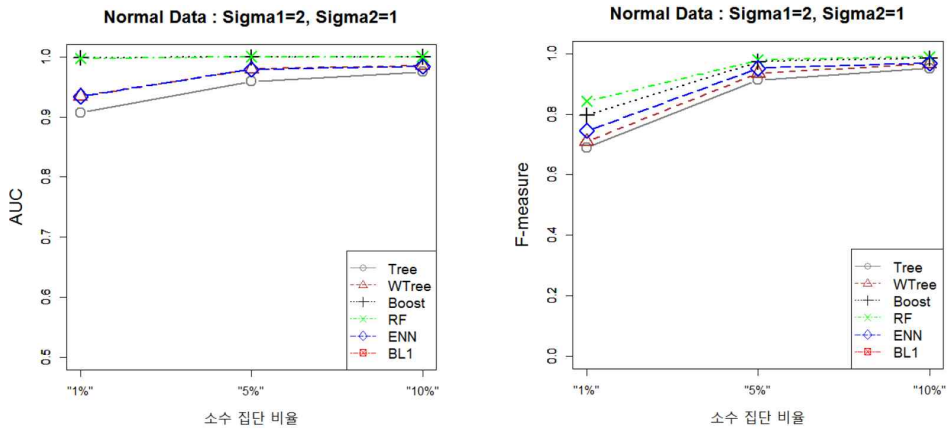
모의실험에서 각 방법들의 분류 예측력 비교를 위해 5-fold Cross Validation(CV)을 사용하였다. 가중치 방법과 자료 수준의 접근 방법의 경우, 훈련용 자료에 자료 수준의 접근 방법 또는 가중치 방법을 적용한 후 의사결정나무 알고리즘을 이용하여 분류 규칙을 결정한 후 테스트용 자료를 이용하여 앞서 구한 분류 규칙을 적용하여 제시한 성능 평가 통계량을 구하였다. 자료 수준의 접근 방법인 Resampling 방법에서 거리를 구하는 방법으로 Euclidean Distance를 사용하였으며, 각 실험은 500번 반복을 하였고 분류 성능 평가 기준값은 500번 반복의 평균값을 사용하였다.

고려대학교 KOREA UNIVERSITY 1905

3.2. 모의실험 결과

3.2.1. 정규 분포 이용

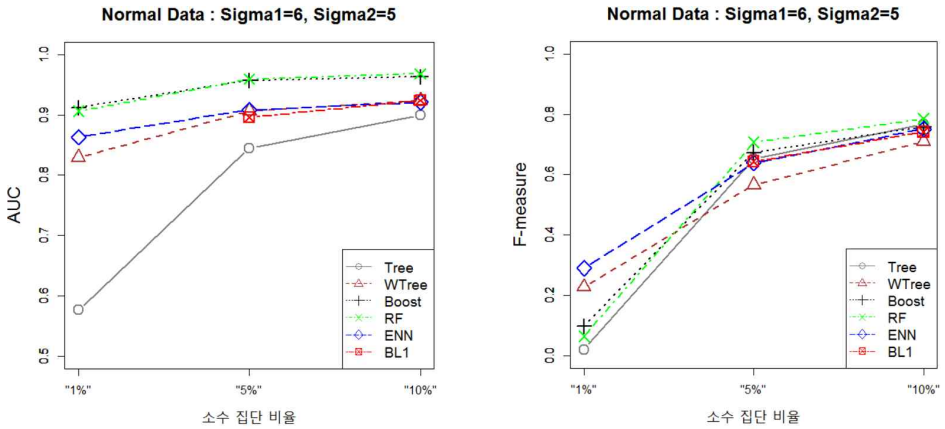
<표 3.3>의 결과를 다음과 같은 그래프로 나타내었다.



[그림 3.3] 정규분포자료(다수집단 분산=2, 소수집단 분산=1)

[그림 3.3]에 나타난 결과는 각 집단의 분산-공분산 행렬이 $\begin{pmatrix}2&0\\0&2\end{pmatrix}$, $\begin{pmatrix}1&0\\0&1\end{pmatrix}$으로 분산이 가장 작은 경우이다. 비교적 분류가 까다롭지 않은 상황이기 때문에 각 방법의 분류 성능이 큰 차이를 보이지 않음을 확인할 수 있으며 소수 집단의 비율이 커짐에 따라 각 방법의 비슷한 성능을 갖는 것을 확인할 수 있다. 각 방법을 자세히 살펴보면, Random Forest 방법의 성능이 가장 좋으며 Tree 방법이 가장 낮은 성능을 갖는다. 가중치 방법과 자료수준의 접근 방법인 SMOTE+ENN 방법이 비슷한 성능을 갖는다.

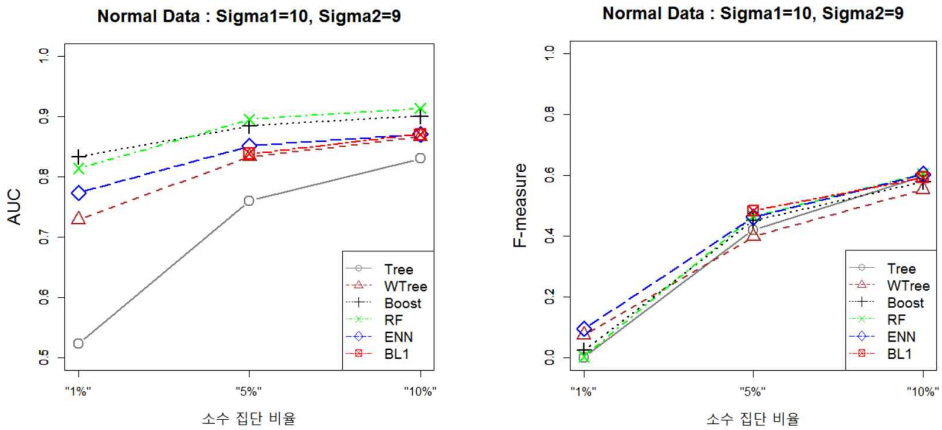
고려대학교 KOREA UNIVERSITY 1905



[그림 3.4] 정규분포자료(다수집단 분산=6, 소수집단 분산=5)

다음으로 [그림 3.4]에 나타난 결과는 각 집단의 분산-공분산 행렬이 각각 $\begin{pmatrix}6&0\\0&6\end{pmatrix}$, $\begin{pmatrix}5&0\\0&5\end{pmatrix}$에 해당하는 경우이다. 앞의 결과에 비해 각 방법사이의 분류 성능 차이가 확연하게 나타나고 있다. 분산이 커지면서 의사결정나무의 성능이 다른 방법에 비해 크게 감소하였으며 Boosting 방법과 Random Forest 방법의 AUC가 가장 높은 값을 갖으며 소수집단의 비율이 5%, 10%인 경우에는 Random Forest의 분류 성능이 가장 높은 것을 확인할 수 있다. 자료수준 접근방법과 가중치 방법을 비교해 보면, 두 방법 모두 의사결정나무만 사용한 경우에 비해 높은 분류 성능을 보이고 있으며 그 중, SMOTE+ENN 방법이 가장 좋은 분류성능을 보이고 있다. 두 집단의 비가 커짐에 따라 가중치 방법과 자료 수준의 접근방법 모두 비슷한 분류 성능을 보이고 있다.

고려대학교 KOREA UNIVERSITY 1905



[그림 3.5] 정규분포자료(다수집단 분산=10, 소수집단 분산=9)

마지막으로 [그림 3.5]에 나타난 결과는 각 집단의 분산-공분산 행렬이 $\begin{pmatrix}10 & 0\\ 0 & 10\end{pmatrix}$, $\begin{pmatrix}9 & 0\\ 0 & 9\end{pmatrix}$으로 분산이 가장 큰 경우이다. 왼쪽의 AUC 값 그래프는 분산-공분산 행렬이 각각 $\begin{pmatrix}6 & 0\\ 0 & 6\end{pmatrix}$, $\begin{pmatrix}5 & 0\\ 0 & 5\end{pmatrix}$에 해당하는 경우와 방법 사이의 비슷한 경향을 보이고 있다. F-measure 값을 나타낸 그래프를 보면 각 방법이 비슷한 값을 갖으며, 자료 수준의 접근 방법이 Random Forest와 Boosting 방법과 비슷하거나 더 높은 값을 갖는 것을 확인할 수 있다. 가중치 방법과 자료 수준의 접근 방법을 비교해보면 자료 수준의 접근 방법 중 SMOTE+ENN 방법이 가장 좋은 분류 성능을 보이고 있다.

정규분포의 자료의 모의실험 결과, 전체적인 추세를 보면 방법 사이의 감소폭의 차이가 존재하지만, 두 집단의 분산이 커질수록 모든 방법의 AUC, F-measure 값이 작아지는 것을 확인할 수 있다. 소수 집단의 비율이 작아질수록 방법의 분류력이 감소하는 추세를 보이지만 분산이 작을 경우 AUC, F-measure의 감소폭이 다른 경우에 비해 크지 않음을 알 수 확

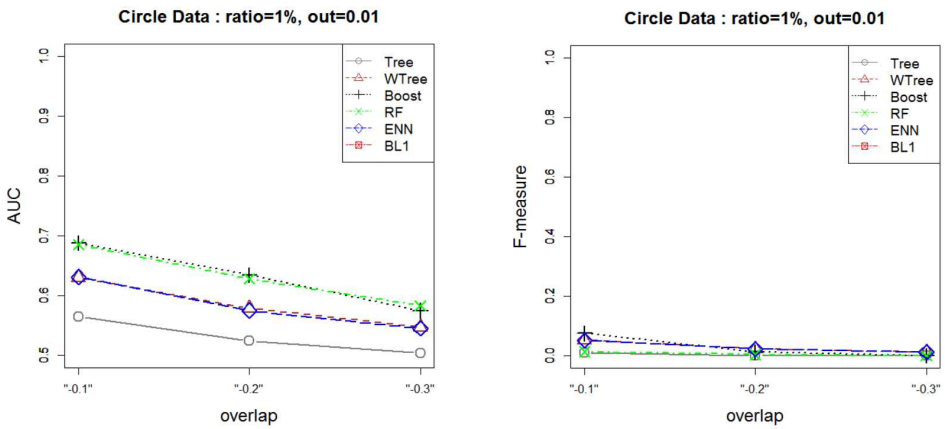
고려대학교 KOREA UNIVERSITY 1905

인할 수 있다.

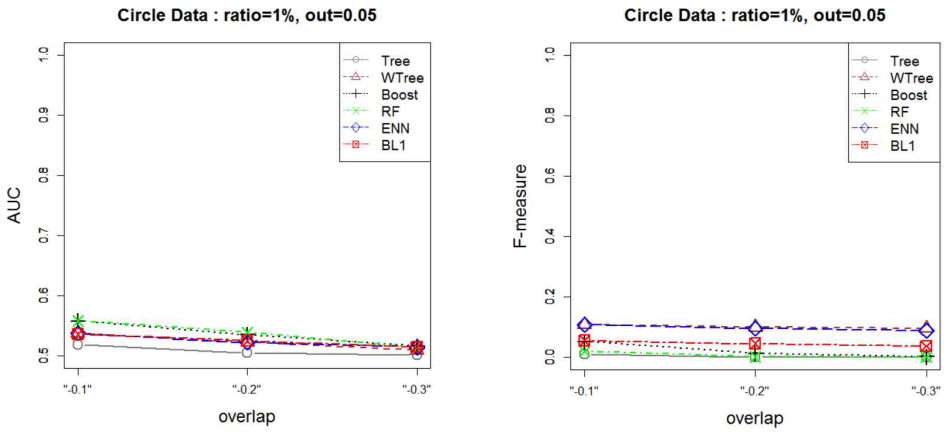
분산이 작은 경우에는 각 방법이 비슷한 성능을 보이지만 분산이 커짐에 따라 의사결정나무(Tree)의 분류 성능이 확연하게 낮아짐을 확인할 수 있다. 또한, Boosting과 Random Forest 방법이 대부분의 경우에서 가장 높은 성능을 보이고 있으며 자료 수준의 접근 방법 중 SMOTE+ENN의 성능이 가중치 방법에 비해 높은 성능을 보이고 있다. 소수집단의 비율이 커지면서, Boosting과 Random Forest 방법의 성능이 거의 비슷해지고 자료 수준의 접근 방법과 가중치 방법의 성능이 거의 비슷해지는 것을 확인할 수 있다.

3.2.2. 원(circle) 자료 이용

<표 3.4>, <표 3.5>, <표 3.6>의 결과는 소수집단의 비율이 각각 1%, 5%, 10%에 해당하는 결과를 나타내며 표의 결과를 다음과 같은 그래프로 나타내었다.

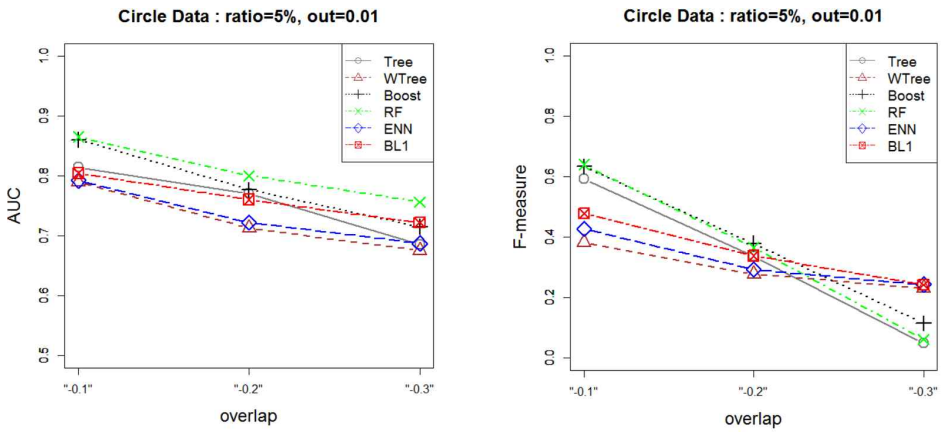


[그림 3.6] 원자료(소수집단 비율=1%, 이상치 비율=0.01)



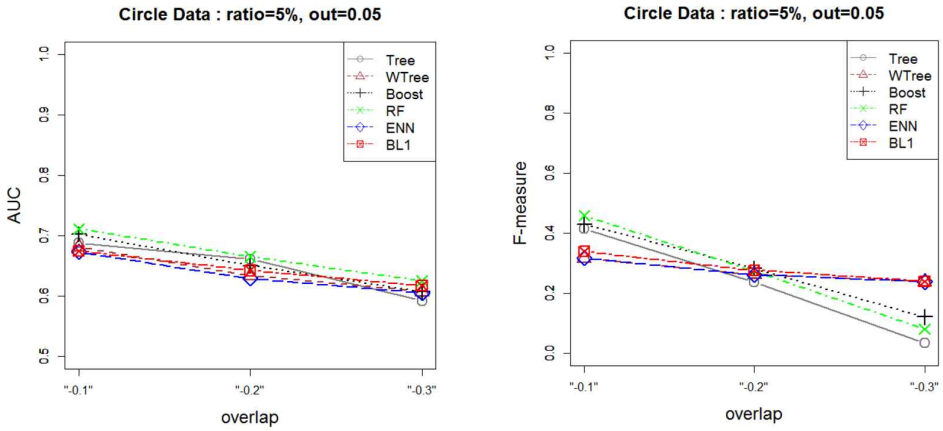
[그림 3.7] 원자료(소수집단 비율=1%, 이상치 비율=0.05)

위의 결과 [그림 3.6]과 [그림 3.7]은 소수 집단의 비율이 1%인 경우의 결과이다. [그림 3.6]은 이상치 비율(out)이 0.01, [그림 3.7]은 0.05에 해당하는 결과이다. 이상치의 비율이 0.01에서 0.05로 증가하면서 모든 방법의 분류 성능이 전반적으로 감소한 것을 확인할 수 있다. AUC 값 기준으로 Boosting과 Random Forest 방법이 항상 높은 값을 갖는다. 가중치 방법과 자료 수준의 접근 방법은 비슷한 AUC 값을 갖는 것을 확인할 수 있다.



[그림 3.8] 원자료(소수집단 비율=5%, 이상치 비율=0.01)

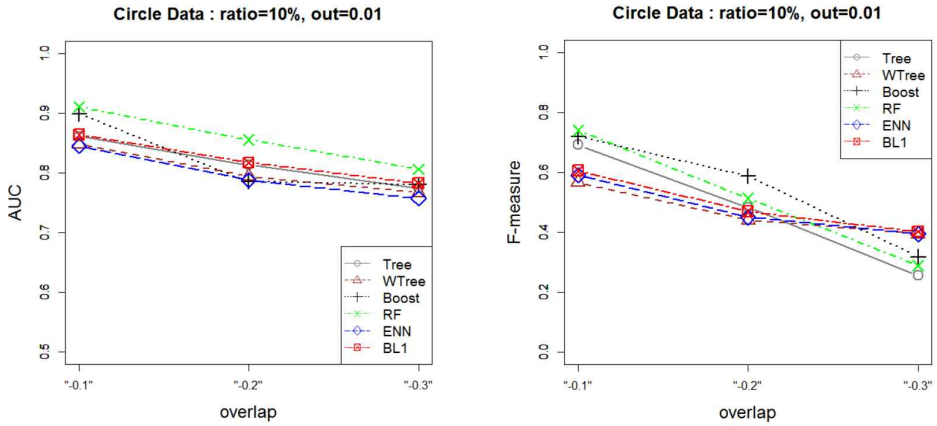
고려대학교 KOREA UNIVERSITY 1905



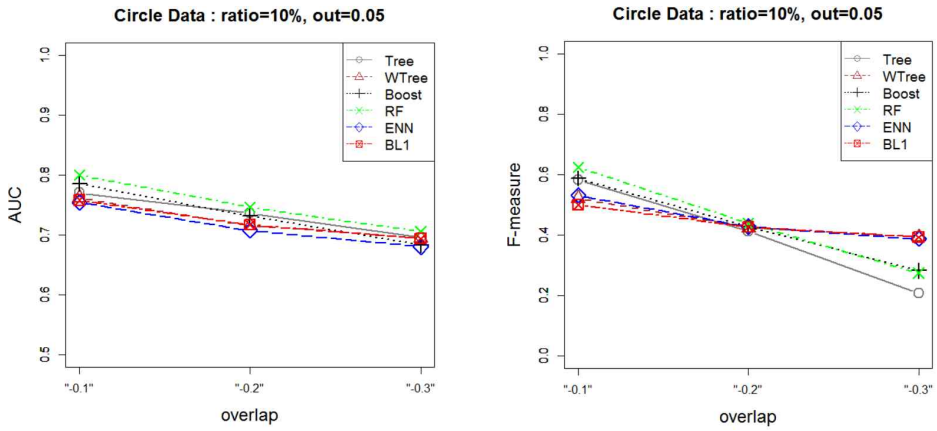
[그림 3.9] 원자료(소수집단 비율=5%, 이상치 비율=0.05)

위의 결과 [그림 3.8]과 [그림 3.9]는 다수 집단의 비율이 5%인 경우의 결과이다. [그림 3.8]은 이상치의 비율이 0.01, [그림 3.9]는 이상치의 비율이 0.05인 경우의 결과이다. 먼저, 두 집단의 이상치 비율(out)이 0.01에서 0.05로 증가하면서 모든 방법에서 AUC 값과 F-measure 값이 감소했다. 두 집단의 겹치는 구간(Overlapping)이 증가하면서 의사결정 나무, Boosting과 Random Forest 방법의 AUC값과 F-measure값이 다른 방법에 비해 큰 감소폭을 갖는다. 따라서, 겹치는 정도가 작은 -0.1~-0.2 에서는 Boosting과 Random Forest 방법이 가장 좋은 분류 성능을 보이지만 겹치는 정도가 -0.3이 되었을 경우, 가중치 사용 방법과 자료 수준의 접근 방법이 Boosting과 의사결정나무 보다 좋은 성능을 보였다. 가중치 방법과 자료 수준의 방법을 비교해보면 Borderline1 SMOTE 방법이 비교적 가장 좋은 성능을 보였으며 가중치 방법과 SMOTE+ENN 방법은 비슷한 성능을 보이고 있다.

고려대학교 KOREA UNIVERSITY 1905



[그림 3.10] 원자료(소수집단 비율=10%, 이상치 비율=0.01)



[그림 3.11] 원자료(소수집단 비율=10%, 이상치 비율=0.05)

위의 결과 [그림 3.10]과 [그림 3.11]은 다수 집단의 비율이 0.99인 경우의 결과이다. [그림 3.10]은 이상치 비율(out)이 0.01, [그림 3.11]은 0.05에 해당하는 결과이다. 앞의 소수 집단이 5%인 경우와 각 방법 사이의 비슷한 경향을 확인할 수 있다. 앞의 결과와 마찬가지로 겹치는 구간이 -0.1~-0.2인 경우 Boosting과 Random Forest 방법의 성능이 가장 좋았고 -0.3인 경우, 가중치 방법과 자료 수준의 접근 방법이 비슷하게 좋은 성능을 보이고 있다.

고려대학교 KOREA UNIVERSITY 1905

원(circle) 자료의 모의실험 결과, 전반적인 추세를 보면 소수 집단의 비율이 감소하고 두 집단 사이의 겹침 정도, 이상치의 비율이 커짐에 따라 각 방법의 분류 성능 감소하고 있었다. 대부분의 상황에서 Boosting과 Random Forest 방법이 가장 좋은 분류 성능을 보이고 있었다.

소수 집단의 비율이 커지면서 AUC값 기준으로 각 방법의 분류 성능이 비슷했으며 F-measure값 기준으로 가중치 방법과 자료 수준의 접근 방법의 성능이 유사했다. 두 집단의 겹치는 구간(overlap)이 넓어지면서 의사결정나무, Boosting과 Random Forest 방법이 가중치 방법과 자료 수준의 접근 방법에 비해 큰 성능 감소폭을 보였다. 그 결과, 겹치는 구간(overlap)이 넓어지면서 가중치 방법과 자료 수준의 접근 방법이 다른 방법에 비해 좋은 성능을 보였다. 가중치 방법과 자료수준 접근 방법을 비교해 보면 각 상황에서 거의 비슷한 추세를 보이고 있었지만, 그 중 Borderline1 SMOTE 방법이 가장 좋은 분류 성능을 보였다.

대부분의 경우에서 자료 수준의 접근 방법이 가중치 방법보다 좋은 성능을 보였고 분산이 크거나 겹치는 구간이 커지는 경우에 다른 방법에 비해 좋은 분류 성능을 확인할 수 있었다. 자료 수준의 접근 방법 중 SMOTE+ENN 방법이 정규분포를 이용한 모의실험 자료에서 좋은 성능을 보였으며 원(circle) 자료에서는 Borderline1 SMOTE 방법이 좋은 성능을 보였다.

고려대학교 KOREA UNIVERSITY 1905

<표 3.3> 정규 분포 실험 결과

그룹 분산		소수 집단 비율	Tree		Weighted Tree		Boost		Random Forest		SMOTE+ ENN		Borderline 1 SMOTE	
다수 집단	소수 집단		AUC	F-measure	AUC	F-measure	AUC	F-measure	AUC	F-measure	AUC	F-measure	AUC	F-measure
$\begin{pmatrix}2&0\\0&2\end{pmatrix}$	$\begin{pmatrix}1&0\\0&1\end{pmatrix}$	1%	0.907	0.689	0.934	0.709	0.998	0.796	0.997	0.843	0.934	0.745	.	.
		5%	0.959	0.913	0.980	0.934	0.999	0.973	1.000	0.978	0.979	0.951	.	.
		10%	0.975	0.951	0.985	0.967	1.000	0.985	1.000	0.989	0.984	0.968	.	.
$\begin{pmatrix}6&0\\0&6\end{pmatrix}$	$\begin{pmatrix}5&0\\0&5\end{pmatrix}$	1%	0.577	0.021	0.829	0.228	0.912	0.098	0.906	0.065	0.863	0.290	.	.
		5%	0.845	0.652	0.906	0.567	0.957	0.673	0.959	0.707	0.908	0.639	0.896	0.644
		10%	0.899	0.767	0.923	0.710	0.963	0.758	0.968	0.785	0.921	0.751	0.924	0.741
$\begin{pmatrix}10&0\\0&10\end{pmatrix}$	$\begin{pmatrix}9&0\\0&9\end{pmatrix}$	1%	0.523	0.001	0.729	0.075	0.833	0.024	0.814	0.004	0.773	0.095	.	.
		5%	0.760	0.420	0.833	0.398	0.884	0.452	0.895	0.465	0.851	0.462	0.838	0.482
		10%	0.830	0.597	0.866	0.552	0.900	0.578	0.913	0.605	0.870	0.603	0.870	0.593

고려대학교 KOREA UNIVERSITY 1905

<표 3.4> 원(circle) 자료 실험 결과 - 소수집단비율 : 1%

Ratio	Out	Overlap	Tree		Weighted Tree		Boost		Random Forest		SMOTE + ENN		Borderline1 SMOTE	
			AUC	F-measure	AUC	F-measure	AUC	F-measure	AUC	F-measure	AUC	F-measure	AUC	F-measure
1%	0.01	-0.1	0.565	0.01	0.63	0.052	0.688	0.075	0.685	0.014	0.631	0.052	.	.
		-0.2	0.524	0.001	0.579	0.023	0.634	0.013	0.628	0.004	0.575	0.023	.	.
		-0.3	0.504	0	0.547	0.013	0.575	0	0.583	0.000	0.545	0.012	.	.
	0.05	-0.1	0.519	0.01	0.537	0.108	0.558	0.052	0.558	0.020	0.538	0.107	0.536	0.055
		-0.2	0.505	0	0.523	0.1	0.535	0.013	0.539	0.002	0.522	0.095	0.525	0.045
		-0.3	0.501	0	0.51	0.096	0.516	0.002	0.514	0.000	0.514	0.088	0.515	0.037

고려대학교 KOREA UNIVERSITY 1905

<표 3.5> 원(circle) 자료 실험 결과 - 소수집단비율 : 5%

Ratio	Out	Overlap	Tree		Weighted Tree		Boost		Random Forest		SMOTE + ENN		Borderline1 SMOTE	
			AUC	F-measure	AUC	F-measure	AUC	F-measure	AUC	F-measure	AUC	F-measure	AUC	F-measure
5%	0.01	-0.1	0.814	0.593	0.789	0.382	0.86	0.634	0.865	0.640	0.792	0.427	0.804	0.478
		-0.2	0.77	0.336	0.713	0.277	0.777	0.379	0.800	0.365	0.722	0.292	0.76	0.338
		-0.3	0.685	0.048	0.675	0.231	0.715	0.115	0.756	0.062	0.687	0.243	0.722	0.242
	0.05	-0.1	0.688	0.414	0.681	0.316	0.702	0.429	0.711	0.458	0.673	0.317	0.674	0.339
		-0.2	0.661	0.237	0.634	0.262	0.65	0.283	0.665	0.274	0.629	0.261	0.642	0.276
		-0.3	0.592	0.034	0.607	0.239	0.607	0.121	0.625	0.081	0.605	0.238	0.617	0.239

고려대학교 KOREA UNIVERSITY 1905

<표 3.6> 원(circle) 자료 실험 결과 - 소수집단비율 : 10%

Ratio	Out	Overlap	Tree		Weighted Tree		Boost		Random Forest		SMOTE + ENN		Borderline1 SMOTE	
			AUC	F-measure	AUC	F-measure	AUC	F-measure	AUC	F-measure	AUC	F-measure	AUC	F-measure
10%	0.01	-0.1	0.861	0.691	0.848	0.567	0.899	0.719	0.911	0.740	0.845	0.59	0.864	0.606
		-0.2	0.813	0.483	0.794	0.44	0.785	0.587	0.855	0.513	0.788	0.45	0.817	0.469
		-0.3	0.774	0.255	0.767	0.395	0.78	0.318	0.806	0.289	0.757	0.394	0.783	0.402
	0.05	-0.1	0.77	0.582	0.762	0.52	0.785	0.587	0.800	0.624	0.754	0.532	0.758	0.5
		-0.2	0.736	0.412	0.715	0.424	0.731	0.427	0.746	0.438	0.707	0.428	0.717	0.426
		-0.3	0.696	0.208	0.694	0.393	0.683	0.282	0.706	0.274	0.68	0.387	0.694	0.393

고려대학교 KOREA UNIVERSITY 1905

제 4 장

실제 자료 분석

실제 다양한 분야에서 불균형 자료가 존재하며 기존의 알고리즘이 소수 집단을 잘 분류하지 못하는 문제가 발생하게 된다. 따라서 실제 자료에서 앞서 제시한 방법들을 적용해보았다. 자료는 UCI 데이터 저장소에서 제공하는 '단백질의 지방화 위치'를 예측을 위한 2가지 자료(Yeast, Ecoli)를 이용하였다.

각 방법의 분류 예측력 평가를 위하여 자료를 이용하여 평가 방법으로 ROC 곡선을 사용하기 위해 약 2:1 비율의 train set: test set으로 나누었다. 각 방법의 통계량을 구하기 위해서 CV의 fold를 3으로 지정한 것 외의 세팅은 앞의 모의실험과 동일하다.

4.1. Yeast4 자료

Yeast4 자료는 Yeast자료에서 'ME1'을 목표 집단(소수집단)으로 설정한 자료이다. 자료의 특성은 <표 5.1>과 같다.

<표 4.1> Yeast4 자료

N	변수 개수	Negative (Y=0)	Positive (Y=1)	#(Positive)/N
1484	8	1433	51	3.44%

앞서 제시한 방법들을 Yeast4 자료에 적용한 결과는 아래의 <표 5.2>와 같다. 전반적인 방법의 분류 성능이 의사결정나무의 분류 성능에 비해 향상되었다. Random Forest 방법의 AUC가 가장 높으며 F-measure는

Borderline1 SMOTE 방법에서 가장 높은 값을 갖는다. 가중치 방법에 비해 자료수준의 접근 방법인 SMOTE+ENN 방법과 Borderline1 SMOTE 방법의 AUC값과 F-measure 값이 높은 것을 확인할 수 있다.

<표 4.2> Yeast4 자료 결과

	Tree	WTree	Boost	RF	ENN	BL1
AUC	0.654	0.733	0.804	0.861	0.833	0.811
F-measure	0.084	0.183	0.205	0.111	0.142	0.355

4.2. Ecoli4 자료

Ecoli4자료는 Ecoli 자료에서 'om'을 목표 집단(소수집단)으로 설정한 자료이다. 자료의 특성은 <표 5.3>과 같다.

<표 4.3> Ecoli4 자료

N	변수 개수	Negative (Y=0)	Positive (Y=1)	#(Positive)/N
336	7	316	20	5.95%

앞서 제시한 방법들을 Ecoli4 자료에 적용한 결과는 아래의 <표 5.4>와 같다. AUC값 기준으로 전반적인 방법의 분류 성능이 의사결정나무의 분류 성능에 비해 향상되었다. Random Forest 방법의 분류 성능이 가장 높으며 그 다음으로 Boosting, Borderline1-SMOTE 방법 순으로 높은 분류 성능을 갖는다.

<표 4.4> Ecoli4 자료 결과

	Tree	WTree	Boost	RF	ENN	BL1
AUC	0.883	0.894	0.978	0.989	0.906	0.914
F-measure	0.708	0.567	0.708	0.833	0.494	0.773

제 5 장

결론

불균형 자료는 다양한 분야에서 빈번하게 발생하는 자료로 이러한 자료를 다루기 위한 많은 연구들이 진행되어왔다. 본 논문에서는 불균형 비율, 두 집단의 겹침 정도 등의 다양한 상황에서 이전에 제안된 가중치 방법, Tree Ensemble 방법, 자료 수준의 접근 방법의 비교에 대해 다루었다. 분류 알고리즘과 독립적이기 때문에 분류 알고리즘에 따라 해석이 가능한 가중치 방법과 자료 수준의 접근 방법과 모형은 복잡하지만 높은 분류 성능을 갖는 Tree Ensemble 방법 사이의 비교를 진행하였다. 비교 결과를 종합하면 다음과 같다.

대부분의 상황에서 Tree Ensemble 방법인 Boosting과 Random Forest 방법이 전반적으로 좋은 분류 성능을 보였다. 하지만, 이상치가 증가하거나 두 집단의 겹치는 구간이 증가함에 따라 가중치 방법 또는 자료 수준의 접근 방법과 비슷하거나 경우에 따라 더 좋지 않은 분류 성능을 보이는 경우도 있었다.

자료의 분산이 커지거나 겹치는 구간이 커지는 등의 자료의 분류가 비교적 어려운 경우에 자료 수준의 접근 방법이 좋은 분류 성능을 보였으며 경우에 따라 Tree Ensemble 방법보다 좋은 분류 성능을 갖는 경우도 있었다. 자료 수준의 접근 방법과 가중치 방법을 비교해보면, 자료 수준의 접근 방법이 가중치 방법에 비해 비슷하거나 더 좋은 분류 성능을 보였다.

Tree Ensemble 방법의 경우, 대부분의 경우에서 다른 방법에 비해 좋은 분류 성능을 보였다. 하지만, 모형이 비교적 복잡하여 해석이 필요한 경우 해석에 어려움이 있다는 단점을 갖는다. 자료 수준의 접근 방법의 경우, 자

고려대학교 KOREA UNIVERSITY 1905

료의 분류가 어려울수록 좋은 분류 성능을 갖는다. 또한, 분류 알고리즘과 독립적이기 때문에 분류 알고리즘의 선택에 따라 해석이 용의하다는 장점을 갖는다.

따라서 본 비교 연구의 결과를 토대로 주어진 상황과 각 방법의 특징에 따라 적합한 분류에 사용할 방법을 선택하는데 유용한 정보를 얻을 수 있었으며, 이렇게 얻어진 정보가 다양한 분야에서 빈번하게 얻어지는 불균형 자료에서의 분류 시 도움이 될 수 있을 것으로 기대하는 바이다.

고려대학교 KOREA UNIVERSITY 1905

참고문헌

[1] Y Sun, AKC Wong, MS Kamel (2009). Classification of imbalanced data: a review, *International Journal of Pattern Recognition and Artificial Intelligence, Vol. 23*(No. 04), pp. 687-719

[2] Elhassan AT, Aljourf M, Al-Mohanna F and Shoukri M (2016). Classification of Imbalance Data using Tomek Link(T-Link) Combined with Random Under-sampling (RUS) as a Data Reduction Method, *Journal of Informatics and Data Mining*

[3] E Ramentol, Y Caballero, R Bello, F Herrera (2012). SMOTE-RSB_*: a hybrid preprocessing approach based on oversampling and undersampling for high imbalanced data-sets using SMOTE and rough sets theory, *Knowledge and Information Systems, Volume 33*(Issue 2), pp. 245 - 265

[4] Han H, Wang WY, Mao BH (2005). Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. *International conference on intelligent computing (ICIC05) LNCS 3644*. Springer, pp. 878 - 887

[5] 이영섭, 오현정, 김미경(2005). 배깅, 부스팅, SVM 분류 알고리즘 비교분석, 응용통계연구 <제 18권 2호>, pp. 343-354

[6] 허명회. 『응용데이터분석』. 자유아카데미, 2014.

고려대학교 KOREA UNIVERSITY 1905