논문 2008-45CI-6-22

클래스 불균형 문제에서 베이지안 알고리즘의 학습 행위 분석

(Learning Behavior Analysis of Bayesian Algorithm Under Class Imbalance Problems)

황 두 성*

(Doosung Hwang)

요 약

본 논문에서는 베이지안 알고리즘이 불균형 데이터의 학습 시 나타나는 현상을 분석하고 성능 평가 방법을 비교하였다. 사전 데이터 분포를 가정하고 불균형 데이터 비율과 분류 복잡도에 따라 발생된 분류 문제에 대해 베이지안 학습을 수행하였다. 실험 결과는 ROC(Receiver Operator Characteristic)와 PR(Precision-Recall) 평가 방법의 AUC(Area Under the Curve)를 계산하여 불균형 데이터 비율과 분류 복잡도에 따라 분석되었다. 비교 분석에서 불균형 비율은 기 수행된 연구 결과와 같이 베이지안 학습에 영향을 주었으며, 높은 분류 복잡도로부터 나타나는 데이터 중복은 학습 성능을 방해하는 요인으로 확인되었다. PR 평가의 AUC는 높은 분류 복잡도와 높은 불균형 데이터 비율에서 ROC 평가의 AUC보다 학습 성능의 차이가 크게 나타났다. 그러나 낮은 분류 복잡도와 낮은 불균형 데이터 비율의 문제에서 두 측정 방법의 학습 성능의 차이는 미비하거나 비슷하였다. 이러한 결과로부터 PR 평가의 AUC는 클래스 불균형 문제의 학습 모델의 설계와 오분류 비용을 고려한 최적의 학습기를 결정하는데 도움을 줄 수 있다.

Abstract

In this paper we analyse the effects of Bayesian algorithm in learning class imbalance problems and compare the performance evaluation methods. The learning performance of the Bayesian algorithm is evaluated over the class imbalance problems generated by priori data distribution, imbalance data rate and discrimination complexity. The experimental results are calculated by the AUC(Area Under the Curve) values of both ROC(Receiver Operator Characteristic) and PR(Precision-Recall) evaluation measures and compared according to imbalance data rate and discrimination complexity. In comparison and analysis, the Bayesian algorithm suffers from the imbalance rate, as the same result in the reported researches, and the data overlapping caused by discrimination complexity is the another factor that hampers the learning performance. As the discrimination complexity and class imbalance rate of the problems increase, the learning performance of the AUC of a PR measure is much more variant than that of the AUC of a ROC measure. But the performances of both measures are similar with the low discrimination complexity and class imbalance rate of the problems. The experimental results show that the AUC of a PR measure is more proper in evaluating the learning of class imbalance problem and furthermore gets the benefit in designing the optimal learning model considering a misclassification cost.

Keywords: Class imbalance problem, Data overlapping, Bayesian algorithm, Performance evaluation

I. 서 론

다양한 분야의 벤치마킹 문제에서 검증된 학습 알고리즘은 새로운 응용에서 숨겨진 분류(classification) 또는 군집화(clustering) 규칙을 찾아내는 데이터마이닝 분야에서 중요한 도구가 되었다. 실 응용에서 학습 알고

* 정회원, 단국대학교 컴퓨터과학과
(Department of computer science, Dankook University)
※ 이 연구는 2008년 단국대학교 대학 연구비의 지원으로 연구되었습니다.
접수일자: 2008년8월25일, 수정완료일: 2008년10월23일

리즘의 높은 일반화 성능을 보장하기 위해서는 데이터 준비, 데이터 전처리와 코딩, 학습 모델 설정, 학습과 테스트 등 일련의 단계들을 밀접하게 연관된다. 기계 학습 응용의 데이터 준비 단계에서 나타나는 클래스 불균형은 데이터 전 처리에서부터 학습과 테스트 단계까지 전체적으로 영향을 미쳐 학습 알고리즘의 높은 성능을 방해하는 요인으로 고려되었다.

학습 단계에서 발생되는 클래스 불균형 문제란 특정 클래스에 속하는 데이터의 비율이 다른 클래스에 비해 높은 차이가 나타나는 분류 문제를 학습하는 경우를 말한다[1~3]. 실 응용에서 의료 진단[4], 스팸 메일 차단[5], 텍스트 마이닝[6] 등은 클래스 불균형 문제의 사례이다. 지금까지 연구와 실험 결과로부터 클래스 불균형 문제는 적용된 학습 알고리즘의 성능을 저하시키는 원인으로 보고되었다[1, 4, 7~8]. 그러나 C4.5 학습 알고리즘을 가지고 실험한 유방암 진단[7]과 보육원[9] 분류 학습의 경우 클래스 간의 데이터 균형(data balancing)을 조절하는 방법을 이용한 학습 성능을 비교 평가하였으나 원래 데이터의 성능과 차이가 미비하였다. 이러한 문제의 학습에서는 클래스 불균형 비율이 학습 알고리즘의 성능 저하에 직접적인 요인이 되지 않은 이유는 학습 데이터가 분류 학습에 필요한 정보를 충분히 가지고 있기 때문으로 분석된다.

본 논문의 목적은 클래스 불균형 문제의 베이지안 학습[10]에서 나타나는 현상을 분석하고 올바른 성능 평가 방법을 파악하는데 있다. 사전 클래스 데이터 분포를 가정하고 불균형 데이터 비율과 분류 복잡도(discrimination complexity)에 따라 분석이 가능한 데이터를 인위적으로 발생시켜 분류 문제가 준비되었다. 베이지안 알고즘리의 성능은 ROC(Receiver Operator Characteristic)[11~13]와 PR(Precision-Recall)[11, 14] 평가 방법으로 측정하였으며, 측정 결과는 기대 성능 AUC (Area Under the Curve)[12, 14]을 계산하여 비교 분석하였다.

이 논문의 구성은 다음과 같다. Ⅱ장에서는 불균형 클래스 문제의 학습에서 나타나는 문제점들을 클래스 데이터 분포, 분류 복잡도, 데이터 중복, 불균형 데이터 비율 등과 관련하여 수행된 연구에 대하여 논의한다. 불균형 클래스 문제의 학습 평가 방법을 Ⅲ장에서 기술하고, Ⅳ장에서는 인위적으로 발생시킨 불균형 분류 문제들에서 베이지안 알고리즘의 학습 성능을 비교 분석한다. 마지막으로 Ⅴ장에서 결론을 기술한다.

Ⅱ. 클래스 불균형 문제

기계 학습의 목표는 주어진 클래스 별 준비된 데이터에 대해 분류 함수를 학습하여 클래스가 알려지지 않은 데이터의 클래스를 예측하는 것이다. 분류 문제에서 학습 알고리즘의 응용은 클래스 별 준비된 데이터의 수가 균등하다고 가정하는 것이 일반적이다. 그러나 실세계에서 나타나는 많은 문제에서 학습 알고리즘의 적용 시 클래스 불균형 데이터의 학습이 이루어지고 있다[4, 7, 9]. 클래스 불균형 문제가 적은 수의 긍정 데이터(positive data)와 대다수의 부정 데이터(negative data)로 구성될 때 학습 알고리즘은 데이터의 참여 비율이 높은 부정 클래스에 치우친 학습이 진행된다. 학습에서 나타나는 이러한 단점은 학습 알고리즘의 목적 함수(objective function)가 학습 데이터의 오류를 전체적으로 최소화시키는 학습을 진행하여 클래스 별 학습 데이터의 분포를 반영시키지 못하기 때문으로 분석되었다[1, 3]. 그러므로 학습 성능 평가에서 데이터 분포 비율이 높은 클래스의 분류 예측율은 높으나 데이터의 비율이 낮은 클래스에 대해 저조한 학습 성능을 보이는 경향이 발생하게 된다.

클래스 불균형 문제에 대한 관련 연구는 벤치마킹 불균형 데이터 문제가 학습 알고리즘에 끼치는 영향 분석 연구, 불균형 데이터 문제의 학습에서 알고리즘의 성능 개선에 대한 연구, 클래스 데이터의 분포를 가정한 분류 문제에서 불균형 데이터 학습이 미치는 영향 분석에 대한 연구, 불균형 클래스 문제의 적절한 학습 성능 평가에 대한 연구 등이 진행되었다. 이러한 연구 결과로부터 기계 학습 연구자들에게 클래스 불균형은 학습 알고리즘의 성능을 방해하는 요인으로 인식되었다.

클래스 불균형 문제가 학습 알고리즘에 미치는 영향을 분석한 사례로는 다양한 벤치마킹 분류 문제에서 의사결정트리 C4.5[7~8, 13], 신경망 모델에서 오류 역전파 알고리즘[4], 지지벡터기계(support vector machine)[5~6], 베이지안 알고리즘[13] 등이 테스트되었다. 샘플링 알고리즘을 이용한 클래스 불균형 문제의 학습 성능 향상에 대한 연구도 병행 진행되었다[5, 9, 16]. 그러나 연구 결과로부터 불균형 데이터 문제가 학습 성능을 방해하는 주원인이 되는가에 대한 해답을 제시하지 못하였다. 이러한 이유는 벤치마킹 문제에서 교차 검증(cross validation)을 이용한 학습과 테스트 단계에서 비교 결과만 가지고 분석되었으며 학습 데이터의 클래스 분포가 고려되지 않았다. 학습 성능의 비교 결과만 가지고 불균형 데이터 문제가 데이터 비율이 낮은 클래스에 대

해 낮은 예측율을 보이는 원인이라고 단정할 수 없다.

클래스 분포를 가정하고 발생시킨 다양한 학습 문제에서 C4.5[2, 15]와 오류 역전파 알고리즘[4]의 적용에서 데이터 불균형이 미치는 현상을 파악하였다. 주로 가우시안 분포(Gaussian distribution)를 가정하고 발생된 불균형 문제의 학습 결과를 비교 평가하였다. 분류 문제의 각 클래스는 평균과 분산을 가지고 발생시켰으며 학습 난이도는 두 클래스의 평균거리의 차로 분류 복잡도이며 분산은 고정시켜 학습 성능이 측정되었다. 학습 성능은 ROC 평가의 AUC 계산으로 비교되었다. 실험 분석으로부터 불균형 데이터와 분류 복잡도로 인한 데이터 중복은 일반화 성능을 방해하는 문제로 파악되었다. 그러나 낮은 분류 복잡도를 보이는 분류 문제의 학습에서 불균형 데이터 비율이 C4.5 알고리즘과 오류 역전파 알고리즘의 학습에 미치는 영향은 미비하다고 보고되었다.

Ⅲ. 클래스 불균형 문제의 학습 평가

이진 분류 문제에서 학습 알고리즘의 예측 결과는 긍정 클래스(positive class)와 부정 클래스(negative class)로 구분된다. 두 개의 분류 클래스로 예측 결과가 나타나는 경우에 표 1의 교차 테이블(confusion table)[11]의 평가 방법이 주로 사용된다. TP(True Positives)는 긍정 클래스의 입력 데이터를 긍정 클래스로 정확히 분류된 데이터의 수, FP(False Positives)는 부정 클래스의 입력 데이터가 긍정 클래스로 오 분류된 데이터의 수, FN(False Negatives)는 부정 클래스의 데이터를 긍정 클래스로 잘못 예측한 데이터의 수, 그리고 TN(True Negatives)는 부정 클래스의 입력 데이터를 부정 클래스로 바르게 분류한 데이터의 수들로부터 여러 형태의 평가가 가능하다.

계산된 교차 테이블로부터 오류율(error rate), 정확률(accuracy rate), ROC와 PR 평가 등 다양한 척도가 이용될 수 있다. 두 클래스 이상을 포함하는 다중 분류 학습의 평가에 주로 사용되는 오류율 *ERR*는 총 입력 데이터의 수에 오 분류된 데이터 수의 비율이며 정확률 *ACC*는 *1-ERR*로 계산된다. 오류율과 정확율을 불균형 데이터로 구성되는 분류 문제의 학습 평가로 사용은 적절하지 않다[11~14]. 대다수 데이터가 속하는 부정 클래스에 대해 편향된 학습이 되면 소수의 긍정 데이터를 부정 클래스로 예측하는 경향이 나타나도 정확률이 높게 계산되기 때문이다. 그러므로 부정 클래스의 높은 참여 비율의 변화가 *ERR*와 *ACC*의 측정에 직접적으로 영향을 주어 높은 정확율을 보장하기 때문이다.

표 1. 교차테이블을 이용한 이진 분류 문제의 학습 평가

Table 1. Learning evaluation of a binary classification problem using confusion table.

예측 \ 실제	긍정	부정
긍정	TP	FP
부정	FN	TN

오류율과 정확률의 단점을 극복하기 위해 교차 테이블로부터 사전 데이터 분포의 비율를 반영하는 각 클래스 내에서 예측율 평가에 ROC와 PR 평가가 이용될 수 있다. ROC 평가 방법은 정확히 예측된 긍정 데이터의 비율 *TPR*(True Positive Rate, 식 (1))과 긍정 클래스로 분류한 부정 클래스의 비율 *FPR*(False Positive Rate, 식 (2))은 두 클래스의 데이터로부터 예측된 긍정 클래스에 참여하는 비율이 된다.

$$TPR = \frac{TP}{TP + FN} \tag{1}$$

$$FPR = \frac{FP}{FP + TN} \tag{2}$$

PR 평가는 정확하게 예측한 긍정 데이터의 REC 비율(Recall rate)과 긍정 클래스로 예측한 데이터 중에서 올바른 긍정 데이터의 비율 *PRE*(Precision rate, 식 (3))로 계산되며 *REC*는 ROC 측정 방법의 *TPR* 식 (1)과 같다.

$$PRE = \frac{TP}{TP + FP} \tag{3}$$

불균형 데이터의 학습 평가에서 ROC 평가의 가시화는 사전 데이터의 분포와 부정 클래스의 데이터가 긍정 클래스로 오 분류된 FP의 비율과 긍정 클래스 내에서 정확히 예측된 TP의 비율로부터 측정된다. FP의 변화는 FPR 계산에는 비례하지만 TPR 측정에는 고려되지 않는다. 그러므로 높은 불균형 데이터 문제의 학습 평가에 ROC 평가는 사전 데이터 분포와 클래스 별 오류율을 이용하여 나타난다. 소수의 긍정 데이터의 비율이 사전 데이터 분포에 반영되나 오 분류된 부정 데이터의 비율이 반영되지 않는다. PR 평가의 PRE는 오 분류된 부정 클래스의 데이터 수 FP를 고려한 TP의 비율을 계산한다. 불균형 데이터 문제가 학습 시 대다수의 학습

표 2. 클래스 간 거리와 불균형 비율에 따른 AUC-ROC 비교(%)
Table 2. AUC-ROC comparisons(%) with class distances and imbalance rates.

불균형 비율 \ 중심거리	0	1	2	3	4	5	6	7	8	9
1%	50.0 (0.00)	52.1 (0.44)	72.8 (1.10)	93.6 (0.98)	98.1 (0.00)	100.0 (0.00)	100.0 (0.00)	100.0 (0.00)	100.0 (0.00)	100.0 (0.00)
5%	50.0 (0.00)	53.4 (0.54)	78.9 (0.48)	98.5 (0.19)	99.1 (0.00)	99.6 (0.00)	100.0 (0.00)	100.0 (0.00)	100.0 (0.00)	100.0 (0.00)
10%	50.0 (0.00)	61.5 (0.45)	87.2 (0.63)	96.6 (0.15)	99.8 (0.11)	100.0 (0.01)	100.0 (0.00)	100.0 (0.00)	100.0 (0.00)	100.0 (0.00)
15%	50.0 (0.00)	64.6 (0.41)	87.1 (0.27)	97.5 (0.16)	99.9 (0.05)	99.9 (0.00)	100.0 (0.00)	100.0 (0.00)	100.0 (0.00)	100.0 (0.00)
20%	50.0 (0.00)	67.5 (0.39)	89.7 (0.25)	97.9 (0.11)	99.6 (0.04)	100.0 (0.00)	100.0 (0.00)	100.0 (0.00)	100.0 (0.00)	100.0 (0.00)
25%	50.0 (0.00)	69.5 (0.38)	90.7 (0.19)	99.7 (0.11)	99.7 (0.03)	100.0 (0.00)	100.0 (0.00)	100.0 (0.00)	100.0 (0.00)	100.0 (0.00)
30%	50.0 (0.00)	72.6 (0.37)	91.4 (0.16)	97.8 (0.07)	99.7 (0.03)	100.0 (0.03)	100.0 (0.00)	100.0 (0.00)	100.0 (0.00)	100.0 (0.00)
35%	50.00 (0.00)	74.1 (0.26)	91.7 (0.16)	98.0 (0.08)	99.7 (0.03)	99.9 (0.00)	100.0 (0.00)	100.0 (0.00)	100.0 (0.00)	100.0 (0.00)
40%	50.0 (0.01)	75.2 (0.17)	91.9 (0.11)	98.5 (0.04)	99.8 (0.04)	100.0 (0.01)	100.0 (0.00)	100.0 (0.00)	100.0 (0.00)	100.0 (0.00)
45%	50.0 (0.00)	74.1 (0.22)	91.7 (0.17)	98.2 (0.06)	99.7 (0.02)	99.9 (0.00)	100.0 (0.00)	100.0 (0.00)	100.0 (0.00)	100.0 (0.00)
50%	50.0 (0.37)	76.1 (0.27)	90.7 (0.17)	98.6 (0.07)	99.7 (0.02)	100.0 (0.00)	100.0 (0.00)	100.0 (0.00)	100.0 (0.00)	100.0 (0.00)

데이터로 구성된 부정 클래스의 학습에 치우치면 FN은 작아지고 FP는 커지게 되어 PRE는 상대적으로 작아지게 된다. 그러므로 PR 평가는 소수의 긍정 데이터의 분포를 성능 평가에 반영한다.

신경망, 지지벡터기계, 베이지안 학습 알고리즘의 예측 결과는 클래스 멤버십 또는 클래스 참여 확률로 고려될 수 있다. 데이터 집합의 예측 결과를 클래스 멤버십에 따라 정열하면 참여 정도에 따라 ROC와 PR 평가가 2차원 그래프로 가시화될 수 있다. 가시화된 2차원의 ROC와 PR 평가로부터 AUC(Area Under the Curve) 계산은 학습 알고리즘의 기대 성능(expected performance)으로 이용되었다[12~13].

지금부터 AUC-ROC 평가를 ROC 평가 방법의 AUC, AUC-PR 평가를 PR 측정 방법의 AUC라 하겠다. 일반적으로 AUC-ROC 평가는 학습 알고리즘의 성능 비교와 클래스 불균형 문제의 객관적인 학습 성능 평가로 채택되었다[2, 4, 7, 9]. AUC-ROC는 불균형 비율이 낮아지면서 나타나는 성능 향상을 제시하지 못한다. AUC-PR 평가가 불균형 데이터 문제의 학습 성능 평가에 더 적절하다는 연구가 수행되었으며 비선형 접근법(nonlinear interpolation)을 이용한 PR 평가의 AUC 계산 알고리즘이 제시되었다[14].

IV. 실 험

불균형 데이터 문제가 학습 성능을 저하시키는 영향을 분석하기 위해 클래스에 속하는 데이터의 분포가 조절되는 이진 분류 문제를 인위적으로 생성하였다. 각 분류 문제는 소수의 클래스 데이터를 갖는 긍정 클래스와 다수 클래스 데이터를 갖는 부정 클래스로 구성시켰으며 불균형 비율에 따라 부정과 긍정 클래스에 속하는 데이터의 수를 결정하여 데이터 분포를 조절하였다.

이러한 사전 데이터 분포에 따라 각 문제의 일반화 성능은 불균형 데이터 비율과 분류 복잡도에 따라 비교 분석하였다. 준비된 문제의 베이지안 알고리즘의 학습 성능은 ROC과 PR 평가로 가시화하여 기대 성능 AUC를 측정하였다. 학습 성능 비교에서 AUC-ROC 계산은 C. Ferri가 제시한 알고리즘[12], AUC-PR 계산은 Jesse Davis의 알고리즘[14]을 구현하였다. 베이지안 알고리즘은 R 패키지 e1071[16]을 사용하였다.

실험 결과를 바탕으로 기 수행된 연구 결과에서 지적한 바와 같이 불균형 문제가 베이지안 학습의 성능 저

하에 미치는 영향이 분석되었다. 불균형 비율과 데이터 중복이 학습 성능 저하에 미치는 영향을 분석하고, 지금까지 분류 문제의 성능 평가에서 주로 사용된 ROC 평가를 불균형 문제의 학습에서 PR 평가와 비교하여 어느 평가가 보다 객관적인 평가 방법인가를 실험 결과를 바탕으로 분석하였다.

1. 학습 데이터의 준비

가우시안 분포를 가정하고 클래스 간의 거리와 불균형 데이터의 비율에 따라 학습 데이터를 준비하였다. 학습 데이터의 차원은 5이며 학습 데이터의 수는 5,000이다. 고정된 불균형 비율에 대해 첫 번째 불균형 문제의 부정과 긍정 클래스는 같은 범위 $(-1,1)$에서 발생시킨 임의 평균과 대각 행렬의 값이 1인 분산으로 준비되었다. 두 번째 불균형 문제의 긍정 클래스는 첫 번째 문제의 평균과 분산을 동일하고, 부정 클래스의 평균은 첫 번째 문제의 평균에 1을 더하여 준비시키는 방법으로 10개의 분류 문제를 구성시켰다. 학습 데이터의 불균형 비율은 1%, 5%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, 50%로 하였다. 각 분류 문제의 긍정과 부정 클래스의 데이터 수가 불균형 데이터의 비율에 따라 결정되어 최소 50개(1%), 최대 2,500개(50%) 긍정 데이터가 분포된다. 11개의 불균형 비율과 중심 거리에 따라 10개의 분류 문제가 만들어져 총 110개의 불균형 데이터 문제가 준비되었다.

발생된 불균형 데이터 문제의 분류 복잡도는 긍정과 부정 클래스의 두 중심 거리이며 학습 난이도가 된다. 첫 번째 발생된 문제의 긍정과 부정 클래스의 중심 거리는 0이며, i 번째 발생된 문제의 부정 클래스의 중심 거리는 $(i-1)$된다. 중심 거리가 같으면 긍정과 부정 데이터가 한 클래스에 모두 속하게 되는 심각한 데이터 중복 문제가 발생하여 학습이 어렵다. 그러나 중심 거리가 증가하면서 데이터 중복 현상은 나타나지 않는 단순 분리가 가능한 학습 문제가 된다.

2. 실험 결과

준비된 불균형 데이터 문제는 베이지안 학습 알고리즘을 가지고 10-교차 검증(10-fold cross validation)으로 성능을 비교 평가하였다. 불균형 비율과 두 클래스의 중심 거리에 따라 ROC 와 PR 평가의 기대 성능 AUC-ROC와 AUC-PR을 계산한 실험 결과가 표 2와 3에 제시되었다. 열은 중심 거리이며, 행은 다수의 부정

표 3. 클래스 간 거리와 불균형 비율에 따른 AUC-PR 비교(%)
Table 3. AUC-PR comparisons(%) with class distances and imbalance rates.

불균형 비율 \ 중심거리	0	1	2	3	4	5	6	7	8	9
1%	1.0 (0.00)	6.8 (1.33)	38.6 (2.05)	85.5 (2.21)	94.4 (0.00)	100.0 (0.00)	100.0 (0.00)	100.0 (0.00)	100.0 (0.00)	100.0 (0.00)
5%	5.0 (0.00)	9.8 (1.00)	54.9 (0.48)	93.0 (0.59)	97.6 (0.29)	99.6 (0.00)	100.0 (0.00)	100.0 (0.00)	100.0 (0.00)	100.0 (0.00)
10%	10.0 (0.00)	27.7 (0.83)	67.6 (1.27)	91.6 (0.42)	99.8 (0.19)	100.0 (0.00)	100.0 (0.00)	100.0 (0.00)	100.0 (0.00)	100.0 (0.00)
15%	15.0 (0.00)	36.6 (0.65)	71.5 (0.36)	93.0 (0.33)	98.6 (0.07)	99.9 (0.24)	100.0 (0.00)	100.0 (0.00)	100.0 (0.00)	100.0 (0.00)
20%	20.0 (0.00)	44.2 (0.59)	77.9 (0.32)	95.8 (0.24)	99.2 (0.14)	100.0 (0.00)	100.0 (0.00)	100.0 (0.00)	100.0 (0.00)	100.0 (0.00)
25%	25.0 (0.00)	50.7 (0.60)	81.8 (0.40)	94.3 (0.22)	99.6 (0.08)	99.9 (0.00)	100.0 (0.00)	100.0 (0.00)	100.0 (0.00)	100.0 (0.00)
30%	30.0 (0.00)	57.9 (0.45)	84.2 (0.21)	95.5 (0.20)	99.7 (0.06)	99.9 (0.00)	100.0 (0.00)	100.0 (0.00)	100.0 (0.00)	100.0 (0.00)
35%	35.0 (0.00)	62.0 (0.34)	86.0 (0.18)	96.5 (0.12)	99.4 (0.07)	99.9 (0.04)	100.0 (0.00)	100.0 (0.00)	100.0 (0.00)	100.0 (0.00)
40%	40.0 (0.06)	65.8 (0.21)	87.4 (0.17)	97.6 (0.07)	99.6 (0.06)	100.0 (0.00)	100.0 (0.00)	100.0 (0.00)	100.0 (0.00)	100.0 (0.00)
45%	46.7 (0.00)	67.3 (0.20)	88.6 (0.22)	97.2 (0.11)	99.5 (0.02)	99.8 (0.00)	100.0 (0.00)	100.0 (0.00)	100.0 (0.00)	100.0 (0.00)
50%	52.7 (0.24)	71.8 (0.26)	88.5 (0.27)	98.2 (0.08)	99.7 (0.03)	100.0 (0.00)	100.0 (0.00)	100.0 (0.00)	100.0 (0.00)	100.0 (0.00)

데이터에 대한 긍정 데이터의 불균형 비율이다. 각 예측값은 불균형 비율과 중심 거리에서 학습한 문제의 기대 성능 AUC 값이며 괄호는 표준편차이다.

AUC-ROC의 기대성능을 보이는 표 2에서 긍정과 부정 클래스의 중심이 같으면 불균형 비율에 관계없이 약 50%의 AUC를 얻었다. 모든 긍정 클래스의 테스트 벡터가 부정 클래스로 예측되었기 때문이다. 고정된 불균형 비율에서 중심거리 0~4까지는 성능 향상이 나타나지만 5이상에서는 차이가 미비하였다. 두 클래스의 중심거리를 고려하면 긍정 데이터가 증가하면서 AUC가 증가하였다. 높은 불균형 비율일 경우 소수의 긍정 데이터와 데이터 중복이 학습을 방해하였으나 긍정 데이터의 수가 많아지면서 학습 성능 향상이 나타났다. 불균형 비율과 관계없이 중심 거리 5~9에서 약 100%의 기대 성능을 나타내 클래스 분포와 데이터 중복의 영향이 미비하였다. 중심 거리가 0인 경우를 제외하면 분류 복잡도에 따라 긍정 데이터의 증가는 일반화 성능을 높일 수 있는 가능성을 제시하고 있다. 측정된 AUC-ROC 값의 범위는 50~100%에서 나타나고 있다. 중심 거리가 증가하면서 같은 불균형 비율에서 표준 편차가 작아지고 있다. 데이터 중복이 사라지면서 단순 학습이 가능하기 때문으로 분석된다.

표 3의 AUC-PR에서 중심 거리가 같을 때 클래스 불균형 비율이 낮아지면서 표 2의 AUC-ROC 보다 성능 향상이 뚜렷이 나타나고 있다. 이러한 증가 추세는 고정된 불균형 비율에서 중심 거리가 3일 때 까지 비슷하게 나타나고 있으며 5이상 실험에서는 성능 향상이 미비하다. 중심 거리가 같은 실험에서 긍정 데이터의 수가 많아지면서 학습 성능이 향상되었다. 중심 거리가 0과 1, 불균형 비율 1%와 5%에서는 10%내의 낮은 AUC을 나타났다. 그러나 중심 거리가 증가하고 불균형 비율이 낮아지면서 AUC 값이 빠르게 증가하고 있다. 이러한 사실은 표 2에서와 동일하게 긍정 데이터의 증가는 일반화 성능을 높일 수 있는 가능성을 제시하여 과도 샘플링(oversampling) 기법을 도입하면 성능 향상이 있을 수 있다는 실험적 근거가 될수 있다.

표 2와 3의 결과를 비교하기 위해 중심 거리와 불균형 비율에 따라 기대 성능을 가시화하였다. 명확한 분석을 위해 선택된 불균형 비율은 1%, 10%, 20%, 30%, 40%, 50%이다. 중심 거리는 ROC와 PR 평가에서 중심 거리의 변화에 따른 성능 향상을 보이는 0, 1, 2, 3, 4, 5로 선택되었다. X-축을 중심 거리로 하는 성능 비교를 보이는 그림 1과 2에서 불균형 비율이 낮아지면서 중심

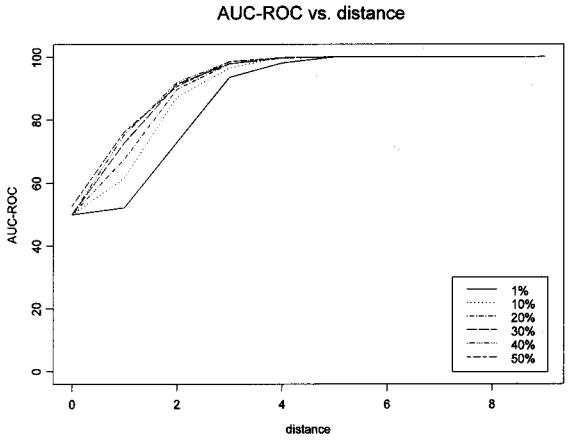

그림 1. 클래스 중심 거리에 따른 AUC-ROC 비교
Fig. 1. Comparisons in AUC-ROC vs. class distances.

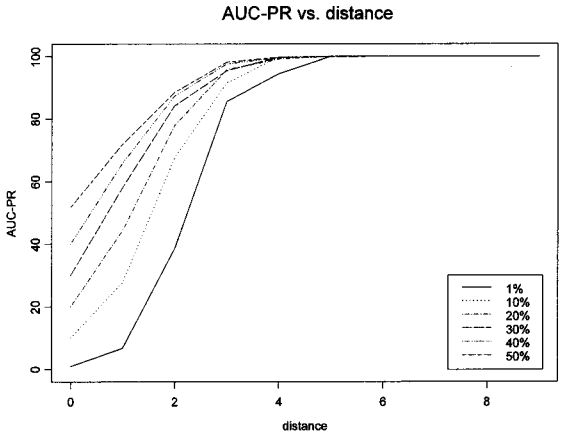

그림 2. 클래스 중심 거리에 따른 AUC-PR 비교
Fig. 2. Comparisons in AUC-PR vs. class center distance.

거리가 4까지 기대 성능이 증가하고 있으나 AUC-PR 평가에서 보다 가파르게 증가하고 있다. 그러나 이러한 학습 성능 증가는 중심 거리의 차가 5에서는 미비하나 6 이상에서는 두 평가 방법이 100%의 성능을 보인다. 그림 2에서 불균형 비율이 1%인 중심거리가 같을 때 AUC-PR은 최저 기대 성능을 갖으나 긍정과 부정 데이터 수가 동등하게 되면 약 0.5(약 50%)까지 성능 향상이 되었다. 이런 성능 개선은 긍정 데이터의 수가 증가하여 나타나지만 그림 1의 AUC-ROC 평가에서는 나타나지 않았다. ROC 평가는 각 클래스에 속하는 데이터에 대한 정확히 예측된 긍정 데이터 비율과 오 분류된 비율만을 고려하여 평가되어 AUC-ROC에서는 나타나지 않았다.

X-축을 불균형 비율로 중심 거리에 대한 두 평가 방법의 기대성능 비교가 그림 3과 4이다. AUC-ROC의 그

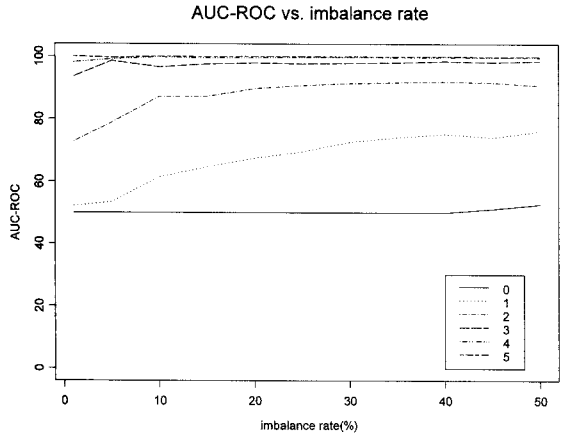

그림 3. 클래스 불균형 비율에 따른 AUC-ROC 비교
Fig. 3. Comparisons in AUC-ROC vs. imbalance rates.

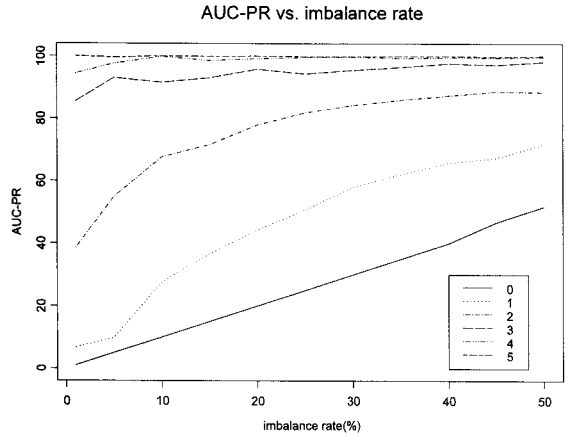

그림 4. 클래스 불균형 비율에 따른 AUC-PR 비교
Fig. 4. Comparisons in AUC-PR vs. imbalance rates

림 3에서 긍정 데이터의 수가 증가하면서 성능 향상이 미비하게 나타나고 있다. 중심거리가 1, 2, 3이고 불균형 비율이 1%에서 증가를 일시적으로 보이고 있으나 불균형 비율 10%를 기점으로 긍정 데이터의 수가 많아져도 AUC 값의 증가가 미비하다.

그림 4의 AUC-PR의 비교에서 AUC-ROC의 그림 3에 비해 불균형 비율이 높더라도 중심 거리가 멀어지면 뚜렷한 AUC 값의 증가를 보이고 있다. 클래스 분포에서 살펴보면 데이터 중복이 높은 중심 거리 0, 1, 2에서 이러한 증가는 불균형 비율이 낮아져도 증가하고 있다. 낮은 데이터 중복으로 인한 분류 복잡도가 단순하기 때문이다. 중심 거리에 따른 AUC-ROC의 그림 1에서 중심거리가 3 이상의 경우 불균형 비율 10%를 기점으로 성능 향상이 나타나지 않는 현상은 불균형 비율에 따른 비교를 보이는 그림 3에서도 비슷하였다. 그러므로 불균형 비율에 관계없이 두 긍정 클래스와 부정 클래스 간의 중심 거리의 차가 3이상일 때 그림 3과 4는 비슷한 학습 성능을 보이고 있다. 균등한 클래스 데이터를 포함하는 50%의 불균형 비율에서는 같은 기대 성능이 측정되었다.

AUC-PR의 기대 성능은 높은 불균형 비율에서 학습 성능의 증가가 AUC-ROC에 비해 높게 나타나 불균형 비율과 분류 복잡도에 따른 성능 향상을 파악할 수 있었다. 그러나 분류 복잡도가 낮은 중간거리가 3이상 그리고 30% 이상의 높은 균형 비율로 구성된 실험에서 학습 성능의 증가율이 비슷하였다. 불균형 문제의 학습에서 AUC-ROC 성능 평가는 긍정 데이터의 증가에 따른 성능 향상의 측정에 어려움이 있다. 그러므로 불균형 데이터 문제의 성능 평가에 AUC-ROC 평가보다는 AUC-PR 평가의 선택이 보다 적절하다.

V. 결 론

클래스 불균형 데이터의 학습은 기계학습 알고리즘의 적용에서 반드시 반영되어야 할 문제로써 연구되었다. 그러나 높은 불균형 비율로 구성되는 분류 문제에서 높은 일반화 성능이 보고되기도 하였다. 이 논문에서는 다양한 클래스 불균형 비율과 분류 복잡도를 가지고 발생된 학습 문제가 베이지안 알고리즘에 끼치는 영향을 파악하고 ROC와 PR 측정을 이용한 학습 성능을 비교 평가 하였다.

실험 분석에서 준비된 클래스 불균형 문제의 베이지안 학습은 지금까지 수행된 연구 결과와 동일하게 불균형 데이터의 학습에 장애가 되었으며, 특히 분류 복잡도 때문에 발생되는 데이터 중복은 학습 성능을 저하시키는 요인이 되었다. 이러한 분석 결과는 C4.5를 가지고 수행된 클래스 불균형과 데이터 중복에 대한 연구 결과와 유사하다[2]. 한편 낮은 분류 복잡도를 갖은 불균형 데이터의 학습에서 미치는 영향은 나타나지 않았다. 그러므로 실 응용에서 데이터 중복과 클래스 불균형의 상관관계의 이해는 높은 일반화 성능을 위해 고려되어야 한다.

ROC와 PR 평가 방법을 이용한 베이지안 학습의 비교에서는 AUC-PR은 높은 분류 복잡도와 높은 불균형 데이터 비율에서 AUC-ROC보다 기대 학습 성능의 증가가 뚜렷하였다. 그러나 낮은 분류 복잡도와 낮은 불균형 데이터 비율의 문제에서 두 측정 방법의 차이는 미비하거나 동일하였다. 이러한 결과로부터 불균형 비율이 심한 분류 문제의 학습 평가에 ROC 평가보다는 PR

평가가 오분류 비용을 고려한 최적의 학습기를 결정하는데 도움을 줄 수 있다.

참 고 문 헌

[1] Japkowicz N. and Stephen S., "The Class Imbalance Problem: A Systematic Study," Intelligent Data Analysis, Vol. 6, no. 5, pp. 429-450, November 2002.

[2] Ronaldo C. Prati, Gustavo E. A. P. A. Batista and Maria Carolina Monard, "Class Imbalances versus Class Overlapping: An Analysis of a Learning System Behavior," MICAI, pp, 312-321, 2004.

[3] Jie Gu, Yuanbing Zhou and Xianqiang Zuo, "Making Class Bias Useful: A Strategy of Learning from Imbalanced Data," Intelligent Data Engineering and Automated Learning(IDEAL), pp.287-295, 2007.

[4] Maciej A. Mazurowski, Piotr A. Habas, Jacek M. Zurada, Joseph Y. Lo, Jay A. Baker and Georgia D. Tourassi, "Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance," Neural Networks, Vol. 21, no. 2-3, pp.427-436, 2008.

[5] Yuchun Tang, Sven Krasser, Paul Judge and Yan-Qing Zhang, "Fast and Effective Spam Sender Detection with Granular SVM on Highly Imbalanced Mail Server Behavior Data," Collaborative Computing: Networking, Applications and Worksharing, pp.1-6, 2006.

[6] Tie-Yan Liu, Yiming Yang, Hao Wan, Hua-Jun Zeng, Zheng Chen and Wei-Ying Ma, "Support Vector Machines Classification with A Very Large-scale Taxonomy," SIGKDD Explorations, Vol. 7, no. 1, 2005.

[7] Gary M. Weiss and Foster J. Provost, "Learning When Training Data are Costly: The Effect of Class Distribution on Tree Induction." J. Artif. Intell. Res.(JAIR), Vol. 19, pp. 315-354, 2003.

[8] Vicente Garca and Ram-n Alberto Mollineda, "An Empirical Study of the Behavior of Classifiers on Imbalanced and Overlapped Data Sets," CIARP, pp. 397-406, 2007.

[9] Gustavo E. A. P. A. Batista, Ronaldo C. Prati and Maria Carolina Monard, "A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data," SIGKDD Explorations, Vol. 6, 2004.

[10] Christopher M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.

[11] Ian H. Witten and Eibe Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd edition, Elsevier, 2005.

[12] C. Ferri, P. Flach and J. Hernndez-Orallo, "Learning Decision Trees Using the Area Under ROC Curve," Proceedings of the 19th International Conference on Machine Learning(ICML-2002), pp. 139-146, 2002.

[13] Jin Huang and Charles X. Ling, "Using AUC and Accuracy in Evaluating Learning Algorithms," IEEE Trans. Knowl. Data Eng., Vol. 17, no. 3, pp. 299-310, 2005.

[14] Jesse Davis and Mark Goadrich, "The relationship between Precision-Recall and ROC curves," Proceedings of the 23th International Conference on Machine Learning(ICML-2006), pp. 233-240, 2006.

[15] Visa, S. and Ralescue, A., "The effect of imbalanced data class distribution on fuzzy classifiers-experimental study," Proceedings of the FUZZ-IEEE Conference, 2005.

[16] Dimitriadou E, Hornik K, Leisch F, Meyer D and Weingessel A, "e1071: Misc Functions of the Department of Statistics(e1071)", Version 1.5-11, TU Wien, 2007.

저 자 소 개

황 두 성(정회원)

1985년 충남대학교 계산통계학과 학사

1990년 충남대학교 계산통계학과 석사

2003년 Wayne State University, Computer Science 박사졸업

1990년~1991년 국토개발연구원 연구원

1991년~1998년 전자통신연구소 선임연구원

2003년~현재 단국대학교 컴퓨터과학과 조교수

<주관심분야 : 데이터 마이닝, 기계학습, 병렬처리, 바이오인포매틱스>