韓國컴퓨터情報學會 論文誌
第15卷 第7號, 2010. 7.

2010-15-7-1-2

불균형 데이터 학습을 위한 지지벡터기계 알고리즘

김 광 성*, 황 두 성**

Support Vector Machine Algorithm for Imbalanced Data Learning

Kwang-seong Kim*, Doosung Hwang**

요 약

본 논문에서는 클래스 불균형 학습을 위한 이차 최적화 문제의 해를 구하는 개선된 SMO 학습 알고리즘을 제안한다. 클래스에 서로 다른 정규화 값이 부여되는 지지벡터기계의 최적화 문제의 구현에 SMO 알고리즘이 적합하며, 제안된 알고리즘은 서로 다른 클래스에서 선택된 두 라그랑지 변수의 현재 해를 구하는 학습 단계를 반복한다. 제안된 학습 알고리즘은 UCI 벤치마킹 문제에서 테스트되어 클래스 불균형 분포를 반영하는 g-mean 평가를 이용한 일반화 성능이 SMO 알고리즘과 비교되었다. 실험 결과에서 제안된 알고리즘은 SMO에 비해 적은 클래스 데이터의 예측율을 높이고 학습시간을 단축시킬 수 있다.

Abstract

This paper proposes an improved SMO solving a quadratic optmization problem for class imbalanced learning. The SMO algorithm is aproporiate for solving the optimization problem of a support vector machine that assigns the different regularization values to the two classes, and the prosoposed SMO learning algorithm iterates the learning steps to find the current optimal solutions of only two Lagrange variables selected per class. The proposed algorithm is tested with the UCI benchmarking problems and compared to the experimental results of the SMO algorithm with the g-mean measure that considers class imbalanced distribution for gerneralization performance. In comparison to the SMO algorithm, the proposed algorithm is effective to improve the prediction rate of the minority class data and could shorthen the training time.

▸ Keyword : 클래스 불균형 분류(class imbalance classification), 지지벡터기계(support vector machine), 성능 평가(performance evaluation)

• 제1저자 : 김광성 교신저자 : 황두성
• 투고일 : 2010. 05. 07, 심사일 : 2010. 06. 05, 게재확정일 : 2010. 06. 25.
* 현대정보기술 ** 단국대학교 공학대학 컴퓨터과학 부교수
※ 이 연구는 2008년 단국대학교 대학원 연구보조장학금의 지원으로 이루어진 것임

Ⅰ. 서 론

다양한 벤치마킹 문제에서 검증된 학습 알고리즘은 새로운 응용에서 숨겨진 분류 또는 군집화 규칙을 찾아내는 데이터마이닝 분야에서 중요한 도구로 활용되고 있다. 실 응용에서 학습 알고리즘의 높은 일반화 성능을 얻기 위해서는 데이터 준비, 데이터 전처리와 코딩, 학습 모델 설정, 학습과 테스트 등 일련의 단계들이 밀접하게 관계된다. 불균형 학습이란 준비된 클래스내 데이터 수가 현저한 차이로 인해 소수로 구성된 클래스의 예측율이 다수 데이터 클래스에 비해 낮게 학습되는 경우를 일컫는다. 클래스 불균형 문제는 데이터 준비 단계에서 나타나 전 처리에서 학습과 테스트 단계까지 전체적으로 영향을 미쳐 학습 알고리즘의 성능을 방해하는 요인으로 보고되었다[1,2,3]. 많은 응용 문제에서 학습 알고리즘의 적용시 클래스 불균형 데이터의 학습이 이루어지고 있다. 의료 진단[4], 스팸 메일 방지[5], 텍스트 마이닝[6], 네트워크 보안[7], 생물정보학[8] 등에서 높은 클래스 불균형 비율로 인해 학습 알고리즘의 일반화 성능에 영향이 나타난 사례들이다. 기계 학습 알고리즘의 적용시 준비된 클래스 데이터의 수는 동등하다고 가정하는 것이 일반적이다. 그러나 클래스에 속하는 데이터들의 사전 분포를 알수 없어 이러한 가정의 적용이 어렵다. 또한 일대다(one-vs-all) 모델을 이용한 다중 분류 문제에서 클래스 불균형 학습은 빈번히 나타난다[9].

본 논문에서는 클래스 불균형 문제를 위한 지지벡터기계 알고리즘과 구현을 다룬다. 클래스 불균형 분류 학습에서 지지벡터기계의 문제점에 대한 수행된 결과를 비교 분석한다. 분석 결과로부터 클래스 불균형 학습에서 SMO(Sequential Minimal Optimization,[10]) 알고리즘의 개선 방법을 제시하고, 테스트을 통하여 개선 방법의 효과를 보인다.

이 논문의 2절에서는 클래스 불균형이 기계학습에서 나타나는 문제를 살펴보고 성능 개선을 위해 수행된 연구에 대해 논의한다. 3절에서는 클래스 불균형 학습을 위한 이차 최적화 문제를 살펴보고, 개선된 SMO 알고리즘을 제안한다. 4절에서는 벤치마킹 분류 문제에서 제안된 알고리즘과와 SMO의 학습 성능을 비교하고, 마지막으로 5절에서 결론을 기술한다.

Ⅱ. 관련연구

클래스 불균형 문제를 위해 수행된 연구는 학습 알고리즘에 끼치는 영향 분석[1,2,11]과 학습 알고리즘의 성능 개선 등이 진행되었다[3,4,5,7]. 학습 알고리즘에 미치는 영향을 분석한 연구는 분포를 가정하고 발생된 문제와 알려진 벤치마킹 문제에서 의사결정트리 C4.5[11], 다층 신경망의 오류 역전파 알고리즘[4], 베이지안 알고리즘, 지지벡터기계[5,7] 등의 일반화 성능을 비교 평가하였다. 비교 결과로부터 클래스 불균형이 지지벡터기계 알고리즘의 학습에 미치는 영향은 타 알고리즘에 비해 미비하였다. 지지벡터기계가 학습한 분리 평면은 서로 다른 클래스에서 결정되는 소수의 지지벡터에 따라 결정되기 때문으로 분석된다. 그러나 클래스 불균형 문제에서 지지벡터기계가 학습한 분리 함수는 클래스의 사전 분포를 반영하지 못하므로 긍정 클래스에 근접한 분리 함수가 학습되어, 테스트에서 긍정 클래스의 예측율이 부정 클래스의 데이터의 예측율보다 낮게 평가되는 경향이 있었다[11,13].

지지벡터기계 알고리즘의 개선은 목적함수의 변형[14,15], 학습시 커널함수의 수정[16], 샘플링 기법을 이용한 긍정 클래스의 데이터를 추가 또는 삭제하는 학습 전략[5,6,7,17] 등이 수행되었다. 목적함수의 변형에서는 이차 최적화 문제의 긍정 클래스의 정규화 변수 값을 부정 클래스의 정규화 변수 값과 다르게 부여하여 부정 클래스에 근접하는 평면을 학습시킬 수 있다. KBA 알고리즘은 지지벡터기계가 학습한 분리 평면을 커널행렬을 수정하여 부정 데이터에 가깝게 이동시킨다[16].

GSVM(Granular-SVM)[5], SMOTE+SVM[7] 등 학습 전략에서는 학습 데이터의 준비 단계에서 과도샘플링(oversampling) 또는 과소샘플링(undersampling) 기법을 도입하여 두 클래스의 데이터 수를 동등하게 만든 후 분리 평면을 학습한다. SMOTE+SVM는 이차 최적화 문제의 학습시 긍정과 부정 클래스의 정규화 값을 다르게 부여하는 학습도 병행하였다. 그러나 과도샘플링 기법의 도입은 새로운 데이터가 증가하며서 지지벡터의 수가 증가하는 경향이 분석되었다.

불균형 문제를 위한 지지벡터기계의 성능 향상을 위한 연구를 요약하면 클래스 불균형이 나타나는 분류 문제의 일반화 성능을 높이기 위해 샘플링 기법의 도입은 적용하기 쉬운나 부정 클래스의 과소샘플링은 정보 손실로 인해 전체적인 성능에 영향을 줄수 있다. 긍정 클래스의 데이터를 추가하는 과도샘플링 기법의 도입은 효과가 실험적으로 입증되었으나 새로운 데이터 생성과 학습 시간의 증가 등 부가적 비용이 요구된다.

본 논문의 목적은 불균형 분류 문제의 이차 최적화 문제에 대한 기 수행된 연구 결과를 분석하고 SMO 알고리즘의 개선 방안을 제안하는데 있다. 개선된 SMO는 최적해의 대상인 두 라그랑쥐 승수의 선택 전략과 근접해의 상한과 하한을 결정하는 방법을 이용한다.

III. 불균형 분류 문제의 지지벡터기계 학습

3.1 이차 최적화 문제

주어진 문제 $S=\{(x_i,y_y)|i=1...l, x_i\in R^d, y_i=\pm 1\}$와 커널함수 $K(x,z)=\phi(x)^t\phi(z)$에 대해 지지벡터기계가 찾는 분리 평면 $h(x)$=0는 직교 거리가 최대 마진 γ내에 위치한 긍정과 부정 데이터로부터 학습된다. $\phi(x)$는 x를 선형 분리가 가능한 고차원의 특징공간으로 매핑하는 함수이다. $h(x)$는 1-norm 소프트 마진 이차 최적화 문제 식 (1)의 라그랑지 승수 벡터 α로부터 결정된다[18]. C는 마진과 오류간 관계를 조절하는 정규화 변수로 경험적으로 선택한다.

$$\begin{aligned}\max_\alpha\ & W(\alpha)=\sum_{i=1}^{l}\alpha_i-\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}y_iy_jK(x_i,x_j)\alpha_i\alpha_j \\ & 0\le\alpha_i\le C\ \forall i \\ s.t\ & \sum_{i=1}^{l}y_i\alpha_i=0\end{aligned} \quad \cdots\cdots (1)$$

지지벡터기계는 마진 γ을 최대로 유지하면서 오류를 최소화하는 분리 평면을 학습한다. C가 크면 마진이 최대화되고, 작은 C 값은 작은 마진을 갖는 분리 평면을 결정한다. 식 (1)의 최적해 라그랑지 승수 벡터 α로부터 결정되는 분리 함수는 식 (2)이며, 테스트 벡터 $x\in R^d$의 예측은 $h(x)>0$이면 긍정 클래스, $h(x)<0$ 이면 부정 클래스로 결정한다. b는 $0<\alpha_i<C$인 (x_i,y_i)로부터 추정한다.

$$h(x)=\sum_{i=1}^{l}y_i\alpha_iK(x_i,x)+b \quad \cdots\cdots (2)$$

불균형 데이터의 학습에서 부정과 긍정 데이터에 동등한 C을 부여하면 지지벡터기계는 대부분 데이터를 부정 클래스로 예측하는 분리 평면을 학습하는 경향이 있다. 학습 과정에서 C값은 긍정과 부정 데이터의 오류의 반영 비율이 동등하기 때문에 분리 평면이 긍정 데이터에 근접하게 위치된다. 이 문제를 위해 불균형 문제의 학습에서 긍정과 부정 클래스의 오류 반영에 차이를 주는 이차 최적화 문제가 제안되었다[14,15]. C^+와 C^-는 긍정과 부정 클래스의 오류 반영 비율, a^+와 a^-는 마진의 반영 비율일때, Yang의 이차 최적화 문제는 식 (3)이다[15].

식 (3)에서 $a^+=a^-=1$이면 식 (3)은 Veropoulos가 제안한 이차 최적화 문제[14]와 같으며, $C^+=C^-$이면 식 (1)과 동일하다. 식 (1)의 해는 IPM(interior point method), LOQO, MINOS, SVQP2 등 이차 최적화 알고리즘[19] 또는 SMO 알고리즘[10] 등으로 부터 구할 수 있다.

$$\begin{aligned}\max_\alpha\ & W(\alpha)=\sum_{y_i=+1}\frac{\alpha_i}{a^+}+\sum_{y_i=-1}\frac{\alpha_i}{a^-} \\ & -\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}y_iy_jK(x_i,x_j)\alpha_i\alpha_j \\ s.t.\ & 0\le\alpha_i\le a^+C^+\ \forall y_i=+1 \\ & 0\le\alpha_i\le a^-C^-\ \forall y_i=-1 \\ & \sum_{i=1}^{l}y_i\alpha_i=0\end{aligned} \quad \cdots\cdots (3)$$

이차 최적화 알고리즘은 학습 데이터의 수가 증가하면 커널행열을 위한 대용량 메모리의 사용이 필요하다. 또한 최적해에 근접하는데 많은 학습시간이 필요하나 SMO 알고리즘은 IPM와 LOQO 등에 비해 적은 메모리가 필요하며 구현이 보다 쉽다. 그러나 라그랑지 승수 a_i 의 상한이 상이하여 SMO 알고리즘을 이용하여 식 (3)의 해를 구할 수 없다.

클래스 불균형 문제의 해는 학습된 분리 평면을 결정하는 긍정 클래스의 지지벡터 SV^+와 부정 클래스의 지지벡터 SV^-에서 결정된 지지벡터의 비율에도 차이를 보였다[17]. 학습에서 결정된 SV^+와 SV^-의 α_i들은 식 (4)을 만족하여야 한다.

$$\sum_{i\in SV^+}\alpha_i=\sum_{j\in SV^+}\alpha_j \quad \cdots\cdots (4)$$

식 (4)으로부터 분리 평면의 결정에 긍정 클래스의 지지벡터의 a_i의 합과 부정 클래스의 지지벡터의 합과 같다. 그러므로 $|SV^+|$보다 $|SV^-|$가 크면 긍정 클래스에서 선택된 지지벡터의 a_i 값이 부정 클래스의 a_j 보다 크게 된다.

3.2 개선된 SMO 알고리즘

SMO 알고리즘은 매 반복시 선택된 두 라그랑지 변수 α_i, α_j에 대해 KKT 조건식을 만족하는 해를 찾는 main_routine(), examineExample(i), takeStep(i,j)의 3개의 함수로 구성된다. main_routine()은 KKT 조건을 만족하는 라그랑지 승수을 학습하기 위해 examineExample(i)를 호출하여 라그랑지 승수를 학습시킨다. examineExample(i)는 α_i와 같이 최적해의 대상인 α_j을 선택하는 함수이다. 선택된 α_i와 α_j의 새로운 최적해을 찾는 과정은 takeStep(i,j)에서 수행된다.

SMO 알고리즘의 학습과정은 Platt의 논문[10]에 상세히 기술되어 있다. SMO가 해에 근접하는데 두 단계의 학습 과정이 반복 수행된다. 첫번째 단계에서 새로운 해의 대상인 두 라그랑지 승수 α_i와 α_j를 선택한다. 두번째 단계는 α_i와 α_j의 새로운 해를 계산한다. 이 두 단계는 모든 라그랑지 승수의 변화가 없을때 까지 반복되어 종료된다.

SMO 알고리즘을 이용하여 식 (3)의 해를 구하는데 위해서는 takeStep(i,j)에서 α_i의 상한이 클래스 별 상이하게 부여되어 파라미터 C^+, C^-내에서 라그랑쥐 승수의 해의 범위가 제한된다. examineExample(i)의 $y_i=+1$일때 선택된 두 라그랑쥐 a_i와 최적해의 하한 L과 상한 H은 식 (5)로부터 결정하며 수정된 α_i와 α_j을 새 근접 해를 결정하는 과정은 SMO 알고리즘의 takeStep(i,j)와 같다. $y_i=-1$이면 C^+와 C^-을 바꾸어 식 (5)으로부터 α_i의 L과 H를 계산한다.

$$y_i = y_j \text{ 이면} \begin{cases} L = \max(0, \alpha_i + \alpha_j - C^+) \\ H = \min(C^+, \alpha_i + \alpha_j) \end{cases} \quad \cdots\cdots\cdots\cdots (5)$$

$$y_i \neq y_j \text{ 이면} \begin{cases} L = \max(0, \alpha_i - \alpha_j) \\ H = \min(C^+, C^- + \alpha_i + \alpha_j) \end{cases}$$

지지벡터기계의 분리 평면이 학습에서 선택된 긍정과 부정 클래스의 지지벡터로부터 결정되는 사실을 바탕으로 개선된 SMO 학습의 examineExample(i)에서는 $a_i(y_i=+1)$와 같이 최적화의 대상을 서로 다른 클래스에서 현재의 오류 E_i와 E_j의 절대값 차이가 가장 큰 $a_j(j = \max_{j, y_j=-1} |E_j - E_i|)$을 선택한다. 이러한 선택 전략은 지지벡터기계가 클래스 불균형 문제의 분리 평면을 결정하는 긍정 클래스의 지지벡터 SV^+와 부정 클래스의 지지벡터 SV^-에서 결정된 지지벡터의 비율을 줄이는 효과가 있다.

그림 1은 불균형 분류 문제를 위한 개선된 SMO 알고리즘에서 main_routine()과 examineExample(i)의 함수이다. main_routine에서 선택되면 a_i와 최적해의 대상인 α_i는 examineExample(i)에서 선택된다. a_i의 $y_i=+1$이면 $y_j=-1$인 a_j을 선택하고 takeStep(i,j)를 호출한다. takeStep(i,j)의 α_i의 상한과 하한은 식 (5)으로부터 결정하며 나머지 학습 단계는 SMO와 같다.

그림 2는 인위적으로 발생시킨 선형 분류 문제에서 SMO와 개선된 SMO 알고리즘이 학습한 분리 평면을 비교하고 있다. 평균이 다르나 동일한 분산을 갖는 가우시안 분포를 가정하고 클래스 불균형 비율이 30%인 분류 문제에서 선형 지지벡터기계의 학습의 예이다. $C=2$은 SMO 알고리즘이 학습한 분리 평면이며, $C1=2(C^-)$, $C2=5(C^+)$는 개선된 SMO가 학습한 분리 평면이다. 두 클래스의 분산이 같다면 이상적인 분리 평면은 부정 클래스에 가까운 분리 평면이 적절하다. 그러나 지지벡터의 학습이 클래스의 분포를 고려하지 않고 최대 마진내에서 학습한 분리 평면이 긍정 클래스에 가깝게 나타났다. 이러한 학습 결과는 긍정 클래스의 데이터가 부정 클래스로 예측될 가능성이 높다. 한편, 학습된 분리 평면을 결정하는 긍정과 부정 클래스의 지지벡터의 비율이 25%로 나타나 불균형 비율이 지지벡터의 수에도 영향을 끼친다는 연구 결과와 유사하다[16,17].

```
//데이터 셋 S={(X[i],y[i])|Xi∈R^d,y[i]=±1,i=1,...,l}
// 라그랑지 승수 a[1...l]
// 긍정 데이터 인덱스 pos = { i|y[i]=+1}
// 부정 데이터 인덱스 neg = {i|y[i]=−1}
// 정규화 파라미터 C[pos] = C+, C[neg] = C−
main_routine(S, a, b){
    a[i] = 0 for all i; b = 0;
    numChanged = 0; examineAll = 1;
    while( numChanged > 0 || examineAll ){
      numChanged = 0;
      if( examineAll )
        for( i in 1..l )
          numChanged += examineExample(i)
      else
        for( 0< a[i] < C[i]) // non−bound 지지벡터 i
          numChanged += examineExample(i)
      if( examineAll == 1 ) examineAll = 0
      elseif( numChanged == 0 ) examineAll = 1
    }
}
examineExample(i){
  if( y[i] = +1 ) nominee = neg; else nominee = pos;
  yi = y[i]; alphai = a[i];
  Ei = SVM output on X[i] − yi;
  ri = Ei * yi;
  if(( ri < −tol && alphai < C− ) // tol = 10^-3~10^-6
   || (ri > tol && alphai >0 )){
    if( a non−bound support vector exists
          in a[nominee]){
       j = max{j in nominee} | yiEi − yiEi |
       if( takeStep(i,j) ) return 1;
    }
    for( j in 1...l)
      if(0 <alpha[j]<C[j])){
         if( takeStep(i,j) ) return 1;
      }
  }
  for( j in random(1,l)) // 1..l에서 임의 j을 선택
        if( takeStep(i,j) ) return 1;
  return 0;
}
```

그림 1. 개선된 SMO 학습알고리즘
Fig 1. The improved SMO learning algorithm

VI. 실험

UCI 벤치마킹 문제로부터 생성시킨 이진 분류 문제를 가지고 SMO와 제안된 알고리즘의 학습 결과를 비교하였다.

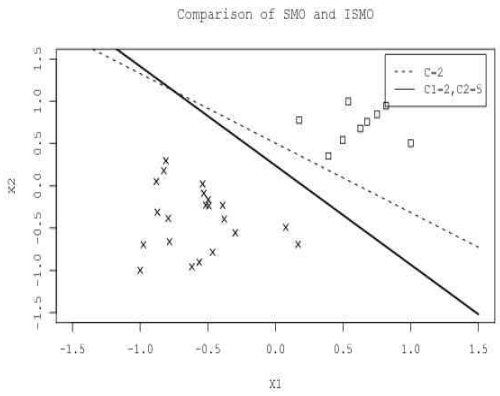

그림 2. 지지벡터기계의 학습의 예
Fig 2. SVM learning example with linear kernel

표 1은 준비된 문제의 긍정 데이터의 수 Pos, 부정 데이터의 수 Neg, 불균형 비율 Im.rate가 제시되었다. Im.rate이 5.0~37.0%까지 다양하게 나타나, 가장 많은 부정 데이터는 Letter7으로 학습 데이터의 약 95.0%이다. 긍정 데이터의 수 백개 이하로 학습 데이터 대비 약 5.0~27.0%로 구성된다.

표 1. UCI 벤치마킹 데이터로부터 이진 분류 문제
Table 1. 2-class classification problems from UCI benchmarking data

Problem	Pos(%)	Neg(%)	Im.rate(%)
Glass7	29(14)	185(86)	16.0
Letter7	226(5)	4,774(95)	5.0
Vehicle1	212(25)	634(75)	33.0
Wine3	48(27)	130(73)	37.0
BalanceBR	49(15)	288(85)	17.0
Adult	395(25)	1,210(75)	33.2

학습 성능은 긍정과 부정 클래스의 학습 예측율을 반영하는 g-mean으로 평가한다. 실험 평가에서 *g-mean*은 긍정 데이터의 예측율은 민감도 *se*(sensitivity)와 부정 데이터의 평가는 특이도 *sp*(specificity)로부터 계산된다. 표 2의 교차테이블로부터 *se, sp, g-mean*의 평가는 식 (4)와 같다.

$$se = TP/(TP+FN), sp = TN/(TN+FN) \quad \cdots\cdots (4)$$

$$g-mean = \sqrt{se*sp}$$

표 2. 교차테이블을 이용한 이진 분류 문제의 학습 평가
Table 2. Learning evaluation of a 2-class classification problem using confusion table

예측 \ 실제	긍정	부정
긍정	TP	FP
부정	FN	TN

표 2으로부터 오류율(error rate) *err*, 정확률(accuracy rate) *acc*, PR(precision-recall) 등 평가가 가능하다[19]. recall은 sp와 같다.

$$err = \frac{FP+FN}{TP+TN+FP+FN}$$
$$acc = 1-err$$
$$pre = \frac{TP}{TP+FP}$$

다중 분류학습의 평가에 주로 사용되는 오류율과 정확율을 불균형 분류 문제의 학습 평가가 사용은 적절하지 않다[4,7]. 클래스 불균형으로 인한 다수 데이터가 속하는 부정 클래스에 치우친 학습은 낮은 오류율과 높은 정확도를 보장하기 때문이다. *g-mean*은 불균형 문제의 평가에 적합하다는 연구로부터 선택되었다.

표 1의 문제의 테스트에서 파라미터 C는 10부터 단계적으로 1까지 감소시켰으며 가장 높은 정확률을 낸 파라미터를 개선된 SMO 학습의 테스트에서 부정 클래스의 파라미터 C^-로 결정하고 긍정 클래스의 파라미터 C^+는 C^-보다 1씩 증가시켜 예측률이 높은 C^+값을 선택하였다. 동일한 테스트을 위해 RBF 커널함수로 선택하고 σ^2=1로 하였다.

표 3은 표 1의 문제에 대한 SMO와 제안한 알고리즘의 비교 결과를 정규화 파라미터 C, C^+, C^-, 민감도 *se*, 특이도 *sp*, g-mean G, 그리고 학습 소요 시간 t으로 제시되었다. SMO의 학습 결과는 뚜렷하게 부정 데이터의 예측율 *sp*가 높이 나타나 학습된 분리 평면이 긍정 클래스에 가까워 긍정데이터에 대한 분류율 *se*이 저조하다. 개선된 SMO의 결과로부터 *se*의 예측율이 높아져 긍정데이터에 대한 정확도가 SMO보다 높게 평가되었다. 이러한 결과는 불균형 데이터의 학습에서 일반화 성능을 높이기 위해서는 긍정데이터의 예측율을 높이는 학습기법이 적절하다는 연구결과[15,17]와 동일하며 불균형 데이터 학습에서 서로 다른 정규화 값의 사용이 효과가 있다.

표 3. SMO와 제안된 알고리즘의 비교
Table 3. The comparision of the SMO and improved SMO al[1] gorithms

Problem	SMO					개선된 SMO					
	C	se	sp	G	t	C+	C-	se	sp	G	t
Glass7	3	0.86	1.00	0.93	1.5	5	3	1.00	1.00	1.00	1.0
Letter7	5	0.00	1.00	0.07	47.5	10	5	0.13	1.00	0.36	25.9
Vehicle1	3	1.00	0.99	0.99	23.5	5	3	0.99	1.00	0.99	20.1
Wine3	1	0.91	0.95	0.95	1.3	3	1	0.98	1.00	0.99	1.8
BalanceBR	2	1.00	1.00	1.00	3.4	5	2	1.00	1.00	1.00	2.3
Adult	5	0.02	1.00	0.14	3.4	10	5	0.04	0.91	0.20	2.9

Letter7의 실험에서 SMO의 긍정데이터의 분류율은 0.0이나 개선된 SMO의 학습에서 분류율이 0.13으로 높아졌으며, Adult의 실험에서도 비슷한 경향이 나타났다. 개선된 알고리즘에서 긍정데이터의 분류율이 높아졌기 때문이다. 학습시간의 비교에서 Wine3의 경우만 제외하고 개선된 SMO의 학습이 빠르게 나타났다. 부정 데이터가 90% 이상 차지하는 Letter7에서 개선된 SMO는 SMO보다 약 2배가 단축되어 긍정과 부정 클래스로부터 현재 최적화 대상 변수의 선택 전략의 효과로 분석된다.

V. 결 론

학습 알고리즘의 응용시 클래스 불균형이 나타나는 문제는 다양한 이유로 나타난다. 클래스 불균형 분류 문제에서 높은 성능을 얻기 위해 샘플링 기법, 학습 알고리즘의 개선 등 관련 연구가 진행되었으나 아직 일반적으로 적용될 수 있는 방법이 밝혀지지 않았다. 이러한 이유는 클래스 불균형 문제는 학습데이터의 준비에서 시작되어 기계학습 응용의 수행 전과정에 영향을 주기 때문이다. 또한 학습을 위해 준비된 다차원의 데이터 자체의 분포 또는 구조화 정보를 학습 알고리즘의 적용시 이용될 수 없기 때문이기도 하다.

본 논문에서는 지지벡터기계의 이차 최적화 문제를 바탕으로 개선된 SMO 학습 알고리즘을 제안하고 벤치마킹 문제를 가지고 SMO 알고리즘과 비교를 하였다. 제안된 알고리즘의 개선은 정규화 파라미터가 클래스 별 다르게 부여되는 이차 최적화 문제의 해결에 SMO 방식의 알고리즘이 적용될 수 있으며, 학습시 두 라그랑지 승수의 선택에서 개선 방안을 제안하였다. 불균형 분류 문제에 대한 실험적 비교 평가에서 개선된 학습 알고리즘은 소수로 구성된 데이터 클래스의 분류율을 높이는데 효과가 있으며, 학습시간의 단축에 효과가 있었다. 제안된 알고리즘은 다중 분류 문제를 위한 학습기 구성에 적용될 수 있을 것으로 기대된다.

참고문헌

[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] Yuchun Tang, Yan-Qing Zhang, N. V. Chawla, and S. Krasser, "SVMs Modeling for Highly Imbalanced Classification," IEEE Transactions on Systems, Man, and Cybernetics, Vol. 39, No. 1, pp. 281-288, 2009.

[8] Yingdong Zhao, Clemencia Pinilla, Danila Valmori, Roland Martin, and Richard Simon, "Application of support vector machines for T-cell epitopes prediction," Bioinformatics Vol. 19, No. 15 2003.

[9] Ryan Rifkin and Aldebaro Klautau, "In Defense of One-vs-All Classification," Journal of Machine Learning Research, Vol. 5, pp. 101-141, 2004.

[10] John C. Platt, "*Fast training of support vector machines using sequential minimal optimization*," Advances in kernel methods: support vector learning, pp. 185-208, MIT Press Cambridge, 1999.

[11] 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.

[12] 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.

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

[14] Veropoulous, C. Campbell, N. Cristianini, "Controlling the Sensitivity of Support Vector Machines," Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence, 1999.

[15] Chan-Yun Yang, Jr-Syu Yang, Jian-Jun Wang, "Margin calibration in SVM class-imbalanced learning," Neurocomputing, Vol. 73, pp.397-411, 2009.

[16] Gang Wu, Edward Y. Chang, "Class-Boundary Alignment for Imbalanced Dataset Learning," ICML, 2003.

[17] Rehan Akbani, Stephen Kwek, and Nathalie Japkowicz, "Applying Support Vector K. Machines to Imbalanced Datasets," Proceedings of 15th European Conference on Machine Learning, pp. 39-50, 2004.

[18] Nello Cristianini and John Showe-Taylor, "*An Introduction to Support Vector Machines and other kernel-based learning methods*," Cambridge University Press, 2000.

[19] L. Bottou and C.-J. Lin. "*Support Vector Machine Solvers*," In Large Scale Kernel Machines, 1-28, MIT Press, 2007.

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

[21] UCI Machine Learning Repository, *http://archive.ics.uci.edu/ml/*

저 자 소 개

김 광 성

2007: 단국대학교 공학사
2009: 단국대학교 전자계산학 석사
2009 - 현재: 현대정보기술
관심분야: 기계학습, 정보검색, 시맨틱웹

황 두 성

1986: 충남대학교 이학사.
1990: 충남대학교 이학석사.
2003: Wayne State University, 박사
2003 - 현재: 단국대학교 컴퓨터과학과 부교수
관심분야: 기계학습, 병렬처리, 데이터베이스, 시맨틱웹