Journal of Korean Institute of Intelligent Systems
Vol. 25, No. 1, February 2015, pp. 015-021
http://dx.doi.org/10.5391/JKIIS.2015.25.1.015

ISSN(Print) 1976-9172
ISSN(Online) 2288-2324

대용량 자료에 대한 서포트 벡터 회귀에서 모수조절

Parameter Tuning in Support Vector Regression for Large Scale Problems

류지열* · 곽민정 · 윤민†

Jee-Youl Ryu, Minjung Kwak, and Min Yoon†

*부경대학교 정보통신공학과, 평택대학교 디지털응용정보학과, † 부경대학교 통계학과

† Department of Statistics, Pukyong National University

요 약

커널에 대한 모수의 조절은 서포트 벡터 기계의 일반화 능력에 영향을 준다. 이와 같이 모수들의 적절한 값을 결정하는 것은 종종 어려운 작업이 된다. 서포트 벡터 회귀에서 이와 같은 모수들의 값을 결정하기 위한 부담은 앙상블 학습을 사용함으로써 감소시킬 수 있다. 그러나 대용량의 자료에 대한 문제에 직접적으로 적용하기에는 일반적으로 시간 소모적인 방법이다. 본 논문에서 서포트 벡터 회귀의 모수 조절에 대한 부담을 감소하기 위하여 원래 자료집합을 유한개의 부분집합으로 분해하는 방법을 제안하였다. 제안하는 방법은 대용량의 자료들인 경우와 특히 불균등 자료 집합에서 효율적임을 보일 것이다.

키워드 : 앙상블 러닝, 서포트 벡터 기계, 부스팅, 대용량 자료 집합, 불균등 자료집합

Abstract

In support vector machine, the values of parameters included in kernels affect strongly generalization ability. It is often difficult to determine appropriate values of those parameters in advance. It has been observed through our studies that the burden for deciding the values of those parameters in support vector regression can be reduced by utilizing ensemble learning. However, the straightforward application of the method to large scale problems is too time consuming. In this paper, we propose a method in which the original data set is decomposed into a certain number of sub data set in order to reduce the burden for parameter tuning in support vector regression with large scale data sets and imbalanced data set, particularly.

Key Words : Ensemble Learning, Support Vector Machine, Boosting, Large Data Set, Imbalanced Data Set.

1. 소 개

최근에 서포트 벡터 기계(support vector machines; SVMs)는 기계학습에서 높은 수행능력을 보이는 것으로 알려져 왔다. 실제 문제들에서 좋은 일반화 능력을 얻기 위하여 서포트 벡터 기계에서 모수의 적절한 값을 선택하는 것이 중요하다.

반면에 SVM에서 모수들의 적절한 값들을 추정하기 위하여 여러 가지 방법들이 제안되었다[2]. 교차타당성(cross validation)방법은 이와 같은 목적을 위하여 일반적으로 많이 사용되어 왔다[1]. 그러나 교차타당성 방법은 시간 소모적이어서 특히 대용량의 자료집합에 적용하기에는 어려움이 따른다.

회귀 목적을 위한 부스팅(boosting)의 선행연구는 [5]에서 확인할 수 있다. 대용량 자료집합에 대한 서포트 벡터 회귀(support vector regression: SVR)에서 모수들의 값들을 결정하는데 대한 부담은 앙상블 학습(ensemble learning)을 이용하여 감소시킬 수 있음이 알려져 있다. 기계학습의 최적화 기법들 중에서 다양한 분야에서 활용되어지고 있는 유전자 알고리즘(genetic algorithm)방법과 비교할 때 제안하는 앙상블 방법을 적용하면 계산 시간을 절약할 수 있는 점 등이 있다. 대용량 자료집합들을 갖는 서포트 벡터 회귀 문제에서 모수 조절에 대한 부담을 감소시키기 위하여 원래의 자료를 몇 개의 부분집합들로 분해한다. 분해된 자료집합에 대해 부스팅 방법의 종류를 적용하면 서포트 벡터 회귀의 모수 조절을 위한 어떤 특별한 부담없이 좋은 근사함수를 쉽게 얻을 수 있다. 더욱이, 배깅(bagging)과 함께 부스팅을 사용하면 이상치(outlier)의 영향력을 감소시킬 수

Received: Dec. 22, 2014
Revised : Feb. 13, 2015
Accepted: Feb. 14, 2015
† Corresponding author(myoon@pknu.ac.kr)
이 논문은 부경대학교 자율창의연구비(2013년)에 의하여 연구되었음.

있음이 알려져 있다[8].

본 논문에서 대용량의 자료집합을 갖는 서포트 벡터 분류(support vector classification; SVC)에 대한 앙상블 방법을 제안한다. 게다가 이 방법은 대용량의 불균등 자료집합으로 확장될 수 있다.

2. 서포트 벡터 회귀

대용량의 자료를 갖는 여러 종류의 서포트 벡터 회귀 방법들은 Schökopf와 Smola에 의하여 제안되었다[7]. SVR의 여러 방법들 중에서 Nakayama 등에 제안된 μ-SVR 모형은 적은 수의 서포트 벡터를 가지므로 이상치에 대하여 덜 민감한 것으로 알려져 왔다[5]. 서포트 벡터 회귀에서 일반적으로 가우지안 커널이 사용된다.

$$K(\boldsymbol{x},\boldsymbol{z}) = \exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{z}\|^2}{2r^2}\right) \tag{1}$$

μ-SVR의 수리계획 모형은 아래와 같다.

$$\begin{aligned} \underset{\boldsymbol{w},b,\xi,\xi'}{\text{minimize}} \quad & \frac{1}{2}\|\boldsymbol{w}\|_2^2 + \mu(\xi+\xi') \\ \text{subject to} \quad & (\boldsymbol{w}^T z_i + b) - y_i \le \epsilon+\xi,\ i=1,\dots,\ell, \\ & y_i - (\boldsymbol{w}^T \boldsymbol{z_i} + b) - y_i \le \epsilon+\xi', \\ & i=1,\dots,\ell, \\ & \epsilon,\ \xi,\ \xi' \ge 0, \end{aligned} \tag{2}$$

여기서 μ는 $\|\boldsymbol{w}\|_2^2$와 최대편차 ξ와 ξ'사이의 트레드 오프(trade-off)를 나타내는 모수이다. 위의 문제에 대한 쌍대 공식은

$$\begin{aligned} \underset{\boldsymbol{\alpha}}{\text{maximize}} \quad & -\frac{1}{2}\sum_{i,j=1}^{\ell}(\alpha_i'-\alpha_i)(\alpha_j'-\alpha_j)K(\boldsymbol{x}_i,\boldsymbol{x}_j) \\ & +\sum_{i=1}^{\ell}(\alpha_i'-\alpha_i)y_i - \epsilon\sum_{i=1}^{\ell}(\alpha_i'+\alpha_i) \\ \text{subject to} \quad & \sum_{i=1}^{\ell}(\alpha_i'-\alpha_i)=0, \\ & \sum_{i=1}^{\ell}\alpha_i' \le \mu, \quad \sum_{i=1}^{\ell}\alpha_i \le \mu, \\ & \alpha_i' \ge 0,\ \alpha_i \ge 0, \quad i=1,\dots,\ell \end{aligned} \tag{3}$$

와 같다. 다른 서포트 벡터 회귀 모형들과 마찬가지로 μ-SVR 또한 일반화 능력의 높은 수행력을 얻기 위하여 가우지안 커널(Gaussian kernel)의 폭 γ의 적절한 값을 결정할 필요가 있다. 예를 들어, $\sin(2\pi x^4)+x,\ x\in[0,2]$와 같은 실제 함수에서 250개의 표본을 무작위로 생성된 회귀문제를 고려하자. μ-SVR의 수행능력은 μ의 값에 민감하지 않음이 알려져 있다[5]. 종종 $\mu=1000$으로 설정한다. 검정용 자료는 구간 $[0,2]$에서 무작위로 1000개를 생성하고 일반화 능력의 성능을 평가하였다. 그림 1에서 점선은 단일 μ-SVR의 가우지안 커널의 폭 γ의 다양한 값에 대한 기준을 제곱근 평균제곱오차(root mean square error; RMSE)를 사용하여 얻어진 성능을 나타내고 실선은 제안한 방법을 이용하여 얻어진 그림이다.

그림 1에서 쉽게 알 수 있듯이, 매우 작은 영역에서 γ의 적절한 값이 위치한다. 따라서 이 실험에서 회귀의 성능은 γ의 값에 매우 민감하다. 교차타당성(cross validation)방법을 이용하여 적절한 γ의 값을 찾기에는 매우 어렵고 시간소모적이다. 이런 어려움을 극복하기 위하여 부스팅(boosting)을 이용하는 방법을 제안한다. 제안된 방법은 대용량 자료집합들에 대하여 현저히 계산시간을 절약하고 좋은 일반화 능력을 얻는 장치이다. 자세한 내용은 다음 절에서 상세하게 설명한다.

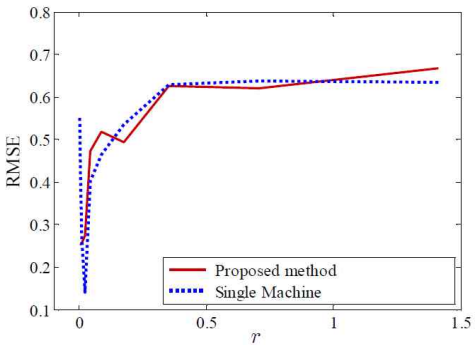

그림 1. 제곱근 평균제곱오차 대 가우지안 커널의 폭
Fig. 1. RMSE v.s. the width of Gaussian kernel($\gamma_1 = \sqrt{2}$)

3. μ-SVR에서 모수조절을 위한 부스팅 방법

대용량 자료집합을 가진 문제들에 대하여 원 자료 집합을 몇 개의 부분집합으로 분해한다. 우선 전체 훈련 자료집합을 $S=\{(\boldsymbol{x}_i, y_i)|i=1,2,\dots,N\}$ 라 두자. S를 $S=S_1\cup S_2\cup\cdots\cup S_n$와 같이 유한개의 부분집합으로 분해하고 여기서 S_k의 크기는 s에 의해 주어진다.

가우지안 커널의 k번째 약 학습기(weak learner)의 폭은 γ_k로 나타낸다. 모수조절을 위한 제안한 방법에 있어 상대적으로 큰 γ_1를 가지고 시작한다. 이후의 모수들은 $\gamma_k < \gamma_{k-1}$의 관계를 유지한다.

다음은, S_k원소의 인덱스 집합을 I_k로 나타내면

$$I_k = \{m \bmod N \mid m=(k-1)s+1,\dots,ks\} \tag{4}$$

와 같고 그러면 I_k와 I_{k-1}의 중첩을 피할 수 있다.

$\hat{f}_k(\boldsymbol{x}|S_k)$는 훈련 자료집합 S_k에 대한 k번째 약 학습기의 출력이라고 하자. 게다가 S_k는

$$\begin{aligned} S_1 &= \{(x_i,y_i) \mid i\in I_1\}, \\ S_k &= \{(x_i, y_i - \hat{f}_{k-1}(x_i|S_{k-1})) \mid i\in I_k\},\ k=2,3,\dots. \end{aligned} \tag{5}$$

에 의하여 주어진다. 그러면, M 폴더(fold) 부스팅 출력값은

$$\hat{f}(x) = \hat{f}_1(x|S_1) + \hat{f}_2(x|S_2) + \dots + \hat{f}_M(x|S_M) \tag{6}$$

와 같이 주어진다.

다음은 위의 (1)에 부스팅 방법을 적용한다. $\gamma_1 = \frac{1}{\sqrt{2}}$ 로 시작하면, γ_k의 값은 다음과 같이 감소된다:

$$\gamma_k = 0.5\gamma_{k-1}, \quad k = 2,3,\ldots \ . \tag{7}$$

전체 훈련자료(training data)의 수는 250개 이고 $s = 115$로 설정한다. 부스팅은 10회 수행하였고 그 결과는 그림 1의 실선으로 나타내었다.

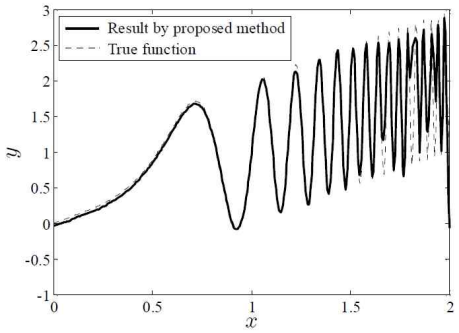

그림 2. 앙상블 서포터 벡터 회귀(RMSE = 0.0283)
Fig. 2. Ensemble Support Vector Regression (RMSE = 0.0283)

그림 3에서 단일 기계에 의해 얻어진 최량의 결과와 비교하여도 정확도에 있어서 만족할만한 9번째 부스팅된 기계에 의해 얻어진 근사함수는 그림 2에서 나타낸 것과 같다.

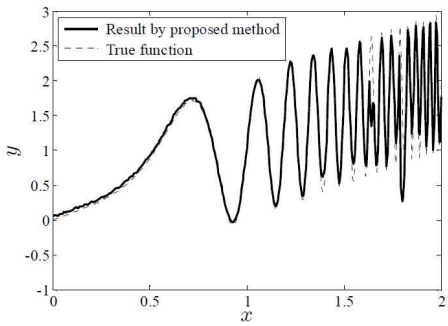

그림 3. $\gamma = 0.0221$에서 최량의 단일 서포터 벡터 회귀 (RMSE = 0.0188)
Fig. 3. Best single Support Vector Regression with $\gamma = 0.0221$ (RMSE = 0.0188)

그림 1에서 알 수 있는 바와 같이 제안한 방법은 γ값의 변화에 대하여 안정적이고 최량의 결과를 얻은 후에 부스팅을 하여도 RMSE의 값은 증가하지 않는다. 달리 말하면 비록 γ 초기 추정값이 대략적이라 하더라도 제안한 방법에 의하여 쉽게 좋은 근사함수를 얻을 수 있다.

게다가 제안된 방법은 좋은 근사함수를 대단히 빨리 얻을 수 있다. 표 1은 훈련자료의 수와 계산시간의 관계를 나타낸다. 전체 훈련 자료집합을 가지고 단일 μ-SVR의 경우는 각 γ값에 대략 19초의 시간이 소요되었다. 그러므로 교차타당성 검정을 10번 시행하여 사용된 총 소요 시간은 대략 190초 정도이다. 한편 부분 자료집합으로 분리된 경우에 계산 소요시간은 2초가 걸렸다. 이는 10개 폴더 부스팅에 대한 계산시간은 대략 20초 전후임을 나타낸다. 일반적으로, 계산시간은 자료집합의 크기 전체에 대해 지수적으로 증가한다. 그러므로. 제안된 방법은 대용량의 자료집합들에 적용이 용이하게 된다.

표 1. 훈련자료의 수 v.s. 계산시간
Table 1. The number of training data v.s. calculation time

Number of training data	250	115
Calculation time	19 seconds	2 seconds

[주의] [6] μ-SVR 자체는 어떤 정도의 이상치에 대하여 강건(robust)하다는 사실이 알려져 있다. 배깅(bagging)을 사용하는 경우에 이상치의 영향력을 좀 더 감소시킬 수 있다.

4. 대용량의 불균등 자료의 분류를 위한 앙상블 학습

본 절에서, 대용량의 불균등(imbalanced) 자료집합을 갖는 분류문제들에 앙상블 학습을 적용하자. 비록 SVM에 부스팅의 적용을 시도한 경우가 여러 차례 있었으나[7], 대용량 자료에 대한 SVM에서 모수조절에 대해 시도된 경우는 없었다. 분류에 대한 가우지안 커널의 폭의 적절한 값을 결정하는 것은 회귀분석과 마찬가지로 중요하고도 어려운 문제이다. 대용량의 자료집합을 유한개의 소용량 부분집합들로 나눈다. 그런 후에, 부스팅과 배깅을 갖는 어떤 앙상블 학습은 효과적으로 불균등 자료집합을 처리할 수 있다. 아래에서 좀 더 자세히 나타낼 것이다.

우선, 저자에 의해 제안되었던 μ-ν-SVM를 대략적으로 소개한다[5]. μ-ν-SVM의 아이디어는 가장 나쁜 여유도(slackness)를 최소화하는 ν-SVM[8]과 가장 나쁜 잉여도(surplus)를 최대화하는 μ-SVM[5]를 결합하여 아래와 같은 공식화가 얻어진다.

$$\underset{\boldsymbol{w},b,\rho,\sigma}{\text{minimize}} \quad \frac{1}{2}\|\boldsymbol{w}\|_2^2 - \nu\rho + \mu\sigma \tag{8}$$

여기서 ν와 μ는 모수이다.

비선형 계획법에서 쌍대성으로부터, 다음의 쌍대 최적화 문제를 얻는다.

$$\begin{aligned} \underset{\boldsymbol{\alpha}}{\text{maximize}} \quad & -\frac{1}{2}\sum_{i,j=1}^{l}\alpha_i\alpha_j y_i y_j K(\boldsymbol{x}_i,\boldsymbol{x}_j) \\ \text{subject to} \quad & \sum_{i=1}^{l}\alpha_i y_i = 0 \\ & \nu \le \sum_{i=1}^{l}\alpha_i \le \mu, \\ & \alpha_i \ge 0, i = 1,\ldots,l. \end{aligned} \tag{9}$$

[주의] [5] μ-ν-SVM의 수행은 ν와 μ의 값들에 많이 민감하지 않다고 관측되어왔고 더 큰 값들의 μ는 커널 폭의 적절한 값의 더 넓은 범위들을 제공한다.

이제, 대용량의 자료집합을 갖는 분류에 대한 앙상블 학습에 μ-ν-SVM을 적용한다. 그림 4에서 보는바와 같이, 배깅과 부스팅 둘을 갖는 앙상블 학습을 제안한다. 부스팅의 각 단계에서 부스팅된 학습 기계는 배깅에 대한 복수의 약 학습기들로 구성한다. 배깅과 부스팅에 대한 두 종류의 약 학습기들을 구별하기 위하여 부스팅의 l-번째 약 학습기는 l-번째 층의 기계에 적용된다. 회귀분석과 유사한 방식으로 부스팅은 가우지안 커널의 폭을 자동적으로 조절하는데 효과적인 역할을 하고, 이는 대용량의 자료집합에 대한 고성능의 일반화를 제공한다.

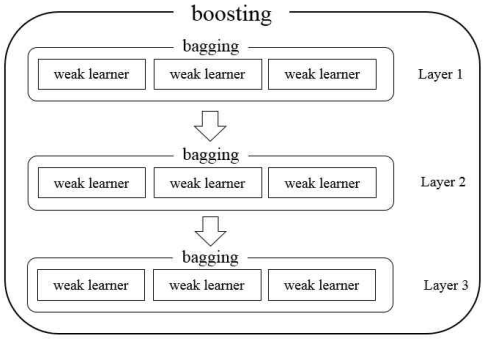

그림 4. 배깅과 부스팅을 이용한 앙상블 학습
Fig. 4. Ensemble Learning using Bagging and Boosting

부스팅의 각 단계에서, D는 주어진 학습 자료집합이라 하자. 배깅을 하기 위해 D를 붓스트랩 표본추출에 의하여 생성된 여러 개의 부분 자료집합 $D_k(k=1,...,K)$로 나눈다. $\hat{f}(\boldsymbol{x}|D_k)$를 D_k의 약 학습기의 출력이라고 나타낸다. 그러면, 배깅에 의한 전체 학습기계의 출력은 이진 분류문제들에 대한 $\text{sign}\left(\hat{f}(\boldsymbol{x}|D_k)\right)$ $(k=1,...,K)$들 중에서 다수결의 원칙(majority rule)에 의해 주어진다.

대용량의 자료집합을 갖는 분류문제에 대한 부스팅의 방법으로 잘 알려진 아다부스트(AdaBoost)[3]를 적용한다. 전체 훈련자료는 이진분류에 대해 $y_n=1$ 또는 -1 을 가지는 $\{(\boldsymbol{x}_1,y_1),\cdots,(\boldsymbol{x}_s,y_s)\}$라 하자. 이전 층에서 얻어진 더 큰 가중치들에서 오분류된(misclassified) 표본들에 대해 조정된 각각의 훈련자료에 대한 가중치를 고려하자. 그 다음의 층은 큰 가중치를 갖는 훈련 자료들의 평균값으로 학습한다.

첫 번째 층에 대해 초기 가중치를 $w_i^{(1)}=\frac{1}{s}$ $(i=1,...,s)$라 하자. 말하자면, 각각의 훈련자료는 부스팅의 초기상태에서 동등하게 처리된다. $h_m(\boldsymbol{x})$를 m번째 층의 출력이라 나타내고 여기서 $h_m(\boldsymbol{x})=1$ 또는 -1이다. $h_m(\boldsymbol{x}_i)=y_i$일 때 $I(h_m(\boldsymbol{x}_i)=y_i)=1$을 나타내고 그렇지 않으면 0이다. 그러면, m번째 층의 분류율은

$$\epsilon_m=\sum_{i=1}^{s}w_i^{(m)}I(h_m(\boldsymbol{x}_i)=y_i) \tag{10}$$

에 의해 주어진다.

더욱이, m번째 층의 신뢰(confidence)는

$$\alpha_m=\frac{1}{2}\log\left(\frac{1-\epsilon_m}{\epsilon_m}\right) \tag{11}$$

에 의해 정의된다.

그러면 각 훈련자료에 대한 가중치는

$$w_i^{(m+1)}=\frac{w_i^{(m)}\exp(-\alpha_m y_i h_m(\boldsymbol{x}_i))}{\sum_{j=1}^{s}w_j^{(m)}\exp(-\alpha_m y_j h_m(\boldsymbol{x}_j))},\ i=1,...,s \tag{12}$$

에 의해 갱신(update)된다.

위의 절차를 M번 반복하면, 마지막 층은

$$Y_M(\boldsymbol{x})=sign\left(\sum_{m=1}^{M}\alpha_m h_m(\boldsymbol{x})\right) \tag{13}$$

와 같이 주어진다.

5. 수치 예제들

본 실험에서는 부스팅의 각 층에서 가중치에 비례한 룰렛 선택(roulette)에 의하여 학습표본들을 선택한다. 말하자면, 그 다음 층에서 학습을 위하여 높은 확률을 갖는 오분류된 표본들이 선택된다. 따라서 $\mu-\nu$-SVM 학습에 대하여 $\mu=1000$ 그리고 $\nu=1$를 가지고 실험을 수행하였다.

5.1 균형 자료집합

실제 분류함수를 $\sin(3x_1)-x_2=0$라 가정하고 범위는 $[-\pi,\pi]\times[-1,1]$이다. 전체 훈련자료 집합은 $[-\pi,\pi]\times[-1,1]$에서 무작위로 생성된 1000개의 실험점으로 구성된 D라 하자. 위와는 분리하여 동일한 영역에서 5000개의 검정용 자료를 생성하였다. 부스팅의 각 단계에서 가우지안 커널 의 폭 γ는 $r_k=0.5r_{k-1}$ $(k=2,3,...)$에 의하여 감소하며 $\gamma_1=5$로 시작한다. 10번의 부스팅 후에 반복이 끝난다. m번째 층에서 훈련용 표본들의 집합 D^m은 원래 훈련 자료집합 D에서 룰렛 선택으로 얻어진 300개의 자료로 구성된다. 그리고 배깅을 위한 훈련 자료집합 $D_j^m,(j=1,2,3)$에 대해 3개의 약 학습기를 고려한다. 각 집합 $D_j^m,(j=1,2,3)$는 붓스트랩에 의해 D^m에서 선택된 100개의 표본점들로 구성된다.

그림 5는 폭 γ의 값에 대한 분류성능을 나타낸다. 회귀분석과 유사하게 분류에 있어서 단일 서포트 벡터기계에서 커널의 폭 γ의 적절한 값을 결정하기는 다소 어렵다. 제안한 앙상블 방법은 γ의 추정치에 특별히 주의할 필요가 없이 좋은 분류성능을 제공한다는 점을 알 수 있었다.

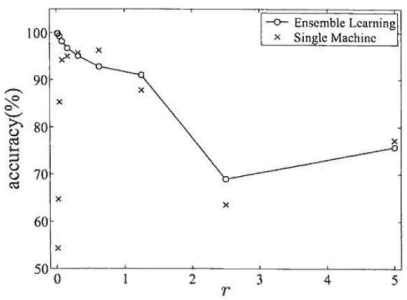

그림 5. 정확성 대 커널의 폭 γ (균형 자료집합)
Fig. 5. Accuracy v.s. γ (well balanced data set)

그림 6과 그림 7은 각각 제안한 방법과 단일 서포트 벡터기계에 의한 분류 경계(점선)를 나타낸다. 배깅을 이용한 앙상블 기계는 이상치에 대하여 강건할 것으로 기대된다. 게다가 제안된 방법은 교차 타당성을 사용한 단일 서포트 벡터기계보다도 시간이 훨씬 적게 소요된다.

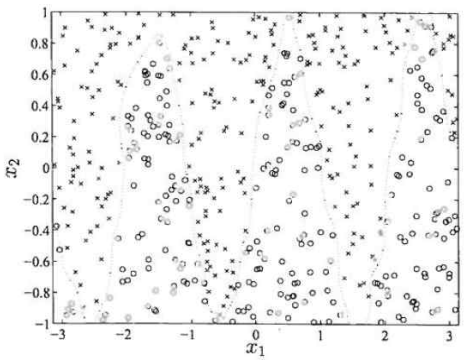

그림 6. 제안된 방법에 의한 분류경계
Fig. 6. Discrimination boundary by the proposed method

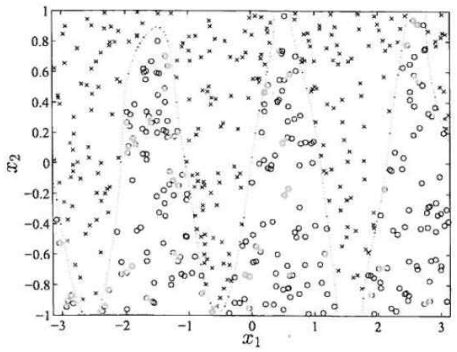

그림 7. 최량의 단일 서포트 벡터 기계에 의한 분류경계 ($\gamma=0.625$)
Fig. 7. Discrimination boundary by the best single support vector machines ($\gamma=0.625$)

[주의] 사전에 적절한 γ의 값에 대한 정보가 없을 때, [5]에서 저자들은 γ의 대략적인 추정값으로

$$r=\frac{d_{max}}{\sqrt[n]{nm}} \tag{14}$$

를 제안하였다. 여기서, d_{max}는 자료들 사이에서의 최대거리, n은 입력 자료의 차원을, m은 표본점들의 개수를 나타낸다. 이와 같은 단순한 추정 값은 제안된 앙상블 학습에서 γ의 초기 값으로 사용될 수 있다.

5.2 불균형 자료집합의 처리

실제 문제에서, 각 범주들 사이에 표본들의 수가 다소 불균형인 분류문제를 처리하는 경우가 빈번하게 발생한다. 서포트 벡터 기계 학습에서 소수 집단의 분류율이 불충분한 경우가 쉽게 발생하게 된다. 제안된 앙상블 학습은 불균형 자료집합과 같은 경우에 적용할 수 있음을 보일 것이다.

본질적인 아이디어는 각 층에서 배깅을 위한 약 학습기들에 소수 집단을 할당하는 것과 같이 다수 집단의 학습 표본들의 크기를 거의 동일하게 만드는 것이다. 불균형 자료집합들을 가지는 경우에 소수 집단의 표본들은 m번째 층의 훈련 자료집합 D^m에서 선택할 필요가 있다. 다수 집합을 고려하면, 훈련 자료들은 룰렛 선택에 의해 결정된다. 각 층에서 약 학습기에 대해, 다수 집단의 훈련자료들의 수는 소수 집단에서와 같이 거의 동일하게 되어야 한다.

5.3 예제 1

실제 분류함수는 $\sin(3x_1)-x_2=0$에 의해 주어진다고 가정하고 범위는 $[-\pi,\pi]\times[-1,1]$이다. 전체 훈련자료들은 $[-\pi,\pi]\times[-1,1]$에서 무작위로 1000개의 점들을 생성하였고, 여기서 한 집단의 자료의 수는 100개이고 다른 집단의 자료의 수는 900개이다.

각 층에서 배깅을 위한 3개의 약 학습기를 설정하였고, 소수 집단의 100개의 표본들은 3개의 약 학습기의 훈련 자료집합에서 선택되어야 한다. 다수 집단의 300개의 표본들은 가중치에 따라서 룰렛 선택에 의해 선택된다. 위의 300개의 표본들은 배깅을 위해 3개의 약 학습기들에 할당된다.

가우지안 커널 γ의 폭은 $r_k=0.5r_{k-1}$ $(k=2,3,...)$에 의해 감소된다. γ의 초기 추정값은 (14)를 이용하여 대략적인 추정값으로서 $r_1=0.0594$로 설정하였다. 10번의 반복 후에 부스팅을 종료한다. 그 결과는 표 2에 표시하였다.

표 2. 예제 1에 대한 정확도(γ의 초기값은 0.0594)
Table 2. Accuracy for example 1(initial value of γ : 0.0594)

	Proposed method		
r	Overall	Majority	Minority
0.0594	86.0%	93.7%	78.3%
0.0297	88.1%	95.7%	80.5%
0.0149	88.6%	95.5%	81.8%
0.0074	88.8%	95.3%	82.4%
0.0037	88.8%	95.2%	82.5%
0.0019	88.8%	95.2%	82.5%
0.0009	88.8%	95.2%	82.5%
0.0005	88.8%	95.2%	82.5%
0.0002	88.8%	95.2%	82.5%
0.0001	88.8%	95.2%	82.5%
	Single SVM		
r	Overall	Majority	Minority
0.0594	64.5%	99.6%	29.5%
0.0297	53.6%	99.9%	7.3%
0.0149	50.9%	100.0%	1.8%
0.0074	50.2%	100.0%	0.4%
0.0037	50.0%	100.0%	0.1%
0.0019	50.0%	100.0%	0.0%
0.0009	50.0%	100.0%	0.0%
0.0005	50.0%	100.0%	0.0%
0.0002	50.0%	100.0%	0.0%
0.0001	50.0%	100.0%	0.0%

단일 서포트 벡터 기계에 대한 γ의 초기 추정치는 표 2에서 나타나듯이 너무 작아 보인다. 그러므로 더 큰 초기 추정치 $\gamma=1.901$를 가지고 또 다른 실험을 수행하였다. 그 결과는 표 3에 나타내었다.

표 3. 예제 1에 대한 정확도(γ의 초기값은 1.901)
Table 3. Accuracy for example 1(initial value of γ : 1.901)

	Proposed method		
r	Overall	Majority	Minority
1.901	82.6%	95.8%	69.4%
0.950	94.0%	93.9%	94.1%
0.475	94.3%	95.4%	93.2%
0.238	94.4%	96.0%	92.9%
0.119	94.4%	96.7%	92.2%
0.059	94.4%	97.3%	91.7%
0.030	94.5%	97.6%	91.6%
0.015	94.6%	97.6%	91.6%
0.007	94.6%	97.6%	91.6%
0.004	94.6%	97.6%	91.6%
	Single SVM		
r	Overall	Majority	Minority
1.901	63.8%	69.1%	58.5%
0.950	94.7%	98.6%	91.0%
0.475	93.3%	99.4%	87.2%
0.238	92.5%	99.5%	85.5%
0.119	86.7%	99.5%	73.9%
0.059	64.4%	99.7%	29.1%
0.030	53.5%	99.9%	7.1%
0.015	50.9%	100.0%	1.8%
0.007	50.2%	100.0%	0.4%
0.004	50.0%	100.0%	0.1%

그림 8은 제안한 방법으로 얻어진 분류경계를 나타내고, 반면에 그림 9는 $\gamma=0.950$을 갖는 단일 서포트 벡터기계에서 최량의 분류경계를 표현했다.

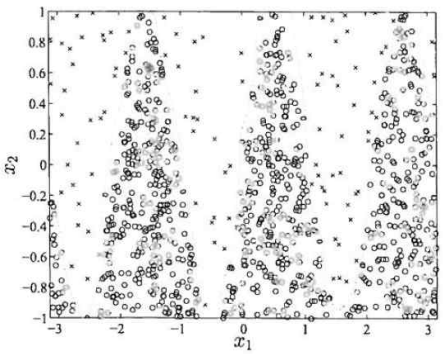

그림 8. 제안한 방법의 분류경계
Fig. 8. Discrimination boundary by the proposed method

다음과 같은 실험들로부터 단일 서포트 벡터기계는 소수집단에 대한 타당한 정확도를 얻기 위하여 커널의 폭 γ의 적절한 값을 결정하기가 쉽지 않다. 그러나 제안된 방법은 γ의 초기 추정치를 얻기 위하여 추가적인 수고를 할 필요없이 다수 집단과 소수 집단 양쪽 모두에 좋은 수행능력을 보인다.

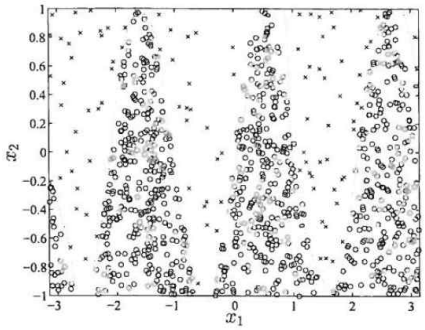

그림 9. 최량의 단일 서포트 벡터 기계에 의한 분류경계 ($\gamma=0.950$)
Fig. 9. Discrimination boundary by the best single support vector machines ($\gamma=0.950$)

5.4 예제 2

http://archive.ics.uci./ml/에서 피마 인디언 당뇨병 자료(PIMA)는 잘 알려진 벤치마크 문제이다. 자료집합은 다수집단으로 500개의 자료가 있고 소수집단의 표본들은 268개로 구성된다. 실험을 위하여 다수집단에서 400개의 자료를 무작위로 추출하고 훈련자료로서 소수집단으로부터 150개의 표본을 선택하였다. 부스팅과 배깅은 예제 1에서와 동일한 방법으로 수행하였다. 그 결과는 표 4에 나타내었으며 γ의 초기값은 (14)에 의해 주어졌다. 이 실험 또한 제안한 앙상블 방법에 의하여 다수집단 및 소수집단 양쪽 모두 타당한 정확도를 얻을 수 있음을 알 수 있었다.

표 4. PIMA 자료에 대한 정확도
Table 4. Accuracy for PIMA data

	Proposed method		
r	Overall	Majority	Minority
4.102	68.1%	67.5%	68.6%
2.051	68.9%	67.1%	70.3%
1.025	71.6%	68.2%	74.5%
0.513	72.4%	65.9%	78.0%
0.256	72.5%	65.7%	78.2%
0.128	72.6%	65.7%	78.4%
0.064	72.6%	65.5%	78.6%
0.032	72.5%	65.2%	78.7%
0.016	72.5%	65.0%	78.8%
0.008	72.5%	65.0%	78.8%
	Single SVM		
r	Overall	Majority	Minority
4.102	62.9%	76.9%	40.8%
2.051	64.1%	78.5%	38.9%
1.025	63.0%	83.0%	33.3%
0.513	48.0%	96.4%	23.7%
0.256	45.8%	99.9%	2.5%
0.128	45.9%	100.0%	0.0%
0.064	45.9%	100.0%	0.0%
0.032	45.9%	100.0%	0.0%
0.016	45.9%	100.0%	0.0%
0.008	45.9%	100.0%	0.0%

6. 결론

본 논문은 대용량 서포트 벡터기계에서 모수 조절에 부스팅과 배깅을 모두 사용하는 앙상블 방법을 제안하였고 또 제안한 방법은 대용량의 불균형 자료집합의 문제에도 효과적으로 확장할 수 있음을 보였다. 여러 개의 수치실험을 통하여 제안한 방법은 대용량 자료집합에 대하여 계산시간을 현저하게 절약할 수 있음을 알 수 있었다. 특히, 제안한 방법은 차례로 커널 폭의 적절한 값의 추정치에 대해 어떠한 노력 없이 타당한 수행능력을 가진 분류기와 회귀 분석을 제공한다는 점을 주목할 수 있었다. 실제 기계학습에서 모수 조절은 중요한 문제이고 추가적으로 실제 문제에서 불균형 자료집합들을 종종 직면하므로 제안한 방법은 많은 실제 문제들에 효과적으로 적용할 수 있을 것으로 기대된다. 향후 연구 과제로써 앙상블 기반의 분류 알고리즘으로 랜덤 포리스트(random forest)와 성능 비교를 통한 연구가 가능하리라 생각된다.

References

[1] C.M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.

[2] O. Chapell, V. Vapnik, O. Bousquet and S. Mukherjee, "Choosing multiple parameters for support vector machines," *Machine Learning*, vol. 46, no. 1-3, pp. 131-159, 1997.

[3] Y. Freund and R. Schapire, "A decision-theoretic generalization of online learning and an application to boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119-139, 1997.

[4] J. M. Moreira, C. Soares, A. B. Jorge and J. F. Sousa, *Ensemble Approaches for Regression: a Survey*, Elsevier, 2007.

[5] H. Nakayama, Y. B. Yun and M. Yoon, *Sequential Approximate Multiobjective Optimization Using Computational Intelligence*, Springer, 2009.

[6] X. Li, L. Wang and E. Sung, "AdaBoost with SVM-based component classifiers," *Engineering Applications of Artificial Intelligence*, vol. 21, no. 5, pp. 785-795, 2008.

[7] B. Schölkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, MIT Press, 2002.

[8] R. Suzuki, H. Nakayama and Y. B. Yun, "Parameter tuning in support vector regression for large scale problems," *International Conference on Optimization, Techniques and Applications*, Shanghai, 2010.

저 자 소 개

류지열(Jee-Youl Ryu)

1993년 : 부경대학교 전자학과 공학사
1997년 : 부경대학교 전자학과 공학석사
2004년 : Arizona대학교 전기공학과 공학박사
2009년 ~ 현재 : 부경대학교 정보통신공학과 부교수

관심분야 : Communication SoC design, RF/Analog IC design
Phone : +82-51-629-6239
E-mail : jyryu@pknu.ac.kr

곽민정(Minjung Kwak)

1988년 : 연세대학교 응용통계과 경제학사
1990년 : 연세대학교 응용통계과 경제학석사
2015년 : Minesota대학교 생물통계학과 생물통계학박사
1998년 ~ 현재 : 평택대학교 디지털응용정보학과 교수

관심분야 : Categorical data analysis, Missing data imputation
Phone : +82-31-659-8355
E-mail : mjkwak@ptu.ac.kr

윤민(Min Yoon)

1994년 : 부경대학교 응용수학과 이학사
1996년 : 부경대학교 응용수학과 이학석사
2002년 : 연세대학교 응용통계학과 통계학박사
2009년 ~ 현재 : 부경대학교 통계학과 부교수

관심분야 : Machine Learning, Evolutional computation, Soft Computing
Phone : +82-51-629-5540
E-mail : myoon@pknu.ac.kr