Journal of the Korean Data & Information Science Society
2010, **21**(3), 461–470

한국데이터정보과학회지

최소제곱 서포터벡터기계 형태의 준지도분류[†]

석경하[1]

[1]인제대학교 데이터정보학과

접수 2010년 4월 8일, 수정 2010년 5월 12일, 게재확정 2010년 5월 19일

요 약

라벨 있는 자료가 분류규칙을 만들 만큼 충분하지 않거나, 라벨 없는 자료가 분류규칙을 만드는데 도움을 줄 수 있는 경우에는 라벨 있는 자료와 라벨 없는 자료를 모두 사용하는 준지도분류가 더 효과적이다. 준지도분류 중 그래프기반 다양체정칙법이 개발되어 최근에 많은 연구가 이루어지고 있다. 본 연구에서는 통계적학습에서 좋은 성능을 보이는 최소제곱 서포터벡터기계를 준지도분류에 적용시키는 방법을 제안한다. 모의실험을 통해 제안된 방법이 라벨 없는 자료를 잘 활용하는 것을 볼 수 있었다.

주요용어: 그래프기반 준지도분류, 그래프 라플라시안, 다양체정칙성, 준지도분류, 최소제곱 서포터벡터기계.

1. 서론

지도학습 (supervised learning)과 자율학습 (unsupervised learning)에 관한 연구는 오래전부터 이루어져 왔다. 회귀분석 (regression analysis)과 분류 (classification analysis)는 대표적인 지도학습이고 군집분석 (clustering)은 자율학습의 대표적인 분석방법이다. 이렇게 분석방법을 나누는 기준은 반응값 (response), 목표값 (target) 혹은 라벨 (label)의 유무에 따른 것이다. 본 연구에서는 분류에 관한 것만 다루기 때문에 '라벨'을 사용하고 라벨이 $\{1, -1\}$의 값을 갖는 이진분류를 고려한다. 최근에 아주 적은 비율의 라벨만 가지는 자료가 많이 만들어지고 있다. 그 이유는 라벨을 만드는 것이 어렵고 시간과 경비가 많이 소요되거나 자료가 너무 커서 모든 자료의 라벨을 만드는 것이 힘들기 때문이다. 폭발적으로 발생하는 스팸메일의 라벨을 모두 정하는 것은 불가능할 것이다. 그리고 환자들로부터 얻어지는 ECG (electrocardiogram) 자료 중 일부분만이 정상 혹은 비정상이라는 라벨을 가질 것이다. 이외에도 음성인식 (speech processing) (시간과 경비문제), 문자분류 (text categorization)와 웹분류 (web categorization) (시간과 자료의 크기), 그리고 생명정보 (bioinformatics) (비용, 시간) 등을 들 수 있다.

이러한 자료를 분석하기 위해 준지도분류 (semi-supervised classification)라는 새로운 방법이 최근에 개발되어 많은 연구가 이루어지고 있다 (Zhu, 2005; Chapelle 등, 2006; Seeger, 2001; Huang 과 Kecman, 2004). 준지도분류는 라벨 있는 자료 (labeled data)와 라벨 없는 자료 (unlabeled data)를 모두 사용하여 라벨 없는 자료 혹은 시험자료 (test data)의 라벨을 추정하는 방법이다. 지도학습의 분류처럼 라벨 있는 자료를 이용하여 분류규칙을 만든 후 라벨 없는 자료의 라벨을 추정할 수 있지만 라벨

[†] 이 논문은 2009년도 정부 (교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (2009-0072369).

[1] (621-749) 경남 김해시 어방동, 인제대학교 데이터정보학과, 통계정보연구소, 교수.
E-mail: statskh@inje.ac.kr

있는 자료가 분류규칙을 만들 만큼 충분하지 않거나, 라벨 없는 자료가 분류규칙을 만드는데 도움을 줄 수 있는 경우에는 준지도분류를 사용하는 것이 더 효과적이다.

준지도분류에 대해 많이 연구되고 있는데 그 중 가장 먼저 개발된 방법은 혼합 분포 (mixture distribution)에 기반한 모형인데 Nigam 등 (2000)이 EM 알고리즘을 이용하여 문자인식에 이 방법을 적용하였다. 그 후에 Holub 등 (2005)은 자료를 특징공간으로 변환하여 이미지분류작업을 하는데 이 방법을 사용하였다.

자신훈련 (self-training) 방법의 가장 큰 단점인 '한번 잘못된 규칙은 계속 다른 자료에 영향을 미칠 수 있다'는 것을 보완할 수 있는 방법으로 개발된 것이 공동훈련 (co-training)이다 (Blum 과 Mitchell, 1998). 이 방법은 변수들이 두 개의 집합으로 나누어 질 수 있고, 각 집합에서 분류기 (classifier)가 충분히 훈련될 수 있다는 가정을 하고 있다. Nigam과 Ghani (2000)과 Zhou과 Goldman (2004)이 이 방법을 더욱 발전시켰는데 최근에는 이 방법을 활용한 준지도회귀분석 (semi-supervised regression analysis)이 개발되었다 (Zhou과 Li, 2007).

최근에 주목을 많이 받고 있는 방법은 그래프기반 (graph-based) 준지도분류 (Zhou 등, 2003; Zhou 등, 2004; Belkin 등, 2006)이다. 이 그래프기반 준지도분류를 해결하기 위한 알고리즘으로는 최소절단 (minicut) 방법, 조화함수 (harmonic function) 사용법, 국소와 전역 일치법 (local and global consistency) 그리고 다양체정칙법 (Belkin 등, 2006)이 있는데 본 연구에서는 그래프기반 방법을 좀 더 일반화한 것으로 알려진 다양체정칙법을 사용한다.

다양체정칙법은 현재까지 많이 사용되면서 서포터벡터기계 (support vector machine, Vapnik, 1998)와 커널방법 (Joachims, 1999) 등을 적용하였다. 대표적인 방법이 전환적 서포터벡터기계 (transductive SVM, Vapnik, 1998)인데 이 방법은 좋은 결과를 보이는 반면 계산이 어렵고 시간이 많이 걸리는 단점이 있을 뿐 아니라 전환적 (transductive, 현재의 자료에 대한 분류 값만 추정하기 때문에 새로운 자료를 분류하려면 알고리즘 전체를 다시 수행해야 한다)이다. 이 외의 여러 가지 방법들도 개발되어 이용되고 있지만 계산이 어려운 단점을 가지고 있을 뿐 아니라, 불균형 되는 결과 (라벨 없는 자료의 라벨 추정값이 하나의 라벨 값에 편중되는 현상)가 가끔씩 나오는 현상을 발견할 수 있다. 이는 추정에 필요한 모수가 잘못 선택되었을 수도 있지만 원천적으로 이러한 현상을 방지하지 못하였기 때문이다 (Zhou 등, 2006).

Suykens과 Vandewalle (1999)는 서포터벡터기계의 이차프로그래밍문제를 선형방정식문제로 해결하는 최소제곱 서포터벡터기계 (least squares support vector machine)를 제안하였다. 최소제곱 서포터벡터기계는선형방정식문제를 사용하므로 서포터벡터기계에 비해 계산시간을 줄일 수 있는 장점이 있다. 최소제곱 서포터벡터기계는 커널함수 (kernel function)를 사용하는 비선형 모형으로서 기존의 선형 모형을 비선형 모형으로 쉽게 확장할 수 있는 이론적 근거를 제공하고 있다. 최소제곱 서포터벡터기계 및 커널방법이 주목받는 이유는 명백한 이론적 근거에 기반을 두므로 결과 해석이 용이하고 실제 응용에 있어서 인공신경망 (neural network)보다 더 높은 성과를 내기 때문이다. 그리고 적은 학습 자료만으로 신속하게 분류 및 회귀학습을 수행할 수 있기 때문이다. 서포터벡터기계, 최소제곱 서포터벡터기계, 커널방법의 내용과 응용에 관한 것은 참고문헌 Schölkopf와 Smola (2002), Vapnik (1998), Suykens와 Vandewalle (1999), Suykens 등 (2002), Shim과 Lee (2009), Shim 등 (2009)을 참고하기 바란다. 그리고 다중분류 (multicategory classification)는 이진분류 (binary classification) 문제를 통해 해결 할 수 있기 때문에 본 연구에서는 이진분류 문제만 다루기로 한다 (Seok, 2008).

본 연구에서는 분류에서 좋은 성능을 보이는 최소제곱 서포터벡터기계를 준지도분류에 적용시키는 방법을 제안한다. 그리고 제안된 방법이 준지도분류에서 좋을 결과를 보인다는 것을 실험을 통해 증명하고자 한다. 2절에서는 준지도분류, 다양체정칙법 그리고 최소제곱 서포터벡터기계를 소개한다. 그리고 3절에서는 최소제곱 서포터벡터기계형태의 준지도학습을 소개하고 4절에서는 제안된 방법이 잘 적용됨

을 실험을 통해 보일 것이다. 그리고 5절에서는 결론 및 향후연구과제에 대해 언급한다.

2. 준지도분류와 최소제곱 서포터벡터기계

2.1. 그래프기반 준지도분류

자료가 $\{(\boldsymbol{x}_L, \boldsymbol{y}_L)\} = \{(x_1, y_1), ..., (x_{n_L}, y_{n_L})\}'$와 $\boldsymbol{x}_U = (x_{n_L+1}, ..., x_{n_L+n_U})'$, $\boldsymbol{x} = \{\boldsymbol{x_U} \cup \boldsymbol{x_L}\}$, $n = n_L + n_U$로 주어졌다. $\boldsymbol{x}_i \in R^d$는 상수 1을 포함하는 입력값이고 y_i는 $\boldsymbol{x}_i$의 라벨인데 -1 혹은 $+1$값을 가진다 ($i = 1, ..., n_L$). n_L과 n_U는 라벨 있는 자료와 라벨 없는 자료의 크기를 각각 나타낸다. 라벨 있는 자료로 라벨 없는 자료의 라벨을 추정하고자 할 때 분류처럼 라벨 있는 자료만 사용하는 것이 아니라 라벨 없는 자료도 사용하는 것이 준지도분류이다. 준지도분류에 대한 많은 연구가 진행되어 좋은 결과를 보이는 방법이 개발되 고 있다.

최근에 주목을 많이 받고 있는 방법은 그래프기반 준지도분류이다. 주어진 자료를 이용하여 정점 (vertex)과 가장자리 (edge)를 결정해 나가는 것이 그래프기반 방법의 주 된 내용이다. y_i값들은 정점에 배치되고, y_i와 y_j를 잇는 가장자리의 가중치 (weight) w_{ij}는 정점들의 유사성에 의해 그 값이 결정된다. 정점 y_i와 y_j이 유사하다면 w_{ij}는 큰 값을 가지고, 그렇지 않으면 작은 값을 가지도록 한다. 그래프에 따라 다양한 w_{ij}를 사용할 수 있는데, 완전연결그래프 (fully connected graph)는 $w_{ij} = \exp(-||\boldsymbol{x}_i - \boldsymbol{x}_j||^2/\sigma^2)$를 사용하고 kNN (k nearest neighborhood) 그래프에서는 $\boldsymbol{x}_j \in \{\boldsymbol{x}|kNN \text{ of } \boldsymbol{x}_i\}$이면 $w_{ij} = 1$을 그렇지 않으면 $w_{ij} = 0$을 사용한다.

그래프를 이용하여 라벨함수 f를 추정할 때 다음의 두 가지 기준을 만족하도록 해야 한다. 1) $f(\boldsymbol{x}_i)$는 y_i와 가까워야 한다. 2) 함수 f는 부드러워야 한다. 이것을 수식으로 표현하여 해결하는 방법 중 하나가 최소절단방법이다. 이 방법은 다음의 벌칙항 있는 손실함수를 최소화 하는 문제로 표현할 수 있다.

$$\min_{f\in\{-1,1\}} C\sum_{i=1}^{n_L}(y_i - f(\boldsymbol{x}_i))^2 + \sum_{i,j=1}^{n} w_{ij}(f(\boldsymbol{x}_i) - f(\boldsymbol{x}_j))^2. \tag{2.1}$$

여기에서 C는 아주 큰 값이다. 이 문제를 해결하는 효율적인 알고리즘이 존재하지만, 이 방법은 전환적이라는 단점을 가지고 있다. 이 방법을 좀 더 일반화시킨 방법이 조화함수를 사용하는 방법이다. 조화함수란 다음의 식과 같이 라벨 있는 자료에서는 추정값이 주어진 라벨과 같으면서 가중평균으로 계산되는 함수를 말한다.

$$f(\boldsymbol{x}_i) = y_i, i = 1, ..., n_L$$
$$f(\boldsymbol{x}_i) = \sum_{k=1}^{n} w_{ik} f(\boldsymbol{x}_k) / \sum_{k=1}^{n} w_{ik}$$

이 식은 (2.1)과 유사한 형태로 표현할 수 있다.

$$\min_{f\in R} \sum_{i,j=1}^{n} w_{ij}(f(\boldsymbol{x}_i) - f(\boldsymbol{x}_j))^2 \tag{2.2}$$
$$\text{s.t. } f(\boldsymbol{x}_i) = y_i,\ i = 1, ..., n_L.$$

조화함수를 이용하는 방법의 최대 장점은 f의 해를 실수 공간에서 찾도록 조건을 완화한 것인데 이렇게 함으로써 최적의 유일한 해를 수식을 통해 구할 수 있다. 그러나 최소절단방법과 마찬가지로 조화함수 방법도 전환적이라는 단점을 가지는데 이를 보완하기 위해 다양체정칙법이 개발되었다.

2.2. 다양체정칙법

먼저 (2.1)과 (2.2)에 나오는 $\sum_{i,j=1}^{n} w_{ij}(f(\boldsymbol{x}_i) - f(\boldsymbol{x}_j))^2$ 항을 그래프 라플라시안 (graph Laplacian)을 이용하여 표현하면 다음과 같다.

$$\frac{1}{2}\sum_{i,j=1}^{n} w_{ij}(f(\boldsymbol{x}_i) - f(\boldsymbol{x}_j))^2 = \boldsymbol{f}'L\boldsymbol{f}.$$

여기에서 $\boldsymbol{f} = (f(\boldsymbol{x}_1), ..., f(\boldsymbol{x}_n))'$이고 $L = D - W$을 그래프 라플라시안이라고 하는데, 여기에서 D는 (i,i)번째 원소가 $\sum_{j=1}^{n} w_{ij}$인 $n \times n$ 크기의 대각행렬이다. 이 식은 큰 가중치를 가지는 관측치 끼리는 유사한 추정치를 갖도록 유도하는 역할을 한다. 그러나 이 항만으로 일반화오류 (generalizarion error)를 줄일 수 있다는 보장을 할 수 없고, 추정된 함수가 충분히 부드럽게 될 수 없다. 이러한 점을 보완하기 위해 $||\boldsymbol{f}||$를 벌칙항에 추가하는 것이 다양체정칙법이다. Belkin 등 (2006)에 의해 개발된 다양체정칙법은 다음과 같이 일반적인 형태로 표현될 수 있다.

$$\boldsymbol{f}^{*} = argmin_{\boldsymbol{f}\in H_K}\sum_{i=1}^{n_L} V(\boldsymbol{x_i}, y_i, \boldsymbol{f}) + \mu_1||\boldsymbol{f}||_K^2 + \mu_2\boldsymbol{f}'L\boldsymbol{f}.$$

여기에서 V는 손실함수로써 잔차제곱 혹은 경첩손실 (hinge loss)등을 사용할 수 있다. 그리고 $||\boldsymbol{f}||_K^2$는 $\boldsymbol{f}$의 부드러운 정도를 조절하기 위한 항이고, μ_1과 μ_2는 각각 추정된 함수의 부드러운 정도를 조절하는 모수와 유사성의 강도를 조절하는 모수이다. 커널 K에 대응되는 RKHS (reproducing kernel Hilbert space)는 H_K로 나타내었다. 커널방법을 부분적으로 적용시킨 연구는 Seok (2007)에서 볼 수 있다.

2.3. 최소제곱 서포터벡터기계

분류를 위해 주어진 자료를 $\{\boldsymbol{x}_i, y_i\}_{i=1}^{n}$라고 표기하기로 한다. 여기서 $\boldsymbol{x}_i \in R^d$는 상수 1을 포함하는 입력변수벡터이고 y_i는 x_i의 라벨인데 -1 혹은 $+1$값을 갖는다. 여기서 $\phi : R^d \to R_f^d$는 입력공간을 고차원 특징공간으로 사상하는 특징함수라고 할 때, 다음과 같이, $\phi(\boldsymbol{x}_i)'\phi(\boldsymbol{x}_j) = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$, 특징공간에서의 내적은 입력공간에서 동등한 커널을 가진다 (Mercer, 1909). 여러 형태의 커널 $k(\cdot,\cdot)$이 자료에 따라 사용될 수 있다. 분류기 $f(\boldsymbol{x}) = \boldsymbol{w}'\phi(\boldsymbol{x})$는 primal 공간에서 다음의 최적화 문제의 해로 정의될 수 있다.

$$\min_{\boldsymbol{w},e} J_P(\boldsymbol{w}, e) = \frac{1}{2}\boldsymbol{w}'\boldsymbol{w} + \mu\sum_{i=1}^{n} e_i^2 \tag{2.3}$$
$$s.t y_i[\boldsymbol{w}\phi(\boldsymbol{x}_i)] = 1 - e_i, i = 1, ..., n.$$

(2.3)을 라그랑제 배수 (α_i)를 사용한 라그랑제 함수로 표현하면 다음과 같다.

$$L_D(\boldsymbol{\alpha}) = J_P - \sum_{i=1}^{n}\alpha_i[y_i\boldsymbol{w}'\phi(\boldsymbol{x_i}) - 1 + e_i],$$

최적 라그랑제 배수는 다음의 조건으로부터 구해진다.

$$\begin{aligned}
&\frac{\partial L_D}{\partial \boldsymbol{w}} = 0 \Rightarrow \boldsymbol{w} = \sum_{i=1}^{n} \alpha_i y_i \phi(\boldsymbol{x_i}), \\
&\frac{\partial L_D}{\partial e_i} = 0 \Rightarrow \alpha_i = \mu e_i, i = 1, ..., n, \\
&\frac{\partial L_D}{\partial \alpha_i} = 0 \Rightarrow y_i[\boldsymbol{w}'\phi(\boldsymbol{x}_i)] = 1 - e_i, i = 1, ..., n.
\end{aligned} \tag{2.4}$$

따라서 (2.4)와 Mercer 정리 (1909)를 이용하면 입력값이 $\boldsymbol{x}$인 자료의 분류기는 다음과 같이 구해지며

$$f(\boldsymbol{x}) = \sum_{i=1}^{n} \alpha_i y_i K(\boldsymbol{x_i}, \boldsymbol{x})$$

분류기의 부호에 따라 주어진 자료가 분류된다. 여기에서 K는 커널함수로서 $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \phi(\boldsymbol{x}_i)'\phi(\boldsymbol{x}_j)$로 정의된다. 많이 사용되는 커널함수는 방사형 기저 함수 (radial basis function)인데 이는 $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left\{-||\boldsymbol{x}_i - \boldsymbol{x}_j||^2/\sigma^2\right\}$이다.

3. 최소제곱 서포터벡터기계 형태의 준지도분류

본 연구에서는 다양한 분야에서 좋은 성과를 보이고 있는 최소제곱 서포터벡터기계를 준지도분류에 적용시키는 새로운 방법을 제안하고자 한다. 먼저 최소제곱 서포터벡터기계 방법을 다양체정칙법에 적용한 다음의 최적화문제를 생각할 수 있다.

$$\begin{aligned}
\min J = \frac{1}{2}\boldsymbol{w}'\boldsymbol{w} + \frac{\mu_1}{2}\sum_{i=1}^{n_L} e_i^2 + \frac{\mu_2}{2}\sum_{i,j=1}^{n} \epsilon_{ij}^{2,} \\
\text{s.t } y_i\left\{\boldsymbol{w}'\phi(\boldsymbol{x}_i) + b\right\} = 1 - e_i, i = 1, ..., n_L, \\
\sqrt{w_{ij}}\left\{(\boldsymbol{w}'\phi(\boldsymbol{x}_i) - \boldsymbol{w}'\phi(\boldsymbol{x}_j)\right\} = \epsilon_{ij}, i, j = 1, ..., n.
\end{aligned} \tag{3.1}$$

여기에서 μ_1은 라벨 있는 자료의 적합도를 조절하는 모수이고, μ_2는 유사한 입력값이 유사한 라벨을 가지도록 조절하는 모수이다. (3.1)를 라그랑제 배수 (α_i, β_{ij})를 사용한 라그랑제 함수로 표현하면 다음과 같다.

$$\min L = J - \sum_{k=1}^{n} \alpha_k[y_k(\boldsymbol{w}'\phi(\boldsymbol{x_k}) + b) - 1 + e_k\ ] - \sum_{k,l=1}^{n} \beta_{kl}\left\{\epsilon_{kl} - \sqrt{w_{kl}}[\boldsymbol{w}'\phi(\boldsymbol{x}_k) - \boldsymbol{w}'\phi(\boldsymbol{x}_l)]\right\} \tag{3.2}$$

(3.2)로부터 다음의 조건식을 얻을 수 있다.

$$\begin{aligned}
&\frac{\partial L}{\partial \boldsymbol{w}} = 0 \Rightarrow \boldsymbol{w} = \sum_{k=1}^{n} \alpha_k y_k \phi(\boldsymbol{x}_k) - \sum_{k,l=1}^{n} \beta_{kl}\sqrt{w_{kl}}[\phi(\boldsymbol{x}_k) - \phi(\boldsymbol{x}_l)], \\
&\frac{\partial L}{\partial b} = 0 \Rightarrow \sum_{k=1}^{n} \alpha_k y_k = 0, \\
&\frac{\partial L}{\partial e_k} = 0 \Rightarrow \mu_1 e_k = \alpha_k, k = 1, ..., n_L \\
&\frac{\partial L}{\partial \epsilon_{kl}} = 0 \Rightarrow \mu_2 \epsilon_{kl} = \beta_{kl}, k, l = 1, ..., n, \\
&\frac{\partial L_D}{\partial \alpha_k} = 0 \Rightarrow y_k[\boldsymbol{w}'\phi(\boldsymbol{x}_k) + b] = 1 - e_k, k = 1, ..., n_L, \\
&\frac{\partial L}{\partial \beta_{kl}} = 0 \Rightarrow \sqrt{w_{kl}}\left\{\boldsymbol{w}'\phi(\boldsymbol{x}_k) - \boldsymbol{w}'\phi(\boldsymbol{x}_l)\right\} = \epsilon_{kl}, k, l = 1, ..., n.
\end{aligned} \tag{3.3}$$

(3.3)로부터 다음의 선형관계식을 얻을 수 있다.

$$\begin{aligned}
&\sum_{k=1}^{n_L} \alpha_k y_k = 0, \\
&y_i \left\{ [\sum_{k=1}^{n} \alpha_k y_k K_{ki} - \sum_{k,l=1}^{n} \beta_{kl}\sqrt{w_{kl}}(K_{ki} - K_{li})] + b \right\} + \frac{\alpha_i}{\mu_1} = 1, i = 1, ..., n_L, \\
&\sqrt{w_{ij}} \left\{ \sum_{k=1}^{n} \alpha_k y_k (K_{ki} - K_{kj}) - \sum_{k,l=1}^{n} \beta_{kl}(K_{ki} - K_{kj} - K_{li} + K_{lj}) \right\} - \frac{\beta_{ij}}{\mu_2} = 0, i, j = 1, ..., n.
\end{aligned} \tag{3.4}$$

여기에서 $K_{ij} = K(\boldsymbol{x}_i, \boldsymbol{x}_j)$, $i, j = 1, ..., n$는 K의 (i, j)번째 원소이다. (3.4)식으로부터 주어진 $\boldsymbol{x}$의 예측값 $\hat{y}(\boldsymbol{x})$는 다음과 같이 얻을 수 있다.

$$\widehat{y}(\boldsymbol{x}) = \ sign \left\{ \sum_{k=1}^{n_L} \alpha_k y_k K(\boldsymbol{x}_k, \boldsymbol{x}) - \sum_{k,l=1}^{n} \sqrt{w_{kl}}\beta_{kl}[K(\boldsymbol{x}_k, \boldsymbol{x}) - K(\boldsymbol{x}_l, \boldsymbol{x})] \right\}.$$

여기에서 $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_{n_L}, \alpha_{n_L+1}, ..., \alpha_n)' = (\boldsymbol{\alpha}_L, \boldsymbol{\alpha}_U)$와 $\boldsymbol{\beta} = (\beta_{11}, \beta_{21}, ..., \beta_{12}, \beta_{22}, ..., \beta_{nn})'$은 다음의 선형시스템으로 구할 수 있다.

$$\begin{bmatrix} 0 & \mathbf{1}_{1\times n_L} & \mathbf{0}_{1\times n^2} \\ \mathbf{1}_{n_L\times 1} & \Omega + I_{n_L} & A_1 \\ \mathbf{0}_{n^2\times 1} & A_2 & A_3 \end{bmatrix} \begin{bmatrix} b \\ \boldsymbol{\alpha}_L \\ \boldsymbol{\beta} \end{bmatrix} = \begin{bmatrix} 0 \\ \mathbf{1}_{n_L\times 1} \\ \mathbf{0}_{n^2\times 1} \end{bmatrix},$$

여기에서 $\mathbf{0}_{m\times n}$ $(\mathbf{1}_{m\times n})$은 모든 원소가 0 (1)인 크기가 $m \times n$인 행렬을 나타낸다. 그리고 $\Omega = (\Omega_{kl})_{n_L\times n_L} = \{ y_k y_l K(\boldsymbol{x}_k, \boldsymbol{x}_l) \}_{n_L\times n_L}$, $k, l = 1, ..., n_L$이고 I_a는 크기가 $a \times a$인 단위행렬이다. $A_1 = (A_{11}, ..., A_{1n})$인데 $A_{1l} = \left\{-(K_{\cdot,(1:n_L)})' + (K_{l,(1:n_L)})'\mathbf{1}_{n_L\times 1}\right\} * (\mathbf{1}_{n_L\times 1}(W_{\cdot,l})')^1/2, l = 1, ..., n$이다. 여기에서 $K_{\cdot,l}$은 행렬 K의 l번째 열을 나타내고, $K_{l,\cdot}$은 K의 l번째 행을 그리고 $K_{(l:m),\cdot}$은 l부터 m번째까지의 행을 나타낸다. 그리고 $*$는 두 행렬의 대응되는 원소끼리 곱하는 연산자이다. $A_2' = (A_{21}', ..., A_{2n}')$인데 $A_{2l} = (\mathbf{1}_{n_L\times 1}\boldsymbol{y}_L^{'} * (W_{l,\cdot}^1/2)^{'}\mathbf{1}_{1\times n_L}) * (\mathbf{1}_{n_L\times 1}(K_{(1:n_L),l})' - K_{(1:n_L),\cdot})$,

$l = 1, ..., n$이다. A_3는 $n \times n$크기의 행렬 $A_3(k,l)$, $k, l = 1, ..., n$가 $n \times n$배열을 이루는데, (k,l)번째 행렬 $A_3(k,l) = (K + K_{k,l}\mathbf{1}_{n\times n} - \mathbf{1}_{n\times 1}(K_{.,k})' - (K_{l,.})'\mathbf{1}_{1\times n}) * (\mathbf{1}_{n\times 1}W_{.,l})^1/2 + B(k,l)$이다. 여기에서 $B(k,l)$은 $n \times n$행렬로서 $(i,j) = (l,k)$이면 $B(k,l)_{ij} = -1/\mu_2$이고 그 외에는 $B(k,l)_{ij}$ =0이다.

다음 절에서는 우리가 제안한 방법이 타당함을 모의실험을 통하여 살펴보기로 한다.

4. 모의실험

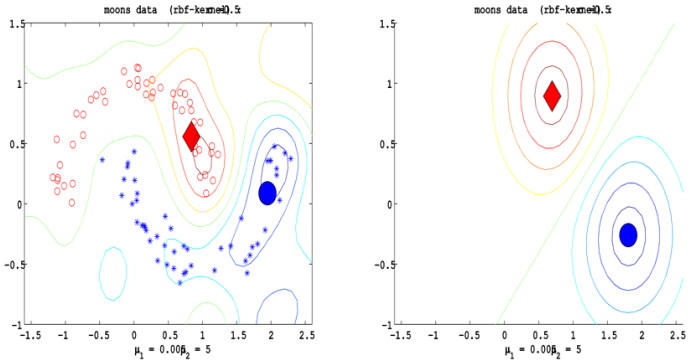

그림 4.1 Moon 자료와 준지도분류의 결과 (왼쪽)와 서포터벡터기계 결과 (오른쪽)

모의실험에 사용된 자료는 준지도학습에서 많이 사용되는 전형적인 자료인 moon 자료와 circles 자료이다. 먼저 moon 자료는 그림 4.1의 왼쪽에 나타나 있는데, 라벨 있는 자료는 +1과 -1이 각각 1개씩 ($n_L = 2$)으로 마름모와 큰 원으로 표시되어 있고 나머지 자료는 라벨 없는 자료이다 ($n_u = 100$, o와 *로 표시됨). 라벨 없는 자료를 제외하고 라벨 있는 자료만을 이용하여 서포터벡터기계로 분류작업을 한 결과 오른쪽 그림과 같은 선형분류선을 얻었다. 이 분류선을 이용하여 라벨 없는 자료의 라벨을 추정하면 그 추정결과가 좋지 않다는 것을 알 수 있을 것이다. 왼쪽 그림은 본 연구에서 제안한 방법으로 라벨 없는 자료도 이용하여 분류작업을 한 결과이다. 모든 자료가 원하는 방향으로 분류된 것을 알 수 있다. 본 연구에서 사용된 가중치 행렬 W는 유클리드 거리, $k = 6$ (라벨 있는 자료만 이용할 때는 1)인 kNN과 이진함수 (binary weight type)를 사용하였다. 그리고 $(\mu_1, \mu_2, \sigma) = (0.005, 5, 0.5)$를 사용하였다.

다음으로 circles 자료는 그림 4.2의 왼쪽에 나타나 있는데, 라벨 있는 자료는 +1과 -1이 각각 1개씩으로 마름모와 큰 원으로 표시되어 있고 나머지 자료는 라벨 없는 자료이다 (o 와 *로 표시됨). 라벨 없는 자료를 제외하고 라벨 있는 자료만을 이용하여 서포터벡터기계로 분류작업을 한 결과 오른쪽 그림과 같은 분류선을 얻었다. 그림 4.1에서와 마찬가지로 이 분류선을 이용하여 라벨 없는 자료의 라벨을 추정하면 그 추정결과가 좋지 않다는 것을 알 수 있을 것이다. 왼쪽 그림은 본 연구에서 제안한 방법으로 라벨 없는 자료도 이용하

여 분류작업을 한 결과이다. 모든 자료가 원하는 방향으로 분류된 것을 알 수 있다. 본 연구에서 사용된 가중치 행렬 W는 유클리드 거리, $k = 6$ (라벨 있는 자료만 이용할 때는 1)인 kNN과 이진함수를 사용하였다. 그리고 $(\mu_1, \mu_2, \sigma) = (0.05, 50, 0.8)$를 사용하였다.

이상의 두 가지 모의실험을 통해 제안된 방법이 준지도분류에 잘 작용하고 있음을 알 수 있다. 즉 라벨을 추정할 때 라벨 없는 자료를 적절하게 이용할 수 있는 하나의 방법을 제시하였다.

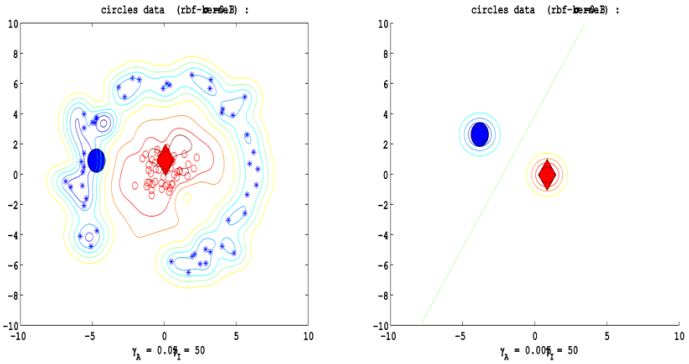

그림 **4.2** Circles 자료와 준지도분류의 결과 (왼쪽)와 서포터벡터기계 결과 (오른쪽)

5. 결론

분류에서 좋은 성능을 보이는 최소제곱 서포터벡터기계를 준지도분류에 적용시키는 방법을 제안한다. 그리고 제안된 방법이 준지도분류에서 좋을 결과를 보인다는 것을 실험을 통해 증명하였다. 그러나 제안된 방법의 선형시스템을 해결하기 위해서 $O((n^2 + n_L + 1)^2)$ 크기의 계산공간이 필요하다는 것이 하나의 단점으로 지적될 수 있으나 컴퓨터 성능향상이 그것을 보완해 줄 수 있다고 본다. 모수 $(k, \mu_1, \mu_2, \sigma)$결정방법과 다른 준지도분류 방법과의 비교는 해결해야 할 차후 연구과제로 남긴다.

참고문헌

Belkin, M., Niyogi, P. and Sindhwani, V. (2006). On manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 1-48.

Blum, A. and Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. *Proceedings of the 11th Annual Conference on Computational Learning Theory*, Madison, 92-100.

Chapelle, O., Schölkopf , B. and Zien, A. (2006). *Semi-supervised learning*, The MIT Press, Cambridge, Massachusetts.

Holub, A., Welling, M. and Perona, P. (2005). Exploiting unlabelled data for hybrid object classification. *NIPS 2005 Workshop in Inter-Class Transfer*.

Huang, T. M. and Kecman, V. (2004). Semi-supervised learning from unbalanced labeled data - An improvement. *Knowledge Based and Emergent Technologies Relied Intelligent Information and Engineering Systems*.

Joachims, T. (1999). Transductive inference for text classification using support vector machines. *Proceedings of the 16th International Conference on Machine Learning*, 200-209.

Mercer, J. (1909). Functions of positive and negative type and their connection with theory of integral equations. *Philosophical Transactions of Royal Society*, A, 415-446.

Nigam, K. and Ghani, R. (2000). Analyzing the effectiveness and applicability of co-training. *Ninth International Conference on Information and Knowledge Management*, 86-93.

Nigam, K., McCallum, A. K., Thrun, S. and Mitchell, T. (2000). Text classification from labeled and unlabeled documents using EM. *Machine Learning*, **39**, 103-134.

Schölkopf , B. and Smola, A. (2002). *Learning with kernels- Support vector machines, regularization, optimizations and beyond*, MIT Press.

Seeger, M. (2001). *Learning with labeled and unlabeled data*, Technical report, University of Edinburgh.

Seok, K. H. (2007). Data-adaptive ECOC for multicategory classification. *Journal of Korean Data & Information Science Society*, **19**, 25 -36.

Seok, K. H. (2007). Semi-supervised learning using kernel estimation. *Journal of Korean Data & Information Science Society*, **18**, 629-636.

Shim , J. and Lee, J. T. (2009). Kernel method for autoregressive data. *Journal of Korean Data & Information Science Society*, **20**, 949-964 .

Shim, J., Park, H. J. and Seok, K. H. (2009). Variance function estimation with LS-SVM for replicated data. *Journal of Korean Data & Information Science Society*, **20**, 925 -931.

Suykens, J. A. K., Gastel, T. V., Bravanter, J. D., Moore, B. D. and Vandewalle, J. (2002). *Least squares support vector machines*, World Scientific.

Suykens, J. A. K. and Vandewalle, J. (1999). Least squares support vector machine classifiers. *Neural Processing Letters*, **9**, 293-300.

Vapnik, V. (1998). *Statistical learning theory*, Wiley, New York.

Zhou, D., Bousquet, T. N., Lal, J . and Schölkopf, B . (2004). Learning with local and global consistency. *Advances in Neural Information Processing Systems*, **16**, 321-328.

Zhou, Y. and Goldman, S. (2004). Democratic co-learning. *Proceedings of the 16th IEEE International Conference on Tools with Artificial Intelligence* (ICTAI2004).

Zhou, X., Ghahramani, Z. and Lafferty, J. (2003). Semi-supervised learning using Gaussian fields and harmonic functions. *Proc. of the 20th International Conference on Machine Learning*, Washington DC.

Zhou, Z., Chen. K., and Dai, H. (2006). Enhancing relevance in image retrieval using unbalance data. *ACM Transactions on Information Systems*, **24**, 219-244.

Zhou, Z. and Li, M. (2007). Semi-supervised regression with co-training style algorithm. *IEEE Transactions on Knowledge and Data Engineering*.

Zhu, D. (2005). *Semi-supervised learning literature survey*, Technical Report Computer Sciences 1530, University of Wisconsin - Madison.

Journal of the Korean Data & Information Science Society
2010, **21**(3), 461–470

한국데이터정보과학회지

Semi-supervised classification with LS-SVM formulation[†]

Kyung Ha Seok[1]

[1] Inje University Department of Data Science

Received 8 April 2010, revised 12 May 2010, accepted 19 May 2010

Abstract

Semi supervised classification which is a method using labeled and unlabeled data has considerable attention in recent years. Among various methods the graph based manifold regularization is proved to be an attractive method. Least squares support vector machine is gaining a lot of popularities in analyzing nonlinear data. We propose a semi supervised classification algorithm using the least squares support vector machines. The proposed algorithm is based on the manifold regularization. In this paper we show that the proposed method can use unlabeled data efficiently.

Keywords: Graph based semi supervised classification, Graph Laplacian, least squares support vector machine, manifold regularization, semi supervised classification.

[†] This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2009-0072369).
[1] Professor, Department of Data Science, Institute of Statistical Information, Inje University, Kimhae 621-749, Korea. E-mail: statskh@inje.ac.kr