Journal of Korean
Data & Information Science Society
2007, Vol. 18, No. 3, pp. 629 ~ 636

Semi-Supervised Learning Using Kernel Estimation

Kyung Ha Seok[1)]

Abstract

A kernel type semi-supervised estimate is proposed. The proposed estimate is based on the penalized least squares loss and the principle of Gaussian Random Fields Model. As a result, we can estimate the label of new unlabeled data without re-computation of the algorithm that is different from the existing transductive semi-supervised learning. Also our estimate is viewed as a general form of Gaussian Random Fields Model. We give experimental evidence suggesting that our estimate is able to use unlabeled data effectively and yields good classification.

Keywords : Gaussian Random Fields Model, Kernel Estimate, Semi-Supervised Learning, Transductive Learning

1. Introduction

지도학습(supervised learning)과 자율학습(unsupervised learning)에 관한 많은 연구가 이루어지고 있다. 회귀분석(regression)과 분류(classification)는 대표적인 지도학습이고 군집분석(clustering)은 자율학습의 대표적인 분석방법이다. 이렇게 분석방법을 나누는 기준은 목표값 유무에 따른 것이다. 특히 분류에서의 목표값을 분류값(label)이라 부른다.

최근에 만들어지는 자료는 분류값을 가지는 비율이 아주 작은 경우가 많이 있다. 그 이유는 분류값을 만드는 것이 어렵거나 혹은 많은 시간과 경비가 소요되거나 혹은 자료가 너무 커서 분류값을 만드는 것이 힘들기 때문이다. 예를 들어 폭발적으로 만들어지는 모든 스팸메일의 분류값을 만드는 것은 불가능한 일이라고 생각된다. 또 다른 예로는 음성인식(speech processing)(시간과 경비문제), 문자분류(text categorization), 웹분류(web categorization)(시간과 자료의 크기), 그리고 생명정보(bioinformatics)(비용, 시간, 자료의 크기)등이 있다.

이러한 자료를 분석하기 위해 준지도학습(Semi-Supervised Learning)이라는 새로운 영역이 개발되어 많은 연구가 이루어지고 있다(Zhu 2005, Chapelle et al. 2006).

1) Professor, Department of Data Science, Inje University, Kyungnam 621-749, Korea, E-mail: statskh@paran.com

준지도학습은 분류값이 있는 자료와 분류값이 없는 자료를 모두 사용하여 분류값이 없는 자료의 분류값을 추정하는 방법이다. 물론 분류값 있는 자료를 이용하여 분류규칙을 만든 후 분류값 없는 자료의 분류값을 추정하는 지도학습도 있으나, 분류값 있는 자료가 지도학습에 사용될 만큼 충분하지 않거나 분류값 없는 자료가 분류규칙을 만드는데 도움을 줄 수 있는 경우에는 준지도학습 방법을 사용한다.

최근에 준지도학습에 대해 많이 연구되고 있는데 가장 많이 사용되고 있는 알고리즘은 자기훈련(self-training)이다. 이는 혼합모형(mixture model) 과 EM알고리즘의 한 부류로 해석이 된다. 이를 이용한 예로는 Rosenberg(2005)를 들 수 있다. 자기훈련의 최대 단점인 '한번 잘못 된 규칙은 계속 영향을 미칠 수 있다'는 것을 줄이기 위해 개발된 것이 공동훈련(co-training)이다. 이 방법은 변수들이 두 개의 집합으로 나누어 질 수 있고, 그리고 각 집합에서 분류기(classifier)가 충분히 훈련될 수 있다는 가정을 하고 있다. 여기에 관한 것은 Nigam and Ghani(2000)과 Zhou and Goldman(2004)에서 볼 수 있다.

최근에 많이 주목 받고 있는 방법으로는 Vapnik(1998)의 TSVM(Transductive Support Vector Machine)을 들 수 있다. 이 방법은 SVM의 경계최대화(margin maximization)개념을 이용한 것으로 밀도가 희박한 지역(sparce region)으로는 분류선이 지나도록 설계하였다(low density separation). 이 원리를 이용한 방법으로는 Gaussian process (Lawrence and Jordan, 2005), Information Regularization(Szummer and Jaakkola , 2002) 등이 있다.

그러나 위에서 소개한 여러 가지 방법들은 해가 전역최적(global optimum)이 아닌 국소최적(local optimum)에 수렴할 수 있다는 결정적인 단점을 가지고 있다. 그 이유는 목적함수가 볼록함수(convex)가 아니기 때문이다. 또 다른 단점은 목적함수가 볼록함수이더라도 해를 찾는데 TSVM 처럼 아주 많은 시간을 요구하는 경우도 있다. 이러한 단점을 극복하기 위해 그래프기반 준지도학습이 개발되었다.

그래프기반 준지도학습에는 여러 가지가 개발되어 있는데 그 중에서 가장 이해가 쉽고 기본이 되는 두 가지 방법은 Zhu et al. (2003)의 GRFM(Gaussian Random Fields Model)과 Zhu et al. (2004) 의 CM(Consistency Model)이다. 이 두 가지 방법은 그래프이론을 기반으로 자료의 다양체구조(manifold structure)를 탐색하기 때문에 전역최적에 수렴하는 좋은 성질을 가지고 있다(Belkin et al. 2006). 이 연구에서 언급하고 있는 방법은 지도학습처럼 모든 가능한 입력자료에 대한 분류규칙을 만드는 것이 아니라 현재 가지고 있는 자료의 분류값만 추정하는 것이다. 그러므로 이런 방법은 새로운 자료를 분류하려면 알고리즘 전체를 다시 수행해야 하는 큰 단점을 지니고 있다.

그래서 본 연구에서는 지도학습처럼 일반화 된 분류규칙을 만드는 방법를 제안하고자 한다. 제안된 방법은 커널추정량과 같은 형태인데 벌칙항 있는 최소제곱 손실함수(penalized least squares loss)와 준지도학습의 원리("같은 군집 안에 있는 개체는 같은 분류값 을 가진다")를 이용한다. 이렇게 구한 분류함수를 전형적인 예제 자료에 적용을 시켜보니 좋은 결과를 보였다.

본 논문은 다음과 같이 구성된다. 2절에는 준지도학습및 GRFM에 대한 간략한 소개를 하고, 3절에서는 새로는 추정법을 제안한다. 그리고 4절에서는 제안한 방법의 정당성을 모의실험을 통해 보여주고, 5절은 결론으로 본 논문의 내용을 정리하고 향후 연구과제를 제안한다.

2. Gaussian Random Fields Model

분류값 있는 자료 $\{x_L, y_L\} = \{(x_1, y_1), ..., (x_{n_L}, y_{n_L})\}$ 와 분류값 없는 자료 $x_U = \{x_{n_L+1}, ..., x_{n_L+n_U}\}$, $x_i \in R^n$, 가 주어졌다. 그리고 $x = \{x_U \cup x_L\}$ 라 한다. 여기에서 y_i, $i = 1, ..., n_L$는 분류값 인데 -1혹은 $+1$값을 갖는다. n_L 과 n_U는 각각 분류값 있는 자료와 분류값 없는 자료의 크기를 나타낸다. 주어진 분류값 있는 자료로 분류값 없는 자료의 분류값 을 추정하고자 할 때 지도학습처럼 분류값 있는 자료만 사용하는 것이 아니라 분류값 없는 자료도 사용하는 것이 준지도학습이다. 최근에 많은 관심을 받고 있으며 활발히 연구되고 있는 방법이 GRFM이다.

GRFM은 정칙이론(regularization theory)을 이용하는 것으로 이해 할 수 있는데 성능이 아주 좋은 것으로 알려져 있다. 이 방법은 x에서 정의되는 그래프 $G = (V, E)$를 이용하여 문제를 해결하는 방법인데, 여기에서 V는 자료점(data points) 에 해당하는 노드집합이고, E는 노드를 연결하는 에지(edge)집합인데 각 에지의 연결강도는 가중행렬(weight matrix) $W = (w_{ij})_{n \times n}$에 의해 정해진다. GRFM의 기본 아이디어는 $f : V \rightarrow R$ on G 인 $f = (f_1, ..., f_n)^T$를 찾는 것이다. 이것을 구할 때 가까이 있는 분류값 없는 자료끼리의 f_i 값의 차이를 적게하고, 분류값 있는 자료의 f_j는 원래의 분류값 y_j 와 같도록 하는 것이다. 이러한 조건의 손실함수(loss function)는 아래와 같다.

$$L_1 = \sum_{i,j=1}^{n} w_{ij}(f_i - f_j)^2 + \lambda \sum_{i=1}^{n}(y_i - f_i)^2$$

$$= f' L f + \lambda (f - y)'(f - y). \quad (1)$$

여기에서 $L = (D - W)$는 라플라시안(Laplacian) 행렬이고, $D = diag(d_i)_{n \times n}$, $d_i = \sum_{j=1}^{n} w_{ij}$, $n = n_L + n_U$ 이다. 그러나 (1) 식에는 f의 복잡도(complexity)에 대한 제약조건이 없기 때문에 이로부터 구해진 f는 과적합(overfitting)될 수도 있다. 그 뿐만 아니라 새로운 자료에 대한 분류를 원할 때에는 모든 자료를 이용하여 알고리즘 전체를 다시 수행해야하는 단점이 있다. 이러한 단점을 보완하기 위하여 통계학과 학습이론에서 많이 이용되는 벌칙항이 있는 최소제곱 손실함수(penalized least squares loss)와 준지도학습의 원리(같은 군집 안에 있는 개체는 같은 분류값 을 가진다)를 이용한 커널추정법을 제안하고자 한다.

3. 커널추정법을 이용한 준지도학습

커널추정법은 Xu et al.(2001)이 제안한 것으로 MSE(Mean Squared Error)를 확장한 개념으로 생각할 수 있다. 이 추정법은 비선형추정법으로써 커널함수를 사용하는데 간단하게 소개하면 다음과 같다. 먼저 특징공간(feature space)에서 가중치 w와 bias b를 사용하여 다음의 선형추정량을 생각할 수 있다.

$$f(x)=\boldsymbol{w}'\psi(x)+b, \tag{2}$$

여기에서 $\psi: R^n \rightarrow R^h$ 는 특징공간으로의 함수이다. (2)식에서 $b=0$으로 하여도 추정량에는 큰 영향을 끼치지 않으므로 본 연구에서는 $b=0$으로 한다. 커널재생(reproducing kernel)이론에 의해 $\boldsymbol{w}$는 다음과 같이 표현이 가능하다.

$$\boldsymbol{w}=\sum_{k=1}^{n}\alpha_k\psi(x_k). \tag{3}$$

(3) 식을 (2)식에 대입하면 다음의 커널추정량

$$\boldsymbol{f}=K\boldsymbol{\alpha}$$

을 얻을 수 있다. 여기에서 K 는 커널행렬인데 (i,j)번째 원소는

$$k_{ij}=\psi(x_i)'\psi(x_j)$$

이다.

커널추정량 $\boldsymbol{f}=K\boldsymbol{\alpha}$은 다음과 같은 벌칙항있는 최소제곱 손실함수를 이용하여 구할 수 있다.

$$L_2=\sum_{i=1}^{n}(y_i-f_i)^2+\gamma_1\,\boldsymbol{\alpha}'\boldsymbol{\alpha}$$

여기에서 $f_i=K(i,:)\boldsymbol{\alpha}$, $K(i,:)$ 는 커널행렬 K 의 i번째 행이다. 커널행렬 K의 (i,j)번째 원소는 RBF(radial basis function)을 사용하여 $k_{ij}=\exp(-\|x_i-x_j\|^2/\sigma^2)$로 쓰기로 한다. γ_1은 함수의 복잡도와 자료의 적합도를 조절하는 모수다. 준지도학습에서 기본적으로 사용하고 있는 가정-가까이 있는 분류값 없는 자료끼리의 f_i 값은 차이를 적게 한다 - 을 위의 식과 결합하여 다음과 같은 손실함수를 제안한다.

$$L_3=\sum_{i=1}^{n}(y_i-f(x_i))^2+\gamma_1\,\boldsymbol{\alpha}'\boldsymbol{\alpha}_1\ \ +\gamma_2\sum_{i,j=1}^{n}w_{ij}(f_i-f_j)^2 \tag{4}$$

여기에서 γ_2는 가까이 있는 분류값 없는 자료의 f_i 값에 대한 차이를 조절하는 모수이고, W의 (i,j)번째 원소는

$$w_{ij}=\begin{cases} e^{-\|x_i-x_j\|^2/\sigma_w^2}, & i\neq j \\ 0 \qquad\qquad\quad , & i=j \end{cases}$$

로 쓰기로 한다. (4)식은 볼록함수이기 때문에 국소최적이 아닌 유일한 전역최적인 해를 찾을 수 있는데, $\boldsymbol{\alpha}$는 간단한 계산으로 다음과 같이 구해진다.

$$\alpha = (I/\gamma_1 + KK + \gamma_2 KLK/\gamma_1)^{-1} Ky. \quad (5)$$

이를 이용하여 분류값 없는 자료의 라벨 $l_i = sign(f_i) = sign(K(i,:)\alpha)$를 계산 할 수 있다. 여기에서 $f_i = K(i,:)\alpha$는 일반화되어 있기 때문에 새로운 자료 x_0에 대한 분류값을 $l_0 = sign(f_0) = sign(\boldsymbol{k}_0\alpha)$, $\boldsymbol{k}_0 = (e^{-\|x_1 - x_0\|^2/\sigma^2}, \ldots, e^{-\|x_n - x_0\|^2/\sigma^2})$ 로 간단하게 계산할 수 있다.

GRFM를 비롯한 다른 준지도학습들을 이용하여 x_0의 분류값을 추정하려면 x_0를 포함하는 분류값 없는 자료와 분류값 있는 자료를 이용하여 전체 알고리즘을 다시 수행해야 하는 것에 비하면 계산시간을 많이 절약할 수 있는 장점이 있다. 그 뿐만 아니라 (4)식의 두 번째 항에 있는 모수 γ_1을 적절히 이용함으로써 추정이 과적합되는 것을 피할 수 있다. 만약 $\gamma_1 = 0$ 일 때는 GRFM 방법과 같은 추정이 되기 때문에 본 연구에서 제안한 방법이 GRFM을 일반화시킨 것이라 할 수 있다.

우리가 제안한 방법을 사용하기 위해서는 (5)식에 포함되어 있는 모수, γ_1, γ_2, σ 그리고 σ_w를 선택해야 한다. 모수 선택을 위해서 여러 가지 방법을 고려할 수 있겠지만 그 중에서 이해가 빠르고 사용하기 쉬운 GCV(generalized cross validation)를 사용할 수 있는데 다음과 같은 식으로 나타난다.

$$GCV = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - f_i}{1 - trace(A)/n}\right)^2$$

여기에서 $A = K(I/\gamma_1 + KK + \gamma_2 KLK/\gamma 1)^{-1} K$ 이다.

다음 절에서는 우리가 제안한 방법이 타당함을 모의실험을 통하여 살펴보기로 한다.

4. 모의실험

모의실험을 위해서 준지도학습에서 많이 사용되는 전형적인 자료인 moon data를 먼저 사용하였다. 이 자료는 그림 1에 나타나 있는데, 분류값 있는 자료는 +1과 -1에 각각 1개씩으로 ○ 으로 표시되어 있다. 나머지 자료는 분류값 없는 자료이다(o 와 + 로 표시됨). 만약 분류값 없는 자료를 제외하고 분류값 있는 자료만을 이용하면 선형 SVM으로 그림위의 직선과 같은 분류선을 얻는다. 그리고 이 분류선을 이용하여 분류값 없는 자료의 분류값을 구하면 원하지 않는 결과가 나온다는 것을 그림에서 알 수 있다. 본 실험에서는 사용된 모수는 3절에서 소개된 GCV를 통하여 구하였다. γ_1, γ_2는 (0.01, 100) 구간에서 σ, σ_w 는 (0.01,3)의 구간에서 최적의 값을 구하였다.

본 연구에서 제안한 방법으로 분류값 없는 자료의 분류값을 구한 결과를 그림에서 볼 수 있다. 분류값이 +1 을 '+'로 -1 을 'o'로 나타내었는데 분류값 있는 자료와 같은 군집에 속하는 자료는 같은 분류값을 가지도록 분류가 잘 되었음을 알 수 다. 이결과는 분류값 없는 자료가 분류하는데 도움을 준 것으로 볼 수 있다. 이 실험에서 사용된 모수, γ_1, γ_2, σ 그리고 σ_w값은 0.03, 1, 0.4, 0.3 이었다.

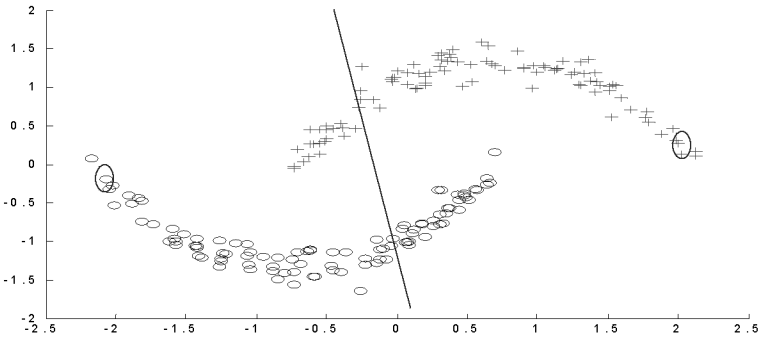

그림 1 Moon data와 분류값 없는 자료의 분류값 추정결과

다음 모의실험을 위해서 사용된 자료는 소위 twobox 자료다. 이 자료는 그림 2에 나타나 있는데 분류값 있는 자료에는 +1과 -1에 각각 3개씩 있다. 본 연구에서 제안한 방법으로 분류를 한 결과를 그림에서 볼 수 있다. 이 결과는 우리가 원하는 것처럼 +1과 -1 의 값을 가진 자료와 같은 군집에 속하는 자료는 같은 분류값을 가지도록 분류가 잘 되었다. 이 실험에서 사용된 모수, γ_1, γ_2, σ 그리고 σ_w값은 0.03, 100, 0.25, 0.3이었다. 그림에 위의 직선은 SVM으로부터 나온 선형분류선이다. 이 실험에 소요된 CPU-시간은 모수선택을 포함해서 각각 12.70 초와 13.12 초이다.

이상의 실험결과에서 알 수 있듯이 본 연구에서 제안한 방법이 좋은 결과를 보이는 것으로 평가된다. 위의 실험에서 분류값 없는 자료를 시험용자료로 하여 실험을 100회씩 반복하여 오분류율을 살펴보았더니 moon 자료에서는 오분류율이 0이 나왔고 twobox 자료에서는 0.002가 나왔다.

이상의 결과를 볼 때 본 연구에서 제안한 방법이 자료에 적응을 잘 하는 것으로 알 수 있다.

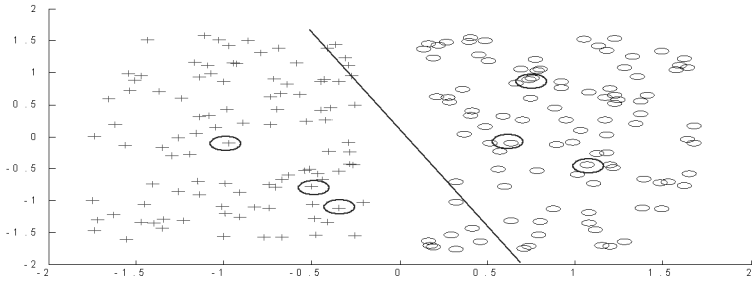

그림 2 TwoBox data와 분류값 없는 자료의 분류값 추정결과

5. 결론

현재까지 개발된 준지도학습은 주어진 분류값 없는 자료의 분류값만을 추정하는 것이다. 그러므로 새로운 자료에 대한 분류값을 추정하고 싶으면 모든 자료를 이용해서

전체 알고리즘을 다시 수행해야하는 단점이 있다. 그래서 본 연구에서는 지도학습처럼 일반화 된 분류함수를 벌칙항이 있는 커널형추정법(regularized kernel estimator)을 이용하여 개발하였다. 개발된 분류함수는 GFRM의 일반형으로 해석이 된다. 두 가지 예제 자료에 적용시켜보니 제안된 방법이 좋은 결과를 보였다. 그러나 다른 준지도학습방법과 비교를 해서 우수성을 보여야 하지만 모수에 많은 영향을 받기 때문에 객관적인 모수선택방법에 대한 연구가 없는 상황에서 비교가 어려울 것 같아 차후 연구과제로 남긴다.

References

1. Belkin, M., Niyogi, P. and Sindhwani, V. (2006) : Manifold regularization: A geometric framework for learning from laveled and unlabeled examples, *Journal of Machine Learning Research,* 1, 1-48.
2. Chapelle, O., Schölkopf, B. and Zien, A. (2006), *Semi-Supervised Learning*. The MIT Press, Cambridge, Massachusetts.
4. Lawrence, N. D., & Jordan, M. I. (2005). Semi-supervised learning via Gaussian processes. In L. K. Saul, Y. Weiss and L. Bottou (Eds.), *Advances in neural information processing systems* 17. Cambridge, MA: MIT Press.
5. Nigam, K., & Ghani, R. (2000). Analyzing the effectiveness and applicability of co-training. *Ninth International Conference on Information and Knowledge Management*, 86 - 93.
7. Niu, Z. Y., Ji, D. H., & Tan, C. L. (2005). Word sense disambiguation using label propagation based semi-supervised learning. *Proceedings of the ACL.*
8. Rosenberg, C., Hebert, M., & Schneiderman, H. (2005). Semi-supervised self-training of object detection models. *Seventh IEEE Workshop on Applications of Computer Vision.*
9. Szummer, M., & Jaakkola, T. (2002). Information regularization with partiallylabeled data. *Advances in Neural Information Processing Systems*, 15.
10. Vapnik, V. N. (1998). *Statistical Learning Theory*, Wiley, NY, 1998.
11. Xu, J., Zhang, X. and Li, Y. (2001). Kernel MSE algorithm: A unified framework for KFD, LS-SVM, *Proceedings of IJCNN'01* 2:1486-1491.
12. Zhou, D., Bousquet, T. N., Lal, J. and Schölkopf(2004), Learning with local and global consistency, *Advances in Neural Information Processing Systems*, 16, 321-328.
13. Zhu, D. (2005). Semi-supervised learning literature survey. *Technical Report Computer Sciences 1530, University of Wisconsin - Madison.*
14. Zhou, X., Ghahramani, Z. and Lafferty, J. (2003). Semi-supervised learning using Gaussian fields and harmonic functions, *In Proc. of the*

20th International Conference on Machine Learning, Washington DC.
15. Zhou, Y., & Goldman, S. (2004). Democratic co-learing. *Proceedings of the16th IEEE International Conference on Tools with Artificial Intelligence* (ICTAI2004).

[2007년 6월 접수, 2007년 8월 채택]