Available online at www.sciencedirect.com
SCIENCE DIRECT®
Pattern Recognition 38 (2005) 1325–1327
PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY
www.elsevier.com/locate/patcog

Rapid and brief communiction

Possibilistic support vector machines

KiYoung Lee[a,b], Dae-Won Kim[b], Kwang H. Lee[a,b], Doheon Lee[b,*]

[a]*Department of Electrical Engineering & Computer Science, Korea Advanced Institute of Science and Technology, Guseong-dong, Yuseong-gu, Daejeon 305-701, Republic of Korea*

[b]*Department of BioSystems and Advanced Information Technology Research Center, Korea Advanced Institute of Science and Technology, Guseong-dong, Yuseong-gu, Daejeon 305-701, Republic of Korea*

Received 1 November 2004; accepted 8 November 2004

Abstract

We propose new support vector machines (SVMs) that incorporate the geometric distribution of an input data set by associating each data point with a possibilistic membership, which measures the relative strength of the self class membership. By using a possibilistic distance measure based on the possibilistic membership, we reformulate conventional SVMs in three ways. The proposed methods are shown to have better classification performance than conventional SVMs in various tests.

Keywords: Classification; Support vector machines; Possibilistic SVMs; Geometric distribution; Possibilistic distance

1. Introduction

Despite many benefits of support vector machines (SVMs) [1,2], there is no mechanism for handling variations in the significance of data points, so that all the data points are treated identically in conventional SVMs. In many real-world problems, however, each data point in the data set for classification problems may differ in the degree of significance due to noise, inaccuracies, or abnormal characteristics; for example, outliers can lead to the inaccuracies in a prediction phase. Hence, if all data are treated as equivalent without considering such differences when finding optimal hyperplanes (OHPs) in SVMs, the OHPs identified are likely to be suboptimal. To solve the aforementioned problems in SVMs, we propose new SVMs to take into account the variation of the geometric significance in a data set by introducing a possibilistic membership for each data point. The membership is the relative strength of the self class membership, through which the inherent geometric distribution of the data set is reflected. By using a possibilistic distance measure based on the membership, we reformulate conventional SVMs.

2. Possibilistic SVMs (PSVMs)

2.1. Possibilistic membership and distance

We propose two methods to extract a possibilistic membership value for each data point from a given data set (Methods I and II). Method I extracts the relative strength of self class membership compared with the non-self class membership. Let us calculate a possibilistic membership value for $\mathbf{x}_k \in i$th class. By using a mean distance between $\mathbf{x}_k$ and $\mathbf{x}_l \in i$th class (mean distance from self-class), and an mean distance between $\mathbf{x}_k$ and $\mathbf{x}_j \in j$th class (mean distance from non-self class), the possibilistic membership value for $\mathbf{x}_k$ is defined as

$$\mu_k = \left(1 + \frac{\sum_{l=1}^{m} \|\mathbf{x}_k - \mathbf{x}_l\|/m}{\sum_{j=1}^{n} \|\mathbf{x}_k - \mathbf{x}_j\|/n}\right)^{-1}, \quad k = 1, \ldots, m, \quad (1)$$

* Corresponding author. Tel.: +82 42 869 4316; fax: +82 42 869 8680.
E-mail address: dhlee@bisl.kaist.ac.kr (D. Lee).

0031-3203/$30.00
doi:10.1016/j.patcog.2004.11.018

where m and n are the number of data of the ith and the jth class, respectively. The possibilistic membership for $\mathbf{x}_j \in j$th class is calculated in a similar manner. In this method, the closer the data point is to the self-class in comparison to non-self-class, the higher membership the data point has.

In Method II, the possibilistic membership for $\mathbf{x}_k$ is defined using the Mahalanobis distance to reflect the geometric shape of the self-class:

$$\mu_k = 1 - \frac{(\mathbf{x}_k - v_i)^{\mathrm{T}} \mathbf{C}_i^{-1} (\mathbf{x}_k - v_i)}{D_{max}}, \quad k = 1, \ldots, m, \tag{2}$$

where v_i $(=1/m \sum_{l=1}^{m} \mathbf{x}_l)$ is the center of the ith class, $\mathbf{C}_i = \sum_{l=1}^{m} (\mathbf{x}_l - v_i)(\mathbf{x}_l - v_i)^{\mathrm{T}}$ is the covariance matrix of data in the ith class, and $D_{max} = \max_l (\mathbf{x}_l - v_i)^{\mathrm{T}} \mathbf{C}_i^{-1} (\mathbf{x}_l - v_i)$. A higher possibilistic membership value is assigned to a data point with lower distance to the center.

To incorporate possibilistic membership values into the search of an optimal hyperplane, we introduce a *possibilistic distance*. Suppose that each data point $\mathbf{x}_k$ has a possibilistic membership value $0 \leqslant \mu_k \leqslant 1$. Then we define a possibilistic distance, δ_k, between $\mathbf{x}_k$ and a hyperplane $(\mathbf{w}, b)$ as

$$\delta_k = \begin{cases} \dfrac{|\mathbf{w} \cdot \mathbf{x}_k + b|}{(\mu_k)^{\tau} \|\mathbf{w}\|} & \text{if } \mu_k \neq 0, \\ \infty & \text{otherwise,} \end{cases} \tag{3}$$

where τ $(\in R)$ is a control coefficient. Note that δ_k decreases with increasing μ_k. Hence, a data point with a larger value of μ_k has a stronger influence on the search of the OHP. When $\tau = 0$, δ_k equals the Euclidean distance so that proposed methods behave like conventional SVMs. As $\tau \to \infty$, δ_k's for data with $\mu_k < 1$ approach infinity, and those data points are neglected in the search for the OHP as a result.

2.2. Formulation for possibilistic support vector machines

2.2.1. Formulation of PSVMs for the linearly separable case

For a linearly separable data set S, we find the OHP by maximizing *the minimum possibilistic distance* from a hyperplane. In this case, a margin can be written as $\min_k [1/(\mu_k)^{\tau}] |\mathbf{w} \cdot \mathbf{x}_k + b| / \|\mathbf{w}\|$, where $1 \leqslant k \leqslant N$ and N is the total number of data points. Because scalar multiplication does not affect an identity, we can normalize the hyperplane $(\mathbf{w}, b)$ to satisfy $\min_k [1/(\mu_k)^{\tau}] |\mathbf{w} \cdot \mathbf{x}_k + b| = 1$. Then, the objective function to maximize the margin is defined as

$$\min\{O = \tfrac{1}{2} \mathbf{w} \cdot \mathbf{w}\}, \tag{4}$$

subject to $[1/(\mu_k)^{\tau}] y_k (\mathbf{w} \cdot \mathbf{x}_k + b) - 1 \geqslant 0$. By using Lagrange multipliers α_k and KKT conditions [2], we obtain the dual representation of Eq. (4):

$$\max \left\{ D(\alpha) = \sum_{k=1}^{N} \alpha_k - \frac{1}{2} \sum_{k=1}^{N} \sum_{l=1}^{N} \frac{1}{(\mu_k)^{\tau}} \times \frac{1}{(\mu_l)^{\tau}} \alpha_k \alpha_l y_k y_l \mathbf{x}_k \cdot \mathbf{x}_l \right\}, \tag{5}$$

subject to $\sum_{k=1}^{N} [1/(\mu_k)^{\tau}] \alpha_k y_k = 0$ $(\alpha_k \geqslant 0)$. When $\tau = 0$, Eq. (5) is equivalent to the separable case in conventional SVMs. The offset b_o of the OHP can be calculated by the complementary KKT condition, $\breve{\alpha}_k ([1/(\mu_k)^{\tau}] y_k (\mathbf{w} \cdot \mathbf{x}_k + b) - 1) = 0$, where $\breve{\alpha}_k$ is the solution of Eq. (5). The decision function for an unseen data point $\mathbf{x}_u$ is

$$f(\mathbf{x}_u) = sgn \left(\sum_{k=1}^{N} \frac{1}{(\mu_k)^{\tau}} \breve{\alpha}_k y_k \mathbf{x}_k \cdot \mathbf{x}_u + b_o \right). \tag{6}$$

2.2.2. Formulation of PSVMs for the soft margin

If a data set S is linearly non-separable, we permit misclassification using a slack variable ζ_k that is the distance from the margin of self class. Using ζ_k, the OHP for this case is found by

$$\min \left\{ O = \frac{1}{2} \mathbf{w} \cdot \mathbf{w} + C \sum_{k=1}^{N} (\mu_k)^{\upsilon\tau} \zeta_k \right\}, \tag{7}$$

subject to $[1/(\mu_k)^{\tau}] y_k (\mathbf{w} \cdot \mathbf{x}_k + b) - 1 + \zeta_k \geqslant 0$ $(\zeta_k \geqslant 0)$, where C $(\in R^+)$ is a regulation parameter. Since ζ_k is related to the penalty for misclassification caused by $\mathbf{x}_k$, the term $(\mu_k)^{\upsilon\tau} \zeta_k$ can be the measure of the margin error with membership-dependent weighting for $\mathbf{x}_k$, where υ is a constant. Note that since ζ_k for misclassified data point is amplified by $1/(\mu_k)^{\tau}$, υ $(\geqslant 1)$ is necessary to reduce the effect of misclassified data on a hyperplane. In a similar manner to the linearly separable case, we obtain the dual representation for this case.

$$\max \left\{ D(\alpha) = \sum_{k=1}^{N} \alpha_k - \frac{1}{2} \sum_{k=1}^{N} \sum_{l=1}^{N} \frac{1}{(\mu_k)^{\tau}} \times \frac{1}{(\mu_l)^{\tau}} \alpha_k \alpha_l y_k y_l \mathbf{x}_k \cdot \mathbf{x}_l \right\}, \tag{8}$$

subject to $\sum_{k=1}^{N} [1/(\mu_k)^{\tau}] \alpha_k y_k = 0$ $(0 \leqslant \alpha_k \leqslant (\mu_k)^{\upsilon\tau} C)$. The decision function for an unseen data point $\mathbf{x}_u$ is also given as Eq. (6).

2.2.3. Formulation of PSVMs for kernelization

As seen in Eqs. (6) and (8), the dual form of the objective function and the decision function of PSVMs are represented entirely in terms of inner products of pairs of input vectors.

Thus, we can kernelize the PSVMs. The kernelized version of the decision function in Eq. (6) for PSVMs is given as

$$f(\mathbf{x}_u) = sgn\left(\sum_{k=1}^{N} \frac{1}{(\mu_k)^{\tau}} \breve{\alpha}_k y_k K(\mathbf{x}_k, \mathbf{x}_u) + b_o\right). \quad (9)$$

3. Experimental results

To evaluate the performance of our methods, we applied conventional SVMs and PSVMs using two possibilistic membership extraction methods (Methods I and II) to Hungarian Heart Disease (13 attributes/2 classes/294 data), Iris (4/3/150), and Wine Recognition (13/3/178) [3], and Leukemia data (50/2/38) [4]. In these tests, we evaluated the prediction accuracies of two-fold cross validations for each data set.

The mean prediction accuracies of five runs for each data set are given in Table 1. Note that PSVMs using Methods I and II have better performance than conventional SVMs for all data sets used. Moreover, PSVMs using Method II showed slightly better accuracies than PSVMs using Method I except for Wine data. Notably, for Leukemia data, PSVMs showed remarkable improvement over conventional SVMs. From this observation, we conjecture that PSVMs have a potential for solving the overfitting problems in SVMs when the number of data is small compared with the dimension of features like Leukemia.

Table 1
Mean accuracies (%) of conventional SVMs and PSVMs

Data set	Conventional SVMs	PSVMs Method I	PSVMs Method II
Hungarian	72.83	75.70 (3.94%)	76.59 (**5.16**%)
Iris	96.33	98.37 (**2.12**%)	98.37 (**2.12**%)
Wine	93.15	95.28 (**2.29**%)	95.17 (2.17%)
Leukemia	65.79	94.21 (43.20%)	96.84 (**47.20**%)

The improvement of the proposed methods over conventional SVMs is specified in the parentheses.

4. Conclusions

We have proposed new SVMs that can reflect the geometric significance in input data. The proposed methods using the two possibilistic membership extraction methods outperformed the conventional SVMs in all test cases.

Acknowledgements

This work was supported by Metabolites Analysis and Function Research Grant (2004-02145) from the Ministry of Science and Technology. We would like to thank CHUNG Moon Soul Center for BioInformation and BioElectronics and the IBM SUR program for providing research and computing facilities.

References

[1] C.J.C. Burges, A tutorial on support vector machines for pattern recognition, Data Mining and Knowledge Discovery 2 (1998) 121–1678.
[2] B. Schölkopf, A.J. Smola, Learning with Kernels, MIT Press, MA, 2002.
[3] C.L. Blake, C.J. Merz, UCI repository of machine learning database, http://www.ics.uci.edu/∼mlearn/MLRepository.html/, 1998.
[4] T. Golub, D. Slonim, P. Tamayo, et al., Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, Science 286 (1999) 531–537.