Available online at www.sciencedirect.com
SCIENCE DIRECT®
Pattern Recognition 38 (2005) 1768–1771
PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY
www.elsevier.com/locate/patcog

Rapid and brief communication

Improving support vector data description using local density degree

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

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

Received 4 March 2005; accepted 22 March 2005

Abstract

We propose a new support vector data description (SVDD) incorporating the local density of a training data set by introducing a local density degree for each data point. By using a density-induced distance measure based on the degree, we reformulate a conventional SVDD. Experiments with various real data sets show that the proposed method more accurately describes training data sets than the conventional SVDD in all tested cases.

Keywords: D-SVDD; Support vector data description; One-class classification; Data domain description; Outlier detection

1. Introduction

The purpose of data domain description is to give a compact description of a set of data referred to as target data. It is usually used for outlier detection or a conventional classification problem especially where one of the classes is undersampled [1]. Recently, a support vector data description (SVDD) inspired by support vector machines [2] was invented by Tax and Duin [3]. In a SVDD, the compact description of target data is given as a hypersphere with minimal volume containing most of the data objects in a high-dimensional feature space using some kernel functions [3]. Despite the usefulness of a SVDD [3], the conventional SVDD (C-SVDD) does not take into account the density distribution of a target data set, since, it only considers the small portion of data that lie around the most outer region in a high-dimensional feature space. In many real world problems, however, each target data point may differ in the degree of significance due to its density: the target data in a higher density region are more significant than those in a lower density region in describing a target data set because the data in a higher density region should be included in the compact description than other data. Hence, if all data are treated as equivalent in describing a target data set, without considering such difference in density degree, the solutions are likely to be less optimal.

To address the above problem in the C-SVDD and find a more robust and more reliable description of a target data set, we propose a new SVDD to reflect the different local density of a target data set by introducing the notion of a local density degree for each data point. By using a density-induced distance measure based on the degree, we generalize the C-SVDD. We refer to the proposed method as a density-induced SVDD (D-SVDD).

* Corresponding author. Tel.: +82 42 8694316; fax: +82 42 8698680.
E-mail address: dhlee@bisl.kaist.ac.kr (D. Lee).

0031-3203/$30.00
doi:10.1016/j.patcog.2005.03.020

2. Density-induced support vector data description

2.1. Extraction of local density degree

In this paper, we propose a method to extract a local density degree for each data point from a target data set using a nearest neighborhood approach. Let us calculate a local density degree ρ_i for a target data point $\mathbf{x}_i$. By using $d(\mathbf{x}_i, \mathbf{x}_i^K)$, the distance between $\mathbf{x}_i$ and $\mathbf{x}_i^K$ (the Kth nearest neighborhood of $\mathbf{x}_i$), and the mean distance of Kth nearest neighborhoods of all target data, $MEAN^K$, the local density degree $\rho_i > 0$ for $\mathbf{x}_i$ is defined by

$$\rho_i = \exp\left\{\omega \times \frac{MEAN^K}{d(\mathbf{x}_i, \mathbf{x}_i^K)}\right\}, \quad i = 1, \ldots, n, \tag{1}$$

where $MEAN^K = (1/n)\sum_{i=1}^n d(\mathbf{x}_i, \mathbf{x}_i^K)$, n is the number of data in a target class, and $0 \leqslant \omega \leqslant 1$ is a weighting factor. Note that this method reports higher local density degree ρ_i for the data in a higher density region: the data with lower $\mathbf{x}_i^K$ have higher ρ_i values. Moreover, a bigger ω produces higher local density degrees.

To incorporate the density degrees into search of the optimal description in a SVDD, we introduce a new geometric distance called a density-induced distance. Suppose each target data point can be expressed as $(\mathbf{x}_i, \rho_i)$. We define a density-induced distance, δ_i, between $\mathbf{x}_i$ and the center of a hypersphere $(\mathbf{a}, R)$ as

$$\delta_i \equiv \{\rho_i(\mathbf{x}_i - \mathbf{a}) \cdot (\mathbf{x}_i - \mathbf{a})\}^{1/2}, \tag{2}$$

where $\mathbf{a}$ and R are the center and the radius of the hypersphere, respectively. Note that δ_i increases with increasing ρ_i. Hence, to enclose the data point with increased δ_i owing to higher local density degree ρ_i, the radius of a minimum-sized hypersphere should be increased; the data point with higher density degree has stronger influence on the search of the minimum-sized hypersphere.

2.2. Mathematical formulation

We first find a hyperspherical model $(\mathbf{a}, R)$ which gives a closed boundary around target data with no training error with regard to the density-induced distance. By minimizing R, we find the optimal hypersphere which includes all the target data. Then, we can obtain the optimal hypersphere $(\mathbf{a}^*, R^*)$ by minimizing the objective function O:

$$O = R^2 \quad \text{subject to } \rho_i(\mathbf{x}_i - \mathbf{a}) \cdot (\mathbf{x}_i - \mathbf{a}) \leqslant R^2, \quad i = 1, \ldots, n. \tag{3}$$

To allow the possibility of training error, and therefore to make the model more robust, the density-induced distance between each target data $\mathbf{x}_i$ and the center $\mathbf{a}$ does not have to be strictly smaller than R, but data points with distance larger than R should be penalized. We handle this case using a slack variable $\zeta_i \geqslant 0$ which is the distance between the boundary Ω and $\mathbf{x}_i$ outside Ω. Using the slack variable for each target data point, we change the objective function in Eq. (3) into

$$O = R^2 + C\sum_{i=1}^n \zeta_i \quad \text{subject to } \rho_i(\mathbf{x}_i - \mathbf{a}) \cdot (\mathbf{x}_i - \mathbf{a}) \leqslant R^2 + \zeta_i, \quad \zeta_i \geqslant 0, \ \ i = 1, \ldots, n, \tag{4}$$

where $C > 0$ is a control parameter which gives the trade-off between the volume of the description and the training errors. Note that ζ_i plays a similar role with the slack variable in the C-SVDD [3], but it has a different meaning. ζ_i equals $\delta_i^2 - R^2$ for a training error data point, otherwise it is 0. It implies that ζ_i also contains the information of local density.

Minimizing Eq. (4) is an optimization problem. Therefore by introducing Lagrange multipliers, we can construct the Lagrangian:

$$\begin{aligned} L(R, \mathbf{a}, \zeta, \alpha, \beta) &= R^2 + C\sum_{i=1}^n \zeta_i - \sum_{i=1}^n \alpha_i\{R^2 + \zeta_i \\ &\quad - \rho_i(\mathbf{x}_i \cdot \mathbf{x}_i - 2\mathbf{a} \cdot \mathbf{x}_i + \mathbf{a} \cdot \mathbf{a})\} - \sum_{i=1}^n \beta_i\zeta_i, \end{aligned} \tag{5}$$

where $\alpha_i \geqslant 0$ and $\beta_i \geqslant 0$ are Lagrange multipliers, from which we can derive the following conditions at the solution point:

$$\sum_{i=1}^n \alpha_i = 1, \quad \mathbf{a} = \frac{1}{T}\sum_{i=1}^n \alpha_i\rho_i\mathbf{x}_i,$$
$$T = \sum_{i=1}^n \alpha_i\rho_i, \quad \beta_i = C - \alpha_i. \tag{6}$$

Combining the conditions with Eq. (5), we obtain the dual representation of the optimization problem: maximize $D(\alpha)$

$$D(\alpha) = \sum_{i=1}^n \alpha_i\rho_i\mathbf{x}_i \cdot \mathbf{x}_i - \frac{1}{T}\sum_{i=1}^n\sum_{j=1}^n \alpha_i\alpha_j\rho_i\rho_j\mathbf{x}_i \cdot \mathbf{x}_j \tag{7}$$

subject to $\sum_{i=1}^n \alpha_i = 1$, $0 \leqslant \alpha_i \leqslant C$, $T = \sum_{i=1}^n \alpha_i\rho_i$, $i = 1, \ldots, n$. Note that the dual form for this case has only the Lagrange multiplier α_i; other variables and Lagrange multiplier β_i have disappeared. Furthermore, when $\rho_i = 1$, this dual representation is equivalent to the formalism of a C-SVDD [3]. Thus, this proposed method is a general extension of the C-SVDD.

This dual representation is a linear constrained optimization problem; hence we can derive the $\breve{\alpha}_k$ that satisfies the Eq. (7). After solving Eq. (7), we can derive the $\mathbf{a}^*$ and the R^* of the solution of the problem from Eqs. (6) and (4), respectively. Unlike in the C-SVDD [1], $\mathbf{a}^*$ is weighted by the local density degree ρ_i. The center of the optimal hypersphere is shifted to a higher density region. Using a

Table 1
The average error rates (%) of ten independent runs for IRIS and LEUKEMIA

Class no.	k-NNDD		C-SVDD			D-SVDD		
	$k=1$	$k=3$	Poly-3	Poly-5	Gaussian	Poly-3	Poly-5	Gaussian
IRIS								
0	4.20	6.73	3.40	31.33	3.47	2.80	2.93	**0.42**
1	13.93	14.20	10.00	33.93	7.73	9.80	9.67	**6.50**
2	18.00	19.20	13.53	35.67	9.33	13.27	13.33	**9.25**
Total	12.04	13.38	8.98	33.64	6.84	8.62	8.64	**5.39**
LEUKEMIA								
0	30.53	23.42	13.68	29.47	18.68	12.63	14.21	**7.37**
1	12.63	12.63	13.42	39.21	18.68	10.26	12.11	**4.74**
Total	21.58	18.03	13.55	34.34	18.68	11.45	13.16	**6.05**

indication function I [3], the decision function for a test data point $\mathbf{x}_t$ can be represented as:

$$f(\mathbf{x}_t) = I\left(\mathbf{x}_t \cdot \mathbf{x}_t - \frac{2}{T}\sum_{i=1}^{n}\rho_i\breve{\alpha}_i\mathbf{x}_t \cdot \mathbf{x}_i - \frac{1}{T^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\rho_i\rho_j\breve{\alpha}_i\breve{\alpha}_j\mathbf{x}_i \cdot \mathbf{x}_j \leqslant R^{*2}\right). \quad (8)$$

As seen in Eqs. (7) and (8), the dual form of the objective function and the decision function of the D-SVDD are represented entirely in terms of inner products of pairs of target data points. Thus, we can kernelize the D-SVDD for flexible description. The kernelized version of the decision function for the D-SVDD is

$$f(\mathbf{x}_t) = I\left(K(\mathbf{x}_t, \mathbf{x}_t) - \frac{2}{T}\sum_{i=1}^{n}\rho_i\breve{\alpha}_i K(\mathbf{x}_t, \mathbf{x}_i) - \frac{1}{T^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\rho_i\rho_j\breve{\alpha}_i\breve{\alpha}_j K(\mathbf{x}_i, \mathbf{x}_j) \leqslant R^{*2}\right), \quad (9)$$

where $K(\cdot,\cdot)$ is a kernel function [2].

3. Experiments and conclusion

To investigate the success of these attempts, we conducted various tests in which three versions of the C-SVDD and three versions of the proposed method were applied to IRIS [4], and LEUKEMIA [5]. Two polynomial kernel functions and a Gaussian kernel function were used for flexible description [2]. The model parameters were found by cross validation to identify optimal solutions of the C-SVDD. The same parameter set with the C-SVDD and $K=3$ in Eq. (1) were used for the proposed method. We conducted the same experiments with two versions of a k-nearest-neighbor data description method [1], k-NNDD.

The average error rates of prediction accuracies of ten independent runs for the data sets are given in Table 1. The label of a target data class is indicated in the first column; the data in other classes are the candidates of negative data that should not be included in a target data description. For the IRIS data set, the two versions of the k-NNDD method showed 4.20 and 6.73% average error rates when the label of a target class is 0. For the same data sets, the C-SVDD showed 3.40, 31.33 and 3.47% error rates for each version. The proposed D-SVDD method, however, showed 2.80, 2.93 and 0.42% error rates. That is the D-SVDD improved the C-SVDD in all versions used, and the D-SVDD with a Gaussian kernel function had the best performance. Moreover, when a degree-5 polynomial kernel function was used, the performance of the D-SVDD was not severely deteriorated, which is not the case with the C-SVDD because the results of the C-SVDD with a higher degree of a polynomial kernel are dominantly determined by the data with larger norms [3]. Similar results were obtained for the LEUKEMIA data set. For the LEUKEMIA data set, the D-SVDD method had 6.05% error rate whereas the error rates with the C-SVDD and the k-NNDD method were respectively 18.68 and 18.03% when a Gaussian kernel function was used or $k=3$ was used. From these results, we draw a conclusion that the proposed method showed better prediction accuracies than the C-SVDD for all the data sets used regardless of the type of kernel functions. Moreover, the best performance was obtained when the D-SVDD with a Gaussian kernel function was used.

Acknowledgements

This work was supported by the Korean Systems Biology Research Grant (2005-00343) 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] D.M.J. Tax, R.P.W. Duin, Support vector domain description, Pattern Recognition Lett. 20 (1999) 1191–1199.
[2] V. Vapnik, The Nature of Statistical Learning Theory, Springer, New York, 1995.
[3] D.M.J. Tax, R.P.W. Duin, Support vector data description, Mach. Learn. 54 (2004) 45–66.
[4] C.L. Blake, C.J. Merz, UCI repository of machine learning database, http://www.ics.uci.edu/~mlearn/MLRepository.html/, 1998.
[5] 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.