Learning the Kernel Parameters in Kernel Minimum Distance Classifier

Daoqiang Zhang[1,2], Songcan Chen[2] and Zhi-Hua Zhou[1*]

[1] National Laboratory for Novel Software Technology

Nanjing University, Nanjing 210093, China

[2] Department of Computer Science and Engineering

Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China

Abstract

Choosing appropriate values for kernel parameters is one of the key problems in many kernel-based methods because the values of these parameters have significant impact on the performances of these methods. In this paper, a novel approach is proposed to learn the kernel parameters in kernel minimum distance (KMD) classifier, where the values of the kernel parameters are computed through optimizing an objective function designed for measuring the classification reliability of KMD. Experiments on both artificial and real-world datasets show that the proposed approach works well on learning kernel parameters of KMD.

Keywords: Kernel minimum distance; kernel parameter optimization; kernel selection

1. Introduction

Minimum distance (MD) and nearest neighbor (NN) are simple but popular techniques in pattern recognition. Recently, both methods have been extended to kernel versions, i.e. the kernel minimum distance (KMD) and kernel nearest neighbor (KNN), for classifying complex and nonlinear patterns such as faces [1], [2]. However, like other kernel-based methods, the performance of KMD and KNN is greatly affected by the selection of kernel parameters values. In this paper, we focus on optimizing the kernel parameters for KMD.

In the literature, there are two widely used approaches in choosing the values of kernel parameters in kernel-based methods [1], [3], [4]. The first approach empirically chooses a series of candidate values for the kernel parameter, executes the concerned method under these values again

* Corresponding author. Email: zhouzh@nju.edu.cn, Tel.: +86-25-8368-6268, Fax: +86-25-8368-6268

and again, and selects the one corresponding to the best performance as the final kernel parameter value. However, this approach suffers from the fact that only a very limited candidate values are considered, therefore the performance of the kernel-based methods may not be optimized. The second approach is the well-known cross-validation, which is also widely used in model selection. Compared with the first approach, cross-validation often yields better performance because it searches the optimal value for kernel parameter in a much wider range. However, performing cross-validation is often time-consuming and hence it cannot be used to adjust the kernel parameters in real time [3]. Furthermore, when there are only a limited number of training examples, the cross-validation approach can hardly ensure robust estimation.

In this paper, a novel approach is proposed to learn the kernel parameters in KMD. At first an objective function is defined to measure the classification reliability of KMD with different kernel parameters. Then, the optimal values of the kernel parameters are chosen through optimizing the above defined objective function. Experiments on both artificial and real-world datasets show the effect of the proposed approach on learning kernel parameters in KMD.

2. Kernel minimum distance classifier

One of the key ingredients of KMD is the definition of kernel-induced distance measures. Given a data set $S = \{x_1, ..., x_l\}$ sampled from the input space X, a kernel $K(x,y)$ and a function Φ in a feature space satisfy $K(x,y) = \Phi(x)^{\mathrm{T}}\Phi(y)$. An important property of the kernel is that it can be directly constructed in the original input space without knowing the concrete form of Φ. That is, a kernel implicitly defines a nonlinear mapping function. There are several typical kernels, e.g. the Gaussian kernel $K(x, y) = \exp\left(\frac{\| x - y \|^2}{2\sigma^2}\right)$, the polynomial kernel $K(x, y) = (x^T y + 1)^d$, etc. The kernel-induced distance between two points defined by a kernel K is shown in Eq. (1).

$$d^2(x, y) = \| \Phi(x) - \Phi(y) \|^2 = K(x, x) - 2K(x, y) + K(y, y). \quad (1)$$

Suppose the training data set S contains c different classes, i.e. $S_1, S_2, ..., S_c$, and each class S_i has l_i samples, satisfying $\sum_{i=1}^{c} l_i = l$. Let $\Phi(S_i) = \{\Phi(x_j) \mid x_j \in S_i\}$ be the image of class S_i under the map Φ, and denote the centre of $\Phi(S_i)$ as

$$\Phi_{S_i} = \frac{1}{l_i} \sum_{x_j \in S_i} \Phi(x_j) . \tag{2}$$

Then, the distance between the image of a new point x and the centre of class Φ_{S_i} can be computed as

$$\begin{aligned} d^2(\Phi(x), \Phi_{S_i}) &= \| \Phi(x) - \Phi_{S_i} \|^2 \\ &= \Phi^T(x)\Phi(x) + \Phi_{S_i}^T \Phi_{S_i} - 2\Phi^T(x)\Phi_{S_i} \\ &= K(x,x) + \frac{1}{l_i^2} \sum_{x_j, x_k \in S_i} K(x_j, x_k) - \frac{2}{l_i} \sum_{x_j \in S_i} K(x, x_j) \end{aligned} \tag{3}$$

According to Eq. (3), the classification rule in KMD is to assign the new point x to the class with the smallest distance:

$$h(x) = \arg\min_{1 \le i \le c} \left\{ d^2(\Phi(x), \Phi_{S_i}) \right\} \tag{4}$$

3. The proposed method

The following objective function is defined to measure the classification reliability of KMD with different kernel parameters:

$$J(\theta) = \sum_{i=1}^{l} \exp\left(\frac{d^2(\Phi(x_i), \Phi_{S_{\pi(i)}})}{\min\limits_{\substack{1 \le j \le c \\ j \ne \pi(i)}} \left(d^2(\Phi(x_i), \Phi_{S_j}) \right)} \right) \tag{5}$$

Here θ denotes the kernel parameters, and $\pi(i)$ denotes the class label of x_i. The intuition behind Eq. (5) is to make the distance between the image of a sample and the centre of its corresponding class as small as possible, while to make the distance between the image of the sample to other classes as large as possible. The smaller the value of the objective function, the higher the classification reliability. Here the exponential function is used for speeding up the convergence of optimization.

Note that when $d^2(\Phi(x_i), \Phi_{S_{\pi(i)}}) < \min\limits_{\substack{1 \le j \le c \\ j \ne \pi(i)}} \left(d^2(\Phi(x_i), \Phi_{S_j}) \right)$, the sample x_i is correctly classified. Equation (5) specifies that the optimal value for a kernel parameter should not only correctly classify the training data, but also make the classification reliability as high as possible. In the extreme case where $d^2(\Phi(x_i), \Phi_{S_{\pi(i)}}) = 0$ and $\min\limits_{\substack{1 \le j \le c \\ j \ne \pi(i)}} \left(d^2(\Phi(x_i), \Phi_{S_j}) \right) = \infty$ for each x_i, the highest classification reliability is obtained.

The optimal values of the kernel parameters can be obtained through minimizing Eq. (5), i.e.

$$\theta^* = \arg\min_{\theta} J(\theta) . \tag{6}$$

In this paper, an iterative algorithm is employed to generate θ^*. According to the general gradient method, the updating equation for minimizing the objective function J is given by

$$\theta^{(n+1)} = \theta^{(n)} + \eta\left(\frac{\partial J}{\partial \theta}\right) \tag{7}$$

Where η is the learning rate and n is the iteration step.

The proposed method KMD-opt is summarized as follows:

Step 1. Set the learning rate η and the maximum iteration number *N*, and set ε to a very small positive number.

Step 2. Initialize the kernel parameters $\theta = \theta^{(0)}$ and set the iteration step $n = 0$.

Step 3. Update the kernel parameters $\theta^{(n)}$ using Eq. (7).

Step 4. If $|\theta^{(n+1)} - \theta^{(n)}| < \varepsilon$ or $n \geq N$, stop. Otherwise, set $n = n+1$, goto Step 3.

4. Experiments

This section evaluates the effectiveness of the proposed KMD-opt method. For comparison, the MD and KMD are also tested. An artificial data set *Circles*, as shown in Fig. 1, and two real-world data sets *Bupa* and *Pid* from UCI Machine Learning Repository [5] are used. For each data set, half of data are used as the training data set, while the remaining data are used as the test data set. The kernel used in the experiments is the Gaussian kernel $K(x, y) = \exp\left(\frac{\| x - y \|^2}{2\sigma^2}\right)$, where σ is the kernel parameter that should be optimized. In this paper, if without explicit explanations, the initial value for the kernel parameter σ is set to $\sigma_0 = \frac{1}{c}\left(\sqrt{\frac{\sum_{j=1}^{l} \| x_j - \bar{x} \|^2}{l}}\right)$, where $\bar{x}$ is the centroid of the total l training data. Specifically, the σ_0 values for *Circles*, *Bupa* and *Pid* are 2.33, 17.26 and 43.48 respectively. The learning rate η is set to 0.05 and ε is set to 0.0001 without extra explanations.

Table 1 shows the test accuracies of MD, KMD and KMD-opt. The σ values are also presented. Table 1 shows that in most cases KMD obtains better test accuracy than MD, but when the kernel parameter is not chosen appropriately its performance deteriorates greatly. In all cases,

KMD-opt achieves the best test accuracy. What is more, from Table 1, it can be found the KMD-opt method is quite robust because on every data set, the final σs it produced are almost with the same value although the method is with different initializations.

As an example, the left part of Fig. 2 plots the test accuracy of KMD under a series of σ values on *Bupa*. It verifies the claim that a good performance of KMD greatly depends on the selection of kernel parameters. The right part of Fig. 2 plots the objective function in Eq. (5) under a series of σ values on *Bupa*. It can be seen from Fig. 2 that the objective function reaches its minimum at similar σ values as those at which KMD achieves its highest accuracy.

5. Conclusions

In this paper, a novel approach for learning the kernel parameters is proposed and successfully applied to the kernel minimum distance (KMD) classifier. An objective function is defined to measure the classification reliability of KMD with different kernel parameters, and then the optimal values of the kernel parameters are obtained by optimizing the objective function. Experiments show the effect of the proposed approach on learning kernel parameters in KMD. In future works, the proposed approach will be extended for other kernel-based learning methods such as support vector machine (SVM) and kernel fisher discriminate (KFD).

References

[1] J. Peng, D.R. Heisterkamp, H.K. Dai, Adaptive quasiconformal kernel nearest neighbor classification, IEEE Trans. PAMI 26 (5) (2004) 656-661.

[2] J. Shawe-Taylor, N. Cristianini, Kernel methods for pattern analysis, Cambridge University Press, 2004.

[3] L. Wang, K.L. Chan, Learning kernel parameters by using class separability measure, NIPS'02 Workshop on Kernel Machines, Canada, 2002.

[4] D.Q. Zhang, S.C. Chen, Clustering incomplete data using kernel-based fuzzy c-means algorithm, Neural Processing Letters 18(3) (2003) 155-162.

[5] C. Blake, E. Keogh, and C.J. Merz, UCI repository of machine learning databases [http://www.ics.uci.edu/~mlearn/MLRepository.html], Department of Information and Computer Science, University of California, Irvine, CA, 1998.

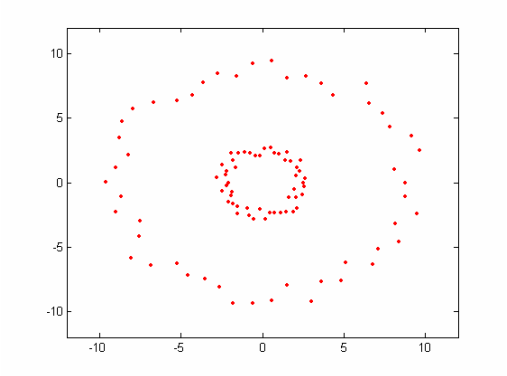

Fig. 1. The *Circles* data set

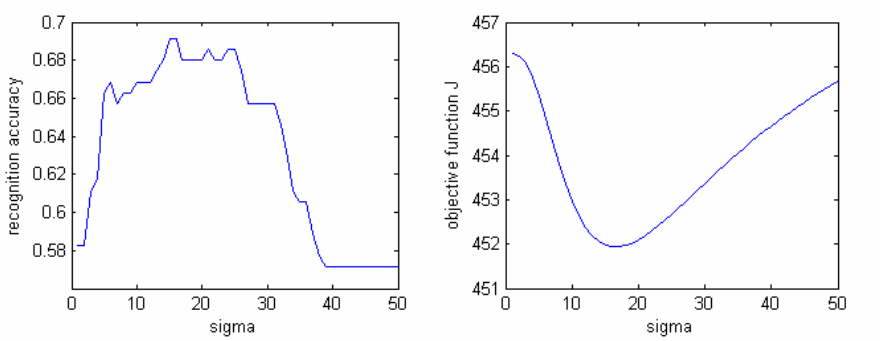

Fig. 2. Test accuracy (left) and objective function values (right) under a series of σ values on *Bupa*.

Table 1. Comparisons of test accuracy (%) of MD, KMD and KMD-opt (the values in the brackets denote the σ values at convergence).

Data sets	MD	KMD					KMD-opt				
		σ_0	$2\sigma_0$	$3\sigma_0$	$\sigma_0/2$	$\sigma_0/3$	σ_0	$2\sigma_0$	$3\sigma_0$	$\sigma_0/2$	$\sigma_0/3$
Circles	50	100	98	50	100	100	100(3.43)	100(3.43)	100(3.43)	100(3.43)	100(3.43)
Bupa	59.43	68	61.14	57.14	66.29	65.14	69.14(16.55)	69.14(16.55)	69.14(16.55)	69.14(16.54)	69.14(16.54)
Pid	62.5	65.1	58.07	50.78	64.32	64.84	65.63(41.45)	65.63(41.45)	65.63(41.45)	65.63(41.45)	65.63(41.45)