Information Sciences 322 (2015) 133–149

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

An efficient Gaussian kernel optimization based on centered kernel polarization criterion

Meng Tian [a,b], Wenjian Wang [a,*]

[a] *School of Computer and Information Technology, Shanxi University, Taiyuan 030006, PR China*
[b] *School of Science, Shandong University of Technology, Zibo 255049, PR China*

ARTICLE INFO

Article history:
Received 20 October 2014
Received in revised form 15 April 2015
Accepted 11 June 2015
Available online 17 June 2015

Keywords:
Gaussian kernel
Kernel selection
Pattern classification
Centered kernel polarization
Local multiclass centered kernel polarization

ABSTRACT

The success of kernel-based learning methods is heavily dependent on the choice of a kernel function and proper setting of its parameters. In this paper, we optimize the Gaussian kernel for binary-class problems by using centered kernel polarization criterion. This criterion is an extension of kernel polarization and a simplified style of centered kernel alignment. Compared with formulated kernel polarization criterion, the proposed criterion has a defined geometrical significance, and it can locate the global optimal point with less influence of threshold selection. Furthermore, the approximate criterion function can be proved to have a determined global minimum point by adopting the Euler–Maclaurin formula under weaker conditions. In addition, taking the preservation of within-class local structure into account, we present an evaluation criterion named local multiclass centered kernel polarization in multiclass classification scenario. Comparative experiments are conducted on some benchmark examples with three Gaussian kernel based learning methods and the results well demonstrate the effectiveness and efficiency of the proposed quality measures.

1. Introduction

Kernel-based learning methods, such as support vector machine (SVM) [27], kernel principal component analysis (KPCA) [23]. and kernel linear discriminant analysis (KLDA) [19], provide high performance for solving a wide range of different problems in machine learning community. These methods work by mapping the input data into a high-dimensional feature space and then build linear algorithms in the feature space to implement nonlinear counterparts in the input space. The key to the success of kernel methods is the incorporation of the "kernel trick" which computes a kernel function as the inner product between each pair of points in the feature space without computing their images directly. Thus these kernel methods combine the advantages of linear and non-linear classifiers in terms of efficient training time, elegant compatibility with high-dimensional data.

It is reasonable to hope that the mapped classes in the feature space possess a better linear separability compared with that obtained in the input space for a classification task. However, the classification performance of kernel methods can be even worse than that of their linear counterparts in the original input space when the kernel functions are not well chosen [33]. So whether kernel methods behave well largely depends on their adopted kernel functions. It is well known that the choice of the kernel function is a challenging problem.

[*] Corresponding author. Tel.: +86 351 7017566; fax: +86 351 7018176.
E-mail addresses: luckywalter@163.com (M. Tian), wjwang@sxu.edu.cn (W. Wang).

http://dx.doi.org/10.1016/j.ins.2015.06.010
0020-0255/© 2015 Elsevier Inc. All rights reserved.

In the literatures, kernel selection is usually tackled by cross validation and leave-one-out method. These two methods are data-independent, but they both suffer heavy computational complexity. To remedy this problem, in the context of SVM, some upper bounds on the generalization error have been proposed [6,7,11]. Of these bounds, the radius-margin bound is most commonly used in practice. However, it still requires the whole learning process for evaluation like cross validation and leave-one-out method.

In order to obtain a better computation efficiency, many universal data-dependent kernel evaluation measures have been derived by optimizing the measure of data separation in the feature space. Based on Fisher discrimination criteria, Refs. [28,31,33] proposed different approaches to optimize the kernel parameters. However, the use of Fisher criteria tends to give undesired results if samples in some class form several separative clusters, especially for the case of multimodally distributed data [25]. By using the measure called "alignment", Ref. [9], for the first time proposed a kernel target alignment criterion to optimize the kernel function. This criterion can measure the similarity between two kernel matrices or the degree of agreement between a kernel and a given target function. Beginning with kernel target alignment, many measurement criteria have been derived for kernel selection, such as kernel polarization [2], feature space based kernel matrix evaluation measure [20] and local kernel polarization [29].

Basically, kernel target alignment is the most commonly used efficient kernel measure criterion. Some researchers found that the sensitivity of kernel target alignment in case of uneven class distribution will drop drastically [14]. Refs. [8,20] showed a kernel matrix with a low kernel alignment value may have a very good performance. This means having a very high kernel target alignment is only a sufficient condition, but not a necessary condition, for kernel function to be a good one for a given task [20]. Therefore, Ref. [8] proposed a new criterion, centered kernel alignment, to modify kernel target alignment by adopting the notion of centering in the feature space. In addition to giving a simple concentration bound for centered kernel alignment, the existence of good predictors for kernel with high alignment both for classification and for regression has been shown. By this criterion, a steepest ascent approach based on forward stagewise additive method has been presented for multiple kernel learning. The approach achieves good performance across a variety of real-world data sets without discretizing the space of base kernels [1]. Multiple kernel clustering based on centered kernel alignment has also been proposed [18].

Recently Ref. [34] proposed an efficient Gaussian kernel optimization method, which works by maximizing the formulated kernel target alignment (in fact, it is the formulated kernel polarization). The contribution of this work lies in obtaining a differentiable objective function having a determined minimum point. More remarkably, the approximate analytical solution of the formulated criterion can be obtained by using the Euler–Maclaurin formula. Furthermore, the optimization has been solved with high computation efficiency by using a Newton-based algorithm with a unique starting point to locate the best local minimum compared with the searching procedure in [28]. However, the objective function curve of alignment value depending on the kernel parameter on some data sets monotonically increases very slowly when the parameter is greater than the optimal parameter, and then the selected parameter may be dependent on the threshold values of the search algorithm. Besides, the proof of having a determined global minimum point for approximate formulated criterion was obtained under strong constraint conditions.

We propose an effective surrogate measure based on kernel polarization, namely, centered kernel polarization. The approximate criterion function can be proved to have a determined global minimum point for two-class pattern classification tasks under weaker constraint conditions than those in [34]. We note that the proposed criterion is similar to the Hilbert–Schmidt Independence Criterion (HSIC) [13], which is a practical criterion for independence test in the context of independent component analysis (ICA). In this paper, we mainly tune the Gaussian kernel parameter on the basis of centered kernel polarization, and study the analytic properties and geometrical significance of the proposed criterion as well. In addition, based on the works in [29,30], we put forward a new multiclass evaluation criterion named local multiclass centered kernel polarization by taking the local structure preservation into account.

The rest of this paper is organized as follows. Section 2 gives a short description of some properties of Gaussian kernel and three criteria, namely, kernel target alignment, kernel polarization and centered kernel alignment. Section 3 discusses the continuous differentiability of the formulated centered kernel polarization and proves that the approximate criterion function has a determined global minimum point. In addition, by exploring the relationship among the centered kernel polarization criterion and two other off the shelf kernel evaluation measures, the geometric meaning of the proposed criterion is revealed. Section 4 describes the proposed local multiclass evaluation criterion in detail. Experimental results are presented in Section 5.

In this paper, all analyses are based on Gaussian kernel function. In the following, K denotes a kernel function, capital-case boldface symbols are used for matrices, $< \cdot, \cdot >$ denotes a dot product, and $< \cdot, \cdot >_F$ denotes a Frobenius inner product.

2. Preliminaries

2.1. The Gaussian kernel for classification

Recently, the use of kernel functions in machine learning and data mining community has received considerable attention. The kinds of kernel K we will be interested in are such that for all samples x_i and x_j, where $x_i, x_j \in \mathcal{X} \subset \mathbb{R}^m$, and $\mathcal{X}$ is the input space:

$$K(x_i, x_j) =< \Phi(x_i), \Phi(x_j) >, K : \mathcal{X} \times \mathcal{X} \to \mathbb{R},$$

where Φ denotes feature map that maps the points to a high dimensional feature space $\mathcal{F}$, i.e. $\Phi : \mathcal{X} \to \mathcal{F}$. In practice, the kernel K is usually defined directly, thus implicitly defining the map Φ and the feature space $\mathcal{F}$. Mercer has shown that a necessary and sufficient condition for a symmetric function $K(x_i, x_j)$ to be a kernel is that it be positive definite.

The well-known kernels include linear kernel, polynomial kernel, Gaussian kernel and so on. In the following we focus on an isotropic Gaussian kernel function which is popular and widely used in various applications. The Gaussian kernel is defined as:

$$K(x_i, x_j) = exp\left(-\frac{||x_i - x_j||^2}{\sigma^2}\right),$$

where σ is the width parameter. As previously discussed [31], the determination of a proper σ is of crucial importance to the Gaussian kernel based method's performance. Different values of σ map the data into different feature spaces. As $||\Phi(x)||^2 = 1$, the Gaussian kernel maps all the input vectors in the feature space with the same length 1. For every pair of different patterns $x_i \neq x_j$, we have $K(x_i, x_j) \in (0, 1)$. When $\sigma \to 0, K(x_i, x_j) \to 0$ holds, namely all the different training data points will be mapped to the orthogonal unit vectors in the feature space, therefore all the training patterns can be separated correctly. However, for any new sample, this classifier may not give right recognition due to "over-fitting" training. On the other hand, when $\sigma \to +\infty, K(x_i, x_j) \to 1$ and all the training data points are regarded as one point. As the result, the classifier cannot recognize any new sample due to "lack of fitting" training. Thus, neither too big σ nor too small σ is suit for a classification target.

2.2. Review of three kinds of alignment criteria

Based on previous results in [2,8,9], we shall propose a new class separability measure criterion. In this section, we first present the notions of kernel target alignment, kernel polarization and centered kernel alignment that will be useful in our quest.

Let $\mathcal{D}$ be the distribution according to which training and test points are drawn. Given a finite sample set $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$ drawn according to $\mathcal{D}$ and the corresponding label vector $y = (y_1, y_2, \ldots, y_n)^T$ where $y_i \in \{+1, -1\}, 1 \leqslant i \leqslant n$. The kernel matrix $\mathbf{K} \in \mathbb{R}^{n \times n}$ is defined by $\mathbf{K}_{ij} = K(x_i, x_j)$ and the label matrix is defined by $\mathbf{Y} = yy^T$ (an ideal target matrix). Kernel target alignment can measure the similarity of the kernel matrix and the target matrix. Mathematically, it is defined as [9]:

$$A(\mathbf{K}, \mathbf{Y}) = \frac{< \mathbf{K}, \mathbf{Y} >_F}{||\mathbf{K}||_F ||\mathbf{Y}||_F}. \tag{1}$$

It has been proved that kernel target alignment is sharply concentrated around its expected value, and the error rate using the kernel with a high empirical alignment can be limited to a certain amount [9]. These theoretical results together with computational efficiency facilitate the application of kernel target alignment in many learning tasks.

Drown from physics, Ref. [2] proposed the kernel polarization criterion:

$$P(\mathbf{K}, \mathbf{Y}) =< \mathbf{K}, \mathbf{Y} >_F \tag{2}$$

Clearly, kernel polarization criterion is a simplified style of kernel target alignment, as it ignores the denominator of kernel target alignment.

Built upon kernel target alignment, centered kernel alignment leverages the notion of centering in the feature space [8]. Following Cortes et al., we define centered kernel alignment between **K** and **Y** by

$$CA(\mathbf{K}, \mathbf{Y}) = \frac{< \mathbf{K}_c, \mathbf{Y}_c >_F}{||\mathbf{K}_c||_F ||\mathbf{Y}_c||_F}. \tag{3}$$

The centered kernel $\mathbf{K}_c$ associated to $\mathbf{K}$ is defined for all $x_i, x_j \in \mathcal{X}$ by

$$\mathbf{K}_c(x_i, x_j) =< \Phi(x_i) - \overline{\Phi}, \Phi(x_j) - \overline{\Phi} >$$

where $\overline{\Phi} = \frac{1}{n}\sum_{i=1}^{n}\Phi(x_i)$. Thus, the centered kernel matrix $\mathbf{K}_c$ is defined for all $i, j \in [1, n]$ by

$$[\mathbf{K}_c]_{ij} = \mathbf{K}_{ij} - \frac{1}{n}\sum_{i=1}^{n}\mathbf{K}_{ij} - \frac{1}{n}\sum_{j=1}^{n}\mathbf{K}_{ij} + \frac{1}{n^2}\sum_{i,j=1}^{n}\mathbf{K}_{ij}.$$

Note that the conception of the centered kernel matrix described above is exactly the KPCA transform [24]. Cortes et al. also gave a compact style of $\mathbf{K}_c$ as: $\mathbf{K}_c = \mathbf{HKH}$, where $\mathbf{H}$ is the so called centering matrix defined by $\mathbf{H} = \mathbf{I}_{n \times n} - \frac{ee^T}{n}$. $\mathbf{I}_{n \times n}$ denotes the $n \times n$ identity matrix and $e = (1, 1, \ldots, 1)^T \in \mathbb{R}^n$.

Centered kernel alignment criterion can be used to measure how well a centered kernel matrix aligns to a centered target matrix. The difference between the centered alignment of two kernel matrices and the alignment of the corresponding kernel functions can be bounded by a term in $O(1/\sqrt{n})$ [8].

Generally speaking, centered kernel polarization, kernel target alignment and kernel polarization have the same property that: the alignment value will increase when we keep within-class data pairs close and between-class data pairs apart in the feature space [8,20,29,34]. The only difference between kernel target alignment and centered kernel alignment, by comparing Eq. (1) with Eq. (3), is the centering operation on kernel matrices. However, this operation is crucial. Without centering operation, kernel target alignment would suffer from ill-conditioned problems. The main reason for the ill-conditioned problems is that the elements of kernel matrix may have almost the same values when the origin is far away from the convex hull of the samples in the feature space [18]. The centering operation cancels mismatches of the mean responses between the two kernels, and effectively cancels the effects caused by the imbalanced class distribution [1]. Thus the centering operation makes it possible that centered kernel alignment has better performance than kernel target alignment.

3. The proposed criterion for binary-class classification

3.1. The "centered kernel polarization" criterion

By importing the notion of centering in the feature space, we propose the centered kernel polarization criterion, which is defined by

$$P_c(\mathbf{K},\mathbf{Y}) = <\mathbf{K}_c,\mathbf{Y}_c>_F. \tag{4}$$

The difference between centered kernel polarization and centered kernel alignment, by comparing Eq. (3) with Eq. (4), is the normalization transformation on kernel matrices. In some sense, centered kernel polarization is like cross-covariance operator between the random variables $\mathbf{K}_c(x_i,x_j)$ and $\mathbf{Y}_c(x_i,x_j)$, while centered kernel alignment can be seen as a standard correlation coefficient between them. Seen from Eq. (4), the proposed criterion may lead to an unconstrained optimization problem. [2] stated that the corresponding feature space geometry can assure the kernel optimization problem is well posed when adopting a bound kernel. Thus, It is easier that the optimization problem can be implemented with the omission of the normalization transformation in the centered kernel polarization.

By definition of $\mathbf{H}$, we can have $\mathbf{H}^2=\mathbf{H}$. Note that, when $\mathbf{U},\mathbf{V}$ are two Gram matrices, $<\mathbf{U},\mathbf{V}>_F=\sum_{ij}\mathbf{U}_{ij}\mathbf{V}_{ij}=tr(\mathbf{UV})$. Hence centered kernel polarization criterion can be rewritten as follows:

$$P_c(\mathbf{K},\mathbf{Y}) = Tr(\mathbf{K}_c\mathbf{Y}_c) = Tr(\mathbf{K}\mathbf{Y}_c) = <\mathbf{K},\mathbf{Y}_c>_F.$$

Centered kernel polarization has two important properties: the concentration bound of the form $|\frac{P_c(\mathbf{K},\mathbf{Y})}{n^2}-E(K_c(K_\mathbf{Y})_c)|\leqslant O(\frac{1}{\sqrt{n}})$, where $K_\mathbf{Y}(x_i,x_j)=y_iy_j$, and the existence of good predictor with high accuracy in the presence of high alignment. Following [8], it is easy to prove the correctness of the claim above.

The centered label matrix $\mathbf{Y}_c$ can be written as $\mathbf{Y}_c=\mathbf{HYH}$. Without loss of generality, let $y_1=\ldots=y_{n_+}=1$ and $y_{n_++1}=\ldots=y_{n_++n_-}=-1$, where n_+ examples belong to class $+1, n_-$ examples belong to class -1, respectively, and $n_++n_-=n$. $\mathbf{Y}_c$ can be expanded and written more explicitly as follows:

$$\mathbf{Y}_c=\begin{pmatrix} 4\frac{n_-^2}{n^2}e_{n_+\times n_+} & -4\frac{n_+n_-}{n^2}e_{n_+\times n_-} \\ -4\frac{n_+n_-}{n^2}e_{n_-\times n_+} & 4\frac{n_+^2}{n^2}e_{n_-\times n_-}\end{pmatrix}, \tag{5}$$

where $e_{l\times l}$ denotes the $l\times l$ matrix whose elements are all equal to unity. The detailed derivation process about the above equation can be found in [33].

For Gaussian kernel function, we express $P_c(\mathbf{K},\mathbf{Y})$ by using Eq. (5), then

$$\begin{aligned}<\mathbf{K},\mathbf{Y}_c>_F&=\sum_{i,j}\mathbf{K}_{ij}[\mathbf{Y}_c]_{ij}\\ &=\frac{4}{n^2}\left[\sum_{y_i=y_j=1,i\neq j}n_-^2exp\left(-\frac{||x_i-x_j||^2}{\sigma^2}\right)+\sum_{y_i=y_j=-1,i\neq j}n_+^2exp\left(-\frac{||x_i-x_j||^2}{\sigma^2}\right)-\sum_{y_i\neq y_j}2n_+n_-exp\left(-\frac{||x_i-x_j||^2}{\sigma^2}\right)\right]+\frac{4n_+n_-}{n}.\end{aligned} \tag{6}$$

The optimal σ is obtained by maximizing the criterion $<\mathbf{K},\mathbf{Y}_c>_F$. Observe that when the number of positive samples equals to the one of negative samples, i.e. $n_+=n_-$, the optimization problem here is identical to that in [34]. Thus, centered kernel polarization is a general extension of kernel polarization.

For convenience, we transform the maximization problem to a minimization problem. By omitting for clarity the constant term and the constant coefficient, $<\mathbf{K},\mathbf{Y}_c>_F$ can be simplified as

$$S(\sigma)=\sum_{y_i\neq y_j}\frac{2n_+n_-}{n^2}exp\left(-\frac{||x_i-x_j||^2}{\sigma^2}\right)-\sum_{y_i=y_j=1,i\neq j}\frac{n_-^2}{n^2}exp\left(-\frac{||x_i-x_j||^2}{\sigma^2}\right)-\sum_{y_i=y_j=-1,i\neq j}\frac{n_+^2}{n^2}exp\left(-\frac{||x_i-x_j||^2}{\sigma^2}\right). \tag{7}$$

And $S(\sigma)$ is the formulated centered kernel polarization criterion. Thus, we can now consider instead:

$$\sigma_{opt} = arg\min_{\sigma} S(\sigma).$$

Obviously, the formulated criterion $S(\sigma)$ is continuously differentiable, the first and the second derivatives of $S(\sigma)$ with respect to σ can be derived easily as follow:

$$\frac{\partial S(\sigma)}{\partial \sigma} = \frac{2}{n^2\sigma^3}\left[\sum_{y_i \neq y_j} 2n_+n_-||x_i - x_j||^2 exp\left(-\frac{||x_i - x_j||^2}{\sigma^2}\right) - \sum_{y_i=y_j=1,i\neq j} n_-^2||x_i - x_j||^2 exp\left(-\frac{||x_i - x_j||^2}{\sigma^2}\right) - \sum_{y_i=y_j=-1,i\neq j} n_+^2||x_i - x_j||^2 exp\left(-\frac{||x_i - x_j||^2}{\sigma^2}\right)\right],$$

$$\frac{\partial^2 S(\sigma)}{\partial \sigma^2} = \frac{4}{n^2\sigma^6}\left[\sum_{y_i \neq y_j} 2n_+n_-||x_i - x_j||^4 exp\left(-\frac{||x_i - x_j||^2}{\sigma^2}\right) - \sum_{y_i=y_j=1,i\neq j} n_-^2||x_i - x_j||^4 exp\left(-\frac{||x_i - x_j||^2}{\sigma^2}\right) - \sum_{y_i=y_j=-1,i\neq j} n_+^2||x_i - x_j||^4 exp\left(-\frac{||x_i - x_j||^2}{\sigma^2}\right)\right] - \frac{6}{n^2\sigma^4}\left[\sum_{y_i \neq y_j} 2n_+n_-||x_i - x_j||^2 exp\left(-\frac{||x_i - x_j||^2}{\sigma^2}\right) - \sum_{y_i=y_j=1,i\neq j} n_-^2||x_i - x_j||^2 exp\left(-\frac{||x_i - x_j||^2}{\sigma^2}\right) - \sum_{y_i=y_j=-1,i\neq j} n_+^2||x_i - x_j||^2 exp\left(-\frac{||x_i - x_j||^2}{\sigma^2}\right)\right].$$

Having a determined global minimum point is a thrilling property for a differentiable separability measure. The property can bring about the decrease of computational cost. Ref. [34] examined the local and global extremal properties of the approximate formulate kernel polarization. However, on four out of thirteen data sets in [34], the objective function curves depending on the kernel parameter were found to monotonically increase very slowly when σ is greater than σ_{opt}. So the choice of stopping criterion of Algorithm 1 (quasi-Newton algorithm in [34]) has a significant impact on the overall kernel method performance. When the stopping criteria are not properly set, the good classification performance of kernel methods cannot be keep. Naturally, we try to find an optimization criterion which has a determined global minimum point with less influence of threshold selection for most datasets.

We note that [22] has stated that the initially normalization of each kernel is necessary. When comparing two kernels with widely different norms, the operation can greatly reduce the caused unfair bias. In the following, we will discuss the analytic properties of the proposed criterion. And we will find the removal of the normalization may cause some instability on kernel performance, while the operation makes it possible for the target function to have some good characteristics.

3.2. Evaluating the local global extremal properties of $S(\sigma)$

For the sake of convenience, we assume $T_{ij} = ||x_i - x_j||^2$, thus, $S(\sigma)$ can be expressed as

$$S(\sigma) = \sum_{y_i \neq y_j} \frac{2n_+n_-}{n^2} exp\left(-\frac{T_{ij}}{\sigma^2}\right) - \sum_{y_i=y_j=1,i\neq j} \frac{n_-^2}{n^2} exp\left(-\frac{T_{ij}}{\sigma^2}\right) - \sum_{y_i=y_j=-1,i\neq j} \frac{n_+^2}{n^2} exp\left(-\frac{T_{ij}}{\sigma^2}\right).$$

The Euler–Maclaurin formula is a very powerful tool in studying the finite series summation problem [16]. It can be expressed as following:

$$\sum_{t=u}^{v} f(t) = \int_u^v f(t)dt + \frac{f(u) + f(v)}{2} + \sum_{k=1}^{+\infty} \frac{B_{2k}}{(2k)!}(f^{(2k-1)}(v) - f^{(2k-1)}(u)), \tag{8}$$

where $f^{(2k-1)}(t), t \in [u, v], k \geqslant 1$ are functions of bounded variation and B_{2k} are the Bernoulli numbers. Let $f(t) = e^t$, we express the first two items of the last expression on the right, and obtain an approximation expression of Eq. (8):

$$\sum_{t=u}^{v} e^t \approx \int_u^v e^t dt + \frac{e^u + e^v}{2} + \frac{1}{12}[e^v - e^u] = \frac{19}{12}e^v - \frac{7}{12}e^u. \tag{9}$$

We define some auxiliary variables as following: $A = max\{T_{ij}|, y_i = y_j = 1, i \neq j\}, B = min\{T_{ij}|, y_i = y_j = 1, i \neq j\}, C = max\{T_{ij}|, y_i = y_j = -1, i \neq j\}, D = min\{T_{ij}|, y_i = y_j = -1, i \neq j\}, E = max\{T_{ij}|, y_i \neq y_j\}, F = min\{T_{ij}|, y_i \neq y_j\}$. Now, these variables, A, B, C, D, E and F, generally have the following relationships for separable data sets:

$$\begin{cases} \textit{the values of } A,\ C \textit{ and } E \textit{ are about } 2 \textit{ orders of magnitude} \\ \textit{larger than those of } B,\ D \textit{ and } F,\ \textit{respectively}, \\ \textit{the value of } max\{A,C,E\} \textit{ is close to the value of } min\{A,C,E\}. \end{cases} \tag{10}$$

The relation restrictions shown in (10) will greatly simplify the analysis below. A binary classification data set may satisfy (10) under the assumption that both the classes are generated from underlying multivariate Normal distributions of common covariance matrix but different means and each class is expressed by a single cluster. Besides this, if each class is not expressed by a single cluster and two classes share the similar cluster structure, the data set may also satisfy (10). In [34], another additional constraint is imposed on separable data sets. The constraint is the minimum distance of within-class sample pairs must less than that of between-class sample pairs in input space. However, many datasets do not meet the constraint condition (see Table 2). Compared with the constraint relationships provided in [34], there are less constraint corresponding to B, D and F here.

The approximation of $S(\sigma)$ can be written as

$$\begin{aligned} S(\sigma) &= \frac{2n_+n_-}{n^2}\sum_{-E/\sigma^2}^{-F/\sigma^2} e^t - \frac{n_-^2}{n^2}\sum_{-A/\sigma^2}^{-B/\sigma^2} e^t - \frac{n_+^2}{n^2}\sum_{-C/\sigma^2}^{-D/\sigma^2} e^t \\ &\approx \frac{1}{12n^2}\left[38n_+n_-e^{-\frac{F}{\sigma^2}} - 19n_-^2e^{-\frac{B}{\sigma^2}} - 19n_+^2e^{-\frac{D}{\sigma^2}} + 7n_-^2e^{-\frac{A}{\sigma^2}} + 7n_+^2e^{-\frac{C}{\sigma^2}} - 14n_+n_-e^{-\frac{E}{\sigma^2}}\right]. \end{aligned}$$

For ease of discussion, let

$$\widehat{S}(\sigma) = \frac{1}{12n^2}\left[38n_+n_-e^{-\frac{F}{\sigma^2}} - 19n_-^2e^{-\frac{B}{\sigma^2}} - 19n_+^2e^{-\frac{D}{\sigma^2}} + 7n_-^2e^{-\frac{A}{\sigma^2}} + 7n_+^2e^{-\frac{C}{\sigma^2}} - 14n_+n_-e^{-\frac{E}{\sigma^2}}\right]. \tag{11}$$

$\widehat{S}(\sigma)$ is the approximate centered kernel polarization criterion function. Hence, the derivative of $\widehat{S}(\sigma)$ with respect to σ is

$$\frac{\partial\widehat{S}(\sigma)}{\partial\sigma} = \frac{1}{6n^2\sigma^3}\left[38n_+n_-Fe^{-\frac{F}{\sigma^2}} - 19n_-^2Be^{-\frac{B}{\sigma^2}} - 19n_+^2De^{-\frac{D}{\sigma^2}} + 7n_-^2Ae^{-\frac{A}{\sigma^2}} + 7n_+^2Ce^{-\frac{C}{\sigma^2}} - 14n_+n_-Ee^{-\frac{E}{\sigma^2}}\right].$$

To minimize $\widehat{S}(\sigma)$, we consider $\frac{\partial\widehat{S(\sigma)}}{\partial\sigma} = 0$. Too large σ reduces the kernel to a constant function, making it impossible to learn any non-trivial classifier. Then we only take the following equation in consideration:

$$\frac{1}{n^2}\left[38n_+n_-Fe^{-\frac{F}{\sigma^2}} - 19n_-^2Be^{-\frac{B}{\sigma^2}} - 19n_+^2De^{-\frac{D}{\sigma^2}} + 7n_-^2Ae^{-\frac{A}{\sigma^2}} + 7n_+^2Ce^{-\frac{C}{\sigma^2}} - 14n_+n_-Ee^{-\frac{E}{\sigma^2}}\right] = 0.$$

According to the values of B, D and F, there are six cases to consider: $B \geqslant D \geqslant F, B \geqslant F \geqslant D, D \geqslant B \geqslant F, D \geqslant F \geqslant B, F \geqslant D \geqslant B, F \geqslant B \geqslant D$. Without loss of generality, we give a detail proof of the first case: $B \geqslant D \geqslant F$. The equation above can be formulated as

$$\frac{1}{n^2}\left[38n_+n_-F - 19n_-^2Be^{\frac{F-B}{\sigma^2}} - 19n_+^2De^{\frac{F-D}{\sigma^2}} + 7n_-^2Ae^{\frac{F-A}{\sigma^2}} + 7n_+^2Ce^{\frac{F-C}{\sigma^2}} - 14n_+n_-Ee^{\frac{F-E}{\sigma^2}}\right] = 0.$$

It can be further expressed as

$$\begin{aligned} &38\tfrac{n_+n_-F}{n^2} - 19\tfrac{n_-^2B+n_+^2D}{n^2}e^{\frac{F-B}{\sigma^2}} \\ &+7\tfrac{n_-^2A+n_+^2C-2n_+n_-E}{n^2}e^{\frac{F-A}{\sigma^2}} \\ &+19\tfrac{n_+^2D}{n^2}\left(e^{\frac{F-B}{\sigma^2}} - e^{\frac{F-D}{\sigma^2}}\right) \\ &-7\tfrac{n_+^2C}{n^2}\left(e^{\frac{F-A}{\sigma^2}} - e^{\frac{F-C}{\sigma^2}}\right) + 14\tfrac{n_+n_-E}{n^2}\left(e^{\frac{F-A}{\sigma^2}} - e^{\frac{F-E}{\sigma^2}}\right) = 0. \end{aligned} \tag{12}$$

Since $A, B, C, D, E,$ and F satisfy (10), the following results can easily be observed: the values $\frac{n_+^2C}{n^2}\left(e^{\frac{F-A}{\sigma^2}} - e^{\frac{F-C}{\sigma^2}}\right), \frac{n_+n_-E}{n^2}\left(e^{\frac{F-A}{\sigma^2}} - e^{\frac{F-E}{\sigma^2}}\right)$, and $\frac{n_-^2A+n_+^2C-2n_+n_-E}{n^2}e^{\frac{F-A}{\sigma^2}}$ are all close to 0. Concisely, let

$$\begin{aligned} \delta_1 &\triangleq \tfrac{n_+^2C}{n^2}\left(e^{\frac{F-A}{\sigma^2}} - e^{\frac{F-C}{\sigma^2}}\right), \\ \delta_2 &\triangleq \tfrac{n_+n_-E}{n^2}\left(e^{\frac{F-A}{\sigma^2}} - e^{\frac{F-E}{\sigma^2}}\right), \\ \delta_3 &\triangleq \tfrac{n_-^2A+n_+^2C-2n_+n_-E}{n^2}e^{\frac{F-A}{\sigma^2}}. \end{aligned}$$

Taken the assumption $B \geqslant D \geqslant F$ in consideration, $|e^{\frac{F-B}{\sigma^2}} - e^{\frac{F-D}{\sigma^2}}|$ is a small variable less than 1. Now let $\eta \triangleq \frac{n_+^2D}{n^2}\left(e^{\frac{F-B}{\sigma^2}} - e^{\frac{F-D}{\sigma^2}}\right)$. With the notation of D, the variable η is a small number close to 0. Plugging these four auxiliary variables $\delta_1, \delta_2, \delta_3$, and η into Eq. (12), then

$$\frac{38n_+n_-F}{n^2}-\frac{19(n_-^2B+n_+^2D)}{n^2}e^{\frac{F-B}{\sigma^2}}+7\delta_3+19\eta-7\delta_1+14\delta_2=0.$$

For clarity, let $\varepsilon=7\delta_3+19\eta-7\delta_1+14\delta_2$, and we have

$$\sigma_0=\sqrt{\frac{F-B}{ln\frac{38n_+n_-F+n^2\varepsilon}{19(n_-^2B+n_+^2D)}}}.$$

The local extremal points of $\widehat{S}(\sigma)$ are $\sigma=0,+\infty$, and σ_0. In the following, we compare the three values of $\widehat{S}(\sigma)$ with $\sigma=0,+\infty$, and σ_0. According to (10), the following equations can easily be obtained based on Eq. (11):

$$\lim_{\sigma\to+\infty}\widehat{S}(\sigma)=-\tfrac{(n_+-n_-)^2}{n^2},$$
$$\lim_{\sigma\to0^+}\widehat{S}(\sigma)=0.$$

Without loss of generality, let $A=min(A,C,E)$. In view of the assumed condition $B\geqslant D\geqslant F$, $\widehat{S}(\sigma_0)$ can be approximate to the next inequations:

$$\begin{aligned}&\widehat{S}(\sigma_0)<\tfrac{1}{12n^2}\left[-19(n_+-n_-)^2e^{-\frac{F}{\sigma^2}}+7(n_+-n_-)^2e^{-\frac{A}{\sigma^2}}\right]\\&=\tfrac{1}{12n^2}\left[7(n_+-n_-)^2(e^{-\frac{A}{\sigma^2}}-e^{-\frac{F}{\sigma^2}})-12(n_+-n_-)^2e^{-\frac{F}{\sigma^2}}\right]\\&<-\tfrac{(n_+-n_-)^2}{n^2},\end{aligned}$$

where the last inequality is easily hold by the inequalities $e^{-\frac{A}{\sigma^2}}<e^{-\frac{F}{\sigma^2}}$ and $e^{-\frac{F}{\sigma^2}}<1$.

In case of $D\geqslant B\geqslant F$, we will have the same result. Based on the analyses above, we obtain the next theorem.

Theorem 1. *Suppose that (10) holds.* $\sigma=\sqrt{\frac{F-B}{ln\frac{38n_+n_-F+n^2\varepsilon}{19(n_-^2B+n_+^2D)}}}$ *is the determined global minimum point of* $\widehat{S}(\sigma)$ *in the case of* $min\{B,D\}\geqslant F$.

Using a similar calculation we can obtain the determined global minimum points of the objective functions corresponding to other four cases: $min\{D,F\}\geqslant B$ *and* $min\{B,F\}\geqslant D$*. Here we only list the results, and the derivation processes are not shown in this section due to space limitation.*

Theorem 2. *Suppose that (10) holds.* $\sigma=\sqrt{\frac{B-F}{ln\frac{19n_-^2B-n^2\varepsilon}{38n_+n_-F-19n_+^2D}}}\left(\text{resp. }\sqrt{\frac{D-F}{ln\frac{19n_+^2D-n^2\varepsilon}{38n_+n_-F-19n_-^2B}}}\right)$ *is the determined global minimum point of* $\widehat{S}(\sigma)$ *in the case of* $min\{D,F\}\geqslant B$ *(resp.* $min\{B,F\}\geqslant D$*), where* $\varepsilon=7\frac{n_-^2A}{n^2}e^{\frac{B-A}{\sigma^2}}+7\frac{n_+^2C}{n^2}e^{\frac{B-C}{\sigma^2}}-14\frac{n_+n_-E}{n^2}e^{\frac{B-E}{\sigma^2}}-19\frac{n_+^2D}{n^2}\left(e^{\frac{B-D}{\sigma^2}}-e^{\frac{B-F}{\sigma^2}}\right)$ $\left(\text{resp. }7\frac{n_-^2A}{n^2}e^{\frac{D-A}{\sigma^2}}+7\frac{n_+^2C}{n^2}e^{\frac{D-C}{\sigma^2}}-14\frac{n_+n_-E}{n^2}e^{\frac{D-E}{\sigma^2}}+19\frac{n_-^2B}{n^2}(e^{\frac{D-F}{\sigma^2}}-e^{\frac{D-B}{\sigma^2}})\right)$.

3.3. Connections with two other kernel criteria

Based on the kernel target alignment [9], a modified kernel target alignment criterion was proposed for uneven data by substituting the target matrix $\mathbf{Y}$ with $\mathbf{Y}_u$ [14]. $\mathbf{Y}_u$ denotes the modified target matrix and $\mathbf{Y}_u=y_uy_u^T$, where $y_u=([y_u]_1,[y_u]_2,\ldots,[y_u]_n)^T$ and

$$[y_u]_i=\begin{cases}\frac{1}{n_+}, & y_i=1,\\-\frac{1}{n_-}, & y_i=-1.\end{cases}$$

The numerator of the modified kernel target alignment criterion can be calculated as

$$<\mathbf{K},\mathbf{Y}_u>_F=\sum_{i,j}\mathbf{K}_{ij}[\mathbf{Y}_u]_{ij}=\sum_{y_i=y_j=1}\frac{1}{n_+^2}exp\left(-\frac{||x_i-x_j||^2}{\sigma^2}\right)+\sum_{y_i=y_j=-1}\frac{1}{n_-^2}exp\left(-\frac{||x_i-x_j||^2}{\sigma^2}\right)-\sum_{y_i\neq y_j}\frac{2}{n_+n_-}exp\left(-\frac{||x_i-x_j||^2}{\sigma^2}\right).$$

The inter-cluster distance in the feature space also was used as an index to choose proper kernel parameters [32]. Experiment results showed that δ_{4F}, an index denotes the distance between two class means in the feature space, can indicate the class separation robustly. The δ_{4F} can be written as

$$\delta_{4F}(\mathcal{X}_+,\mathcal{X}_-)=d(\hat{x}_+,\hat{x}_-)=\sqrt{\frac{\sum_{y_i=y_j=1}exp\left(-\frac{||x_i-x_j||^2}{\sigma^2}\right)}{n_+^2}+\frac{\sum_{y_i=y_j=-1}exp\left(-\frac{||x_i-x_j||^2}{\sigma^2}\right)}{n_-^2}-\frac{2\sum_{y_i\neq y_j}exp\left(-\frac{||x_i-x_j||^2}{\sigma^2}\right)}{n_+n_-}},$$

where $\hat{x}_+$ and $\hat{x}_-$ are the class means of the mapped positive class $\mathcal{X}_+$ and the mapped negative class $\mathcal{X}_-$. Compared with Eq. (6), we have

$$< \mathbf{K}, \mathbf{Y}_u >_F = \delta^2_{4F}(\mathcal{X}_+, \mathcal{X}_-) = \frac{n^2}{4n_+^2 n_-^2} < \mathbf{K}, \mathbf{Y}_c >_F.$$

Based on above discussion, centered kernel polarization is similar to the criterion proposed in [14], but without the normalization. Furthermore, the optimization problem that maximizing the empirical estimate of centered kernel polarization is totally equivalent to the maximization with the measure δ_{4F} introduced in [32] although the derivation approaches this criterion in a completely different way. In other words, the proposed alignment maximization problem, from a geometrical point of view, can be regarded as the maximization of the distance between the class mean locations. We note that Ref. [32] chose the index δ_{4F} as optimization criterion only based on the testing performances in experiments. The discussion on the local and global extremal properties of objective function in Section 3.2 exactly makes up for the theory deficiency of [32].

4. The “local multiclass centered kernel polarization” criterion

The multiclass classification problems are usually divided into binary classification sub-problems. And several methods, such as one-versus-rest method [21] and one-versus-one method [15], have been proposed. For SVM, Ref. [21] proposed that one-versus-rest method and one-versus-one method usually have no significant difference in classification accuracy when the underlying binary classifiers are well tuned. As one-versus-rest method has lower computation cost and conceptual simplicity, our discussion here are based on one-versus-rest method.

Given a problem with L classes, one-versus-rest method constructs L binary classifiers, in which each classifier is trained to separate one class from the other classes. One always optimizes kernel parameter by using the sum function of corresponding index values or alignment values of all pairs of classes [32]. Recently, the multiclass kernel polarization criterion for Gaussian kernel function was proposed [30]. The criterion can be measured as

$$P_m(\mathbf{K}, \mathbf{Y}_m) =< \mathbf{K}, \mathbf{Y}_m >_F \tag{13}$$

where

$$(\mathbf{Y}_m)_{ij} = \begin{cases} 1, & y_i = y_j = 1, 2, \ldots, L, \\ -1, & y_i \neq y_j, \end{cases}$$

and L denotes the class numbers.

Multiclass kernel polarization discards the restriction of binary classification, and it can encode the multiclass information and directly address the multiclass problems simultaneously. Detailed description of multiclass kernel polarization can be found in [30]. Compared with previous methods, multiclass kernel polarization has better computation efficiency. The optimal parameter can be obtained by maximizing $P_m(\mathbf{K}, \mathbf{Y}_m)$, i.e.

$$\sigma_{opt} = arg \max_{\sigma} P_m(\mathbf{K}, \mathbf{Y}_m).$$

Clearly, kernel polarization can be seen as a special case of multiclass kernel polarization when all samples belong to two classes.

We note that the optimal alignment happens when $\mathbf{K}_{ij} = 1, y_i = y_j = 1, 2, \ldots, L$, which implies that all examples of the same class are mapped into the same point in the feature space. In other words, within-class structure penalizes the alignment value. Similar problems can be seen in kernel target alignment and kernel polarization [20,29]. For binary-class classification problems, Refs. [20,29] have put forward different schemes to remedy this problem for kernel target alignment and kernel polarization, respectively. To the best of our knowledge, there was no previous work to remedy this problem for multiclass classification problems.

In this section, we present a feasible multiclass evaluation criterion, which combines the advantages of previous criteria. Let $\mathbf{Y}_{lm}$ be an aggregation target matrix, i.e., the $n \times n$ matrix with the element $(\mathbf{Y}_{lm})_{ij}$ being the aggregation degree between x_i and x_j,

$$(\mathbf{Y}_{lm})_{ij} = \begin{cases} exp(-t||x_i - x_j||^2), & y_i = y_j = 1, 2, \ldots, L, \\ -1, & y_i \neq y_j, \end{cases} \tag{14}$$

where $t \in \mathbb{R}$ and $t \in (0, +\infty)$. The new measure criterion, namely local multiclass centered kernel polarization, is defined as follows:

$$L_c(\mathbf{K}, \mathbf{Y}_{lm}) =< \mathbf{K}_c, (\mathbf{Y}_{lm})_c >_F. \tag{15}$$

And the optimal parameter is obtained by

$$\sigma_{opt} = arg \max_{\sigma} L_c(\mathbf{K}, \mathbf{Y}_{lm})$$

Table 1
The specification of selected data sets.

Data set	Number of features	Number of samples	Number of classes
Sonar	60	208	2
Heart	13	270	2
Liverdisorder	6	345	2
Ionosphere	34	351	2
Wdbc	30	569	2
Australian	14	690	2
Ringnorm	20	1000	2
Twonorm	20	1000	2
German	24	1000	2
Splice	60	1000	2
Yeast	8	1136	2
A1a	123	1605	2
Mushrooms	112	2031	2
W1a	300	2477	2
Phoneme	5	5404	2
Iris	4	150	3
Wine	13	178	3
Glass	9	214	6
Vowel	10	990	11
Satimage	36	2000	6
IJK	16	2241	3
Segment	19	2310	7
Waveform	21	6000	3

Compared with multiclass kernel polarization (Eq. (12)), the target values for within-class pairs in Eq. (13) are weighted by the aggregation degree. This means that all samples of the same class are not forced to map into the same point any more. The farther away the pairs in the same class are, the smaller target values are and the less influences on the value of this measure are. The principle of this method is same as that of local kernel polarization [29]. Local kernel polarization adopts an affinity matrix weighting the kernel values for within-class pairs, while the new presented criterion adopts an aggregation matrix weighting the target values for the consistency of text.

Identical to local kernel polarization, far-apart points in the same class are not made close as the multiclass kernel polarization does. Thus the local structure of the data of the same class tends to be preserved. At the same time, points in different classes are also made apart. We note that it adds a hyperparameter t which must be tuned. For convenience, in the following experiments, the value of t is fixed, $i.e., t = \frac{1}{2}$.

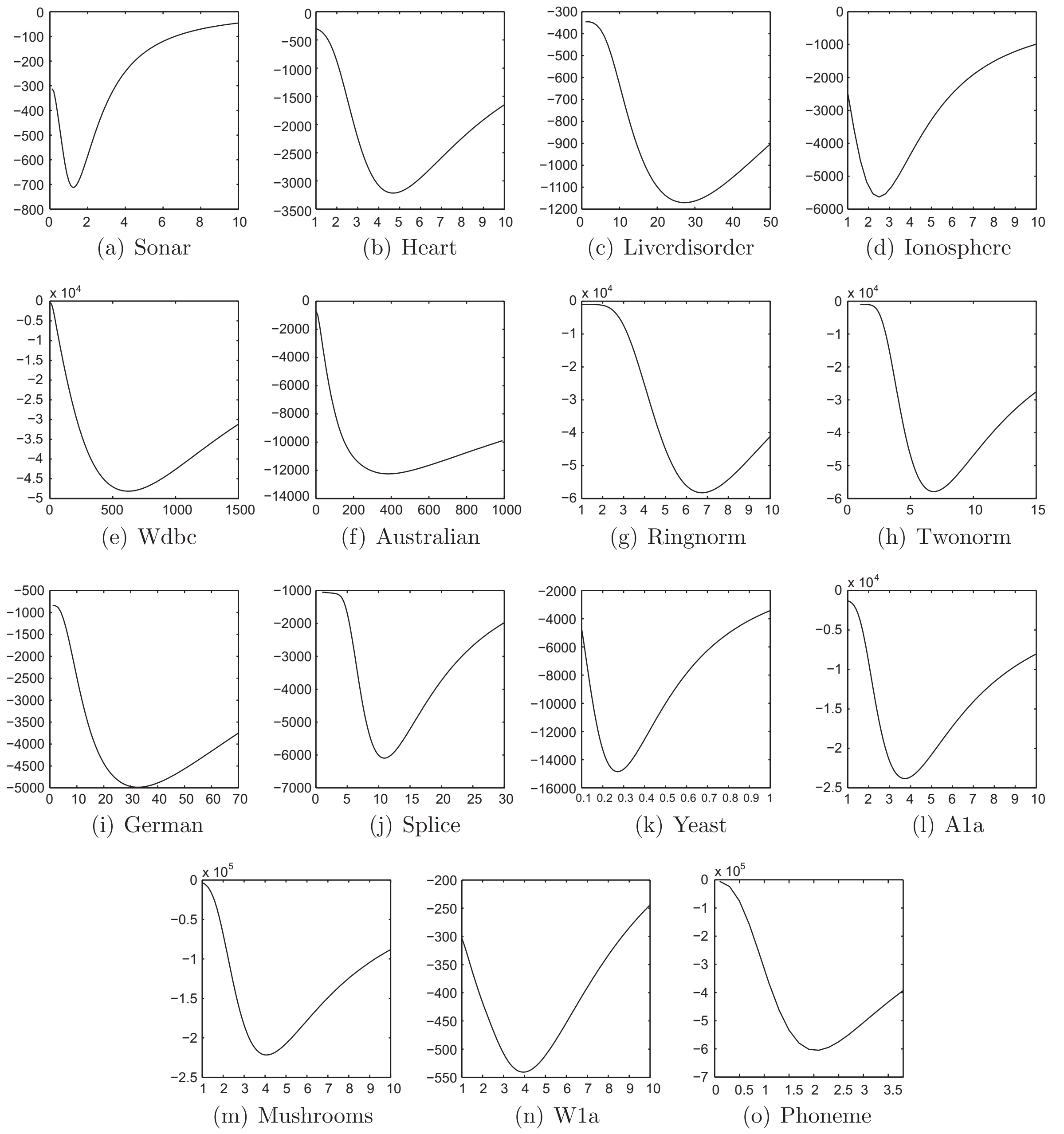
5. Experimental results

5.1. Experimental setup

In experiments, we selected 23 popular data sets in which 15 data sets for the binary-class classification and others for the multiclass classification [5,10,12]. The specifications of these data sets are listed in Table 1. All the benchmark examples considered in Table 1 are small databases ranging from 150 to 6000, in feature number from 4 to 300, and in class number from 2 to 11. The Yeast dataset is took as the same as [29], and it is a binary classification dataset between 'CYT' and 'NUC &

Table 2
The variables $A - F$ for fifteen data sets.

Data set	A	B	C	D	E	F	Satisfy (10)?
Sonar	10.8569	0.0392	12.4571	0.0532	11.5366	0.2573	Yes
Heart	82.5611	0.0717	98.9115	0.0609	96.9528	1.2065	Yes
Liverdisorder	4.3783e+04	12.25	8.5878e+04	12.000	8.7305e+04	6.0000	Yes
Ionosphere	59.8720	0.0100	95.0000	0.0382	73.9753	0.2178	Yes
Wdbc	1.8927e+07	153.2981	1.7832e+06	14.5616	2.2459e+07	119.2926	Yes
Australian	1.0000e+10	6.8081	3.4803e+07	0.3906	1.0002e+10	6.3397	Yes
Ringnorm	114.6558	6.4984	476.7291	21.1897	310.7860	17.1179	No
Twonorm	114.9259	5.9452	134.0748	6.6845	171.0332	7.7641	No
German	2.7609e+04	1.0000	3.4262e+04	5.0000	3.5004e+04	6.0000	Yes
Splice	303.0000	49.0000	278.0000	1.0000	300.0000	52.0000	No
Yeast	1.5245	0.0001	1.4411	0.0002	1.8918	0.0003	Yes
A1a	26.0000	1.0000	28.0000	2.0000	28.0000	2.0000	No
Mushrooms	36.0000	2.0000	34.0000	2.0000	36.0000	4.0000	No
W1a	55.0000	1.0000	126.0000	1.0000	112.0000	1.0000	No
Phoneme	27.6965	1.0000e-06	38.1036	1.0000e-06	34.8110	5.3700e-04	Yes



(a) Sonar (b) Heart (c) Liverdisorder (d) Ionosphere

(e) Wdbc (f) Australian (g) Ringnorm (h) Twonorm

(i) German (j) Splice (k) Yeast (l) A1a

(m) Mushrooms (n) W1a (o) Phoneme

Fig. 1. $S(\sigma)$ on different σ values on fifteen datasets for bivariate classification.

MIT'. The IJK dataset is a subset problem of the letter data set corresponding to the classes I, J and K. For each data set, we partition it into a training set and a test set by stratified sampling: 50% of the data set serves as training set and the left 50% as test set. For multiclass problem, the original training data are normalized to have zero mean and unit variance.

We denote centered kernel polarization, kernel polarization, centered kernel alignment and cross validation method, as 'CKP', 'KP', 'CKA', and 'CV', respectively.

The purpose of these experiments is mainly to provide empirical proof for the following two hypotheses: (H1) For binary-class classification problems, the formulated centered kernel polarization criterion has a determined global minimum point. And this criterion can lead competitive performance compared with KP, CKA and CV method on test accuracy, Cohen's kappa statistic [4] and time efficiency. (H2) For multiclass classification problems, the proposed multiclass criterion is a universal one for Gaussian kernel selection.

Three Gaussian kernel based method, kernel direct discriminant analysis (KDDA), generalized discriminant analysis (GDA) and SVM, are applied to each data set. KDDA and GDA are two popular kernel-based feature extraction algorithms.

Table 3
Comparison of KDDA + KNN using the optimized σ obtained by KP, CKP, CKA and CV.

Data set	Accuracy (%)			
	KP	CKP	CKA	CV
Sonar	75.58 ± 0.0724	76.35 ± 0.0947	**82.60** ± 0.0629	**83.94** ± 0.0542
Heart	**77.04** ± 0.0296	76.56 ± 0.0242	76.67 ± 0.0297	**80.59** ± 0.0253
Liverdisorder	56.82 ± 0.0321	**59.19** ± 0.0392	58.50 ± 0.0551	**63.12** ± 0.0234
Ionosphere	80.97 ± 0.0385	85.29 ± 0.0242	**85.80** ± 0.0253	**86.93** ± 0.0328
Wdbc	**86.42** ± 0.0119	**86.04** ± 0.0133	85.89 ± 0.0201	82.07 ± 0.0808
Australian	**58.32** ± 0.0294	58.09 ± 0.0353	57.85 ± 0.0251	**61.60** ± 0.0341
Ringnorm	97.64 ± 0.0069	**97.68** ± 0.0059	97.59 ± 0.0058	**97.84** ± 0.0067
Twonorm	95.96 ± 0.0080	95.94 ± 0.0068	**96.14** ± 0.0073	**96.64** ± 0.0055
German	59.14 ± 0.0169	**67.00** ± 0.0204	63.50 ± 0.0427	**70.24** ± 0.0131
Splice	**76.20** ± 0.0253	74.40 ± 0.0425	72.93 ± 0.0335	**75.47** ± 0.0298
Yeast	58.10 ± 0.0213	**59.05** ± 0.0094	58.54 ± 0.0217	**65.94** ± 0.0121
A1a	74.72 ± 0.0132	**75.47** ± 0.0067	74.76 ± 0.0125	**78.01** ± 0.0104
Mushrooms	92.88 ± 0.0110	**93.15** ± 0.0106	91.39 ± 0.0063	**99.89** ± 0.0021
W1a	89.19 ± 0.0180	**89.58** ± 0.0140	89.44 ± 0.0137	**90.11** ± 0.0121
Phoneme	70.54 ± 0.0063	**72.63** ± 0.0117	71.94 ± 0.0072	**80.66** ± 0.0043
W-T-L	2-13-0	–	3-11-1	0-3-12
	Kappa (%)			
Sonar	51.19 ± 0.1424	52.73 ± 0.1883	**65.37** ± 0.1239	**67.87** ± 0.1091
Heart	**53.27** ± 0.0567	52.34 ± 0.4244	52.20 ± 0.0625	**60.40** ± 0.0509
Liverdisorder	11.69 ± 0.0616	**16.40** ± 0.0698	14.11 ± 0.0816	**20.63** ± 0.0598
Ionosphere	58.63 ± 0.0782	68.23 ± 0.0545	**68.99** ± 0.0541	**71.33** ± 0.0783
Wdbc	**70.53** ± 0.260	**69.81** ± 0.0287	69.50 ± 0.0437	58.59 ± 0.2044
Australian	**15.32** ± 0.0591	15.18 ± 0.0734	14.68 ± 0.0505	**23.05** ± 0.0882
Ringnorm	95.27 ± 0.0117	**95.36** ± 0.0139	95.15 ± 0.0119	**95.68** ± 0.0135
Twonorm	91.91 ± 0.0161	91.87 ± 0.0136	**92.27** ± 0.0147	**93.27** ± 0.0109
German	**4.95** ±0.0463	**7.38** ± 0.018	2.67 ± 0.0311	1.44 ± 0.0045
Splice	**52.49** ± 0.0495	48.92 ± 0.0792	46.08 ± 0.0642	**51.08** ± 0.0587
Yeast	13.87 ± 0.0401	**15.29** ± 0.0173	14.40 ± 0.0494	**22.78** ± 0.0421
A1a	**31.60** ± 0.0315	**32.94** ± 0.0174	31.56 ± 0.0290	30.20 ± 0.0562
Mushrooms	85.52 ± 0.0226	**86.27** ± 0.0214	82.75 ± 0.0125	**99.78** ± 0.0042
W1a	72.07 ± 0.0618	**77.75** ± 0.0271	**77.16** ± 0.0284	67.82 ± 0.0434
Phoneme	28.86 ± 0.0134	**34.26** ± 0.0270	32.55 ± 0.0147	**50.22** ± 0.0440
W-T-L	2-12-1	–	3-11-1	2-9-4
	Training time (s)			
Sonar	0.3086 ± 0.0508	**0.2827** ± 0.0087	0.3514 ± 0.0568	5.2321 ± 0.1247
Heart	**0.4042** ± 0.0134	0.4222 ± 0.0189	0.6527 ± 0.2050	8.0488 ± 0.1881
Liverdisorder	2.2038 ± 0.1179	**1.0303** ± 0.0636	1.4290 ± 0.0703	20.3363 ± 0.0598
Ionosphere	**0.7314** ± 0.0373	0.8835 ± 0.0214	0.9497 ± 0.0284	16.4965 ± 0.4424
Wdbc	**2.9051** ± 0.3840	3.3507 ± 0.3199	4.8683 ± 0.3922	39.7987 ± 0.2883
Australian	6.5856 ± 0.4119	**5.1851** ± 0.6938	7.2073 ± 1.6030	52.5616 ± 0.6351
Ringnorm	**7.1405** ± 0.3056	8.0671 ± 0.2768	9.7672 ± 0.1526	129.0443 ± 1.7096
Twonorm	**5.6451** ± 0.1829	6.4238 ± 0.2274	9.9196 ± 0.4202	129.0104 ± 0.5327
German	24.3285 ± 6.5150	**8.2415** ± 1.2194	16.6576 ± 6.7238	130.6449 ± 0.1917
Splice	9.7992 ± 0.5823	**7.7126** ± 0.3486	8.7010 ± 0.4616	139.3011 ± 0.1349
Yeast	29.8939 ± 1.5521	**9.8473** ± 0.1724	11.7641 ± 0.6443	110.6901 ± 1.0258
A1a	83.6737 ± 2.3772	**19.3573** ± 0.7923	75.8727 ± 2.2965	341.9878 ± 3.2165
Mushrooms	**26.6821** ± 0.0226	30.3166 ± 1.0439	116.2290 ± 4.7105	601.6218 ± 3.2430
W1a	612.6792 ± 24.2753	**134.4480** ± 15.7834	321.9465 ± 11.8732	1047.9342 ± 4.7235
Phoneme	982.2161 ± 20.8242	**257.9333** ± 8.6783	521.2031 ± 11.3165	2450.5577 ± 3.6971
W-T-L	9-5-1	–	12-3-0	15-0-0

They can be viewed as the implementation of the well-known LDA method in the kernel feature space. Detailed descriptions of KDDA and GDA can be found in [3,17]. In experiments KDDA and GDA are followed by a K-Nearest Neighbor (K-NN) classifier [26] to perform the recognition. The parameter K in K-NN is set as 1 and the regularization parameter C in SVM is set with the values $\{2^{-3}, 2^{-2}, \ldots, 2^{2}, 2^{3}\}$. For cross validation method, we use the 10-fold cross validation to find the best σ within the given set $\{2^{-5}, 2^{-4}, \ldots, 2^{4}, 2^{5}\}$. The recognition performance is evaluated by the best classification accuracy rate and the Cohens kappa meter obtained by average 10 randomly independent performances.

One particular line search algorithm is adopted here. First comes the bracketing phase by advance-retreat method with the starting point $\sigma_0 = d_0$, which denotes the average pairwise Euclidean distance between the training samples. Default step $h = 1$ and $L_{stop} = 1.0e - 6$. The approximate solution to the objective function is subsequently found by using *fminbnd* function of Matlab.

Table 4
Comparison of GDA + KNN using the optimized σ obtained by KP, CKP, CKA and CV.

Data set	Accuracy (%)			
	KP	CKP	CKA	CV
Sonar	87.98 ± 0.0321	87.98 ± 0.0351	**88.17** ± 0.0475	**89.81** ± 0.0348
Heart	75.55 ± 0.0382	**75.93** ± 0.0352	74.37 ± 0.0363	**80.59** ± 0.0214
Liverdisorder	60.52 ± 0.0443	61.33 ± 0.0273	**61.85** ± 0.0288	**63.93** ± 0.0233
Ionosphere	90.40 ± 0.0262	91.93 ± 0.0167	**93.16** ± 0.0114	**93.18** ± 0.0114
Wdbc	90.07 ± 0.0199	89.89 ± 0.0132	**90.49** ± 0.0157	**91.93** ± 0.0118
Australian	60.55 ± 0.0239	**64.03** ± 0.0236	60.43 ± 0.0354	**66.24** ± 0.0234
Ringnorm	95.88 ± 0.0104	95.86 ± 0.0108	**96.62** ± 0.0086	**97.36** ± 0.0052
Twonorm	**94.26** ± 0.0082	94.22 ± 0.0094	93.88 ± 0.0064	**96.40** ± 0.0056
German	68.18 ± 0.0207	**69.70** ± 0.0154	67.96 ± 0.0097	**71.30** ± 0.0111
Splice	78.60 ± 0.2540	**78.93** ± 0.0236	78.47 ± 0.0272	**78.93** ± 0.0209
Yeast	59.36 ± 0.0128	**62.08** ± 0.0128	61.39 ± 0.0149	**65.55** ± 0.0148
A1a	78.56 ± 0.0137	**78.75** ± 0.0165	77.95 ± 0.0151	**80.37** ± 0.0124
Mushrooms	**99.92** ± 0.0016	**99.92** ± 0.0016	**99.92** ± 0.0016	**99.92** ± 0.0016
W1a	96.08 ± 0.0220	**96.35** ± 0.0236	96.12 ± 0.0239	**96.88** ± 0.0217
Phoneme	71.33 ± 0.0069	**79.88** ± 0.0072	79.14 ± 0.0060	**86.31** ± 0.0065
W-T-L	2-13-0	–	3-11-1	0-3-12
	Kappa (%)			
Sonar	75.95 ± 0.0629	75.95 ± 0.0691	**76.33** ± 0.0935	**79.57** ± 0.0689
Heart	50.59 ± 0.0759	**51.37** ± 0.0701	48.05 ± 0.0667	**60.55** ± 0.0467
Liverdisorder	19.16 ± 0.0943	**21.04** ± 0.0606	17.62 ± 0.0588	**25.53** ± 0.0598
Ionosphere	78.71 ± 0.0578	**82.65** ± 0.0342	81.49 ± 0.0541	**85.40** ± 0.0255
Wdbc	78.54 ± 0.436	78.20 ± 0.0301	**79.34** ± 0.0355	**82.60** ± 0.2256
Australian	20.52 ± 0.0506	**27.33** ± 0.0481	19.86 ± 0.0710	**32.45** ± 0.0452
Ringnorm	91.77 ± 0.0207	91.73 ± 0.0216	**93.24** ± 0.0171	**97.72** ± 0.0105
Twonorm	**88.50** ± 0.0164	88.42 ± 0.0187	87.75 ± 0.0147	**92.79** ± 0.0112
German	**24.12** ± 0.0457	**27.56** ± 0.0381	16.48 ± 0.0947	17.72 ± 0.1567
Splice	57.24 ± 0.0504	**57.95** ± 0.0466	56.98 ± 0.0548	**57.87** ± 0.0422
Yeast	16.40 ± 0.0286	**21.60** ± 0.0283	20.38 ± 0.0306	**22.49** ± 0.0443
A1a	47.31 ± 0.0331	**48.22** ± 0.0296	47.41 ± 0.0429	**49.43** ±0.0356
Mushrooms	**99.84** ± 0.0033	**99.84** ± 0.0033	**99.84** ± 0.0033	**99.84** ± 0.0033
W1a	38.29 ± 0.1575	**46.17** ± 0.0489	26.99 ± 0.2561	**47.23** ± 0.1718
Phoneme	30.95 ± 0.0162	**51.44** ± 0.0196	49.79 ± 0.0135	**66.67** ± 0.0157
W-T-L	3-12-0	–	3-11-1	0-8-7
	Training time (s)			
Sonar	0.3240 ± 0.0788	**0.2810** ± 0.0099	0.3445 ± 0.0421	3.4088 ± 0.0726
Heart	**0.4019** ± 0.0142	0.4062 ± 0.0156	0.6706 ± 0.2805	6.3427 ± 0.1151
Liverdisorder	2.3794 ± 1.3249	**1.0200** ± 0.0577	1.4660 ± 0.0696	10.8155 ± 0.2700
Ionosphere	**0.7348** ± 0.0640	0.8767 ± 0.0225	0.9450 ± 0.0322	9.7742 ± 0.3106
Wdbc	**2.9490** ± 0.4003	3.3387 ± 0.3040	5.0233 ± 0.3910	38.2645 ± 1.3036
Australian	7.0546 ± 0.3014	**5.3624** ± 0.8108	7.6061 ± 1.7542	48.1913 ± 3.4289
Ringnorm	**7.6908** ± 0.1917	8.4821 ± 0.1922	10.3493 ± 0.1672	123.8194 ± 3.9560
Twonorm	**5.7686** ± 0.1360	6.4664 ± 0.2195	9.8936 ± 0.4141	117.7365 ± 0.6655
German	25.9599 ± 7.7235	**8.0211** ± 1.2957	17.1543 ± 7.2221	125.3052 ± 6.3187
Splice	28.8380 ± 3.6742	**27.7168** ± 3.0496	32.6939 ± 3.0717	130.8596 ± 4.2892
Yeast	28.5955 ± 0.9887	**8.8009** ± 0.1577	10.7370 ± 0.5659	72.0570 ± 0.4623
A1a	82.2572 ± 1.3668	**19.5349** ± 0.6760	75.8738 ± 3.4381	410.5585 ± 3.8702
Mushrooms	**26.4290** ± 0.9476	29.5990 ± 1.0690	118.5738 ± 2.6458	740.9735 ± 18.7118
W1a	618.7053 ± 23.6417	**138.0171** ± 15.6908	325.0727 ± 15.2622	1391.0913 ± 26.4485
Phoneme	1051.2190 ± 30.9665	**405.8839** ± 11.6355	496.6071 ± 11.1552	4156.6233 ± 45.4712
W-T-L	8-4-3	–	12-3-0	15-0-0

All the compared calculations are carried out by using Matlab (V2008, the Mathworks, Inc.) and the SVM toolbox developed by Gunn from http://www.isis.ecs.soton.ac.uk/isystems/kernel/. All experiments are conducted on a PC with 2.93 GHz CPU and 2G RAM.

5.2. Comparisons for binary-class problems

Table 2 shows the variables A, B, C, D, E and F in Section 3.2 of the fifteen binary-class data sets. And Table 2 also shows that these variables on 9 out of 15 datasets satisfy the relationships shown in (10). Fig. 1 depicts $S(\sigma)$ on different σ of the fifteen datasets. It is clear from Fig. 1 that $S(\sigma)$ has the determined global minimum point for all fifteen data sets. We note that the objective function on every data set monotonically increases quickly when the value of σ is greater than the optimal value.

Table 5
Comparison of SVM using the optimized σ obtained by KP, CKP, CKA and CV.

Data set	Accuracy (%)			
	KP	CKP	CKA	CV
Sonar	88.27 ± 0.0323	88.46 ± 0.0327	**88.85** ± 0.0321	**89.91** ± 0.0327
Heart	**85.41** ± 0.0197	85.19 ± 0.0218	82.30 ± 0.0870	**86.18** ± 0.0627
Liverdisorder	59.54 ± 0.0736	**68.38** ± 0.0237	63.93 ± 0.0275	**69.56** ± 0.0162
Ionosphere	90.00 ± 0.0178	**90.23** ± 0.0153	89.89 ± 0.0131	**91.18** ± 0.0294
Wdbc	93.65 ± 0.0118	**93.68** ± 0.0105	92.84 ± 0.0118	**94.29** ± 0.0131
Australian	**70.17** ± 0.0227	69.80 ± 0.0246	**70.06** ± 0.0251	67.24 ± 0.0159
Ringnorm	96.54 ± 0.0068	**96.56** ± 0.0065	**96.64** ± 0.0066	96.11 ± 0.0081
Twonorm	96.90 ± 0.0044	96.92 ± 0.0043	**97.06** ± 0.0042	**97.38** ± 0.0034
German	70.08 ± 0.0213	**75.46** ± 0.0167	71.08 ± 0.0305	**75.40** ± 0.0718
Splice	82.18 ± 0.1183	**85.88** ± 0.0083	**85.86** ± 0.0079	85.54 ± 0.0116
Yeast	59.28 ± 0.0133	67.36 ± 0.0107	**67.41** ± 0.0147	**68.89** ± 0.0459
A1a	78.55 ± 0.0072	**82.82** ± 0.0075	79.55 ± 0.0072	**82.48** ± 0.0858
Mushrooms	**99.92** ± 0.0016	**99.92** ± 0.0016	**99.92** ± 0.0016	**99.92** ± 0.0016
W1a	97.13 ± 0.0038	**97.88** ±0.0039	97.49 ± 0.0057	**97.70** ± 0.0261
Phoneme	70.24 ± 0.1223	**85.31** ± 0.4222	82.09 ± 0.3254	**88.16** ± 0.0252
W-T-L	6-9-0	–	6-8-1	0-4-11
	Kappa (%)			
Sonar	76.48 ± 0.0641	76.85 ± 0.0649	**77.58** ± 0.0639	**79.82** ± 0.0217
Heart	**69.75** ± 0.0451	69.30 ± 0.0479	62.07 ± 0.2201	**71.21** ± 0.0684
Liverdisorder	11.23 ± 0.1809	**34.74** ± 0.0424	25.67 ± 0.0554	**26.67** ± 0.0372
Ionosphere	77.40 ± 0.0398	**77.91** ± 0.0342	77.07 ± 0.0294	**78.62** ± 0.0257
Wdbc	85.97 ± 0.0251	**86.03** ± 0.0223	84.13 ± 0.0270	**86.60** ± 0.0270
Australian	**38.45** ± 0.0501	36.85 ± 0.0545	37.67 ± 0.0577	**38.27** ± 0.0415
Ringnorm	93.08 ± 0.0135	93.12 ± 0.0130	**93.28** ± 0.0133	**93.22** ± 0.0127
Twonorm	93.79 ± 0.0089	93.83 ± 0.0086	**94.11** ± 0.0084	**94.24** ± 0.0431
German	21.64 ± 0.0487	**35.51** ± 0.0528	27.91 ± 0.1328	**32.89** ± 0.0337
Splice	64.68 ± 0.2279	**71.79** ± 0.0165	**71.76** ± 0.0155	70.89 ± 0.0165
Yeast	21.45 ± 0.0123	**31.96** ± 0.0191	31.95 ± 0.0259	**32.12** ± 0.0204
A1a	56.66 ± 0.0752	**72.82** ± 0.0175	69.44 ± 0.0672	**76.45** ± 0.0158
Mushrooms	**99.84** ± 0.0033	**99.84** ± 0.0033	**99.84** ± 0.0033	**99.84** ± 0.0033
W1a	77.16 ± 0.0035	**85.43** ± 0.0039	79.49 ± 0.0045	**88.90** ± 0.0061
Phoneme	72.52 ± 0.4571	**81.61** ± 0.1991	80.52 ± 0.0739	**86.16** ± 0.0382
W-T-L	7-8-0	–	3-11-1	0-6-9
	Training time (s)			
Sonar	0.5030 ± 0.1130	0.5710 ± 0.1811	**0.3926** ± 0.1433	29.7634 ± 0.1645
Heart	**0.6101** ± 0.0139	0.6159 ± 0.0230	1.0997 ± 0.3746	46.8753 ± 0.9234
Liverdisorder	2.9291 ± 0.9388	**1.6098** ± 0.1358	1.7327 ± 0.0070	64.8736 ± 1.0567
Ionosphere	**1.2020** ± 0.1174	1.2129 ± 0.0378	0.9001 ± 0.6232	84.9854 ± 2.9077
Wdbc	**8.4472** ± 1.2597	9.3869 ± 1.4757	11.8714 ± 1.3829	163.5435 ± 3.6543
Australian	17.0070 ± 2.8063	**16.3856** ± 0.9510	17.9830 ± 0.9842	248.7530 ± 4.7996
Ringnorm	**9.8213** ± 0.3204	10.7194 ± 0.2995	12.4638 ± 0.3672	676.2098 ± 8.5660
Twonorm	**8.1341** ± 0.2399	8.8409 ± 0.2095	17.4327 ± 0.1241	567.7665 ± 3.6543
German	31.2003 ± 4.0011	**26.9586** ± 4.1780	27.7083 ± 3.5367	652.6702 ± 6.3186
Splice	12.4352 ± 1.0405	**11.4572** ± 0.8385	13.6532 ± 1.2923	772.8125 ± 11.5496
Yeast	32.7664 ± 1.0457	**18.2827** ± 3.5188	42.2034 ± 1.4669	1278.5952 ± 14.7876
A1a	90.2697 ± 2.3220	**28.0686** ± 1.0399	89.3656 ± 1.4047	1743.7843 ± 23.8002
Mushrooms	**45.7491** ± 1.2988	48.8908 ± 1.3310	137.8919 ± 0.1291	3756.6428 ± 35.7118
W1a	605.0344 ± 17.8913	**155.8632** ± 12.7367	331.8414 ± 16.2334	1590.8452 ± 28.8695
Phoneme	3782.3433 ± 300.2365	**1293.2213** ± 33.8734	2446.6113 ± 399.1243	12523.7352 ± 112.5732
W-T-L	7-5-3	–	11-3-1	15-0-0

The formulated approximate criterion functions of 'CKP' and 'KP' have been proved to have a determined global minimum point. Thus, when the approximate criterion function is a sufficient approximation of the criterion function, without repeating the searching procedure with different starting points we can locate the best local minimum. For 'CKP' and 'KP', the run time is the sum of the time for looking for σ_{opt} and the recognition time spent on classification corresponding to the selected σ_{opt}. For the criterion 'CKA', we repeat the optimization procedure three times with different starting points $d_0/50, d_0, d_0 \times 50$, The final σ is thus the one with the largest alignment value. The run time of CKA is the time for looking for the 'best' σ from different start points and the recognition time spent on classification corresponding to the selected parameter.

The average classification accuracies with standard deviations, the average kappa statistic with standard deviations, and the mean running time with standard deviations over 10 trials of KDDA + KNN, GDA + KNN and SVM using the optimal σ obtained by KP, CKP, CKA and CV are summarized in Tables 3–5, respectively. The bold font denotes the best two recognition performance and the best time efficiency across the methods compared. On each data set, the test accuracy, Cohen's kappa statistics and the elapsed time are compared by using the paired t test according to the resampling scheme used. The significance level, α, is taken as 0.05 for all statistical tests. Win-tie-loss (W-T-L) summarizations based on t-test are attached at the bottoms of Tables 3–5. A win or a loss means that CKP is better or worse than other criterion on a data set. A tie means that both criteria have the same performance.

From Tables 3–5, we found the cross validation method always gives the best accuracy and the highest kappa in most cases, but it costs the most time, almost 10 times more than the other three criteria.

According to the prediction accuracy in Tables 3,4, we note that CKP obtains the best or next best accuracy on 9 out of 15 datasets. KP and CKA fall behind, giving better performance on no more than 6 out of 15 datasets in Tables 3,4. Meanwhile, CKP provides the best or next best accuracy on 10 out of 15 datasets in Table 5. KP and CKA present better performance on 3 datasets and 7 datasets in terms of test accuracy in Table 5, respectively. From Tables 3,4, CKP gives a comparable performance to KP and CKA on all but the Twonorm data set, and CKP is significantly better than KP and CKA on 2 and 3 out of 15 datasets, respectively. Table 5 shows that CKP is statistically significantly more accurate than the other two criteria on 6 datasets. There is no significant difference between CKP and KP on at least 8 out 15 datasets as well as CKA.

In terms of the Cohen's kappa statistic, the similar results as those in testing accuracy can be found in Tables 3 and 4. From Table 5, we note that CKP obtains the best or next best performance on 10 out of 15 datasets. CKP is more accurate (statistically significantly) than KP and CKA on 7 and 3 datasets, respectively. On most of the rest of the datasets, CKP obtains kappa comparable to KP as well as CKA.

For the training time, the KP criterion gives the shortest training time on 6 datasets in Tables 3–5. In these three tables, CKA only wins out on the Sonar dataset in Table 5. CKP gains the shortest running time on 9, 9 and 8 datasets in Tables 3–5, respectively. The W-T-L summarization shows that CKP has an obvious advantage compared with KP on 9 datasets, 8 datasets and 7 datasets in Tables 3–5, respectively. It implies a lot of time locating the optimal parameter may be saved since the proposed objective function curve has a trough. And on at least eleven datasets, CKP gives a comparable performance to KP as well as CKA in Tables 3–5, and part of the reasons may be the omission of the calculation of the denominator of CKA.

In a nutshell, CKP is a robust and efficient indication of the goodness of the Gaussian kernel compared with KP and CKA for binary-class problems.

5.3. Comparisons for multiclass problems

We compare the average accuracy, Cohen's kappa statistic and the time efficiency of the multiclass kernel polarization, the local centered multiclass kernel polarization and the multiclass centered kernel alignment. The Multiclass centered

Table 6
Comparison of KDDA + KNN using the optimized σ obtained by MKP, LMCKP, MCKA and CV.

Data set	Accuracy (%)			
	MKP	LMCKP	MCKA	CV
Iris	**94.00** ± 0.0398	93.33 ± 0.0251	92.67 ± 0.0261	**95.20** ± 0.0128
Wine	96.07 ± 0.0152	**96.29** ± 0.0167	95.51 ± 0.0129	**97.08** ± 0.0095
Glass	43.08 ± 0.0390	59.91 ± 0.0647	**62.58** ± 0.0388	**65.51** ± 0.0290
Vowel	72.44 ± 0.0282	**88.79** ± 0.0196	87.96 ± 0.0218	**93.48** ± 0.0991
Satimage	83.29 ± 0.0099	**83.76** ± 0.0081	83.74 ± 0.0058	**84.78** ± 0.0032
IJK	92.52 ± 0.0109	**93.08** ± 0.0088	81.41 ± 0.0191	**95.62** ± 0.0062
Segment	92.70 ± 0.0068	94.03 ± 0.0069	**95.33** ± 0.0083	**96.75** ± 0.0052
Waveform	**85.41** ± 0.0052	85.26 ± 0.0049	84.80 ± 0.0040	**86.21** ± 0.0039
W-T-L	4-4-0	–	2-5-1	0-1-7
	Kappa (%)			
Iris	**90.99** ± 0.0598	90.00 ± 0.0376	89.00 ± 0.0389	**92.79** ± 0.0192
Wine	93.99 ± 0.0234	**94.34** ± 0.0258	93.15 ± 0.0194	**95.53** ± 0.0148
Glass	28.16 ± 0.0345	45.15 ± 0.0886	**48.97** ± 0.0463	**52.84** ± 0.0403
Vowel	70.42 ± 0.0301	**87.98** ± 0.0209	87.09 ± 0.0233	**93.00** ± 0.0107
Satimage	79.44 ± 0.0124	**80.02** ± 0.0097	80.01 ± 0.0072	**81.29** ± 0.0037
IJK	88.77 ± 0.0163	**89.61** ± 0.0131	72.11 ± 0.0285	**93.43** ± 0.0093
Segment	91.48 ± 0.0079	93.03 ± 0.0081	**94.55** ± 0.0097	**96.20** ± 0.0060
Waveform	**78.11** ± 0.0078	77.89 ± 0.0073	77.19 ± 0.0060	**79.32** ± 0.0059
W-T-L	4-4-0	–	2-5-1	0-1-7
	Training time (s)			
Iris	0.1468 ± 0.0138	**0.1329** ± 0.0108	0.1687 ± 0.0251	3.6969 ± 0.3420
Wine	0.2631 ± 0.0336	**0.1697** ± 0.0195	0.2250 ± 0.0140	4.7448 ± 0.2455
Glass	0.4177 ± 0.0304	**0.2656** ± 0.0149	0.3443 ± 0.0238	6.7508 ± 0.3545
Vowel	8.2948 ± 0.3131	**5.8637** ± 0.4464	9.5296 ± 0.3620	105.9906 ± 0.6406
Satimage	34.5680 ± 1.0524	**25.6267** ± 0.9466	46.4461 ± 0.3009	379.7771 ± 1.3004
IJK	42.5332 ± 0.1432	**31.2267** ± 0.6550	52.0300 ± 3.1791	534.2616 ± 1.3926
Segment	47.0254 ± 1.3095	**35.1885** ± 1.3435	66.8769 ± 2.0267	578.8990 ± 1.1585
Waveform	**334.7509** ± 2.4818	381.0420 ± 11.1378	591.9844 ± 18.7130	1781.3229 ± 3.9356
W-T-L	6-1-1	–	8-0-0	8-0-0

Table 7
Comparison of GDA + KNN using the optimized σ obtained by MKP, LMCKP, MCKA and CV.

Data set	Accuracy (%)			
	MKP	LMCKP	MCKA	CV
Iris	93.87 ± 0.0210	94.00 ± 0.0211	**94.26** ± 0.0244	**96.40** ± 0.0167
Wine	**97.76** ± 0.0191	97.64 ± 0.0134	97.19 ± 0.0199	**99.10** ± 0.0116
Glass	63.46 ± 0.0544	**64.30** ± 0.0510	60.93 ± 0.0486	**68.60** ± 0.0402
Vowel	73.29 ± 0.0105	**90.08** ± 0.0190	88.16 ± 0.0198	**93.29** ± 0.0105
Satimage	83.38 ± 0.0096	**87.20** ± 0.0087	85.86 ± 0.0067	**87.42** ± 0.0089
IJK	91.81 ± 0.0055	**97.65** ± 0.0030	96.99 ± 0.0066	**97.83** ± 0.0044
Segment	**96.71** ± 0.0043	95.85 ± 0.0070	96.24 ± 0.0035	**97.22** ± 0.0026
Waveform	84.65 ± 0.0055	**85.95** ± 0.0039	84.73 ± 0.0053	**86.67** ± 0.0036
W-T-L	3-4-1	–	3-5-0	0-2-6
	Kappa (%)			
Iris	90.78 ± 0.0314	90.98 ± 0.0315	**91.38** ± 0.0364	**94.58** ± 0.0251
Wine	**96.58** ± 0.0291	96.41 ± 0.0203	95.73 ± 0.0304	**98.63** ± 0.0177
Glass	50.27 ± 0.0759	**50.72** ± 0.0689	43.80 ± 0.0829	**56.89** ± 0.0510
Vowel	70.81 ± 0.0113	**89.36** ± 0.0203	87.33 ± 0.0212	**92.81** ± 0.0113
Satimage	79.57 ± 0.0116	**84.26** ± 0.0107	82.63 ± 0.0082	**84.51** ± 0.0109
IJK	87.71 ± 0.0082	**96.47** ± 0.0045	95.49 ± 0.0099	**96.75** ± 0.0065
Segment	**96.16** ± 0.0049	95.16 ± 0.0082	95.62 ± 0.0041	**96.76** ± 0.0030
Waveform	84.65 ± 0.0055	**85.95** ± 0.0039	84.73 ± 0.0053	**86.67** ± 0.0036
W-T-L	3-4-1	–	3-5-0	0-2-6
	Training time (s)			
Iris	**0.1713** ± 0.0056	0.1905 ± 0.0109	0.3839 ± 0.0346	1.7028 ± 0.0939
Wine	0.3758 ± 0.0654	**0.2855** ± 0.0097	0.5558 ± 0.0243	2.3430 ± 0.0409
Glass	0.3714 ± 0.0439	**0.3067** ± 0.0214	0.3091 ± 0.0256	3.5148 ± 0.1955
Vowel	8.2494 ± 0.2636	**7.5083** ± 0.1406	9.5156 ± 0.3847	84.5411 ± 5.0346
Satimage	35.2815 ± 1.1129	**34.3526** ± 0.2059	45.8101 ± 0.3554	455.0871 ± 23.3761
IJK	**42.8871** ± 0.2815	48.4904 ± 1.5779	51.1759 ± 3.2792	629.4805 ± 18.6996
Segment	46.2836 ± 1.3052	**42.5259** ± 0.6966	64.0953 ± 2.120	723.9044 ± 49.6376
Waveform	**415.4985** ± 14.5660	550.6463 ± 35.6439	623.0710 ± 32.1195	3243.3522 ± 19.3611
W-T-L	3-3-2	–	6-2-0	8-0-0

kernel alignment is the multiclass extension of centered kernel alignment, without considering the local structure of within-class samples. It is defined by $CA_m(\mathbf{K},\mathbf{Y}) = \frac{<\mathbf{K}_c,[\mathbf{Y}_m]_c>_F}{||\mathbf{K}_c||_F||[\mathbf{Y}_m]_c||_F}$. The definition of $\mathbf{Y}_m$ can be seen in Eq. (12). In this section, MKP, LMKP, MCKA and CV denote the multiclass kernel polarization, the local multiclass kernel polarization, the multiclass centered kernel alignment and cross validation, respectively. The same algorithm is used to search the optimal parameter as described in Section 5.1. The starting points of MKP, LMCKP and MCKA are set in the same way as those of KP, CKP, and CKA in Section 5.2. The average performance results, in terms of the classification accuracy, Cohen's kappa and time efficiency, over 10 trials are summarized in Tables 6–8. The W-T-L summarizations based on *t*-test are attached at the bottoms of Tables 6–8. For all statistical tests, the significance level is taken as 0.05, which is as same as the default value of Section 5.2. A win or a loss means that MKP is better or worse than other criterion on a data set. A tie means that both criteria have the same performance.

Seen from Tables 6–8, the CV method always yields both the best accuracy and the best kappa in most cases with expensive computational cost compared with the other three criteria. The result is similar to that obtained in binary-class classification scenario.

From Tables 6–8, we observe that LMCKP obtains the best or next best test accuracy and kappa statistic on at least 4 datasets. Meanwhile, the MKP criterion obtains the best or next best accuracy on no more than 3 datasets in Tables 6–8. Except for the kappa statistic in Table 8, MCKA provides the best or next best accuracy on no more than 3 datasets in Tables 6–8. Table 6 shows that LMCKP is statistically significantly more accurate than MKP and MCKA on 4 and 2 datasets about test accuracy and kappa statistic, respectively. Table 7 shows that LMCKP is statistically significantly more accurate than MKP and MCKA on 3 out 8 datasets in terms of test accuracy and kappa statistic. And LMCKP works poorly on 1 dataset in Tables 6 and 7, respectively. In Table 8, LMCKP has a poorer performance compared with MKP on 1 dataset, and a better classification performance compared with MCKA on 2 dataset. Besides this, there is no significant difference between LMCKP and MKP as well as MCKA. Therefore the proposed criterion LMCKP can compete with MKP and MCKA in terms of correct recognition rate and Kappa. On most datasets, the difference of the accuracy of these three criteria is less than three per cent. However, for the Glass data set and the Vowel data set in Tables 6, it is obvious that MKP works poorly while LMCKP works considerably well: LMCKP outperforms MKP by more than ten per cent. And for the IJK dataset in Table 6, LMCKP outperforms MCKA by more than ten per cent too.

Seen form Tables 6–8, choosing the optimal parameter σ by measure criteria can save the recognition time with an acceptable test accuracy cost as expected. In Table 6, LMCKP has an obvious advantage over MKP and MCKA on 6 and 8

Table 8
Comparison of SVM using the optimized σ obtained by MKP, LMCKP, MCKA and CV

Data set	Accuracy (%)			
	MKP	LMCKP	MCKA	CV
Iris	**96.13** ± 0.0098	95.99 ± 0.0126	**96.00** ± 0.0177	94.53 ± 0.0222
Wine	96.18 ± 0.0213	**96.63** ± 0.0159	94.72 ± 0.0225	**97.19** ± 0.0143
Glass	65.70 ± 0.0335	**67.38** ± 0.0312	67.29 ± 0.0394	**67.42** ± 0.0298
Vowel	95.41 ± 0.0110	**96.02** ± 0.0106	**96.04** ± 0.0096	95.88 ± 0.0103
Satimage	**89.70** ± 0.0063	89.09 ± 0.0102	88.93 ± 0.0106	**89.30** ± 0.0098
IJK	97.87 ± 0.0044	**97.92** ± 0.0049	97.88 ± 0.0044	**97.94** ± 0.0039
Segment	**96.74** ± 0.0036	**96.74** ± 0.0037	96.68 ± 0.0039	96.50 ± 0.0033
Waveform	87.60 ± 0.0077	**87.94** ± 0.0080	**87.98** ± 0.0019	87.79 ± 0.0037
W-T-L	0-7-1	–	2-6-0	0-5-3
	Kappa (%)			
Iris	**94.18** ± 0.0149	**93.98** ± 0.0190	**93.98** ± 0.0286	91.79 ± 0.0332
Wine	94.14 ± 0.0327	**94.83** ± 0.0244	91.90 ± 0.0341	**95.71** ± 0.0221
Glass	51.94 ± 0.0518	**54.04** ± 0.0467	**53.81** ± 0.0563	51.63 ± 0.0381
Vowel	94.95 ± 0.0120	**95.62** ± 0.0117	**95.62** ± 0.0106	95.46 ± 0.0113
Satimage	**87.33** ± 0.0075	86.59 ± 0.0124	86.39 ± 0.0128	**86.84** ± 0.0119
IJK	96.80 ± 0.0065	**96.88** ± 0.0073	**96.81** ± 0.0066	96.60 ± 0.0058
Segment	**96.18** ± 0.0042	**96.19** ± 0.0044	96.12 ± 0.0046	95.92 ± 0.0039
Waveform	81.40 ± 0.0069	81.91 ± 0.0075	**81.98** ± 0.0074	**83.68** ± 0.0066
W-T-L	0-7-1	–	2-6-0	0-5-3
	Training time (s)			
Iris	**0.4157** ± 0.0127	0.4235 ± 0.0069	0.4410 ± 0.0185	8.6320 ± 0.0851
Wine	0.5497 ± 0.0240	**0.5065** ± 0.0145	0.5224 ± 0.0254	10.6354 ± 0.0563
Glass	1.3536 ± 0.0521	**1.2722** ± 0.0243	1.9833 ± 0.0151	2.2846 ± 0.0941
Vowel	67.7033 ± 0.802	**66.0308** ± 0.3900	68.6372 ± 0.8532	1721.6003 ± 0.5664
Satimage	**132.9432** ± 7.1129	134.3526 ± 8.9201	155.8101 ± 0.3554	2455.0871 ± 13.3761
IJK	**92.8871** ± 0.2815	98.4904 ± 1.5779	151.1339 ± 12.2792	1729.6605 ± 17.3422
Segment	**179.8733** ± 11.2001	292.5359 ± 9.6766	335.8269 ± 12.3120	9825.9044 ± 42.4466
Waveform	**276.4825** ± 8.5945	461.8040 ± 2.4190	572.3492 ± 17.8516	16538.5017 ± 93.8646
W-T-L	2-4-2	–	7-1-0	8-0-0

datasets out of 8 datasets, respectively. In Tables 7,8, LMCKP wins out MCKA on more than 6 datasets. And the differences between LMCKP and MKP is little in Tables 7, 8.

In one-versus-one setting, experimental results show that the proposed criterion can also perform well and the results are similar to those obtained in one-versus-rest setting. The experimental results are not shown in this paper due to space limitations. Therefore, LMCKP criterion is a better and robust indication of the Gaussian kernel compared with MKP criterion and MCKA criterion for multiclass problems.

6. Conclusion

Centered kernel polarization is put forward to determine kernel parameter for Gaussian kernel based methods. For binary-class classification problems the proposed criterion is differentiable, which means a series of efficient line search methods can be used to locate the optimal parameter. Compared with kernel polarization, the proposed criterion has an intuitive geometric meaning, and it can locate the optimal parameter with less dependence on the threshold of algorithm. The approximate objective function can be proved to have a determined global minimum point under some weaker constraint conditions. In addition, we present a new multiclass evaluation measure to encode the multiclass information and preserve the local structure of within-class data simultaneously. Experiment results show that two proposed criteria can achieve good overall classification performance and efficient training time.

In this paper, we focus on optimizing the isotropic Gaussian kernel function since it is a successfully used kernel function in various applications. Based on the good analytic properties of exponential function, a closed-form approximate solution to the objective function is proposed by adopting the Euler–Maclaurin formula. Because of this, the analysis method is only applicable for some particular kernels, such as the Gaussian kernel and the exponential radial basis function. How to evaluate the local or global extremal properties of the formulated centered kernel polarization to other kernels, such as linear kernel, polynomial kernel and wavelet kernel, is beyond our reach at present. Further investigation is needed to determine the applicability of the introduced criteria for other kernel functions. In addition, we will study the extensions and the applications of the proposed criteria in multiple kernel learning and feature selection for some larger data in our future work.

Acknowledgements

The work described in this paper was partially supported by the National Natural Science Foundation of China (No. 61273291, 71031006), Research Project Supported by Shanxi Scholarship Council of China (No. 2012-008), Open Project Foundation of Information Technology Research Base of Civil Aviation Administration of China (No. CAAC-ITRB-201305).

References

[1] A. Afkanpour, C. Szepesvári, M. Bowling, Alignment based kernel learning with a continuous set of base kernels, Machine Learn. 91 (3) (2013) 305–324.

[2] Y. Baram, Learning by kernel polarization, Neural Comput. 17 (6) (2005) 1264–1275.

[3] G. Baudat, F. Anouar, Generalized discriminant analysis using a kernel approach, Neural Comput. 12 (10) (2000) 2385–2404.

[4] A. Ben-David, A lot of randomness is hiding in accuracy, Eng. Appl. Artif. Intell. 20 (7) (2007) 875–885.

[5] C.-C. Chang, C.-J. Lin, LIBSVM Data: Classification, Regression, and Multilabel, 2013. <http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets>.

[6] O. Chapelle, V. Vapnik, S. Mukherjee, Choosing multiple parameters for support vector machines, Machine Learn. 46 (1) (2002) 131–159.

[7] K.-M. Chung, W.-C. Kao, C.-L. Sun, L.-L. Wang, C.-J. Lin, Radius margin bounds for support vector machines with the RBF kernel, Neural Comput. 15 (11) (2003) 2463–2681.

[8] C. Cortes, M. Mohri, A. Rostamizadeh, Two-stage learning kernel algorithms, in: Proceedings of the 27th International Conference on Machine Learning, 2010, pp. 239–246.

[9] N. Cristianini, J. Shawe-Taylor, A. Elisseeff, J. Kandola, On kernel-target alignment, in: Proceedings of Advances in Neural Information Processing Systems 14, 2002, pp. 367–373.

[10] Delve, Delve Datasets, 2013. <http://www.cs.toronto.edu/~delve/data/datasets.html>.

[11] K. Duan, S.S. Keerthi, A.N. Poo, Evaluation of simple performance measures for tuning SVM hyperparameters, Neurocomputing 51 (2003) 41–59.

[12] A. Frank, A. Asuncion, UCI machine learning repository, University of california, School of Information and Computer Science, Irvine, CA, 2010. <http://archive.ics.uci.edu/ml/>.

[13] A. Gretton, O. Bousquet, A. Smola, B. Schölkopf, Measuring statistical dependence with Hilbert–Schmidt norms, in: Proceedings of the 16th International Conference on Algorithmic Learning Theroy, 2005, pp. 63–77.

[14] J. Kandola, J. Shawe-Taylor, Refining kernels for regression and uneven classification problems, in: Proceedings of the 9th International Workshop on Artifical Intelligence and Statistics, 2003.

[15] U.H.-G. Kreβel, Pairwise classification and support vector machines, in: B. Schölkopf, C.J.C. Burges, A.J. Smola (Eds.), Advances in Kernel Methods-Support Vector Learning, MIT Press, Cambridge, MA, 1999.

[16] V. Lampret, The Euler–Maclaurin and Taylor formulas: twin, elementary derivations, Math. Mag. 74 (2001) 109–122.

[17] J. Lu, K.N. Plataniotis, A.N. Venetsanopoulos, Face recognition using kernel direct discriminant analysis algorithms, IEEE Trans. Neural Netw. 14 (1) (2003) 117–126.

[18] Y. Lu, L. Wang, J. Lu, J. Yang, C. Shen, Multiple kernel clustering based on centered kernel alignment, Pattern Recogn. 47 (2014) 3656–3664.

[19] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, K.-R. Müller, Fisher discriminant analysis with kernels, neural networks for signal processing IX, in: Proceedings of the 1999 IEEE Signal Processing Society Workshop, 1999, pp. 41–48.

[20] C.H. Nguyen, T.B. Ho, An efficient kernel matrix evaluation measure, Pattern Recogn. 41 (11) (2008) 3366–3372.

[21] R. Rifkin, A. Klautau, In defense of one-vs-all classification, Machine Learn. Res. 5 (2004) 101–141.

[22] A. Rostamizadeh, Theoretical foundations and algorithms for learning with multiple kernels, PhD diss., New York University, 2010, pp. 71–72.

[23] B. Schölkopf, A.J. Smola, K.-R.Müller, Kernel principal component analysis, in: M. Press (Ed.), Advances in Kernel Methods: Support Vector Learning, 1999, pp. 327–352.

[24] J. Shawe-Taylor, N. Cristianini, Kernel Methods for Pattern Analysis, Cambridge University Press, New York, USA, 2004. pp. 150–151.

[25] M. Sugiyama, Local Fisher discriminant analysis for supervised dimensionality reduction, in: Proceedings of the 23rd International Conference on Machine Learning, ACM, 2006, pp. 905–912.

[26] S. Theodoridis, K. Koutroumbas, Pattern Recognition, Elsevier (Singapore) Pte Ltd., 2006. pp. 44–46.

[27] V.N. Vapnik, The Nature of Statistical Learning Theory, second ed., Springer-Verlag, New York, USA, 2000. pp. 138–145.

[28] J. Wang, H. Lu, K.N. Plataniotis, J. Lu, Gaussian kernel optimization for pattern classification, Pattern Recogn. 42 (7) (2009) 1237–1247.

[29] T. Wang, S. Tian, H. Huang, D. Deng, Learning by local kernel polarization, Neurocomputing 72 (13–15) (2009) 3077–3084.

[30] T. Wang, D. Zhao, Y. Feng, Two-stage multiple kernel learning with multiclass kernel polarization, Knowl.-Based Syst. 48 (2013) 10–16.

[31] W. Wang, Z. Xu, W. Lu, X. Zhang, Determination of the spread parameter in the gaussian kernel for classification and regression, Neurocomputing 55 (3–4) (2003) 643–663.

[32] K.-P. Wu, S.-D. Wang, Choosing the kernel parameters for support vector machine by the inter-cluster distance in the feature space, Pattern Recogn. 42 (5) (2009) 710–717.

[33] H. Xiong, M.N.S. Swamy, M.O. Ahmad, Optimizing the kernel in the empirical feature space, IEEE Trans. Neural Netw. 16 (2) (2005) 460–474.

[34] S. Zhong, D. Chen, Q. Xu, T. Chen, Optimizing the gaussian kernel function with the formulated kernel target alignment criterion for twoclass pattern classification, Pattern Recogn. 46 (7) (2013) 2045–2054.