Kernel MSE Algorithm: A Unified Framework for KFD, LS-SVM and KRR

Jianhua Xu, Xuegong Zhang, and Yanda Li

Dept. of Automation, Tsinghua University / State Key Lab of Intelligent Technology and Systems
Beijing 100084, China
xujianhua99@mails.tsinghua.edu.cn, zhangxg@mail.tsinghua.edu.cn

Abstract

In this paper, we generalize the conventional minimum squared error (MSE) method to yield a new nonlinear learning machine by using the kernel idea and adding different regularization terms. We name it as kernel minimum squared error or KMSE algorithm, which can deal with linear and nonlinear classification and regression problems. With proper choices of the output coding schemes and regularization terms, we prove that KMSE is identical to the kernel Fisher discriminant (KFD) except for an unimportant scale factor, and it is directly equivalent to the least square version for support vector machine (LS-SVM). For continuous real output values, we find that KMSE is the kernel ridge regression (KRR) with a bias. Therefore KMSE can act as a general framework that includes KFD, LS-SVM and KRR as its particular cases. In addition, we simplify the formula to estimate the projecting direction of KFD. Experiments on artificial and real world data sets in numerical computation aspects demonstrate that KMSE is a class of powerful kernel learning machines.

I. INTRODUCTION

In the last few years, support vector machine (SVM) is one of the most influential developments in the machine learning [1-4][14]. One of its prominent advantages is the idea of using kernels to realize the nonlinear transforms without knowing the detailed transforms. According to this idea, other authors proposed a class of kernel-based algorithms, such as the kernel Fisher discriminant analysis or KFD [5], the least square version for support vector machines or LS-SVM [6], and the kernel ridge regressions without bias term or KRR [7].

In classical linear classifiers, minimum squared error algorithm (MSE) and Fisher linear discriminant (FLD) are still widely used in practice since they are simple and they can tackle linearly separable and non-separable cases [8-11]. It had been proved that with the proper choice of the output coding scheme, MSE is equivalent to FLD, and that MSE approaches a minimum mean-squared-error approximation to the Bayesian discriminant function as the number of samples approaches infinity [8][9]. Therefore MSE can be viewed as a more general classifier, which can fulfill other types of classifiers. Furthermore, for continuous real outputs, MSE directly is the least squares linear regression algorithm.

In this paper, we generalize the conventional MSE method to yield a new type of nonlinear learning machine, by using the kernel idea and adding different regularization terms. Since the different regularization term gives the solution different properties, two regularization terms are used to generate two different algorithms. We name the proposed learning machines as kernel minimum squared error or KMSE algorithm. With properly chosen output coding schemes and regularization terms, we prove that KMSE is identical to KFD except for an unimportant scale factor and is directly equivalent to LS-SVM. For the continuous output values, we prove that KMSE is KRR with a bias. Therefore, KMSE can be viewed as a class of more general kernel algorithms, which can implement KFD, LS-SVM and KRR as its three special cases. Also, we simplify the formula to estimate the projecting direction of KFD. In order to evaluate the performance of KMSE and the equivalence between different methods in computational aspects, we took three experiments (the two spirals problem, an image classification and a cancer classification). The results demonstrate that KMSE is a powerful kernel algorithm.

This paper is organized as in the following way: In section II, KMSE classifier is defined by using kernel ideas and defining different objective functions. The equivalence between KMSE and KFD is proved in section III. In section IV, the equivalence between KMSE and LS-SVM is discussed. Section V analyses the relation between KMSE and KRR. The experiment results of several artificial and real world data sets are reported and analyzed in section VI. Finally we present the conclusions and discussions.

II. THE KERNEL MSE ALGORITHM

In this section, we present the kernel MSE or KMSE algorithm using the kernel idea and defining different objective functions. For simplicity, we first consider the binary classification problem.

Let $\boldsymbol{x}_1 = \{\mathbf{x}_1^1, \ldots, \mathbf{x}_{l_1}^1\}$ and $\boldsymbol{x}_2 = \{\mathbf{x}_1^2, \ldots, \mathbf{x}_{l_2}^2\}$ be the

This work is supported by Natural Science Foundation of China, project No.69885004.

0-7803-7044-9/01/$10.00 ©2001 IEEE

training samples from two different classes ω_1, ω_2, and denote $x = x_1 \cup x_2 = \{\mathbf{x}_1, ...\mathbf{x}_l\}$. Let $\mathbf{y} = [y_1, ..., y_l]^T$ be the output coding scheme of samples, where $\mathbf{x}_i^1, \mathbf{x}_j^2, \mathbf{x}_k \in R^n$, $y_k \in R$, $i = 1,...,l_1$, $j = 1,...,l_2$, $k = 1,...,l$, $l = l_1 + l_2$ and l_1, l_2 are the number of samples of classes ω_1, ω_2 respectively.

In the classical MSE approach, the objective function of the training phase is defined as the summation of squared errors between the output code and actual output for samples, i.e.,

$$E_0(\mathbf{w}_M, b_M) = \frac{1}{2}(\mathbf{y} - \mathbf{A}\mathbf{w}_M - b_M\mathbf{u})^T(\mathbf{y} - \mathbf{A}\mathbf{w}_M - b_M\mathbf{u}) \quad (1)$$

where

$$\mathbf{A} = [\mathbf{x}_1^1 \quad ... \quad \mathbf{x}_{l_1}^1 \quad \mathbf{x}_1^2 \quad ... \quad \mathbf{x}_{l_2}^2]^T \quad (2)$$

$\mathbf{w}_M \in R^n$ and $b_M \in R$ stand for the weight vector and threshold respectively, and $\mathbf{u}$ is a column vector with l ones. The conventional MSE classifier is the solution of a set of linear equations derived from functional (1). Obviously, MSE solution depends on the output coding schemes and different choices arrive at solutions with different properties [8]. There exist two well-studied choices for the output coding schemes. One is

$$y_i = \begin{cases} +1, & if \ \mathbf{x}_i \in \omega_1 \\ -1, & if \ \mathbf{x}_i \in \omega_2 \end{cases}, \quad (3)$$

which results in that the MSE solution approaches an optimal mean-squared-error approximation to the Bayesian discriminant function as the number of samples approaches infinity. Another choice is

$$y_i = \begin{cases} +l/l_1, & if \ \mathbf{x}_i \in \omega_1 \\ -l/l_2, & if \ \mathbf{x}_i \in \omega_2 \end{cases}, \quad (4)$$

which cause the MSE is identical to FLD except for an unimportant scale factor. If the output of MSE is continuous values but not class labels, MSE becomes a linear regression algorithm.

Now we generalize the classical MSE algorithm by applying some kernel functions and adding a suitable regularization term in objective functional.

Assume Φ is a nonlinear mapping ($\Phi: R^n \rightarrow F$), which transform the vectors in the input space into vectors in some new feature space F. In the F space, we build a linear MSE whose weight vector and threshold are denoted by $\mathbf{w}_\Phi$ and β_M respectively. From the theory of reproducing kernels we know that any solution in the feature space must lie in the span of all training samples in the feature space [5][14]. Therefore we can construct an expansion for $\mathbf{w}_\Phi$ in the form,

$$\mathbf{w}_\Phi = \sum_{i=1}^{l} \alpha_i^M \Phi(\mathbf{x}_i) \quad (5)$$

where $\alpha_i^M \in R, i = 1,2,....l$ are coefficients which describe significance of each sample in the weight vector. Thus, by using the expansion (5) and the kernel function [1-3][14]

$$k(\mathbf{x}_i, \mathbf{x}_j) = (\Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)) \quad (6)$$

We can define the objective function of the MSE algorithm in the feature space F as,

$$E_0^\Phi(\boldsymbol{\alpha}_M, \beta_M) = \frac{1}{2}(\mathbf{y} - \mathbf{K}^T\boldsymbol{\alpha}_M - \beta_M\mathbf{u})^T(\mathbf{y} - \mathbf{K}^T\boldsymbol{\alpha}_M - \beta_M\mathbf{u}) \quad (7)$$

where $\boldsymbol{\alpha}_M = [\alpha_1^M, ..., \alpha_l^M]^T$. $(\mathbf{K})_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$, $i, j = 1,..l$, is the positive semi-definite kernel matrix satisfying the Mercer condition. A linear set of equations can be derived from (6), i.e.,

$$\begin{bmatrix} \mathbf{K}\mathbf{K}^T & \mathbf{K}\mathbf{u} \\ (\mathbf{K}\mathbf{u})^T & l \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha}_M \\ \beta_M \end{bmatrix} = \begin{bmatrix} \mathbf{K}\mathbf{y} \\ \mathbf{u}^T\mathbf{y} \end{bmatrix} \quad (8)$$

Note that this coefficient matrix is always singular since we want to estimate $l+1$ parameters from l samples, which will cause multiple solutions.

According to statistical learning theory, if two classifiers have the same training error, the classifier with smallest capacity is more likely to perform better [12]. In an effort to choose one solution among the many solutions of (8), additional regularization term can be added [12]. Smola and Scholkopf [13] pointed out that the regularization term can effectively reduce the model space and thereby control the complexity of the solution (i.e. control capacity and generalization). There exist two usual regularization terms: $\boldsymbol{\alpha}_M^T\boldsymbol{\alpha}_M$ in KFD [5], and $\mathbf{w}_\Phi^T\mathbf{w}_\Phi$ in SVM [1][13][14], LS-SVM [6] and ridge regression [7].

Now, we add these terms in objective function (7) and construct different regularized objective functions, i.e.,

$$E_1^\Phi(\boldsymbol{\alpha}_M, \beta_M) = \frac{1}{2}\mu_1\boldsymbol{\alpha}_M^T\boldsymbol{\alpha}_M + E_0^\Phi(\boldsymbol{\alpha}_M, \beta_M) \quad (9)$$

$$E_2^\Phi(\boldsymbol{\alpha}_M, \beta_M) = \frac{1}{2}\mu_2\mathbf{w}_\Phi^T\mathbf{w}_\Phi + E_0^\Phi(\boldsymbol{\alpha}_M, \beta_M) \quad (10)$$

where μ_1 *and* μ_2 are positve constants or regularization parameters and

$$\mathbf{w}_\Phi^T\mathbf{w}_\Phi = \boldsymbol{\alpha}_M^T\mathbf{K}\boldsymbol{\alpha}_M \quad (11)$$

Minimizing these objective functions, we obtian two new sets of linear equations,

$$\begin{bmatrix} \mathbf{K}\mathbf{K}^T + \mu_1 \mathbf{I} & \mathbf{K}\mathbf{u} \\ (\mathbf{K}\mathbf{u})^T & l \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha}_M \\ \beta_M \end{bmatrix} = \begin{bmatrix} \mathbf{K}\mathbf{y} \\ \mathbf{u}^T\mathbf{y} \end{bmatrix} \tag{12}$$

and

$$\begin{bmatrix} \mathbf{K} + \mu_2 \mathbf{I} & \mathbf{u} \\ \mathbf{u}^T\mathbf{K}^T & l \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha}_M \\ \beta_M \end{bmatrix} = \begin{bmatrix} \mathbf{y} \\ \mathbf{u}^T\mathbf{y} \end{bmatrix} \tag{13}$$

Note that we asuumed the kernel matrix is not singular in derivation of equation (13). From the viewpoint of numerical stability, if the constants (μ_1 and μ_2) are large enough, the coefficient matrices in (12) and (13) become positve definite and the problem can be made more stable [13].

Now we obtained two linear machines in the feature space, or two nonlinear machines with kernels in the original input space, which are the solutions of two linear sets of equations. We gave these two algorithms one name, i.e., KMSE, to emphasize that they are two versions of implementation of the same idea.

Like MSE, KMSE solutions depend on the output coding schemes and the different choices give the solutions different properties. In the next section, we'll prove that KMSE is identical to KFD when we choose (4), (8) and (12), and in section IV, we'll prove the equivalence between KMSE and LS-SVM with (3) and (13) chosen. In section V, we'll see that for continuous real outputs and equation (13), KMSE is the KRR with a bias. Therefore KMSE can be viewed as a general class of kernel learning machines which includes KFD, LS-SVM and KRR as its specific cases.

III. EQUIVALENCE BETWEEN KMSE AND KFD

For two class problem, the basic idea of FLD is to find an orientation for which the projected samples are well separated [8][9]. Mika et al [5] generalized the classical FLD using kernel idea and defined the kernel Fisher discriminant (KFD). The basic conception of this technique is that the features in the input space are transformed into some feature space nonlinearly and in this feature space an optimal projected direction is found by using FLD.

The objective function in KFD [5] to be maximized is,

$$J(\boldsymbol{\alpha}_F) = \frac{\boldsymbol{\alpha}_F^T \mathbf{M} \boldsymbol{\alpha}_F}{\boldsymbol{\alpha}_F^T \mathbf{N} \boldsymbol{\alpha}_F} \tag{14}$$

where

$$\boldsymbol{\alpha}_F = (\alpha_1^F, ..., \alpha_l^F)^T \tag{15}$$

$$\mathbf{M} = (\mathbf{M}_1 - \mathbf{M}_2)(\mathbf{M}_1 - \mathbf{M}_2)^T \tag{16}$$

$$(\mathbf{M}_i)_j = \frac{1}{l_i} \sum_{k=1}^{l_i} \mathrm{k}(\mathbf{x}_j, \mathbf{x}_k^i), \quad i = 1,2; j = 1,..l \tag{17}$$

$$\mathbf{N} = \sum_{j=1}^{2} \mathbf{K}_j (\mathbf{I} - \mathbf{1}_{l_j}) \mathbf{K}_j^T \tag{18}$$

$$(\mathbf{K}_j)_{nm} = \mathrm{k}(\mathbf{x}_n, \mathbf{x}_m^j), \quad j = 1,2; n = 1,..,l; m = 1,...,l_j \tag{19}$$

and $\mathbf{I}$ is the identity matrix, $\mathbf{1}_{l_j}$ is the matrix with all entries as $\frac{1}{l_j}$ $(j = 1,2)$.

In the work of Mika et al [5], the solution vector $\boldsymbol{\alpha}_F$, which maximizes the functional (14), is to find the leading eigenvector of $\mathbf{N}^{-1}\mathbf{M}$. In fact, like the derivation of FLD, we can simplify this computation and obtain,

$$\boldsymbol{\alpha}_F = \mathbf{N}^{-1}(\mathbf{M}_1 - \mathbf{M}_2) \tag{20}$$

Obviously, since we estimate the l dimensional covariance structures from l samples, the proposed setting is ill posed [5]. In order to cope with numerical stability problem or to control the capacity, Mika et al simply substituted $\mathbf{N}_\mu = \mathbf{N} + \mu \mathbf{I}$ for $\mathbf{N}$, where μ is a positive constant. The threshold β_F usually can be represented as,

$$\beta_F = -\boldsymbol{\alpha}_F^T \frac{l_1 \mathbf{M}_1 + l_2 \mathbf{M}_2}{l} \tag{21}$$

KFD is to find an optimal linear projected direction in some feature space. However such a projected orientation is nonlinear in the original input space.

Now, we prove that the KMSE algorithm (8) and (12) is equivalent to KFD when choosing the output codes as (4). Let $\mathbf{K}=[\mathbf{K}_1 \quad \mathbf{K}_2]$, $\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix}$, $\mathbf{u} = \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{bmatrix}$, we can rewrite equation (8) as

$$\begin{bmatrix} \mathbf{K}_1\mathbf{K}_1^T + \mathbf{K}_2\mathbf{K}_2^T & \mathbf{K}_1\mathbf{u}_1 + \mathbf{K}_2\mathbf{u}_2 \\ (\mathbf{K}_1\mathbf{u}_1 + \mathbf{K}_2\mathbf{u}_2)^T & l \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha}_M \\ \beta_M \end{bmatrix} = \begin{bmatrix} \mathbf{K}_1 & \mathbf{K}_2 \\ \mathbf{u}_1^T & \mathbf{u}_2^T \end{bmatrix} \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} \tag{22}$$

Let the output coding scheme be equation (4), i.e., $\mathbf{y}_1 = \frac{l}{l_1}\mathbf{u}_1, \mathbf{y}_2 = -\frac{l}{l_2}\mathbf{u}_2$, the linear set of equations (22) becomes

$$\begin{bmatrix} \mathbf{K}_1\mathbf{K}_1^T + \mathbf{K}_2\mathbf{K}_2^T & \mathbf{K}_1\mathbf{u}_1 + \mathbf{K}_2\mathbf{u}_2 \\ (\mathbf{K}_1\mathbf{u}_1 + \mathbf{K}_2\mathbf{u}_2)^T & l \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha}_M \\ \beta_M \end{bmatrix} = \begin{bmatrix} l(\mathbf{M}_1 - \mathbf{M}_2) \\ 0 \end{bmatrix} \tag{23}$$

From this equation set, firstly we can obtain

$$\beta_M = -\frac{1}{l}(l_1\mathbf{M}_1 + l_2\mathbf{M}_2)^T \boldsymbol{\alpha}_M = \beta_F \tag{24}$$

The definition of formula (18) can be described in the form,

$$\mathbf{N} = \sum_{j=1}^{2}(\mathbf{K}_j\mathbf{K}_j^T - l_j\mathbf{M}_j\mathbf{M}_j^T) \tag{25}$$

From (23), we obtain

$$\frac{1}{l}(\mathbf{N} + \frac{l_1 l_2}{l^2}(\mathbf{M}_1 - \mathbf{M}_2)(\mathbf{M}_1 - \mathbf{M}_2)^T)\boldsymbol{\alpha}_M = (\mathbf{M}_1 - \mathbf{M}_2) \tag{26}$$

Since $\gamma = (\mathbf{M}_1 - \mathbf{M}_2)^T\boldsymbol{\alpha}_M$ is a scalar, this equation can be further simplified,

$$\boldsymbol{\alpha}_M = l(1 - \frac{l_1 l_2}{l^2}\gamma)\mathbf{N}^{-1}(\mathbf{M}_1 - \mathbf{M}_2) \tag{27}$$

which except for an unimportant scale factor is identical to the solution for KFD (20). Especially when we substitute $\mathbf{N}_\mu$ for $\mathbf{N}$, such KFD is identical to the solution of a linear set of equations (12).

In this section, we have proved that KMSE is equivalent to KFD with choosing the output coding scheme (4) and objective functions (8) and (10). This means that KFD is a special case of KMSE.

IV. Equivalence Between KMSE and LS-SVM

The least square version of support vector machines [6] by formulating the classification problem can be described as

$$\min_{\mathbf{w}_L, b_L, e} L(\mathbf{w}_L, b_L, e) = \frac{1}{2}\mathbf{w}_L^T\mathbf{w}_L + \gamma\frac{1}{2}\sum_{k=1}^{l} e_k^2 \tag{28}$$

subject to the equality constrains

$$y_k(\mathbf{w}_L^T\Phi(\mathbf{x}_k) + \beta_L) = 1 - e_k,\ k = 1,...,l \tag{29}$$

Thus the optimal problem can be turned into a linear set of equations,

$$\begin{bmatrix} \Omega + \gamma^{-1}\mathbf{I} & \mathbf{y} \\ -\mathbf{y}^T & 0 \end{bmatrix}\begin{bmatrix} \boldsymbol{\alpha}_L \\ \beta_L \end{bmatrix} = \begin{bmatrix} \mathbf{u} \\ 0 \end{bmatrix} \tag{30}$$

where

$$\Omega = \mathbf{Z}\mathbf{Z}^T \tag{31}$$

$$\mathbf{Z} = [y_1\Phi(\mathbf{x}_1),..., y_l\Phi(\mathbf{x}_l)]^T \tag{32}$$

$$\boldsymbol{\alpha}_L = [\alpha_1^L,..., \alpha_l^L]^T \tag{33}$$

Note that there exist some mistakes in definition of symbols in paper [6]. In the least square version of support vector machine, only a linear set of equations has to be solved instead of the quadratic programming problem in original SVM.

Now, we prove the equivalence between KMSE and LS-SVM with the outputs (3) and the linear set of equations (13).

We define a diagonal matrix,

$$\mathbf{Y} = diag(y_1, y_2,..., y_l) \tag{34}$$

This matrix is symmetric and always non-singular. Again, we rewrite

$$\boldsymbol{\alpha}_M = [y_1\alpha_1^S,..., y_l\alpha_l^S]^T = \mathbf{Y}\boldsymbol{\alpha}_S \tag{35}$$

where $\boldsymbol{\alpha}_S = [\alpha_1^S, \cdots, \alpha_l^S]^T$. From the linear set of equations (13), we have

$$(\mathbf{K} + \mu_2\mathbf{I})\boldsymbol{\alpha}_M + \mathbf{u}\beta_M = \mathbf{y} \tag{36}$$

and

$$\mathbf{u}^T(\mathbf{K}^T\boldsymbol{\alpha}_M + \mathbf{u}\beta_M) = \mathbf{u}^T\mathbf{y} \tag{37}$$

By multiplying $\mathbf{Y}$ in (36) and applying (35), we obtain

$$(\Omega + \mu_2\mathbf{I})\boldsymbol{\alpha}_S + \mathbf{y}\beta_M = \mathbf{u} \tag{38}$$

where $\Omega = \mathbf{YKY}$, $\mathbf{YY} = \mathbf{I}$, $\mathbf{Yu} = \mathbf{y}$ and $\mathbf{Yy} = \mathbf{u}$.

Now eliminating $\mathbf{u}\beta_M$ from (36) and (37), we obtain

$$-\mathbf{y}^T\boldsymbol{\alpha}_S = 0 \tag{39}$$

From (38) and (39), a new linear set of equations is constructed

$$\begin{bmatrix} \Omega + \mu_2\mathbf{I} & \mathbf{y} \\ -\mathbf{y}^T & 0 \end{bmatrix}\begin{bmatrix} \boldsymbol{\alpha}_S \\ \beta_M \end{bmatrix} = \begin{bmatrix} \mathbf{u} \\ 0 \end{bmatrix} \tag{40}$$

Comparing equation (40) with (30), we find out that the linear set of equations (30) is equivalent to (40) when $\mu_2 = \gamma^{-1}$. This equivalence between KMSE and LS-SVM indicates that LS-SVM can be viewed as a special case of KMSE too.

V. Relationship Between KMSE and KRR

In [7], Saunders et al proposed one dual form of the ridge regression, which does not involve a threshold. A linear set of equations is built as,

$$(\mathbf{K} + \mu_3\mathbf{I})\boldsymbol{\alpha}_R = \mathbf{y} \tag{41}$$

In linear case, if we add a dimension in sample vectors and weight vector, the threshold or bias term can be hidden in the weight vector thus need not be considered in derivation procedure. However in its dual form, if there is not a threshold, the dual form cannot be degenerated into the linear one by using the linear kernel. When adding a threshold, the dual of ridge regression is the linear set of equations (13).

Therefore for the continuous value output, KMSE is directly the dual form of ridge regressions. Moreover the regression function with a threshold is more comprehensive.

VI. Experiments

Since KMSE can approach to the performance of KFD, LS-SVM and KRR, we designed several experiments on artificial and real-world data sets to evaluate the performance of KMSE in the computation aspects.

A. The Two Spiral Problem

For the two spiral problem, our task is to discriminate between two sets of sample points which lie in two spirals in a plane. As shown in Figs. 1 and 2, in our experiment each category includes 108 samples. The samples of the two classes are illustrated as "+"s and "."s respectively.

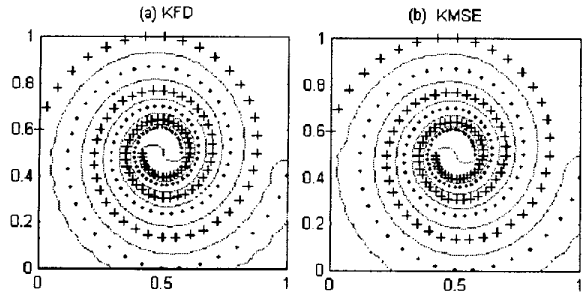

Fig.1 The separation hyperplane obtained with KFD (a) and KMSE (b). The KMSE algorithm was tuned to similate KFD.

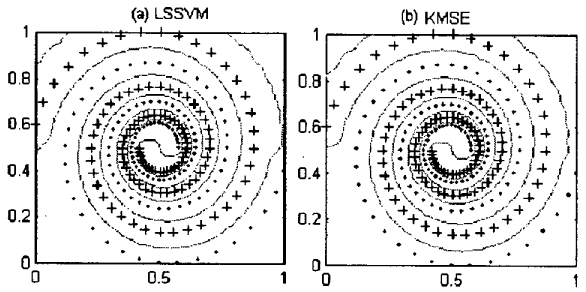

Fig.2 The separation hyperplane obtained with LS-SVM (a) and KMSE (b). The KMSE algorithm was tuned to simulate LS-SVM.

The performance of KMSE and KFD is illustrated in Fig.1, where Fig.1(a) shows the separation line by KFD and Fig.1(b) shows that of KMSE. We adopted the RBF kernel function with $\sigma = 0.02$. Both KMSE and KFD classify all samples correctly and obtain the central and smooth hyperplanes. Fig.2 comparies the result of KMSE (Fig.2a) and that of LS-SVM (Fig.2b). All samples are correctly classified too. Again two smooth and centered hyperplanes between two category samples are found.

Therefore for the two spiral problem, we can obtain very good decision functions by KMSE, which can approach the results of LS-SVM and KFD.

B. An Image Segmentation Data Set

The image segmentation data sets from the DELVE repositories [16] include seven classes: cement, brick face, grass, foliage, sky, path and window. Each class consists of 30 training samples and 300 test samples. Every sample is characterized by eighteen attributes extracted from original images.

We compared the correct rate of KFD with that of KMSE using RBF kernel. In the numerical computation, we divided the seven-class problem into six binary classification problems. When the RBF parameter σ varies from 0.1 to 1.0 (with step 0.1), and μ and μ_1 are two fixed constants respectively, at $\sigma = 0.25$ KMSE and KFD attain the maximum 85.86% and 87.29% respectively, and at $\sigma = 0.2$ the same value 85.67. The maximal difference between two classifiers is less than 4%, which possibly results from the fixed μ and μ_1.

Also using RBF kernel, we compared the correct rates of LS-SVM and KMSE. With the RBF parameter σ increasing from 0.1 to 2.0 (with step 0.1), and $\lambda = 1/\mu_2$, the maximal distinction of correct rates between KMSE and LS-SVM is less than 1%, which maybe result from the numerical computation. At $\sigma = 1.0$, they reach the maximum 93.38 and 93.43 respectively. Particularly the correct rates of two algorithms are lager than 90% when σ ranges from 0.4 to 2.0.

These experiments again proved our argument that KMSE can be viewed as a general classifier which can fulfill KFD and LS-SVM as its special cases.

C. The Cancer (Leukemia) Classification Problem

In [15], Golub et al introduced a generic approach to cancer classification based on gene expression monitoring by DNA microarrays and used a data set included 38 training samples and 34 test samples from two categories: acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML). Since there are 6817 genes, i.e., 6817 attributes, 50 genes are selected to design and verify their classification approach. The result of their classifier is two samples were rejected in the training procedure, and five samples were rejected in the test procedure. (The decision would be error if these samples were not rejected).

We also use the 38 samples as training set and the 34 samples as test set to evaluate the performance of KFD, LS-SVM and KMSE. The linear kernel function is used in this experiment. The results are listed in table 1. When KMSE and KFD classify all samples correctly, there is only a misclassified test sample. For LS-SVM and the corresponding KMSE, there is one misclassified sample in the training set, and two misclassified among the test samples.

Table 1. Number of rejected/misclassified samples for the leukemia data set with different approaches

Data Set	Golub's Approach	KMSE & KFD	KMSE & LS-SVM
Training Set	2	0	1
Test Set	5	1	2

VII. Conclusion

In this paper, we extended the traditional MSE algorithm to nonlinear cases with kernels, and proposed the Kernel MSE or KMSE method. A proper regularization term is added to the objective function besides the summation of squared errors between the actual output of kernel neuron and the desired output. This can make the method more stable in the numerical computation and control its generalization ability. The relationships of KMSE with KFD, LS-SVM and KRR are discussed in detail, leading to the conclusion that KMSE can be viewed as a unified framework for the other methods. With these results, a better understanding of the kernel method family can be achieved.

References

[1] C. Cortes, V. N. Vapnik. Support Vector Networks. Ma-chine Learning, 20(3), 273-297, 1995.

[2] V. N. Vapnik. Statistical Learning Theory. Wiley, New York, 1998.

[3] V. N. Vapnik. An overview of statistical learning theory. IEEE Trans. on Neural Networks, 10, 988-999, 1999.

[4] V. N. Vapnik. The Nature of Statistical Learning Theory (2nd ed.). Springer-Verlag, New York, 1999.

[5] S. Mika, G. Ratsch, et al. Fisher discriminant analysis with kernels. Neural Networks for Signal Processing IX, 41-48. IEEE Press, New York, 1999.

[6] J. A. K. Suykens and J. Vandewalle. Least squares support vector machines. Neural Processing Letters, 9, 293-300, 1999.

[7] C. Saunders, A. Gammerman and V. Vovk. Ridge regression learning algorithm in dual variables. In J. Shavlik (editor), Machine Learning Proceedings of the Fifteenth International Conference, Morgan Kaufmann, 1998.

[8] R. O. Duda and P. E. Hart. Pattern Classification and Scene Analysis. John Wiley & Sons, New York, 1973.

[9] Z. Bian, X. Zhang,, et al. Pattern Recognition (2nd ed., in Chinese). Tsinghua University Press, Beijing, 2000.

[10] J. T. Tou and R.C. Gonzadez. Pattern Recognition Principles. Addison-Wesly, Reading, 1974.

[11] S. Theodoridis and K. Koutroumbas. Pattern Recognition. Academic Press, San Diego, 1999.

[12] I. Guyon and D. G. Stork. Linear discriminant and sup-port vector classifiers. In A. J. Smola, P. Bartlett, et al. (editors). Advances in Large Margin Classifiers. MIT Press, 2000.

[13] A. J. Smola and B. Scholkopf. On a kernel-based method for pattern recognition, regression, approximation and operator inversion. Algorithmica, 22, 211-231, 1998.

[14] N. Cristianini and J. S. Taylor. An introduction to sup-port vector machines and other kernel-based learning methods. Cambridge University Press, U.K, 2000.

[15] T. R. Golub, D. K. Slonim, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science, 286, 531-537, 1999.

[16] Data source: www.cs.toronto.edu/~delve/data/image-seg/desc.html