2007 IEEE International Conference on Control and Automation
Guangzhou, CHINA - May 30 to June 1, 2007

A Learning Approach to Derive Sparse Kernel Minimum Square Error Model

Yong Xu
Shenzhen Graduate School
Harbin Institute of Technology
Shenzhen, China
laterfall@hitsz.edu.cn

Jing-Yu Yang, Zhong Jin, Chuancai Liu
College of Computer Science & Technology
Nanjing University of Science & Technology
Nanjing, China
{yangjy,zhongjin}@mail.njust.edu.cn,
chuancailiu@yahoo.com.cn

***Abstract*—Kernel minimum square error (KMSE) model is computationally more tractable than other nonlinear methods, but it still has some drawbacks in theory and computational problems. Moreover, the characteristic that the classification efficiency of KMSE decreases as the size of the training sample set increases makes KMSE yield low classification efficiency for classification problems with a large number of training samples. In this paper, several methods which are developed for improving the classification efficiency of KMSE are assessed and their shortcomings are indicated. Then, KMSE is presented as a regression model. Taking advantage of local ridge regression, we develop an efficient KMSE classification technique. The proposed technique can sufficiently exploit the theoretical merit of local ridge regression which may produce more stable estimates with smaller variance than the least square error technique. This technique can also determine local regularization parameters properly and automatically, and then construct an improved KMSE model with lower structure complex which leads to a more efficient classification process. Experiments show that the improved KMSE model not only classifies much more efficiently but also obtains higher classification accuracy than KMSE, while outperforming several existing improved KMSE models.**

Keywords- Least square error; regression; classification

I. INTRODUCTION

The minimum square error (MSE) technique, which is theoretically equivalent to Fisher discriminant analysis, has received much attention in recent years. The so-called kernel minimum squares error (KMSE) method, is a late development of MSE. KMSE bases on the MSE technique and kernel functions. It would appear that the implementation of KMSE is theoretically equivalent to sequential implementations of the following two procedures: to transform the original sample space (input space) into a new high-dimensional space (feature space) and then to construct the MSE model using the data in the feature space. On the other hand, KMSE is much more mathematically tractable than the two phases above. By using the kernel functions, one may not explicitly carry out the procedure of transforming the input space into the feature space in implementing of KMSE. Nevertheless, for the implementation of an ordinary nonlinear method, the two procedures above must be explicitly carried out, and consequently it is computationally more expensive than the implementation of KMSE. Another theoretical property of KMSE is that KMSE [1] may be still equivalent to LS-SVM or Fisher discriminant analysis [2,3].

KMSE has been proposed and applied for years yet some theoretical and computational drawbacks exist in this model. Moreover, the characteristic that the classification efficiency of KMSE decreases as the size of the training sample set increases is also a disadvantage for its real-world applications. Sometime this makes KMSE incompetent for applications which have a strict efficiency requirement. Therefore, it is very significant to improve KMSE for efficient classification. Though several methods for improving KMSE are available, they have some weaknesses. We will discuss this in detail in the next section.

II. THEORETICAL ANALYSIS ON KMSE

A Description of KMSE

We consider two-class problems in which category labels for the two classes are 1 and –1, respectively. Suppose that the input space is mapped into a high-dimensional feature space F by a nonlinear function ϕ. And suppose l_1 samples, $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{l_1}$, are in class 1, while l_2 samples, $\mathbf{x}_{l_1+1}, \mathbf{x}_{l_1+2}, \ldots, \mathbf{x}_l$, are in class –1($l_1 + l_2 = l$). The MSE model for an l-training-samples set in the feature space can be formulated as

$$\mathbf{\Phi W} + \mathbf{e} = \mathbf{B}, \tag{1}$$

where

$$\mathbf{W} = \begin{bmatrix} w_0 \\ \mathbf{w} \end{bmatrix}, \quad \mathbf{B} = [1 \ . \ . \ . \ -1]^T,$$

$$\mathbf{\Phi} = \begin{bmatrix} 1 \ . \ . \ . \ . \ 1 \\ \boldsymbol{\varphi}(\mathbf{x}_1) \ . \ . \ . \ \boldsymbol{\varphi}(\mathbf{x}_l) \end{bmatrix}^T. \tag{2}$$

The vector $\mathbf{e}$ denotes the error, and the ith entry of the vector $\mathbf{B}$ is the class label of the ith training sample x_i. We call $\mathbf{w}$, w_0 discriminant vector and threshold in the feature space F, respectively. Because w can be expressed in terms of $\mathbf{w} = \sum_{i=1}^{l} \alpha_i \boldsymbol{\varphi}(\mathbf{x}_i)$ [5], $\mathbf{W}$ can be rewritten in the form of

$$\mathbf{W} = \begin{bmatrix} w_0 \\ \sum_{i=1}^{l} \alpha_i \boldsymbol{\varphi}(\mathbf{x}_i) \end{bmatrix}. \quad (3)$$

By introducing the kernel function $k(\mathbf{x}_i, \mathbf{x}_j) = \boldsymbol{\varphi}(\mathbf{x}_i)^T \boldsymbol{\varphi}(\mathbf{x}_j)$, we can obtain

$$\mathbf{KA} + \mathbf{e} = \mathbf{B}, \quad (4)$$

where $\mathbf{A} = [w_0 \quad \alpha_1 \ . \ . \ . \ \alpha_l]^T$,

$$(\mathbf{K})_{ij} = k(\mathbf{x}_i, \mathbf{x}_{j-1}), i = 1,2,...,l, j = 2,3,...l+1,$$

$$(\mathbf{K})_{i1} = 1, i = 1,2,...,l, \quad (5)$$

$\mathbf{B}$ is also defined as (2). In this paper, the model formulated using (4) is called KMSE and $\mathbf{K}$ is the kernel matrix of KMSE. If $\mathbf{A}$ is evaluated correctly, classification can be carried out easily as follows. The projection onto $\mathbf{W}$, of a test sample $\boldsymbol{\varphi}(\mathbf{x})$ in the feature space, is computed by

$$l_p(\mathbf{x}) = w_0 + \sum_{i=1}^{l} \alpha_i k(\mathbf{x}, \mathbf{x}_i). \quad (6)$$

If $l_p(\mathbf{x}) > 0$, $\mathbf{x}$ will be classified into class 1; otherwise $\mathbf{x}$ will be classified into class -1. $l_p(\mathbf{x})$ is called the classification function of KMSE. The computational cost of KMSE is much lower than that of ordinary nonlinear models and the classification decision based on this model is very simple.

B. *Drawbacks of KMSE*

Since KMSE bases its classification for every test sample on all the kernel functions which are determined by this test sample and all training samples, the classification efficiency of KMSE will be in inverse proportion to the size of the training sample set. This may make KMSE unsuitable for some applications with a large number of training samples, especially for ones with high efficiency requirement. Other kernel methods also suffer the same problem [4-12].

Moreover, KMSE is also a model that seeks to determine $l+1$ parameters $w_0, \alpha_1, ..., \alpha_l$ using l equations. As a result, these problems follow KMSE.

There is no unique solution to these parameters. Because the available equations are fewer than the unknown parameters, there are many possible solutions and we do not know which one is the true solution. It would appear that one solution, may be formally available, taking the following form:

$$\mathbf{A} = (\mathbf{K}^T\mathbf{K})^{-1}\mathbf{K}^T Y \quad (7\text{-}1)$$

The maximum possible rank of the $(l+1)\times(l+1)$ matrix $\mathbf{K}^T\mathbf{K}$ is l and $\mathbf{K}^T\mathbf{K}$ is singular.

Thus, if $\mathbf{A} = (\mathbf{K}^T\mathbf{K})^{-1}\mathbf{K}^T Y$ was numerically available, it would be quite numerically unstable. If this $\mathbf{A}$ is further exploited to classify new samples, it will generalize badly and the classification accuracy will be low. We explain it in brief as follows, analyzing the simplest case. Generally, there are errors in observation data. When these observation data are treated as inputs of a mathematical model, we may say that this model works well in controlling computational error if the errors in the observation data are not amplified. For the KMSE model, the matrix $\mathbf{K}$ is directly related to observation data with error. Suppose that $\mathbf{K}_0$ is associated with imaginary observation data without error, and $\mathbf{K} = \mathbf{K}_0 + \varepsilon\mathbf{I}$. In other words, the magnitude order of the observation error is ε. We have the following two outputs of KMSE: $\mathbf{Y}_0 = \mathbf{K}_0\mathbf{A}$ and $\mathbf{Y} = \mathbf{KA}$. The norm of the deviation between these two outputs is $\| \mathbf{Y} - \mathbf{Y}_0 \| = | \varepsilon | \cdot \| \mathbf{A} \|$, which shows that there is a deviation, proportional to the norm of $\mathbf{A}$, between the real output calculated using KMSE and the output associated with the imaginary observation data without error. As a result, the larger the norm of $\mathbf{A}$ is, the greater this deviation. It is certain that the $\mathbf{A}$ obtained using (7-1) must have a large norm because of the singularity of $\mathbf{K}^T\mathbf{K}$. Consequently, this solution $\mathbf{A}$ must generalize badly and lead to a high classification error rate.

It would appear that the computation of $\mathbf{A}$ can be formally improved and expressed as

$$\mathbf{A} = (\mathbf{K}^T\mathbf{K} + \mu\mathbf{I})^{-1}\mathbf{K}^T\mathbf{Y}, \quad (7\text{-}2)$$

where $\mathbf{I}$ is an identity matrix, and μ a positive constant. It may be regarded as an approach that artificially assigns one solution, which is directly related to μ, to the KMSE model. However, it is not guaranteed that the artificially assigned solution is the most suitable for KMSE and so the problem of how to properly set the value of μ should be addressed.

The drawbacks of KMSE presented above can be analyzed in another way. The solution to KMSE determined in (7-1) is the least square error solution. However, the least square error solution is generally applied to a regression model which aims to determine m parameters using n equations, where $n > m$. The KMSE model does not belong to this class of model. Hence, from the viewpoint of regression analysis, the KMSE model has too many parameters to be determined relative to the available equations.

C. *Improve KMSE*

As presented in section 2.2, KMSE has some drawbacks. On the other hand, if the KMSE model can be improved by containing fewer unknown parameters in the improved model, then not only more stable numerical solution can be expected but also a more efficient classification can be achieved.

Some literates improve kernel methods for efficient classification from the point of view of numerical approximation. For example, the numerical approximation approach proposed by B. Scholkopf et al. [4] is one of them. This approach bases on the supposition that one or more training samples can be expressed as a linear combination of the others in the feature space. Provided that this is true, an improved kernel model must be capable to be constructed to obtain a more efficient classification. Nevertheless, the

supposition is not always available for any case. An obvious exception to the supposition is the feature spaces associated with the Gaussian kernel, in which none of the training samples can be expressed in terms of the others.

The orthogonal least square error method (OLS) [5] was also used to improve KMSE for achieving more efficient classification. This method is usually based on Gram-Schmidt orthogonalization [13,14], which has been shown to be numerically unstable. More importantly, the kernel matrix of KMSE is ill-conditioned, therefore the expectation of exactly orthogonalizing the column vectors of the kernel matrix of KMSE and obtaining the best improved KMSE model cannot be satisfied [15,16].

Also, we have developed several improved KMSE models. The method presented in ref. [6] is not only very simple but also computationally cheap. However, the strategy employed in this method is so simple that it also cannot obtain the optimal solution to the problem of improving KMSE. Another drawback of the method is that the value of the regularization parameter is artificially assigned and it is usually not optimal. Compared with the method in ref.[6], the ref. [7] presents a deliberated method to improve KMSE. However, the condition of terminating the procedure for selecting training samples to construct the classification function should be set using some experience. Moreover, because a system of linear equations should be solved in each iteration step, the method in ref. [7] has a high computational cost.

In this paper, we propose a novel improved KMSE model, using the ideas of local ridge regression analysis and model selection. The improved model is capable to achieve higher classification accuracy, whereas its structure complex is lower than the KMSE model. Experimental results show that the improved model classifies much more efficiently than KMSE and the greatest improvement of classification efficiency can reach 93.5%. The experiments also show that the improved KMSE model can achieve better classification result than KMSE and can outperform several existing methods for improving KMSE.

D. Ridge regression and a routine to improve KMSE

We regard $\mathbf{B}$ as the output of KMSE by viewing Eqs.(4) as a input-output model. Thus, $k(\cdot,\mathbf{x}_1),k(\cdot,\mathbf{x}_2),...,k(\cdot,\mathbf{x}_l)$ can be regarded as l predictor variables, while the elements of $\mathbf{B}$ are the values of the predicted variable. Each output of the regression model is decided by one observation of the l predictor variables. There are l observations of these variables. For example, the ith observation is $\left(k(\mathbf{x}_i,\mathbf{x}_1),k(\mathbf{x}_i,\mathbf{x}_2),...,k(\mathbf{x}_i,\mathbf{x}_l)\right)$, which is the ith row of the matrix $\mathbf{K}$, too.

Further, solving (4) using (7-2) may be regarded as a ridge regression approach. The ridge regression can produce more stable estimate with smaller standard deviation than the least square error technique. Although refs. [6,7] also improve KMSE using ridge regression, the value of the regularization parameter μ is only simply artificially set. Indeed, different values of μ will lead to quite different results. Therefore, a logical routine to improve KMSE should be the one that can first properly set the value of μ using some available means and then can construct new KMSE model with lower structure complex.

E. More discussion on improving KMSE

One of the main characteristics of KMSE is that the numbers of the predictor variables and the training samples are identical. We also say that the complexity of the model structure increases as the size of the training sample set increases. This in turn means that the computational cost of KMSE classification is proportional to the size of the training sample set. Moreover, the accuracy of the model will become lower and lower, and the variances of the estimates of the parameters obtained by solving the model will get larger and larger as the training samples become more and more. Another potential problem is that "overfitting" usually follows high model structure complexity. It is thus significant to construct a new KMSE model with fewer predictor variables (i.e. lower structure complexity) and higher model accuracy. The optimal model construction method should achieve a model which has a suitable structure complexity and satisfactory model accuracy. SVM also has lower structure complexity than KMSE, but it seems that SVM fails to obtain the optimal structure complexity. In fact, reformative kernel methods have theoretical advantages over SVM [8].

This paper extends theories and methodologies of improving KMSE for efficient classification with a novel point of view. The method developed in this paper is on a basis of reliable theoretical background and able to achieve an improved model with acceptable structure complexity and high classification accuracy.

III. Algorithm

We here exploit ridge regression, a technique with solid theoretical background, to obtain new KMSE model for efficient classification. The method to be developed uses both the global ridge regression and local ridge regression techniques. The problem of setting the optimal global regularization parameter is discussed and an algorithm is proposed at first. Then, the optimal value of the global regularization parameter is used as initial values of all local regularization parameters. After that, an improved KMSE model can be constructed.

A. Determine global regularization parameters

The key to ridge regression is to determine the regularization parameters. In practice, the optimal value of μ can be determined by using an iteration procedure based on the following formula [17]:

$$\hat{\mu} = \frac{\mathbf{Y}^T\mathbf{P}^2\mathbf{Y} \quad trace(\mathbf{G}^{-1} - \mu\mathbf{G}^{-1}\mathbf{G}^{-1})}{\hat{\mathbf{A}}^T\mathbf{G}^{-1}\hat{\mathbf{A}} \quad trace(\mathbf{P})} \tag{8}$$

, where $\mathbf{G} = \mathbf{K}^T\mathbf{K} + \mu\mathbf{I}$, $\hat{\mathbf{A}} = \mathbf{G}^{-1}\mathbf{K}^T\mathbf{Y}$, $\mathbf{P} = \mathbf{I} - \mathbf{K}\mathbf{G}^{-1}\mathbf{K}^T$. $\mathbf{P}$ is the projection matrix.

In implementing the procedure, we set an initial value for the μ and then calculate $\hat{\mu}$ using (8) . Then, the μ is newly set to $\hat{\mu}$ and (8) is repeatedly carried out in the same way till $\hat{\mu}$ convergences. The final $\hat{\mu}$ is taken as the value of the ridge regularization parameter, called global regularization parameter.

B. Improve KMSE using local ridge regression

Local ridge regression can well present local characteristic of data, while effectively eliminating the side-effects of outliers. Thus the use of the local ridge technique allows the improved KMSE to generalize well. Moreover, the use of the model selection technique combined with local ridge regression allows an improved KMSE with lower structure complexity to be obtained. Therefore, the classification process associated with the improved KMSE model is much more efficient than that associated with KMSE.

In this subsection, we develop an algorithm for improving KMSE using local ridge regression. Local ridge regression derives $\mathbf{A} = (\mathbf{K}^T\mathbf{K} + \Lambda)^{-1}\mathbf{K}^T\mathbf{Y}$ and $\Lambda = diag(\lambda_1, \lambda_2, ..., \lambda_l)$ from KMSE formulated as (4). Usually, λ_j is called local regularization parameters. Obviously, how to determine λ_j properly is the key of the algorithm. We take the final value of $\hat{\mu}$ obtained using (8) as initial values of local regularization parameters. Define that $a = \mathbf{Y}^T\mathbf{P}_j^2\mathbf{Y}$, $b = \mathbf{Y}^T\mathbf{P}_j^2\mathbf{K}_j\mathbf{Y}^T\mathbf{P}_j\mathbf{K}_j$, $c = \mathbf{K}_j^T\mathbf{P}_j^2\mathbf{K}_j(\mathbf{Y}^T\mathbf{P}_j\mathbf{K}_j)^2$, $\alpha = trace(\mathbf{P}_j)$, $\beta = \mathbf{K}_j^T\mathbf{P}_j^2\mathbf{K}_j$, where $\mathbf{P}_j$ is the projection matrix, the jth column vector $\mathbf{K}_j$ of $\mathbf{K}$ has been eliminated from $\mathbf{K}$, and $\mathbf{P}_j^2$ the matrix operation $\mathbf{P}_j^2 = \mathbf{P}_j\mathbf{P}_j$. The procedure of eliminating some column vectors of the matrix $\mathbf{K}$ is also the procedure of constructing the improved KMSE model. From the point of view, the column vectors associated with positive infinite λ_j should be eliminated from the matrix $\mathbf{K}$. Then λ_j can be calculated using

$$\lambda_j = \frac{c\alpha - b\beta}{b\alpha - a\beta} - \mathbf{K}_{j+1}^T\mathbf{P}_j\mathbf{K}_{j+1} . \quad (9)$$

There are the following possible computation results [18]. If $b\alpha - a\beta = 0$, the value of λ_j will be almost infinite and the corresponding column of the $\mathbf{K}$ will be eliminated. If $\lambda_j < 0$ and $a\beta > b\alpha$, the effect of the corresponding column vector of the $\mathbf{K}$ will be equivalent to the column vector associated with infinite local regularization parameter. It means that the current column vector should be eliminated from the $\mathbf{K}$. If $\lambda_j < 0$ and $a\beta < b\alpha$, the effect of the corresponding column vector of the $\mathbf{K}$ will be equivalent to the column vector associated with the local regularization parameter of "0", and the current column vector should not be eliminated from the $\mathbf{K}$ while the parameter should be set to 0. For other cases, values of local regularization parameters should not be changed and no column vectors of $\mathbf{K}$ should be eliminated. In conclusion, after we evaluate each λ_j using (9), we should revaluate the value of the λ_j defined as follows:

$$\lambda_j = \begin{cases} \infty & \text{if} \quad a\beta = \alpha b \\ \infty & \text{if} \quad \mathbf{K}_j^T\mathbf{P}_j\mathbf{K}_j > \frac{c\alpha - b\beta}{b\alpha - a\beta} \quad \text{and} \quad a\beta > \alpha b \\ 0 & \text{if} \quad \mathbf{K}_j^T\mathbf{P}_j\mathbf{K}_j > \frac{c\alpha - b\beta}{b\alpha - a\beta} \quad \text{and} \quad a\beta < \alpha b \\ \frac{c\alpha - b\beta}{b\alpha - a\beta} - \mathbf{K}_j^T\mathbf{P}_j\mathbf{K}_j & \text{if} \quad \mathbf{K}_j^T\mathbf{P}_j\mathbf{K}_j \le \frac{c\alpha - b\beta}{b\alpha - a\beta} \end{cases} \quad (10)$$

If the revaluated λ_j has infinite value, the corresponding column vectors of the matrix $\mathbf{K}$ should be eliminated. Superficially, it would appear that for each λ_j , there is an analytic solution and no re-estimation is necessary. However, there are other parameters to optimize and while λ_j is optimized the optimal values of the others also changes. Thus, optimizing all the parameters together has to be done as a class of re-estimation, doing one at a time and then repeating until they all converge.

In the procedure of constructing the improved KMSE model, we use the same approach of minimizing generalized cross-validation (GCV) error [18] to determine optimal values of the global and local regularization parameters. Consequently the two steps on ridge regression and local ridge regression are technically compatible. Furthermore, the use of these two steps can allow the improved KMSE model to meet our expectation. After the procedure, the obtained matrix $\mathbf{K}$ is denoted by $\mathbf{K}_s$. Thus, the resulting improved KMSE is $\mathbf{K}_s\mathbf{A}_s + \mathbf{e} = \mathbf{Y}$ and the solution is $\mathbf{A}_s = (\mathbf{K}_s^T\mathbf{K}_s + \Lambda_s)^{-1}\mathbf{K}_s^T\mathbf{Y}$, $\Lambda_s = diag(\lambda_1 \quad \lambda_2 \ldots \lambda_s)$. $\mathbf{K}_s$ has only s column vectors. Suppose that $\mathbf{K}_s$ is in the following form :

$(\mathbf{K}_s)_{ij} = k(\mathbf{x}_i, \mathbf{x}'_{j-1}), i = 1,2,...,l, j = 2,3,...s+1,$

$(\mathbf{K}_s)_{i1} = 1, i = 1,2,...,l,$ where $\mathbf{x}'_1, \mathbf{x}'_2, \ldots \mathbf{x}'_s$ are s samples selected from the total training samples. Here we also call each of $\mathbf{x}'_1, \mathbf{x}'_2, \ldots \mathbf{x}'_s$ a node , a term used in [5]. Then, the consequent classification function for a sample $\mathbf{x}$ is $l'_p(\mathbf{x}) = w'_0 + \sum_{i=1}^{l} \alpha'_i k(\mathbf{x}, \mathbf{x}'_i)$, where $\mathbf{A}_s = [w'_0 \quad \alpha'_1 \quad \alpha'_2 \ldots \alpha'_s]^T$. Because $s < l$, the improved KMSE model will have a more efficient classification process than the KMSE model.

IV. EXPERIMENTS

We conduct experiments on four benchmark datasets: Cancer, Diabetis, Heart and German. Each of them contains 100 training and test subsets. For every dataset, the first training subset is used as the training sample set and testing is

carried out for all the test subsets, respectively. The kernel function exploited in this experiment is $k(\mathbf{x},\mathbf{y}) = \exp(-\|\mathbf{x}-\mathbf{y}\|^2/2\sigma^2)$. We calculate the square sum of the standard deviation of each data component of the first training subset and regard it as the value of σ^2. The μ in KMSE is set to 0.1, while the initial value of μ associated with improved KMSE is also set 0.1.

TABLE 1. Classification results on each benchmark dataset (the mean and standard deviation of the percentage classification error rates on all the test subsets of one dataset)

	Diabetis	Cancer	Heart	German
KMSE	*20.1±1.9*	*19.4±3.8*	*18.4±3.2*	*23.9±2.1*
The method presented in ref.[5]	*23.3±1.9*	*25.5±4.6*	*16.3±3.3*	*24.1±2.6*
The method presented in ref.[6]	*21.4±1.8*	*20.6±4.0*	*16.3±2.9*	*19.1±2.1*
Our method	*20.0±1.7*	*20.6±4.1*	*16.3±3.4*	*19.3±2.1*

TABLE 2. The number of the total training samples in one training subset and the number of the "nodes" associated with three improved KMSE

	Diabetis	Cancer	Heart	German
The total number of training samples	*468*	*200*	*170*	*700*
The nodes obtained using the method presented in ref.[5]	*12*	*6*	*7*	*8*
The nodes obtained using the method presented in ref.[6]	*141*	*60*	*51*	*210*
The nodes obtained using Our method	*34(7.3%)*	*15(7.5%)*	*13(7.6%)*	*46(6.5%)*

From Table 1, we can see that for all the datasets except "Cancer", our method obtains lower classification error rate than KMSE. This in turn means that reducing the model structure complexity and properly setting the values of regularization parameters do improve the model accuracy. Moreover, Table 2 shows that our method uses fewer training samples than the total training samples to construct the classification function and therefore our method classifies much more efficiently than KMSE. The ratio of the number of the nodes determined using our method to that of the total training samples is shown within the brackets in the last row of Table 2. The highest and lowest ratios are 7.6% and 6.5%, respectively. In other words, the classification time of our method is below ten percent of that of KMSE and the least classification time of our method is only 6.5 percent of the classification of KMSE.

Superficially, it would appear that the method in ref. [5] can obtain an improved KMSE with lower structure complexity i.e. with fewer "nodes" in comparison with our method. However, this is not to say that the method in ref.[5] classifies more accurately than our method. In fact, the method in ref.[5] has the highest classification error rate among the four methods. It indicates that the method in ref.[5] cannot achieve the optimal balance between model structure complexity and classification accuracy. Theoretically, the kernel matrix of KMSE is not a matrix with full column rank and consequently the true orthogonal vector set of its column vectors cannot be exactly obtained. Nonetheless, the ref. [5] attempts to achieve an improved KMSE model based on this orthogonal vector set. Because of this fault, the KMSE model cannot be optimized by the method in ref. [5]. Consequently, it is not strange that the method in ref.[5] does not do the best in improving KMSE.

The classification error rate of our method is very close to that of the method in ref.[6], whereas fewer nodes are determined using our method than using the method in ref.[6] and our method classifies more efficiently than the method in ref.[6]. This comparison also demonstrates that our method can achieve the optimal balance between model structure complexity and classification accuracy.

V. CONCLUSION

The drawbacks in theory and numerical computational problems of KMSE, which make KMSE suffer from the problems of overfitting and generalizing poorly, are first analyzed in detail. These problems can be overcome by using model structure stabilization and the regularization technique. The target of stabilizing model structure is to construct a model with a low structure complexity i.e. few nodes, which can generalize well. The regularization technique penalizes the structure complexity of a model by using the regularization parameter.

The novel method for improving KMSE may be regarded as some combination of the method of stabilizing model structure and the regularization technique. By contrast with the least square error technique applied to KMSE, the use of the local ridge regression technique allows more stable estimate with smaller variance of a model to be obtained. Our method can automatically determine not only the model structure complexity but also the optimal values of the regularization parameters. The technical routine of taking the optimal value of the global regularization parameter as initial values of local regularization parameters is very effective and efficient for obtaining the convergence values of local regularization parameters.

The experiments on benchmark datasets show that our method developed in this paper does powerfully improve KMSE to achieve more efficient classification as the theoretical analysis predicts. For example, in the best case, the classification time of the improved KMSE is only 6.5% of the classification time of KMSE. Moreover, the improved KMSE obtained using our method can achieve higher classification accuracy than KMSE. The experiments also show that our method outperforms several existing methods for improving KMSE, getting higher classification accuracy or classification efficiency.

ACKNOWLEDGEMENTS

This work was supported by Natural Science Foundation of China (Nos. 60602038 and 60620160097) and Natural Science Foundation of Guangdong Province, China (No. 06300862)

REFERENCES

[1] XU J, ZHANG X, LI Y. Kernel MSE algorithm: A unified framework for KFD, LS-SVM and KRR[A]. In: Proceedings of the International Joint Conference on Neural Networks(IJCNN-2001) [C], Washington, D.C: 2001

[2]Yong Xu, Jing-yu Yang, Zhong Jin. Theory analysis on FSLDA and ULDA[J]. Pattern Recognition,2003,36(12): 3031-3033.

[3]Yong Xu, Jing-yu Yang, Zhong jin. A novel method for Fisher discriminant Analysis[J]. Pattern Recognition, 2004, 37 (2): 381-384.

[4] B. Scholkopf, S. Mika, CJC Burges, P. Knirsch, K.-R. Muller, G. Ratsch, and A. Smola, Input space vs. feature space in kernel-based methods. IEEE Transactions on Neural Networks , 10(5):1000-1017, 1999.

[5] S.A. Billings, K.L. Lee, Nonlinear Fisher discriminant analysis using a minimum square error cost function and the orthogonal least squares algorithm. Neural Networks, 15(1)(2002)263-270.

[6] Y. Xu, J.-Y. Yang, J.-F. Lu, An efficient kernel-based nonlinear regression method for two-class classification, Proceedings of 2005 International Conference on Machine Learning and Cybernetics, Guangzhou, China, August, 2005, pp.4442-4445.

[7] Y. Xu, D. Zhang, Z. Jin, M. Li, J.-Y. Yang, A fast kernel-based nonlinear discriminant analysis for multi-class problems, Pattern Recognition,2006, 39(6) : 1026-1033.

[8] Tipping M. E.. Sparse Bayesian Learning and the Relevance Vector Machine. Journal of Machine Learning Research, (1), (2001)211-244.

[9] SCHOLKOPF B, SMOLA A, MULLER K R. Nonlinear component analysis as a kernel eigenvalue problem[J]. Neural Computation, 1998,10(5):1299-1319.

[10] Mika S., Ratsch G., Weston J., etal. Fisher discriminant analysis with kernels[J]. In: IEEE Neural Networks for Signal Processing Workshop[C], Madison， IEEE Press, 1999, 41-48

[11] MULLER K R, S MIKA, RÄTSCH G, et al. An introduction to kernel-based learning algorithms[J]. IEEE Trans. On Neural Network, 2001,12(2):181-201.

[12] MIKA S, SMOLA A J, SCHÖLKOPF B. An improved training algorithm for kernel fisher discriminants[A]. In: T JAAKKOLA, T RICHARDSON, eds. Proceedings AISTATS[C], Morgan Kaufmann:2001

[13] S. Chen, S. A. Billings, and W. Luo, Nonlinear least squares methods and their application to nonlinear system identifications, International Journal of control, Vol. 50, No. 5: (1989)1873-1896

[14]S. Chen, E. S. Chng, K. Alkadhimi, Regularized orthogonal least squares algorithm for constructing radial basis function networks, Int. J. control, 64(5)(1996), 829-837

[15] J. R. Rice， Experiments on Gram-Schmidt Orthogonalization. Math. Comp. 20 (1966), 325-328.

[16] A. BjÄorck, Solving linear least squares problems by Gram-Schmidt orthogonalization. BIT
7 (1967), 1-21.

[17] B. Efron and R.J. Tibshirani. An Introduction to the Bootstrap. Chapman and Hall, 1993.

[18] M.J.L. Orr. Local Smoothing of Radial Basis Function Networks (long version). In International Symposium on Artificial Neural Networks, Hsinchu, Taiwan, 1995.