[10] L. Cao, "Practical method for determining the minimum embedding dimension of a scalar time series," *Physica D*, vol. 110, pp. 43–50, 1997.
[11] A. Savran, "Multifeedback-layer neural network," *IEEE Trans. Neural Netw.*, vol. 18, no. 2, pp. 373–384, Mar. 2007.
[12] Z. Hou, M. Gupta, P. Nikiforuk, M. Tan, and L. Cheng, "A recurrent neural network for hierarchical control of interconnected dynamic systems," *IEEE Trans. Neural Netw.*, vol. 18, no. 2, pp. 466–481, Mar. 2007.
[13] M. Chen, T. Gautama, M. M. Van Hulle, and D. P. Mandic, "On nonlinear modular neural filters," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2005, vol. 5, pp. 317–320.
[14] D. P. Mandic, J. Baltersee, and J. A. Chambers, "Non-linear adaptive prediction of speech with a pipelined recurrent neural network and advanced learning algorithms," in *Signal Analysis and Prediction*, A. Prochazka, J. Uhlir, P. W. Rayner, and N. G. Kingsbury, Eds. Boston, MA: Birkhauser, 1998, vol. 5.

Recursive Support Vector Machines for Dimensionality Reduction

Qing Tao, Dejun Chu, and Jue Wang

***Abstract*—The usual dimensionality reduction technique in supervised learning is mainly based on linear discriminant analysis (LDA), but it suffers from singularity or undersampled problems. On the other hand, a regular support vector machine (SVM) separates the data only in terms of one single direction of maximum margin, and the classification accuracy may be not good enough. In this letter, a recursive SVM (RSVM) is presented, in which several orthogonal directions that best separate the data with the maximum margin are obtained. Theoretical analysis shows that a completely orthogonal basis can be derived in feature subspace spanned by the training samples and the margin is decreasing along the recursive components in linearly separable cases. As a result, a new dimensionality reduction technique based on multilevel maximum margin components and then a classifier with high accuracy are achieved. Experiments in synthetic and several real data sets show that RSVM using multilevel maximum margin features can do efficient dimensionality reduction and outperform regular SVM in binary classification problems.**

***Index Terms*—Classification, dimensionality reduction, feature extraction, projection, recursive support vector machines (RSVMs), support vector machines (SVMs).**

I. Introduction

Dimensionality reduction is an important preprocessing step in many applications of data mining, machine learning, and pattern recognition, due to the so-called curse of dimensionality [1], [2]. Now, principal component analysis (PCA, [3]) and linear discriminant analysis (LDA, [4]) are regarded as the most fundamental and powerful tools of dimensionality reduction for extracting effective features from high-dimensional vectors of input data. From the point of view of mathematics, PCA is an orthogonal transformation of the coordinate system in which we describe our data. The new coordinate values by which we represent the data are called principal components. Usually, a small number of principal components is sufficient to account for most of the structure in the data. From the viewpoint of pattern recognition, LDA aims to find the optimal discriminant vectors (and then, an orthogonal transformation) by maximizing the ratio of the between-class distance to the within-class distance, thus achieving the maximum class discrimination. LDA is the benchmark for the linear discrimination between two classes in multidimensional space. One of the most obvious differences between PCA and LDA is that the former does not employ the labels of all samples while the latter does.

Around 1997, several comparative studies between LDA and PCA on the face recognition problems were reported independently by numerous authors [5], [6], in which LDA outperformed PCA significantly. So far, LDA has proven to be a more efficient approach for extracting features for many pattern classification problems as compared to PCA. However, there exists a serious limitation for using LDA to solve high-dimensional recognition with finite samples. Usually, LDA requires the so-called total scatter matrix to be nonsingular. In many applications, especially in face recognition, all scatter matrices in question can be singular since the data points are from a very high-dimensional space and, in general, the sample size does not exceed this dimensionality. This is known as the *singularity* or *undersampled problem* [8] and inevitably gives rise to a problem of unstable numerical computation. In recent years, many approaches have been proposed to deal with such high-dimensional undersampled problems, including null space LDA and orthogonal LDA, and their detailed computational and theoretical analysis can be seen in [9]. Recently, a recursive LDA for calculating the discriminant features was suggested in [10]. This new algorithm incorporates the same fundamental idea behind LDA of seeking the projection that best separates the data corresponding to different classes, while in contrast to regular LDA, the features are obtained recursively and the number of features that may be derived is independent of the number of the classes to be recognized. Extensive experiments of comparing the recursive LDA algorithm with the traditional approaches have been carried out on face recognition problems, in which the resulting improvement of the performances by the new feature extraction scheme is significant. Obviously, how to employ the recursive idea to get a dimensionality reduction approach without undersampled problems is very interesting.

In the last few years, there have been very significant developments in the understanding of support vector machines (SVMs) and statistical learning theory [11]–[13]. In appearance, the geometric interpretation of a linear SVM, known as the maximum margin algorithm, is very clear. In theory, increasing margin has been shown to improve the generalization performance. In [14], an SVM-like framework was established for LDA and it was proved that the general framework of LDA is based on the simplest and most intuitive LDA with zero within-class variance. Further, it can be found that LDA and SVM are closely related. Commonly, they all try to seek the projection that best separates the data in terms of a specific objective function. Along the former direction, the within-class variance is minimized while between-class variance is maximized. Along the latter, the between-class distance is maximized with the large margin while within-class distance is not considered. Since the usual dimensionality reduction technique in supervised learning is mainly based on using a small number of orthogonal

Manuscript received November 11, 2006; revised February 7, 2007; accepted July 2, 2007. This work was supported by the National Basic Research Program 2004CB318103 and the National Science Foundation of China under Grant 60575001.

Q. Tao is with the Key Laboratory of Complex Systems and Intelligence Science, Institute of Automation, Chinese Academy of Sciences, Beijing 100080, P.R. China and also with the New Star Research Institute of Applied Technology, Hefei 230031, P.R. China (e-mail: qing.tao@mail.ia.ac.cn; taoqing@gmail.com).

J. Wang is with the Key Laboratory of Complex Systems and Intelligence Science, Institute of Automation, Chinese Academy of Sciences, Beijing 100080, P.R. China (e-mail: jue.wang@mail.ia.ac.cn).

D. Chu is with the New Star Research Institute of Applied Technology, Hefei, 230031 P.R. China.

Digital Object Identifier 10.1109/TNN.2007.908267

LDA vectors, the theoretical analysis in [14] naturally enlightens us to use multilevel maximum margin features to reduce the dimensionality. As the computation of reverse of matrix is not concerned in SVM formulation, the fatal disadvantage of singularity problem can be avoided.

On the other hand, although it was reported that even 1-D SVM classifier could outperform LDA and PCA for several particular cases,[1] it may not be the case for most of the other two-class classification problems since it is too naive to believe that only one single direction of maximum margin would suffice for all. Therefore, it is desirable to eliminate this constraint completely if possible such that SVM can make full use of the multidimensional maximum margin. It is for the motivation of both of dimensionality reduction and accuracy improvement that we wish to suggest a recursive procedure for extracting multilevel margin features, recursive support vector machine (RSVM), which constitutes the main contribution of this letter. Our main idea is *recursively deriving new maximum margin features by discarding all the information represented by the old maximum margin features*. Compared with RLDA [10] in the viewpoints of mathematical transformations, a completely orthogonal basis of feature subspace spanned by the training samples can be derived. In fact, our RSVM dimensionality reduction approach is based on an intuitive idea that not a single but a small number of maximum margin components is sufficient to account for most of the differences in the classifications. Naturally, a terminate criterion, which is identical to that in PCA, is introduced and this is also different from that in RLDA [10].

In order to evaluate whether the proposed multidimensional maximum margin approach would bring any advantages over regular SVM, we choose to carry out experiments on SVM benchmark data sets.[2] All of the experimental results have demonstrated that RSVM can do efficient dimensionality reduction and the performance of SVM classifiers can be improved using the same parameters.

The remainder of this letter is arranged as follows. In Section II, some preliminaries about the relationship between SVM and LDA are stated. In Section III, the proposed RSVM algorithm is described and discussed in detail. Several numerical classification examples are given in Section IV. Finally, Section V concludes this letter with a brief remark on conclusions and future work.

II. SVM AND LDA

Basically, a regular SVM considers the following two-category classification problem:

$$(x_1, y_1), (x_2, y_2), \ldots, (x_l, y_l) \in R^N \times Y, Y = \{-1, 1\} \quad (1)$$

where x_i is independently drawn and identically distributed, and y_i is the label of x_i. For linearly separable cases, an SVM classifier attempts to optimize the generalization bound by separating data with a maximal margin. Geometrically speaking, the margin of a classifier is the minimal distance of training points from the decision boundary. The maximal margin classifier is the one with the maximum distance from the *nearest patterns* to the boundary, called *support vectors* [12].

For linearly separable cases, we can assume $y_i(w^T x_i + b) \geq 1$, $1 \leq i \leq l$. Thus, the margin is $2/\|w\|$ and the maximum margin algorithm can then be described as follows:

$$\begin{cases} \min \dfrac{1}{2}\|w\|^2 \\ y_i(w^T x_i + b) \geq 1, \qquad 1 \leq i \leq l. \end{cases} \quad (2)$$

For LDA, we assume that there exist w and b such that $y_i(w^T x + b) = 1, 1 \leq i \leq l$. In such a simple case, the within-class distance can be regarded as zero and the between-class distance is $2/\|w\|$. Thus, the intuitive Fisher classifier $w^T x + b = 0$ satisfies

$$\begin{cases} \min \dfrac{1}{2}\|w\|^2 \\ y_i(w^T x_i + b) = 1, \qquad 1 \leq i \leq l. \end{cases} \quad (3)$$

Equation (3) works only under the specific assumption. Nonetheless, it is the easiest algorithm to understand, and it is proved to form the main building block for more complex LDA and exhibits the key features that characterize the complete framework of LDA [14]. From (2) and (3), it is easy to find out that SVM and LDA are closely related in that a 1-D direction is sought for classification but in terms of different objective functions. Since multidimensional LDA plus k-nearest neighbor methods are widely and effectively applied in face recognition, the deep relationship between SVM and LDA naturally motivates us to incorporate multidimensional idea with SVM. Specifically in this letter, we employ the recursive method in [10] to do binary classifications.

In many real applications, the classification problems are usually linearly inseparable. To get a classifier with good performance, we often solve the dual optimization problems for nonlinear soft margin algorithms and adjust the parameters using cross-validation strategy. By introducing slack variables to relax the hard margin constraints, the following optimization problem for a soft margin is proposed by Cortes and Vapnik [15] and Vapnik [11] and [12]:

$$\begin{cases} \min \dfrac{1}{2}\|w\|^2 + C \sum\limits_{i=1}^{l} \xi_i \\ \xi_i \geq 0,\ y_i(w^T \phi(x_i) + b) \geq 1 - \xi_i, \qquad 1 \leq i \leq l \end{cases} \quad (4)$$

where C is a predefined positive real number, ϕ is a feature map associated with kernel k, and ξ_i's are slack variables.

The dual optimization problem of (4) is [15], [11], [12]

$$\begin{cases} \max \sum\limits_{i=1}^{l} \alpha_i - \dfrac{1}{2} \sum\limits_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j k(x_i, x_j) \\ C \geq \alpha_i \geq 0, \qquad 1 \leq i \leq l; \qquad \sum\limits_{i=1}^{l} \alpha_i y_i = 0. \end{cases} \quad (5)$$

III. RSVM

The recursive LDA algorithm in [10] consists of the following two steps: 1) determine the discriminant direction for separating different classes by maximizing the generalized Rayleigh quotient and 2) generate a new sample set by projecting the samples into a subspace that is orthogonal to the direction calculated in 1) and continue with step 2). Similarly in appearance, our RSVM works as follows.

---

1. Determine the vector $\tilde{w}_1 = \sum_{i=1}^{l} \alpha_i^1 \phi(x_i)$ by solving problem (5).

2. Let $w_{t-1} = \tilde{w}_{t-1}/\|\tilde{w}_{t-1}\|$ and generate the following training set for SVM problem (5) by projecting the samples into a subspace that is orthogonal to w_{t-1}:

$$\phi(x_i^t) = \phi(x_i^{t-1}) - \langle \phi(x_i^{t-1}), w_{t-1} \rangle w_{t-1}, \qquad 1 \leq i \leq l. \quad (6)$$

3. Terminate if $\max\{\|\phi(x_i^t)\| : 1 \leq i \leq l\} < \epsilon$.

4. Increment t by 1 and go back to Step 2.

To clearly show that the previous recursive algorithm can conduct efficient dimensionality reduction by using the multilevel margin components, the following three theorems are given.

[1] See http://ida.first.fhg.de/projects/bench/benchmarks.htm

[2] See http://ida.first.fhg.de/projects/bench/benchmarks.htm

Theorem 1: Let w_1 and w_2 be the solutions of the corresponding optimization problems in the aforementioned RSVM algorithm, then w_1 is orthogonal to w_2.

Proof: From (6), $\phi(x_i^2) = \phi(x_i) - \langle\phi(x_i), w_1\rangle w_1$, $i = 1, 2, \ldots, l$. Then, $\langle\phi(x_i^2), w_1\rangle = \langle\phi(x_i), w_1\rangle - \langle\phi(x_i), w_1\rangle\langle w_1, w_1\rangle = 0$.

From SVM, w_2 is a linear combination of $\phi(x_i^2)$, $i = 1, 2, \ldots, l$.

Therefore, $\langle w_1, w_2\rangle = 0$ and the proof is completed.

Similarly, it can be proved that w_i and w_j $(i \neq j)$ are orthogonal.

Theorem 2: Each $\|\phi(x_i^t)\|$ is decreasing about t and there exists a positive number $m \leq l$ such that $\|\phi(x_i^m)\| = 0$, $i = 1, 2, \ldots, l$.

Proof: From Theorem 1, it is not difficult to get

$$\begin{aligned}\|\phi(x_i^t)\|^2 &= \|\phi(x_i^{t-1})\|^2 - (\langle\phi(x_i^{t-1}), w_{t-1}\rangle)^2 \\ &= \|\phi(x_i)\|^2 - \sum_{j=1}^{t-1}(\langle\phi(x_i), w_j\rangle)^2.\end{aligned}$$

This means that $\|\phi(x_i^t)\|$ is decreasing about t.

Let $S = span\{\phi(x_1), \phi(x_2), \ldots, \phi(x_l)\}$. Obviously, S is a finite-dimensional space.

Recursively from optimization problem (5), w_1 is a linear combination of $\phi(x_i)$, $i = 1, 2, \ldots, l$. Easily, $w_1 \in S$.

From (6), $\phi(x_i^t) \in S$ for each t and i.

Since the dimension of S is at most l, a completely orthogonal basis of S can be obtained by RSVM and there must exist a positive number $m \leq l$ such that

$$\|\phi(x_i^t)\|^2 = \|\phi(x_i)\|^2 - \sum_{j=1}^{m}(\langle\phi(x_i), w_j\rangle)^2 = 0.$$

The proof is now completed.

If the aforementioned RSVM algorithm works using the terminate criterion $\max\{\|\phi(x_i^t)\| : 1 \leq i \leq l\} = 0$, Theorems 1 and 2 imply that a completely orthogonal basis of feature subspace spanned by $\{\phi(x_1), \phi(x_2), \ldots, \phi(x_l)\}$ can be derived. Intuitively, the dimensionality reduction can be carried out by using partial elements of the acquired orthogonal basis. Theorem 2 also implies that $\|\phi(x_i^m)\| < \epsilon$ if m is sufficiently large. Therefore, the RSVM algorithm will terminate within finite steps. In fact, $\max\{\|\phi(x_i^t)\| : 1 \leq i \leq l\} < \epsilon$ in our RSVM is a criterion for evaluating the extent of approximation which is identical to that in PCA. This theoretically ensures that RSVM is an approximation method in terms of margin–direction transformation. Further, RSVM dimensionality reduction is based on the viewpoint that only one single maximum margin direction is not enough but a small number of maximum margin directions are sufficient to account for most of the differences in the classifications.

Note in [10] that the recursive LDA process naturally stops when the between-class scatter is zero and cannot be further maximized by projection. Obviously, the disadvantage is that all the discriminate vectors found by RLDA may not form a complete basis of even finite-dimensional feature space. It means that RLDA-based dimensionality reduction may not be an orthogonal transformation. At this point, RLDA and RSVM are different.

Theorem 3: Let $\{\tilde{w}_1, b_1, \xi_1^1, \xi_2^1, \ldots, \xi_l^1\}$ and $\{\tilde{w}_2, b_2, \xi_1^2, \xi_2^2, \ldots, \xi_l^2\}$ be the solutions of the corresponding optimization problems in the previously described RSVM algorithm, then

$$\frac{1}{2}\|\tilde{w}_2\|^2 + C\sum_{i=1}^{l}\xi_i^2 \geq \frac{1}{2}\|\tilde{w}_1\|^2 + C\sum_{i=1}^{l}\xi_i^1.$$

Proof: Obviously, $\{\tilde{w}_2, \xi_1^2, \xi_2^2, \ldots, \xi_l^2\}$ satisfies the following constraints:

$$\xi_i^2 \geq 0,\ y_i(\tilde{w}_2^T\phi(x_i^2) + b_2) \geq 1 - \xi_i^2, \qquad i = 1, 2, \ldots, l.$$

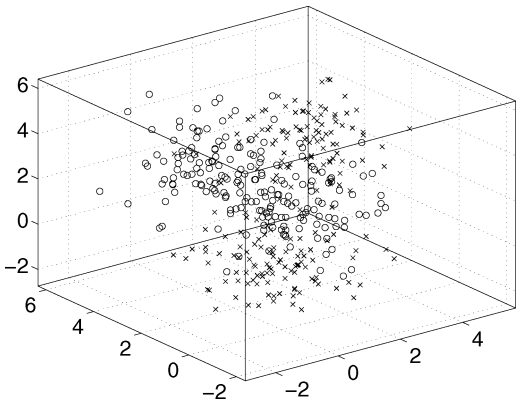

Fig. 1. Toy data in 3-D space.

By (6) and Theorem 1

$$y_i(\tilde{w}_2^T\phi(x_i^2) + b_2) = y_i(\tilde{w}_2^T\phi(x_i) + b_2) \geq 1 - \xi_i^2.$$

This means that $\{\tilde{w}_2, b_2, \xi_1^2, \xi_2^2, \ldots, \xi_l^2\}$ also satisfies the constraints of optimization problem (4) for training set $(\{(\phi(x_1), y_1), (\phi(x_2), y_2), \ldots, (\phi(x_l), y_l)\}$. Since $\{\tilde{w}_1, b_1, \xi_1^1, \xi_2^1, \ldots, \xi_l^1\}$ is the optimal solution, thus

$$\frac{1}{2}\|\tilde{w}_2\|^2 + C\sum_{i=1}^{l}\xi_i^2 \geq \frac{1}{2}\|\tilde{w}_1\|^2 + C\sum_{i=1}^{l}\xi_i^1$$

and the proof is completed.

If the considered classification problem is linearly separable, Theorem 3 tells us that the margin decided by w_i is decreasing when the index $i, 1 \leq i \leq n$, increases. Similarly, like the description of components in PCA, the vector found by SVM about the original samples $(\phi(x_1), y_1), (\phi(x_2), y_2), \ldots, (\phi(x_l), y_l)$ can naturally be regarded as the first margin component and the other levels of margin components can be similarly defined.

The following remarks are helpful to understand the implementation of the proposed RSVM.

- All the involved inner product computations can be based on kernel evaluation instead of the explicit $\phi(x_i)$, which may be infinite-dimensional. For example, $\langle\phi(x_j), w_1\rangle$ can be computed by $\sum_{i=1}^{l}\alpha_i^1 k(x_j, x_i)$, where $w_1 = \sum_{i=1}^{l}\alpha_i^1\phi(x_i)$ is a solution of (5) for training points $(x_1, y_1), (x_2, y_2), \ldots, (x_l, y_l)$.
- Note that only $k(x, y)$ computation is concerned in optimization problem (5). To solve (5) about the new generated samples $\{x_1^{t-1}, x_2^{t-1}, \ldots, x_l^{t-1}\}$, $k(x_i^t, x_j^t)$ can be recursively computed by using (6) and $k(x_i^{t-1}, x_j^{t-1})$.
- All the involved norm computations can be based on kernel evaluations instead of the explicit $\phi(x_i)$, which may be infinite-dimensional. For example, $\|\phi(x_i) - \phi(x_j)\|$ can be computed by $\sqrt{k(x_i, x_i) - 2k(x_i, x_j) + k(x_j, x_j)}$.

IV. Experiments

Example 1: To intuitively illustrate the efficiency of multilevel features for linear dimensionality reduction, the following synthetic example is designed. As can be seen in Fig. 1, four 100-points are generated, respectively, by the normal distribution with mean vector $(0, 0, 0)^T$, $(3, 3, 3)^T$, $(3, 0, 3)^T$, and $(3, 3, 0)^T$ and covariance matrix

$$\begin{bmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{bmatrix}.$$

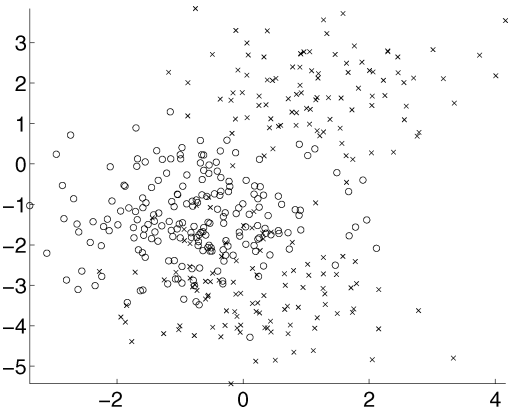

Fig. 2. Projected data in 2-D space using linear RSVM with $C = 11$.

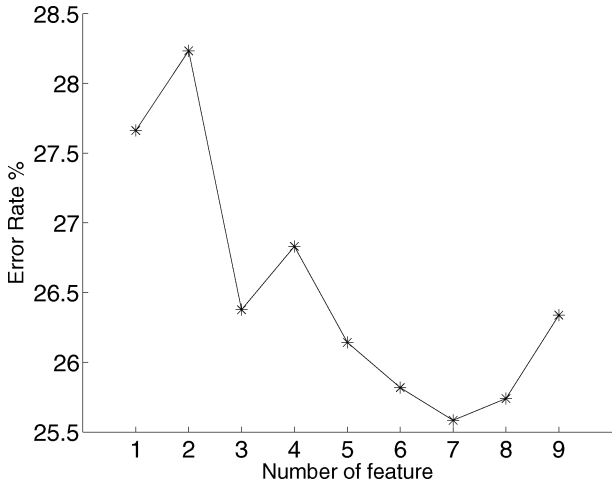

Fig. 3. Relationship between the number of feature and classification accuracy on breast cancer data set.

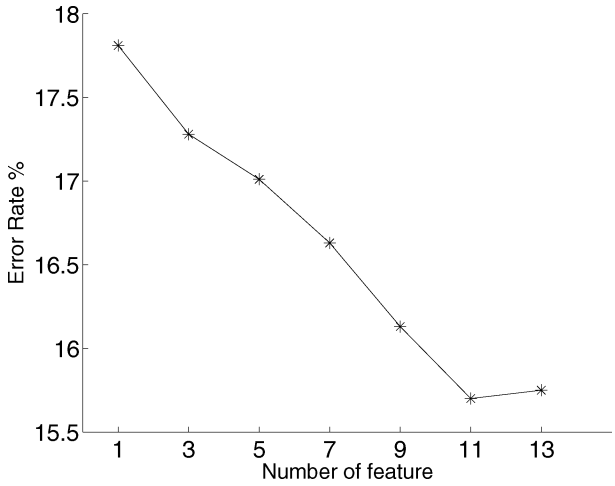

Fig. 4. Relationship between the number of feature and classification accuracy on heart data set.

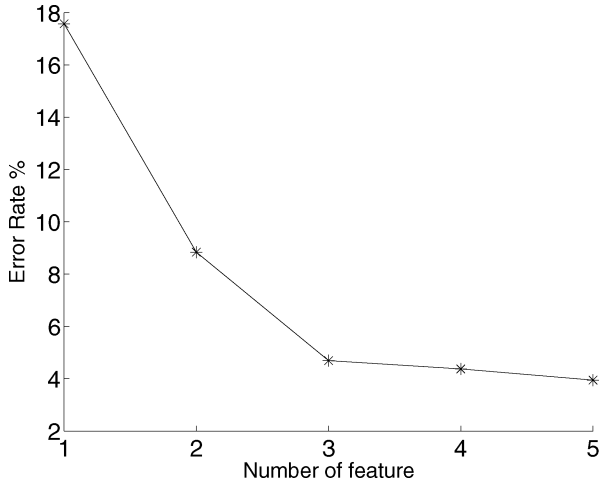

Fig. 5. Relationship between the number of feature and classification accuracy on thyroid data set.

Each sample in the first two parts has label 1 and the left has label -1. From Fig. 2, it is easy to find out that not a single but two of the maximum margin components plus k-nearest neighbor methods are sufficient to account for most of the differences in the classifications.

Example 2: To show the relationship between the number of feature and classification accuracy, several real binary classification problems are conducted.[3] For the convenience of comparison, RSVM adopts the same parameters as that in SVM. In the previously described experiments, we first select different number of features using RSVMs and then do the classification employing k-nearest neighbor methods. The relationship between the number of features and classification accuracy on several data sets are illustrated in Figs. 3–5. The optimal dimensions with their average test errors are summarized in Table I, where m is the number of feature and k is the number of nearest neighbors. To certain extent, this example demonstrates that multidimensional margin methods with the same parameters can produce better results than one single direction margin SVM.

Example 3: To further demonstrate the performance of RSVM for dimensionality reduction and against regular SVM, several other real classification problems are conducted.[4] In these binary classification experiments, the same comparison strategy as that in Example 2 is adopted. The only difference is that we set $\varepsilon = 1e - 6$ in Step 3 of RSVM without fixing the dimension. The dimension of the original data in ringnorm, twonorm, and waveform is 20, 20, and 21, respectively. After RSVM is used, the maximum numbers of feature on these training data sets are 8, 4, and 5 respectively. Even with such a small ε, it is not difficult to find that RSVMs can still conduct efficient dimensionality reduction. Table II presents the average classification test errors by SVMs and RSVMs. This example also indicates that multidimensional margin methods with the same parameters can outperform regular SVMs.

Note that it is highly recommended that Mahalanobis-distance-based nearest neighbors methods be more efficient in the feature reduced space [16]. However, the comparison study between the Mahalanobis

TABLE I
AVERAGE TEST ERRORS ERROR RATE WITH OPTIMAL DIMENSION

Data Sets	SVM	RSVM(m, k)	Decreasing
Breast cancer	26.04 $\pm$ 4.74	25.58 $\pm$ 4.34($m = 7$,k=5)	0.46
Heart	15.95$\pm$3.26	15.70$\pm$3.34($m = 11$,k=3)	0.25
Thyroid	4.80$\pm$2.19	3.95$\pm$2.47($m = 5$,k=1)	0.85

[3]The data sets, cross-validation strategy, test errors, and kernel functions of SVM algorithms are taken from http://ida.first.fhg.de /projects /bench/ benchmarks.htm

[4]Taken from http://ida.first.fhg.de/projects/bench/benchmarks.htm

TABLE II
AVERAGE TEST ERRORS ERROR RATE (IN PERCENT) WITH $\varepsilon = 1e-6$

Data Sets	SVM	RSVM(k)	Decreasing
Ringnorm	1.66 $\pm$ 0.12	1.48 $\pm$ 0.11(k=5)	0.18
Twonorm	2.96$\pm$0.23	2.40$\pm$0.12(k=4)	0.56
Waveform	9.88$\pm$0.43	9.48$\pm$0.53(k=4)	0.40

and Euclidean distances is inconclusive from our experiments, and we then decided to only report the results by Euclidean distance. It should also be pointed out that better classification accuracy can be achieved if we search all the optimal parameters in RSVM using the cross-validation strategy.

V. CONCLUSION

In this paper, a multidimensional maximum margin feature extraction approach for constructing a completely orthogonal basis and thus conducting efficient dimensionality reduction, called RSVM, is presented. Theoretical analysis shows that the SVM objective function is decreasing along the recursive components. In contrast to PCA, we use supervised information (labels) to conduct dimensionality reduction. Compared with LDA and regular SVM, the proposed method has no singularity problems and can further improve the accuracy. The general multilevel margin direction idea in this letter can be easily extended to SVM regression and several weighted SVM cases [17], [18] helping us to achieve more accurate results. Our future work will focus on using the recursive and multidimensional maximum margin idea to solve multiclassifications, especially face recognition problems. it may be that a new representing and recognizing approach for face patterns can be expected.

ACKNOWLEDGMENT

The authors would like to thank the Associate Editor and the referees for their valuable comments.

REFERENCES

[1] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York: Wiley, 2001.
[2] T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. New York: Springer-Verlag, 2001.
[3] I. T. Jolliffe, *Principal Component Analysis*. New York: Springer-Verlag, 1986.
[4] K. Fukunaga, *Introduction to Statistical Pattern Classification*. San Diego, CA: Academic, 1990.
[5] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 711–720, Jul 1997.
[6] D. L. Swets and J. Weng, "Using discriminant eigenfeatures for image retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 8, pp. 831–836, Aug. 1996.
[7] X. Wang and X. Tang, "A unified framework for subspace face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 9, pp. 1222–1228, Sep. 2004.
[8] W. J. Krzanowski, P. Jonathan, W. V. McCarthy, and M. R. Thomas, "Discriminant analysis with singular covariance matrices: Methods and applications to spectroscopic data," *Appl. Statist.*, vol. 44, pp. 101–115, 1995.
[9] J. Ye and T. Xiong, "Computational and theoretical analysis of null space and orthogonal linear discriminant analysis," *J. Mach. Learn. Res.*, vol. 7, pp. 1183–1204, 2006.
[10] C. Xiang, X. A. Fan, and T. H. Lee, "Face recognition using recursive fisher linear discriminant," *IEEE Trans. Image Process.*, vol. 15, no. 8, pp. 2097–2105, Aug. 2006.
[11] V. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer-Verlag, 1995.
[12] V. Vapnik, *Statistical Learning Theory*. Reading, MA: Addison-Wiley, 1998.
[13] N. Cristianini and J. Schawe-Taylor, *An Introduction to Support Vector Machines*. Cambridge, U.K.: Cambridge Univ. Press, 2000.
[14] Q. Tao, G. Wu, and J. Wang, "The theoretical analysis of FDA and applications," *Pattern Recognit.*, vol. 39, no. 6, pp. 1199–1204, 2006.
[15] C. Cortes and V. Vapnik, "Support vector networks," *Mach. Learn.*, vol. 20, pp. 273–297, 1995.
[16] C. Liu and H. Wechsler, "Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition," *IEEE Trans. Image Process.*, vol. 11, no. 4, pp. 467–476, Apr. 2002.
[17] Q. Tao, G. Wu, F. Y. Wang, and J. Wang, "Posterior probability support vector machines for unbalanced data," *IEEE Trans. Neural Netw.*, vol. 16, no. 6, pp. 1561–1573, Nov. 2005.
[18] Q. Tao and J. Wang, "A new fuzzy support vector machine based on the weighted margin," *Neural Process. Lett.*, vol. 20, pp. 139–150, 2004.

A Forward-Constrained Regression Algorithm for Sparse Kernel Density Estimation

Xia Hong, Sheng Chen, and Chris J. Harris

***Abstract*—Using the classical Parzen window (PW) estimate as the target function, the sparse kernel density estimator is constructed in a forward-constrained regression (FCR) manner. The proposed algorithm selects significant kernels one at a time, while the leave-one-out (LOO) test score is minimized subject to a simple positivity constraint in each forward stage. The model parameter estimation in each forward stage is simply the solution of jackknife parameter estimator for a single parameter, subject to the same positivity constraint check. For each selected kernels, the associated kernel width is updated via the Gauss–Newton method with the model parameter estimate fixed. The proposed approach is simple to implement and the associated computational cost is very low. Numerical examples are employed to demonstrate the efficacy of the proposed approach.**

***Index Terms*—Cross validation, jackknife parameter estimator, Parzen window (PW), probability density function (pdf), sparse modeling.**

I. INTRODUCTION

The estimation of the probability density function (pdf) from observed data samples is a fundamental problem in many machine learning and pattern recognition applications [1]–[3]. The Parzen window (PW) estimate is a simple yet remarkably accurate nonparametric density estimation technique [2]–[4]. A general and powerful approach to the problem of pdf estimation is the finite mixture model [5]. The finite mixture model includes the PW estimate as a special case in that equal weights are adopted in the PW, with the number of mixtures equal to the number of training data samples. A disadvantage associated with the PW estimate is its high computational cost of the point density estimate for a future data sample in the cases whereby the training data set is very large. Clearly, by taking a much smaller

Manuscript received January 12, 2007; revised April 20, 2007 and May 25, 2007; accepted July 2, 2007.

X. Hong is with the School of Systems Engineering, University of Reading, Hampshire RG6 6AY, U.K. (e-mail: x.hong@reading.ac.uk).

S. Chen and C. J. Harris are with the School of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, U.K.

Digital Object Identifier 10.1109/TNN.2007.908645