Expert Systems with Applications 38 (2011) 11796–11803

Contents lists available at ScienceDirect

Expert Systems with Applications

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

A new hybrid method based on local fisher discriminant analysis and support vector machines for hepatitis disease diagnosis

Hui-Ling Chen, Da-You Liu *, Bo Yang, Jie Liu, Gang Wang

College of Computer Science and Technology, Jilin University, Changchun 130012, China
Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun 130012, China

ARTICLE INFO

Keywords:
Hepatitis disease diagnosis
Local fisher discriminant analysis
Support vector machines
Feature extraction

ABSTRACT

In this paper, a novel hybrid method named the LFDA_SVM, which integrates a new feature extraction method and a classification algorithm, has been introduced for diagnosing hepatitis disease. The two integrated methods are the local fisher discriminant analysis (LFDA) and the supporting vector machine (SVM), respectively. In the proposed LFDA_SVM, the LFDA is employed as a feature extraction tool for dimensionality reduction in order to further improve the diagnostic accuracy of the standard SVM algorithm. The effectiveness of the LFDA_SVM has been rigorously evaluated against the hepatitis dataset, a benchmark dataset, from UCI Machine Learning Database in terms of classification accuracy, sensitivity and specificity respectively. In addition, the proposed LFDA_SVM has been compared with three existing methods including the SVM based on principle component analysis (PCA_SVM), the SVM based on fisher discriminant analysis (FDA_SVM) and the standard SVM in terms of their classification accuracy. Experimental results have demonstrated that the LFDA_SVM greatly outperforms other three methods. The best classification accuracy (96.77%) obtained by the LFDA_SVM is much higher than that of the compared ones. Promisingly, the proposed LFDA_SVM might serve as a new candidate of powerful methods for diagnosing hepatitis with excellent performance.

© 2011 Elsevier Ltd. All rights reserved.

1. Introduction

The liver is the heaviest organ in the body and is one of the largest. The main functions of the liver are to process nutrients from food, make bile, remove toxins from the body and build proteins. It is easy to see how inflammation of the liver, or hepatitis, interferes with these important functions and can lead to poor health. Hepatitis is now one of the most important causes of chronic liver disease in the world, and millions of people are at risk for its complications. (http://hepatitis.about.com/od/overview/a/numbers.htm, last accessed December 2009). People who have the liver disease can be characterized by four common symptoms: jaundice, loss of appetite, fatigue and muscle and joint aches.

Expert systems and machine learning techniques are increasingly introduced to help the medical diagnosis. With the help of diagnostic systems, the possible errors experts made in the course of diagnosis can be avoided, and the medical data can be examined in shorter time and more detailed as well. Feature extraction as an important component of a pattern recognition system has been greatly used in medical diagnostic systems (Dogantekin, Dogantekin, & Avci, in press, 2009; Polat & Gunes, 2007a, 2007b, 2008). It performs two tasks: transforming input parameter vector into a feature vector and reducing its dimensionality. Two popular methods for feature extraction are fisher discriminant analysis (FDA) and principal component analysis (PCA) (Duda, Hart, & Stork, 2001). Both of them extract features by projecting the original parameter vectors into a new feature space through a linear transformation matrix. PCA seeks to find the largest variations in the original feature space, while FDA pursues the largest ratio of between-class variation and within-class variation when projecting the original feature to a subspace. In all, a well-defined feature extraction algorithm makes the classification process more effective and efficient.

1.1. Related work

Many feature extraction methods have been proposed to deal with the automated diagnosis of hepatitis disease problem, and most of them have achieved high classification accuracies. In Polat and Gunes (2007a, 2007b), an artificial immune recognition system (AIRS) based on principal component analysis (PCA) via 10-fold cross-validation was used for classification, the reported accuracy

* Corresponding author at: Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun 130012, China.
E-mail address: dillon5291@gmail.com (D.-Y. Liu).

0957-4174/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2011.03.066

was up to 94.12%. In Dogantekin et al. (2009), an adaptive network based on fuzzy inference system combining with linear discriminant analysis (LDA-ANFIS) was applied for automatic hepatitis diagnosis, and an accuracy of 94.16% was obtained. Ster and Dobnikar (1996) have obtained 86.4%, 85.3%, 83.2%, respectively, using LDA (linear discriminant analysis), QDA (quadratic discriminant analysis), Fisher discriminant analysis.

Aiming at improving the efficiency and effectiveness of the classification accuracy for hepatitis disease diagnosis, in this study, a new feature extraction method, local fisher discriminant analysis (LFDA) (Sugiyama, 2007) is examined. LFDA is an extension of conventional fisher discriminant analysis (FDA), which localizes the evaluation of the within-class scatter, and thus works well even when within-class multimodality or outliers exist. In addition, LFDA overcomes a critical limitation of the original FDA in dimensionality reduction, namely the dimension of the FDA embedding space should be less than the number of classes, while LFDA does not suffer from this restriction in general. The objective of the proposed method is to explore the performance of hepatitis diagnosis using a two-stage hybrid modeling procedure in integrating LFDA with SVM. SVM as a relatively new machine learning technique was first introduced by Vapnik (1995, 1998). It has been proven advantageous in handling classification tasks with excellent generalization performance (Cortes & Vapnik, 1995; Joachims, Nedellec, & Rouveirol, 1998; Osuna, Freund, & Girosi, 1997).

The rationale underlying the proposed method (LFDA_SVM) is firstly to use LFDA in reducing the dimension of the hepatitis dataset, and then the obtained reduced feature subset is served as the input into the designed SVM classifier. The effectiveness of LFDA_SVM is examined in terms of classification accuracies, sensitivity and specificity. Moreover, the superior classification capability of the proposed method can be observed by comparing the results with those using SVM based on PCA (PCA_SVM), SVM based on FDA (FDA_SVM) and the standard SVM. Experimental results have shown that LFDA_SVM outperforms the other methods significantly and has achieved the best predicative classification accuracy with the reduced feature subset.

1.2. The organization of the paper

The remainder of this paper is organized as follows. Section 2 offers brief background knowledge on local fisher discriminant analysis and supporting vector machines. The research design is described in Section 3. Section 4 presents the experimental results and discussion of the proposed method. Finally, Conclusions and recommendations for future work are summarized in Section 5.

2. The theoretical backgrounds of the related methodologies

2.1. Fisher discriminant analysis

Fisher discriminant analysis (FDA) is a well-known linear technique for reducing dimensions and pattern classification. It seeks to find the Fisher optimal discriminant vectors by maximizing the scatter between the classes while minimizing the scatter within each class (Duda et al., 2001).

Let $x_i \in \Re^d$ $(i = 1, 2, \ldots, n)$ be d-dimensional samples and $y_i \in \{1, 2, \ldots, c\}$ be associated class labels, where n is the number of samples and c is the number of classes. Let n_l be the number of samples in class l : $\sum_{l=1}^{c} n_l = n$. Let $z_i \in \Re^r (1 \leqslant r \leqslant d)$ be low-dimensional representations of x_i, where r is the reduced dimension. In the cases of linear dimensionality reduction, using a $d \times r$ transformation matrix T, the reduced sample z_i can be given by $z_i = T^T x_i$, where A^T denotes the transpose of the matrix A. Let $S^{(w)}$ and $S^{(b)}$ be the within-class scatter matrix and the between-class scatter matrix, respectively (Sugiyama, 2007):

$$S^{(w)} = \sum_{l=1}^{c} \sum_{i:y_i=l} (x_i - \mu_l)(x_i - \mu_l)^T, \tag{1}$$

$$S^{(b)} = \sum_{l=1}^{c} n_l (\mu_l - \mu)(\mu_l - \mu)^T, \tag{2}$$

where $\sum_{i:y_i=l}$ denotes the summation over i such that $y_i = l$, μ_l is the mean of the samples in class l, and μ is the mean of all samples:

$$\mu_l = \frac{1}{n_l} \sum_{i:y_i=l} x_i, \tag{3}$$

$$\mu = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{1}{n} \sum_{l=1}^{c} n_l \mu_l. \tag{4}$$

Here, we assume that $S^{(w)}$ has full rank. The FDA transformation matrix $T^{(FDA)}$ is defined as

$$T^{(FDA)} = \arg\max_{T \in \mathbb{R}^{d \times r}} \left[tr(T^T S^{(b)} T (T^T S^{(w)} T)^{-1}) \right]. \tag{5}$$

That is, FDA seeks a transformation matrix T such that the between-class scatter is maximized while the within-class scatter is minimized. In the above formulation, the within-class scatter in the embedding space, denoted as $T^T S^{(w)} T$, is assumed to be invertible. This implies that the above optimization is subject to

$$rank(T) = r. \tag{6}$$

Let $\{\varphi_k\}_{k=1}^{d}$ be the generalized eigenvectors associated with the generalized eigenvalues $\lambda_1 \geqslant \cdots \geqslant \lambda_d$ of the following generalized eigenvalue problem:

$$S^{(b)} \varphi = \lambda S^{(W)} \varphi. \tag{7}$$

Then a solution $T^{(FDA)}$ of the above maximization problem is analytically given by

$$T^{(FDA)} = (\varphi_1 | \varphi_2 | \cdots | \varphi_r). \tag{8}$$

2.2. Local fisher discriminant analysis

Local fisher discriminant analysis (LFDA) is a new linear supervised dimensionality reduction method proposed by Sugiyama (2007). It can be considered to be a natural localized variant of fisher discriminant analysis, which seeks to maximize between-class separability and preserves within-class local structure at the same time. Since the LFDA evaluates the levels of the between-class scatter and the within-class scatter in a local manner, it works well even when within-class multimodality or outliers exist. LFDA was shown to compare favorably with other supervised dimensionality reduction methods through extensive experiments in Sugiyama (2007).

Let $S^{(lb)}$ and $S^{(lw)}$ be the local between-class scatter matrix and the local within-class scatter matrix, which are respectively defined as (Sugiyama, Ide, Nakajima, & Sese, 2010):

$$S^{(lb)} = \frac{1}{2} \sum_{i,j=1}^{n'} W_{i,j}^{(lb)} (x_i - x_j)(x_i - x_j)^T, \tag{9}$$

$$S^{(lw)} = \frac{1}{2} \sum_{i,j=1}^{n'} W_{i,j}^{(lw)} (x_i - x_j)(x_i - x_j)^T, \tag{10}$$

where $W^{(lb)}$ and $W^{(lw)}$ are the $n' \times n'$ matrices with

$$W_{i,j}^{(lb)} = \begin{cases} A_{i,j}(1/n' - 1/n'_{yi}) & \text{if } y_i = y_j, \\ 1/n' & \text{if } y_i \neq y_j, \end{cases} \tag{11}$$

$$W_{i,j}^{(lw)} = \begin{cases} A_{i,j}/n'_{yi} & \text{if } y_i = y_j, \\ 0 & \text{if } y_i \neq y_j, \end{cases} \tag{12}$$

where n'_{yi} denotes the number of labeled samples in class $y_i \in \{1, 2, \ldots, c\}$. $A_{i,j}$ is the affinity value between x_i and x_j based on the local scaling heuristic(Zelnik-Manor & Perona, 2004), which is defined as follow:

$$A_{i,j} = \exp\left(-\frac{\|x_i - x_j\|^2}{\sigma_i \sigma_j}\right). \tag{13}$$

the parameter σ_i represents the local scaling around x_i defined by

$$\sigma_i = \|x_i - x_i^{(k)}\|. \tag{14}$$

where $x_i^{(k)}$ is the kth nearest neighbor of x_i.

And the local scaling is computed in a classwise manner in order to preserve the within-class local structure (Sugiyama, 2007). According to the affinity $A_{i,j}$, the values for the sample pairs in same class were weighted. The LFDA transformation matrix $T^{(LFDA)}$ is defined as:

$$T^{(LFDA)} \equiv \underset{T \in \mathbb{R}^{d \times r}}{\arg\max} \left[tr(T^T S^{(lb)} T (T^T S^{(lw)} T)^{-1}) \right]. \tag{15}$$

In other words, LFDA seeks a transformation matrix T such that the local between-class scatter in the embedding space, i.e. $T^T S^{(lb)} T$, will be maximized, and the local within-class scatter in the embedding space, i.e. $T^T S^{(lw)} T$, will be minimized. The solution of Eq. (13) is equivalent to solving a generalized eigenvalue problem of $S^{(b)}$ and $S^{(w)}$. When $A_{i,j} = 1$ for all sample pairs, $S^{(lb)}$ and $S^{(lw)}$ can be reduced to $S^{(b)}$ and $S^{(w)}$. Thus, LFDA can be regarded as a localized variant of the FDA.

2.3. Support vector machines for classification

Support vector machine (SVM), originally developed by Boser, Guyon, and Vapnik (1992), Vapnik (1995), is based on the Vapnik–Chervonenkis (VC) theory and structural risk minimization (SRM) principle (Vapnik, 1995, 1998), which is known to have high generalization performance. Another key feature of SVM is that training SVM is equivalent to solving a linear constrained quadratic programming problem. Thus it is unlikely to be trapped in the local minimum (Cristianini & Shawe-Taylor, 2000). For more details, one can refer to (Cristianini & Shawe-Taylor, 2000; Vapnik, 1995), which give a complete description of the SVM theory. In this section we will be concentrated on the basic SVM concepts for typical binary-classification problems.

2.3.1. Linearly separable case – hard margin SVM

Let us consider a binary classification task:$\{x_i, y_i\}$, $i = 1, \ldots l$, $y_i \in \{-1, 1\}$, $x_i \in R^d$, where x_i are data points and y_i are corresponding labels. They are separated with a hyper plane given by $w^T x + b = 0$, where w is a d-dimensional coefficient vector which is normal to the hyper plane and b is the offset from the origin.

There will be many hyper planes that can separate the two classes, whereas the decision boundary should be as far away from the data of both classes as possible. The support vector algorithm aims to look for an optimal separating hyper plane that will maximize the separating margin between the two classes of data since the wider margin can achieve the better generalization ability. We can define a canonical hyper plane (Vapnik, 1995) such that $H_1 : \mathbf{w}^T \mathbf{x}^+ + b = 1$ for the closet points on one side and $H_2 : \mathbf{w}^T \mathbf{x}^- + b = -1$ for the closest on the other. Now maximizing the separating margin is equivalent to maximizing the distance between hyper plane H_1 and H_2. Hence we can get the maximal width between them $m = (\mathbf{x}^+ - \mathbf{x}^-) \cdot \frac{\mathbf{w}}{\|\mathbf{w}\|} = \frac{2}{\|\mathbf{w}\|}$. To maximize the margin the task is therefore:

$$\text{Minimize} \quad g(w) = \frac{1}{2}\|\mathbf{w}\|^2, \tag{16}$$

$$\text{Subject to} : y_i(\mathbf{w}^T \mathbf{x}_i + b) \geqslant 1. \tag{17}$$

By introducing Lagrangian multipliers $\alpha_i (i = 1, 2, \ldots, n)$ for the constraint, the primal problem can be reduced to finding the saddle point of Lagrange. Therefore, the dual Lagrangian becomes:

$$\text{Maximize} \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j, \tag{18}$$

$$\text{Subject to} : \sum_{i=1}^{n} \alpha_i y_i = 0, \ \alpha_i \geqslant 0. \tag{19}$$

Obviously, it is a quadratic optimization problem (QP) with linear constraints. From Karush Kuhn–Tucker (KKT) condition, we know:

$$\alpha_i (y_i(\mathbf{w}^T \mathbf{x}_i + b) - 1) = 0. \tag{20}$$

If $\alpha_i > 0$,the corresponding data points are called support vectors. Hence the solution has the form:

$$\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i, \tag{21}$$

where n is the number of support vectors. Now we can get b from $y_i(\mathbf{w}^T \mathbf{x}_i + b) - 1 = 0$, where $\mathbf{x}_i$ is support vector. After w and b are determined, the linear discriminant function can be given by:

$$g(\mathbf{x}) = \text{sgn}\left(\sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i^T \mathbf{x} + b \right). \tag{22}$$

2.3.2. Approximately linearly separable case – soft margin SVM

In practice, the data is always subject to noise or outliers, and thus it is impossible to accurately classify two classes. In order to extend the SVM to solve imperfect separation, positive slack variables ξ_i, $i = 1, \ldots l$ (Cortes & Vapnik, 1995) are introduced to allow misclassification of noisy data points, and a penalty value C is introduced for the points that cross the boundaries to take into account the misclassification errors. In fact, parameter C can be viewed as a way to control over-fitting.

Hence the new optimization problem can be reformulated as follows:

$$\text{Minimize} \quad g(w, \xi) = \frac{1}{2}\|\mathbf{w}\|^2 + C \sum_{i=1}^{n} \xi_i, \tag{23}$$

$$\text{Subject to} : y_i(\mathbf{w}^T \mathbf{x}_i + b) \geqslant 1 - \xi_i, \ \xi_i \geqslant 0. \tag{24}$$

Translate this problem into a Lagrangian dual problem

$$\text{Maximize} \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j, \tag{25}$$

$$\text{Subject to} : 0 \leqslant \alpha_i \leqslant C, \sum_{i=1}^{n} \alpha_i y_i = 0. \tag{26}$$

The solution to this minimizations problem is identical to the separable case except for the upper bound C on the Lagrange multipliers α_i.

2.3.3. Non-linearly separable case – kernel trick

In most cases, the two classes cannot be linearly separated. In order to make the linear learning machine work well in non-linear

cases, a general idea is introduced. That is, the original input space can be mapped into some higher-dimensional feature space where the training set is linearly separable. The mapping is based on the kernel function. In general, any positive semi-definite functions that satisfy the Mercer's condition can be as kernel functions (Scholkopf & Smola, 2002). Some most widely used kernels are listed here:

Linear kernel	$K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \mathbf{x}_j$
Polynomial kernel	$K(\mathbf{x}_i, \mathbf{x}_j) = (r + \gamma \mathbf{x}_i^T \mathbf{x}_j)^p$
Gaussian kernel	$K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2)$
Sigmoid kernel	$K(\mathbf{x}_i, \mathbf{x}_j) = tanh(r + \gamma \mathbf{x}_i^T \mathbf{x}_j)$

where r, p and γ are kernel parameters. Hence, now the decision function takes the form of:

$$g(\mathbf{x}) = \mathrm{sgn}\left(\sum_{i=1}^{n} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b\right). \tag{27}$$

3. Dataset and methods

3.1. Data collection

In this section, we have performed our experiments on the hepatitis database taken from UCI machine learning repository (http://archive.ics.uci.edu/ml/datasets/Hepatitis). The purpose of the dataset is to predict the presence or absence of hepatitis given the results of various medical tests carried out on a patient. The dataset contains 155 samples, of which 32 cases belong to "die" class and the remaining 123 cases belong to "live" class. Each sample in the dataset has 19 attributes besides the label. The whole 19 attributes are listed in Table 1, in which 13 attributes with binary values and 6 attributes with 6 to 9 discrete values.

3.2. The rationale of the LFDA_SVM

The rationale of LFDA_SVM which combines the feature extraction and parameter optimization is shown in Fig. 1. The feature extraction is done by local fisher discriminant analysis (LFDA), and the parameter optimization is performed by a grid search method using 10-fold cross-validation. All input variables are normalized before applying the feature extraction method. The main purpose of normalization is to avoid attributes in greater numerical ranges dominating those in smaller numerical ranges. Additionally, the normalization could avoid numerical difficulties during the calculation (Hsu, Chang, & Lin, 2003). Usually, the data could be normalized by scaling them into the interval of [0, 1] according to the Eq. (31), in which x is the original value, x' is the scaled value, $\max_a$ is the maximum value of feature a, and $\min_a$ is the minimum value of feature a. After the data is normalized, LFDA is performed to reduce the dimensionality. In the second phase, the reduced dataset is divided into three training–testing partitions, namely 80–20%, 70–30% and 50–50% respectively, via the stratified sampling method. In the third phase, a grid search using 10-fold cross-validation is carried out on each training set to find the optimal parameter pair (C, γ), where C is a penalty parameter, and γ is the kernel width of RBF kernel. In the fourth phase, the classifier is trained on each training subset with the obtained optimal parameter pair (C, γ) to get a predictor model. In the last phase, the obtained predictor model is used to predict the instances in each testing set.

$$x' = \frac{x - \min_a}{\max_a - \min_a}. \tag{28}$$

Table 1
The details of the 19 attributes of hepatitis data.

Label	Attribute	Domain
1	Age	10, 20, 30, 40, 50, 60, 70, 80
2	Sex	Male, Female
3	Steroid	No, Yes
4	Antivirals	No, Yes
5	Fatigue	No, Yes
6	Malaise	No, Yes
7	Anorexia	No, Yes
8	Liver Big	No, Yes
9	Liver Firm	No, Yes
10	Spleen Palpable	No, Yes
11	Spiders	No, Yes
12	Ascites	No, Yes
13	Varices	No, Yes
14	Bilirubin	0.39, 0.80, 1.20, 2.00, 3.00, 4.00
15	Alk Phosphate	33, 80, 120, 160, 200, 250
16	Sgot	13, 100, 200, 300, 400, 500
17	Albumin	2.1, 3.0, 3.8, 4.5, 5.0, 6.0
18	Protime	10, 20, 30, 40, 50, 60, 70, 80, 90
19	Histology	No, Yes

3.2.1. Feature extraction

When using SVM, one should bear in mind that the choice of optimal input feature subset and the optimal parameters play a crucial role for building a predictor model with high prediction accuracy and stability, and both of them are important because the feature subset choice will influence the appropriate kernel parameters and vice versa (Frohlich, Chapelle, & Scholkopf, 2003). Like feature selection, feature extraction is an alternative method of dimensionality reduction, which seeks to find one reduced representation set of features containing the most relevant information of the original data through transforming the input data into the set of reduced features. From the medical point of view, this aims at identifying the most relevant information influencing the treatment of patients (patient or normal). Thus, it plays an important role in building the classifier systems. In this study, local fisher discriminant analysis (LFDA), one new feature extraction method, is investigated for the hepatitis dataset. After normalizing the data into the interval [0, 1] according to the Eq. (31), LFDA is executed to eliminate the irrelevant or useless features. The algorithm of LFDA is implemented in Matlab (Sugiyama, 2007), and the feature number of hepatitis dataset is reduced from 19 to 2. The scatter plot of the reduced two new feature subset is given in Fig. 2.

3.2.2. Model parameters setting

In addition to the feature extraction, proper model parameters setting can improve the SVM classification accuracy. Values of parameters in SVM have to be carefully chosen in advance. These parameters include the followings: (1) regularization parameter C, which determines the tradeoff cost between minimizing the training error and the complexity of the model; (2) parameter gamma (γ or d) of the kernel function which defines the non-linear mapping from the input space to some high-dimensional feature space; (3) a kernel function used in SVM, which constructs a non-linear decision hyperplane in an input space. The sigmoid kernel behaves like the RBF for certain parameters. However, it is not valid under some parameters (Vapnik, 1998). The polynomial kernel has more hyperparameters to adjust than the RBF kernel, and takes a longer time in the training stage of SVM. Moreover, it may go to infinity or zero while the degree is large. Thus, this investigation only considers the Gaussian kernel, and a grid-search technique (Hsu et al., 2003) is employed using 10-fold cross-validation to find out the optimal parameter values of RBF kernel function. Because the computational time to find the optimal parameter values by the grid-search is not much more than those

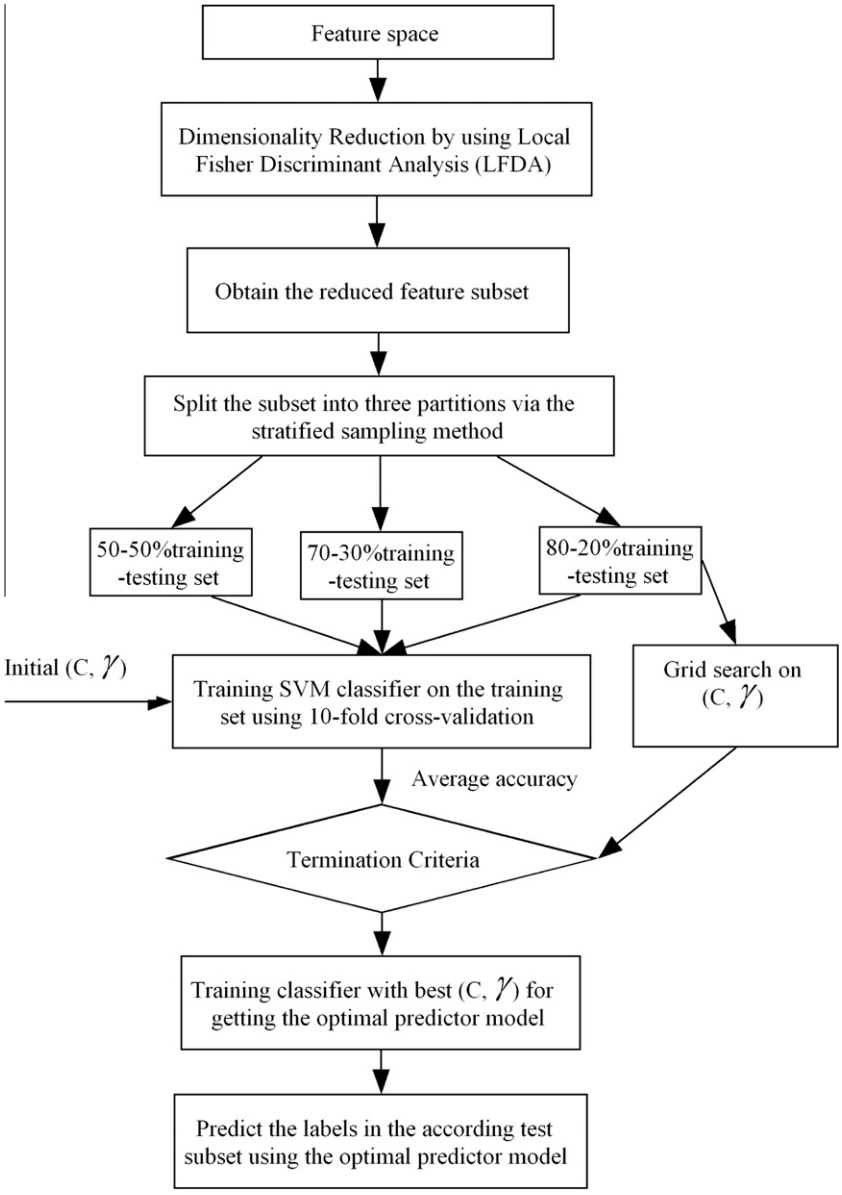

Fig. 1. The rationale of the LFDA_SVM method in terms of block diagram.

by advanced methods since there are only two parameters to be considered. Furthermore, the grid-search can be easily parallelized because each pair (C, γ) is independent (Hsu et al., 2003).

In order to ensure the same class distribution in the subset, the data set is randomly partitioned into three training–testing partitions (80–20%, 70–30% and 50–50% respectively) via a stratified sampling in which the sample proportion in each data subset is the same as that in the population. The detail of the division is represented in Table 2. Before building the Classifier, datasets are scaled. With training and testing data together, we scale each feature to the interval of [0, 1] according to the Eq. (31). Then we perform the 10-fold cross-validation on the 80%, 70% and 50% training set to choose the proper parameters of $C = \{2^{-5}, 2^{-3}, \ldots, 2^{15}\}$ and $\gamma = \{2^{-15}, 2^{-13}, \ldots, 2^{1}\}$, respectively. There will be $11 \times 10 = 110$ parameter combinations of (C, γ) are tried and the one with the best cross-validation accuracy is chosen as the parameter values of the RBF kernel. Then the best parameter pair (C, γ) is used to create the model for training. After obtain the predictor model, we conduct the prediction on each testing set accordingly.

3.2.3. Measure for performance evaluation

In order to evaluate the prediction performance of LFDA_SVM classifier, we define and compute the classification accuracy, sensitivity, specificity and confusion matrix respectively. The formulations are as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \times 100\%, \tag{29}$$

$$Sensitivity = \frac{TP}{TP + FN} \times 100\%, \tag{30}$$

$$Specificity = \frac{TN}{FP + TN} \times 100\%. \tag{31}$$

In Eq. (29)–(31), *TP* is the number of true positives; *FN* is the number of false negatives; TN is the number of true negatives; and FP is the number of false positives. They are defined as a confusion matrix in Table 3.

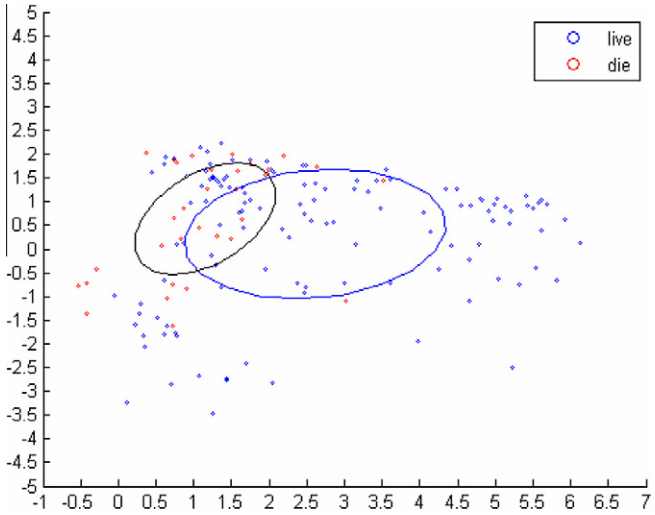

Fig. 2. The scatter plot of the reduced feature subset (where two ellipses corresponding to the two dimensions of the reduced subset defined by the mean vector and the covariance matrix).

Table 2
Training set and testing set.

Training–testing partition (%)	No. of records in the subset	
	Training set	Testing set
50–50	78	77
70–30	109	46
80–20	124	31

Table 3
Confusion matrix.

	Predicted positive	Predicted negative
Actual positive	TP	FN
Actual negative	FP	TN

Table 4
Classification accuracies for different testing sets.

Classification accuracy (%)		
50–50% training–testing	70–30% training–testing	80–20% training–testing
92.21	95.65	96.77

4. Experimental results and discussions

4.1. Experimental results

To evaluate the effectiveness of the proposed method, we conduct experiments on the hepatitis database. The classification accuracy on the testing data for the reduced feature subset is shown in Table 4. As we can see from Table 4, the highest classification accuracy, namely, 96.77% has been achieved for the 80–20% training–testing partition. The best parameter pairs (C, γ) and the numbers of support vectors (SVs) of each training–testing partition are presented in Table 5.

In addition, we present values of sensitivity and specificity for each partition in Table 6.

Classification results are displayed using a confusion matrix in Table 7. As we can see from Table 7, the number of false positives and false negatives decrease with the increase of the training set size. Especially, there are no false negative for 80–20% training–testing partition.

Table 5
The best parameter pairs (C, γ) and number of SVs of each subset.

Partition (%)	C	γ	Number of SVs
50–50	2	2^{-1}	30
70–30	2^5	2^{-1}	39
80–20	2^{-1}	2	43

For comparison purposes, Table 8 gives the classification accuracies of our method and previous methods.

As shown in Table 8, the LFDA_SVM diagnosis system has obtained the highest classification accuracy, 96.77%, reported so far.

4.2. Comparative study

In this experiment, we attempt to compare the proposed system with that of principle component analysis (PCA) based SVM, fisher discriminant analysis (FDA) based SVM and standard SVM. The whole hepatitis data is normalized to the interval [0, 1] according to Eq. (31). After the normalization of the data, PCA is used to reduce the dimensionality, and the first 7, 8, 9 and 10 principle components (PCs) are extracted from the original 19 features, respectively. As shown in Fig. 3, these numbers of PCs account for more than 76% information of the data, and all the four reduced feature subsets are adopted as the inputs to the SVM Classifier. The classification results of using the first 7–10 PCs of normalized data in SVM is presented in Table 9. The classification accuracies are found to be 87.12–93.55%. Among them, the one using nine PCs performs slightly better than the others; the highest classification accuracy of 93.55% is achieved on the 80–20% of training–testing partition. Meanwhile, 19 features are reduced to one through FDA, and then the subset with one feature is used as the input to the SVM model. The classification result of FDA_SVM is presented in Table 10, the classification accuracy is among 90.91% until 93.55%, which shows a slightly superiority over the results of PCA_SVM. In addition, the numbers of support vectors (SVs) are much lower than those produced by PCA_SVM. The result of the standard SVM using the original features is shown in Table 11, and the classification accuracy of this process is among 84.42% until 87.10%. The relatively bad performance of this classification is due to the existence of irrelevant and useless features, which leads to decreasing the performance of the classifier. Moreover, the numbers of SVs produced by this method are relatively higher than those of the other three methods. Among them, the LFDA_SVM has gained the fewest numbers of SVs. It indicates the proposed method has the best generalization ability as compared with the other three methods, since the number of support vectors is proportional to the generalization error of the SVM classifier (Vapnik, 1995). Both PCA and FDA are implemented with Matlab, and the SVM model is developed by using a simple Matlab interface to LIBSVM (Chang & Lin, 2001). The detail comparison of the four methods is shown in Fig. 4.

Table 6
Sensitivity, specificity for each subset.

Metrics	50–50% training–testing partition	70–30% training–testing partition	80–20% training–testing partition
Sensitivity (%)	98.36	97.22	100
Specificity (%)	81.25	80	85.71

Table 7
Confusion matrixes for each subset.

Actual	Predicted		Partitions
	Normal	Patient	
Normal	60	1	50–50% training–testing partition
Patient	3	13	
Normal	35	1	70–30% training–testing partition
Patient	2	8	
Normal	24	0	80–20% training–testing partition
Patient	1	6	

Table 8
Classification accuracies obtained with our method and other methods.

Author	Method	Classification accuracy (%)
Ozyildirim, Yildirim, et al.	MLP	74.37
Ozyildirim, Yildirim, et al.	RBF	83.75
Ozyildirim, Yildirim, et al.	GRNN	80.0
Adamczak	FSM with rotations	89.7
Adamczak	FSM without rotations	88.5
Adamczak	RBF (ToolDiag)	79
Adamczak	MLP + BP (ToolDiag)	77.4
Stern and Dobnikar	LDA	86.4
Stern and Dobnikar	Naive Bayes and Semi-NB	86.3
Stern and Dobnikar	QDA	85.8
Stern and Dobnikar	1-NN	85.3
Stern and Dobnikar	ASR	85
Stern and Dobnikar	Fisher discriminant analysis	84.5
Stern and Dobnikar	LVQ	83.2
Stern and Dobnikar	CART (decision tree)	82.7
Stern and Dobnikar	ASI	82.0
Stern and Dobnikar	LFC	81.9
Stern and Dobnikar	MLP with BP	82.1
Grudzinski	Weighted 9-NN	92.9
Grudzinski	18-NN, stand. Manhattan	90.2
Grudzinski	15-NN, stand. Euclidean	89.0
Jankowski	IncNet	86.0
Bascil and Temurtas	MLNN (MLP) + LM	91.87
Polat and Gunes	FS-AIRS with fuzzy res	92.59
Polat and Gunes	PCA-AIRS	94.12
Dogantekin, Avci, et al.	LDA-ANFIS	94.16
This Study	LFDA_SVM	96.77

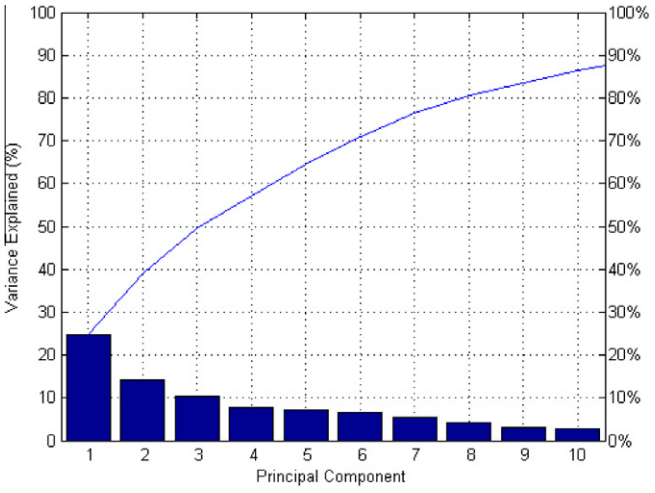

Fig. 3. The principal component of the hepatitis data and its according variance.

Table 9
Classification results with different number of principle components with SVM (PCA_SVM).

Number of PCs	Partition (%)	Classification accuracy (%)	Number of SVs
7	50–50	87.01	45
	70–30	86.96	48
	80–20	90.32	52
8	50–50	89.61	37
	70–30	89.13	40
	80–20	90.32	45
9	50–50	87.01	42
	70–30	91.30	40
	80–20	93.55	48
10	50–50	87.01	40
	70–30	86.96	44
	80–20	90.32	50

Table 10
Classification results of fisher discriminant analysis with SVM (FDA_SVM).

Partition (%)	Classification accuracy (%)	Number of SVs
50–50	90.91	28
70–30	93.48	36
80–20	93.55	44

Table 11
Classification results of the standard SVM algorithm using original features.

Partition (%)	Classification accuracy (%)	Number of SVs
50–50	84.42	48
70–30	86.96	52
80–20	87.10	64

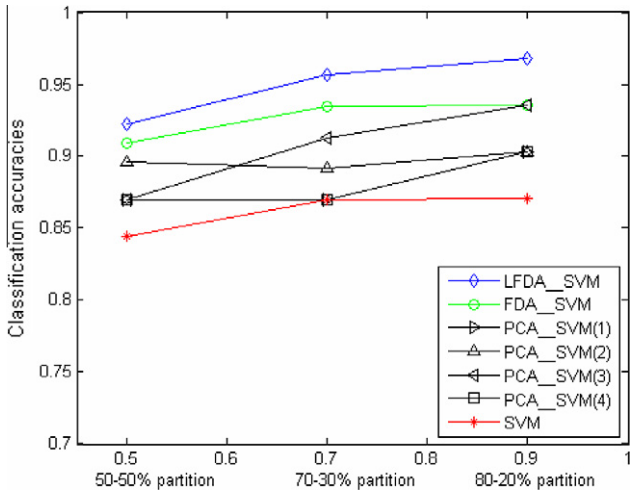

Fig. 4. The classification accuracies of different methods on different training–testing partitions.

(The horizontal axis is each partition; the vertical axis is classification accuracy of each method, where PCA_SVM(1), PCA_SVM(2), PCA_SVM(3) and PCA_SVM(4) represent PCA using 7, 8, 9 and 10 principle component combined with SVM, respectively).

As can be seen from Fig. 4, the feature extraction using LFDA shows the best performance among them, because LFDA seeks to

localize the evaluation of the within-class scatter, and thus works well even when within-class multimodality or outliers exist in the data set. In addition, LFDA does not suffer from the restriction of the original FDA in dimensionality reduction, namely the dimension of the FDA embedding space should be less than the number of classes. And the performance of FDA_SVM is slightly superior to that of PCA_SVM, it can be attributed to the characterization of the hepatitis data. When the whole data is projected into one dimension, most of them can be separately much more easily, compared with the uncorrelated components in some higher subspace extracted by PCA. Since the entire features without feature extraction were used to train SVM, the solely SVM approach performed relatively poorly in comparison with the other three methods. Obviously, from above comparative empirical study, we can see clearly that LFDA is a much more appropriate dimensionality reduction tool for hepatitis diagnosis problem compared with the other two feature extraction methods. Consequently, it make us be more convinced that the proposed diagnostic system can be very helpful in assisting the physicians to make the accurate diagnosis on the patients and will show great potential in the area of clinical hepatitis disease diagnosis.

5. Conclusion and future work

In this work, we have developed a new medical diagnostic method, LFDA_SVM, for addressing hepatitis diagnosis problem. Experiments on different portions of the hepatitis dataset demonstrated that the proposed method performed significantly well in distinguishing the live liver from the dead one. It was observed that LFDA_SVM achieved the best classification accuracies (96.77% for 80–20% training–testing partition) for a reduced feature subset that contained two features. Meanwhile, comparative study was conducted on the methods of the PCA_SVM, the FDA_SVM and the SVM. The experimental results showed that the LFDA_SVM performed advantageously over the other three methods in terms of the classification accuracy.

We believe the promising results demonstrated by the LFDA_SVM can ensure that the physicians make very accurate diagnostic decision. Future investigation will pay much attention to evaluate the proposed LFDA_SVM in other medical diagnosis problems. In addition, since the performance of SVM greatly depends on the model parameters, developing a more efficient approach to identify the optimal model parameters should also be examined in our future work.

Acknowledgement

This research is supported by the National Natural Science Foundation of China (NSFC) under Grant Nos. 60873149, 60973088, 60773099 and the National High-Tech Research and Development Plan of China under Grant Nos. 2006AA10Z245, 2006AA10A309. This work is also supported by the Open Projects of Shanghai Key Laboratory of Intelligent Information Processing in Fudan University under the Grand No. IIPL-09-007, the Open Project Program of the National Laboratory of Pattern Recognition (NLPR) and the basic scientific research fund of Chinese Ministry of Education.

References

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Fifth annual workshop on computational learning theory*. Pittsburgh: ACM.

Chang, C. C., & Lin, C. J. (2001). LIBSVM: a library for support vector machines. Software available at: <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning, 20*(3), 273–297.

Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines: And other kernel-based learning methods*. Cambridge, UK: Cambridge University Press.

Dogantekin, E., Dogantekin, A., & Avci, D. (2009). Automatic hepatitis diagnosis system based on linear discriminant analysis and adaptive network based on fuzzy inference system. *Expert Systems with Applications, 36*(8), 11282–11286.

Dogantekin, Esin., Dogantekin, Akif., & Avci, Derya. (2009). Automatic hepatitis diagnosis system based on linear discriminant analysis and adaptive network based on fuzzy inference system. *Expert Systems with Applications, 36*(8), 11282–11286.

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern classification*. New York: Wiley.

Frohlich, H., Chapelle, O., & Scholkopf, B. (2003). Feature selection for support vector machines by means of genetic algorithms, In *Proceedings of the 15th IEEE international conference on tools with artificial intelligence, Sacramento, CA, USA* (pp. 142–148).

Hsu, C. W., Chang, C. C., & Lin, C. J. (2003). A practical guide to support vector classification, Technical report, Department of Computer Science and Information Engineering, National Taiwan University, Taipei, 2003. Available from <http://www.csie.ntu.edu.tw/cjlin/libsvm/>.

Joachims, T., Nedellec, C., & Rouveirol, C. (1998). Text categorization with support vector machines: Learning with many relevant. In *Proceedings of the 10th European conference on machine learning* (pp.137–142).

Osuna, E., Freund, R., & Girosi, F. (1997). Training support vector machines: Application to face detection. In *Proceedings of computer vision and pattern recognition, Puerto Rico* (pp. 130–136).

Polat, K., & Gunes, S. (2007). An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease. *Digital Signal Processing, 17*(4), 702–710.

Polat, K., & Gunes, S. (2007). Prediction of hepatitis disease based on principal component analysis and artificial immune recognition system. *Applied Mathematics and Computation, 189*(2), 1282–1291.

Polat, K., & Gunes, S. (2008). Computer aided medical diagnosis system based on principal component analysis and artificial immune recognition system classifier algorithm. *Expert Systems with Applications, 34*(1), 773–779.

Scholkopf, B., & Smola, A. J. (2002). *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. MIT Press.

Ster, B., & Dobnikar, A. (1996). Neural networks in medical diagnosis: Comparison with other methods. In *Proceedings of the international conference on engineering applications of neural networks* (pp. 427–430).

Sugiyama, M. (2007). Dimensionality reduction of multimodal labeled data by local Fisher discriminant analysis. *Journal of Machine Learning Research, 8*, 1027–1061.

Sugiyama, M., Ide, T., Nakajima, S., & Sese, J. (2010). Semi-supervised local Fisher discriminant analysis for dimensionality reduction. *Machine Learning, 78*(1–2), 35–62.

Vapnik, V. (1998). *Statistical learning theory*. New York: Wiley.

Vapnik, V. (1995). *The nature of statistical learning theory*. New York: Springer.

Zelnik-Manor, L., & Perona, P. (2004). Self-tuning spectral clustering. *Advances in Neural Information Processing Systems, 17*, 1601–1608.