Hindawi Publishing Corporation
The Scientific World Journal
Volume 2014, Article ID 536434, 12 pages
http://dx.doi.org/10.1155/2014/536434

Research Article

New Fuzzy Support Vector Machine for the Class Imbalance Problem in Medical Datasets Classification

Xiaoqing Gu, Tongguang Ni, and Hongyuan Wang

School of Information Science and Engineering, Changzhou University, Changzhou 213164, China

Correspondence should be addressed to Hongyuan Wang; tiddyddd@163.com

Received 25 November 2013; Accepted 20 February 2014; Published 23 March 2014

Academic Editors: V. Bhatnagar and Y. Zhang

In medical datasets classification, support vector machine (SVM) is considered to be one of the most successful methods. However, most of the real-world medical datasets usually contain some outliers/noise and data often have class imbalance problems. In this paper, a fuzzy support machine (FSVM) for the class imbalance problem (called FSVM-CIP) is presented, which can be seen as a modified class of FSVM by extending manifold regularization and assigning two misclassification costs for two classes. The proposed FSVM-CIP can be used to handle the class imbalance problem in the presence of outliers/noise, and enhance the locality maximum margin. Five real-world medical datasets, breast, heart, hepatitis, BUPA liver, and pima diabetes, from the UCI medical database are employed to illustrate the method presented in this paper. Experimental results on these datasets show the outperformed or comparable effectiveness of FSVM-CIP.

1. Introduction

Computer techniques such as machine learning and pattern recognition have been widely adopted by modern medicine. One reason is that an enormous amount of data has to be gathered and analyzed which is very hard or even impossible without making use of computer techniques. The other reason is that computer techniques have led toward digital analysis of pathological diagnosis, automatic classification differentiating, and detecting diseases. In some cases, an early symptom of some diseases is lighter and gives no obvious pointer to a possible diagnosis; moreover, many symptoms look very similar to each other, though they are caused by different diseases. So it may be difficult even for experienced doctors to make correct diagnosis. Therefore, an automatic classification system can help doctor diagnose accurately, assess disorders remotely and evaluate the treatment process [1].

In recent years, researchers have proposed a lot of approaches for medicine classification, such as neural network, Bayesian network, and support vector machine (SVM). Among them SVM is considered to be one of the most successful ones [2]. For example, to improve time and accuracy in differentiating diffuse interstitial lung disease for computer-aided quantification, a hierarchical SVM is introduced which shows promise for various real-time and online image-based classification applications in clinical fields [3]. SVM as a classifier is used for liver disorders and its correct classification rate is highly successful compared to the other results attained [4]. A two-stage approach is proposed for medical datasets classification, in which the artificial bee colony algorithm is used for feature selection and SVM is used for classification [5].

The support vector machine (SVM) proposed by Vapnik [6, 7] is a novel approach for solving pattern recognition problems. SVM maps the sample points into a high-dimensional feature space to seek for an optimal separating hyperplane through maximizing the margin between two classes. In addition, SVM is a quadratic programming (QP) problem that assures that its solution is obtained once it is the global unique solution, and the sparsity of solution assures better generalization. However, most of the real-world medical datasets usually contain some outliers and noisy examples. The classical SVM is very sensitive to outliers/noise. To solve this problem, fuzzy support vector machine (FSVM) [8] is proposed, in which each sample is given a fuzzy membership that denotes the attitude of the corresponding point toward

one class. The membership represents how important the sample is to the decision surface.

Nevertheless, many medical datasets are composed of "normal" samples with only a small percentage of "abnormal" ones, which leads to the so-called class imbalance problems. FSM does not take into consideration the class distribution and can be sensitive to the class imbalance problem. As a result, the hyperplane of FSVM can be skewed towards the minority class, and this skewness can degrade the performance of FSVM with respect to the minority class. To tackle this problem, Veropoulos et al. [9] have proposed a method called different error costs (DEC), where the SVM objective function has been modified to assign two different misclassification cost values. It is noticed that One-Class Classification [10, 11] is sometimes used in novelty detection, and it only uses the normal training data. However, in many real medical datasets, abnormal examples exist, although they are very few. Furthermore, in classification tasks, the scatter matrix can play an important role when incorporated with local intrinsic geometry structures of samples [12]. Some methods have been recently proposed to incorporate the structure of the data distribution into SVM. A linear manifold learning method named locality preserving projection (LPP) is proposed in [13, 14], which aims at preserving the local manifold structure of the samples space. Although LPP considers enhancing the local data compactness with each manifold, it does not separate manifolds with different class labels.

In this paper, we propose a new FSVM method for the class imbalance problem (FSVM-CIP) which can be used to address both the problem of class imbalance and outliers/noise. FSVM-CIP not only considers the fuzziness of each training sample but also extends manifold regularization and maximizes the localized relative margin. It takes the positive samples and negative samples into consideration with different misclassification costs according to their unbalanced distributions. We systematically evaluated the FSVM-CIP on five real-world medical datasets and compared its performance with four different SVM methods for classification. The results showed that the proposed method can improve the classification accuracy and handle the classification problems with outliers/noise and imbalanced datasets more effectively.

The rest of this paper is organized as follows. Section 2 briefly reviews the related works. Section 3 presents the details of FSVM-CIP in the linear case. Section 4 presents FSVM-CIP in the nonlinear case in detail. The experimental results on five medical datasets are reported in Section 5, and some concluding remarks are given in Section 6.

2. Related Works

2.1. Fuzzy Support Vector Machines (FSVMs). In traditional SVM, all the data points are considered with equal importance and assigned the same penal parameter in its objective function. However, in many real-world classification applications, some sample points, such as the outliers or noises, may not be exactly assigned to one of these two classes, and each sample point does not have the same meaning to the decision surface. To solve this problem, the theory of fuzzy support vector machine was originally proposed in [8]. Fuzzy membership to each sample point is introduced such that different sample points can make different contributions to the construction of decision surface.

Suppose the training samples are

$$S = \left\{\left(\mathbf{x}_i, y_i, s_i\right),\ i = 1, \ldots, N\right\}, \tag{1}$$

where $\mathbf{x}_i \in \mathbf{R}^n$ is the n-dimension sample point, $y_i \in \{-1, +1\}$ represents its class label, and s_i $(i = 1, \ldots, N)$ is a fuzzy membership which satisfies $\sigma \le s_i \le 1$ with a sufficiently small constant $\sigma > 0$. The quadratic optimization problem for classification is considered as follows:

$$\begin{aligned} \min_{\mathbf{w},s,\xi} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i-1}^{l} s_i\xi_i \\ \text{s.t.} \quad & y_i\left(\mathbf{w}^T\mathbf{x}_i + b\right) \ge 1 - \xi_i, \quad \xi_i \ge 0,\ i = 1, \ldots, l, \end{aligned} \tag{2}$$

where $\mathbf{w}$ is a normal vector of the separating hyperplane, b is a bias term, and C is a parameter which has to be determined beforehand to control the tradeoff between the classification margin and the cost of misclassification error. Since s_i is the attitude of the corresponding point $\mathbf{x}_i$ towards one class and the slack variables ξ_i are a measure of error, then the term $s_i\xi_i$ can be considered a measure of error with different weights. It is noted that the bigger the s_i is, the more importantly the corresponding point is treated; the smaller the s_i is, the less importantly the corresponding point is treated; thus, different input points can make different contributions to the learning of decision surface. Therefore, FSVM can find a more robust hyperplane by maximizing the margin by letting some misclassification of less important points.

In order to solve the FSM optimal problem, (2) is transformed into the following dual problem by introducing Lagrangian multipliers α_i:

$$\begin{aligned} \max_{\alpha} \quad & \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j \mathbf{x}_i\mathbf{x}_j \\ \text{s.t.} \quad & \sum_{i=1}^{N} y_i\alpha_i = 0, \quad 0 \le \alpha_i \le s_i C,\ i = 1, \ldots, N. \end{aligned} \tag{3}$$

Compared with the standard SVM, the above statement only has a little difference, which is the upper bound of the values of α_i. By solving this dual problem in (3) for optimal α_i, $\mathbf{w}$ and b can be recovered in the same way as in the standard SVM.

2.2. Locality Preserving Projections (LPP). Locality preserving projection (LPP) [13, 14] is a linear dimensionality reduction algorithm by feature extraction or projection. It builds an adjacency graph incorporating neighborhood information of the data set using the Laplacian graph and then computes a transformation matrix which maps the data points into a subspace. This linear transformation optimally preserves local neighborhood information in a certain sense. The representation map generated by this method can be

viewed as a linear discrete approximation to a continuous map that naturally arises from the geometry of the manifold.

For a set $X = \{\mathbf{x}_i\}$ ($i \in [1, N]$), let $N_k(\mathbf{x}_i)$ denote k nearest neighbors of node i, and let G denote the adjacency graph of dataset X. Here, the ith node corresponds to the data point x_i and nodes i and j are connected by an edge if node i is among the k nearest neighbors of node j or if node j is among the k nearest neighbors of node i; that is, $\mathbf{x}_i \in N_k(\mathbf{x}_j)$ or $\mathbf{x}_j \in N_k(\mathbf{x}_i)$. The adjacency graph G can be weighed as follows:

$$W_{ij} = \begin{cases} \exp\left(\dfrac{-\left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2}{t}\right) & \text{if } \mathbf{x}_i \in N_k\left(\mathbf{x}_j\right) \\ & \text{or } \mathbf{x}_j \in N_k\left(\mathbf{x}_i\right) \\ 0 & \text{otherwise,} \end{cases} \tag{4}$$

where $\exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/t)$ is called the heart kernel function and t is a constant. $\|\mathbf{x}_i - \mathbf{x}_j\|$ is the Euclidean distance in $\mathbf{R}^n$ between point i and point j. LPP tries to find the transformation vector $\mathbf{w} \in \mathbf{R}^n$ by minimizing the following objective function:

$$\begin{aligned} \min_{\mathbf{w} \neq 0} \quad & \mathbf{w}^T \mathbf{X} \mathbf{L} \mathbf{X}^T \mathbf{w} \\ \text{s.t.} \quad & \mathbf{w}^T \mathbf{X} \mathbf{D} \mathbf{X}^T \mathbf{w} = 1, \end{aligned} \tag{5}$$

where $\mathbf{D}$ is a diagonal matrix whose entries are column sum of $\mathbf{W}$ and $D_{ii} = \sum_j W_{ij}$ normalizes each weight. $\mathbf{L} = \mathbf{D} - \mathbf{W}$ is the Laplacian matrix. The transformation vector $\mathbf{w}$ in the objective function in (5) is given by the minimum eigenvalue solution to the generalized eigenvalue problem. LPP preserves the intrinsic geometry and local structure of the data by minimizing the objective function.

3. FSVM for the Class Imbalance Problem in the Linear Case

In this section, we first define the local within-class preserving scatter matrix in the linear case. Secondly, the optimization problem formulation of FSVM-CIP in the linear case is given. Moreover, the fuzzy membership functions for linear FSVM-CIP are defined. Finally, the algorithm of linear FSVM-CIP is summarized.

3.1. The Local within-Class Preserving Scatter Matrix in the Linear Case. Following the idea of [15], we build the nearest within-class neighbor graph to model intrinsic geometry and local structure of the data. The graph preserves local neighborhood information in a certain sense and it can be viewed as a linear discrete approximation to a continuous map that naturally arises from the geometry of the manifold.

Considering the fact that we have a binary classification problem, one class denoted as C_1 contains sample points $\mathbf{x}_i$ with $y_i = 1$ and the other class denoted as C_2 contains sample points $\mathbf{x}_i$ with $y_i = -1$. Set $|C_1| = m_1$ and $|C_2| = N - m_1$, and the total number of sample points is N.

Definition 1. For each data $\mathbf{x}_i$, suppose its k nearest within-class neighbors set $N_k(\mathbf{x}_i)$ and an edge is put between $\mathbf{x}_j$ and its neighbors. The corresponding weight matrix W_{ij} is

$$W_{ij} = \begin{cases} \dfrac{1}{D_{ii}} \exp\left(\dfrac{-\left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2}{t}\right) & \text{if } \mathbf{x}_i \in N_k\left(\mathbf{x}_j\right) \\ & \text{or } \mathbf{x}_j \in N_k\left(\mathbf{x}_i\right), y_i = y_j \\ 0 & \text{otherwise,} \end{cases} \tag{6}$$

where $D_{ii} = \sum_j W_{ij}$ normalizes each weight.

Definition 2. The local within-class preserving scatter matrix

$$\begin{aligned} & \mathbf{S}_{lw} \\ & = \sum_{k=1}^{2} \sum_{\mathbf{x}_i \in C_k} \left(\mathbf{x}_i - \sum_{\mathbf{x}_i \in N(\mathbf{x}_j) \text{ or } \mathbf{x}_j \in N(\mathbf{x}_i)} W_{ij} \mathbf{x}_j\right) \\ & \qquad \times \left(\mathbf{x}_i - \sum_{\mathbf{x}_i \in N(\mathbf{x}_j) \text{ or } \mathbf{x}_j \in N(\mathbf{x}_i)} W_{ij} \mathbf{x}_j\right)^T \\ & = \sum_{k=1}^{2} \mathbf{X}^{(k)} \left(\mathbf{I}^{(k)} - \mathbf{W}^{(k)}\right)^T \left(\mathbf{I}^{(k)} - \mathbf{W}^{(k)}\right) \mathbf{X}^{(k)T}, \end{aligned} \tag{7}$$

where $\mathbf{I}^{(k)}$ is an $N_k \times N_k$ diagonal matrix. In this case, the obtained nearest within-class neighbor graph attempts to preserve the local structure of the data set and $(\mathbf{I}^{(k)} - \mathbf{W}^{(k)})^T(\mathbf{I}^{(k)} - \mathbf{W}^{(k)})$ preserves locality of nearby points with same class label in the embedding space during the unfolding process of nonlinear structures [15]. In fact, a heavy penalty is applied to the objective function through the weight W_{ij} if the neighboring data $\mathbf{x}_i$ and $\mathbf{x}_j$ are mapped far apart. Hence, the minimization criterion is an attempt to ensure points y_i and y_j close to each other as well as $\mathbf{x}_i$ and $\mathbf{x}_j$ being close.

It is worthwhile to note that the local within-class scatter matrix $\mathbf{S}_{lw}$ is symmetric and positive semidefinite. $\mathbf{S}_{lw}$ looks similar to the within-class scatter matrix $\mathbf{S}_w$ [16, 17] and the Laplacian matrix $\mathbf{L}$ in LPP. However, $\mathbf{S}_{lw}$ reflects the intrinsic geometry and local structure of the data, and $\mathbf{S}_w$ only considers the mean value of samples in different classes. $\mathbf{S}_{lw}$ carries the class label information and discriminating information but $\mathbf{L}$ only considers the information of nearest neighbors for each data point in the input space, without considering the class labels.

3.2. FSVM-CIP in the Linear Case. To tackle the imbalance classification problem with noise and outliers, we integrate FSVM, the ideas of imbalance classification problem, and the local within-class preserving scatter. On one hand, as shown in Figure 1, the linear classifier presented by the hyperplane is $(\mathbf{w}^T\mathbf{x} + b = 0)$ and defines a field for majority-class examples $(\mathbf{w}^T\mathbf{x}+b > 1-\xi)$ and another field for minority-class examples $(\mathbf{w}^T\mathbf{x} + b > -(1 + \rho - \xi))$ which is used to weaken the skewness towards the minority class and enhance the locality

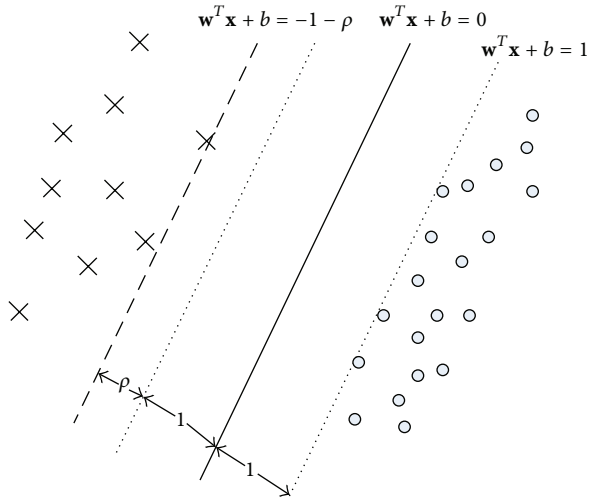

Figure 1: The hyperplanes of linear FSVM-CIP.

maximum margin. On the other hand, by assigning a higher misclassification cost for the minority class examples than the majority class examples, the effect of class imbalance could be reduced. In addition, to minimize the amount of misclassifications, the local within-class scatter matrix $\mathbf{S}_{lw}$ is used to preserve intrinsic geometry and local structure of the data.

Due to this, we define the primal problem of FSVM-CIP as follows:

$$\begin{aligned}
\min_{w,b,\rho,\xi} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} - \nu\rho \\
& + \frac{1}{\nu_1 m_1}\sum_{i=1}^{m_1}\mu_i\xi_i + \frac{1}{\nu_2 m_2}\sum_{j=m_1+1}^{N}\mu_j\xi_j + \frac{\eta}{2}\mathbf{w}^T\mathbf{S}_{lw}\mathbf{w} \\
\text{s.t.} \quad & \mathbf{w}^T\mathbf{x}_i + b \geq 1 - \xi_i, \quad i = 1, \ldots, m_1 \\
& -\left(\mathbf{w}^T\mathbf{x}_j + b\right) \geq 1 + \rho - \xi_j, \quad j = m_1 + 1, \ldots, N \\
& \xi_k \geq 0, \quad k = 1, \ldots, N, \ \rho \geq 0,
\end{aligned} \tag{8}$$

where m_1, m_2 denote the number of positive (normal class or majority class) and negative (abnormal class or minority class) training points, and $m_2 = N - m_1$. ρ is a nonnegative number, and $\rho + 1$ is the margin between the hyperplane and the minority class examples. η is a nonnegative regulation constant which is the tradeoff between the local within-class scatter and the margin. Variables ν_1, ν_2 are positive penalty parameters, which tune penalty cost of the training error for positive and negative training data, respectively. $\xi_i, \xi_j \geq 0$ are the slack variables, and μ_i, μ_j are fuzzy memberships for two-class examples.

Obviously, $\mathbf{w}^T\mathbf{S}_{lw}\mathbf{w}$ provides prior geometrical information into the penalty terms based on manifold regularization. Minimizing $\mathbf{w}^T\mathbf{S}_{lw}\mathbf{w}$ means that close data originally in the same class in the input space are likely to be close in the output place. Therefore, $\mathbf{w}^T\mathbf{S}_{lw}\mathbf{w}$ aims to preserve the local information of the manifold structure.

It is noted that, in FSVM-CIP, we assign different fuzzy membership values for training examples to reflect their different classes of importance. We also showed that it is similar to assign different misclassification costs $\mu_i/\nu_1 m_1 (\mu_j/\nu_2 m_2)$ for different training examples. In order to reduce the effect of class imbalance, we can assign higher membership values μ_j or lower parameter ν_2 for the minority class examples, while we assign lower membership values μ_i or higher ν_1 for the majority class. That is, our proposed method would not tend to skew the separating hyperplane towards the minority class examples as the minority class examples are now assigned with a higher misclassification cost. By means of setting $\mu_i/\nu_1 m_1 (\mu_j/\nu_2 m_2)$ and extending manifold regularization, the learned optimal separating hyperplane enhances the relative maximum margin and FSVM-CIP will be less sensitive to imbalanced class problems.

Then, we transform this problem into its corresponding dual problem as follows.

The primal Lagrangian is

$$\begin{aligned}
& L\left(\mathbf{w}, b, \rho, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\gamma}, s\right) \\
& = \frac{1}{2}\mathbf{w}^T\mathbf{w} - \nu\rho + \frac{1}{\nu_1 m_1}\sum_{i=1}^{m_1}\mu_i\xi_i + \frac{1}{\nu_2 m_2}\sum_{j=m_1+1}^{N}\mu_j\xi_j \\
& \quad + \frac{\eta}{2}\mathbf{w}^T\mathbf{S}_{lw}\mathbf{w} - \sum_{i=1}^{m_1}\alpha_i\left(\mathbf{w}^T\mathbf{x}_i + b - 1 + \xi_i\right) \\
& \quad + \sum_{j=m_1+1}^{N}\alpha_j\left(\mathbf{w}^T\mathbf{x}_j + b + 1 + \rho - \xi_j\right) - \sum_{i=1}^{N}\gamma_i\xi_i - s\rho,
\end{aligned} \tag{9}$$

with Lagrangian multipliers $\alpha_i \geq 0$, $\gamma_i \geq 0$, and $s \geq 0$. The derivatives of $L(\mathbf{w}, b, \rho, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\gamma}, s)$ with respect to the primal variables using the Karush-Kuhn-Tucker (KKT) conditions should vanish. Consider

$$\frac{\partial L}{\partial b} = \sum_{i=1}^{N}\alpha_i y_i = 0, \tag{10}$$

$$\frac{\partial L}{\partial \mathbf{w}} = \mathbf{I}\mathbf{w} + \eta\mathbf{S}_{lw}\mathbf{w} - \sum_{i=1}^{N}\alpha_i y_i \mathbf{x}_i = 0, \tag{11}$$

$$\frac{\partial L}{\partial \rho} = -\nu + \sum_{j=m_1+1}^{N}\alpha_j - s = 0, \tag{12}$$

$$\frac{\partial L}{\partial \xi_i} = \frac{\mu_i}{\nu_1 m_1} - \alpha_i - \gamma_i = 0, \quad i = 1, \ldots, m_1, \tag{13}$$

$$\frac{\partial L}{\partial \xi_j} = \frac{\mu_j}{\nu_2 m_2} - \alpha_j - \gamma_j = 0, \quad j = m_1 + 1, \ldots, N, \tag{14}$$

where $\mathbf{I}$ is an N-dimensional vector of ones, and $\mathbf{I} = [1, \ldots, 1]^T$. We have $\mathbf{w} = (\mathbf{I} + \eta\mathbf{S}_{lw})^{-1}\sum_{i=1}^{N}\alpha_i y_i \mathbf{x}_i$.

Substituting (10)–(14) into (9), we obtain the dual form of the optimization problem:

$$\begin{aligned}
\min_{\boldsymbol{\alpha}} \quad & \frac{1}{2}\boldsymbol{\alpha}^T \mathbf{H}\boldsymbol{\alpha} \\
\text{s.t.} \quad & \sum_{i=+1}^{m_1} \alpha_i = \nu \\
& \sum_{j=m_1+1}^{N} \alpha_j = \nu \\
& 0 \le \alpha_i \le \frac{\mu_i}{\nu_1 m_1}, \quad i = 1, \ldots, m_1 \\
& 0 \le \alpha_j \le \frac{\mu_j}{\nu_2 m_2}, \quad j = m_1 + 1, \ldots, N,
\end{aligned} \tag{15}$$

where $\mathbf{H}$ is a matrix with entry $H_{ij} = y_i y_j \mathbf{x}_i^T (\mathbf{I} + \eta \mathbf{S}_{lw})^{-1} \mathbf{x}_j$, and vectors $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_N]^T$.

Equation (15) is a typical convex quadratic programming problem which is easy to be numerically solved. Suppose $\boldsymbol{\alpha}^* = [\alpha_1^*, \ldots, \alpha_N^*]$ can be used to solve the above optimization problem, and then the optimal weight vector is

$$\mathbf{w}^* = (\mathbf{I} + \eta \mathbf{S}_{lw})^{-1} \sum_{i=1}^{N} \alpha_i^* y_i \mathbf{x}_i. \tag{16}$$

Denote a training sample $\mathbf{x}_i$ $(1 \le i \le N)$ called a support vector (SV) if the corresponding Lagrange multiplier $\alpha_i > 0$. Denote the SV sets as $\mathrm{SV}_1 = \{\mathbf{x}_i \mid 0 < \alpha_i \le \mu_i/\nu_1 m_1,\ 1 \le i \le m_1\}$ and $\mathrm{SV}_2 = \{\mathbf{x}_j \mid 0 < \alpha_j \le \mu_j/\nu_2 m_2,\ 1 + m_1 \le j \le N\}$ while s^+ and s^- denote the number of SVs in SV_1 and SV_2, respectively. According to KKT condition, (15) becomes equations for the input data in SV_1 and SV_2, respectively, with slack variables ξ_i and ξ_j being 0. Thus, the optimal thresholds b^* and ρ^* can be calculated. However, from the numerical perspective, it is better to take the mean value of b^* and ρ^* resulting from all such data. Therefore, the optimal thresholds b^* and ρ^* are computed by the following formula:

$$b^* = 1 - \frac{1}{s^+} \sum_{\mathbf{x}_i \in \mathrm{SV}_1} \left(\mathbf{w}^*\right)^T \mathbf{x}_i, \tag{17}$$

$$\rho^* = -\frac{1}{s^+} \sum_{\mathbf{x}_i \in \mathrm{SV}_1} \left(\mathbf{w}^*\right)^T \mathbf{x}_i + \frac{1}{s^-} \sum_{\mathbf{x}_j \in \mathrm{SV}_2} \left(\mathbf{w}^*\right)^T \mathbf{x}_j. \tag{18}$$

As a result, the corresponding decision function of the linear FSVM-CIP will be

$$\begin{aligned}
f(\mathbf{x}) &= \operatorname{sgn}\left(\mathbf{w}^T \mathbf{x} + b^*\right) \\
&= \operatorname{sgn}\left(\sum_{i=1}^{N} \alpha_i^* y_i \left(\mathbf{x}_i^T (\mathbf{I} + \eta \mathbf{S}_{lw})^{-1} \mathbf{x}\right) + b^*\right).
\end{aligned} \tag{19}$$

Note that, to deal with the small sample size problem, $(\mathbf{I} + \eta\mathbf{S}_{lw})$ is regularized by adding a scale multiple η of the identity matrix $\mathbf{S}_{lw}$ with $\mathbf{I}$ before any inversion takes place. Hence, $(\mathbf{I} + \eta\mathbf{S}_{lw})$ is always nonsingular, and the inverse of $(\mathbf{I} + \eta\mathbf{S}_{lw})$ exists.

Following the terminology in [18], a training sample $\mathbf{x}_i$ $(1 \le i \le N)$ is called a margin error (ME) if the corresponding slack variable $\boldsymbol{\xi}_i > 0$. We give the following theorem for parameter selection later.

Theorem 3. *Let m^+ and m^- denote the number of MEs in the positive and negative classes; s^+ and s^- denote the number of SVs in the positive and negative classes, respectively. Then one has*

$$\overline{\mu_m^+} m^+ \le \nu\nu_1 m_1 \le \overline{\mu_s^+} s^+, \tag{20}$$

$$\overline{\mu_m^-} m^- \le \nu\nu_2 m_2 \le \overline{\mu_s^-} s^-, \tag{21}$$

where $\overline{\mu_m^+}$ and $\overline{\mu_m^-}$ denote the mean fuzzy membership of MEs in the positive and negative classes; $\overline{\mu_s^+}$ and $\overline{\mu_s^-}$ denote the mean fuzzy membership of SVs in the positive and negative classes, respectively.

A proof of the above theorem can be found in Appendix.

3.3. Fuzzy Membership Functions in the Linear Case. In FSVM, the fuzzy membership is used to reduce the effects of outliers or noises and different fuzzy membership functions have different influences on the fuzzy algorithm. Basically, the rule to assign proper membership values to data points can depend on the relative importance of date points to their own classes. In this paper, we consider two fuzzy membership functions given in [19].

Given the sequence of training points, denote the mean of positive class and negative class as $\overline{x}_+$ and $\overline{x}_-$.

Definition 4. The μ_{lin} is called the linear fuzzy membership and μ_{lin} can be defined as

$$\mu_{\text{lin}} = \begin{cases} \dfrac{1 - \left\|\mathbf{x}_i - \overline{\mathbf{x}}_+\right\|}{\left(\max_j \left(\left\|\mathbf{x}_j - \overline{\mathbf{x}}_+\right\|\right) + \delta\right)} & \text{if } y_i = 1 \\ \dfrac{1 - \left\|\mathbf{x}_i - \overline{\mathbf{x}}_-\right\|}{\left(\max_j \left(\left\|\mathbf{x}_j - \overline{\mathbf{x}}_-\right\|\right) + \delta\right)} & \text{if } y_i = -1, \end{cases} \tag{22}$$

where δ is a small positive value, which is used to avoid μ_{lin} becoming zero. $\|\cdot\|$ is the Euclidean distance.

Definition 5. The μ_{exp} is called the exponential fuzzy membership and μ_{exp} can be defined as

$$\mu_{\text{exp}} = \begin{cases} \dfrac{2}{1 + \exp\left(\lambda \left\|\mathbf{x}_i - \overline{\mathbf{x}}_+\right\|\right)} & \text{if } y_i = 1 \\ \dfrac{2}{1 + \exp\left(\lambda \left\|\mathbf{x}_i - \overline{\mathbf{x}}_-\right\|\right)} & \text{if } y_i = -1, \end{cases} \tag{23}$$

where parameter $\lambda \in [0, 1]$ determines the steepness of the decay.

3.4. Solution. Based on the above, we can state the approach of proposed FSVM-CIP in the linear case as Algorithm 1.

Input:
Training samples $\{(\mathbf{x}_i, y_i), i = 1, \ldots, N\}$
Testing samples $\{\mathbf{x}_j, j = 1, \ldots, U\}$
Output:
The predicted labels y_j of data $\{\mathbf{x}_j, j = 1, \ldots, U\}$
Procedure:
(1) Compute fuzzy membership μ_i using (22) or (23) for the data $\{(\mathbf{x}_i, y_i), i = 1, \ldots, N\}$
(2) Construct data adjacency graph G using k nearest neighbors and compute the edge weights matrix W_{ij} with N examples
(3) Construct local within-class preserving scatter matrix $\mathbf{S}_{\text{lw}}$ using (8)
(4) Choose parameters t (6); η, ν, ν_1 and ν_2 (8)
(5) Compute α^* using (15) and b^* using (17) with a QP Solver
(6) Using decision function (19) with samples $\mathbf{x}_j$, and output the final class labels

Algorithm 1: FSVM-CIP in the linear case.

4. FSVM for the Class Imbalance Problem in the Nonlinear Case

In this section, we extend the local within-class preserving scatter matrix and FSVM-CIP into feature space. Moreover, the fuzzy membership functions in feature space are defined. Finally, the algorithm of kernel FSVM-CIP is summarized.

4.1. Kernel Extension. In order to handle nonlinear classification, the kernelization trick [20] is used to map the n-dimensional date points into an arbitrary reproducing kernel Hilbert space (RKHS) [21] via a mapping function $\phi : \mathbf{R}^n \mapsto \mathbf{H}$; that is, $\mathbf{x_i} \mapsto \phi(\mathbf{x_i})$. Then a linear hyperplane $f(\mathbf{v}) = \boldsymbol{\alpha}^T\phi(\mathbf{v})+b$ in feature space $\mathbf{H}$ would correspond to a nonlinear hyperplane in the original space $\mathbf{R}^n$ where $\boldsymbol{\alpha}, \phi(\mathbf{v}) \in \mathbf{H}$, $\mathbf{v} \in \mathbf{R}^n$, and $b \in \mathbf{R}$.

Let $\phi(\mathbf{X})$ denote the date matrices in feature space $\mathbf{H}$, $\phi(\mathbf{X}) = [\phi(\mathbf{x}_1), \phi(\mathbf{x}_2), \ldots, \phi(\mathbf{x}_n)]$; then the kernel function $\mathbf{K}$ is a matrix with entry $K_{ij} = K(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^T\phi(\mathbf{x}_j)$.

Here the kernel local within-class scatter matrix $\mathbf{S}_{lw}^{\phi}$ in feature space is

$$\begin{aligned}
&\mathbf{S}_{lw}^{\phi} \\
&= \sum_{k=1}^{2}\sum_{i=1}^{N_k}\left(\phi(\mathbf{x}_i) - \sum_{j=1}^{N_k}\mathbf{W}_{ij}^{\phi k}\phi(\mathbf{x}_j)\right) \\
&\qquad \times \left(\phi(\mathbf{x}_i) - \sum_{j=1}^{N_k}\mathbf{W}_{ij}^{\phi k}\phi(\mathbf{x}_j)\right)^T \qquad (24)\\
&= \mathbf{K}^{(1)}\left(\mathbf{I}^{(1)} - \mathbf{W}^{\phi(1)}\right)^T\left(\mathbf{I}^{(1)} - \mathbf{W}^{\phi(1)}\right)\mathbf{K}^{(1)T} \\
&\quad + \mathbf{K}^{(2)}\left(\mathbf{I}^{(2)} - \mathbf{W}^{\phi(2)}\right)^T\left(\mathbf{I}^{(2)} - \mathbf{W}^{\phi(2)}\right)\mathbf{K}^{(2)T},
\end{aligned}$$

where $\mathbf{I}^{(1)}$, $\mathbf{I}^{(2)}$ are N_1-order, N_2-order identity matrixes, respectively. Based on the above notations, $\mathbf{K}^{(1)}$, $\mathbf{K}^{(2)}$ are $N \times m_1$, $N \times (N - m_1)$ matrixes, respectively; thus $\mathbf{K} = [\mathbf{K}^{(1)}, \mathbf{K}^{(2)}]$.

The weight matrixes $\mathbf{W}^{\phi(1)}$ and $\mathbf{W}^{\phi(2)}$ are the nonlinear version of $\mathbf{W}^{(1)}$ and $\mathbf{W}^{(2)}$, respectively. $\mathbf{W}^{\phi(1)}$ and $\mathbf{W}^{\phi(2)}$ could be built by W_{ij}^{ϕ}, and the nonlinear version of W_{ij}^{ϕ} is

$$W_{ij}^{\phi} = \begin{cases} \dfrac{1}{D_{ii}^{\phi}}\exp\left(\dfrac{-\left(K_{ii} + K_{jj} - 2K_{ij}\right)}{t}\right) & \text{if } \mathbf{x}_i \in N_k\left(\mathbf{x}_j\right) \\ & \text{or } \mathbf{x}_j \in N_k\left(\mathbf{x}_i\right), \\ & y_i = y_j \\ 0 & \text{otherwise,} \end{cases} \qquad (25)$$

where $D_{ii}^{\phi} = \sum_j W_{ij}^{\phi}$ is a normalizer.

Thus, the kernel FSVM-CIP can be easily achieved by solving the following quadratic problem:

$$\begin{aligned}
\min_{w,b,\rho,\xi} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} - \nu\rho + \frac{1}{\nu_1 m_1}\sum_{i=1}^{m_1}\mu_i\xi_i + \frac{1}{\nu_2 m_2}\sum_{j=m_1+1}^{N}\mu_j\xi_j \\
& + \frac{\eta}{2}\mathbf{w}^T\mathbf{S}_{lw}^{\phi}\mathbf{w} \\
\text{s.t.} \quad & \mathbf{w}^T\phi\left(\mathbf{x}_i\right) + b \geq 1 - \xi_i, \quad i = 1, \ldots, m_1 \\
& \mathbf{w}^T\phi\left(\mathbf{x}_j\right) + b \geq 1 + \rho - \xi_j, \quad j = m_1 + 1, \ldots, N \\
& \xi_k \geq 0, \quad k = 1, \ldots, N, \ \rho \geq 0.
\end{aligned} \qquad (26)$$

Like its linear counterpart, the solution to this optimization problem can be easily found using Lagrange multipliers. By using the representer theorem, $\mathbf{w}$ can be given by $\mathbf{w} = \sum_{i=1}^{N}\beta_i\phi(\mathbf{x}_i)$. We obtain the dual form of the optimization problem:

$$\begin{aligned}
\min_{\alpha} \quad & \frac{1}{2}\boldsymbol{\alpha}^T\mathbf{M}\boldsymbol{\alpha} \\
\text{s.t.} \quad & \sum_{i=1}^{m_1}\alpha_i = \nu \\
& \sum_{j=m_1+1}^{N}\alpha_j = \nu
\end{aligned}$$

$$0 \le \alpha_i \le \frac{\mu_i}{\nu_1 m_1}, \quad i = 1, \ldots, m_1$$
$$0 \le \alpha_j \le \frac{\mu_j}{\nu_2 m_2}, \quad j = m_1 + 1, \ldots, N, \tag{27}$$

where $\mathbf{M} = \mathbf{Y}\mathbf{K}^T\mathbf{Q}^{-1}\mathbf{K}\mathbf{Y}$ and $\mathbf{Q} = \mathbf{K} + \eta\mathbf{K}^{(1)}(\mathbf{I}^{(1)} - \mathbf{W}^{\phi^{(1)}})^T(\mathbf{I}^{(1)} - \mathbf{W}^{\phi^{(1)}})\mathbf{K}^{(1)T} + \eta\mathbf{K}^{(2)}(\mathbf{I}^{(2)} - \mathbf{W}^{\phi^{(2)}})^T(\mathbf{I}^{(2)} - \mathbf{W}^{\phi^{(2)}})\mathbf{K}^{(2)T}$. Vectors $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_N]^T$, and $Y = \mathrm{diag}(y_1, y_2, \ldots, y_n)$ is a diagonal matrix.

Equation (27) is a typical convex quadratic programming problem which is easy to be numerically solved. Suppose $\boldsymbol{\alpha}^* = [\alpha_1^*, \ldots, \alpha_N^*]^T$ can be used to solve the above optimization problem; then the optimal weight vector $\boldsymbol{\beta}^* = \mathbf{Q}^{-1}\mathbf{K}\mathbf{Y}\boldsymbol{\alpha}^*$. Therefore, the optimal thresholds b^* and ρ^* are computed by the following formula:

$$b^* = 1 - \frac{1}{s^+} \sum_{\mathbf{x}_i \in \mathrm{sv}_1} \sum_{j=1}^{N} \beta_j^* y_j K\left(\mathbf{x}_i, \mathbf{x}_j\right), \tag{28}$$

$$\rho^* = -\frac{1}{s^+} \sum_{\mathbf{x}_i \in \mathrm{SV}_1} \sum_{j=1}^{N} \beta_j^* y_j K\left(\mathbf{x}_i, \mathbf{x}_j\right) + \frac{1}{s^-} \sum_{\mathbf{x}_i \in \mathrm{SV}_2} \sum_{j=1}^{N} \beta_j^* y_j K\left(\mathbf{x}_i, \mathbf{x}_j\right). \tag{29}$$

Finally, a more robust decision function of kernel FSVM-CIP will be

$$f(\mathbf{x}) = \mathrm{sgn}\left(\sum_{i=1}^{N} \beta_i^* K(\mathbf{x}, \mathbf{x}_i) + b^*\right). \tag{30}$$

Theorem 6. *The matrix* $\mathbf{M}$ *in* (27) *is symmetric and positive semidefinite.*

A proof of the above theorem can be found in Appendix.

Next, we consider fuzzy membership functions in feature space.

Definition 7. The $\mu_{\mathrm{lin}}^{\phi}$ is called the linear fuzzy membership in feature space and $\mu_{\mathrm{lin}}^{\phi}$ can be defined as

$$\mu_{\mathrm{lin}}^{\phi} = \begin{cases} \dfrac{1 - \left\|\phi(\mathbf{x}_i) - \phi(\overline{\mathbf{x}}_+)\right\|}{\left(\max_j \left(\left\|\phi(\mathbf{x}_j) - \phi(\overline{\mathbf{x}}_+)\right\|\right) + \delta\right)} & \text{if } y_i = 1 \\ \dfrac{1 - \left\|\phi(\mathbf{x}_i) - \phi(\overline{\mathbf{x}}_-)\right\|}{\left(\max_j \left(\left\|\phi(\mathbf{x}_j) - \phi(\overline{\mathbf{x}}_-)\right\|\right) + \delta\right)} & \text{if } y_i = -1, \end{cases} \tag{31}$$

where δ is a small positive value. $\|\cdot\|$ is the Euclidean distance.

Definition 8. The $\mu_{\mathrm{exp}}^{\phi}$ is called the exponential fuzzy membership in feature space and $\mu_{\mathrm{exp}}^{\phi}$ can be defined as

$$\mu_{\mathrm{exp}}^{\phi} = \begin{cases} \dfrac{2}{1 + \exp\left(\lambda \left\|\phi(\mathbf{x}_i) - \phi(\overline{\mathbf{x}}_+)\right\|\right)} & \text{if } y_i = 1 \\ \dfrac{2}{1 + \exp\left(\lambda \left\|\phi(\mathbf{x}_i) - \phi(\overline{\mathbf{x}}_-)\right\|\right)} & \text{if } y_i = -1, \end{cases} \tag{32}$$

where parameter $\lambda \in [0, 1]$ determines the steepness of the decay. Consider

$$\phi(\overline{\mathbf{x}}_+) = \frac{1}{m_1} \sum_{x_i \in c_1} \phi(\mathbf{x}_i),$$
$$\phi(\overline{\mathbf{x}}_-) = \frac{1}{N - m_1} \sum_{x_i \in c_2} \phi(\mathbf{x}_i). \tag{33}$$

Thus, the distance $\|\phi(\mathbf{x}_i) - \phi(\overline{\mathbf{x}}_+)\|$ can be given by

$$\left\|\phi(\mathbf{x}_i) - \phi(\overline{\mathbf{x}}_+)\right\| = \sqrt{K(\mathbf{x}_i, \mathbf{x}_i) - \frac{2}{m_1} \sum_{\mathbf{x}_j \in C_1} K\left(\mathbf{x}_i, \mathbf{x}_j\right) + \frac{1}{m_1^2} \sum_{\mathbf{x}_s \in C_1} \sum_{\mathbf{x}_t \in C_1} K\left(\mathbf{x}_s, \mathbf{x}_t\right)}. \tag{34}$$

Likewise, the $\|\phi(\mathbf{x}_i) - \phi(\overline{\mathbf{x}}_-)\|$ can be given in a similar manner.

4.2. Solution. Based on the above, we can state the approach of kernel FSVM-CIP as Algorithm 2.

5. Experiments and Discussions

To evaluate the performance of our proposed FSVM-CIP, in this section, FSVM-CIP is evaluated compared with other related representative methods, such as standard FSVM [8], SVDD [11], FSVM for class imbalance learning (FSVM-CIL) [22], and FSVM with minimum within-class scatter (WCS-FSVM) [23]. We implement FSVM-CIP using the linear fuzzy membership and the exponential fuzzy membership, respectively, which are represented as FSVM-CIP$_{\mathrm{lin}}$ and FSVM-CIP$_{\mathrm{exp}}$. All the experiments are performed in Matlab (R2010a) on personal computer, whose configuration is as follows: CPU 2.99 GHz, 4.0 G RAM, and Microsoft Windows XP.

5.1. Data Preparation. In this section, we use five real-world medical datasets from the UCI repository of machine learning database [24], to demonstrate the classification performance of the method proposed in this paper. These five medical datasets are breast, heart, hepatitis, BUPA liver, and pima diabetes. It is highly likely that these real-world datasets contain some outliers and noisy examples in different amounts [22]. In each of them, the positive class consists of the data corresponding to the healthy, normal, or benign cases, while the negative class contains the data for diseased, abnormal, or malignant cases. Further details of these datasets are provided in Table 1. This contains the total number of positive data #pos, the total number of negative data #neg, the number of positive training examples $m1$, the number of negative training examples $m2$, the positive-to-negative imbalance ratio Ratio, and the data dimensionality d.

5.2. Performance Measure and Experimental Settings. We used the geometric mean of sensitivity (sensitivity = proportion of the positives correctly recognized), specificity (specificity = proportion of the negatives correctly recognized),

Input:
training samples $\{(\mathbf{x}_i, y_i), i = 1, \ldots, N\}$
Testing samples $\{\mathbf{x}_j, j = 1, \ldots, U\}$
Output:
The predicted labels y_j of data $\{\mathbf{x}_j, j = 1, \ldots, U\}$
Procedure:
(1) Choose a kernel function **K**. Compute the Gram matrix $K_{ij} = K(\mathbf{x}_i, \mathbf{x}_j)$.
(2) Compute fuzzy membership μ_i^{ϕ} using (31) or (32) for the data $\{(\mathbf{x}_i, y_i), i = 1, \ldots, N\}$
(3) Construct data adjacency graph G using k nearest neighbors and compute the edge weights matrix W_{ij}^{ϕ} with N examples
(4) Construct local within-class preserving scatter matrix $\mathbf{S}_{\text{lw}}^{\phi}$ using (24)
(5) Choose parameters t (25); η, ν, ν_1 and ν_2 (26)
(6) Compute α^* using (27) and b^* using (28) with a QP Solver
(7) Using decision function (30) with samples $\mathbf{x}_j$, and output the final class labels

Algorithm 2: Kernel FSVM-CIP.

Table 1: Characteristics of the selected datasets.

Datasets	#pos	#neg	$m1$	$m2$	Ratio	d
Breast	458	241	240	120	2 : 1	9
Heart	120	150	80	20	4 : 1	13
Hepatitis	123	32	100	10	10 : 1	19
BUPA liver	200	145	150	10	15 : 1	6
Pima diabetes	268	500	180	10	18 : 1	8

and accuracy (accuracy = proportion of correctly classified instances) for the classifier performance evaluation in experiments, as commonly used in medical datasets classification research [7].

Like the existing SVM and FSVM algorithms, the solution is sensitive to the setting of the parameters. In order to evaluate the performance, a strategy is that a set of the parameters is given first and then the best cross-validation mean rate among the set is used to estimate the generalized accuracy. We adopt this strategy in this paper. For FSVM-CIP, the parameter ν is searched in $\{1, 5, 10, 15, \ldots, 80\}$, while ν_1 and ν_2 are selected from $\{0.001, 0.005, 0.01, 0.05\}$. η is selected from $\log_2\eta \in \{-5, -4.5, -4, \ldots, 5.5, 6\}$. The heat kernel parameter t is searched in $\{0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0\}$ and the neighborhood parameter k is searched in $\{3, 5, 7, 9, 11, 13, 15\}$. In addition, when the linear fuzzy function is used, we set $\delta = 10^{-6}$. When the exponential fuzzy function is used, the optimal value of λ is chosen from the range $\lambda = \{0.1, 0.2, 0.3, \ldots, 1\}$.

The regularization parameter C for FSVM, SVDD, FSVM-CIL, and WCS-FSVM is selected from the set $\{0.001, 0.01, 0.1, 1, 10, 100\}$. In WCS-FSVM, β is selected from $\log_2\beta \in \{-5, -4.5, -4, \ldots, 5.5, 6\}$. For FSVM-CIL, the fuzzy membership is based on the distance from the actual hyperplane and uses the exponential fuzzy membership λ. λ is chosen from the range $\lambda = \{0.1, 0.2, 0.3, \ldots, 1\}$.

For the kernel-based methods, we use a Gaussian RBF kernel, that is, $\exp(-(u - v)^T(u - v)/\sigma)$, where σ is the spread of Gaussian kernel, and σ is searched in $\{\tau^2/16, \tau^2/8, \tau^2/4, \tau^2/2, \tau^2, 2\tau^2, 4\tau^2, 8\tau^2, 16\tau^2\}$, where τ^2 is the mean norm of the training data.

For parameter selection, we conduct fivefold cross-validation in a stratified manner so that each validation set has the same positive to negative ratio as in the training set. Finally, the experiment is repeated 10 times independently of each dataset.

5.3. Experimental Results. FSVM-CIP method test results developed for the breast, heart, hepatitis, BUPA liver, and pima diabetes datasets are given both in the linear case and nonlinear case. Tables 2, 3, 4, 5, and 6 display the comparison results with the other methods on these five databases, respectively.

The main observations from the performance comparisons include the following.

(1) We can see that, in many real-world applications, a linear classifier seems powerless. In terms of accuracy, kernel method can improve the classification performance for all five medical datasets.

(2) We can clearly observe that the FSVM-CIP outperforms other methods on almost datasets both in the linear case and nonlinear case, which gives higher accuracy. This fortifies the fact that the locality maximum margin and the local structure information presented by local within-class preserving scatter could improve classification performance; furthermore, the method of different misclassification costs based on the number of two classes is a sensitive learning solution to overcome the imbalance problem in SVMs.

(3) It is noted that, for all the datasets considered, the classification accuracy given by the FSVM-CIP$_{\text{exp}}$ setting is higher than the FSVM-CIP$_{\text{lin}}$ setting. Therefore, we can state that FSVM-CIP$_{\text{exp}}$ setting with the appropriate selection

Table 2: Comparison of the classification results (%) on breast dataset.

	Method	Sensitivity	Specificity	Accuracy
Linear	FSVM	95.87 ± 0.017	95.04 ± 0.043	95.58 ± 0.035
	SVDD	97.71 ± 0.065	90.90 ± 0.013	95.28 ± 0.052
	FSVM-CIL	95.87 ± 0.024	95.87 ± 0.015	95.81 ± 0.028
	WCS-FSVM	96.33 ± 0.067	95.04 ± 0.056	95.87 ± 0.047
	FSVM-CIP$_{lin}$	96.98 ± 0.039	96.49 ± 0.022	**96.76 ± 0.040**
	FSVM-CIP$_{exp}$	96.68 ± 0.011	96.69 ± 0.042	**96.76 ± 0.037**
Gaussian kernel	FSVM	96.33 ± 0.023	95.87 ± 0.051	96.17 ± 0.050
	SVDD	97.30 ± 0.065	91.25 ± 0.013	95.44 ± 0.052
	FSVM-CIL	96.79 ± 0.059	95.87 ± 0.042	96.46 ± 0.055
	WCS-FSVM	96.97 ± 0.030	96.69 ± 0.093	96.76 ± 0.067
	FSVM-CIP$_{lin}$	97.25 ± 0.055	96.29 ± 0.032	97.05 ± 0.042
	FSVM-CIP$_{exp}$	97.25 ± 0.055	97.52 ± 0.045	**97.34 ± 0.033**

Table 3: Comparison of the classification results (%) on heart dataset.

	Method	Sensitivity	Specificity	Accuracy
Linear	FSVM	87.50 ± 0.080	80.77 ± 0.069	82.35 ± 0.069
	SVDD	87.03 ± 0.021	77.69 ± 0.005	80.00 ± 0.051
	FSVM-CIL	85.00 ± 0.046	82.04 ± 0.110	82.35 ± 0.072
	WCS-FSVM	87.30 ± 0.071	81.54 ± 0.089	82.94 ± 0.088
	FSVM-CIP$_{lin}$	85.00 ± 0.063	82.31 ± 0.083	82.84 ± 0.054
	FSVM-CIP$_{exp}$	87.50 ± 0.025	82.31 ± 0.083	**83.53 ± 0.055**
Gaussian kernel	FSVM	86.70 ± 0.099	82.61 ± 0.087	83.35 ± 0.042
	SVDD	90.35 ± 0.022	80.77 ± 0.034	82.80 ± 0.070
	FSVM-CIL	87.05 ± 0.034	81.54 ± 0.067	82.94 ± 0.044
	WCS-FSVM	91.00 ± 0.076	81.73 ± 0.083	84.12 ± 0.085
	FSVM-CIP$_{lin}$	90.00 ± 0.045	82.31 ± 0.086	84.12 ± 0.052
	FSVM-CIP$_{exp}$	86.05 ± 0.023	83.08 ± 0.078	**84.71 ± 0.066**

Table 4: Comparison of the classification results (%) on hepatitis dataset.

	Method	Sensitivity	Specificity	Accuracy
Linear	FSVM	82.60 ± 0.053	22.73 ± 0.087	53.33 ± 0.073
	SVDD	73.91 ± 0.071	45.45 ± 0.011	60.00 ± 0.046
	FSVM-CIL	77.66 ± 0.026	45.46 ± 0.082	61.02 ± 0.070
	WCS-FSVM	79.56 ± 0.107	27.27 ± 0.062	53.33 ± 0.059
	FSVM-CIP$_{lin}$	78.26 ± 0.046	45.46 ± 0.032	62.22 ± 0.023
	FSVM-CIP$_{exp}$	78.26 ± 0.068	50.00 ± 0.086	**64.44 ± 0.071**
Gaussian kernel	FSVM	73.91 ± 0.038	31.82 ± 0.012	53.33 ± 0.025
	SVDD	82.60 ± 0.053	42.86 ± 0.025	63.64 ± 0.030
	FSVM-CIL	77.26 ± 0.041	50.00 ± 0.086	63.84 ± 0.064
	WCS-FSVM	78.26 ± 0.015	36.36 ± 0.074	57.78 ± 0.056
	FSVM-CIP$_{lin}$	73.51 ± 0.064	54.55 ± 0.037	64.44 ± 0.058
	FSVM-CIP$_{exp}$	73.91 ± 0.050	59.10 ± 0.011	**66.67 ± 0.036**

of λ value would be an effective choice applied to any medical dataset. In other words, when dealing with medical datasets classification, the performance of the exponential fuzzy membership is better than linear fuzzy membership in FSVM-CIP.

(4) For breast and heart datasets, the class imbalance is not obviously shaped; WCS-FSVM yielded standard FSVM, SVDD, and FSVM-CIL. We can say that the performance can indeed be improved when the structure of the data is taken into consideration. For the other three datasets, the class

Table 5: Comparison of the classification results (%) on BUPA liver dataset.

	Method	Sensitivity	Specificity	Accuracy
Linear	FSVM	88.10 ± 0.008	66.42 ± 0.073	72.19 ± 0.057
	SVDD	87.27 ± 0.021	68.05 ± 0.063	72.72 ± 0.042
	FSVM-CIL	88.00 ± 0.004	67.44 ± 0.042	73.19 ± 0.015
	WCS-FSVM	84.00 ± 0.360	67.15 ± 0.068	71.66 ± 0.051
	FSVM-CIP$_{lin}$	88.00 ± 0.004	67.88 ± 0.063	73.26 ± 0.031
	FSVM-CIP$_{exp}$	86.00 ± 0.048	69.34 ± 0.072	**73.80 ± 0.054**
Gaussian kernel	FSVM	96.00 ± 0.057	66.67 ± 0.026	74.60 ± 0.038
	SVDD	95.43 ± 0.033	71.24 ± 0.050	77.23 ± 0.017
	FSVM-CIL	95.00 ± 0.045	72.59 ± 0.052	78.37 ± 0.050
	WCS-FSVM	90.08 ± 0.070	67.44 ± 0.083	73.73 ± 0.062
	FSVM-CIP$_{lin}$	94.00 ± 0.049	74.10 ± 0.045	79.46 ± 0.048
	FSVM-CIP$_{exp}$	94.00 ± 0.049	73.33 ± 0.084	**79.92 ± 0.074**

Table 6: Comparison of the classification results (%) on pima diabetes dataset.

	Method	Sensitivity	Specificity	Accuracy
Linear	FSVM	91.91 ± 0.022	49.98 ± 0.053	55.36 ± 0.051
	SVDD	88.65 ± 0.081	53.43 ± 0.062	58.45 ± 0.029
	FSVM-CIL	86.36 ± 0.064	55.10 ± 0.059	59.86 ± 0.060
	WCS-FSVM	87.50 ± 0.043	52.65 ± 0.024	57.96 ± 0.030
	FSVM-CIP$_{lin}$	85.23 ± 0.021	57.76 ± 0.064	**61.94 ± 0.043**
	FSVM-CIP$_{exp}$	84.09 ± 0.009	57.96 ± 0.062	**61.94 ± 0.053**
Gaussian kernel	FSVM	93.18 ± 0.031	51.02 ± 0.073	57.44 ± 0.053
	SVDD	91.76 ± 0.025	56.86 ± 0.052	62.57 ± 0.028
	FSVM-CIL	90.91 ± 0.047	58.78 ± 0.084	63.67 ± 0.077
	WCS-FSVM	92.05 ± 0.010	54.69 ± 0.066	60.38 ± 0.053
	FSVM-CIP$_{lin}$	88.84 ± 0.040	61.38 ± 0.063	**65.57 ± 0.063**
	FSVM-CIP$_{exp}$	88.64 ± 0.029	61.43 ± 0.074	**65.57 ± 0.070**

imbalance strikingly improved, the results given by standard FSVM and WCS-FSVM for datasets are biased towards the majority class represented as lower specificity and lower accuracy. These results justify the fact that these two methods are sensitive to the class imbalance problem. Meanwhile, SVDD and FSVM-CIL yielded standard FSVM and WCS-FSVM. BY assigning different misclassification costs for the minority class and majority class, the effect of class imbalance could be reduced.

5.4. Parameter Selection for Kernel FSVM-CIP$_{exp}$. The parameter $\eta > 0$ is an essential parameter in our proposed method which controls the tradeoff between the local within-class scatter and the margin. Figure 2 shows the impact of parameter η on the classification accuracy of FSVM-CIP$_{exp}$ in kernel case with each value of η selected from $\log_2\eta \in \{-5, -4.5, -4, \ldots, 5.5, 6\}$. It can be seen that the best accuracy is obtained for all the datasets and therefore η is searched in a reasonable range.

Compared with standard FSVM, the additional neighbor parameter k is employed in FSVM-CIP. To evaluate the influence of this parameter on the performance, the classification accuracy of kernel FSVM-CIP$_{exp}$ for five medical databases is recorded for each value of k in $\{3, 5, 7, 9, 11, 13, 15\}$. Figure 3 shows the results. It can be seen that the classification accuracy is not high when k value is small and, by increasing k, the classification accuracy increases; however, if k continues to increase, the classification accuracy begins to drop severely down. It is because, when k is too small, the number of nearest neighbors is sparse; when k is too large, the number of nearest neighbors is excessive, so to preserve so much local relation may be inappropriate.

6. Conclusion

Computer tools have improved the medical practice implementation to a greater extent. Although computer tools cannot replace the doctors, they can make their work easier and more effective. In this paper, a new fuzzy support machine called FSVM-CIP, used for medical datasets classification, is proposed. The proposed method is based on local within-class preserving scatter and assigned two misclassification costs in the SVM objective function, which is for learning from imbalance datasets in the presence of outliers/noise and enhancing the locality maximum margin. Experiments were performed on several UCI medical datasets with a comparison of the proposed method with several other related methods such as standard FSVM, SVDD, FSVM-CIL, and

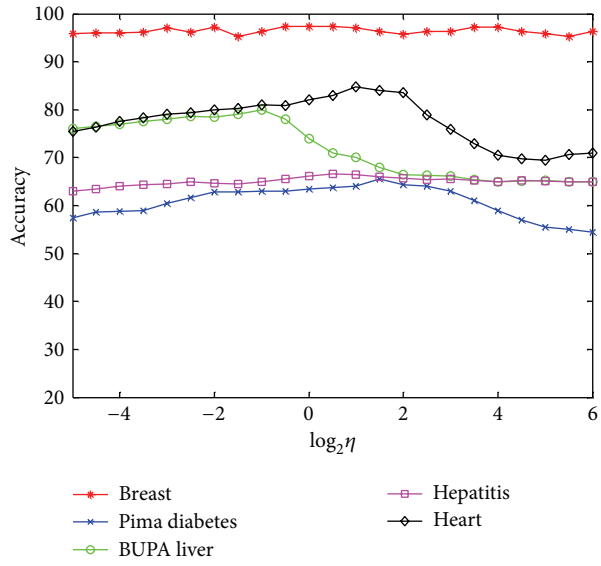

Figure 2: The effect of the parameter η on kernel FSVM-CIP$_{exp}$.

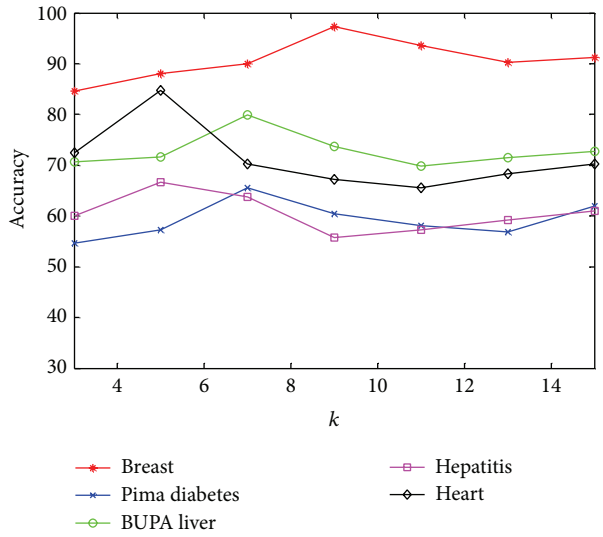

Figure 3: The effect of the parameter k on kernel FSVM-CIP$_{exp}$.

WCS-FSVM. Obtained results show that the performance of the proposed method is highly successful compared to other results attained and seems very promising. Finally, we can recommend that FSVM-CIP$_{exp}$ which uses the exponential fuzzy membership would be an effective choice for medical datasets classification applications. In future work, we intend to perform investigations to large-scale classification problems.

Appendix

Proof of Theorem 3 in Section 3.2.

Proof. According to the dual form of the optimization problem (15), we can derive

$$\sum_{i=1}^{m_1}\alpha_i = \nu. \tag{A.1}$$

Likewise, according to the KKT conditions, $\sum_{i=1}^{N}\alpha_i = \nu$ with $\rho > 0$ satisfy $s = 0$ by (12). According to (11), all samples with $\boldsymbol{\xi}_i > 0$ satisfy $\gamma_i = 0$. In view of (13), this implies that $\alpha_i = \mu_i/\nu_1 m_1$ holds for every positive ME. Summing up α_i over the positive MEs using (A.1), we have

$$\frac{\overline{\mu_m^+}m^+}{\nu_1 m_1} \le \sum_{i=1}^{m_1}\alpha_i = \nu. \tag{A.2}$$

Furthermore, in view of (15), each SV in the positive class can control at most $1/\nu_1 m_1$ to the $\sum_{i=1}^{m_1}\alpha_i$; as a result,

$$\sum_{i=1}^{m_1}\alpha_i \le \frac{\overline{\mu_s^+}s^+}{\nu_1 m_1}. \tag{A.3}$$

Combining (A.2) and (A.3), inequality (20) can hold true. Likewise, inequality (21) can be proven in a similar manner. □

Proof of Theorem 6 in Section 4.1.

Proof. We know that $\mathbf{M} = \mathbf{Y}\mathbf{K}^T\mathbf{Q}^{-1}\mathbf{K}\mathbf{Y}$, and $\mathbf{K}$ is a Gram matrix, so $\mathbf{K}$ is symmetric and positive semidefinite. The transpose of the matrix $\mathbf{Q}$ is

$$\begin{aligned}\mathbf{Q}^T = &\left(\mathbf{K} + \eta\mathbf{K}^{(1)}\left(\mathbf{I}^{(1)} - \mathbf{W}^{\phi^{(1)}}\right)^T\left(\mathbf{I}^{(1)} - \mathbf{W}^{\phi^{(1)}}\right)\mathbf{K}^{(1)T}\right.\\ &\left.+\eta\mathbf{K}^{(2)}\left(\mathbf{I}^{(2)} - \mathbf{W}^{\phi^{(2)}}\right)^T\left(\mathbf{I}^{(2)} - \mathbf{W}^{\phi^{(2)}}\right)\mathbf{K}^{(2)T}\right)^T = \mathbf{Q}.\end{aligned} \tag{A.4}$$

So $\mathbf{Q}$ is a symmetric matrix and then $\mathbf{M}$ is symmetric. Set $\mathbf{Q} = \mathbf{K} + \eta\mathbf{R}$, where

$$\begin{aligned}\mathbf{R} = &\mathbf{K}^{(1)}\left(\mathbf{I}^{(1)} - \mathbf{W}^{\phi^{(1)}}\right)^T\left(\mathbf{I}^{(1)} - \mathbf{W}^{\phi^{(1)}}\right)\mathbf{K}^{(1)T}\\ &+\mathbf{K}^{(2)}\left(\mathbf{I}^{(2)} - \mathbf{W}^{\phi^{(2)}}\right)^T\left(\mathbf{I}^{(2)} - \mathbf{W}^{\phi^{(2)}}\right)\mathbf{K}^{(2)T}.\end{aligned} \tag{A.5}$$

For any nonzero vector $\mathbf{u} = (u_1, u_2, \ldots, u_N)^T$,

$$\begin{aligned}\mathbf{u}^T\mathbf{R}\mathbf{u} = &\mathbf{u}^T\mathbf{K}^{(1)}\left(\mathbf{I}^{(1)} - \mathbf{W}^{\phi^{(1)}}\right)^T\left(\mathbf{I}^{(1)} - \mathbf{W}^{\phi^{(1)}}\right)\mathbf{K}^{(1)T}\mathbf{u}\\ &+\mathbf{u}^T\mathbf{K}^{(2)}\left(\mathbf{I}^{(2)} - \mathbf{W}^{\phi^{(2)}}\right)^T\left(\mathbf{I}^{(2)} - \mathbf{W}^{\phi^{(2)}}\right)\mathbf{K}^{(2)T}\mathbf{u}\\ = &\boldsymbol{\varsigma}^T\mathbf{S}_{lw}^{\phi}\boldsymbol{\varsigma} \ge 0,\end{aligned} \tag{A.6}$$

where $\boldsymbol{\varsigma} = \sum_{i=1}^{N}u_i\phi(\mathbf{x}_i)$. The local within-class scatter matrix $\mathbf{S}_{lw}^{\phi}$ is semidefinite, so the matrix $\mathbf{R}$ is semidefinite. That is, the matrix $\mathbf{Q}$ is semidefinite, and then $\mathbf{M}$ is semidefinite. □

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

This work was supported by the National Natural Science Foundation of China under Contact (61070121).

References

[1] S. Ghumbre, C. Patil, and A. Ghatol, "Heart disease diagnosis using support vector machine," in *Proceedings of the International Conference on Computer Science and Information Technology (ICCSIT '11)*, Pattaya, Thailand, 2011.

[2] M. Tong, K. Liu, C. Xu, and W. Ju, "An ensemble of SVM classifiers based on gene pairs," *Computers in Biology and Medicine*, vol. 43, no. 6, pp. 729–737, 2013.

[3] Y. Chang, N. Kim, Y. Lee, J. Lim, and J. B. Seo, "Fast and efficient lung disease classification using hierarchical one-against-all support vector machine and cost-sensitive feature selection," *Computers in Biology and Medicine*, vol. 42, no. 12, pp. 1157–1164, 2012.

[4] D. C. Li, C. W. Liu, and S. C. Hu, "A fuzzy-based data transformation for feature extraction to increase classification performance with small medical data sets," *Artificial Intelligence in Medicine*, vol. 52, no. 1, pp. 45–52, 2011.

[5] M. Serter, U. N. Yilmaz, and O. Inan, "Feature selection method based on artificial bee colony algorithm and support vector machines for medical datasets classification," *The Scientific World Journal*, vol. 2013, Article ID 419187, 10 pages, 2013.

[6] V. N. Vapnik, *The Natural of Statistical Learning Theory*, Springer, New York, NY, USA, 1995.

[7] V. N. Vapnik, *Statistical Learning Theory*, Wiley, New York, NY, USA, 1998.

[8] C. F. Lin and S. D. Wang, "Fuzzy support vector machines," *IEEE Transactions on Neural Networks*, vol. 13, no. 2, pp. 464–471, 2002.

[9] K. Veropoulos, C. Campbell, and N. Cristianini, "Controlling the sensitivity of support vector machines," in *Proceedings of the International Joint Conferences on Artificial Intelligence*, pp. 55–60, Stockholm, Sweden, 1999.

[10] B. Schölkopf and A. J. Smola, *Learning with Kernels*, MIT Press, 2002.

[11] D. M. J. Tax and R. P. W. Duin, "Support vector data description," *Machine Learning*, vol. 54, no. 1, pp. 45–66, 2004.

[12] E. Kokiopoulou and Y. Saad, "Orthogonal neighborhood preserving projections: a projection-based dimensionality reduction technique," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 12, pp. 2143–2156, 2007.

[13] X. He and P. Niyogi, "Locality preserving projections," in *Proceedings of the Conference on Advances in Neural Information Processing Systems*, pp. 585–591, 2003.

[14] X. He, S. Yan, Y. Hu, P. Niyogi, and H. J. Zhang, "Face recognition using Laplacianfaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 3, pp. 328–340, 2005.

[15] X. Wang and Y. Niu, "New one-versus-all ν-SVM solving intra-inter class imbalance with extended manifold regularization and localized relative maximum margin," *Neural Computing*, vol. 115, no. 9, pp. 106–121, 2013.

[16] S. Zafeiriou, A. Tefas, and I. Pitas, "Minimum class variance support vector machines," *IEEE Transactions on Image Processing*, vol. 16, no. 10, pp. 2551–2564, 2007.

[17] I. Kotsia, S. Zafeiriou, and I. Pitas, "Novel multiclass classifiers based on the minimization of the within-class variance," *IEEE Transactions on Neural Networks*, vol. 20, no. 1, pp. 14–34, 2009.

[18] M. Wu and J. Ye, "A small sphere and large margin approach for novelty detection using training data with outliers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 11, pp. 2088–2092, 2009.

[19] X. Jiang, Z. Yi, and J. C. Lv, "Fuzzy SVM with a new fuzzy membership function," *Neural Computing and Applications*, vol. 15, no. 3-4, pp. 268–276, 2006.

[20] K. R. Müller, S. Mika, G. Rätsch, K. Tsuda, and B. Schölkopf, "An introduction to kernel-based learning algorithms," *IEEE Transactions on Neural Networks*, vol. 12, no. 2, pp. 181–201, 2001.

[21] G. Wahba, "Support vector machines, reproducing kernel hilbert spaces and the randomized gacv," in *Advances in Kernel Methods: Support Vector Learning*, MIT Press, Cambridge, Mass, USA, 1998.

[22] R. Batuwita and V. Palade, "FSVM-CIL: fuzzy support vector machines for class imbalance learning," *IEEE Transactions on Fuzzy Systems*, vol. 18, no. 3, pp. 558–571, 2010.

[23] W. An and M. Liang, "Fuzzy support vector machine based on within-class scatter for classification problems with outliers or noises," *Neural Computing*, vol. 110, no. 6, pp. 101–110, 2013.

[24] "UCI Repository of machine learning database," http://www.ics.uci.edu/%20mlearn/MLRepository.html.