Neural Computing and Applications (2019) 31:7153–7164
https://doi.org/10.1007/s00521-018-3551-9

ORIGINAL ARTICLE

A fuzzy twin support vector machine based on information entropy for class imbalance learning

Deepak Gupta[1] · Bharat Richhariya[1] · Parashjyoti Borah[1]

Received: 27 March 2017 / Accepted: 14 May 2018 / Published online: 24 May 2018

Abstract

In real-world binary class datasets, the total number of samples may not be the same in both the classes, i.e. size of the majority class is much larger than minority class which is called as imbalance problem. In various classification problems, the main interest is to correctly classify the samples belonging to the minority class. Since support vector machine (SVM) and twin support vector machine (TWSVM) obtain the resultant classifier by giving same importance to all the training samples, it results in a biased classifier towards the majority class in imbalanced datasets. In this paper, by considering the fuzzy membership value for each sample, we have proposed an efficient approach, entropy-based fuzzy twin support vector machine for class imbalanced datasets (EFTWSVM-CIL) where fuzzy membership values are assigned based on the entropy values of samples. Here, we give more importance to the minority class by assigning relatively larger fuzzy memberships to the minority class samples. Further, it solves a pair of smaller-size quadratic programming problems (QPPs) rather than a large one as in the case of SVM. Experiments are performed on various real-world imbalanced datasets, and results of our proposed EFTWSVM-CIL are compared with twin support vector machine (TWSVM), fuzzy twin support vector machine (FTWSVM) and entropy-based fuzzy SVM (EFSVM) for imbalanced datasets.

Keywords Information entropy · Imbalanced dataset · Fuzzy membership · K-nearest neighbour (K-NN) · Twin support vector machine (TWSVM)

1 Introduction

In recent years, many machine learning and data mining techniques have been introduced to solve the classification and regression problems. If a particular dataset is having equal number of samples of each class, then it is called a balanced dataset; otherwise, it is an imbalanced dataset. It is not easy to solve the imbalance problem for classification. Support vector machine (SVM) is one of the most popular machines learning approach which is based on structural risk minimization (SRM) principle [1–3]. It solves a quadratic programming problem and always provides a globally optimal, relatively robust and sparse solution, whereas techniques like artificial neural network (ANN) is based on empirical risk minimization (ERM) principle and has local minima problem. SVM has been used in applications such as face recognition [4–6], pattern recognition [7, 8], speaker verification [9], intrusion detection [10] and various other classification problems [11–14].

SVM finds the resultant classifier by maximizing the margin between the support vectors and decision boundary, meanwhile improving the generalization ability. One can notice that SVM provides better generalization performance, but the training cost of SVM is very high i.e. $O(m^3)$ where m is the total number of training samples [15]. Recently, an efficient approach twin support vector machine (TWSVM) is proposed by Jayadeva et al. [15] to decrease the training cost of SVM. In TWSVM, two quadratic programming problems of smaller size are solved

✉ Deepak Gupta
deepakjnu85@gmail.com; deepak.cse@nitap.in

Bharat Richhariya
bharatrichhariya2@yahoo.com

Parashjyoti Borah
parashjyoti@hotmail.com

[1] Department of Computer Science & Engineering, National Institute of Technology, Papum Pare, Arunachal Pradesh, India

to find the solution rather than a single large problem as in SVM.

SVM is a supervised machine learning algorithms which constructs a model depending on the available number of samples of each class. Due to some imbalance in the dataset, samples belonging to the minority class get misclassified since they cannot contribute much in the training phase of the method. Thus, the classifier becomes biased towards the majority class. Here, the class of interest is the minority class; therefore, giving more weights to the data points of minority class resolves this problem to some extent. In applications such as fault detection and disease detection, more emphasis is on correctly identifying the faults in machinery and abnormalities in the patients data which are present in very few samples.

To address this problem, Lin et al. [16] proposed a support vector machine based on fuzzy membership values (FSVM). Similar to SVM, FSVM also suffers from the problem of class imbalance. Batuwita and Palade [17] have presented a new model as FSVMs for class imbalance learning (FSVM-CIL) to handle the problem of class imbalance which is less sensitive to outliers and noise. Here, the smaller fuzzy membership values are assigned to support vectors to reduce the effect of support vectors on the resultant decision surface based on class centres. In a similar manner, a new efficient approach fuzzy support vector machine for non-equilibrium data is proposed [18] to reduce the misclassification accuracy of minority class in FSVM. A new approach, Bilateral-weighted FSVM (B-FSVM) is proposed [19] where membership of each sample is calculated by considering the samples as belonging to minority and majority class with different membership values. To solve bankruptcy prediction problem, a new fuzzy SVM is proposed by Chaudhuri and De [20]. In order to reduce the complexity of TWSVM for large-scale data, Shao et al. [21] proposed a weighted linear loss twin support vector machine for imbalanced probelm (WLTSVM) where linear equations are solved and lesser weights are given to the points having high loss values. A fuzzy-based Lagrangian twin parametric-margin support vector machine (FLTPMSVM) is proposed by Gupta et al. [22] to deal with noisy data. Tomar et al. [12] assigned weights to the data points on the basis of number of samples in each class and proposed a weighted least squares twin support vector machine (WLSTVM). In this, all the samples of each class are assigned the same weight. For more efficient classification methods, reader may see [23, 24].

Recently, Fan et al. [25] proposed an entropy-based fuzzy support vector machine (EFSVM) for class imbalance problem in which fuzzy membership is computed based on the class certainty of samples. Motivated by the work of Fan et al. [25] and Jayadeva et al. [15], we propose a new approach termed as entropy-based fuzzy twin support vector machine (EFTWSVM-CIL) to solve the class imbalance problem. One can notice that EFTWSVM-CIL solves a pair of smaller-size QPPs to find the resultant decision surface rather than solving a single large one in case of SVM. Hence, EFTWSVM-CIL improves the generalization of the decision surface for minority class samples based on class certainty and also takes less training time.

In this paper, all vectors are considered as column vectors. Suppose x and z are the vector in $n-$ dimensional real space R^n then the inner product of two vectors is denoted as: $x^t z$ where x^t is the transpose of x. $||x||$ and $||Q||$ will be the 2-norm of a vector x and a matrix Q, respectively. The identity matrix of appropriate size and the vector of dimension m are denoted by I and e, respectively.

The paper is organized as follows: Sect. 2 is to give a review on the work related to Support Vector Machine discussing Twin Support Vector Machine (TWSVM), Fuzzy Twin Support Vector Machine (FTWSVM) and Entropy Fuzzy Support Vector Machine (EFSVM). The proposed method is discussed in Sect. 3. Several numerical experiments have been performed on well-known real-world dataset for the discussed and proposed variant of SVM in Sect. 4. In Sect. 5, we conclude the paper with future work.

2 Related Work

In this section, we briefly describe the formulations of twin support vector machine (TWSVM), fuzzy twin support vector machine (FTWSVM) and entropy support vector machine (EFSVM).

2.1 Twin Support Vector Machine (TWSVM)

Mangasarian and Wild [26] extended the idea of proximal SVM (PSVM) [27] to a new approach termed as multi-surface proximal SVM via generalized eigenvalues (GEPSVM) for binary classification. In order to improve the learning efficiency, Jayadeva et al. [15] suggested a novel approach as Twin Support Vector Machine (TWSVM) in the light of GEPSVM. In TWSVM, two non-parallel hyperplanes are obtained instead of one hyperplane such that each of them is nearer to one of the class and as far as possible from the other class. Here, two optimization problems of smaller size are solved in form of QPPs instead of solving a large QPP as in the case of standard SVM. The running time of TWSVM is given as $\left\{2 \times \left(\frac{m}{2}\right)^3 = \frac{m^3}{4}\right\}$ which is a reduction of four times as compared to standard SVM.

Let us consider the input matrices X_1 and X_2 of size $p \times n$ and $q \times n$ where p is the total number of data point belonging to 'Class 1' and q are the total number of data points belonging to 'Class 2' such that total number of data samples $m = p + q$ and n is the dimension of each data points. In nonlinear case, twin support vector machine finds a pair of non-parallel hyperplanes $f_1(x) = K(x^t, D^t)w_1 + b_1 = 0$ and $f_2(x) = K(x^t, D^t)w_2 + b_2 = 0$ from the solution of the following QPPs as

$$\min \frac{1}{2}||K(X_1, D^t)w_1 + e_1 b_1||^2 + C_1 e_2^t \xi$$

subject to

$$\begin{aligned} &-(K(X_2, D^t)w_1 + e_2 b_1) + \xi \geq e_2, \xi \geq 0 \\ &\min \frac{1}{2}||K(X_2, D^t)w_2 + e_2 b_2||^2 + C_2 e_1^t \eta \end{aligned} \tag{1}$$

subject to

$$(K(X_1, D^t)w_2 + e_1 b_2) + \eta \geq e_1, \eta \geq 0 \tag{2}$$

where ξ,η represent slack variables; C_1, C_2 are penalty parameters; $D = [\mathrm{X}_1; \mathrm{X}_2]$; $\boldsymbol{e_1}$,$\boldsymbol{e_2}$ are vectors of suitable dimension having all values as 1's; and $K(x^t, D^t) = (k(x, x_1), \ldots, k(x, x_m))$ is a row vector in R^m.

The Lagrangian of problems (1) and (2) is written as

$$\begin{aligned} L_1 = &\frac{1}{2}||K(X_1, D^t)w_1 + e_1 b_1||^2 + C_1 e_2^t \xi \\ &+ \alpha_1^t((K(X_2, D^t)w_1 + e_2 b_1) - \xi + e_2) - \beta_1^t \xi \end{aligned} \tag{3}$$

$$\begin{aligned} L_2 = &\frac{1}{2}||K(\mathrm{X}_2, \mathrm{D}^t)w_2 + e_2 b_2||^2 + C_2 e_1^t \eta \\ &+ \alpha_2^t((-K(\mathrm{X}_1, \mathrm{D}^t)w_2 - e_1 b_2) - \eta + e_1) - \beta_2^t \eta \end{aligned} \tag{4}$$

where $\alpha_1 = (\alpha_{11}, \ldots, \alpha_{1q})^t$, $\beta_1 = (\beta_{11}, \ldots, \beta_{1q})^t$, $\alpha_2 = (\alpha_{21}, \ldots, \alpha_{2p})^t$, and $\beta_2 = (\beta_{21}, \ldots, \beta_{2p})^t$ are the vectors of Lagrange multipliers. The Wolfe dual of Eqs. (3) and (4) is written by applying the Karush–Kuhn–Tucker (K.K.T) necessary and sufficient conditions [28] as

$$\max e_2^t \alpha_1 - \frac{1}{2}\alpha_1^t T(S^t S)^{-1} T^t \alpha_1 \tag{5}$$

subject to

$$0 \leq \alpha_1 \leq C_1$$

$$\max e_1^t \alpha_2 - \frac{1}{2}\alpha_2^t S(T^t T)^{-1} S^t \alpha_2 \tag{6}$$

subject to

$$0 \leq \alpha_2 \leq C_2$$

where $S = [K(X_1, D^t) \;\; e_1]$ and $T = [K(X_2, D^t) \;\; e_2]$.

We compute the nonlinear hyperplanes $K(x^t, D^t)w_1 + b_1 = 0$ and $K(x^t, D^t)w_2 + b_2 = 0$ by computing the value of w_1,w_2, b_1 and b_2 using Eqs. (7) and (8)

$$\begin{bmatrix} w_1 \\ b_1 \end{bmatrix} = -(S^t S + \delta I)^{-1} T^t \alpha_1 \tag{7}$$

$$\begin{bmatrix} w_2 \\ b_2 \end{bmatrix} = (T^t T + \delta I)^{-1} S^t \alpha_2 \tag{8}$$

Each new data point $x \in R^n$ is assigned to a given class $'i'$ by using the following formula depending on which plane is closest to that data point.

$$\text{class } i = \min|K(x^t, D^t)w_i + b_i| \quad \text{for } i = 1, 2. \tag{9}$$

2.2 Fuzzy twin support vector machine (FTWSVM)

In the case of FTWSVM, a weighting parameter is used based on fuzzy membership values. For comparison, we choose the fuzzy membership for each data points based on its distance from the centroid [17]. The membership values are used for giving weights to the error tolerance, i.e. C for every data point in FTWSVM.

The fuzzy membership function is given as

$$\text{mem} = 1 - \frac{d_{\text{cen}}}{\max(d_{\text{cen}}) + \delta}$$

where d_{cen} is the Euclidean distance of each data point from the centroid of its class and δ is a small positive integer for making the denominator non-zero.The formulation of FTWSVM in primal is written as

$$\min \frac{1}{2}||K(X_1, D^t)w_1 + e_1 b_1||^2 + C_1 s_2^t \xi$$

subject to

$$\begin{aligned} &-(K(X_2, D^t)w_1 + e_2 b_1) + \xi \geq e_2, \;\; \xi \geq 0 \\ &\min \frac{1}{2}||K(X_2, D^t)w_2 + e_2 b_2||^2 + C_2 s_1^t \eta \end{aligned} \tag{10}$$

subject to

$$(K(X_1, D^t)w_2 + e_1 b_2) + \eta \geq e_1, \;\; \eta \geq 0 \tag{11}$$

where ξ, η represent slack variables; C_1, C_2 are penalty parameters; $K(,)$ is the kernel function, s_1,s_2 are vectors having the membership values of the data samples in the constraints.

The Lagrangian of the problems (10) and (11) is written as

$$\begin{aligned} L_1 = &\frac{1}{2}||K(X_1, D^t)w_1 + e_1 b_1||^2 + C_1 s_2^t \xi \\ &+ \alpha_1^t((K(X_2, D^t)w_1 + e_2 b_1) - \xi + e_2) - \beta_1^t \xi \end{aligned} \tag{12}$$

$$L_2 = \frac{1}{2}||K(X_2, D^t)w_2 + e_2b_2||^2 + C_2s_1^t\eta \\ + \alpha_2^t((-K(X_1, D^t)w_2 - e_1b_2) - \eta + e_1) - \beta_2^t\eta \quad (13)$$

where $\alpha_1 = (\alpha_{11}, \ldots, \alpha_{1q})^t$, $\beta_1 = (\beta_{11}, \ldots, \beta_{1q})^t$, $\alpha_2 = (\alpha_{21}, \ldots, \alpha_{2p})^t$ and $\beta_2 = (\beta_{21}, \ldots, \beta_{2p})^t$ are the vectors of Lagrange multipliers. Now, we apply the Karush–Kuhn–Tucker (K.K.T) necessary and sufficient conditions [28] to find the Wolfe dual of Eqs. (12) and (13) as

$$\min \frac{1}{2}\alpha_1^t T(S^tS)^{-1}T^t\alpha_1 - e_2^t\alpha_1$$

subject to

$$0 \le \alpha_1 \le s_2C_1 \\ \min \frac{1}{2}\alpha_2^t S(T^tT)^{-1}S^t\alpha_2 - e_1^t\alpha_2 \quad (14)$$

subject to

$$0 \le \alpha_2 \le s_1C_2 \quad (15)$$

where $S = [K(X_1, D^t)\ e_1]$ and $T = [K(X_2, D^t)\ e_2]$.

We compute the nonlinear hyperplanes $K(x^t, D^t)w_1 + b_1 = 0$ and $K(x^t, D^t)w_2 + b_2 = 0$ by computing the values of w_1,w_2, b_1 and b_2 by using Eq. (16) as

$$\begin{bmatrix} w_1 \\ b_1 \end{bmatrix} = -(S^tS + \delta I)^{-1}T^t\alpha_1 \quad \text{and} \\ \begin{bmatrix} w_2 \\ b_2 \end{bmatrix} = (T^tT + \delta I)^{-1}S^t\alpha_2 \quad (16)$$

Similarly, the resultant classifier is obtained by using Eq. (9).

3 Proposed Entropy-based Fuzzy Twin Support Vector Machine for class imbalance learning (EFTWSVM-CIL)

Recently, Fan et al. [25] proposed a novel fuzzy membership evaluation to improve the effectiveness and generalization ability of fuzzy support vector machine where memberships of the samples are computed based on class certainty. In information theory, entropy is a measure of the information carried by a sample. Chen et al. [29] used information entropy to find the uncertainty measure of a neighbourhood system. In case of class imbalance problem, most of the noisy data points of the majority class lie at the boundary of the two classes. So, for the majority class, the information of every data point is calculated based on its probability of belonging to any of the classes. This information is higher for the noisy samples as compared to rest of the samples in that class. The probability of a sample belonging to a particular class is based on class certainty. To find the class certainty, we can use entropy which is one of the effective-measuring approaches. Hence, one can assign the fuzzy membership to the data points by using the information entropy as the weighted parameter. Thus, the noisy samples of the majority class get lesser weights as compared to the other samples of the class. The traditional approach of giving weights [16] does not take into account the noise at the boundary of the two classes and do not incorporate the information about the probability distribution. Moreover, in most of the weighting strategies used for class imbalance problems, measures like distance from the centroid are used which do not give any information about the data points at the boundary of the two classes. In the proposed approach, to enhance the participation of the minority class in the decision classifier, the samples of majority class with lower entropy get larger fuzzy membership values. The entropy of any sample x_i is calculated as:

$$E_i = -P_{\text{pos}\_x_i} * \ln(p_{\text{pos}\_x_i}) - p_{\text{neg}\_x_i} * \ln(p_{\text{neg}\_x_i})$$

where $P_{\text{pos}\_x_i}$ and $P_{\text{neg}\_x_i}$ are the probability of minority class and majority class of sample x_i, respectively. Further, we calculate the K nearest neighbours of sample x_i and assign the values to $P_{\text{pos}\_x_i}$ and $P_{\text{neg}\_x_i}$ based on count of total minority and majority class neighbours.

Further, the data points of the majority class are divided into n subsets based on increasing order of entropy. The fuzzy membership of samples in each subset are calculated as

$$F_{\text{q}} = 1.0 - \beta * (q - 1),\ q = 1, 2, \ldots, n$$

where F_{q} is the fuzzy membership for samples distributed in qth subset with fuzzy membership parameter $\beta \in \left(0, \frac{1}{n-1}\right]$ which controls the scale of the fuzzy values of samples. The fuzzy membership function is written as

$$s_i = \begin{cases} 1 - \beta * (q - 1), & \text{if} \quad y_i = -1 \,\&\, x_i \in q\text{th subset} \\ 1, & \text{if} \quad y_i = 1 \end{cases}$$

Fan et al. [25] considered this approach to find the fuzzy membership of the sample and proposed a new approach termed as entropy-based fuzzy support vector machine for imbalance datasets. Motivated by the work of Fan et al. [25] and Jayadeva et al. [15], in this paper, we propose a new fuzzy twin support vector machine based on information entropy for class imbalance learning where information entropy is used for the fuzzy membership. The data points which have highest entropy are those present on the boundary between the classes. So, the data points of the majority class get their membership value based on their entropy and all the minority class samples get full membership value equal to 1. EFTWSVM-CIL finds two nonparallel hyperplanes such that each one is closer to the two classes and as far as possible from the other, whereas EFSVM finds separating hyperplanes that maximizes the margin between two classes. Due to this approach, the proposed EFTWSVM-CIL gives better generalization performance in comparison with EFSVM. Further, one can notice that we consider a pair of QPP of smaller size to find the decision surface of our proposed EFTWSVM-CIL, instead of solving a single large QPP as in the case of EFSVM. This makes our proposed EFTWSVM-CIL faster than EFSVM in terms of training time. Thus, it is very well suited for training on large imbalanced data. Now, we discuss the linear and nonlinear formulations of our EFTWSVM-CIL.

3.1 Linear EFTWSVM-CIL

In linear case, the EFTWSVM-CIL finds the resultant classifier by solving the following pair of QPPs

$$\min \frac{1}{2}||X_1 w_1 + e_1 b_1||^2 + C_1 s_2^t \xi$$

subject to

$$\begin{aligned} &-(X_2 w_1 + e_2 b_1) + \xi \geq e_2,\ \xi \geq 0 \\ &\min \frac{1}{2}||X_2 w_2 + e_2 b_2||^2 + C_2 s_1^t \eta \end{aligned} \tag{17}$$

subject to

$$(X_1 w_2 + e_1 b_2) + \eta \geq e_1,\ \eta \geq 0 \tag{18}$$

where ξ, η represent slack variables, $C_1, C_2 > 0$ are penalty parameters and s_1,s_2 are vectors containing the entropy-based fuzzy membership values of minority as well as majority class, respectively. The Lagrangian of problems (17) and (18) in primal is written as

$$\begin{aligned} L_1 = &\frac{1}{2}||X_1 w_1 + e_1 b_1||^2 + C_1 s_2^t \xi + \alpha_1^t((X_2 w_1 + e_2 b_1) - \xi \\ &+ e_2) - \beta_1^t \xi \end{aligned} \tag{19}$$

$$\begin{aligned} L_2 = &\frac{1}{2}||X_2 w_2 + e_2 b_2||^2 + C_2 s_1^t \eta + \alpha_2^t((-X_1 w_2 - e_1 b_2) - \eta \\ &+ e_1) - \beta_2^t \eta \end{aligned} \tag{20}$$

where $\alpha_1 = (\alpha_{11}, \ldots, \alpha_{1q})^t$, $\beta_1 = (\beta_{11}, \ldots, \beta_{1q})^t$, $\alpha_2 = (\alpha_{21}, \ldots, \alpha_{2p})^t$ and $\beta_2 = (\beta_{21}, \ldots, \beta_{2p})^t$ are the vectors of Lagrange multipliers. Applying the KKT conditions to (19), we get

$$\frac{\partial L}{\partial w_1} = 0 \Rightarrow X_1^t(X_1 w_1 + e_1 b_1) + X_2^t \alpha_1 = 0 \tag{21}$$

$$\begin{aligned} &\frac{\partial L}{\partial b_1} = 0 \Rightarrow e_1^t(X_1 w_1 + e_1 b_1) + e_2^t \alpha_1 = 0 \\ &\frac{\partial L}{\partial \xi} = 0 \Rightarrow C_1 s_2 - \beta_1 - \alpha_1 = 0 \\ &-(X_2 w_1 + e_2 b_1) + \xi \geq e_2,\ \xi \geq 0 \\ &\alpha_1^t(-(X_2 w_1 + e_2 b_1) + \xi - e_2) = 0 \\ &\beta_1^t \xi = 0,\ \alpha_1 \geq 0,\ \beta_1 \geq 0 \end{aligned} \tag{22}$$

Combining (21) and (22), we get

$$\begin{bmatrix} X_1^t \\ e_1^t \end{bmatrix} [X_1 \quad e_1] \begin{bmatrix} w_1 \\ b_1 \end{bmatrix} + \begin{bmatrix} X_2^t \\ e_2^t \end{bmatrix} \alpha = 0 \tag{23}$$

One can rewrite (23) as

$$u_1 = -(A^t A)^{-1} B^t \alpha_1$$

where $A = [X_1 \quad e_1]$, $B = [X_2 \quad e_2]$ and the augmented vector $u_1 = \begin{bmatrix} w_1 \\ b_1 \end{bmatrix}$.

Here, we introduce the regularization term δI where $\delta > 0$ and I is the identity matrix of appropriate size to handle the ill-conditioning of $\mathrm{A}^{\mathrm{t}}\mathrm{A}$ in finding the inverse. Thus, we get,

$$u_1 = -(A^t A + \delta I)^{-1} B^t \alpha_1 \tag{24}$$

Using the above KKT conditions and (19), the dual of the optimization problem in (17) can be written in the form of following QPP

$$\min \frac{1}{2} \alpha_1^t B (A^t A)^{-1} B^t \alpha_1 - e_2^t \alpha_1$$

subject to

$$0 \leq \alpha_1 \leq s_2 C_1 \tag{25}$$

In similar manner, one can find the dual of (18) as

$$\min \frac{1}{2} \alpha_2^t A (B^t B)^{-1} A^t \alpha_2 - e_1^t \alpha_2$$

subject to

$$0 \leq \alpha_2 \leq s_1 C_2 \tag{26}$$

The values of w_2 and b_2 are calculated as

$$u_2 = (B^t B + \delta I)^{-1} A^t \alpha_2 \tag{27}$$

where $u_2 = \begin{bmatrix} w_2 \\ b_2 \end{bmatrix}$.

After calculating the value of u_1 and u_2, we find the non-parallel hyperplanes $f_1(x) = w_1^t x + b_1$ and $f_2(x) = w_2^t x + b_2$. Every new data point $x \in R^n$ is assigned to a given class $'i'$ by using the following formula depending on the distance from the two planes.

$$\text{class } i = \min|x^t w_i + b_i| \text{ for } i = 1, 2. \tag{28}$$

3.2 Nonlinear EFTWSVM-CIL

For classifying nonlinear separable data points, we used kernel function to transform the data points in the higher-dimensional feature space [30]. The nonlinear TWSVM is formulated in the primal form as

$$\min \frac{1}{2}||K(X_1, D^t)w_1 + e_1 b_1||^2 + C_1 s_2^t \xi$$

subject to

$$\begin{aligned} &-(K(X_2, D^t)w_1 + e_2 b_1) + \xi \geq e_2,\ \xi \geq 0 \\ &\min \frac{1}{2}||K(X_2, D^t)w_2 + e_2 b_2||^2 + C_2 s_1^t \eta \end{aligned} \tag{29}$$

subject to

$$(K(X_1, D^t)w_2 + e_1 b_2) + \eta \geq e_1, \eta \geq 0 \tag{30}$$

where ξ,η represent slack variables, C_1, C_2 are penalty parameters, $D = [X_1; X_2]$, and s_1,s_2 are vectors containing the entropy-based fuzzy membership values. The Lagrangian function of the problems (29) and (30) is written as

$$\begin{aligned} L_1 = \frac{1}{2}||K(X_1, D^t)w_1 + e_1 b_1||^2 + C_1 s_2^t \xi + \alpha_1^t((K(X_2, D^t)w_1 \\ + e_2 b_1) - \xi + e_2) - \beta_1^t \xi \end{aligned} \tag{31}$$

$$\begin{aligned} L_2 = \frac{1}{2}||K(X_2, D^t)w_2 + e_2 b_2||^2 + C_2 s_1^t \eta \\ + \alpha_2^t((-K(X_1, D^t)w_2 - e_1 b_2) - \eta + e_1) - \beta_2^t \eta \end{aligned} \tag{32}$$

where $\alpha_1 = (\alpha_{11}, \ldots, \alpha_{1q})^t$, $\beta_1 = (\beta_{11}, \ldots, \beta_{1q})^t$, $\alpha_2 = (\alpha_{21}, \ldots, \alpha_{2p})^t$ and $\beta_2 = (\beta_{21}, \ldots, \beta_{2p})^t$ are the vectors containing the Lagrange multipliers.

Following the same procedure as in the linear case, we compute the nonlinear hyperplanes $K(x^t, D^t)w_1 + b_1 = 0$ and $K(x^t, D^t)w_2 + b_2 = 0$ by computing the value of w_1,w_2, b_1 and b_2 using Eqs. (33) and (34)

$$u_1 = \begin{bmatrix} w_1 \\ b_1 \end{bmatrix} = -(P^t P + \delta I)^{-1} Q^t \alpha_1 \tag{33}$$

$$u_2 = \begin{bmatrix} w_2 \\ b_2 \end{bmatrix} = (Q^t Q + \delta I)^{-1} P^t \alpha_2 \tag{34}$$

where $P = [K(X_1, D^t)\ e_1]$,$Q = [K(X_2, D^t)\ e_2]$.

For each new data point $x \in R^n$, it is assigned to a given class $'i'$ by using the following formula depending on which of the planes is closest to that point.

$$\text{class } i = \min|K(x^t, D^t)w_i + b_i| \text{ for } i = 1,\ 2. \tag{35}$$

4 Numerical Experiments

In this section, to check the effectiveness of the proposed EFTWSVM-CIL with TWSVM, FTWSVM and EFSVM, we performed experiments on several imbalanced datasets from KEEL imbalanced datasets [31] and UCI repository [32] for binary classification. All computations were carried out on a PC running on Windows 7 OS with 64 bit, 3.20 GHz Intel® core™ i5-2400 processor having 2 GB of RAM under MATLAB R2008b environment. We used MOSEK optimization toolbox to solve the SVM formulations which is taken from http://www.mosek.com. For selecting the optimum parameters, we used fivefold cross-validation technique. To construct nonlinear classifier, we have used Gaussian kernel $k(a, b) = \exp(-\sigma||a - b||^2)$ where vector $a, b \in R^m$.

We have taken the value of the parameter $C = C_1 = C_2$ from the set $\{2^{-5}, \ldots, 2^5\}$ for all the cases. For FTWSVM, δ is taken as 0.5. For EFTWSVM and EFSVM the value of K for k-NN is chosen from {5, 10} and β is taken as 0.05. The value of σ is calculated as per the following formula [33] in all methods,

$$\sigma = \frac{1}{N^2} \sum_{i,j=1}^{N} ||x_i - x_j||^2$$

All the results for TWSVM, FTWSVM, EFSVM and proposed method EFTWSVM-CIL are shown in terms of prediction accuracy, i.e. the area under the ROC curve (AUC) [34] and training time for both linear and nonlinear cases in Tables 1 and 3. One can observe from Tables 1 and 3 that EFTWSVM-CIL is much superior to TWSVM, FTWSVM, and EFSVM in terms of better generalization performance. Our proposed EFTWSVM-CIL takes very less training time in comparison with EFSVM because EFTWSVM-CIL solves a pair of smaller-size QPPs instead of solving a large one as in the case of EFSVM.

It is observable from Table 1 that our proposed method EFTWSVM-CIL has not performed better in case of all the datasets for linear kernel. Further, we analyse the comparative performance of EFTWSVM-CIL with TWSVM, FTWSVM, and EFSVM based on the average ranks of all the methods which are presented in Table 2 for the linear case. One can clearly observe form Table 2 that the average rank of proposed EFTWSVM-CIL is lowest among all the methods. We perform the Friedman test with the corresponding post hoc test [35] in the case of linear kernel for statistical comparison on the performance of the 4 algorithms using 24 datasets. We assume all the methods are equivalent under null hypothesis, and the Friedman statistic is computed from Table 2 as

Table 1 Performance comparison of EFTWSVM-CIL with TWSVM, FTWSVM and EFSVM using linear kernel for classification on imbalance datasets

Dataset (train size, test size)	Imbalance ratio	TWSVM (C) time	FTWSVM (C) time	EFSVM (C) time	EFTWSVM-CIL (C) time
Vehicle2 (400 × 19, 446 × 19)	2.88	97.086 (2^1) 0.04562	94.9503 (2^4) 0.11147	92.6208 (2^5) 3.589	**97.2317** (2^2) 0.66708
Pima (300 × 9, 468 × 9)	1.87	74.9586 (2^− 2) 0.01878	74.94 (2^− 1) 0.02499	66.4364 (2^2) 1.98326	**76.2822** (2^− 2) 0.36165
Ripley (600 × 3, 650 × 3)	1	89.0905 (2^3) 0.09075	89.5369 (2^4) 0.09528	84.5788 (2^5) 7.97929	**89.644** (2^3) 1.4067
Ecoli-0-2-3-4_vs_5 (100 × 8, 102 × 8)	9.1	96.8421 (2^− 1) 0.00926	**97.3684** (2^− 1) 0.01176	91.2782 (2^4) 0.22948	**97.3684** (2^− 1) 0.05114
Ecoli-0-4-6_vs_5 (100 × 7, 103 × 7)	9.15	88.37 (2^− 4) 0.00928	**91.1172** (2^0) 0.01212	83.3333 (2^3) 0.22982	**91.1172** (2^− 2) 0.10835
Led7digit-0-2-4-5-6-7-8-9_vs_1 (220 × 8, 223 × 8)	10.97	88.0081 (2^− 1) 0.02177	87.7642 (2^0) 0.02808	**89.9593** (2^− 1) 1.07733	**88.9837** (2^0) 0.21258
Yeast-0-5-6-7-9_vs_4 (250 × 9, 278 × 9)	9.35	75.5413 (2^− 2) 0.03031	77.1161 (2^− 1) 0.03492	68.3563 (2^4) 1.37908	77.313 (2^− 1) 0.26397
Yeast-2_vs_4 (250 × 9, 264 × 9)	9.08	85.4895 (2^− 2) 0.02832	**87.2803** (2^− 2) 0.03222	79.7908 (2^5) 1.38427	**85.4895** (2^− 2) 0.27136
Ecoli-0-1-4-6_vs_5 (150 × 7, 130 × 7)	13	**98.3871** (2^− 4) 0.01378	**98.3871** (2^− 3) 0.01608	82.9301 (2^4) 0.50532	**98.3871** (2^− 3) 0.10047
Transfusion (350 × 5, 398 × 5)	3.2	50 (2^0) 0.02644	50 (2^0) 0.03757	50 (2^− 5) 2.71225	**51.2761** (2^0) 0.48061
*Ecoli*2 (150 × 8, 186 × 8)	5.46	**87.7574** (2^0) 0.02669	87.1691 (2^0) 0.01462	73.5294 (2^1) 0.50558	**87.4632** (2^0) 0.10691
Vowel (500 × 11, 488 × 11)	9.98	**90.2744** (2^2) 0.13623	89.138 (2^2) 0.15027	81.7056 (2^5) 5.59896	**90.2744** (2^1) 1.06284
Wisconsin (300 × 10, 383 × 10)	1.86	96.1125 (2^− 3) 0.01996	97.1226 (2^1) 0.02818	**98.0634** (2^− 5) 2.01953	**96.6176** (2^− 3) 0.36869
Vehicle 1 (400 × 19, 446 × 19)	2.9	79.9439 (2^− 4) 0.04893	**81.2462** (2^− 4) 0.05378	64.1104 (2^5) 3.57224	80.6877 (2^− 4) 0.65652
Shuttle-c0-vs-c4 (900 × 10, 929 × 10)	13.87	**100** (2^− 5) 0.91874	**100** (2^− 5) 0.91797	99.1803 (2^− 3) 18.3282	**100** (2^− 5) 3.80893
Ecoli-0-1_vs_2-3-5 (120 × 8, 124 × 8)	9.17	**85.2679** (2^− 2) 0.05941	**85.2679** (2^− 1) 0.08346	66.6667 (2^1) 0.39566	**85.2679** (2^− 1) 0.13547
New-thyroid1 (100 × 6, 115 × 6)	5.14	97.0588 (2^− 5) 0.00818	**98.0392** (2^− 4) 0.01866	95.6637 (2^5) 0.23029	**98.0392** (2^− 4) 0.06304
*Ecoli*0137vs26 (180 × 8, 131 × 8)	39.14	**93.1818** (2^− 3) 0.01421	90.9091 (2^− 3) 0.02264	84.0909 (2^5) 0.71967	**93.1818** (2^− 3) 0.13426
*Yeast*5 (500 × 9, 984 × 9)	32.73	94.5178 (2^− 2) 0.19316	**97.0126** (2^− 2) 0.20238	60 (2^5) 5.57902	**97.0126** (2^− 2) 1.10171
Cleve (150 × 14, 147 × 14)	1.18	82.4026 (2^− 3) 0.00975	82.4026 (2^− 1) 0.0126	78.961 (2^3) 0.51493	**83.1818** (2^− 3) 0.09639
Wpbc (100 × 34, 94 × 34)	3.22	64.8649 (2^− 3) 0.00846	62.8378 (2^− 2) 0.01052	**70.2703** (2^4) 0.23212	**66.4865** (2^− 3) 0.0498
Votes (200 × 17, 235 × 17)	1.59	**95.6311** (2^− 1) 0.01406	**95.6311** (2^0) 0.01797	**95.6311** (2^− 2) 0.90898	**95.6311** (2^− 1) 0.17495
Ecoli-0-1_vs_5 (120 × 7, 120 × 7)	11	**93.3501** (2^− 2) 0.01205	**93.3501** (2^-1) 0.01828	84.6154 (2^3) 0.32806	**93.3501** (2^− 2) 0.06821
Shuttle-6_vs_2-3 (100 × 10, 130 × 10)	22	**100** (2^− 5) 0.00988	**100** (2^− 5) 0.01197	**100** (2^2) 0.22768	**100** (2^− 5) 0.05254

Bold values indicate the best result

Table 2 Average ranks of TWSVM, FTWSVM, EFSVM and EFTWSVM-CIL for imbalance datasets using linear kernel for classification on imbalance datasets

Dataset	Imbalance Ratio	TWSVM	FTWSVM	EFSVM	EFTWSVM-CIL
Vehicle2	2.88	2	3	4	1
Pima	1.87	2	3	4	1
Ripley	1	3	2	4	1
Ecoli-0-2-3-4_vs_5	9.1	3	1.5	4	1.5
Ecoli-0-4-6_vs_5	9.15	3	1.5	4	1.5
Led7digit-0-2-4-5-6-7-8-9_vs_1	10.97	3	4	1	2
Yeast-0-5-6-7-9_vs_4	9.35	3	2	4	1
Yeast-2_vs_4	9.08	2.5	1	4	2.5
Ecoli-0-1-4-6_vs_5	13	2	2	4	2
Transfusion	3.2	3	3	3	1
Ecoli 2	5.46	1	3	4	2
Vowel	9.98	1.5	3	4	1.5
Wisconsin	1.86	4	2	1	3
Vehicle1	2.9	3	1	4	2
Shuttle-c0-vs-c4	13.87	2	2	4	2
Ecoli-0-1_vs_2-3-5	9.17	2	2	4	2
New-thyroid1	5.14	3	1.5	4	1.5
0137vs26	39.14	1.5	3	4	1.5
Yeast 5	32.73	3	1.5	4	1.5
Cleve	1.18	2.5	2.5	4	1
Wpbc	3.22	2	4	1	3
Votes	1.59	2.5	2.5	2.5	2.5
Ecoli-0-1_vs_5	11	2	2	4	2
Shuttle-6_vs_2-3	22	2.5	2.5	2.5	2.5
Average ranks		2.4583	2.3125	3.4583	**1.7708**

Bold value indicates the best result

$$\chi_F^2 = \frac{12 \times 24}{4 \times (4+1)} \Big[(2.4583^2 + 2.3125^2 + 3.4583^2 + 1.7708^2) - \frac{4 \times (4+1)^2}{4} \Big] \cong 21.4051$$

$$F_F = \frac{(24-1) \times 21.4051}{24 \times (4-1) - 21.4051} \cong 9.7306$$

where F_F is distribution according to the F-distribution with $(3, 3 \times 23) = (3, 69)$ being degree of freedom with 4 methods and 24 datasets. The critical value of $F(3, 69)$ is 2.7375 for the level of significance at $\alpha = 0.05$. Since the value of $F_F = 9.7306 > 2.7375$, we reject the null hypothesis. Further, Nemenyi post hoc test is performed for pair-wise comparison of methods and the significant difference between them is checked by computing the critical difference (CD) at $P = 0.10$ which should differ by at least $2.291\sqrt{\frac{4\times(4+1)}{6\times 24}} \approx 0.8539$.

Since the difference between the averages ranks of EFSVM with EFTWSVM-CIL $(3.4583 - 1.7708 = 1.6875)$ is greater than 0.8538, we conclude that EFTWSVM-CIL is significantly better than EFSVM. Since the differences in the average rank of TWSVM and FTWSVM with EFTWSVM-CIL are $(2.4583 - 1.7708 = 0.6875)$ and $(2.3125 - 1.7708 = 0.5417)$, respectively, which are less than 0.8539, this shows that there is no significant difference between EFTWSVM-CIL with TWSVM and FTWSVM.

For the Gaussian kernel, the accuracy values are shown with the training time for the proposed EFTWSVM-CIL with TWSVM, FTWSVM and EFSVM in Table 3. One can observe from Table 3 that EFTWSVM shows the better or equal generalization performance in 18 cases. The training speed of our proposed EFTWSVM-CIL is better than EFSVM and comparable to TWSVM and FTWSVM. The average ranks of all the methods based on accuracy values are shown in Table 2. One can conclude that among all the methods our proposed EFTWSVM-CIL has the lowest average rank. It is noticeable from the table that the proposed EFTWSVM is not always better in terms of accuracy for all the datasets, so further Friedman statistical test is performed with the post hoc tests.

Now, the Friedman statistic is computed for nonlinear kernel under null hypothesis by using Table 4:

Table 3 Performance comparison of EFTWSVM-CIL with TWSVM, FTWSVM and EFSVM using Gaussian kernel for classification on imbalance datasets

Dataset (train size, test size)	Imbalance ratio	TWSVM (C, σ) time	FTWSVM (C, σ) time	EFSVM (C, σ) time	EFTWSVM-CIL (C, σ) time
WPBC (100 × 34, 94 × 34)	3.22	65.3378 (2^− 3, 1.41946) 0.0263	65.3378 (2^− 2, 1.41946) 0.02894	62.9054 (2^3, 1.41946) 0.27326	**67.1622** (2^− 2, 1.41946) 0.06876
Votes (200v17, 235 × 17)	1.59	**96.6728** (2^− 2, 2.65705) 0.08304	95.9783 (2^− 3, 2.65705) 0.08706	95.6311 (2^1, 2.65705) 1.04555	**96.6728** (2^− 2, 2.65705) 0.24102
Australian Credit (300 × 15, 390 × 15)	1.25	86.3194 (2^1, 1.56989) 0.17582	87.0563 (2^1, 1.56989) 0.18497	86.1177 (2^− 4, 1.56989) 2.35807	**87.2902** (2^− 3, 1.56989) 0.52862
Transfusion (350 × 5, 398 × 5)	3.2	65.2657 (2^− 4, 0.43636) 0.23416	66.1324 (2^− 2, 0.43636) 0.24351	64.3031 (2^5, 0.43636) 3.16117	**66.189** (2^− 4, 0.43636) 0.70044
Ecoli-0-2-3-4_vs_5 (100 × 8, 102 × 8)	9.1	98.4211 (2^− 3, 0.72706) 0.02564	97.8947 (2^− 5, 0.72706) 0.02821	**98.9474** (2^5, 0.72706) 0.26689	98.4211 (2^− 2, 0.72706) 0.07516
Ionosphere (200 × 34, 151 × 34)	0.56	90.1355 (2^− 3, 2.17532) 0.08439	**90.702** (2^− 3, 2.17532) 0.09039	80.5665 (2^4, 2.17532) 1.06399	**90.702** (2^− 4, 2.17532) 0.24762
Ecoli-0-4-6_vs_5 (100 × 7, 103 × 7)	9.15	**87.5** (2^− 2, 0.769994) 0.02586	**87.5** (2^− 1, 0.769994) 0.02828	86.9505 (2^2, 0.769994) 0.27972	**87.5** (2^− 2, 0.769994) 0.0689
CMC (700 × 10, 773 × 10)	0.75	**64.7811** (2^− 5, 1.32239) 1.02449	64.2043 (2^− 5, 1.32239) 1.07252	63.1802 (2^5, 1.32239) 12.8557	64.3685 (2^− 4, 1.32239) 2.95445
Ecoli-0-1_vs_2-3-5 (120 × 8, 124 × 8)	9.17	**79.1667** (2^− 4, 0.742547) 0.03562	78.7202 (2^− 3, 0.742547) 0.03845	78.2738 (2^3, 0.742547) 0.39122	78.7202 (2^− 3, 0.742547) 0.09247
Pima Indians (300 × 9, 468 × 9)	1.87	72.0924 (2^2, 0.64933) 0.17657	72.5019 (2^− 2, 0.64933) 0.18144	**75.8603** (2^2, 0.64933) 2.32454	75.1489 (2^− 4, 0.64933) 0.5208
Ecoli 0137vs26 (180 × 8, 131 × 8)	4.76	97.2686 (2^− 1, 0.658638) 0.0689	94.9958 (2^1, 0.658638) 0.0709	97.2686 (2^5, 0.658638) 0.845	**97.7273** (2^− 1, 0.658638) 0.19469
Ecoli 3 (150 × 8, 186 × 8)	8.6	**90.5147** (2^− 2, 0.663699) 0.0508	88.4559 (2^− 4, 0.663699) 0.05444	89.0441 (2^5, 0.663699) 0.5916	89.3382 (2^− 1, 0.663699) 0.13885
Heart-statlog (130 × 14, 140 × 14)	0.8	83.3887 (2^− 2, 1.72389) 0.03757	81.9361 (2^− 1, 1.72389) 0.04085	81.822 (2^− 1, 1.72389) 0.44879	**84.4677** (2^− 3, 1.72389) 0.10503
Yeast-0-2-5-6_vs_3-7-8-9 (500 × 9, 504 × 9)	9.14	70.5123 (2^− 2, 0.494682) 0.57998	**73.9021** (2^− 1, 0.494682) 0.58797	71.3597 (2^2, 0.494682) 6.51619	**73.9021** (2^− 1, 0.494682) 1.52481
Yeast 5 (500 × 9, 984 × 9)	32.73	69.8428 (2^− 4, 0.466753) 0.61213	71.5094 (2^− 3, 0.466753) 0.63864	60 (2^1, 0.466753) 6.51586	**71.5618** (2^− 3, 0.466753) 1.57086
Ecoli-0-6-7_vs_3-5 (110 × 8, 112 × 8)	10	85.7143 (2^− 4, 0.730104) 0.03051	85.7143 (2^− 3, 0.730104) 0.03268	**88.7755** (2^4, 0.730104) 0.2736	85.7143 (2^− 3, 0.730104) 0.08083
Yeast-0-5-6-7-9_vs_4 (250 × 9, 278 × 9)	9.35	70.1608 (2^− 4, 0.573935) 0.1347	69.7671 (2^− 4, 0.573935) 0.1379	**73.0315** (2^3, 0.573935) 1.37647	72.8346 (2^− 2, 0.573935) 0.12798
Yeast-0-3-5-9_vs_7-8 (250 × 9, 256 × 9)	9.12	60.5605 (2^− 5, 0.583643) 0.13121	59.8968 (2^− 5, 0.583643) 0.13718	61.0029 (2^− 1, 0.583643) 1.63235	**61.3422** (2^− 4, 0.583643) 0.37476
Glass4 (150 × 10, 64 × 10)	15.46	**79.1525** (2^1, 0.735447) 0.05211	**79.1525** (2^1, 0.735447) 0.05524	**79.1525** (2^5, 0.735447) 0.58844	**79.1525** (2^1, 0.735447) 0.13987
Vehicle2 (400 × 19, 446 × 19)	2.88	96.7463 (2^4, 1.17148) 0.33379	**97.6209** (2^0, 1.17148) 0.34313	97.3775 (2^4, 1.17148) 4.18954	97.1836 (2^4, 1.17148) 0.95295
Ecoli-0-1-4-7_vs_5-6 (150 × 7, 182 × 7)	12.28	**91.0784** (2^− 5, 0.763992) 0.05167	**91.0784** (2^− 4, 0.763992) 0.05396	**91.0784** (2^3, 0.763992) 0.58755	**91.0784** (2^− 5, 0.763992) 0.13791
Ecoli 4 (150 × 8, 186 × 8)	15.8	**91.092** (2^− 2, 0.662858) 0.05138	**91.092** (2^− 2, 0.662858) 0.05429	90.5172 (2^5, 0.662858) 0.59294	**91.092** (2^− 1, 0.662858) 0.14004
Ecoli-0-1_vs_5 (120 × 7, 120 × 7)	11	**88.4615** (2^− 5, 0.697633) 0.03527	**88.4615** (2^− 4, 0.697633) 0.03871	84.6154 (2^1, 0.697633) 0.38122	**88.4615** (2^− 5, 0.697633) 0.09448
Yeast-2_vs_4 (250 × 9, 264 × 9)	9.08	82.954 (2^− 1, 0.472432) 0.13305	82.7448 (2^− 1, 0.472432) 0.13726	82.954 (2^3, 0.472432) 1.37384	**84.7448** (2^− 1, 0.472432) 0.12722
Ecoli-0-6-7_vs_5 (110 × 7, 110 × 7)	10	**88.3938** (2^− 5, 0.691324) 0.03014	**88.3938** (2^− 5, 0.691324) 0.03254	**88.3938** (2^4, 0.691324) 0.3205	**88.3938** (2^− 5, 0.691324) 0.0778
Cleveland (150 × 14, 147 × 14)	1.17	79.0909 (2^− 3, 1.75769) 0.04825	78.4416 (2^− 3, 1.75769) 0.05159	**80.6494** (2^0, 1.75769) 0.59991	80.5195 (2^− 3, 1.75769) 0.13771

Table 3 (continued)

Dataset (train size, test size)	Imbalance ratio	TWSVM (C, σ) time	FTWSVM (C, σ) time	EFSVM (C, σ) time	EFTWSVM-CIL (C, σ) time
Monk2 (300 × 8, 301 × 8)	1.92	77.0989 (2^− 5, 1.54371) 0.176	**78.4111** (2^− 5, 1.54371) 0.18269	73.3664 (2^5, 1.54371) 2.32977	76.7295 (2^− 5, 1.54371) 0.524
Shuttle-c0-vs-c4 (900 × 10, 929 × 10)	13.87	**99.1803** (2^− 5, 0.474217) 2.18748	**99.1803** (2^− 5, 0.474217) 2.23814	**99.1803** (2^− 5, 0.474217) 21.6004	**99.1803** (2^− 5, 0.474217) 5.28983

Bold values indicate the best result

Table 4 Average ranks of TWSVM, FTWSVM, EFSVM and EFTWSVM-CIL for imbalance datasets using Gaussian kernel for classification of imbalance datasets

Dataset	Imbalance Ratio	TWSVM	FTWSVM	EFSVM	EFTWSVM-CIL
WBPC	3.22	2.5	2.5	4	1
Votes	1.59	1.5	3	4	1.5
Australian credit	1.25	3	2	4	1
Transfusion	3.2	3	2	4	1
Ecoli-0-2-3-4_vs_5	9.1	2.5	4	1	2.5
Ionosphere	0.56	3	1.5	4	1.5
Ecoli-0-4-6_vs_5	9.15	3	3	1	3
CMC	0.75	1	3	4	2
Ecoli-0-1_vs_2-3-5	9.17	1	2.5	4	2.5
Pima Indians	1.87	4	3	1	2
Ecoli 0137vs26	4.76	2.5	4	2.5	1
Ecoli 3	8.6	1	4	3	2
Heart-statlog	0.8	2	3	4	1
Yeast-0-2-5-6_vs_3-7-8-9	9.14	4	1.5	3	1.5
Yeast 5	32.73	3	2	4	1
Ecoli-0-6-7_vs_3-5	10	3	3	1	3
Yeast-0-5-6-7-9_vs_4	9.35	3	4	1	2
Yeast-0-3-5-9_vs_7-8	9.12	3	4	2	1
Glass4	15.46	2.5	2.5	2.5	2.5
Vehicle2	2.88	4	1	2	3
Ecoli-0-1-4-7_vs_5-6	12.28	2.5	2.5	2.5	2.5
Ecoli 4	15.8	2	2	4	2
Ecoli-0-1_vs_5	11	2	2	4	2
Yeast-2_vs_4	9.08	2.5	4	2.5	1
Ecoli-0-6-7_vs_5	10	2.5	2.5	2.5	2.5
Cleveland	1.17	3	4	1	2
Monk2	1.92	2	1	4	3
Shuttle-c0-vs-c4	13.87	2.5	2.5	2.5	2.5
Average rank		2.5536	2.7143	2.8214	**1.9107**

Bold value indicates the best result

$$\chi_F^2 = \frac{12 \times 28}{4 \times (4+1)} \left[(2.5536^2 + 2.7143^2 + 2.8214^2 + 1.9107^2) - \frac{4 \times (4+1)^2}{4} \right] \cong 8.3894$$

$$F_F = \frac{(28-1) \times 8.3894}{28 \times (4-1) - 8.3894} \cong 2.9958$$

The critical value of $F(3, 84)$ i.e. 2.7132 for the level of significant $\alpha = 0.05$ is less than the value of F_F. Thus, it rejects the null hypothesis. Further, the Nemenyi post hoc test is used to find the significant difference between the pair-wise comparisons. We computed the critical difference (CD) at $p = 0.10$ which should differ by at least $2.291\sqrt{\frac{4\times(4+1)}{6\times28}} \approx 0.7905$.

The difference between the average ranks of EFTWSVM-CIL with EFSVM and FTWSVM are $(2.8214 - 1.9107 = 0.9107)$ and $(2.7143 - 1.9107 = 0.8036)$, respectively, which are greater than 0.7905. Hence, proposed EFTWSVM-CIL is significantly better than EFSVM and FTWSVM.

One can verify that the performance of our proposed EFTWSVM-CIL is not sensitive to the values of its parameters C and K. After extensive simulations, it is found that EFTWSVM-CIL is not very sensitive to the user-specified parameter K. To illustrate this result, the performance of EFTWSVM-CIL with Gaussian RBF kernel on Australian Credit, WPBC, *Yeast*-0-3-5-9_vs_7-8 and *Yeast*-2_vs_4 datasets is shown in Fig. 1. From the figures, one can observe that better accuracy could be achieved for smaller values of C.

5 Conclusions and future work

In this paper, we proposed a new variant of SVM as EFTWSVM-CIL to solve class imbalance problem in binary class datasets where the fuzzy membership values are calculated based on entropy values of samples. Here, our proposed EFTWSVM-CIL solves the two smaller-size QPPs rather than a single large one as in case of EFSVM to find the decision surface. So, one can conclude from the results that EFTWSVM-CIL shows better generalization performance as compared to TWSVM, FTWSVM and EFSVM which clearly illustrates its efficacy and applicability. It has been found that EFTWSVM-CIL outperforms in terms of learning speed in comparison with EFSVM for both linear and nonlinear kernels. Here, the performance of EFTWSVM-CIL also depends on the optimal parameters.

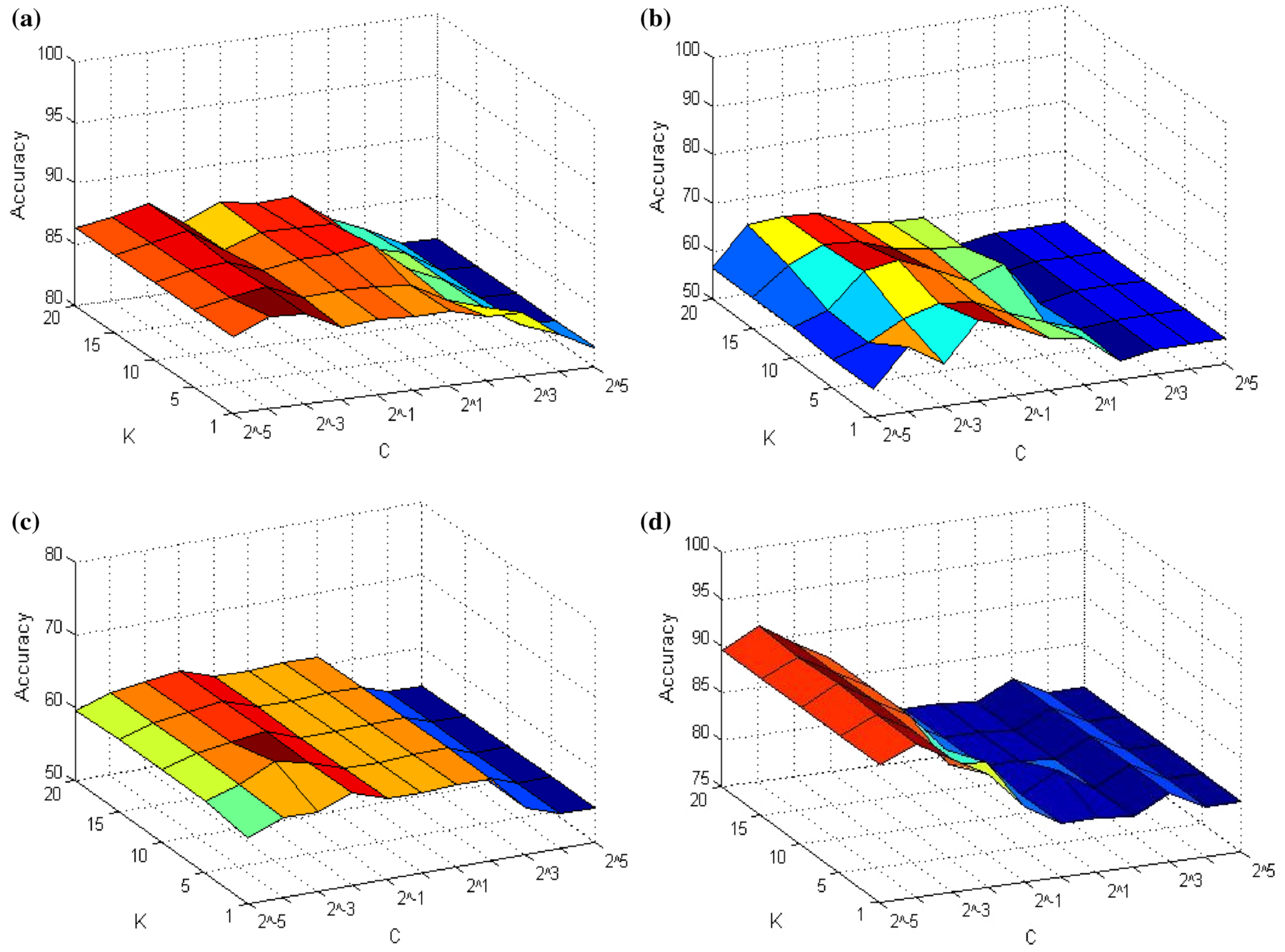

Fig. 1 Insensitivity performance of EFTWSVM-CIL for classification to the user-specified parameters (C, K) on imbalance datasets using Gaussian kernel. **a** Australian Credit, **b** WPBC, **c** *Yeast*-0-3-5-9_versus_7–8, **d** *Yeast*-2_vs_4

So, in future the proper selection of parameters for EFTWSVM-CIL may improve the performance of our proposed model. Some heuristic approaches can also be used to improve the method for parameter selection which may result into the better performance.

Compliance with ethical standards

Conflict of interest All authors declare that they have no conflict of interest.

References

1. Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20(3):273–297
2. Vapnik VN (1998) Statistical learning theory. Wiley, New York
3. Vapnik VN (2000) The nature of statistical learning theory, 2nd edn. Springer, New York
4. Osuna E, Freund, R, Girosi F (1997) Training support vector machines: an application to face detection. In: Computer vision and pattern recognition, 1997. Proceedings., IEEE computer society conference on (pp 130–136)
5. Phillips PJ (1998) Support Vector Machines Applied to Face recognition. Proc Conf Adv Neural Inf Process Syst 11:803–809
6. Michel P, El Kaliouby R (2003) Real time facial expression recognition in video using support vector machines. In: Proceedings of the 5th International Conference on Multimodal Interfaces, pp 258–264, ISBN: 1-58113-621-8
7. Borovikov E (2005) An evaluation of support vector machines as a pattern recognition tool. University of Maryland at College Park. http://www.umiacs.umd.edu/users/yab/SVMForPatternRecognition/report.pdf. Accessed 1 Dec 2016
8. Kumar MA, Gopal M (2009) Least squares twin support vector machines for pattern classification. Expert Syst Appl 36(4):7535–7543
9. Schmidt M, Gish H (1996) Speaker identification via support vector classifiers, acoustics, speech, and signal processing, 1996. ICASSP-96. In: Conference Proceedings, 1996 IEEE International Conference on, vol. 1. Atlanta, GA, pp 105–108
10. Khan L, Awad M, Thuraisingham B (2007) A new intrusion detection system using support vector machines and hierarchical clustering. VLDB J 16:507–521
11. Tomar D, Ojha D, Agarwal S (2014) An emotion detection system based on multi least squares twin support vector machine. Adv Artif Intell 2014:8
12. Tomar D, Agarwal S (2015) Hybrid feature selection based weighted least squares twin support vector machine approach for diagnosing breast cancer, hepatitis, and diabetes. Adv Artif Neural Syst 2015:1, Article ID 265637
13. Zhang J, Liu Y (2004) Cervical cancer detection using SVM-based feature screening. In: Proceedings of Seventh Int'l Conference Medical Image Computing and Computer Aided Intervention, pp 873–880
14. Balasundaram S, Gupta D, Prasad SC (2017) A new approach for training Lagrangian twin support vector machine via unconstrained convex minimization. Appl Intell 46:124–134
15. Jayadeva, Khemchandani R, Chandra S (2007) Twin support vector machines for pattern classification. IEEE Trans Pattern Anal Mach Intell (TPAMI) 29:905–910
16. Lin C-F, Wang S-D (2002) Fuzzy support vector machines. IEEE Trans Neural Netw 13(2):464–471
17. Batuwita R, Palade V (2010) FSVM-CIL: fuzzy support vector machines for class imbalance learning. IEEE Trans Fuzzy Syst 18(3):558–571
18. Tian D-Z, Peng G-B, Ha M-H (2012) Fuzzy support vector machine based on non-equilibrium data. In: International Conference on Machine Learning and Cybernetics, Xi'an, China, pp 15–17
19. Wang Y, Wang S, Lai KK (2005) A new fuzzy support vector machine to evaluate credit risk. IEEE Trans Fuzzy Syst 13(6):820–831
20. Chaudhuri, De K (2010) Fuzzy support vector machine for bankruptcy prediction. Appl Soft Comput 11(2):2472–2486
21. Shao YH, Chen WJ, Zhang JJ, Wang Z, Deng NY (2014) An efficient weighted Lagrangian twin support vector machine for imbalanced data classification. Pattern Recogn 47(9):3158–3167
22. Gupta D, Borah P, Prasad M (2017) A fuzzy based Lagrangian twin parametric-margin support vector machine (FLTPMSVM). In: Computational intelligence (SSCI), 2017 IEEE symposium series on pp 1–7 https://doi.org/10.1109/ssci.2017.8280964
23. Balasundaram S, Gupta D (2016) On optimization based extreme learning machine in primal for regression and classification by functional iterative method. Int J Mach Learn Cybern Springer 7(5):707–728
24. Balasundaram S, Gupta D, Kapil (2014) 1-norm extreme learning machine for regression and multiclass classification using Newton method. Neurocomputing, Elsevier 128:4–14
25. Fan Qi, Wang Zhe, Li Dongdong, Gao Daqi, Zha Hongyuan (2017) Entropy-based fuzzy support vector machine for imbalanced datasets. Knowl-Based Syst 115:87–99
26. Mangasarian OL, Wild EW (2006) Multisurface proximal support vector classification via generalized eigenvalues. IEEE Trans Pattern Anal Mach Intell 28(1):69–74
27. Fung G, Mangasarian OL (2001) Proximal support vector machine classifiers. In: Proceedings Internation Conference Knowl. Discov. Data Mining, pp 77–86
28. Mangasarian OL (1994) Nonlinear programming. SIAM Philadelphia, PA
29. Chen Y, Wu K, Chen X, Tang C, Zhu Q (2014) An entropy-based uncertainty measurement approach in neighborhood systems. Inf Sci 279:239–250
30. Burges CJC (1998) Geometry and invariance in kernel based methods. In: Cristopher JCB, Alexander JS (eds) Advances in kernel methods-support vector learning, Bernhard Scholkopf. MIT Press, Cambridge
31. Alcalá-Fdez J, Fernandez A, Luengo J, Derrac J, García S, Sánchez L, Herrera F (2011) KEEL data-mining software tool: data set repository, integration of algorithms and experimental analysis framework. J Multiple-Valued Logic Soft Comput 17(2–3):255–287
32. Murphy PM, Aha DW (1992) UCI repository of machine learning databases, University of California, Irvine. http://www.ics.uci.edu/~mlearn. Accessed 1 Dec 2016
33. Tsang I, Kocsor A, Kwok J (2006) Efficient kernel feature extraction for massive data sets. In: International Conference on Knowledge Discovery and Data Mining
34. Huang J, Ling CX (2005) Using AUC and accuracy in evaluating learning algorithms. IEEE Trans Knowl Data Eng 17(3):299–310
35. Demšar J (2006) Statistical comparisons of classifiers over multiple data sets. J Mach Learn Res 7:1–30