Journal of Engineering and Applied Sciences 13 (21): 9065-9077, 2018
ISSN: 1816-949X

Improved Performance of Support Vector Machine for Imbalanced Data Sets Using Oversampling and Optimization

Sana Saeed and Hong Choon Ong
School of Mathematical Sciences, University Sains Malaysia,
11700 Gelugor Penang, Malaysia

Abstract: Classification of imbalanced data sets particularly in the presence of noise is a significant problem in machine learning and data mining. Support Vector Machine (SVM) is one of the most renowned supervised classification algorithm. However, its performance becomes limited for imbalanced data sets. To improve the performance of SVM for imbalanced data sets including noisy borderline and real data sets, a methodology based on oversampling and optimization algorithm is proposed for two-class classification problems. By generating the synthetic samples in the minority class and searching the best choices of the parameters of SVM after minimizing the objective function, the performance of SVM is improved. To confirm the validity of the proposed methodology, an experimental study including noisy borderline and real imbalanced data sets was conducted. SVM was applied by using the proposed methodology, two optimization algorithms and one oversampling algorithm on all the data sets. The performance of SVM with all methods was evaluated using sensitivity, G mean and F-measure. A significantly improved performance of SVM was observed by using the proposed methodology.

Key words: Support vector machines, oversampling, optimization algorithm, noisy borderline imbalanced data sets, real imbalanced data sets, proposed methodology

INTRODUCTION

In many real-life data sets an uneven distribution of samples/instances is observed among the classes of a data. For two-class classification problems, this uneven distribution leads to one class having many more samples than the other class. A class having many more samples/instances is usually called a majority class and frequently represented by a negative class. The other class is called the minority class and represented by a positive class. However, this minority class is more significant for a researcher, from data mining perspective and its rareness needs a highly focused and intelligent approach (Krawczyk, 2016). These datasets are usually known as imbalanced data sets or skewed data sets. The skewness in these data sets usually exhibits the trouble in the canonical machine learning algorithms resulting in biased results by accommodating only the majority class. Major reasons for poor performance by these algorithms, on the imbalanced data sets are the assumptions of equal distribution of classes and equally distributed cost among these classes (Ganganwar, 2012).

Imbalanced data sets are getting more and more attention from researchers now a days. Three main approaches can be found in the literature for dealing with these data sets namely, sampling, algorithm level, feature selection approaches (Napierala and Stefanowski, 2016; Longadge and Dongre, 2013). There are two major issues in learning these types of data sets namely, lack of samples/instances and complexity of data. When learning these data sets, lack of samples/instances is a big issue. For the classification task, size of the dataset has an important role in building a good classifier. Therefore, lack of instances is a big obstacle for a good classifier. The second issue with these data sets is the complexity of data which can be defined as the level of separability of classes in a data set. If the data is linearly separable, the class Imbalance Ratio (IR) does not affect the performance of a classifier. However, class (IR) starts its influence on classifiers with the increasing degree of complexity in data set. The high complexity of data usually exhibits high noise, non-separability and overlapping classes (Phung *et al*., 2009). The class imbalance problem becomes more significant in the presence of noise. Noise

Corresponding Author: Sana Saeed, School of Mathematical Sciences, Institute of Post graduate (IPS), University Sains Malaysia, 11700 Gelugor Penang, Malaysia

in data sets can arise from many sources such as incorrectly labeling or the insufficient number of features during the data collection and data preparation stages, etc., Noise has an adverse effect on the machine learning algorithm. However, its effect gets worse in the presence of data imbalances because standard machine learning algorithms tend to treat minority class samples as noise (Weiss, 2004). So, noise has more influence on the minority class/rare instances than the majority class. Since, the presence of noise in the imbalanced datasets adds more problems in the classification task, therefore, learning from these, noisy imbalanced datasets is a big challenge for the researchers. The motivation of this research is also, accomplished from this prevailing machine learning and data mining issue.

SVM is a popular supervised machine learning algorithm, successfully handling classification task in many real-world applications, for example for credit scoring, text classification and bankruptcy prediction (Chaudhuri and De, 2011; Shin *et al*., 2005; Sun *et al*., 2009; Huang *et al*., 2007). SVM has a strong theoretical and mathematical background and a high generalization capability for finding global and non-linear classification solutions (Ben-Hur and Weston, 2010). However, its performance becomes limited for imbalanced noisy and borderline datasets (Imam *et al*., 2006; Eitrich and Lang, 2006). Other reasons behind the poor performance of SVM for imbalanced data sets will be discussed.

Classification performance of SVM is completely dependent on the optimal choice of its parameters (Bhadra *et al*., 2012; Alwan and Ku-Mahamud, 2017). Traditionally grid search approach was used, however, it is really time-consuming and laborious research to search parameter via. this approach. Therefore, the research question to be addressed in this study is how to improve the performance of SVM for imbalanced data sets?

The modern era has seen a lot of advancements in the form of new nature-inspired optimization algorithms which are more efficient and faster than the classical optimization algorithms. Therefore, it is planned to use optimization algorithm for the parameters selection of SVM instead of the traditional grid search. For the given classification task of imbalanced data sets, a methodology based on an oversampling and optimization algorithm for SVM is proposed.

Many studies found in the literature also employed different existing optimization algorithms based on metaheuristics for SVM whereas this study will use our own proposed optimization algorithm. However, to handle the imbalanced classes of the data sets a well-known oversampling algorithm, Synthetic Minority Over-Sampling Technique (SMOTE) will be engaged.

Literatur review: In data mining, prediction or classification problem has a significant role. However, this study area has to face problems in the presence of rare events because of their less frequency in the data sets. Misclassification in many situations can result in high costs. In addition to it, minority classes are usually considered as noise in the presence of imbalances by many traditional algorithms. For the binary class imbalanced problem, there are two classes in a data set usually termed as the majority and minority class. Minority class has less frequency and in fact a point of interest for the researchers (Chawla *et al*., 2004; He and Garcia, 2009; Galar *et al*., 2013; Sun *et al*., 2015). For handling imbalanced data sets, different methodologies have been proposed including algorithm level methodologies or using optimization techniques. The detailed reviews of all these methodologies and applications have been discussed by different researchers (Kotsiantis *et al*., 2006; Haixiang *et al*., 2017).

SVM has a strong theoretical and mathematical background and a high generalization capability for finding global and nonlinear classification solutions. Although this algorithm research efficiently on the balanced, it generates suboptimal solutions for the imbalanced datasets (Batuwita and Palade, 2010). The separating hyperplane for SVM Model developed with imbalanced data sets may be skewed towards the minority class. Another issue is if the training data sets get more imbalanced, support vectors also become imbalanced (Pant *et al*., 2011; Lessmann, 2004; Lee *et al*., 2015; Wu and Chang, 2003).

To improve the performance of SVM for imbalanced data sets, many types of research have been conducted sharing different ideas. Three types of methodologies are usually available in the literature; External learning, Internal learning and hybrid methodologies. The external learning methodologies include all the preprocessing data techniques like resampling and ensemble learning methods whereas algorithm level methodologies are recognized as internal learning and a combination of these two is recognized as hybrid methodologies. Since, SVM has a significant place among machine learning algorithms, so, it always remains a focused research area for scholars.

In 2005 an optimization for SVM was proposed in which derivative free numerical optimizer was applied for

this purpose (Eitrich and Lang, 2006). This study also introduced a new sensitive quality measure. An idea of optimized cost-sensitive SVM was proposed in which an effective wrapper framework incorporating AUC (Area Under the Curve) and G Mean into the objective function of SVM was introduced to gain a better performance of SVM. A subset of feature selection, parameters and misclassification cost were simultaneously optimized (Cao *et al*., 2013). For the optimal selection of SVM parameters using three metrics Area Under the roc Curve (AUC) accuracy and balanced accuracy were employed using computational data. Different levels of separability, different levels of unbalance and different levels of training sets were engaged (Jiang *et al*., 2014). Another study proposed an idea based on the fuzzy systems, i.e., entropy based fuzzy SVM (FSVM) for imbalanced data sets in which fuzzy class membership was determined based on the class certainty of samples (Fan *et al*., 2017).

To enhance the performance of SVM for imbalanced data sets, different resampling approaches were also, proposed by researchers. Most of them Suggested a use of Oversampling Technique (SMOTE) in combination with SVM. Different kinds of sampling techniques were proposed for example, a combined sampling approach using SMOTE and Tomek link with SVM for binary classification, a hybrid sampling approach using under and oversampling an ensemble method namely Bagging of Extrapolation Borderline SMOTE (BEBS) can be studied from the available literature (Sain and Purnami, 2015; Wang, 2014; Wang *et al*., 2017).

Since, the optimal performance of SVM is based on the good choice of its parameters with kernel settings, therefore, the model selection problem basically includes the search for the best values of the slack variable penalty weight (C) and kernel parameters which are supposed to be used for the classification task. Traditionally, a grid search selection is adopted. However, this method is time-consuming and does not perform well (Hsu and Lin, 2002). Due to the emerging use of metaheuristic techniques for optimization problems, the use of these techniques for SVM can also, be justified. For example, a genetic algorithm based feature selection and parameter optimization procedure for SVM can be found in the literature (Huang and Wang, 2006; Wang *et al*., 2011). Genetic Algorithm (GA) and Particle Swarm Optimization (PSO) for SVM and ant Colony Optimization (ACO) for SVM Model selection can also, be seen in the available studies on SVM optimization (Alwan and Ku-Mahamud, 2013; Ren and Bai, 2010; Blondin and Saad, 2010; Liu and Fu, 2014). An efficient memetic algorithm based on Particle Swarm Optimization (PSO) and Pattern Search (PS) was proposed for SVM parameter optimization. PSO was used for the exploration purpose while for exploitation PS was applied (Bao *et al*., 2013). Besides single objective optimization techniques, the use of multi-objective optimization for the SVM can also be found in the existing literature for SVM optimization. Few of them are discussing here: L1 norm SVM approach based on optimization of three objective functions, incorporating the error sum of two classes was proposed for the classification of imbalanced data sets (Askan and Sayin, 2014).

In a study on SVM, Model selection problem using multi-objective optimization method was proposed by Rosales-Perez *et al*. (2015) in which two famous terms of statistics, bias and variance were minimized. A surrogate-assisted evolutionary multi-objective was used to explore hyper parameter space. Another study proposed a combination of optimization and classification algorithm for SVM by using SMOTE and PSO (Cervantes *et al*., 2017). Wu *et al*. (2017) applied Two-phase Sequential Minimal Optimization (TSMO) and Differential Learning Particle Swarm Optimization (DPSO) for SVM.

SVM is a significant machine learning algorithm, therefore no one can deny its role as a significant classifier. However, its performance becomes limited in the presence of imbalances between the two classes of a data set. Therefore, many studies, discussed above, on SVM have presented many ideas, few of them shared the ideas of different optimizations in SVM and few of them presented different types of sampling techniques with SVM to enhance its performance for imbalanced data sets. However, we can able to find only a single study for SVM which proposed an idea of the combination of optimization and oversampling (Cervantes *et al*., 2017). The researcher of this study combined the standard PSO with SMOTE for SVM whereas our study will combine two fields (sampling and optimization) by using our own hybrid algorithm (CSCMAES) with SMOTE and thereafter the performance of SVM will be studied. So, this proposed methodology will be an addition to the existing literature on SVM.

MATERIALS AND METHODS

To improve the performance of SVM for binary imbalanced data sets including noisy borderline and real data sets, this study is going to introduce a new methodology which is a combination of the oversampling algorithm (SMOTE) and an optimization algorithm.

Before introducing the new methodology, few details on Support Vector Machines (SVM) an optimization algorithm and oversampling algorithm (SMOTE) are given.

Support vector machines: SVM as a nonlinear classifier can offer a better precision in many real-world applications. The way of making linear classifiers to nonlinear is to map the data from input space X to feature space F using a nonlinear function (Abe, 2005). In the feature space F the discriminant function can be written as:

$$g(x)=\theta^{T}\varnothing(x)+\theta_{0} \quad (1)$$

Kernel methods provide the best way of tackling this problem of mapping data to the high dimensional feature space instead of computing their dot products. Suppose that the weight vector may be expressed as a linear combination of training examples as:

$$\theta=\sum_{i}^{T_r}\beta_i x_i \quad (2)$$

Therefore, in terms of a discriminant function, it can be written as:

$$g(x)=\sum_{i}^{T_r}\beta_i x^{T}{}_i x_i+\theta_0 \quad (3)$$

In the feature space, Eq.1 can be written as:

$$g(x)=\sum_{i}^{T_r}\beta_i\varnothing(x^{T}{}_i)(x_i)+\theta_0 \quad (4)$$

In terms of kernel function Eq. 4 can be rewritten as:

$$g(x)=\sum_{i}^{T_t}\beta_i k(x,x_i)+\theta_o \quad (5)$$

The values of the set of parameters (β_i, θ_0) are determined during the training process. SVM has another set of parameters called hyperparameters, the soft margin constant C and any parameter of the involved kernel function (width of the Gaussian function or degree of the polynomial). Kernel parameters also have a significant influence on the decision boundary. The parameters of kernel function can control the tractability of the resulting classifier. Imbalanced data sets are usually not linearly separable, so, for these data sets nonlinear version of SVM with kernel functions are applied.

Hybrid algorithm: In this study, a hybrid algorithm is engaged which is based on the algorithm proposed by Saeed and Ong (2018). This hybrid algorithm took the advantages of Evolution Strategies (ES) and Swarm Intelligence (SI). Covariance Matrix Adaptation Evolution Strategy (CMA-ES) and Cuckoo Search (CS) are hired for this purpose. This hybrid algorithm is named as CSCMAES.

Covariance matrix adaptation evolution strategy: Evolutionary strategies are the most powerful evolutionary methods used for black box optimization problems without having the knowledge of the derivative. They are based on the evolutionary principles introduced by Darwin. Covariance Matrix Adaptation Evolution Strategy (CMA-ES) is a type of evolutionary strategies used for optimization of a continuous function. CMA-ES is one of the most powerful evolutionary strategy proposed by Hansen and Ostermeier (1997). The key idea of CMA-ES lies in its invariance properties which can be achieved by carefully designed variation, selection operators and its efficient self-adaptation of mutation distribution (Igel *et al*., 2007). CMA-ES research with three operations; Sampling of the new solutions with multivariate normal distribution Selection and recombination and Adapting of the covariance matrix. Adaptation of the covariance matrix with the help of rotation and scale of the mutation distribution provides guidance to the population (Hu and Qiao, 2015).

Cuckoo search: Cuckoo Search (CS), one of the most renowned nature inspired metaheuristic proposed by Yang and Deb (2009), for the continuous optimization problems. CS has gained popularity due to its fast convergence to a global solution. This algorithm is inspired by the Cuckoos, the fascinating birds not only due to their sounds but also because of their hostile reproductive approach. Ani and Guira are the two famous species of this bird (Cuckoo) which usually lay their eggs in shared nests. Usually, these species eradicate the other eggs to increase identification probability of their own eggs. The behavior of obligate brood parasitism and laying eggs in the nest of other species (other host birds) is adopted by many species.

If the host bird discovers that eggs are not their own they will get rid of these alien eggs or simply abandon its nest and built a new nest elsewhere (Yang and Deb, 2009). Based on the egg-laying behavior of Cuckoos, Cuckoo Search (CS) algorithm has the following three rules defined by the research in their proposed procedure for this algorithm; Each Cuckoo lay one egg at a time and dumps its egg in a randomly chosen nest. For the next

generation, the best nest with high-quality eggs are approved only. The number of host nests (n) is fixed and the egg laid by a Cuckoo is discovered by the host bird with a probability Pa. In this case, the host bird has two choices either to get rid of the egg or simply abandon the nest to build a completely new nest.

For the algorithm proposed by Saeed and Ong (2018) setting the objective function in the Cuckoo Search algorithm (CS) and after generating the initial solution, the best solution (X^s_{CS}) are produced at sth iteration. Then with the recombination operator of Covariance Matrix Adaptation Evolution Strategy (CMA-ES), the weighted means (m^s) are produced. Before moving to the next iteration in order to produce the new solution the best solution obtained from Cuckoo search and the weighted mean are plugged in into this new solution:

For the next iteration (s+1), X^s is used to get new solutions. Then the procedure of discovery and randomization are completed. All details are provided in the pseudocode of CSCMAES algorithm (Algorithm 1).

Synthetic minority over sampling technique: Synthetic Minority Over-Sampling Technique (SMOTE) is an oversampling algorithm for imbalanced datasets proposed by Chawla *et al*. (2002). Ignoring the previous concept of with replacement sampling, this sampling technique uses oversampling of the minority class by creating synthetic samples. Subject to the amount of oversampling requirement, neighbors from the k nearest are selected. For example, if the amount of oversampling needed 200% then only two of the five nearest neighbors are selected and produce one sample in the direction of each. For synthetic samples following steps are applied:

- Compute the difference between nearest neighbor and feature vector (sample)
- Generate a random number between 0 and 1, multiple the difference by this random number
- Add it to the feature vector under consideration. This will originate the selection of a random point beside the line segment between two specific features

This approach successfully forces the decision area of the minority class to become broader. The implementation of this algorithm requires five nearest neighbors (Han *et al*., 2005; Lusa, 2013).

Algorithm 1: Hybrid Algorithm

Begin
1. Setting the initial parameters number of nests n , no of solutions N_d
2. Setting the objective function f (x), x = (x_1, x_2, x_d) adjusting lower and upper bound of function and constraints (if any)
3. Initialize the Cuckoo search by generating the random initial solution of n host nests
4. Find the best solution (X^s_{CS}) from Cuckoo search at sth iteration
5. Initialize CMA-ES algorithm and generate m^s weighted means at sth iteration with the help of recombination operators
6. Generate the new solution X^s
7. Set the number of iterations and maximum number of iterations
8. While (number of iteration <Maximum iteration) or (stop criterion)
9. Produce the new solution at (s+1)th iteration by levy flights
10. Evaluate its quality or fitness
11. Choose a nest among n say (j) randomly
12. If ($f_i > f_j$) then
13. Replace j with the new solution
14. End if
15. A fraction (P_a) using adaptation procedure, abandoned the new nest and new ones are built
16. Keep the best solution
17. Rank the solution and find the current best
18. End while
19. Post process results and visualization

End

Proposed methodology: To improve the performance of SVM for the imbalanced dataset, this study presents a new methodology for the given task. Our proposed methodology is based on SMOTE and algorithm. SMOTE is used to generate synthetic instances in the minority classes of data sets whereas optimization algorithm is used to search optimal parameters of SVM in a random search space because of the nonlinear relationship between dependent and independent variables (Li *et al*., 2018). After applying an oversampling algorithm (SMOTE), the whole data set is randomly divided into three parts by using the ratios (60:20:20). The 60% data set is considered for training process while other two data sets are considered for validation and testing process. During the training process an optimal search for SVM parameters is made using a hybrid algorithm. SVM has two parameters to be optimized, C and kernel parameter. The first parameter of SVM, C, trades of between the misclassification of training examples and the simplicity of the decision surface. Small values of C makes the decision surface smooth while a large value of C aims at classifying all training examples correctly by allowing a model to choose more samples/instances as support vector. The second parameter of SVM, i.e., kernel parameter is basically the parameter of the function which is involved in the study. In this study, we are using only a single standard kernel function, radial basis kernel function (rbf) which can be defined as:

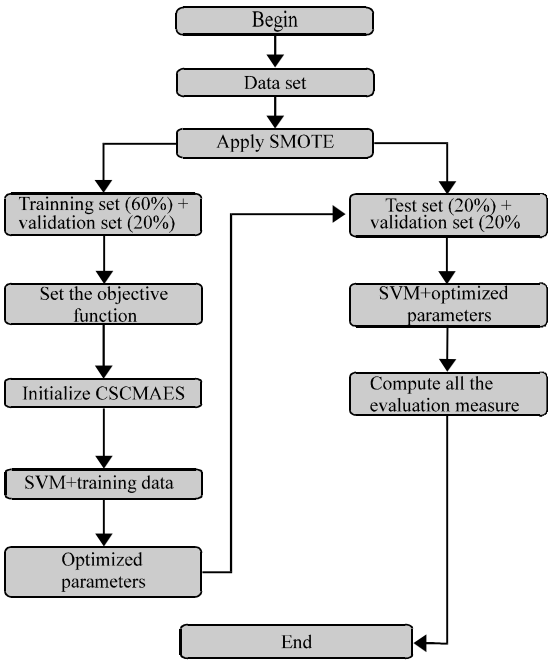

Fig. 1: Flow chart of the proposed methodology

$$k(x,x_i) = \exp\left(\gamma \| x\text{-}x_1 \|^2\right)$$

Rbf has only one parameter which defines the influence of a training example. Therefore two significant parameters, slack variable (C) and rbf parameter (γ) are optimized by minimizing the objective function during this training process. The objective function to be minimized in this study is the error of the minority (positive) class which can be computed by dividing the number of misclassified instances of the positive class by the total number of instances. After getting the optimized parameters of SVM, these are the plugin in test data to complete the whole classification task. Evaluation measures, sensitivity, G Mean and F measure are also computed on test data. A complete classification process of the proposed methodology is shown in a given flow chart in Fig. 1.

RESULTS AND DISCUSSION

To study the performance of the proposed methodology for binary classification of the imbalanced dataset an experimental study is conducted. Since, the proposed methodology is a combination of an oversampling algorithm and optimization algorithm therefore to justify an improved performance of SVM using the proposed methodology, its performance is compared with other optimization algorithms and oversampling algorithms separately. To see the performance of SVM by using only optimization algorithms, Cuckoo Search (CS) and Particle Swarm Optimization (PSO) with SVM, i.e., CS+SVM and PSO+SVM are employed. However, to study the performance of SVM using oversampling algorithm only, SMOTE is applied with SVM (SMOTE+SVM). Standard settings of PSO, CS and SMOTE algorithms are applied. To make a fair comparison of the proposed methodology with optimization and oversampling algorithms, three established performance evaluation measures for imbalanced data sets, namely Sensitivity (Sen), G Mean (G), F measure (F) are computed (Bekkar *et al*., 2013). All evaluation measures are computed with the help of the following four measures from a confusion matrix Table 1.

Table 1: Confusion matrix

Actual class	Predicted class negative	Positive
Minority class	True Negative (TN)	False Negative (FN)
Majority class	False Positive (FP)	True Positive (TP)

- True Negative = Negative examples are predicted truly negative
- False Positive = Negative examples are predicted to be positive
- False Negative = Positive examples are predicted to be negative
- True Positive = Positive examples are predicted truly positive

The above mentioned measures are computed with the help of the following equations:

$$\text{Sensitivity} = \frac{TP}{TP+FP}$$

$$\text{Specificity} = \frac{TN}{TN+FP}$$

$$\text{GMean} = \sqrt{\text{Sensitivity} \times \text{Specificity}}$$

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{F-measure} = \frac{2 \times \text{Sensitivity} \times \text{Specificity}}{\text{Sensitivity}+\text{Specificity}}$$

Two types of data sets are taken from a well-known datasets repository KEEL (Alcala-Fdez *et al*., 2011). Noisy borderline imbalanced datasets and Real imbalanced data sets.

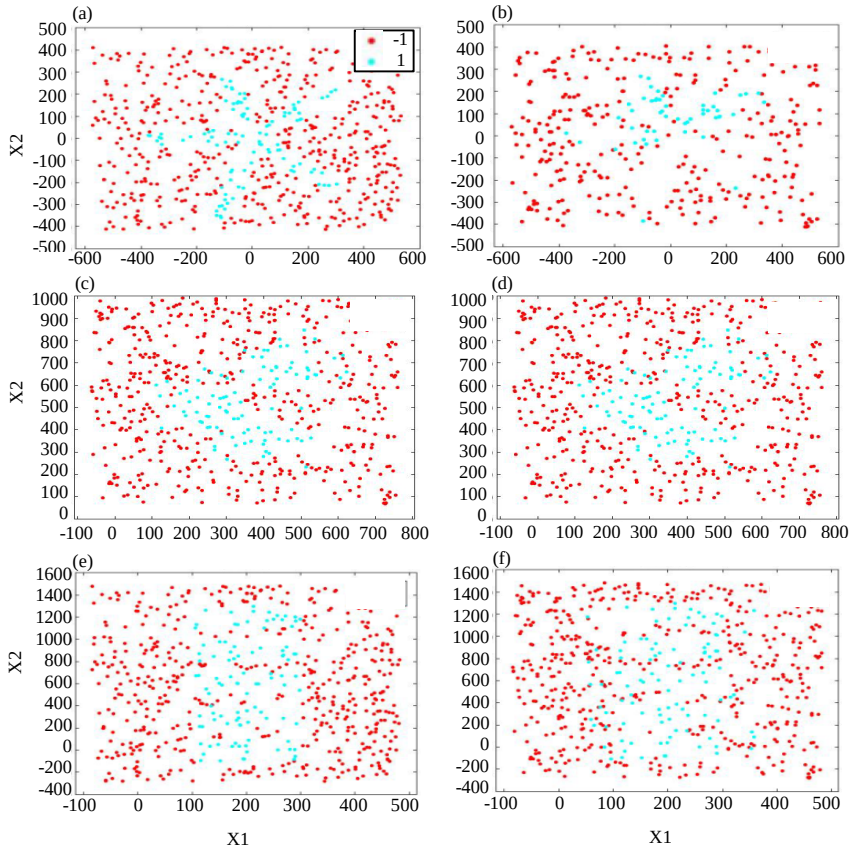

Fig. 2: Scatter plots of noisy borderline data sets: a) Scatter plot of clove 0; b) Scatter plot of clove 30; c) Scatter plot of paw 0; d) Scatter plot of paw 30; e) Scatter plot of subclus 0 and f) Scatter plot of subclus 3

Table 2: Performance of svm using all methods on noisy borderline imbalanced data sets

Data sets	Clove 0				Clove 30			
Methods	Sen	G	F	Time (sec)	Sen	G	F	Time (sec)
Proposed method	1.000	0.980	0.325	2.408	1.000	1.000	1.000	2.2500
PSO+SVM	0.000	0.000	0.000	1.076	0.784	0.885	0.879	0.8382
CS+SVM	0.000	0.000	0.000	0.766	0.686	0.822	0.8406	0.6140
SMOTE+SVM	0.0133	0.111	0.026	1.847	0.007	0.081	0.013	1.7860
Data sets	Paw 0				Paw 30			
Proposed method	1.000	1.000	1.000	3.184	0.961	0.980	0.980	3.107
PSO+SVM	0.070	0.264	0.129	0.688	0.000	0.000	0.000	0.779
CS+SVM	0.000	0.000	0.000	1.167	0.000	0.000	0.000	0.824
SMOTE+SVM	0.007	0.081	0.013	1.481	0.002	0.116	0.026	1.747
Data sets	Subclus 0				Subclus 30			
Proposed method	1.000	1.000	1.000	2.512	1.000	1.000	1.000	2.021
PSO+SVM	0.000	0.000	0.000	0.785	0.000	0.000	0.000	0.579
CS+SVM	0.000	0.000	0.000	0.775	0.000	0.000	0.000	0.643
SMOTE+SVM	0.020	0.141	0.039	1.9109	0.013	0.115	0.002	1.462

Noisy borderline data sets: Three types of synthetic noisy borderline imbalanced datasets (Clover, Paw and Subclus) with two different levels of disturbance ratios (0 and 30) are used. Each dataset contains 600 instances and with Imbalance Ratio (IR = 5). More details on these datasets can be found from (Lopez *et al*., 2013). The scatter plot of synthetic noisy borderline imbalanced data sets, Clover 0, 30, Paw 0, 30, Subclus 0 and 30 are displayed in Fig. 2. To study the performance of the proposed methodology in enhancing the classification ability of SVM in the presence of noise, the methodology and other three methods based on optimization and oversampling technique are applied. Performances of SVM with all methods are evaluated using three points criteria including sensitivity, G mean and F-measure. All results are given in Table 2.

For Clove 0 and 30 as the minority classes in these data sets look like five petal flowers (Fig. 2a,b and classification task of these types of data sets is really a challenging issue. Therefore, for Clover 0 data set,optimization algorithms with SVM (PSO+SVM, CS+SVM) only showed a poor performance on this data whereas oversampling algorithm with SVM (SMOTE+SVM) performed well as compared to optimization algorithms. However, its performance is not much outstanding (Table 2). A significant performance of the proposed methodology based on a combination of optimization and oversampling algorithm, can be observed from the same table (Sen = 1.000, G Mean = 0.980). However, F measure for the proposed methodology is only 0.325. On the other hand for Clover 30, both optimization algorithms and the proposed methodology all performed very well, resulting in high values of all three evaluation measures. However, for this data set, the oversampling algorithm in combination with SVM (SMOTE+SVM) having grid searched parameters, failed to show a good classification performance.

Paw 0 and 30 datasets are borderline and less noisy as compared to subclus data sets. Therefore, for Paw 0, PSO+SVM and SMOTE+SVM showed a moderate performance on the minority class whereas CS+SVM showed a very poor performance. However, for Paw 30, only the proposed methodology worked well. But a reduced performance of SVM using the proposed methodology is observed as compared to the performance on Paw 0 data set. This same happened with Subclus data sets with two disturbance levels, 0 and 30.

Sublcus 0 and 30, data sets are both highly noisy. In this situation where instances are noisy, only a combination of oversampling and optimization algorithm, i.e., proposed methodology remained successful in improving the performance of SVM. A low performance of SVM using oversampling is observed, resulting in minimum values of all evaluation measures. For all data sets, the time taken by the proposed methodology for the execution of the whole process is longer than the other methods.

Real imbalanced data sets: Seven real data sets with varying imbalance ratios are employed in this experimental study in which the validation of the proposed methodology for the improved performance of SVM is observed. These datasets are taken from a data set repository KEEL (Alcala-Fdez *et al*., 2011). Details of these data sets are provided in Table 3. These datasets are presented according to their imbalance ratio, from the lowest to the highest. Results of evaluation measures (sensitivity, G mean and F measure) from all methods are given in Table 4.

Table 3: Data sets description

Data sets	Data sets name	Imbalance Ratio (majority/minority) IR	Total instances
D1	Glass	1.82	214
D2	Pima	1.87	768
D3	Haberman	2.78	306
D4	Throid1	5.14	215
D5	Yeast1	9.08	514
D6	Cleveland	12.62	177
D7	wine1	29.17	1599

Testing time in seconds for all methods are also shown in the same table. The obtained results are compared with a recent study conducted by Li *et al.* (2018) for most of the datasets.

A decent performance of SVM is observed using all methods for approximately all real imbalanced data sets as compared to noisy borderline data sets. For the first data set (D1) with Imbalance Ratio, IR = 1.82, better performances of all methods are observed with high values of sensitivities, G Mean and F measure. Sensitivity by PSO+SVM and CS+SVM are same, i.e., 0.729. A slightly increased sensitivity is observed by SMOTE+SVM, i.e., (Sen = 0.794).

For two datasets D4 (IR = 5.14) and D6 (IR = 12.2), the same performances of the proposed method and two optimization algorithms (PSO+SVM, CS+SVM) are observed resulting in maximum values of all measures which can also, be justified by the study (Li *et al*., 2018). However, oversampling for these two data sets did not show an outstanding improved performance in SVM resulting in moderate values of evaluation measures.

For three data sets, D2 (IR = 1.87), D3 (IR = 2.78) and D5 (IR = 9.08), PSO+SVM, CS+SVM and SMOTE+SVM, showed their average performances whereas for these data sets the combined methodology performed well. A very close value to the G mean 0.960, reported by one of the applied methods in the first round of experiments in a study conducted by Li *et al.* (2018) is observed for D2 (G mean = 0.961 by the proposed method). G Mean for D3 given by the proposed methodology is 0.760 which is near to the G mean (0.79) reported in the previously mentioned study, produced during the first round of their experiments Execution time taken by PSO+SVM, CS+SVM and SMOTE+SVM are very near to each other. Since, the last data set D7 has the highest imbalance ratio (IR = 29.17), SVM performance using optimization algorithms (PSO, CS) becomes decreased whereas SVM using SMOTE moderately performed on this data set. For this dataset, the proposed methodology remained successful in producing maximum values of evaluation measures. Out of seven data sets, same performances are observed for two data sets whereas the proposed methodology took the leading position for five data sets (Table 4). However, the testing time taken by the proposed methodology is slightly longer than the other methods.

Table 4: Performance Of svm using all methods on real imbalanced data sets

Data sets/	D1				D2			
Methods	Sen	G	F	Time (sec)	Sen	G	F	Time (sec)
Proposed method	0.982	0.968	0.960	1.579	1.000	0.961	0.928	1.897
PSO+SVM	0.729	0.839	0.814	0.754	0.210	0.454	0.424	1.211
CS+SVM	0.729	0.844	0.823	3.697	0.857	0.893	0.400	1.513
SMOTE+SVM	0.794	0.891	0.855	0.435	0.685	0.000	0.000	1.322
Data sets	D3				D4			
Proposed method	0.978	0.760	0.825	1.295	1.000	1.000	1.000	0.925
PSO+SVM	0.234	0.471	0.333	1.194	1.000	1.000	1.000	0.643
CS+SVM	0.212	0.448	0.308	0.953	1.000	1.000	1.000	0.613
SMOTE+SVM	0.330	0.574	0.985	0.442	0.330	0.574	0.496	0.442
Data sets	D5				D6			
Proposed method	0.981	0.990	0.990	2.088	1.000	1.000	1.000	0.782
PSO+SVM	0.364	0.603	0.533	1.160	1.000	1.000	1.000	0.849
CS+SVM	0.181	0.426	0.307	1.001	1.000	1.000	1.000	0.819
SMOTE+SVM	0.194	0.440	0.325	0.430	0.013	0.115	0.002	0.334
Data sets	D7							
Proposed method	1.000	1.000	1.000	1.125				
PSO+SVM	0.000	0.000	0.000	0.803				
CS+SVM	0.000	0.000	0.000	0.504				
SMOTE+SVM	0.194	0.440	0.325	0.231				

Table 5: Average ranking of SVM performances by using all methods on noisy borderline data sets

	With respect to sensitivity		With respect to G mean	
Methods	Friedman test	Quade test	Friedman test	Quade test
Proposed method	1	1	1	1
PSO+SVM	3	3.03	3	3.03
CS+SVM	3.5	3.55	3.5	3.55
SMOTE+SVM	2.50	2.40	2.5	2.40
Test statistics	12.60	7.27	12.60	7.27
Critical values	7.60	3.29	7.60	3.29
Decision (5%)	Sig.	Sig.	Sig.	Sig.
	With respect to	F-measure	With respect to	Testing time
Proposed method	1	1	4	4
PSO+SVM0	3	2.92	1.33	1.23
CS+SVM	3.5	3.54	1.67	1.76
SMOTE+SVM	2.5	2.52	3	3
Test statistics	12.60	6.64	16.40	14.93
Critical values	7.60	3.29	7.60	3.29
Decision (5%)	Sig.	Sig.	Sig.	Sig.

Table 6: Average ranking of SVM performances by using all methods on real imbalanced data sets

	With respect to sensitivity		With respect to G mean	
Methods	Friedman test	Quade test	Friedman test	Quade test
Proposed method	1.2857	1.1786	1.2857	1.250
PSO+SVM	2.8571	2.9286	2.7857	2.7321
CS+SVM	2.8571	3.0357	2.7857	2.7671
SMOTE+SVM	2.8571	2.750	3.00	3.1071
Test statistics	4.80	3.44	4.93	2.94
Critical values	7.80	3.16	7.80	3.16
Decision (5%)	Insig	Sig	Insig	Insig
	With respect to	F-measure	With respect to	testing time
Proposed method	1.4286	1.3214	3.5714	3.6071
PSO+SVM	2.6429	2.5893	2.7143	2.6071
CS+SVM	2.9286	3.1607	2.5714	2.6786
SMOTE+SVM	2.8571	2.8214	1.1429	1.1071
Test statistics	3.21	2.69	12.77	6.33
Critical values	7.80	3.16	7.80	3.16
Decision (5%)	Insig	Insig	Sig.	Sig.

Average ranking of methods: Two rank tests (Friedman and Quade) are used to study the statistical ranking of all methods (proposed method, PSO+SVM, CS+SVM, SMOTE+SVM). These are very common tests in literature for determining the average ranking of non-normal datasets (Demsar, 2006; Garcia *et al*., 2007; Graczyk *et al*., 2010). The null hypothesis to be tested is on the average (median) the performances of all methods are equal. Friedman test follows an approximately a Chi-square distribution for a large number of blocks (data sets) and treatments (methods). As for this study, we have a small number of treatments and data sets, therefore, its critical values are taken from Friedman's table values. Table 5 is available in different books, one can easily find it on the internet also, Quade test is another rank test which follows an F distribution.

As for this study, the number of treatments is k = 4, the number of noisy borderline data sets (number of blocks) is b = 6 and the number of real data sets is b = 7. Therefore, the critical value for noisy borderline data sets at 5% level of significance by using Friedman test is 7.6 and by using Quade test it is 3.29. For real data sets, the critical value is 7.8 by using Friedman test and 3.16 by using Quade test. Ranks of all methods are determined with respect to four criterions: with respect to sensitivity, G mean, F-measure and testing time.

For noisy borderline datasets, average ranking by two non-parametric tests is shown in Table 5. Both tests showed significant results with respect to all criterions. The method producing the maximum values takes the first position, the second highest takes the second position and so on. But when the ranks are assigned with respect to time, the method which takes the minimum execution time takes the first position and the second minimum takes the sec rank and so, on. The same ranking scheme is adopted for real imbalanced data sets, presented in Table 6. These ranking and their positions with respect to

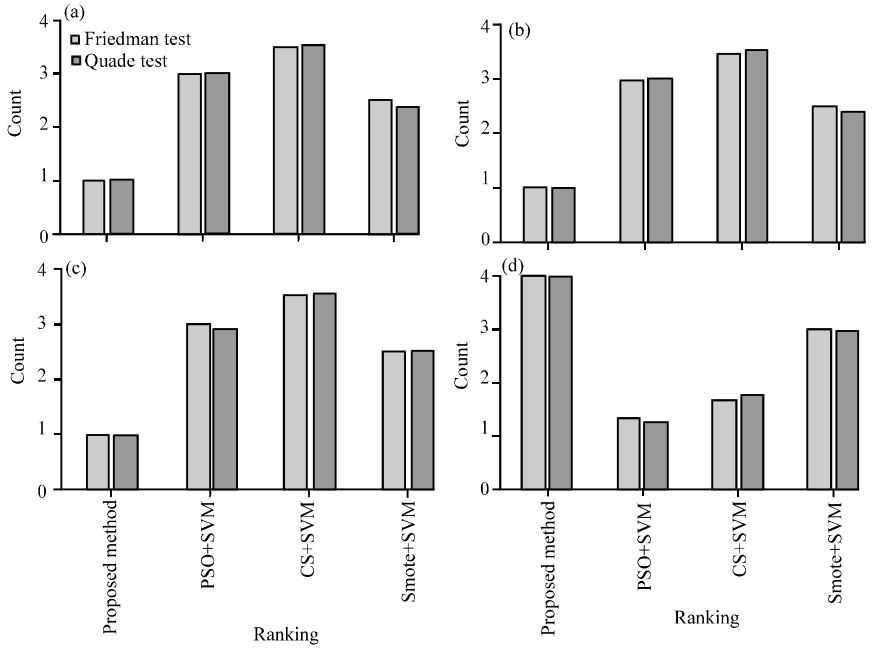

Fig. 3: Average ranking of SVM performances by using all methods on noisy borderline imbalanced data sets: a) With respect to sensitivity; b) With respect to g mean; c) With respect to f measure and d) With respect to time (sec)

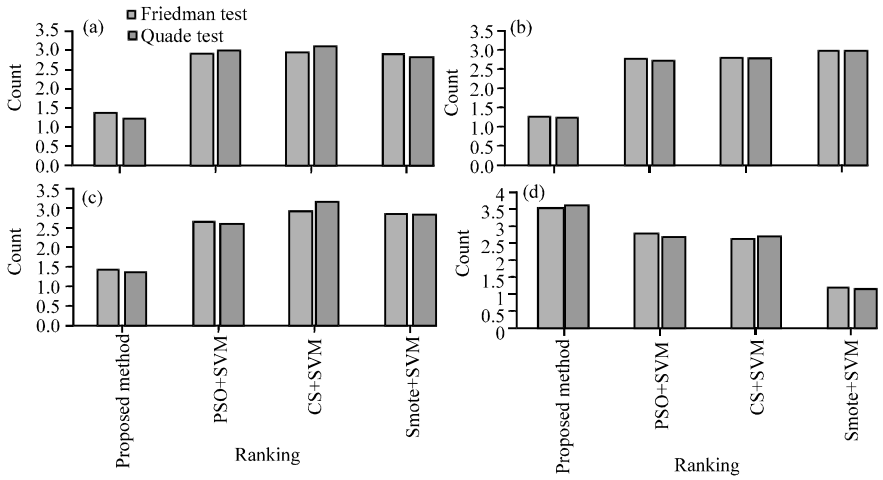

Fig. 4: Average ranking of SVM performances by using all methods on real imbalanced data sets: a) With respect to sensitivity; b) With respect to g mean; c) With respect to f measure and d) With respect to time (sec)

sensitivity, G Mean and F measure are also, shown graphically (Fig. 3), for noisy borderline data sets and for real imbalanced datasets (Fig. 4).

CONCLUSION

To improve the performance of SVM for imbalanced, noisy borderline data sets, a methodology based on a combined effect of oversampling (SMOTE) and optimization algorithm has been proposed. Because of the dependence of SVM on its optimal choices of parameters, these parameters were planned to search in a random search space by assigning the initial search directions to the algorithm after minimizing the error of the positive class as an objective function. In order to overcome the problems of imbalanced data sets, it was planned to generate synthetic samples by applying oversampling to minority class by using SMOTE.

To prove the efficiency of the proposed methodology an experimental study was conducted. Experiments were performed on the noisy borderline and real imbalanced data sets. SVM performance on imbalanced datasets was observed by applying the proposed methodology, two optimization algorithms (PSO+SVM, CS+SVM) and an oversampling algorithm (SMOTE+SVM).

Six synthetic noisy borderline and seven real imbalanced datasets were engaged. It was clearly seen that SVM showed a significantly improved performance by using the proposed methodology for the noisy borderline and real imbalanced data sets as compared to the performances of SVM by applying the optimization and oversampling algorithm separately.

REFERENCES

Abe, S., 2005. Support Vector Machines for Pattern Classification. Springer, Berlin, Germany, ISBN-13:978-1-85233-929-6, Pages: 345.

Alcala-Fdez, J., A. Fernandez, J. Luengo, J. Derrac and S. Garcia *et al.*, 2011. Keel data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework. J. Multiple Valued Logic Soft Comput., 17: 255-287.

Alwan, H.B. and K.R. Ku-Mahamud, 2013. Solving support vector machine model selection problem using continuous ant colony optimization. Intl. J. Inf. Process. Manage., 4: 86-97.

Alwan, H.B. and K.R. Ku-Mahamud, 2017. Mixed-variable ant colony optimisation algorithm for feature subset selection and tuning support vector machine parameter. Intl. J. Bio Inspired Comput., 9: 53-63.

Askan, A. and S. Sayin, 2014. SVM classification for imbalanced data sets using a multiobjective optimization framework. Ann. Oper. Res., 216: 191-203.

Bao, Y., Z. Hu and T. Xiong, 2013. A PSO and pattern search based memetic algorithm for SVMs parameters optimization. Neurocomput., 117: 98-106.

Batuwita, R. and V. Palade, 2010. Efficient resampling methods for training support vector machines with imbalanced datasets. Proceedings of the 2010 International Joint Conference on Neural Networks (IJCNN'10), July 18-23, 2010, IEEE, Barcelona, Spain, ISBN:978-1-4244-6916-1, pp: 1-8.

Bekkar, M., H.K. Djemaa and T.A. Alitouche, 2013. Evaluation measures for models assessment over imbalanced datasets. J. Inf. Eng. Appl., 3: 27-39.

Ben-Hur, A. and J. Weston, 2010. A User's Guide to Support Vector Machines. In: Data Mining Techniques for the Life Sciences, Carugo, O. and F. Eisenhaber (Eds.). Humana Press, New York, USA., ISBN:978-1-60327-240-7, pp: 223-239.

Bhadra, T., S. Bandyopadhyay and U. Maulik, 2012. Differential evolution based optimization of SVM parameters for meta classifier design. Procedia Technol., 4: 50-57.

Blondin, J. and A. Saad, 2010. Metaheuristic techniques for support vector machine model selection. Proceedings of the 10th International Conference on Hybrid Intelligent Systems (HIS'10), August 23-25, 2010, IEEE, Atlanta, Georgia, ISBN:978-1-4244-7363-2, pp: 197-200.

Cao, P., D. Zhao and O. Zaiane, 2013. An optimized cost-sensitive SVM for imbalanced data learning. Proceedings of the 17th Pacific-Asia Conference on Knowledge Discovery and Data Mining, April 14-17, 2013, Springer, Gold Coast, Australia, ISBN:978-3-642-37455-5, pp: 280-292.

Cervantes, J., F. Garcia-Lamont, L. Rodriguez, A. Lopez and J.R. Castilla *et al.*, 2017. PSO-based method for SVM classification on skewed data sets. Neurocomput., 228: 187-197.

Chaudhuri, A. and K. De, 2011. Fuzzy support vector machine for bankruptcy prediction. Appl. Soft Comput., 11: 2472-2486.

Chawla, N.V., K.W. Bowyer, L.O. Hall and W.P. Kegelmeyer, 2002. SMOTE: Synthetic minority Over-sampling technique. J. Artificial Intell. Res., 16: 321-357.

Chawla, N.V., N. Japkowicz and A. Kolcz, 2004. Editorial: Special issue on learning from imbalanced data sets. SIGKDD Explorations, 6: 1-6.

Demsar, J., 2006. Statistical comparisons of classifiers over multiple data sets. J. Mach. Learn. Res., 7: 1-30.

Eitrich, T. and B. Lang, 2006. Efficient optimization of support vector machine learning parameters for unbalanced datasets. J. Comput. Appl. Math., 196: 425-436.

Fan, Q., Z. Wang, D. Li, D. Gao and H. Zha, 2017. Entropy-based fuzzy support vector machine for imbalanced datasets. Knowl. Based Syst., 115: 87-99.

Galar, M., A. Fernandez, E. Barrenechea and F. Herrera, 2013. EUSBoost: Enhancing ensembles for highly imbalanced data-sets by evolutionary undersampling. Pattern Recognit., 46: 3460-3471.

Ganganwar, V., 2012. An overview of classification algorithms for imbalanced datasets. Int. J. Emerging Technol. Adv. Eng., 2: 42-47.

Garcia, S., A.D. Benitez, F. Herrera and A. Fernandez, 2007. Statistical comparisons by means of non-parametric tests: A case study on genetic based machine learning. Algorithms, 13: 95-104.

Graczyk, M., T. Lasota, Z. Telec and B. Trawinski, 2010. Nonparametric statistical analysis of machine learning algorithms for regression problems. Proceedings of the 14th International Conference on Knowledge-Based and Intelligent Information and Engineering Systems, September 8-10, 2010, Springer, Cardiff, Wales, UK., ISBN:978-3-642-15386-0, pp: 111-120.

Haixiang, G., L. Yijing, J. Shang, G. Mingyun and H. Yuanyue *et al*., 2017. Learning from class-imbalanced data: Review of methods and applications. Expert Syst. Appl., 73: 220-239.

Han, H., W.Y. Wang and B.H. Mao, 2005. Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning. Proceedings of the International Conference on Intelligent Computing, August 23-26, 2005, Hefei, China, pp: 878-887.

Hansen, N. and A. Ostermeier, 1997. Convergence properties of evolution strategies with de-randomized covariance matrix adaptation: The (μ/μ1, ?)-CMA-ES. Proceedings of the EUFIT'97: 5th Europe Congress on Intelligent Techniques and Soft Computing, September 8-11, 1997, Aachen, Germany, pp: 650-654.

He, H. and E.A. Garcia, 2009. Learning from imbalanced data. IEEE Trans. Knowledge Data Eng., 21: 1263-1284.

Hsu, C.W. and C.J. Lin, 2002. A simple decomposition method for support vector machines. Mach. Learn., 46: 291-314.

Hu, G.Y. and P.L. Qiao, 2015. An efficient improvement of CMA-ES algorithm for the network security situation prediction. Open Autom. Control Syst. J., 7: 1499-1517.

Huang, C.L. and C.J. Wang, 2006. A GA-based feature selection and parameters optimizationfor support vector machines. Expert Syst. Applic., 31: 231-240.

Huang, C.L., M.C. Chen and C.J. Wang, 2007. Credit scoring with a data mining approach based on support vector machines. Exp. Syst. Applic., 33: 847-856.

Igel, C., N. Hansen and S. Roth, 2007. Covariance matrix adaptation for multi-objective optimization. Evolut. Comput., 15: 1-28.

Imam, T., K.M. Ting and J. Kamruzzaman, 2006. Z-SVM: An SVM for improved classification of imbalanced data. Proceedings of the 19th Australasian Joint Conference on Artificial Intelligence, December 4-8, 2006, Springer, Hobart, Australia, ISBN:978-3-540-49787-5, pp: 264-273.

Jiang, P., S. Missoum and Z. Chen, 2014. Optimal SVM parameter selection for non-separable and unbalanced datasets. Struct. Multidiscip. Optim., 50: 523-535.

Kotsiantis, S., D. Kanellopoulos and P. Pintelas, 2006. Handling imbalanced datasets: A review. GESTS. Intl. Trans. Comput. Sci. Eng., 30: 25-36.

Krawczyk, B., 2016. Learning from imbalanced data: Open challenges and future directions. Prog. Artif. Intell., 5: 221-232.

Lee, J., Y. Wu and H. Kim, 2015. Unbalanced data classification using support vector machines with active learning on scleroderma lung disease patterns. J. Appl. Stat., 42: 676-689.

Lessmann, S., 2004. Solving imbalanced classification problems with support vector machines. Proceedings of the International Conference on Artificial Intelligence (ICAI'04), June 21-24, 2004, CSREA Press Publisher, Las Vegas, Nevada, pp: 214-220.

Li, J., S. Fong, R.K. Wong and V.W. Chu, 2018. Adaptive multi-objective swarm fusion for imbalanced data classification. Inf. Fusion, 39: 1-24.

Liu, X. and H. Fu, 2014. PSO-based support vector machine with Cuckoo search technique for clinical disease diagnoses. Sci. World J., 2014: 1-7.

Longadge, R. and S. Dongre, 2013. Class imbalance problem in data mining review. Intl. J. Comput. Sci. Netw., 2: 1-6.

Lopez, V., A. Fernandez, M.J. del Jesus and F. Herrera, 2013. A hierarchical genetic fuzzy system based on genetic programming for addressing classification with highly imbalanced and borderline data-sets. Knowl. Based Syst., 38: 85-104.

Lusa, L., 2013. SMOTE for high-dimensional class-imbalanced data. BMC. Bioinf., 14: 106-121.

Napierala, K. and J. Stefanowski, 2016. Types of minority class examples and their influence on learning classifiers from imbalanced data. J. Intell. Inf. Syst., 46: 563-597.

Pant, R., T.B. Trafalis and K. Barker, 2011. Support vector machine classification of uncertain and imbalanced data using robust optimization. Proceedings of the 15th WSEAS International Conference on Computers, July 15-17, 2011, WSEAS, Stevens Point, Wisconsin, USA., ISBN:978-1-61804-019-0, pp: 369-374.

Phung, S.L., A. Bouzerdoum and G.H. Nguyen, 2009. Learning Pattern Classification Tasks with Imbalanced Data Sets. In: Pattern Recognition, Yin, P. (Ed.). IntechOpen, Vukovar, Croatia, pp: 193-208.

Ren, Y. and G. Bai, 2010. Determination of optimal SVM parameters by using GA/PSO. J. Comput., 5: 1160-1168.

Rosales-Perez, A., J.A. Gonzalez, C.A.C. Coello, H.J. Escalante and C.A. Reyes-Garcia, 2015. Surrogate-assisted multi-objective model selection for support vector machines. Neurocomput., 150: 163-172.

Saeed, S. and H.C. Ong, 2018. A bi-objective hybrid algorithm for the classification of imbalanced noisy and borderline data sets. Patt. Anal. Appl., 1: 1-20.

Sain, H. and S.W. Purnami, 2015. Combine sampling support vector machine for imbalanced data classification. Procedia Comput. Sci., 72: 59-66.

Shin, K.S., T.S. Lee and H.J. Kim, 2005. An application of support vector machines in bankruptcy prediction model. Exp. Syst. Applic., 28: 127-135.

Sun, A., E.P. Lim and Y. Liu, 2009. On strategies for imbalanced text classification using SVM: A comparative study. Decis. Support Syst., 48: 191-201.

Sun, Z., Q. Song, X. Zhu, H. Sun and B. Xu *et al*., 2015. A novel ensemble method for classifying imbalanced data. Pattern Recognit., 48: 1623-1637.

Wang, L., G. Xu, J. Wang, S. Yang and L. Guo *et al*., 2011. GA-SVM based feature selection and parameters optimization for BCI research. Proceedings of the 7th International Conference on Natural Computation (ICNC'11) Vol. 1, July 26-28, 2011, IEEE, Shanghai, China, ISBN:978-1-4244-9950-2, pp: 580-583.

Wang, Q., 2014. A hybrid sampling SVM approach to imbalanced data classification. Abstr. Appl. Anal., 2014: 1-7.

Wang, Q., Z. Luo, J. Huang, Y. Feng and Z. Liu, 2017. A novel ensemble method for imbalanced data learning: Bagging of extrapolation-SMOTE SVM. Comput. Intell. Neurosci., 2017: 1-11.

Weiss, G.M., 2004. Mining with rarity: A unifying framework. ACM SIGKDD Explorations Newsl., 6: 7-19.

Wu, G. and E.Y. Chang, 2003. Class-boundary alignment for imbalanced dataset learning. Proceedings of the ICML 2003 Workshop on Learning from Imbalanced Data Sets II, August 21, 2003, ICML, Washington, DC., USA., pp: 49-56.

Wu, S.J., V.H. Pham and T.N. Nguyen, 2017. Two-phase optimization for support vectors and parameter selection of support vector machines: Two-class classification. Appl. Soft Comput., 59: 129-142.

Yang, X.S. and S. Deb, 2009. Cuckoo search via. Levy flights. Proceedings of the World Congress on Nature and Biologically Inspired Computing (NaBIC), December 9-11, 2009, IEEE, Coimbatore, India, ISBN:978-1-4244-5053-4, pp: 210-214.