Journal of Big Data

RESEARCH

Open Access

Improved cost-sensitive representation of data for solving the imbalanced big data classification problem

Mahboubeh Fattahi, Mohammad Hossein Moattar[*] and Yahya Forghani

*Correspondence:
moattar@mshdiau.ac.ir
Department of Computer Engineering, Mashhad Branch, Islamic Azad University, Mashhad, Iran

Abstract

Dimension reduction is a preprocessing step in machine learning for eliminating undesirable features and increasing learning accuracy. In order to reduce the redundant features, there are data representation methods, each of which has its own advantages. On the other hand, big data with imbalanced classes is one of the most important issues in pattern recognition and machine learning. In this paper, a method is proposed in the form of a cost-sensitive optimization problem which implements the process of selecting and extracting the features simultaneously. The feature extraction phase is based on reducing error and maintaining geometric relationships between data by solving a manifold learning optimization problem. In the feature selection phase, the cost-sensitive optimization problem is adopted based on minimizing the upper limit of the generalization error. Finally, the optimization problem which is constituted from the above two problems is solved by adding a cost-sensitive term to create a balance between classes without manipulating the data. To evaluate the results of the feature reduction, the multi-class linear SVM classifier is used on the reduced data. The proposed method is compared with some other approaches on 21 datasets from the UCI learning repository, microarrays and high-dimensional datasets, as well as imbalanced datasets from the KEEL repository. The results indicate the significant efficiency of the proposed method compared to some similar approaches.

Keywords: Feature selection, Feature extraction, Imbalanced data, Big data classification, Cost sensitive, Optimization

Introduction

In recent decades, with the increase in the size of data in medical science, data called microarray data have emerged. Microarray data is data that is extracted from tissue and cell samples. This type of data are important in diagnosing the disease and types of cancerous masses in medicine [1]. This increase in dimensions increases the computational cost of the system and leads to a decrease in the classification accuracy rate [2].

The high size of the data set makes feature selection one of the most fundamental and important topics in machine learning. As the number of features increases, the efficiency of learning algorithms initially increases, but from a certain point onwards, increasing the number of features not only does not improve the performance of the machine learning algorithm, but sometimes reduces the performance of these algorithms. In addition to this problem, with increasing the number of features, the need for more data samples increases, which increases the temporal and spatial complexity of the problem [3]. The features in the data can be divided into three general categories:

> Irrelevant features: These are features that have little information load and have nothing to do with data mining goals, so they usually reduce the performance of data mining algorithms.
> Redundant Features: Features that relate to other features and are not directly unrelated, such as features that can be used to compute other features.
> Relevant features: These are features that have a great impact on the data classification accuracy and are the main purpose of feature selection methods.

The dimensionality reduction techniques can be classified into two main groups [4–6]: feature selection and feature extraction. In feature selection, best features are selected based on their contribution to the final performance of the approach. In this type of dimensionality reduction some information can be lost. On the other hand, feature extraction or mapping approaches aim to find another representation of the data so that the new representation (i.e. features) improves some specific criteria such as maintaining the information [i.e. Principle Component Analysis (PCA)], discriminating the classes [i.e. Linear Discriminant Analysis (LDA)] or maintaining the local or global structure of the data (i.e. manifold learning methods). In feature extraction, the dimension can be decreased without losing much initial feature information [4, 7–9].

High dimensionality and imbalance are common problems in microarray data. Imbalance is one of the major crises in classification and the challenge becomes more acute when the data set has a large number of features. Traditional classification usually favors the majority class for attribute selection, leading to poor performance for parameter setting or selecting attributes that better describe the majority class. In order to solve the problem of imbalance, there are different solutions that can be divided into two general categories based on data and model. In data-based methods, an attempt is made to strike the expected balance by reducing the majority class data or generating data from the minority class distribution. In model-based methods, an attempt is made to build a model that is sensitive to the cost of incorrectly classifying minority class data. These methods are called cost-sensitive methods or model-based methods for short.

In this manuscript, we look for a space in which data that are similar in nature are inherently close. We do not reduce the number of our samples, but try to have a new space in which we can better represent the data. A new space-based method in which data is better separated and the problem of data imbalance in the reduced space is considered. The general purpose of this paper is to provide a model by considering the problem of data balance, solving the optimization problem in a combined method of selecting and extracting features simultaneously and improving accuracy and precision.

The organization of the article is as follows. Related works are studied in "Related works" section. The proposed method is presented in "The proposed method" section. "Experiments" section evaluates the proposed method and also compares its performance with other methods. Finally, "Conclusion" section concludes with a discussion and conclusion.

Related works

Feature reduction as a preprocessing step can remove irrelevant data, noise and additional features. Feature reduction is based on two main methods of feature selection and feature extraction. In this section, we discuss some important related works with regard to this classification [4].

Feature extraction

Feature extraction methods extract new features from the original dataset, and is very useful when we want to reduce the number of resources required for processing without losing the relevant feature dataset [4]. Instead of deleting a few feature, the input data space changes. When data is mapped from the input space to a space with a smaller dimension by a transformation; the nature of the basic features are changed.

Principle Component Analysis (PCA) is a simple feature extraction and embedding method. PCA[1] is an unsupervised method which is widely used as a baseline feature mapping approach but it performs the task in an unsupervised manner. In contrast, our method is actually an supervised method that uses labels in the feature reduction process. Authors of [13] present a dimension reduction algorithm for information spaces. This algorithm reduces the dimensions of space by maintaining a simplex structure, and is able be used as a black-box method to speed up algorithms which operate in information divergence spaces. It shows how to embed information distances like the x^2 and Jensen Shannon divergences efficiently in low dimensional spaces while preserving all pairwise distances. Other than the definition of the feature space, the main difference between this method and the proposed

[1] Principal component analysis.

approach is that we consider the miss-classification cost in feature extraction which aims to solve the imbalanced data problem.

In [10] a method is presented that is effective in dealing with data with multiple views. In this type of data, there are several different views for each data at the same time. It searches for a space called x with a lower dimension than the input space, which has information from all views and can also be returned from that space to all views.

Feature selection

The main purpose of feature selection is to select the appropriate number of attributes to perform the classification tasks [12].

Conventional feature selection methods perform feature selection operations over the entire sample space. The filter-based local feature selection algorithm [14] is proposed based on the artificial immune system, which determines a subset of the relevant local feature for each area adjacent to the sample space. This algorithm introduces a selection algorithm to optimize the search space for attribute subsets and adopts the idea of local clustering as an evaluation criterion that maximizes between class distances and minimizes within class distance.

Ref. [9] tries to select good features by optimizing multivariate criteria based on sparse representation. To measure the complexity of classification under different feature spaces, first a feature evaluation criterion is proposed, called counting region covering (CRC). Then, by simultaneously optimizing the classification error rate and the separation boundary complexity, a feature selection framework is provided. The proposed approach in [15] performs the feature selection process based on the minimization of the generalized error limit. This method simultaneously performs the classification feature reduction. In this method, a linear model inspired by the support vector machine is used, which performs the classification operation on the reduced dimensional data. Therefore, feature selection and classification are done simultaneously.

In [10], multiple kernel learning feature selection (MKL-FS) uses kernel methods to search for the complex properties of each feature. However, the available kernels are usually limited to positive constraints. In fact, certain negative kernels can often perform better in real applications. However, due to the non-convexity of indeterminate kernels, most methods are usually not practical and relevant researched are relatively limited. Also, a two-step algorithm for optimizing indefinite kernel support vector machine (IKSVM) and kernel combination coefficients is proposed. In [11] an approach is proposed to classify web text documents using the benefits of a hierarchical structure to remove words from attribute vectors that are not related to the WordNet lexical categories.

In [16], a criterion for evaluating the selected features based on the quality of the features is presented. The main idea is to use a sparse representation to test each feature independently. Also, the feature-based classification method has been used to evaluate the proposed method. Authors of [17] propose an effective distance-based feature selection (ED-Relief) method, which is used as a complex distance measurement to deal with the simultaneous optimization of within class and between class distances also in [18], an algorithm for feature selection based on associative rules and an integrated classification algorithm based on random sampling are proposed.

Authors of [19], introduced a new sample augmentation method called MAHAKIL. They believed that selecting samples very close to their neighbors would result in little variation in the samples produced in the minority class. Therefore, they used the characteristics of the two samples as parent samples to produce a new sample. A new feature selection method based on interaction information (II) is proposed in [32] to provide high-level interaction analysis and improve the search method in the feature space.

Ref. [20] propose a new hybrid feature selection called the IGIS algorithm for selecting features based on interaction information. This algorithm uses the JMI[2] criterion to find candidate attributes to add to the attribute set and adds one attribute to the currently selected subset at any time. By adding an attribute to the selected attribute set, the attribute list is recalculated.

Interacting features are those that appear to be irrelevant or weakly relevant with the class individually, but when it combined with other features, it may highly correlate to the class. Interacting features are feature that, among other features, may be related to the class. Those appear to be separately irrelevant or weak with the class. Discovering feature interaction is a challenging task in feature selection. In [21], a novel feature selection algorithm considering feature interaction is proposed. Mutual information-based feature selection algorithms, although performing well in many cases, currently suffer from two drawbacks: (1) Ignoring feature interaction. (2) Over-estimation of some features. To overcome these shortcomings, [22] proposes a new filter feature selection algorithm based on WJMI-weighted mutual information. Prevents over-estimation of some features by considering feature interaction.

Imbalance learning

One of the main and simple methods of data reduction for imbalanced datasets is presented in [23], which accidentally deletes some of the majority class data. The Condensed Nearest Neighbor Rule (CoNN) method [24] uses a similar approach to remove samples that are farther apart than the majority of the data. Unlike [24], the [25] eliminates noisy and near-boundary specimens. In [26], a clustering method is used to maintain the distribution of minority and majority class data after deleting the data. An approach based on evolutionary algorithms has also been performed in [27] in which the selection of samples for deletion is done as a search problem.

One of the most popular data generation methods for the minority class is called the SMOTE method [28]. In this method, samples for the minority class are generated through the interpolation of neighboring data. Some approaches have been proposed to address SMOTE weaknesses. In [29], an SMOTE-inspired method to reduce the tendency to overlap between majority and minority classes is used which is called

[2] joint mutual information.

Borderline-SMOTE. Also LN-SMOTE [30] and safe-level SMOTE [31] are research methods development [32].

In [33] a method called cost-sensitive high-margin support vector machines is proposed. In this method, the goal is to increase the margin of the minority class and reduce the margin of the majority class. This operation is performed by manipulating the cost parameter C in the support vector machine and dividing it into C_+ for positive class data and C_- for negative class data. If the positive class is in the minority, the C_+ parameter is selected as a larger number and vice versa. In [34], a cost-sensitive issue is included in the extreme learning machine (ELM) classification with a similar approach.

In [35], a new risk forecasting method is proposed as imbalanced classification and solves the feature selection problem. In particular, a high-margin loss function is presented in which the weight of the samples is involved. Accordingly, an optimization objective function is designed with a soft tuning of one to improve performance, which is solved in an iterative context.

SMOTE-based class-specific learning is proposed in [36] and uses minority sampling in the kernel space to solve the class imbalance problem. Motivated by weighted kernel-based SMOTE (WKSMOTE), this method proposes a SMOTE class-specific extreme learning machine (SMOTECSELM), a class-specific extreme learning machine (CS-ELM), which takes advantage of minority and class-specific sampling.

Other works

In [37], a low-rank regression model is proposed for feature extraction and feature selection from images without vectorization. To effectively solve the objective function, an optimization-based alternative to Lagrangian coefficients has been developed. In [38], by extracting the features and selecting the features in a cascading manner, and the Pigeon Inspired based Optimization (PIO) method is used to select the features. In [39], a hybrid approach is performed simultaneously by reducing the majority data with the rough set theory and increasing the minority data using the SMOTE method simultaneously. These methods and methods similar to [39] and [40] are among the data-based hybrid methods. In [41], a method has been designed to select a feature that emphasizes two issues. One is the problem of class imbalances and the other is the large size of the data. Also, [39] proposes a cost-sensitive approach in the context of concave optimization problem and proposes the solution through a Newtonian-like process. In order to prevent the explosion of data space dimensions while maintaining the statistical coherence of a part of the data set selected for teaching, [42] has developed an approach for selecting training data based on Pareto analysis performed on classification descriptors. It also provides empirical evidence that this approach retains its validity, even when compared to traditional space-reduction methods and classical machine learning algorithms.

Ref. [43] proposes a new framework that makes it possible to identify anomalous data points in large volumes of data with high dimensional problems. Authors of

Table 1 Frequently used notations and descriptions

Notations	Descriptions
Z	The result of the feature extraction operation on X
X	Data matrix
Q	The mapping matrix
Y	Data label
n_+	The number of positive class data
n_-	The number of negative class data
b	The bias
ξ	The slack variable
n	The whole number of data points
ε	The reconstruction error
ω	Denotes the coefficients of separating hyper-plane
v_j^-	The empirical mean of the second order moment of the jth feature in the negative classes respectively
v_j^+	The empirical mean of the second order moment of the jth feature in the positive classes respectively

[44] proposed methods for reducing the number of variables that include more information and for reliable classification, as well as several methods for reducing dimensions and classification.

The tasks reviewed generally use only one feature selection or extraction operation to reduce the size. Of course, hybrid methods were also available, but their efforts are mainly aimed at combining filter, embedded and wrapper approaches, and do not combine the main categories of selection and extraction at the same time. In this paper, by introducing the proposed method in the form of an optimization problem that simultaneously solves the feature selection and extraction problem, a context for using the benefits of both approaches is provided. The proposed optimization problem also includes a cost-sensitive function that is designed to make the model resistant to data with imbalanced labels, while creating a balance without manipulating the data.

The term cost-sensitive refers to creating resistance in the feature reduction process to imbalanced data. The existence of this resistor is embedded within the proposed optimization problem. Therefore, the proposed method is used based on the feature space to solve the problem of imbalance.

The proposed method

Assume that $X \in \mathbb{R}^{d \times n}$ is a data matrix that represents n data d dimensional in which x_i is equal to i-th data point. The above data label is represented by the vector $Y = \{y_1 \ldots y_n\}$ where $y_i \in \{-1. +1\}$. To clarify the rest of the proposed approach, Table 1 is included which summarizes the frequently used notations. Matrix $Z \in \mathbb{R}^{m \times n}$ is a reduced latent representation of Z where $m \ll d$. In other words, the Z matrix is the result of the feature extraction operation on X, which can be generated by the following mapping:

$$X = QZ + \varepsilon, \tag{1}$$

where $Q \in \mathbb{R}^{d\times m}$ is the mapping matrix and $\varepsilon \in \mathbb{R}^{d\times n}$ is the reconstruction error. In order to minimize the reconstruction error, a simple solution can be the soft minimization of the following equation:

$$\min_{z.Q} \frac{1}{n}\sum_{i=1}^{n} \|x_i - Qz_i\|^2. \tag{2}$$

Suppose that the feature selection is done on input space through a diagonal matrix σ such that $diag(\sigma)$ consists of 0 and 1 where $\sigma_j = 1$, denotes that the jth feature is selected from a dataset and vice versa. Assuming feature selection by σ, Eq. (2) will be written as follows:

$$\min_{z.Q} \frac{1}{n}\sum_{i=1}^{n} \|x_i\sigma - Qz_i\|^2. \tag{3}$$

Adding the regularization term to the above optimization problem, the following equation is obtained:

$$\min_{z.Q} \frac{1}{n}\sum_{i=1}^{n} \|x_i\sigma - Qz_i\|^2 + C\sum_{i=1}^{n}\|z_i\|^2, \tag{4}$$

where the non-negative parameter C indicated the penalty for the regularization term. In addition, to control the Q matrix and prevent large values, $Q^TQ = I$ constraint is added to the objective function. Therefore, the following optimization problem is obtained:

$$\begin{aligned}&\min_{Z.Q} \frac{1}{n}\sum_{i=1}^{n} \|x_i\sigma - Qz_i\|^2 + C\sum_{i=1}^{n}\|z\|^2\\ &s\,t \quad Q^TQ = I.\end{aligned} \tag{5}$$

In order to increase the discriminability of classes and preserving the primitive geometry structure in the reduced space Z, another term is also added to objective function named as locally alignment. The following problem is attained by adding this term to an objective function. $LA(Z)$ will be discussed later.

$$\begin{aligned}&\min_{Z.Q} \frac{1}{n}\sum_{i=1}^{n} \|x_i\sigma - Qz_i\|^2 + C\sum_{i=1}^{n}\|z_i\|^2 + \boldsymbol{LA}(Z)\\ &s.\,t. \quad Q^TQ = I.\end{aligned} \tag{6}$$

where $LA(Z)$ would be as:

$$\boldsymbol{LA}(Z) = \frac{1}{nk_1}\sum_{i=1}^{n}\sum_{j=1}^{k_1}\left\|z_i - z_{ij}\right\|^2 - \frac{\beta}{nk_2}\sum_{i=1}^{n}\sum_{p=1}^{k_2}\left\|z_i - z^{ip}\right\|^2. \tag{7}$$

Minimization of Eq. (7) means that distance of reduced data z_i from k_1 intra-class nearest neighbors should be decreased and the distance from k_2 farther neighbors from the opposite class should be increased. In Eq. (7), z_{ij} denotes the jth nearest neighbor for ith data point with the similar label; and z^{ip} is the pth nearest neighbor for ith data from the opposite class. Also, the parameter β indicates the importance level of the second term in the equation. It must be noted that LA makes the algorithm supervised because the class labels must be known to compute intra-class or inter-class neighbors.

The optimization problem is similar to the support vector machine (SVM) optimization [10] and is formulated as follows which performs simultaneous feature selection and classification.

$$\begin{aligned}
&\min_{\omega.b.\sigma.\xi}\ \mathbf{1}^T\xi \\
&s.\ t. \\
&y_i\left(\sum_{j=1}^{m} z_{ij}\omega_j\sigma_j + b\right) \geq 1-\xi_i \quad i=1\ldots n \\
&\omega^T\omega \leq 1 \\
&R_+ \geq \sum_{j=1}^{m} v_j^+\sigma_j^2 \\
&R_- \geq \sum_{j=1}^{m} v_j^-\sigma_j^2 \\
&\xi.\sigma \succcurlyeq 0.
\end{aligned} \tag{8}$$

In (8), ξ is the slack variable and $\mathbf{1}$ is a vector of ones having the length equal to ξ, the matrix multiplication of $\mathbf{1}^T\xi$ denotes the summation of the ξ values. Also, ω denotes the coefficients of separating hyper-plane, b is the bias, and z_{ij} is the jth feature of the ith data. Also,

$$v_j^- = \frac{1}{n_-}\sum_{i\in I_-}\left(z_{ij}\right)^2 \quad and \quad v_j^+ = \frac{1}{n_+}\sum_{i\in I_+}\left(z_{ij}\right)^2. \tag{9}$$

Which means the empirical mean of the second order moment of the jth feature in the negative and positive classes respectively. n_- and n_+ are the number of negative and positive class data, respectively.

One of the constraints in (8), is to prevent from exorbitant growth of $\sum_{j=1}^{m} v_j^+\sigma_j^2$ and $\sum_{j=1}^{m} v_j^-\sigma_j^2$ which have the upper bound R_+ and R_- respectively. It is evident that by choosing a low value for R_+ and R_-, the optimum solution for σ would include more zero values and consequently it results in a higher feature reduction rate.

Another constraint in the problem (8), i.e. $\omega^T\omega \leq 1$, is the bounding over l_2-norm of ω which means a regularization role. Now the final optimization problem resulting from combining the two relations (6) and (8) is as follows:

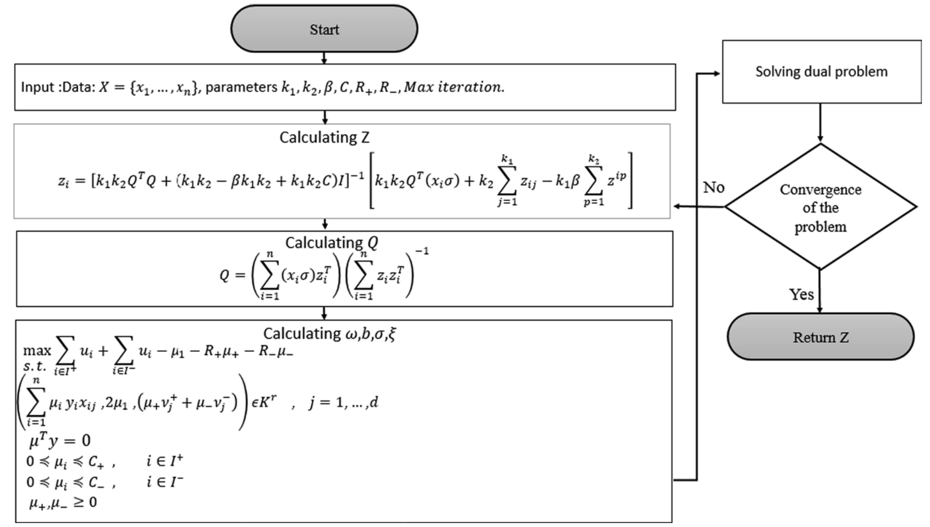

Start
Input :Data: $X = \{x_1, \ldots, x_n\}$, parameters $k_1, k_2, \beta, C, R_+, R_-, Max\ iteration.$
Calculating Z
$z_i = [k_1k_2Q^TQ + (k_1k_2 - \beta k_1k_2 + k_1k_2C)I]^{-1}\left[k_1k_2Q^T(x_i\sigma) + k_2\sum_{j=1}^{k_1} z_{ij} - k_1\beta\sum_{p=1}^{k_2} z^{ip}\right]$
Calculating Q
$Q = \left(\sum_{i=1}^{n}(x_i\sigma)z_i^T\right)\left(\sum_{i=1}^{n} z_iz_i^T\right)^{-1}$
Calculating ω,b,σ,ξ
$\max \sum_{i\in I^+} u_i + \sum_{i\in I^-} u_i - \mu_1 - R_+\mu_+ - R_-\mu_-$
$s.t.$
$\left(\sum_{i=1}^{n} \mu_i\, y_ix_{ij}\ ,2\mu_1\ ,(\mu_+v_j^+ + \mu_-v_j^-)\right)\epsilon K^r\ ,\ j = 1, \ldots, d$
$\mu^Ty = 0$
$0 \preccurlyeq \mu_i \preccurlyeq C_+\ ,\ i \in I^+$
$0 \preccurlyeq \mu_i \preccurlyeq C_-\ ,\ i \in I^-$
$\mu_+,\mu_- \geq 0$
Solving dual problem
Convergence of the problem
No
Yes
Return Z

Fig. 1 The flowchart of the proposed algorithm

$$
\begin{aligned}
&\min_{Q.Z.\omega.b.\sigma.\xi}\ 1^T\xi + \frac{1}{n}\sum_{i=1}^{n}\|x_i\sigma - Qz_i\|^2 + \frac{1}{nk_1}\sum_{i=1}^{n}\sum_{j=1}^{k_1}\left\|z_i - z_{ij}\right\|^2 \\
&- \frac{\beta}{nk_2}\sum_{i=1}^{n}\sum_{p=1}^{k_2}\left\|z_i - z^{ip}\right\|^2 + C\sum_{i=1}^{n}\|z_i\|^2 \\
&s.\ t. \\
&Q^TQ = I \\
&y_i\left(\sum_{j=1}^{m} z_{ij}\omega_j\sigma_j + b\right) \geq 1 - \xi_i \quad and \quad i = 1\ldots n \\
&\omega^T\omega \leq 1 \\
&R_+ \geq \sum_{j=1}^{m} v_j^+\sigma_j^2 \\
&R_- \geq \sum_{j=1}^{m} v_j^-\sigma_j^2 \\
&\xi.\sigma \succcurlyeq 0.
\end{aligned}
\tag{10}
$$

In the above optimization problem, the input data is assumed to be balanced. This means that if there are two classes of data, approximately equal amounts of data are available from each of them. Since this assumption is not always true in real-world datasets, a cost-sensitive function is added to reinforce the above optimization problem, such as [21]. Therefore, the following optimization problem will be resulted with the same constraints in Eq. (8):

$$
\min_{\omega.b.\sigma.\xi} C_+\sum_{i\in I_+}\xi_i + C_-\sum_{j\in I_-}\xi_j. \tag{11}
$$

In problem (11), the slack values for the positive and negative class data are added with different coefficients. Suppose the positive class is minor; therefore, to prevent the model from deviating towards that class, the C_+ coefficient can be selected as a larger number than C_. This means that since the slack of the positive samples will be more fined, the model will not deviate towards them.

The cost-sensitive optimization problem now becomes:

$$\begin{aligned}
&\min_{Q.z.\omega.b.\sigma.\xi} C_+ \sum_{i\in I_+} \xi_i + C_- \sum_{j\in I_-} \xi_j + \frac{1}{n}\sum_{i=1}^{n} \|x_i\sigma - Qz_i\|^2 + \frac{1}{nk_1}\sum_{i=1}^{n}\sum_{j=1}^{k_1} \left\|z_i - z_{ij}\right\|^2 - \frac{\beta}{nk_2}\sum_{i=1}^{n}\sum_{p=1}^{k_2} \left\|z_i - z^{ip}\right\|^2 + C\sum_{i=1}^{n} \|z_i\|^2 \\
&s.\,t. \\
&Q^T Q = I \\
&y_i\left(\sum_{j=1}^{d} x_{ij}\omega_j\sigma_j + b\right) \geq 1 - \xi_i \quad and \quad i = 1\ldots n \\
&\omega^T\omega \leq 1 \\
&R_+ \geq \sum_{j=1}^{d} v_j^+\sigma_j^2 \\
&R_- \geq \sum_{j=1}^{d} v_j^-\sigma_j^2 \\
&\xi.\sigma \succcurlyeq 0.
\end{aligned} \tag{12}$$

Solving the proposed cost-sensitive optimization problem, Eq. (12), leads to the fact that in addition to feature extraction, feature weighting is performed simultaneously. The object of the problem is to find z_i, which is the result of extracting the property on $x_i\sigma$. Also, since the optimization problem is cost sensitive due to the lack of proper classification of positive and negative classes, this leads the feature reduction process to the output properties that are not only suitable for the majority class but also for the minority class.

The proposed problem, causes more separation in the reduced space in two ways. One for the existence of the expression LA and the other for the existence of the constraint $y_i\left(\sum_{j=1}^{d} x_{ij}\omega_j\sigma_{jj} + b\right) \geq 1 - \xi_i$.

Feature extraction as a solution to a reduced manifold learning optimization problem is based on error reduction and maintaining geometric relationships between data. Also, in order to select the features, optimization problems based on the minimization of the above the generalization error have been adopted. Finally, the optimization problem combined from the above two problems is solved by adding a cost-sensitive expression to create a balance without manipulating the data in the imbalanced data. The flowchart of the proposed approach in illustrated in Fig. 1.

The pseudo code

The steps of the optimization algorithm are denoted in Alg. 1.

Algorithm I An iterative solution to optimize the proposed problem

Table 2 Data sets used in experiments

	Data set	Samples	Features	Classes	Imbalance ratio
1	Cancer	699	9	2	1.9
2	Wine	178	13	3	1.4
3	Ionosphere	351	34	2	1.7
4	Pima Indian diabetes	768	8	2	1.8
5	Iris	150	4	3	1
6	Wdbc	569	30	2	1.6
7	Cleveland	303	13	5	12.6
8	Musk	476	166	2	1.2
9	Dermatology-6	366	34	6	5.6
10	FuelCons	1764	37	4	13.08
11	Movement_libras	270	90	15	1
12	Sonar	208	60	2	1.14
13	SPECTF	267	44	2	3.85
14	Colon tumor	62	166	2	1.81
15	DLBCL77	77	5469	2	3.05
16	Mnist	10,000	784	10	5.99
17	Caltech101	8671	784	101	25.74
18	Kddcup-rootkit-imap-vs-back	2225	41	2	100.13
19	Kddcup-buffer-overflow-vs-back	2233	41	2	73.44
20	Kddcup-guess-passwd-vs-satan	1642	41	2	29.98
21	Kddcup-land-vs-satan	1610	41	2	75.66

Table 3 The number of original features and the number of features after dimension reduction

	Name of data set	Main feature difference	Features after dimension reduction
1	Cancer	9	4
2	Wine	13	6
3	Ionosphere	34	17
4	Pima Indian diabetes	8	4
5	Iris	4	2
6	Wdbc	30	15
7	Cleveland	13	6
8	Musk	166	83
9	Dermatology-6	34	17
10	FuelCons	37	18
11	Movement_libras	90	45
12	Sonar	60	30
13	SPECTF	44	22
14	Colon tumor	166	83
15	DLBCL	5469	2734
16	Mnist	784	392
17	Caltech101	784	392
18	Kddcup-rootkit-imap-vs-back	41	20
19	Kddcup-buffer-overflow-vs-back	41	20
20	Kddcup-guess-passwd-vs-satan	41	20
21	Kddcup-land-vs-satan	41	20

Table 4 The best values of the parameters obtained by the PSO evolutionary algorithm

Hyper parameters	Value	Definition
R_+	20	It is evident that by choosing a low value for R_+, the optimum solution for σ would include more zero values and consequently it results in a higher feature reduction rate
R_-	50	It is evident that by choosing a low value for R_-, the optimum solution for σ would include more zero values and consequently it results in a higher feature reduction rate
C	0.1	The non-negative parameter *C* indicated the penalty for the regularization term
β	0.0001	The parameter β indicates the importance level of the second term in the equation
k_1	3	The number of neighbors are considered
k_2	3	The number of neighbors are considered
max iteration	10	Repeat of optimization steps

```
Input: data matrix $X = \{x_1 \ldots x_n\}$ where $x_i \in \mathbb{R}^d$ $i = 1 \ldots n$ and labels $y_i \in \{-1. + 1\}$ Output: dimensionality reduced data $z_i\ \forall i = 1 \ldots n, \omega.$ and $b.$

1: Initialize $Q$ and $\sigma$ with random values.
2: For $t = 1 \ldots max\_iteration$ do
3: Solve the problem associated with step 1 to oain $z_i$ at the $(t + 1)$th iteration by:
       $z_i^{(t+1)} = \left[k_1 k_2 Q^{(t)T} Q^{(t)} + (k_1 k_2 - \beta k_1 k_2 + k_1 k_2 C) I\right]^{-1} \left[k_1 k_2 Q^{(t)T}\left(x_i \sigma^{(t)}\right) + k_2 \sum_{j=1}^{k_1} \left(z_{ij}\right)^{(t)} - k_1 \beta \sum_{p=1}^{k_2} \left(z^{ip}\right)^{(t)}\right],$
4: Solve the problem associated with step 2 to obtain $Q$ by:
       $Q^{(t+1)} = \left(\sum_{i=1}^{n} \left(x \sigma^{(t)}\right) z_i^{(t+1)T}\right) \left(\sum_{i=1}^{n} z_i^{(t+1)} z_i^{(t+1)T}\right)^{-1},$
5: Solve the dual problem (12) using $z_i^{(t+1)}, Q^{(t+1)}$
6: Exit, if convergence criterion meets.
7: end for
```

As it may be inferred from the above algorithm, there is a loop in line 2 which is iterated max_iteration times. In the loop, z_i is first computed which as seen, has a constant time with respect to the number of samples and features. Therefore, calculating the whole Z has a time complexity of O(n). Calculating Q is the most complex stage of the approach. The first parenthesis on Q is calculated in O(nm) while the second parenthesis has the same complexity. Therefore, the complexity of finding Q is O(2nm). For solving Eq. (12), we have six summations which are computed over n. Therefore, considering approximately constant operations in each summation, the computational complexity of the last stage in O(cn).

Table 5 The accuracy of the proposed cost-sensitive method compared to other methods

Dataset	**Classification Accuracy (%**						
	Baseline SVM	**S-MVML-LA**	**GMEB**	**WJMI**	**IGIS**	**IWFS**	**Proposed**
Iris	77.33	83.11 (4)	96.44 (2)	NA	88 (3)	NA	**97.33** (1)
Cancer	95.13	96.71 (3)	95.42 (5)	**100 (1)**	94.56 (6)	**100 (1)**	96.44 (4)
Ionosphere	64.67	76.92 (4)	85.76 (2)	NA	76.91 (3)	NA	**86.30 (1)**
Wine	59.91	88.33 (2)	97.37 (4)	NA	93.31 (3)	NA	**98.86 (1)**
Pid	64.67	64.97 (4)	76.82 (2)	NA	75.53 (3)	NA	**77.07** (1)
Wdbc	85.93	89.80 (4)	94.89 (3)	NA	95.78 (2)	NA	**97.01** (1)
Musk	83.61	85.50 (3)	80.87 (6)	84.99 (4)	82.56 (5)	85.79 (2)	**87.81 (1)**
Dermatology-6	94.41	96.53 (3)	99.44 (2)	88.87 (5)	91.57 (4)	87.28 (6)	**99.45** (1)
FuelCons	86.45	89.54 (3)	89.73 (2)	83.06 (4)	78.12 (6)	82.59 (5)	90.73 (1)
Movement_libras	91.30	91.90 (3)	97.13 (2)	68.31 (6)	80.62 (5)	81.79 (4)	97.18 (1)
Sonar	81.21	79.80 (4)	**86.09 (1)**	76.8 (5)	75.41 (6)	81.6 (3)	**86.09 (1)**
SPECTF	79.42	79.35 (3)	79.35 (3)	78.93 (5)	79.40 (2)	**80.4 (1)**	71.51 (6)
Mnist	98.51	95.97 (3)	98.36 (2)	75.17 (5)	NA	79.35 (4)	**98.58 (1)**
Colon	83.84	83.97 (3)	69.48 (5)	97.25 (2)	NA	**98.26 (1)**	83.97 (3)
Caltch101	99.26	**99.19 (1)**	99.14 (3)	37.91 (4)	NA	32.86 (5)	**99.19 (1)**
DLBCL77	88.23	75.33 (5)	89.41 (4)	**100 (1)**	NA	**100 (1)**	97.41 (3)
Kddcup-rootkit-imap-vs-back	99.01	98.88 (3)	**100 (1)**	NA	NA	NA	**100 (1)**
Kddcup-buffer-overflow-vs-back	98.65	97.93 (3)	99.81 (2)	NA	NA	NA	**100 (1)**
Kddcup-guess-passwd-vs-satan	99.87	98.78 (3)	**99.95 (1)**	NA	NA	NA	**99.95 (1)**
Kddcup-land-vs-satan	99.19	99.31 (3)	**100 (1)**	NA	NA	NA	**100 (1)**
Cleveland	81.64	81.65 (3)	81.91 (2)	NA	NA	NA	82.96 (1)
Average rank		3.23	2.73	3.90	4.36	3.09	1.71

The numbers in bold are the best accuracies achieved on the dataset

Having all these complexities for max_iteration times, the total time complexity of the algorithm will be O(max_iteration*(n + 2 nm + cn)) in which n is the number of samples and m is the number of final extracted features. Therefore, as seen, the time complexity of the algorithm is a linear function of n and m which is not much as compared to similar approaches and implementing the approach in parallel will decrease the run time of the algorithm.

Experiments

In this section, we simulate the proposed method and evaluate the performance of the proposed algorithm. This test needs criteria to be used to measure the performance of the algorithm. Here, after stating the conditions and implementation environment, a description of the evaluated data set, setting the algorithm parameters, the desired criteria are introduced, and then, the efficiency of the proposed method in terms of these different criteria is compared with other methods.

Table 6 The F-score of the proposed cost-sensitive method compared to other methods

Dataset	**F-score (%)**						
	Baseline SVM	**S-MVML-LA**	**GMEB**	**WJMI**	**IGIS**	**IWFS**	**Proposed**
Iris	66.16	74.23 (4)	94.98 (2)	NA	88.30 (3)	NA	**95.60 (1)**
Cancer	94.50	96.28 (3)	95.19 (5)	**100 (1)**	93.96 (6)	**100 (1)**	96.02 (4)
Ionosphere	40.81	68.14 (4)	82.37 (2)	NA	74.53 (3)	NA	**83.97 (1)**
Wine	18.98	82.74 (4)	95.54 (2)	NA	93.57 (3)	NA	**98.31 (1)**
Pid	40.81	39.37 (4)	72.18 (2)	NA	71.7 (3)	NA	**72.57** (1)
Wdbc	82.99	88.34 (4)	94.42 (3)	NA	95.51 (2)	NA	**96.73 (1)**
Musk	83.11	84.91 (4)	80.21 (5)	85.85 (3)	82.34 (6)	87.06 (2)	**87.86 (1)**
Dermatology-6	48.56	82.95 (6)	96.33 (2)	87.51 (4)	89.41 (3)	86.2 (5)	**97.90 (1)**
FuelCons	46.05	54.73 (6)	67.21 (4)	**77.49 (1)**	66.68 (5)	76.92 (2)	72.51 (3)
Movement_libras	33.58	41.08 (6)	79.76 (3)	67.99 (5)	**81.80 (1)**	80.71 (2)	76.23 (4)
Sonar	80.38	78.81 (4)	**85.85 (1)**	77.77 (5)	75.28 (6)	82.61 (3)	85.76 (2)
SPECTF	44.16	44.17 (5)	44.17 (5)	62.39 (3)	44.26 (4)	67.39 (2)	**66.36 (1)**
Mnist	91.01	76.94 (3)	90.25 (2)	70.28 (5)	NA	75.85 (4)	**91.62 (1)**
Colon	82.13	82.14 (4)	65.95 (5)	96.18 (2)	NA	**98.1 (1)**	82.72 (3)
Caltch101	52.13	**48.59 (1)**	45.26 (3)	22.27 (5)	NA	17.75 (4)	48.43 (2)
DLBCL77	77.27	42.78 (5)	80.48 (4)	**100 (1)**	NA	**100 (1)**	96.73 (3)
Kddcup-rootkit-imap-vs-back	49.75	69.21 (3)	**98.87 (1)**	NA	NA	NA	96.65 (2)
Kddcup-buffer-overflow-vs-back	49.66	79.13 (3)	**100 (1)**	NA	NA	NA	**100 (1)**
Kddcup-guess-passwd-vs-satan	99.06	85.12 (3)	98.27 (2)	NA	NA	NA	**100 (1)**
Kddcup-land-vs-satan	78.03	86.04 (3)	**100 (1)**	NA	NA	NA	**100 (1)**
Cleveland	14.01	13.93 (3)	23.89 (2)	NA	NA	NA	**29.51 (1)**
Average rank		3.88	2.78	3.27	3.66	2.68	1.76

The numbers in bold are the best accuracies achieved on the dataset

Experimental setup

In order to evaluate the effectiveness of the proposed method, the data collections from the UCI machine learning repository, high-dimensional microarrays datasets as well as imbalanced datasets from the KEEL repository have been used.

Attempts are made to use data that is large in size so that a number of appropriate features are reached when the dimension reduction is performed. Also, the use of data with more imbalance in their labels will lead to a better evaluation of the performance of the proposed method in the face of imbalanced data. Specifications of the test classification datasets are mentioned in Table 2.

The size of the datasets in this article varies from 62 to 10,000, the number of their features varies from 4 to 7129, and the number of classes from 2 to 101. The imbalance ratio in Table 2 is calculated by the following equation [45]:

$$Imbalance\ ratio = \frac{size\ of\ \ majority\ class}{size\ of\ \ minority\ class}. \tag{13}$$

Also in Table 3, the number of selected features for each data set is specified. As a heuristic, the integer value of half the number of attributes in each data set is used as the number of selected features.

Table 7 The proposed cost-sensitive method with other methods in term of RCL

Dataset	RCL (%)		
	S-MVML-LA	GMEB	Proposed cost sensitive
Iris	84.40 (3)	95.14 (2)	**96.52 (1)**
Cancer	**96.28 (1)**	94.90 (3)	96.02 (2)
Ionosphere	68.14 (3)	82.37 (2)	**83.97 (1)**
Wine	90.04 (3)	96.49 (2)	**98.55 (1)**
Pid	32.52 (3)	71.01 (2)	**76.55** (1)
Wdbc	92.54 (3)	92.55 (2)	**96.00 (1)**
Musk	85.75 (2)	81.19 (3)	**87.69 (1)**
Dermatology-6	**98.15 (1)**	**98.15 (1)**	89.85 (3)
FuelCons	54.29 (3)	77.67 (2)	**80.88** (1)
Movement_libras	44.70 (3)	**79.72(1)**	79.04 (2)
Sonar	82.35 (3)	86.99 (2)	**85.41 (1)**
SPECTF	39.69 (2)	39.69 (2)	**67.38 (1)**
Mnist	79.28 (3)	91.00 (2)	**91.84 (1)**
Colon	82.63 (2)	71.71 (3)	**83.21 (1)**
Caltch101	**56.67 (1)**	50.14 (3)	52.69 (2)
DLBCL77	37.66 (3)	78.94 (2)	**98.32(1)**
Kddcup-rootkit-imap-vs-back	71.92 (3)	**99.97 (1)**	**99.97 (1)**
Kddcup-buffer-overflow-vs-back	82.03 (3)	**100 (1)**	**100 (1)**
Kddcup-guess-passwd-vs-satan	83.68 (3)	96.83 (2)	**100 (1)**
Kddcup-land-vs-satan	80.93 (3)	**100 (1)**	**100 (1)**
Cleveland	10.82 (3)	22.34 (2)	30.63 (1)
Average rank	2.61	2.04	1.30

The numbers in bold are the best accuracies achieved on the dataset

Evaluation method

To evaluate the results of the approach, the multi-class linear SVM classifier is used on the reduced data. K-fold cross validation method is used to perform the experiments. MATLAB libsvm library is used for SVM implementation and evaluation and linear kernel with $C=1$ is used as the SVM settings. The one-vs-one model is used to classify the proposed method for multiclass data.

Parameter settings

The best values of the parameters are obtained by the Particle Swarm Optimization (PSO) algorithm on the validation data set. Given that there are different parameters in the proposed method, the parameter values β is set as 0.0001, C is 0.1 and the number of neighbors are considered as 3, and also the values of the super-parameters of the problem is set as 20 and 50 in the experiments. The values for all hyper-parameters used in the proposed method are presented in Table 4.

Table 8 The proposed cost-sensitive method with other methods in term of PRC

Dataset	**PRC (%)**		
	S-MVML-LA	**GMEB**	**Method proposed**
Iris	79.86 (3)	**95.61** (1)	95.14 (2)
Cancer	**96.23 (1)**	94.83 (3)	95.86 (2)
Ionosphere	67.77 (3)	80.13 (2)	**82.31 (1)**
1Wine	81.90 (3)	96.49 (2)	**98.19 (1)**
Pid	49.89 (3)	71.01 (2)	**71.42 (1)**
Wdbc	86.69 (3)	93.46 (2)	**96.20 (1)**
Musk	84.55 (2)	80.06 (3)	**87.95 (1)**
Dermatology-6	70 (3)	**95 (1)**	88.33 (2)
FuelCons	56.12 (3)	65.67 (2)	**68.82 (1)**
Movement_libras	47.49 (3)	76.60(2)	**79.40 (1)**
Sonar	78.90(2)	78.90 (2)	**84.67 (1)**
SPECTF	50(2.5)	50 (2)	**75.93 (1)**
Mnist	76.15 (3)	89.74 (2)	**91.55 (1)**
Colon	84.29 (2)	**84.50 (1)**	84.29 (2)
Caltch101	46.84 (2)	45.26 (3)	**48.25 (1)**
DLBCL77	50 (3)	82.70 (2)	**95.50(1)**
Kddcup-rootkit-imap-vs-back	68.01(3)	**98 (1)**	95 (2)
Kddcup-buffer-overflow-vs-back	79.27 (3)	**100 (1)**	**100 (1)**
Kddcup-guess-passwd-vs-satan	90.65 (3)	99.90 (2)	**100 (1)**
Kddcup-land-vs-satan	80.93 (3)	**100 (1)**	**100 (1)**
Cleveland	20 (3)	28.65 (2)	31.79 (1)
Average Rank	2.69	1.95	1.30

The numbers in bold are the best accuracies achieved on the dataset

Experimental design

MATLAB 2018 software has been used to implement the proposed algorithm. The tests were also evaluated on a PC with Intel Core i3 processor, 4 gigabytes of RAM, and Windows 8.1 operating system.

Evaluation criteria

To compare the performance of the proposed method F-score and accuracy criteria are used.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{14}$$

$$F{-}score = \frac{2TP}{2TP + FN + FP}, \tag{15}$$

where TP denotes true positive, TN is true negatives, FP is false positives, and FN denotes false negatives.

Two other criterions are used in the experiments which are:

Table 9 The proposed cost-sensitive method with other methods in term of G.Mean

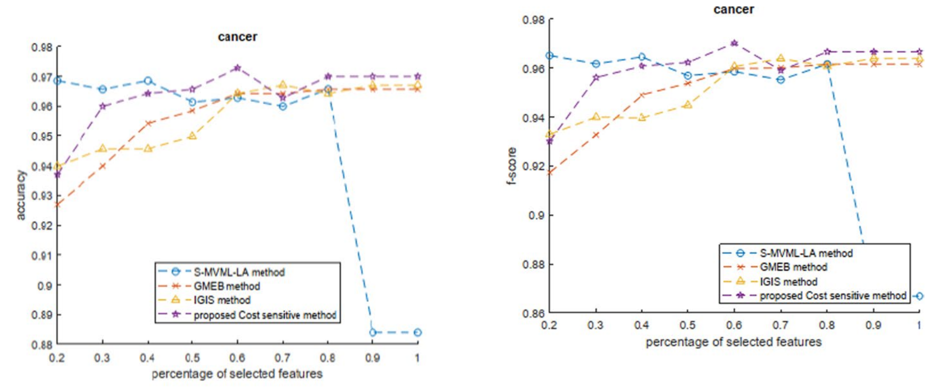
Dataset	**g-mean (%)**		
	S-MVML-LA	**GMEB**	**Method proposed**
Iris	85.49 (3)	96.05 (2)	**97.30 (1)**
Cancer	**96.33** (1)	94.98 (3)	96.22 (2)
Ionosphere	85.79 (3)	**90.50 (1)**	87.88 (2)
Wine	90.62 (3)	97.16(2)	**98.85 (1)**
Pid	0 (3)	75.74(2)	**76.41 (1)**
Wdbc	**99.26 (1)**	95.92 (3)	97.51 (2)
Musk	85.71 (2)	81.06(3)	**87.54 (1)**
Dermatology-6	79.85 (3)	98.13 (2)	**99.70 (1)**
FuelCons	60.23 (3)	84.53 (2)	**86.50 (1)**
Movement_libras	53.70 (3)	85.06(2)	**85.43 (1)**
Sonar	78.90(3)	78.90 (2)	**84.67 (1)**
SPECTF	50(2.5)	50 (2.5)	**75.93 (1)**
Mnist	87.29 (3)	94.94 (2)	**95.42 (1)**
Colon	81.63(2)	67.88 (3)	**82.55 (1)**
Caltch101	**70.12 (1)**	66.09 (3)	68.26 (2)
DLBCL77	50 (3)	82.70 (2)	**95.50(1)**
Kddcup-rootkit-imap-vs-back	57.13(3)	**99.97 (1)**	**99.97 (1)**
Kddcup-buffer-overflow-vs-back	79.22 (3)	**100 (1)**	**100 (1)**
Kddcup-guess-passwd-vs-satan	81.15 (3)	96.73 (2)	**100** (1)
Kddcup-land-vs-satan	77.74 (3)	**100 (1)**	**100 (1)**
Cleveland	0 (3)	32.97 (2)	**40.51 (1)**
Average rank	2.59	2.09	1.23

The numbers in bold are the best accuracies achieved on the dataset

Table 10 Average runtime of the proposed method compared with other methods (in seconds)

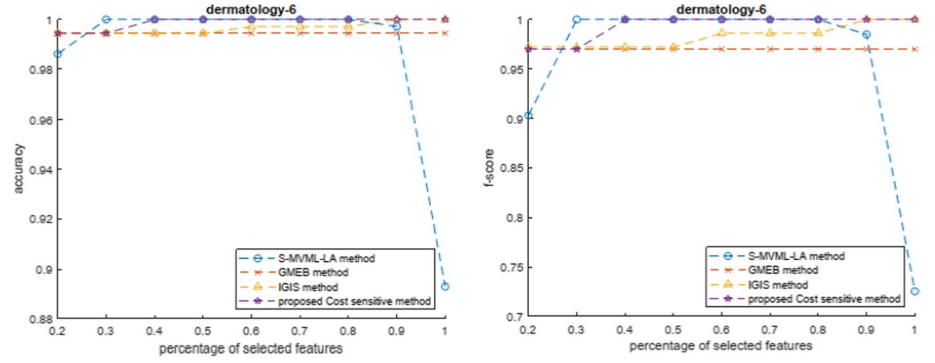
Dataset	**S-MVML-LA**	**GMEB**	**IGIS**	**Proposed**
Iris	0.8	**0.6**	3.2	1.2
Cancer	1.8	**0.4**	1.3	4.2
Ionosphere	0.9	**0.8**	4.0	6.4
Wine	0.7	**0.5**	2.2	2.7
Pid	1.1	**0.3**	1.6	3.7
Wdbc	1.6	**0.6**	3.3	8.2
Musk	3.1	**1.8**	37.2	288.7
Dermatology-6	0.8	**0.5**	3.3	6.3
FuelCons	3.4	**3.2**	15.5	84.9
Movement_libras	**6.2**	14.5	319.0	219.6
Sonar	0.8	**0.4**	7.5	10.6
SPECTF	0.8	**0.4**	4.1	10.9
Mnist	1751.7	1767.2	168,000	**1461.1**
Colon	19.3	3.0	2310.8	204.6
Caltch101	2247.0	9794.4	840,000	**94.0**
DLBCL77	28.4	23.4	1958.9	**20.3**
Kddcup-rootkit-imap-vs-back	11.0	**1.3**	14.0	85.3
Kddcup-buffer-overflow-vs-back	2.7	**0.8**	7.9	87.7
Kddcup-guess-passwd-vs-satan	2.8	**0.7**	7.0	59.9
Kddcup-land-vs-satan	4.5	0.7	6.9	48.3
Cleveland	1.7	0.9	5.4	7.2

The numbers in bold are the best accuracies achieved on the dataset



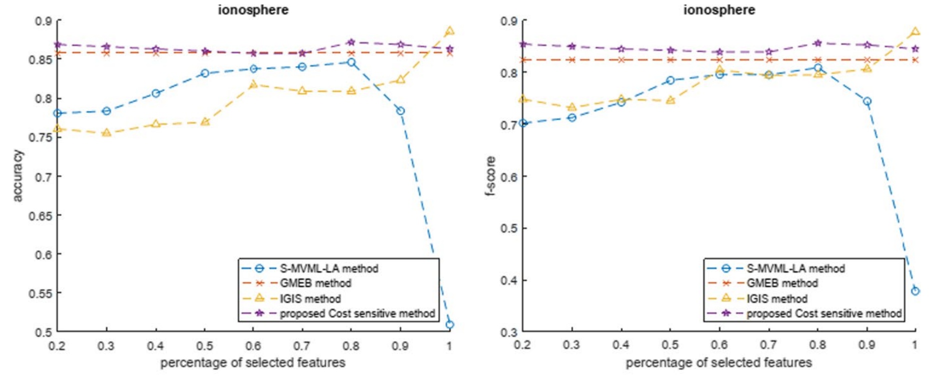
(a) (b)

Fig. 2 The performance of the proposed cost sensitive method, S-MVML-LA [15], GMEB [10] and IGIS [20] versus the changes in percentage of selected features numbers over the cancer dataset. **a** Accuracy, **b** f-score



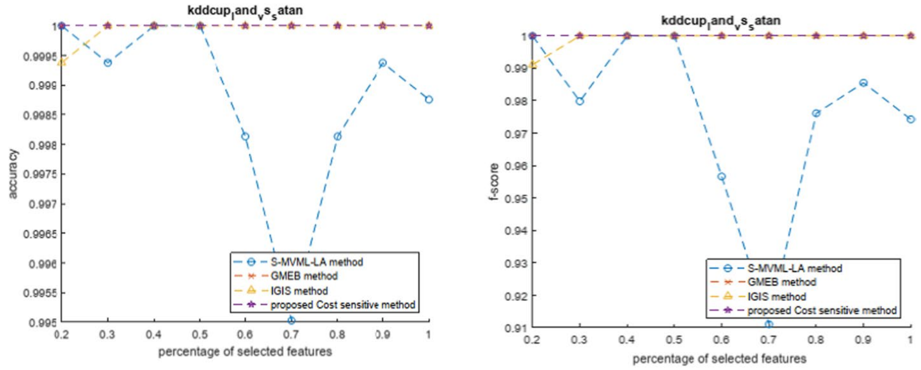
(a) (b)

Fig. 3 The performance of the proposed cost sensitive method, S-MVML-LA [15], GMEB [10] and IGIS [20] versus the changes in percentage of selected features numbers on the deramatology-6 dataset. **a** Accuracy, **b** f-score



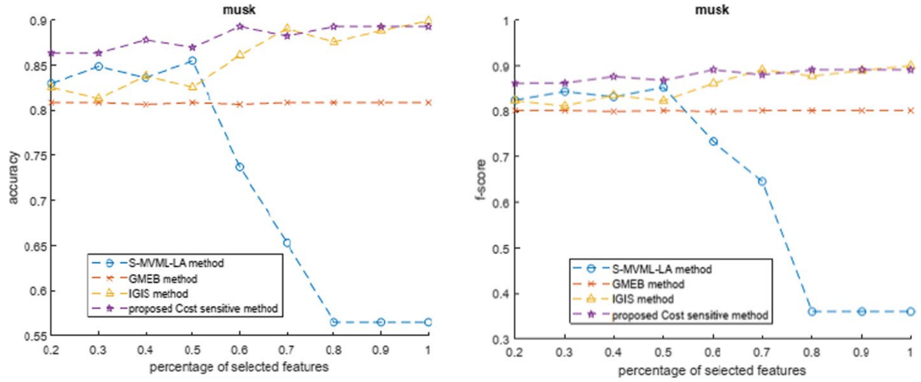
(a) (b)

Fig. 4 The performance of the proposed cost sensitive method, S-MVML-LA [15], GMEB [10] and IGIS [20] versus the changes in percentage of selected features numbers on the Ionosphere dataset. **a** Accuracy, **b** f-score



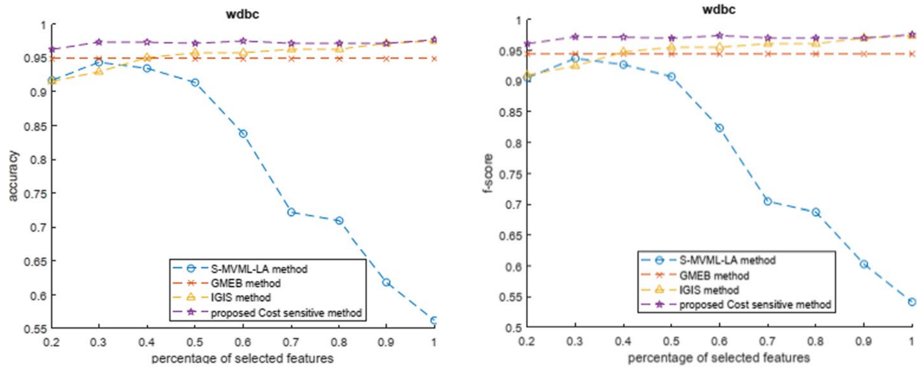
(a) (b)

Fig. 5 The performance of the proposed cost sensitive method, S-MVML-LA [15], GMEB [10] and IGIS [20] versus the changes in percentage of selected features numbers on the kddcup_land_vs_satan dataset. **a** Accuracy, **b** f-score



(a) (b)

Fig. 6 The performance of the proposed cost sensitive method, S-MVML-LA [15], GMEB [10] and IGIS [20] versus the changes in percentage of selected features numbers on the musk dataset. **a** Accuracy, **b** f-score



(a) (b)

Fig. 7 The performance of the proposed cost sensitive method, S-MVML-LA [15], GMEB [10] and IGIS [20] versus the changes in percentage of selected features numbers on the WDBC dataset. **a** Accuracy, **b** f-score

$$RCL = \frac{TP}{TP + FN}, \tag{16}$$

$$PRC = \frac{TP}{TP + FP}. \tag{17}$$

Due to the imbalance in the data set, one of the RCL and PRC criteria may be too high or too low, so their geometric mean is considered in our experiments.

$$G_{mean} = \sqrt{(RCL * PRC)}. \tag{18}$$

Feature reduction and classification methods

Results of classification operations after feature reduction were studies using the proposed cost-sensitive method compared with S-MVML-LA [15], GMEB [10], IGIS [20], WJMI [46] And IWFS [21] methods.

As in Table 5, the proposed method has performed better than the other methods in most cases in term of the accuracy. The numbers in parenthesis denote the rank of the approach on each dataset and the final row denote the average rank of each method. As seen in the table, the average rank of the proposed method has the highest value among other methods. After the proposed method, the GMEB method was able to obtain a better ranking, and finally the IGIS method is lower than the others. Not available (NA) values in the following tables, is due to the fact that the corresponding datasets are not evaluated in the reference articles.

In Table 6, the proposed cost-sensitive method has the highest average ranking. The IWFS method average rank is 0.92 worse than the proposed method and is ranked second in total.

Again in Table 7, the proposed method has a better result than other methods. The proposed method has the highest average rank value. In the three datasets Kddcup-buffer-overflow-vs-back, Kddcup-land-vs-satan, Kddcup-rootkit-imap-vs-back, the proposed method had similar performance with GMEB.

As shown in Table 8, the performance of the proposed method is better in most cases. Comparing the proposed cost-sensitive method with other methods, the GMEB approach has superior performance on iris dataset, Kddcup-rootkit-imap-vs-back and Dermatology-6 while the S-MVML-LA approach cannot compete with the two other approaches. However, on several data sets, the proposed method has a high performance and the average rank of the method has the highest value.

In Table 9, the methods are evaluated with the g-mean criterion. This criterion is one of the most accurate evaluation criteria for imbalanced data sets. As seen in the results, the cost-sensitive method performed better on most datasets and the proposed approach has again the highest average rank.

In Table 10, the average execution time of our method compared to other approaches is shown. As seen and could be predicted, the proposed approach has a relatively high execution cost on most of the datasets as compared to other evaluated approaches. However, the approach is still better than the IGIS method on Movement_libras, Colon, and Iris datasets. Also, on the Caltch101 (with 784 original features), Mnist (with 5469 original features) and DLBCL77 (with 784 original features), which are among the highest dimensional evaluation datasets, our method has the best convergence time than all other methods. These observations show that when it comes to high dimensional data, the other compared approaches degrade and our algorithm is more effective than them. But with lower dimensional data, the proposed approach has higher running time which is not considerable when the whole execution time is in the scale of 1 or 2 s. As shown in Table 10, GMEB approach has the lowest execution time on 14 datasets most of which has the original dimensionality of 100 or less. Therefore, these evaluations also demonstrate the effectiveness of the approach in high dimensionality.

To better demonstrate the results of the experiments, the performance of the approaches versus the percentage of the selected features are depicted in Figs. 2, 3, 4, 5, 6 and 7.

On the deramatology-6 dataset, the proposed cost-sensitive method has performed better than other methods from the beginning. It still performs well when the number of features increases. With this interpretation, it can be concluded that our method has selected relatively good features that show a better result. Also, on the ionosphere dataset, the cost-sensitive method initially performed better than other methods with less specificity. In the case of the IGIS method, it performs badly at the beginning (i.e. when the number of selected features is low) and then gets better. This means that when a feature increases, it works well. It can be concluded that the features it has selected must not have been good features. On the other hand, if 20% of the features are selected using the proposed approach, the performance is well and stable. With the S-MVML-LA approach, the diagram slope has gradually increased, and then suddenly drops strangely. The performance contour of the proposed method is above the others from the beginning, and maintains the performance which shows the superiority of the approach.

Conclusion

In this paper, a hybrid was proposed in order to reduce the data dimensionality, which combines feature selection and feature extraction in the context of an optimization problem solving while creating a balance without manipulating the data. In this method, it uses the advantages of feature selection and feature extraction together. In feature extraction, it tries to solve a Manifold learning optimization problem and does feature selection as an optimization problem based on minimization of the general error boundary. In evaluations the accuracy and f-score results are reported on the test data. Comparison results of the proposed method with other methods on 21 datasets from the UCI machine learning repository, microarrays and high-dimensional datasets as well as imbalanced datasets from KEEL repository are reported. The evaluations indicate the superiority of the proposed model over other methods. As the future works, evaluating the proposed approach on real world problems and applications is suggested.

Acknowledgements

All the contributors are mentioned in the manuscript.

Author contributions

MF performed the main implementations and evaluations. MHM and YF were thesis supervisor and advisor respectively and the idea was originated by them. All authors read and approved the final manuscript.

Funding

The authors approve that there was no funding in preparing the manuscript.

Availability of data and materials

The datasets and codes will be available after acceptance.

Declarations

Ethics approval and consent to participate

This manuscript does not include studies involving human participants.

Consent for publication

This manuscript does not contain any individual person's data.

Competing interests

The authors approve that there is no conflict of interest in the manuscript.

Received: 22 November 2021 Accepted: 11 April 2022
Published online: 03 May 2022

References

1. Rakkeitwinai S, et al. New feature selection for gene expression classification based on degree of class overlap in principal dimensions. Comput Biol Med. 2015;64:292–8.
2. Kabir MM, Shahjahan M, Murase K. A new local search based hybrid genetic algorithm for feature selection. Neurocomputing. 2011;74(17):2914–28.
3. Vieira SM, Sousa JM, Runkler TA. Two cooperative ant colonies for feature selection using fuzzy models. Expert Syst Appl. 2010;37(4):2714–23.
4. Zebari R, et al. A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction. J Appl Sci Technol Trends. 2020;1(2):56–70.
5. Cheng Z, Lu Z. A novel efficient feature dimensionality reduction method and its application in engineering. Complexity. 2018. https://doi.org/10.1155/2018/2879640.
6. Zebari DA, et al. A simultaneous approach for compression and encryption techniques using deoxyribonucleic acid. In: 2019 13th international conference on software, knowledge, information management and applications (SKIMA). IEEE; 2019.
7. Ayesha S, Hanif MK, Talib R. Overview and comparative study of dimensionality reduction techniques for high dimensional data. Inf Fusion. 2020;59:44–58.
8. Abd-Alsabour N. On the role of dimensionality reduction. J Comput. 2018;13(5):571–9.
9. Verleysen M, François D. The curse of dimensionality in data mining and time series prediction. In: International work-conference on artificial neural networks. Springer; 2005.
10. Peleg D, Meir R. A feature selection algorithm based on the global minimization of a generalization error bound. In: Advances in neural information processing systems. 2005.
11. Elhadad MK, Badran KM, Salama GI. A novel approach for ontology-based dimensionality reduction for web text document classification. Int J Softw Innov. 2017;5(4):44–58.
12. Luo W. Face recognition based on laplacian eigenmaps. In: 2011 International conference on computer science and service system (CSSS). IEEE; 2011.
13. Abdullah A, et al. Sketching, embedding and dimensionality reduction in information theoretic spaces. In: Artificial intelligence and statistics. PMLR; 2016.
14. Wang Y, Li T. Local feature selection based on artificial immune system for classification. Appl Soft Comput. 2020;87: 105989.
15. Zhao Y, et al. Multi-view manifold learning with locality alignment. Pattern Recogn. 2018;78:154–66.
16. Xu J, et al. Feature selection based on sparse imputation. In: The 2012 international joint conference on neural networks (IJCNN). IEEE; 2012.
17. Shahee SA, Ananthakumar U. An effective distance based feature selection approach for imbalanced data. Appl Intell. 2020;50(3):717–45.
18. Chenxi H, et al. Sample imbalance disease classification model based on association rule feature selection. Pattern Recognit Lett. 2020;133:280–6.
19. Bennin KE, et al. Mahakil: diversity based oversampling approach to alleviate the class imbalance issue in software defect prediction. IEEE Trans Softw Eng. 2017;44(6):534–50.
20. Nakariyakul S. High-dimensional hybrid feature selection using interaction information-guided search. Knowl Based Syst. 2018;145:59–66.
21. Zeng Z, et al. A novel feature selection method considering feature interaction. Pattern Recogn. 2015;48(8):2656–66.
22. Qi X, et al. WJMI: a new feature selection algorithm based on weighted joint mutual information. In: 2015 3rd international conference on mechatronics and industrial informatics (ICMII 2015). Atlantis Press; 2015.
23. Japkowicz N. The class imbalance problem: significance and strategies. In: Proc. of the Int'l Conf. on artificial intelligence. 2000. Citeseer.
24. Hart P. The condensed nearest neighbor rule (corresp.). IEEE Trans Inf Theory. 1968;14(3):515–6.
25. Tomek I. Two modifications of CNN. IEEE Trans Syst Man Cybern. 1976;6:769–72.

26. Yen S-J, Lee Y-S. Cluster-based under-sampling approaches for imbalanced data distributions. Expert Syst Appl. 2009;36(3):5718–27.
27. García S, Herrera F. Evolutionary undersampling for classification with imbalanced datasets: proposals and taxonomy. Evol Comput. 2009;17(3):275–306.
28. Chawla NV, et al. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res. 2002;16:321–57.
29. Han H, Wang W-Y, Mao B-H. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In: International conference on intelligent computing. Springer; 2005.
30. Maciejewski T, Stefanowski J. Local neighbourhood extension of SMOTE for mining imbalanced data. In: 2011 IEEE symposium on computational intelligence and data mining (CIDM). IEEE; 2011.
31. Bunkhumpornpat C, Sinapiromsaran K, Lursinsap C. Safe-level-smote: safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem. In: Pacific-Asia conference on knowledge discovery and data mining. Springer; 2009.
32. Ramentol E, et al. SMOTE-RS B*: a hybrid preprocessing approach based on oversampling and undersampling for high imbalanced data-sets using SMOTE and rough sets theory. Knowl Inf Syst. 2012;33(2):245–65.
33. Cheng F, et al. Large cost-sensitive margin distribution machine for imbalanced data classification. Neurocomputing. 2017;224:45–57.
34. Xiao W, et al. Class-specific cost regulation extreme learning machine for imbalanced classification. Neurocomputing. 2017;261:70–82.
35. Du G, et al. Joint imbalanced classification and feature selection for hospital readmissions. Knowl Based Syst. 2020;200: 106020.
36. Raghuwanshi BS, Shukla S. SMOTE based class-specific extreme learning machine for imbalanced learning. Knowl Based Syst. 2020;187: 104814.
37. Yuan H, et al. Low-rank matrix regression for image feature extraction and feature selection. Inf Sci. 2020;522:214–26.
38. Buvana M, Muthumayil K, Jayasankar T. Content-based image retrieval based on hybrid feature extraction and feature selection technique pigeon inspired based optimization. Ann Roman Soc Cell Biol. 2021;25:424–43.
39. Wang Q. A hybrid sampling SVM approach to imbalanced data classification. In: Abstract and applied analysis. 2014. Hindawi.
40. Prachuabsupakij W. CLUS: a new hybrid sampling classification for imbalanced data. In: 2015 12th international joint conference on computer science and software engineering (JCSSE). IEEE; 2015.
41. Maldonado S, López J. Dealing with high-dimensional class-imbalanced datasets: embedded feature selection for SVM classification. Appl Soft Comput. 2018;67:94–105.
42. Roccetti M, et al. An alternative approach to dimension reduction for pareto distributed data: a case study. J Big Data. 2021;8(1):1–23.
43. Thudumu S, et al. A comprehensive survey of anomaly detection techniques for high dimensional big data. J Big Data. 2020;7(1):1–30.
44. Badaoui F, et al. Dimensionality reduction and class prediction algorithm with application to microarray Big Data. J Big Data. 2017;4(1):1–11.
45. Amin A, et al. Comparing oversampling techniques to handle the class imbalance problem: a customer churn prediction case study. IEEE Access. 2016;4:7940–57.
46. Qi X, et al. WJMI: a new feature selection algorithm based on weighted joint mutual information. 2015.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.