Neural Computing and Applications (2018) 30:987–1013
https://doi.org/10.1007/s00521-018-3471-8

ORIGINAL ARTICLE

Similarity-based attribute weighting methods via clustering algorithms in the classification of imbalanced medical datasets

Kemal Polat[1]

Received: 28 November 2017 / Accepted: 27 March 2018 / Published online: 3 April 2018

Abstract

In the fields of pattern recognition and machine learning, the use of data preprocessing algorithms has been increasing in recent years to achieve high classification performance. In particular, it has become inevitable to use the data preprocessing method prior to classification algorithms in classifying medical datasets with the nonlinear and imbalanced data distribution. In this study, a new data preprocessing method has been proposed for the classification of Parkinson, hepatitis, Pima Indians, single proton emission computed tomography (SPECT) heart, and thoracic surgery medical datasets with the nonlinear and imbalanced data distribution. These datasets were taken from UCI machine learning repository. The proposed data preprocessing method consists of three steps. In the first step, the cluster centers of each attribute were calculated using *k*-means, fuzzy *c*-means, and mean shift clustering algorithms in medical datasets including Parkinson, hepatitis, Pima Indians, SPECT heart, and thoracic surgery medical datasets. In the second step, the absolute differences between the data in each attribute and the cluster centers are calculated, and then, the average of these differences is calculated for each attribute. In the final step, the weighting coefficients are calculated by dividing the mean value of the difference to the cluster centers, and then, weighting is performed by multiplying the obtained weight coefficients by the attribute values in the dataset. Three different attribute weighting methods have been proposed: (1) similarity-based attribute weighting in *k*-means clustering, (2) similarity-based attribute weighting in fuzzy *c*-means clustering, and (3) similarity-based attribute weighting in mean shift clustering. In this paper, we aimed to aggregate the data in each class together with the proposed attribute weighting methods and to reduce the variance value within the class. Thus, by reducing the value of variance in each class, we have put together the data in each class and at the same time, we have further increased the discrimination between the classes. To compare with other methods in the literature, the random subsampling has been used to handle the imbalanced dataset classification. After attribute weighting process, four classification algorithms including linear discriminant analysis, *k*-nearest neighbor classifier, support vector machine, and random forest classifier have been used to classify imbalanced medical datasets. To evaluate the performance of the proposed models, the classification accuracy, precision, recall, area under the ROC curve, κ value, and *F*-measure have been used. In the training and testing of the classifier models, three different methods including the 50–50% train–test holdout, the 60–40% train–test holdout, and tenfold cross-validation have been used. The experimental results have shown that the proposed attribute weighting methods have obtained higher classification performance than random subsampling method in the handling of classifying of the imbalanced medical datasets.

Keywords Imbalanced medical dataset classification · Data preprocessing · Attribute weighting · Clustering algorithms

1 Introduction

In this study, new data preprocessing methods have been proposed to classify medical datasets that cannot be classified as linear and have an imbalanced data distribution. In classifying imbalanced datasets, various suggestions are presented but excellent results cannot be obtained. An

✉ Kemal Polat
kpolat@ibu.edu.tr

[1] Department of Electrical and Electronics Engineering, Faculty of Engineering and Architecture, Abant Izzet Baysal University, 14280 Bolu, Turkey

imbalanced dataset is that for a classification problem of two classes, the data in one class are greater than the number in the other class. This causes overfitting in the classification algorithms. In the literature, different approaches are presented in the classification of imbalanced datasets. We can summarize these approaches in three basic categories: (a) algorithm approach, (b) data preprocessing, and (c) attribute (attribute) selection. Figure 1 gives a schematic representation of an imbalanced dataset. In addition, Fig. 2 shows the class distributions according to the attributes of datasets that can be separated linearly [1].

Many algorithms and methods have been proposed to solve the problem of classifying imbalanced datasets. These approaches are basically divided into three groups as sampling, algorithms, and attribute selection. Sampling techniques are known artificially re-sampling the dataset. Sampling can be achieved in two ways: under-sampling for large-class datasets, and oversampling for low-class datasets [4]. In addition, new algorithms have been developed to solve the class imbalance problem. The goal of these approaches is to optimize the performance of the learning algorithm on unknown data. One-class learning methods recognize instances belonging to that class and reject others. Under certain conditions, such as multidimensional datasets, one-class learning provides better performance than others [4]. Another way to improve the performance of the classifier is to apply the cost to the decision instead of the class distribution variance [4]. An imbalanced data classifier is another method of selecting an attribute. In the imbalanced data class, another solution method is the method of selecting the attributes called attribute selection. In general, the goal of the attribute (attribute) selection method is to select a subset from the *j* attributes that allow a classifier to achieve optimal performance. For high-dimensional datasets, attribute selection uses filters that score each attribute based on a rule. Attribute selection is an important data preprocessing step for machine learning algorithms when the dataset has a high size. Because the class imbalance problem is generally associated with high-dimensional datasets, the application of attribute selection methods is an important preprocessing step [4].

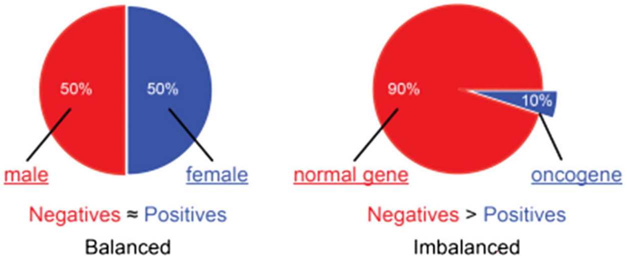

Fig. 1 Schematic representation of balanced and imbalanced datasets. Reproduced with permission from [1, 2]

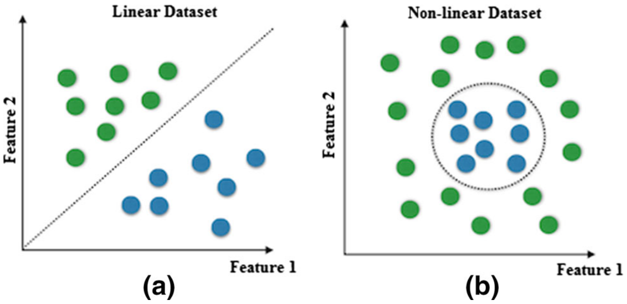

Fig. 2 Schematic representation of linear and nonlinear datasets. Reproduced with permission from [3]

Recently, a lot of methods have been proposed to classify class imbalance problems in the literature. Some of these methods are summarized as follows. Gong and Kim have proposed an effective ensemble classification method called RHSBoost to address the imbalance classification problem. This classification rule employs randomly sparse sampling and uses the ROSE sampling under a boosting scheme. According to the results of the researchers, the RHSBoost method can be used as an effective method to classify imbalanced datasets [5]. Swati Shilaskar, Ashok Ghato, and Prashant Chatur suggested a synthetic sampling technique to balance a dataset along with modified particle swarm optimization technique. They performed a comparative study based on grid selection, hybrid attribute selection, genetic algorithm, and modified particle swarm optimization [6]. XiaoWendong et al. proposed a novel ELM, class-specific cost regulation extreme learning machine for binary and multiclass classification problems with imbalanced data distributions. CCR-ELM has defined the class summary overhead cost for the misclassification of each class in a performance index as the relationship between the structural risk and the experimental risk [7]. SilviaCateni, Valentina Colla, and MarcoVannucci proposed a novel re-sampling method on the basis of the combination of an oversampling and an under-sampling technique to solve the classification problem on especially binary classification problems on imbalanced datasets [7].

There are several studies in the literature about the diagnosis and molecular structure of Parkinson's disease. These studies include treatment, diagnosis, cerebral examination of Parkinson's disease [20–22].

Apart from the literature, a new attribute weighting method has been proposed to classify medical data clusters that have both class imbalance problem and cannot be separated linearly. The used medical datasets are: Parkinson, hepatitis, Pima Indians, SPECT heart, and thoracic surgery medical datasets. Thanks to the proposed attribute weighting method, high classification performances have been achieved without size reduction or data reduction or re-sampling.

The outline of this paper is given as follows. Section 2 gives the used medical datasets. Section 3 explains the proposed method and its algorithms. In Sect. 4, the obtained results are given. Finally, Sect. 5 concludes the results and findings of this paper.

2 Imbalanced and nonlinear separable medical datasets

In this paper, we have used five imbalanced medical datasets including Parkinson, hepatitis, Pima Indians, SPECT heart, and thoracic surgery datasets to see the performance of the proposed data weighting methods. These datasets have been explained in the following.

2.1 Thoracic surgery dataset

The dataset was created by Marek Lubicz, Konrad Pawelczyk, Adam Rzechonek, and Jerzy Kolodziej at the Wroclaw Thoracic Surgery Center in Poland for patients diagnosed with the first lung cancer from 2007 to 2011. The dataset was taken from the UCI machine learning repository [8]. The dataset is shown as: rows show the patient number (470 samples) and columns show attributes (16 attributes).

The samples were labeled with ground truth values, indicating that a given patient was alive or dead. There are two labels (classes) in this dataset:

- A “false” label shows which the patient lived 1 year post the surgery (400 samples).
- A “true” label indicates the patient died within 1 year after the surgery (70 samples).

The attributes in this dataset comprise both continuous and classification data regarding the patient’s health conditions at the time of the surgery. Each patient data consist of 16 attributes. The attributes used in the dataset are listed in Table 1. In addition, the class distribution graph in the dataset is given in Fig. 3.

2.2 Parkinson disease dataset

The dataset was created by Max Little, a member of the University of Oxford, in partnership with the National Center for Voice and Speech in Denver, Colorado, recording speech signals. The Parkinson disease dataset was taken from the UCI machine learning repository [8]. The dataset consists of 195 consecutive vowel pronunciations from 23 women and 31 men diagnosed with Parkinson’s disease (PD). The age ranges in the dataset ranged from 46 to 85 (mean = 65.8 and SD = 9.8). Averages of six different voices, ranging from one to 36 s in length, have been recorded for each individual. The audio signals have been recorded directly to the computer using CSL 4300B hardware (Kay Elemetrics) with 16-bit resolution and 44.100-kHz sampling [8]. Attributes have been obtained using these sound signals. In the dataset, 22 linear and nonlinear attributes have been extracted. Table 2 shows the attributes used in the dataset. Also, the class data distribution graph in the dataset is given in Fig. 4.

2.3 Hepatitis disease dataset

The hepatitis dataset was taken from the UCI machine learning database [8]. The dataset consists of 20 attributes, including class values. The class values of the hepatitis dataset indicate whether patients with hepatitis are dead or alive. The dataset consists of a total of 155 datasets, including 23 dead cases and 123 live classes [8]. The attribute information of hepatitis dataset is shown in Table 3. Also, the class data distribution graph in hepatitis dataset is given in Fig. 5.

2.4 Pima Indians disease dataset

The Pima Indians dataset was taken from the UCI machine learning database [8]. The dataset consists of 9 attributes, including class values. The class values of the Pima Indians dataset indicate whether or not there are patients with diabetes. The dataset consists of a total of 768 datasets, including 500 having “no” class and 268 having “yes” class [8]. The attribute information of Pima Indians dataset is shown in Table 4. The class data distribution graph in Pima Indians dataset is given in Fig. 6.

2.5 Single proton emission computed tomography (SPECT) heart disease dataset

The cardiac single proton emission computed tomography (SPECT) heart dataset was taken from the UCI machine learning database [8]. The SPECT dataset explains the diagnosing of cardiac SPECT images. Each of the patients is classified into two categories: normal and abnormal. The dataset consists of a total of 267 datasets, including 55 having “no” class and 212 having “yes” class [8]. The class data distribution graph in Pima Indians dataset is given in Fig. 7. The attributes of this dataset composed of pixel values. There are 22 attributes in the dataset.

3 Method

In this paper, three new data preprocessing methods have been proposed to classify medical datasets that cannot be classified as linear and have imbalanced data distribution.

Table 1 Attribute information about thoracic surgery dataset. Reproduced with permission from [8]

Number of attribute	Name of the attribute in the dataset	Type of data	Some values of these attributes
1	DGN: Diagnosis—specific combination of ICD-10 codes for primary and secondary as well multiple tumors if any (DGN3, DGN2, DGN4, DGN6, DGN5, DGN8, DGN1)	Nominal	DGN1: 1 DGN2: 2 DGN3: 3 DGN4: 4 DGN5: 5 DGN6: 6 DGN8: 8
2	PRE4: Forced vital capacity—FVC	Numeric	Numeric values
3	PRE5: Volume that has been exhaled at the end of the first second of forced expiration—FEV1	Numeric	Numeric values
4	PRE6: Performance status—Zubrod scale (PRZ2, PRZ1, PRZ0)	Nominal	PRZ0: 0 PRZ1: 1 PRZ2: 2
5	PRE7: Pain before surgery (T, F)	Nominal	F: 0 T: 1
6	PRE8: Hemoptysis before surgery (T, F)	Nominal	F: 0 T: 1
7	PRE9: Dyspnoea before surgery (T, F)	Nominal	F: 0 T: 1
8	PRE10: Cough before surgery (T, F)	Nominal	F: 0 T: 1
9	PRE11: Weakness before surgery (T, F)	Nominal	F: 0 T: 1
10	PRE14: T in clinical TNM—size of the original tumor, from OC11 (smallest) to OC14 (largest) (OC11, OC14, OC12, OC13)	Nominal	OC11: 1 OC12: 2 OC13: 3 OC14: 4
11	PRE17: Type 2 DM—diabetes mellitus (T, F)	Nominal	F: 0 T: 1
12	PRE19: MI up to 6 months (T, F)	Nominal	F: 0 T: 1
13	PRE25: PAD—peripheral arterial diseases (T, F)	Nominal	F: 0 T: 1
14	PRE30: Smoking (T, F)	Nominal	F: 0 T: 1
15	PRE32: Asthma (T, F)	Nominal	F: 0 T: 1
16	AGE: Age at surgery (numeric)	Numeric	Numeric values

As a data-processing method, attribute weighting process has been proposed before classification algorithms. After attribute weighting, three different classification algorithms have been used to classify medical datasets with weighted imbalanced data distribution. These classification algorithms are: linear discriminant analysis (LDA), *k*-NN (*k*-nearest neighbor) classifier, support vector machine (SVM), and random forest classifier. To compare with other methods in the literature, the random subsampling has been used to handle the imbalanced dataset classification.

In order to classify medical datasets that cannot be classified as linear and have imbalanced data distribution, the flowchart of the proposed method is given in Fig. 8. The proposed attribute weighting methods include: (1) similarity-based attribute weighting in *k*-means clustering

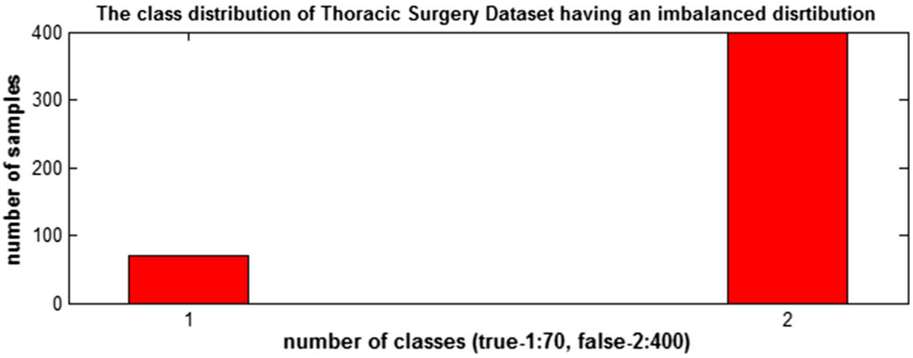

Fig. 3 Class distribution of thoracic surgery dataset

Table 2 Attribute descriptions of Parkinson disease dataset. Reproduced with permission from [8]

Description	Attribute label
Average vocal fundamental frequency	MDVP: Fo (Hz)
Maximum vocal fundamental frequency	MDVP: Fhi (Hz)
Minimum vocal fundamental frequency	MDVP: Flo (Hz)
Several measures of variation in fundamental frequency	MDVP: Jitter (%)
	MDVP: Jitter (Abs)
	MDVP: RAP
	MDVP: PPQ
	Jitter: DDP
Several measures of variation in fundamental amplitude	MDVP: Shimmer
	MDVP: Shimmer (db)
	Shimmer: APQ 3
	Shimmer APQ 5
	MDVP: APQ
	Shimmer: DDA
Two measures of ratio of noise to tonal components in the voice	NNR
	HNR
Two nonlinear dynamical complexity measures	RPDE
	D2
Signal fractal scaling exponent	DFA
Three nonlinear measures of fundamental frequency variation	Spread 1
	Spread 2
	PPE

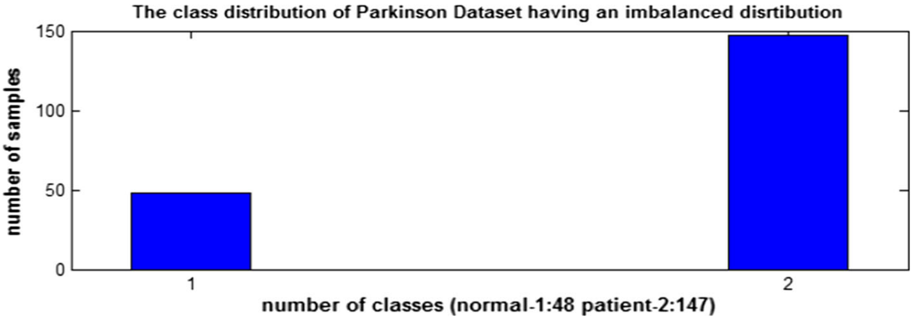

Fig. 4 Class distribution of PD dataset

(SBAWKMC), (2) similarity-based attribute weighting in fuzzy *c*-means clustering (SBAWFCM), and (3) similarity-based attribute weighting in mean shift clustering (SBAWMSC). The algorithms in the proposed hybrid systems have been explained in the following subsections.

Table 3 Attribute descriptions of hepatitis disease dataset. Reproduced with permission from [8]

Number of attribute in the dataset	Attribute information	Values
1	Age	10–80
2	Sex	Male or female
3	Steroid	No or yes
4	Antiviral	No or yes
5	Fatigue	No or yes
6	Malaise	No or yes
7	Anorexia	No or yes
8	Liver big	No or yes
9	Liver firm	No or yes
10	Spleen palpable	No or yes
11	Spiders	No or yes
12	Ascites	No or yes
13	Varices	No or yes
14	Bilirubin	0.39, 0.80, 1.20, 2.00, 3.00, 4.00
15	Alk phosphate	33, 80, 120, 160, 200, 250
16	Sgot	13, 100, 200, 300, 400, 500
17	Albumin	2.1, 3.0, 3.8, 4.5, 5.0, 6.0
18	Protime	10, 20, 30, 40, 50, 60, 70, 80, 90
19	Histology	No or yes
20	Class label	Die or live

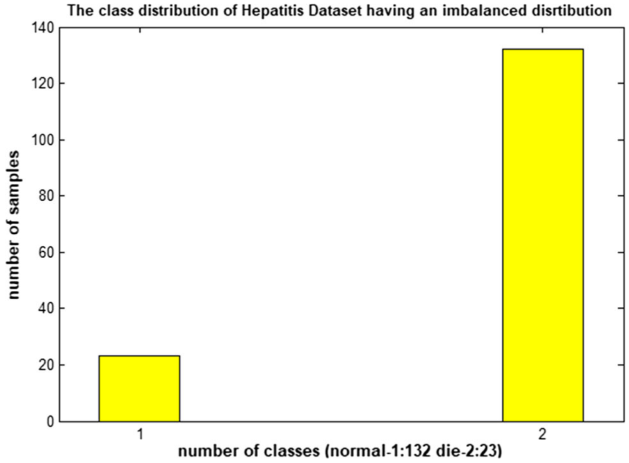

Fig. 5 Class distribution of hepatitis dataset

3.1 The proposed attribute weighting methods

Various methods have been proposed to classify data clusters with the nonlinear and imbalanced data distribution. These methods can be generalized as attribute reduction, data reduction, and random sampling. Apart from these methods, it is emphasized that a preprocessing step can be used to classify medical datasets with the nonlinear and imbalanced data distribution using attribute weighting algorithms.

The proposed attribute weighting method consists of similarity, clustering, weighted averaging, and multiplication. As a similarity measure, the absolute difference between the data values and the cluster center has been used. As the clustering method, three clustering algorithms have been used to calculate the cluster center of each attribute in each class. These clustering algorithms are *k*-means clustering (KMC), fuzzy *c*-means clustering (FCM), and mean shift clustering (MSC). The average of the difference values for each attribute has been taken as the weighted average, and then, they divided into cluster centers of each attribute. And these divided values have been determined as the weight coefficients. The dataset has been weighted by multiplying the weight coefficients with the data in each attribute separately. The presentation of the proposed attribute weighting schemes has been given as flowchart in Fig. 9.

The work of proposed attribute weighting methods is as follows:

1. First, the cluster center values of each attribute in each dataset have been computed using the clustering algorithms that include *k*-means clustering, fuzzy *c*-means clustering, and mean shift clustering, separately.
2. Subsequently, the absolute differences between the data values in each attribute and the cluster center value of the class have been calculated.
3. The weighted difference mean values for each attribute have been calculated by averaging of these difference values.

Table 4 Attribute descriptions of Pima Indians disease dataset. Reproduced with permission from [8]

Number of attribute in the dataset	Attribute information	Values
1	Number of times pregnant	Numeric values
2	Plasma glucose concentration at 2 h in an oral glucose tolerance test	Numeric values
3	Diastolic blood pressure (mm Hg)	Numeric values
4	Triceps skin fold thickness (mm)	Numeric values
5	2 H serum insulin (μU/ml)	Numeric values
6	Body mass index	Numeric values
7	Diabetes pedigree function	Numeric values
8	Age (years)	Numeric values
9	Class label	0 or 1

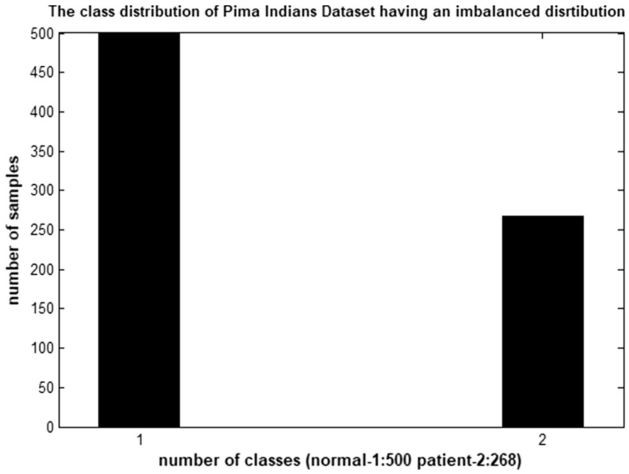

Fig. 6 Class distribution of Pima Indians dataset

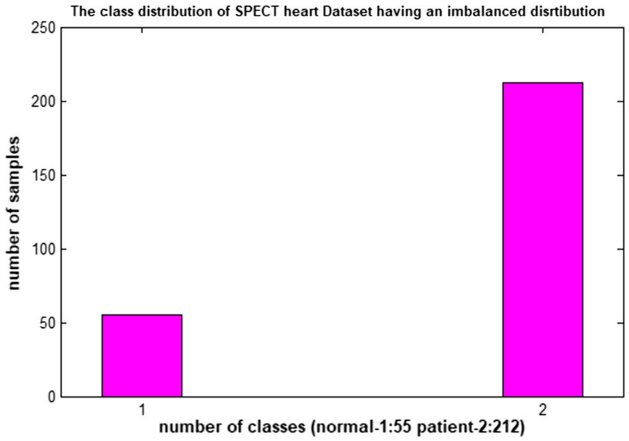

Fig. 7 Class distribution of SPECT heart dataset

4. The weighting coefficients have been calculated by dividing the average difference values calculated for each attribute to the cluster center value of the relevant class.
5. The datasets have been weighted by multiplying the weight coefficients calculated in each attribute by the data values in each class.

The pseudo-code of the proposed attribute weighting methods is given in Fig. 10.

To cluster the nominal type data, nominal data were first converted to categorical numeric data types. To show the working of the proposed attribute weighting methods in the nominal datasets, for example, the thoracic surgery dataset has been given as the schematic representation. For the first attribute in the thoracic surgery dataset, we have given the categorical numeric values of this attribute in Table 5.

For other attributes in this dataset, there are two nominal data including T and F. With the same idea, these nominal data were converted to categorical numeric data as 0 (F) and 1 (T). The all new values of attributes in the thoracic surgery dataset are given in Table 1. After the nominal data has been converted to the categorical numeric data type, routinely the attribute weighting methods have been applied to datasets having the categorical numeric data.

After weighting process, the obtained numeric values are again converted to nominal categorical data using the following equalization in Fig. 11. In Fig. 11, since there are 7 categories in the first attribute in the dataset, the difference between the maximum value and the minimum value is divided by 7. After the above procedure applied to 1 attribute of thoracic surgery dataset, the obtained categorical nominal values are given in Table 6.

In the thoracic surgery dataset, there are also ten nominal data having T and F labels. For these attribute, we have used the same method, but the interval is changed from 7 to 2. With the same method, the obtained weighted numeric values are again converted to nominal data type using the below and above structures and formulas. Figure 12 shows the converting formula from numeric values to nominal data type for the attributes having two labels in the dataset.

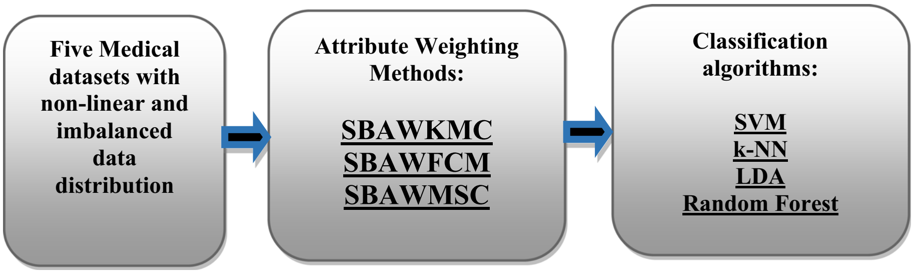
Five Medical datasets with non-linear and imbalanced data distribution
Attribute Weighting Methods:
SBAWKMC
SBAWFCM
SBAWMSC
Classification algorithms:
SVM
k-NN
LDA
Random Forest

Fig. 8 Flowchart of the proposed hybrid methods (combination of classifier algorithms and attribute weighting methods)

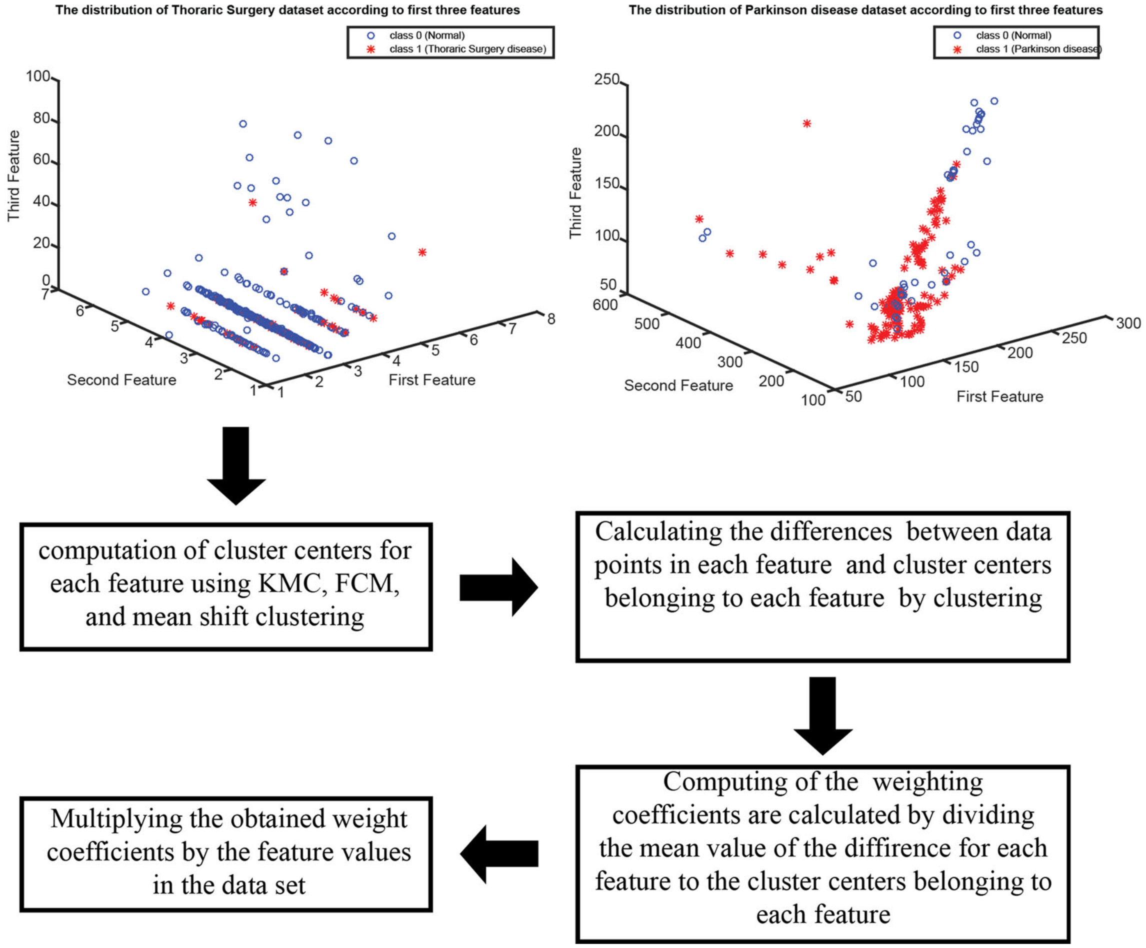
The distribution of Thoraric Surgery dataset according to first three features
class 0 (Normal)
class 1 (Thoraric Surgery disease)
Third Feature
Second Feature
First Feature
The distribution of Parkinson disease dataset according to first three features
class 0 (Normal)
class 1 (Parkinson disease)
Third Feature
Second Feature
First Feature
computation of cluster centers for each feature using KMC, FCM, and mean shift clustering
Calculating the differences between data points in each feature and cluster centers belonging to each feature by clustering
Computing of the weighting coefficients are calculated by dividing the mean value of the diffirence for each feature to the cluster centers belonging to each feature
Multiplying the obtained weight coefficients by the feature values in the data set

Fig. 9 Flowchart of the proposed attribute weighting methods

3.2 Random subsampling

Random subsampling, also known as Monte Carlo cross-validation, can be defined as the random subdivision of the data into subsets, where the size of the subclasses is defined by the user [23]. Random splitting of the data can often be repeated arbitrarily. In contrast to the cross-validation procedure, Random subsampling showed that the test samples had more predictable results. The estimates of the test data give the true value of the estimates of the external verification data [23].

3.3 Classification algorithms

After weighting process of imbalanced medical datasets, we have used four classifier algorithms including linear discriminant analysis (LDA), *k*-NN (*k*-nearest neighbor) classifier, support vector machine (SVM), and random forest to classify the weighted medical datasets. These classifier algorithms have been used in the following subsections.

```
algorithm SBFWKMC - SBFWFCM- SBFWMSC is
input: Parkinson dataset (195x22: 195 data, 22 attributes) and Thoracic
Surgery medical dataset (470x16: 470 data, 16 attributes).
input matrix A=m x n matrix: m← number of data, n ← number of attributes

output: W, weighted dataset (m x n matrix)

for each attribute in datasets with two classes do
       cluster centers ← finding of the cluster centers for each attribute
                         in the datasets using k-means clustering
                         (SBFWKMC), FCM (SBFWFCM), and mean shift
                         clustering (SBFWMSC)
end for
%%%%% the following process was done for first class in the dataset. Here,
the difference values between each data in each attribute and related
cluster center (first one)

for j←1 to n do
   for i←1 to m do
           difference_1(j,i)=abs(A(j,i)- cluster_centers(1,i));
end for
end for

for each attribute in the dataset do
             mean_difference1=mean(difference_1)
end for

%%%%% calculating of weight coefficients for the first class using
following ratios.
 for i←1 to m do

  if cluster_centers (1,i)==0,
            cluster_centers (1,i)=1;
        end if

 weight1(1,i)= mean_difference1 (1,i)/ cluster_centers (1,i);

end for

 %%%%% weighting process: the multiplication of weight coefficients
 with A matrix

for i←1 to m do

         weighted_A(:,i)=(A(:,i)* weight1(:,i));
end for
%%%%%%%%%%% all the above operations are repeated for the second class in
the dataset.
```

Fig. 10 Pseudo-code of the proposed attribute weighting methods including SBAWKMC, SBAWFCM, and SBAWMSC

3.3.1 *k*-NN classifier

k-NN (nearest neighbors) is a simple classification algorithm that loads all appropriate states and classifies new states based on the similarity (distance functions) between the data. *k*-NN was used as a nonparametric method in pattern recognition and statistical estimation problems at the beginning of the 1970s. *k*-NN is a nonparametric lazy learning type. For more information on *k*-NN classifier, the readers can refer to [9–11].

Table 5 Categorical numeric values of the first attribute in the thoracic surgery dataset

1	DGN: Diagnosis—specific combination of ICD-10 codes for primary and secondary as well multiple tumors if any (DGN3, DGN2, DGN4, DGN6, DGN5, DGN8, DGN1)	Nominal	DGN1: 1 DGN2: 2 DGN3: 3 DGN4: 4 DGN5: 5 DGN6: 6 DGN8: 8

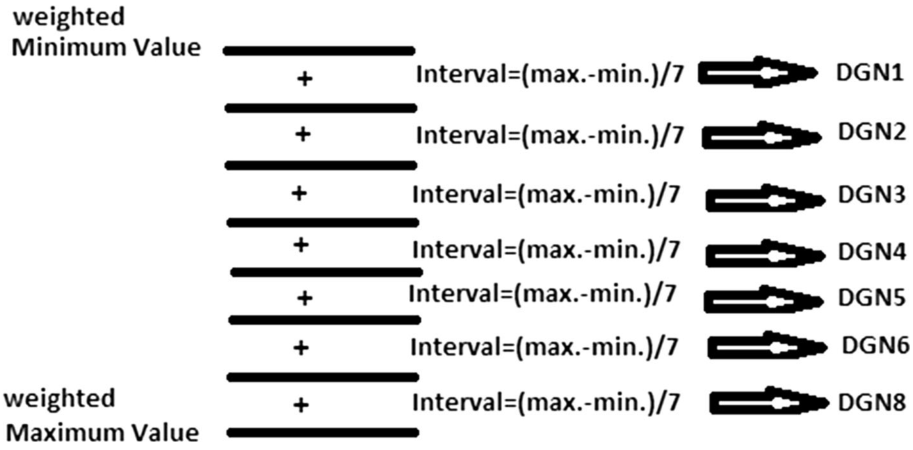

Fig. 11 Converting formula from numeric values to nominal data type for the attributes having seven labels in the dataset

Table 6 New categorical nominal values of 1. Attribute of thoracic surgery dataset after weighting process

Range using the above formula (minimum value: 0.106, maximum value: 1.601) interval = 0.213	New nominal acquired value
0.106–0.319	DGN1
0.320–0.533	DGN2
0.534–0.747	DGN3
0.748–0.961	DGN4
0.962–1.175	DGN5
1.176–1.389	DGN6
1.390–1.601	DGN8

3.3.2 Random forest classifier

Random forest classification algorithm is a consultative classification method. This algorithm produces a forest with many trees. In general, the more the trees in the forest, the stronger the forest. In a similar way in the random forest class, the growth of the number of trees in the forest gives a higher classification accuracy. The random forest classification algorithm can be used in both classification and regression problems. For more information on random forest classifier, the readers can refer to [12, 13].

3.3.3 Linear discriminant analysis (LDA)

Linear discriminant analysis (LDA) is a generalized form of Fisher's linear discriminant, a method used in statistics, pattern recognition, and machine learning to find a linear combination of attributes that distinguishes or characterizes two or more classes or objects. The resulting attribute association can also be used for dimensionality reduction before it is classified as a linear classifier or more generally. For more information on LDA, the readers can refer to [14–16].

3.3.4 Support vector machines (SVM)

In machine learning, support vector machines (SVMs) are the supervised learning models used for classification and regression analysis. SVM is a statistical-based learning algorithm proposed by Vapnik in 1995. In general, it has been designed to solve two-class classification problems. An SVM is created based on a model that allocates new samples to one category or another by giving a set of training samples in which each value is labeled in one or the other of the two categories. An SVM model is an illustration of examples as space points. Samples of discrete categories are split up as wide as possible into a space. New samples are then mapped in the same space and

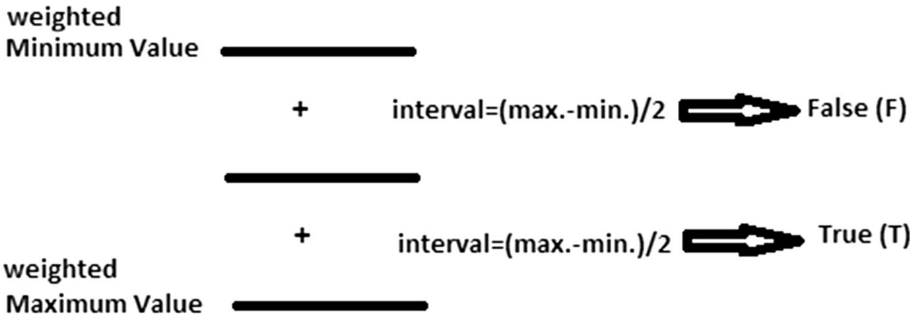

Fig. 12 Converting formula from numeric values to nominal data type for the attributes having two labels in the dataset

Table 7 Confusion matrix. Reproduced with permission from [23]

Actual result			
	Yes	No	Total
Predicted result			
Yes	True positive (t_{p})	False positive (f_{p})	t_{Poz}
No	False negative (f_{n})	True negative (t_{n})	t_{Neg}
Total	Positive	Negative	m

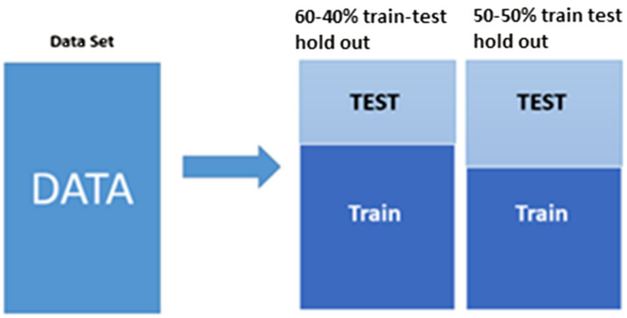

Fig. 13 Test and training set separation representations with 50–50% training–testing partition and 60–40% training–testing partition methods

are estimated to belong to a category depending on where the samples fall. In addition to performing linear classifications, SVMs can efficiently perform a nonlinear classification by transforming the input space into a higher dimensionality space with a function called kernel trick. In this work, RBF kernel functions have been used as kernel functions. For more information on SVM, the readers can refer to [17–19].

4 The experimental results and discussion

4.1 The performance evaluation metrics

To evaluate the classification performance of the proposed models, the confusion matrix has been used. And also, the κ value and AUC (area under the ROC curve) have been used as the comparing parameter.

In the confusion matrix, the actual values and the values predicted by the classification algorithm are shown in Table 7. Performance evaluation criteria of classification algorithms are shown in Table 7.

The accuracy of the model generated by the classification algorithms according to Table 7 is given by Eq. (1) [23]:

$$\text{Accuracy} = \frac{t_{\mathrm{p}} + \mathrm{t}_n}{m} \tag{1}$$

The effectiveness of estimating the positive class labels of the classifier is called precision (P). The P formula is shown in Eq. (2) [23].

$$P = \frac{T_{\mathrm{p}}}{T_{\mathrm{p}} + F_{\mathrm{p}}} \tag{2}$$

The ratio of the positive samples to the predicted samples is called precision. The precision equation is shown in Eq. (3) [23]:

$$Precision = \frac{t_{\mathrm{p}}}{t_{\mathrm{Poz}}} = \frac{t_{\mathrm{p}}}{t_{\mathrm{p}} + f_{\mathrm{p}}} \tag{3}$$

The F-measure equation is shown in Eq. (4) [23]:

$$F\text{-measure} = \frac{2 * \text{Precision} * \text{Sentivity}}{\text{Precision} + \text{Sensitivity}} \tag{4}$$

In the training and testing of the classifier models, three different methods including the 50–50% train–test holdout,

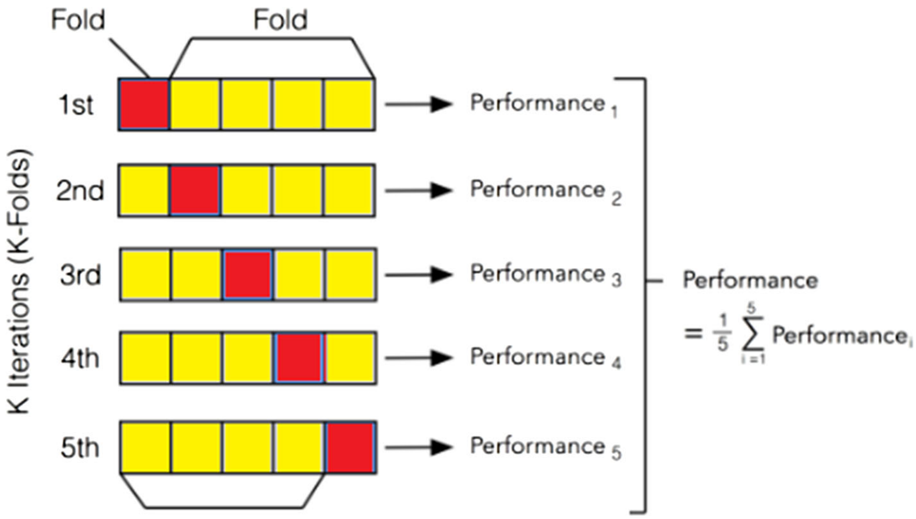

Fig. 14 Test and training set separation representation with the fivefold cross-validation method

Table 8 Obtained test results in the classification of thoracic surgery disease dataset based on the various combinations of attribute weighting methods (SBAWFCM, SBAWKMC, and SBAWMSC) and random subsampling with *LDA* classifier with 50–50% training–testing holdout, 60–40% training–testing holdout, and tenfold cross-validation

Main folder name	Data partition	Accuracy	Precision	Recall	AUC	κ value	F-measure
Weighted dataset with SBAWFCM	50–50% train–test partition	95.744	0.957	0.957	0.989	0.849	0.957
Weighted dataset with SBAWKMC		95.744	0.957	0.957	0.988	0.849	0.957
Weighted dataset with SBAWMSC		94.893	0.952	0.949	0.972	0.799	0.945
In Raw dataset		75.319	0.700	0.753	0.595	− 0.036	0.724
Re-sampling approach to raw dataset		82.127	0.789	0.821	0.731	0.231	0.797
Weighted dataset with SBAWFCM	60–40% train–test partition	96.276	0.962	0.963	0.990	0.878	0.963
Weighted dataset with SBAWKMC		96.276	0.962	0.962	0.988	0.878	0.962
Weighted dataset with SBAWMSC		95.744	0.958	0.957	0.997	0.853	0.956
In Raw dataset		75.000	0.644	0.750	0.611	− 0.098	0.693
Re-sampling approach to raw dataset		79.787	0.773	0.798	0.763	0.252	0.781
Weighted dataset with SBAWFCM	Tenfold cross-validation	95.531	0.954	0.955	0.982	0.811	0.954
Weighted dataset with SBAWKMC		95.744	0.956	0.957	0.981	0.853	0.956
Weighted dataset with SBAWMSC		97.23	0.972	0.972	0.993	0.884	0.972
In Raw dataset		81.914	0.720	0.819	0.653	− 0.055	0.766
Re-sampling approach to raw dataset		86.170	0.839	0.862	0.758	0.322	0.843

Table 9 Obtained test results in the classification of thoracic surgery disease dataset based on the various combinations of attribute weighting methods (SBAWFCM, SBAWKMC, and SBAWMSC) and random subsampling with **SVM** classifier with 50–50% training–testing holdout, 60–40% training–testing holdout, and tenfold cross-validation

Main folder name	Data partition	Accuracy	Precision	Recall	AUC	κ value	F-measure
Weighted dataset with SBAWFCM	50–50% train–test partition	95.744	0.958	0.957	0.888	0.840	0.956
Weighted dataset with SBAWKMC		95.319	0.954	0.953	0.875	0.822	0.951
Weighted dataset with SBAWMSC		93.617	0.941	0.936	0.817	0.741	0.930
In Raw dataset		82.553	0.682	0.826	0.500	0.000	0.747
Re-sampling approach to raw dataset		82.553	0.682	0.826	0.500	0.000	0.747
Weighted dataset with SBAWFCM	60–40% train–test partition	95.744	0.958	0.957	0.899	0.853	0.956
Weighted dataset with SBAWKMC		95.212	0.953	0.952	0.886	0.833	0.950
Weighted dataset with SBAWMSC		94.680	0.950	0.947	0.861	0.807	0.943
In Raw dataset		80.851	0.654	0.809	0.500	0.000	0.723
Re-sampling approach to raw dataset		80.851	0.654	0.809	0.500	0.000	0.723
Weighted dataset with SBAWFCM	Tenfold cross-validation	95.531	0.955	0.955	0.868	0.806	0.953
Weighted dataset with SBAWKMC		94.255	0.941	0.943	0.837	0.748	0.939
Weighted dataset with SBAWMSC		96.170	0.962	0.962	0.877	0.833	0.960
In Raw dataset		85.106	0.724	0.851	0.500	0.000	0.783
Re-sampling approach to raw dataset		85.106	0.724	0.851	0.500	0.000	0.783

the 60–40% train–test holdout, and tenfold cross-validation have been used. In the holdout data partition method, two different partitions have been used. These are 50–50% training–testing partition and 60–40% training–testing partition. The schematic representations of these methods are given in Fig. 13. Also, the working of *k*-fold cross-validation is given in Fig. 14.

4.2 The results

In this study, a new data preprocessing method has been proposed to classify Parkinson, hepatitis, Pima Indians, SPECT heart, and thoracic surgery medical datasets having imbalanced data distribution. Before classifying five medical datasets with nonlinear and imbalanced data

Table 10 Obtained test results in the classification of thoracic surgery disease dataset based on the various combinations of attribute weighting methods (SBAWFCM, SBAWKMC, and SBAWMSC) and random subsampling with *k*-**NN** (for *k* = 1) classifier with 50–50% training–testing holdout, 60–40% training–testing holdout, and tenfold cross-validation

Main folder name	Data partition	Accuracy	Precision	Recall	AUC	κ value	*F*-measure
Weighted dataset with SBAWFCM	50–50% train–test partition	94.042	0.939	0.940	0.868	0.780	0.939
Weighted dataset with SBAWKMC		95.744	0.957	0.957	0.897	0.843	0.956
Weighted dataset with SBAWMSC		94.893	0.952	0.949	0.854	0.799	0.945
In Raw dataset		76.595	0.706	0.766	0.493	− 0.018	0.732
Re-sampling approach to raw dataset		85.957	0.845	0.860	0.758	0.430	0.846
Weighted dataset with SBAWFCM	60–40% train–test partition	94.148	0.940	0.941	0.890	0.804	0.941
Weighted dataset with SBAWKMC		95.212	0.952	0.952	0.896	0.836	0.951
Weighted dataset with SBAWMSC		95.744	0.960	0.957	0.889	0.849	0.955
In Raw dataset		75.000	0.686	0.750	0.496	− 0.011	0.713
Re-sampling approach to raw dataset		84.574	0.832	0.846	0.730	0.436	0.835
Weighted dataset with SBAWFCM	Tenfold cross-validation	94.680	0.947	0.947	0.892	0.788	0.947
Weighted dataset with SBAWKMC		95.531	0.955	0.955	0.903	0.820	0.955
Weighted dataset with SBAWMSC		97.446	0.975	0.974	0.920	0.893	0.974
In Raw dataset		75.744	0.739	0.757	0.486	− 0.029	0.748
Re-sampling approach to raw dataset		90.851	0.906	0.909	0.811	0.628	0.907

Table 11 Obtained test results in the classification of thoracic surgery disease dataset based on the various combinations of attribute weighting methods (SBAWFCM, SBAWKMC, and SBAWMSC) and random subsampling with random forests classifier with 50–50% training–testing holdout, 60–40% training–testing holdout, and tenfold cross-validation

Main folder name	Data partition	Accuracy	Precision	Recall	AUC	κ value	*F*-measure
Weighted dataset with SBAWFCM	50–50% train–test partition	100.000	1.000	1.000	1.000	1.000	1.000
Weighted dataset with SBAWKMC		99.574	0.996	0.996	1.000	0.985	0.996
Weighted dataset with SBAWMSC		98.723	0.987	0.987	1.000	0.954	0.987
In Raw dataset		81.702	0.680	0.817	0.580	− 0.016	0.742
Re-sampling approach to raw dataset		88.5106	0.886	0.885	0.714	0.487	0.866
Weighted dataset with SBAWFCM	60–40% train–test partition	100.000	1.000	1.000	1.000	1.000	1.000
Weighted dataset with SBAWKMC		100.000	1.000	1.000	1.000	1.000	1.000
Weighted dataset with SBAWMSC		99.468	0.995	0.995	1.000	0.982	0.995
In Raw dataset		79.255	0.651	0.793	0.583	− 0.030	0.715
Re-sampling approach to raw dataset		87.766	0.878	0.878	0.778	0.505	0.859
Weighted dataset with SBAWFCM	Tenfold cross-validation	99.787	0.998	0.998	1.000	0.991	0.998
Weighted dataset with SBAWKMC		100.000	1.000	1.000	1.000	1.000	1.000
Weighted dataset with SBAWMSC		99.574	0.996	0.996	1.000	0.985	0.996
In Raw dataset		84.468	0.723	0.845	0.652	− 0.012	0.779
Re-sampling approach to raw dataset		93.404	0.934	0.934	0.925	0.694	0.928

distribution, similarity-based attribute weighting methods have been proposed based on the similarity between cluster centers and data points in each attribute. As the clustering algorithms, *k*-means clustering, fuzzy *c*-means clustering, and mean shift clustering algorithms have been used. After medical datasets weighting, Parkinson, hepatitis, Pima Indians, SPECT heart, and thoracic surgery medical datasets with imbalanced data distribution have been classified with a high classification accuracy using four different classification algorithms including *k*-NN classifier, LDA, random forests classifier, and SVM. The proposed attribute weighting methods are: (1) similarity-based attribute weighting in *k*-means clustering-SBAWKMC, (2) similarity-based attribute weighting in fuzzy *c*-means

Table 12 Obtained test results in the classification of Parkinson disease dataset based on the various combinations of attribute weighting methods (SBAWFCM, SBAWKMC, and SBAWMSC) and random subsampling with **LDA** classifier with 50–50% training–testing holdout, 60–40% training–testing holdout, and tenfold cross-validation

Main folder name	Data partition	Accuracy	Precision	Recall	AUC	κ value	*F*-measure
Weighted dataset with SBAWFCM	50–50% train–test partition	92.783	0.935	0.759	0.982	0.815	0.925
Weighted dataset with SBAWKMC		92.783	0.935	0.928	0.990	0.815	0.925
Weighted dataset with SBAWMSC		100.000	1.000	1.000	1.000	1.000	1.000
In Raw dataset		77.319	0.763	0.773	0.820	0.3861	0.753
Re-sampling approach to raw dataset		81.443	0.835	0.814	0.861	0.474	0.789
Weighted dataset with SBAWFCM	60–40% train–test partition	93.589	0.941	0.936	0.996	0.83	0.933
Weighted dataset with SBAWKMC		96.153	0.963	0.962	0.997	0.900	0.961
Weighted dataset with SBAWMSC		100.000	1.000	1.000	1.000	1.000	1.000
In Raw dataset		80.769	0.799	0.808	0.886	0.490	0.799
Re-sampling approach to raw dataset		85.897	0.868	0.859	0.912	0.602	0.846
Weighted dataset with SBAWFCM	Tenfold cross-validation	94.359	0.943	0.944	0.991	0.844	0.943
Weighted dataset with SBAWKMC		94.871	0.949	0.949	0.992	0.855	0.948
Weighted dataset with SBAWMSC		100.000	1.000	1.000	1.000	1.000	1.000
In Raw dataset		85.128	0.845	0.851	0.867	0.559	0.843
Re-sampling approach to raw dataset		86.153	0.857	0.862	0.885	0.589	0.854

Table 13 Obtained test results in the classification of Parkinson disease dataset based on the various combinations of attribute weighting methods (SBAWFCM, SBAWKMC, and SBAWMSC) and random subsampling with **SVM** classifier with 50–50% training–testing holdout, 60–40% training–testing holdout, and tenfold cross-validation

Main folder name	Data partition	Accuracy	Precision	Recall	AUC	κ value	*F*-measure
Weighted dataset with SBAWFCM	50–50% train–test partition	85.567	0.880	0.856	0.759	0.600	0.839
Weighted dataset with SBAWKMC		84.536	0.873	0.845	0.741	0.566	0.826
Weighted dataset with SBAWMSC		100.000	1.000	1.000	1.000	1.000	1.000
In Raw dataset		84.536	0.973	0.845	0.741	0.566	0.826
Re-sampling approach to raw dataset		82.474	0.860	0.825	0.707	0.497	0.798
Weighted dataset with SBAWFCM	60–40% train–test partition	87.179	0891	0.872	0.773	0.632	0.858
Weighted dataset with SBAWKMC		87.179	0.891	0.872	0.773	0.632	0.858
Weighted dataset with SBAWMSC		100.000	1.000	1.000	1.000	1.000	1.000
In Raw dataset		84.615	0.873	0.846	0.727	0.544	0.825
Re-sampling approach to raw dataset		85.897	0.882	0.859	0.750	0.589	0.842
Weighted dataset with SBAWFCM	Tenfold cross-validation	89.230	0.894	0.892	0.802	0.675	0.885
Weighted dataset with SBAWKMC		90.256	0.904	0.903	0.823	0.711	0.897
Weighted dataset with SBAWMSC		100.000	1.000	1.000	1.000	1.000	1.000
In Raw dataset		86.153	0.869	0.862	0.733	0.555	0.845
Re-sampling approach to raw dataset		88.205	0.898	0.882	0.760	0.621	0.868

clustering-SBAWFCM, and (3) similarity-based weighting in mean shift clustering- SBAWMSC. After attribute weighting, four classification algorithms including *k*-NN classifier, LDA, random forests classifier, and SVM have been used.

Six different performance measures have been used to test the performance of the proposed hybrid systems (combination of attribute weighting methods and classifier algorithms). These performance measures: classification accuracy (%), precision (%), recall (%), AUC (area under the ROC curve), κ value, and *F*-measure.

The classification results obtained in thoracic surgery dataset, which is not classified as linear and having imbalanced class distribution, are given in Table 8 using

Table 14 Obtained test results in the classification of Parkinson disease dataset based on the various combinations of attribute weighting methods (SBAWFCM, SBAWKMC, and SBAWMSC) and random subsampling with *k*-NN (for *k* = 1) classifier with 50–50% training–testing holdout, 60–40% training–testing holdout, and tenfold cross-validation

Main folder name	Data partition	Accuracy	Precision	Recall	AUC	κ value	*F*-measure
Weighted dataset with SBAWFCM	50–50% train–test partition	92.783	0.930	0.928	0.889	0.818	0.926
Weighted dataset with SBAWKMC		93.814	0.939	0.938	0.906	0.846	0.937
Weighted dataset with SBAWMSC		100.000	1.000	1.000	1.000	1.000	1.000
In Raw dataset		92.783	0.927	0.928	0.909	0.826	0.927
Re-sampling approach to raw dataset		98.969	0.990	0.990	0.986	0.975	0.990
Weighted dataset with SBAWFCM	60–40% train–test partition	93.589	0.936	0.936	0.900	0.834	0.934
Weighted dataset with SBAWKMC		94.871	0.949	0.949	0.923	0.869	0.948
Weighted dataset with SBAWMSC		100.000	1.000	1.000	1.000	1.000	1.000
In Raw dataset		94.871	0.949	0.949	0.923	0.869	0.948
Re-sampling approach to raw dataset		97.435	0.974	0.974	0.976	0.936	0.974
Weighted dataset with SBAWFCM	Tenfold cross-validation	96.923	0.969	0.969	0.944	0.915	0.969
Weighted dataset with SBAWKMC		97.948	0.980	0.979	0.976	0.945	0.980
Weighted dataset with SBAWMSC		100.000	1.000	1.000	1.000	1.000	1.000
In Raw dataset		96.410	0.965	0.964	0.963	0.905	0.964
Re-sampling approach to raw dataset		98.974	0.990	0.990	0.984	0.972	0.990

Table 15 Obtained test results in the classification of Parkinson disease dataset based on the various combinations of attribute weighting methods (SBAWFCM, SBAWKMC, and SBAWMSC) and random subsampling with *random forests* classifier with 50–50% training–testing holdout, 60–40% training–testing holdout, and tenfold cross-validation

Main folder name	Data partition	Accuracy	Precision	Recall	AUC	κ value	*F*-measure
Weighted dataset with SBAWFCM	50–50% train–test partition	90.721	0.912	0.907	0.952	0.762	0.903
Weighted dataset with SBAWKMC		92.783	0.935	0.928	0.962	0.815	0.925
Weighted dataset with SBAWMSC		100.000	1.000	1.000	1.000	1.000	1.000
In Raw dataset		86.597	0.870	0.866	0.919	0.649	0.858
Re-sampling approach to raw dataset		86.597	0.877	0.866	0.980	0.641	0.855
Weighted dataset with SBAWFCM	60–40% train–test partition	93.589	0.941	0.936	0.952	0.830	0.933
Weighted dataset with SBAWKMC		93.589	0.941	0.936	0.970	0.830	0.933
Weighted dataset with SBAWMSC		100.000	1.000	1.000	1.000	1.000	1.000
In Raw dataset		88.461	0.884	0.885	0.952	0.694	0.880
Re-sampling approach to raw dataset		91.025	0.913	0.910	0.988	0.762	0.906
Weighted dataset with SBAWFCM	Tenfold cross-validation	93.846	0.939	0.938	0.979	0.824	0.936
Weighted dataset with SBAWKMC		96.923	0.970	0.969	0.981	0.913	0.969
Weighted dataset with SBAWMSC		100.000	1.000	1.000	1.000	1.000	1.000
In Raw dataset		92.307	0.922	0.923	0.959	0.782	0.921
Re-sampling approach to raw dataset		95.897	0.959	0.959	0.992	0.889	0.959

the new hybrid systems by the combination of LDA classifier and attribute weighting methods (with random subsampling method). Table 9 shows the classification results obtained in the classification of the thoracic surgery dataset using the new hybrid systems by the combination of SVM classifier and attribute weighting methods (with random subsampling method). The results obtained in the classification of the thoracic surgery dataset using the new hybrid systems with the combination of the *k*-NN classifier and the attribute weighting methods (with random subsampling method) are given in Table 10. Table 11 gives the obtained test results in the classification of thoracic surgery disease dataset based on the various combinations of attribute weighting methods (SBAWFCM, SBAWKMC,

Table 16 Obtained test results in the classification of hepatitis disease dataset based on the various combinations of attribute weighting methods (SBAWFCM, SBAWKMC, and SBAWMSC) and random subsampling with **LDA** classifier with 50–50% training–testing holdout, 60–40% training–testing holdout, and tenfold cross-validation

Main folder name	Data partition	Accuracy	Precision	Recall	AUC	κ value	*F*-measure
Weighted dataset with SBAWFCM	50–50% train–test partition	98.701	0.987	0.987	1.000	0.963	0.987
Weighted dataset with SBAWKMC		100.000	1.000	1.000	1.000	1.000	1.000
Weighted dataset with SBAWMSC		98.701	0.988	0.987	1.000	0.964	0.987
In Raw dataset		80.519	0.787	0.805	0.870	0.371	0.787
Re-sampling approach to raw dataset		75.324	0.758	0.753	0.766	0.324	0.756
Weighted dataset with SBAWFCM	60–40% train–test partition	98.387	0.984	0.984	0.998	0.049	0.984
Weighted dataset with SBAWKMC		100.000	1.000	1.000	1.000	1.000	1.000
Weighted dataset with SBAWMSC		98.387	0.985	0.984	1.000	0.952	0.984
In Raw dataset		91.935	0.918	0.919	0.942	0.749	0.918
Re-sampling approach to raw dataset		80.645	0.818	0.806	0.694	0.447	0.811
Weighted dataset with SBAWFCM	Tenfold cross-validation	99.354	0.994	0.994	1.000	0.980	0.994
Weighted dataset with SBAWKMC		99.354	0.994	0.994	1.000	0.980	0.994
Weighted dataset with SBAWMSC		98.064	0.982	0.981	1.000	0.942	0.981
In Raw dataset		81.935	0.808	0.819	0.792	0.407	0.812
Re-sampling approach to raw dataset		82.580	0.820	0.826	0.836	0.449	0.823

Table 17 Obtained test results in the classification of hepatitis disease dataset based on the various combinations of attribute weighting methods (SBAWFCM, SBAWKMC, and SBAWMSC) and random subsampling with **SVM** classifier with 50–50% training–testing holdout, 60–40% training–testing holdout, and tenfold cross-validation

Main folder name	Data partition	Accuracy	Precision	Recall	AUC	κ value	*F*-measure
Weighted dataset with SBAWFCM	50–50% train–test partition	100.000	1.000	1.000	1.000	1.000	1.000
Weighted dataset with SBAWKMC		100.000	1.000	1.000	1.000	1.000	1.000
Weighted dataset with SBAWMSC		100.000	1.000	1.000	1.000	1.000	1.000
In Raw dataset		76.623	0.587	0.766	0.500	0.000	0.665
Re-sampling approach to raw dataset		80.519	0.809	0.805	0.738	0.466	0.807
Weighted dataset with SBAWFCM	60–40% train–test partition	100.000	1.000	1.000	1.000	1.000	1.000
Weighted dataset with SBAWKMC		100.000	1.000	1.000	1.000	1.000	1.000
Weighted dataset with SBAWMSC		100.000	1.000	1.000	1.000	1.000	1.000
In Raw dataset		83.871	0.833	0.839	0.644	0.371	0.809
Re-sampling approach to raw dataset		72.580	0.749	0.726	0.629	0.237	0.736
Weighted dataset with SBAWFCM	Tenfold cross-validation	100.000	1.000	1.000	1.000	1.000	1.000
Weighted dataset with SBAWKMC		100.000	1.000	1.000	1.000	1.000	1.000
Weighted dataset with SBAWMSC		99.354	0.994	0.994	0.984	0.980	0.994
In Raw dataset		81.935	0.798	0.819	9.643	0.342	0.799
Re-sampling approach to raw dataset		85.161	0.841	0.852	0.722	0.494	0.842

and SBAWMSC) and random subsampling with random forests classifier with 50–50% training–testing holdout, 60–40% training–testing holdout, and tenfold cross-validation.

As for the classification of Parkinson disease (PD) dataset, the obtained results are given in Tables 12, 13, 14, and 15. Table 12 shows the obtained classification performance results using the new hybrid systems by the combination of LDA classifier and attribute weighting methods in the classification of PD dataset. Table 13 gives the classification results obtained in the classification of the Parkinson disease (PD) dataset using the combination of SVM classifier and attribute weighting methods. The results obtained in the classification of the PD dataset using

Table 18 Obtained test results in the classification of hepatitis disease dataset based on the various combinations of attribute weighting methods (SBAWFCM, SBAWKMC, and SBAWMSC) and random subsampling with *k*-NN (for *k* = 1) classifier with 50–50% training–testing holdout, 60–40% training–testing holdout, and tenfold cross-validation

Main folder name	Data partition	Accuracy	Precision	Recall	AUC	κ value	*F*-measure
Weighted dataset with SBAWFCM	50–50% train–test partition	94.805	0.951	0.948	0.889	0.842	0.946
Weighted dataset with SBAWKMC		100.000	1.000	1.000	1.000	1.000	1.000
Weighted dataset with SBAWMSC		98.701	0.988	0.987	0.993	0.964	0.987
In Raw dataset		76.6234	0.749	0.766	0.635	0.292	0.755
Re-sampling approach to raw dataset		80.519	0.802	0.805	0.761	0.445	0.803
Weighted dataset with SBAWFCM	60–40% train–test partition	93.548	0.940	0.935	0.846	0.780	0.931
Weighted dataset with SBAWKMC		100.000	1.000	1.000	1.000	1.000	1.000
Weighted dataset with SBAWMSC		98.387	0.985	0.984	0.990	0.952	0.984
In Raw dataset		80.645	0.806	0.806	0.708	0.416	0.806
Re-sampling approach to raw dataset		79.023	0.774	0.790	0.660	0.308	0.780
Weighted dataset with SBAWFCM	Tenfold cross-validation	96.129	0.961	0.961	0.911	0.876	0.960
Weighted dataset with SBAWKMC		99.354	0.994	0.994	0.976	0.980	0.994
Weighted dataset with SBAWMSC		97.419	0.974	0.974	0.980	0.921	0.974
In Raw dataset		81.290	0.811	0.813	0.688	0.422	0.812
Re-sampling approach to raw dataset		92.258	0.921	0.923	0.851	0.758	0.922

Table 19 Obtained test results in the classification of hepatitis disease dataset based on the various combinations of attribute weighting methods (SBAWFCM, SBAWKMC, and SBAWMSC) and random subsampling with **random forests** classifier with 50–50% training–testing holdout, 60–40% training–testing holdout, and tenfold cross-validation

Main folder name	Data partition	Accuracy	Precision	Recall	AUC	κ value	*F*-measure
Weighted dataset with SBAWFCM	50–50% train–test partition	100.000	1.000	1.000	1.000	1.000	1.000
Weighted dataset with SBAWKMC		100.000	1.000	1.000	1.000	1.000	1.000
Weighted dataset with SBAWMSC		100.000	1.000	1.000	1.000	1.000	1.000
In Raw dataset		81.818	0.808	0.818	0.918	0.371	0.790
Re-sampling approach to raw dataset		84.415	0.835	0.844	0.860	0.508	0.832
Weighted dataset with SBAWFCM	60–40% train–test partition	100.000	1.000	1.000	1.000	1.000	1.000
Weighted dataset with SBAWKMC		100.000	1.000	1.000	1.000	1.000	1.000
Weighted dataset with SBAWMSC		100.000	1.000	1.000	1.000	1.000	1.000
In Raw dataset		88.709	0.882	0.887	0.943	0.627	0.881
Re-sampling approach to raw dataset		82.258	0.810	0.823	0.851	0.415	0.814
Weighted dataset with SBAWFCM	Tenfold cross-validation	100.000	1.000	1.000	1.000	1.000	1.000
Weighted dataset with SBAWKMC		100.000	1.000	1.000	1.000	1.000	1.000
Weighted dataset with SBAWMSC		100.000	1.000	1.000	1.000	1.000	1.000
In Raw dataset		83.225	0.818	0.832	0.857	0.421	0.820
Re-sampling approach to raw dataset		90.967	0.908	0.910	0.961	0.717	0.909

the combination of the *k*-NN (for *k* = 1) classifier and the attribute weighting methods are explained in Table 14. Table 15 shows the obtained test results in the classification of Parkinson disease dataset based on the various combinations of attribute weighting methods (SBAWFCM, SBAWKMC, and SBAWMSC) and random subsampling with random forests classifier with 50–50% training–testing holdout, 60–40% training–testing holdout, and tenfold cross-validation.

The results obtained in the hepatitis disease dataset, another unbalanced medical dataset, are given in the following tables. Table 16 depicts the obtained test results in the classification of hepatitis disease dataset based on the various combinations of attribute weighting methods

Table 20 Obtained test results in the classification of Pima Indians disease dataset based on the various combinations of attribute weighting methods (SBAWFCM, SBAWKMC, and SBAWMSC) and random subsampling with LDA classifier with 50–50% training–testing holdout, 60–40% training–testing holdout, and tenfold cross-validation

Main folder name	Data partition	Accuracy	Precision	Recall	AUC	κ value	*F*-measure
Weighted dataset with SBAWFCM	50–50% train–test partition	84.375	0.841	0.844	0.905	0.633	0.841
Weighted dataset with SBAWKMC		89.322	0.894	0.893	0.916	0.745	0.890
Weighted dataset with SBAWMSC		94.010	0.943	0.940	0.986	0.858	0.939
In Raw dataset		76.302	0.756	0.763	0.821	0.438	0.757
Re-sampling approach to raw dataset		80.468	0.800	0.805	0.849	0.526	0.797
Weighted dataset with SBAWFCM	60–40% train–test partition	84.364	0.854	0.844	0.896	0.624	0.839
Weighted dataset with SBAWKMC		89.250	0.895	0.893	0.903	0.739	0.889
Weighted dataset with SBAWMSC		93.811	0.942	0.938	0.985	0.851	0.936
In Raw dataset		79.804	0.793	0.798	0.843	0.509	0.790
Re-sampling approach to raw dataset		82.084	0.817	0.821	0.863	0.568	0.815
Weighted dataset with SBAWFCM	Tenfold cross-validation	84.114	0.839	0.841	0.906	0.638	0.838
Weighted dataset with SBAWKMC		87.760	0.878	0.878	0.926	0.720	0.875
Weighted dataset with SBAWMSC		94.531	0.947	0.945	0.982	0.876	0.944
In Raw dataset		77.474	0.770	0.775	0.832	0.478	0.767
Re-sampling approach to raw dataset		79.036	0.786	0.790	0.856	0.524	0.787

Table 21 Obtained test results in the classification of Pima Indians disease dataset based on the various combinations of attribute weighting methods (SBAWFCM, SBAWKMC, and SBAWMSC) and random subsampling with SVM classifier with 50–50% training–testing holdout, 60–40% training–testing holdout, and tenfold cross-validation

Main folder name	Data partition	Accuracy	Precision	Recall	AUC	κ value	*F*-measure
Weighted dataset with SBAWFCM	50–50% train–test partition	85.156	0.849	0.852	0.815	0.651	0.849
Weighted dataset with SBAWKMC		88.802	0.888	0.888	0.851	0.733	0.885
Weighted dataset with SBAWMSC		93.750	0.939	0.938	0.912	0.852	0.936
In Raw dataset		76.822	0.760	0.768	0.706	0.438	0.759
Re-sampling approach to raw dataset		80.729	0.803	0.807	0.750	0.532	0.799
Weighted dataset with SBAWFCM	60–40% train–test partition	84.364	0.841	0.844	0.797	0.624	0.839
Weighted dataset with SBAWKMC		90.228	0.905	0.902	0.862	0.763	0.899
Weighted dataset with SBAWMSC		93.811	0.941	0.938	0.909	0.852	0.937
In Raw dataset		79.804	0.793	0.798	0.740	0.509	0.790
Re-sampling approach to raw dataset		82.410	0.821	0.824	0.772	0.575	0.818
Weighted dataset with SBAWFCM	Tenfold cross-validation	83.854	0.838	0.839	0.801	0.628	0.834
Weighted dataset with SBAWKMC		89.062	0.892	0.891	0.861	0.750	0.888
Weighted dataset with SBAWMSC		95.052	0.951	0.951	0.936	0.889	0.950
In Raw dataset		76.562	0.760	0.766	0.711	0.449	0.755
Re-sampling approach to raw dataset		78.906	0.785	0.789	0.749	0.517	0.784

(SBAWFCM, SBAWKMC, and SBAWMSC) and random subsampling with LDA classifier with 50–50% training–testing holdout, 60–40% training–testing holdout, and tenfold cross-validation. Table 17 gives the obtained test results for the combinations of SVM classifier and weighting methods. In Table 18, the obtained test results for the combinations of *k*-NN classifier and weighting methods are given. Table 19 presents the obtained test results in the classification of hepatitis disease dataset based on the various combinations of attribute weighting methods (SBAWFCM, SBAWKMC, and SBAWMSC) and random subsampling with random forests classifier with

Table 22 Obtained test results in the classification of Pima Indians disease dataset based on the various combinations of attribute weighting methods (SBAWFCM, SBAWKMC, and SBAWMSC) and random subsampling with *k*-NN (for $k = 1$) classifier with 50–50% training–testing holdout, 60–40% training–testing holdout, and tenfold cross-validation

Main folder name	Data partition	Accuracy	Precision	Recall	AUC	κ value	*F*-measure
Weighted dataset with SBAWFCM	50–50% train–test partition	84.114	0.841	0.841	0.818	0.637	0.841
Weighted dataset with SBAWKMC		90.364	0.903	0.904	0.881	0.776	0.903
Weighted dataset with SBAWMSC		94.531	0.873	0.945	0.945	0.930	0.945
In Raw dataset		72.656	0.725	0.727	0.686	0.373	0.726
Re-sampling approach to raw dataset		84.895	0.847	0.849	0.812	0.651	0.848
Weighted dataset with SBAWFCM	60–40% train–test partition	83.387	0.833	0.834	0.806	0.616	0.833
Weighted dataset with SBAWKMC		90.228	0.902	0.902	0.875	0.770	0.901
Weighted dataset with SBAWMSC		95.765	0.958	0.958	0.942	0.901	0.957
In Raw dataset		73.615	0.734	0.736	0.694	0.391	0.735
Re-sampling approach to raw dataset		84.690	0.844	0.847	0.807	0.637	0.844
Weighted dataset with SBAWFCM	Tenfold cross-validation	82.942	0.828	0.829	0.812	0.620	0.828
Weighted dataset with SBAWKMC		90.104	0.900	0.901	0.892	0.779	0.900
Weighted dataset with SBAWMSC		94.270	0.943	0.943	0.930	0.872	0.942
In Raw dataset		70.052	0.697	0.701	0.658	0.331	0.698
Re-sampling approach to raw dataset		90.625	0.906	0.906	0.892	0.792	0.906

Table 23 Obtained test results in the classification of Pima Indians disease dataset based on the various combinations of attribute weighting methods (SBAWFCM, SBAWKMC, and SBAWMSC) and random subsampling with random forests classifier with 50–50% training–testing holdout, 60–40% training–testing holdout, and tenfold cross-validation

Main folder name	Data partition	Accuracy	Precision	Recall	AUC	κ value	*F*-measure
Weighted dataset with SBAWFCM	50–50% train–test partition	88.020	0.879	0.880	0.949	0.723	0.879
Weighted dataset with SBAWKMC		94.010	0.940	0.940	0.981	0.862	0.940
Weighted dataset with SBAWMSC		94.270	0.942	0.943	0.992	0.868	0.942
In Raw dataset		76.302	0.758	0.763	0.806	0.447	0.760
Re-sampling approach to raw dataset		87.239	0.877	0.872	0.933	0.688	0.867
Weighted dataset with SBAWFCM	60–40% train–test partition	88.599	0.885	0.886	0.949	0.734	0.885
Weighted dataset with SBAWKMC		95.439	0.954	0.954	0.991	0.895	0.954
Weighted dataset with SBAWMSC		95.439	0.954	0.954	0.990	0.894	0.954
In Raw dataset		79.153	0.790	0.792	0.841	0.520	0.791
Re-sampling approach to raw dataset		88.599	0.893	0.886	0.949	0.719	0.880
Weighted dataset with SBAWFCM	Tenfold cross-validation	87.890	0.878	0.879	0.953	0.728	0.878
Weighted dataset with SBAWKMC		94.010	0.940	0.940	0.986	0.867	0.940
Weighted dataset with SBAWMSC		96.614	0.966	0.966	0.994	0.925	0.966
In Raw dataset		76.171	0.756	0.762	0.825	0.458	0.757
Re-sampling approach to raw dataset		91.666	0.917	0.917	0.959	0.813	0.916

50–50% training–testing holdout, 60–40% training–testing holdout, and tenfold cross-validation.

The classification results of the Pima Indians disease dataset are given in the following tables. Table 20 presents the obtained test results in the classification of Pima Indians disease dataset based on the various combinations of attribute weighting methods (SBAWFCM, SBAWKMC, and SBAWMSC) and random subsampling with LDA classifier with 50–50% training–testing holdout, 60–40% training–testing holdout, and tenfold cross-validation. Table 21 shows the test results of combinations based on attribute weighting methods and SVM classifier. In

Table 24 Obtained test results in the classification of SPECT heart dataset based on the various combinations of attribute weighting methods (SBAWFCM, SBAWKMC, and SBAWMSC) and random subsampling with LDA classifier with 50–50% training–testing holdout, 60–40% training–testing holdout, and tenfold cross-validation

Main folder name	Data partition	Accuracy	Precision	Recall	AUC	κ value	*F*-measure
Weighted dataset with SBAWFCM	50–50% train–test partition	79.699	0.817	0.797	0.843	0.440	0.805
Weighted dataset with SBAWKMC		80.451	0.848	0.805	0.879	0.503	0.817
Weighted dataset with SBAWMSC		80.451	0.786	0.805	0.675	0.343	0.792
In Raw dataset		74.436	0.758	0.744	0.767	0.269	0.750
Re-sampling approach to raw dataset		82.706	0.812	0.827	0.886	0.410	0.814
Weighted dataset with SBAWFCM	60–40% train–test partition	78.504	0.802	0.785	0.847	0.424	0.792
Weighted dataset with SBAWKMC		80.373	0.835	0.804	0.881	0.502	0.814
Weighted dataset with SBAWMSC		80.373	0.784	0.804	0.667	0.349	0.786
In Raw dataset		75.700	0.772	0.757	0.781	0.340	0.763
Re-sampling approach to raw dataset		81.308	0.795	0.813	0.863	0.369	0.794
Weighted dataset with SBAWFCM	Tenfold cross-validation	85.393	0.851	0.854	0.884	0.544	0.852
Weighted dataset with SBAWKMC		86.516	0.869	0.865	0.921	0.598	0.867
Weighted dataset with SBAWMSC		81.647	0.794	0.816	0.649	0.252	0.777
In Raw dataset		81.273	0.803	0.813	0.816	0.394	0.807
Re-sampling approach to raw dataset		84.269	0.835	0.843	0.850	0.491	0.838

Table 25 Obtained test results in the classification of SPECT heart dataset based on the various combinations of attribute weighting methods (SBAWFCM, SBAWKMC, and SBAWMSC) and random subsampling with SVM classifier with 50–50% training–testing holdout, 60–40% training–testing holdout, and tenfold cross-validation

Main folder name	Data partition	Accuracy	Precision	Recall	AUC	κ value	*F*-measure
Weighted dataset with SBAWFCM	50–50% train–test partition	82.706	0.834	0.827	0.760	0.499	0.830
Weighted dataset with SBAWKMC		84.210	0.872	0.842	0.835	0.584	0.851
Weighted dataset with SBAWMSC		82.706	0.858	0.827	0.589	0.255	0.775
In Raw dataset		78.947	0.813	0.789	0.736	0.426	0.789
Re-sampling approach to raw dataset		82.706	0.834	0.827	0.760	0.499	0.830
Weighted dataset with SBAWFCM	60–40% train–test partition	83.177	0.827	0.832	0.743	0.501	0.829
Weighted dataset with SBAWKMC		85.046	0.867	0.850	0.830	0.605	0.856
Weighted dataset with SBAWMSC		80.373	0.843	0.804	0.563	0.181	0.738
In Raw dataset		82.243	0.825	0.822	0.752	0.497	0.824
Re-sampling approach to raw dataset		81.308	0.803	0.813	0.702	0.429	0.807
Weighted dataset with SBAWFCM	Tenfold cross-validation	87.265	0.871	0.873	0.799	0.605	0.872
Weighted dataset with SBAWKMC		89.513	0.898	0.895	0.853	0.687	0.896
Weighted dataset with SBAWMSC		82.022	0.853	0.820	0.760	0.188	0.564
In Raw dataset		83.895	0.829	0.839	0.717	0.468	0.832
Re-sampling approach to raw dataset		83.146	0.826	0.831	0.726	0.466	0.828

Table 26 Obtained test results in the classification of SPECT heart dataset based on the various combinations of attribute weighting methods (SBAWFCM, SBAWKMC, and SBAWMSC) and random subsampling with *k*-NN classifier with 50–50% training–testing holdout, 60–40% training–testing holdout, and tenfold cross-validation

Main folder name	Data partition	Accuracy	Precision	Recall	AUC	κ value	*F*-measure
Weighted dataset with SBAWFCM	50–50% train–test partition	78.947	0.827	0.789	0.760	0.452	0.802
Weighted dataset with SBAWKMC		83.458	0.875	0.835	0.859	0.579	0.845
Weighted dataset with SBAWMSC		87.969	0.874	0.880	0.777	0.607	0.874
In Raw dataset		74.436	0.786	0.744	0.715	0.335	0.759
Re-sampling approach to raw dataset		82.706	0.851	0.827	0.823	0.534	0.835
Weighted dataset with SBAWFCM	60–40% train–test partition	83.177	0.850	0.832	0.847	0.556	0.838
Weighted dataset with SBAWKMC		86.915	0.885	0.869	0.899	0.654	0.874
Weighted dataset with SBAWMSC		86.915	0.863	0.869	0.778	0.587	0.862
In Raw dataset		75.700	0.780	0.757	0.726	0.358	0.766
Re-sampling approach to raw dataset		83.177	0.843	0.832	0.865	0.543	0.836
Weighted dataset with SBAWFCM	Tenfold cross-validation	83.520	0.862	0.835	0.838	0.556	0.844
Weighted dataset with SBAWKMC		88.015	0.913	0.880	0.896	0.688	0.888
Weighted dataset with SBAWMSC		89.138	0.888	0.891	0.827	0.651	0.889
In Raw dataset		76.404	0.785	0.764	0.731	0.336	0.773
Re-sampling approach to raw dataset		89.138	0.907	0.891	0.932	0.698	0.896

Table 27 Obtained test results in the classification of SPECT heart dataset based on the various combinations of attribute weighting methods (SBAWFCM, SBAWKMC, and SBAWMSC) and random subsampling with random forests classifier with 50–50% training–testing holdout, 60–40% training–testing holdout, and tenfold cross-validation

Main folder name	Data partition	Accuracy	Precision	Recall	AUC	κ value	*F*-measure
Weighted dataset with SBAWFCM	50–50% train–test partition	94.736	0.958	0.947	0.980	0.855	0.949
Weighted dataset with SBAWKMC		94.736	0.958	0.947	0.990	0.855	0.949
Weighted dataset with SBAWMSC		90.225	0.901	0.902	0.962	0.702	0.902
In Raw dataset		80.451	0.800	0.805	0.776	0.396	0.802
Re-sampling approach to raw dataset		85.714	0.855	0.857	0.890	0.564	0.856
Weighted dataset with SBAWFCM	60–40% train–test partition	95.327	0.961	0.953	0.986	0.875	0.955
Weighted dataset with SBAWKMC		95.327	0.961	0.953	0.991	0.875	0.955
Weighted dataset with SBAWMSC		89.719	0.894	0.897	0.969	0.690	0.895
In Raw dataset		81.308	0.799	0.813	0.775	0.410	0.803
Re-sampling approach to raw dataset		85.981	0.853	0.860	0.882	0.564	0.854
Weighted dataset with SBAWFCM	Tenfold cross-validation	96.254	0.968	0.963	0.989	0.892	0.964
Weighted dataset with SBAWKMC		96.629	0.971	0.966	0.990	0.902	0.967
Weighted dataset with SBAWMSC		95.131	0.953	0.951	0.980	0.854	0.952
In Raw dataset		82.771	0.817	0.828	0.798	0.435	0.821
Re-sampling approach to raw dataset		91.011	0.913	0.910	0.943	0.732	0.911

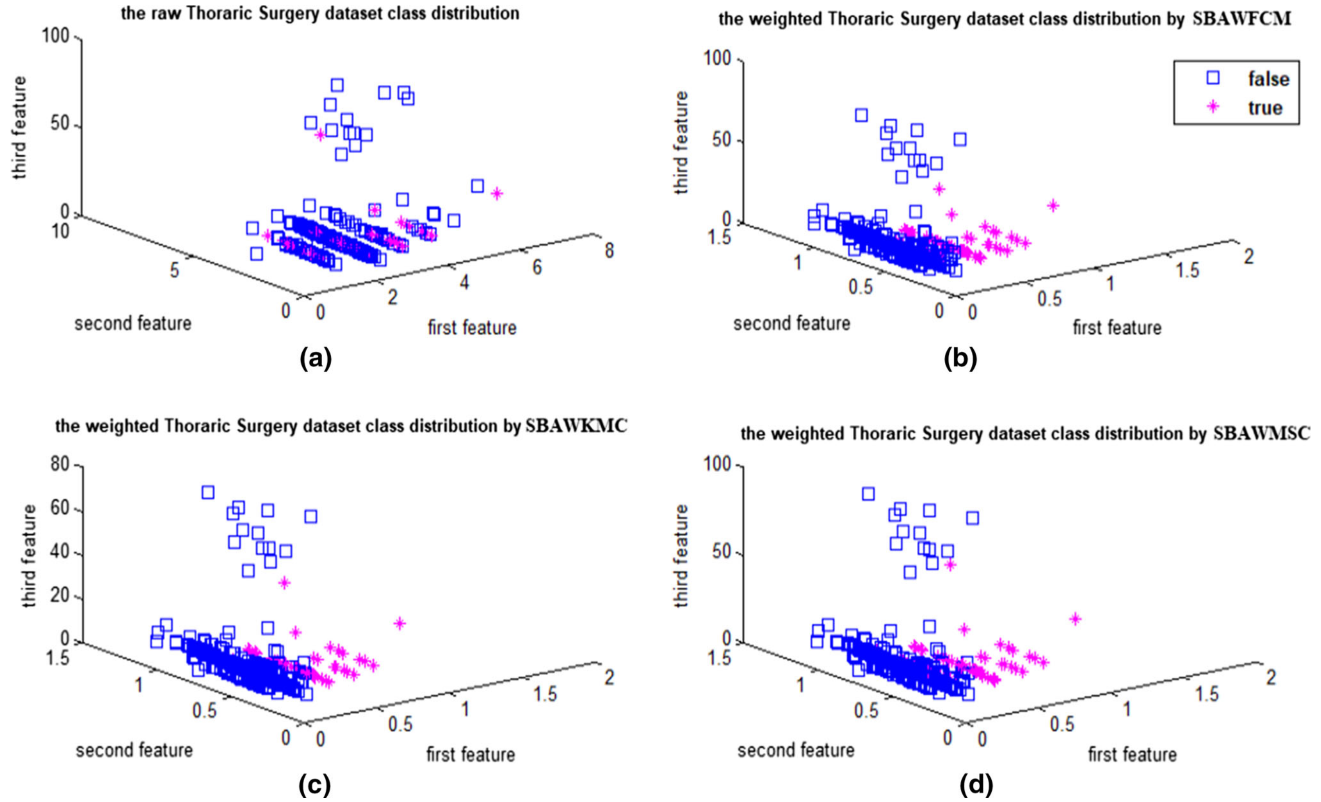

Fig. 15 Class distributions of both raw and weighted thoracic surgery datasets according to the first three attributes of the dataset. **a** The class distribution of raw thoracic surgery dataset, **b** the class distribution of weighted thoracic surgery dataset by SBAWFCM, **c** the class distribution of weighted thoracic surgery dataset by SBAWKMC, and **d** the class distribution of weighted thoracic surgery dataset by SBAWMSC

Table 22, the test results of combinations based on attribute weighting methods and *k*-NN classifier are shown. Table 23 gives the results of combinations based on attribute weighting methods and random forests classifier.

The results obtained in the classification of the SPECT heart dataset, which cannot be separated linearly and have an unbalanced data distribution, are given in the following tables. Table 24 shows the obtained test results in the classification of SPECT heart dataset based on the various combinations of attribute weighting methods (SBAWFCM, SBAWKMC, and SBAWMSC) and random subsampling with LDA classifier with 50–50% training–testing holdout, 60–40% training–testing holdout, and tenfold cross-validation. Table 25 denotes the classification test performance results of attribute weighting methods and SVM classifier. Table 26 presents the test results of combinations based on attribute weighting methods and *k*-NN classifier. In Table 27, the obtained test results in the classification of SPECT heart dataset based on the various combinations of attribute weighting methods (SBAWFCM, SBAWKMC, and SBAWMSC) and random subsampling with random forests classifier with 50–50% training–testing holdout, 60–40% training–testing holdout, and tenfold cross-validation are shown.

The class distributions of the medical datasets used in this study have been given in order to show the effect of the proposed attribute weighting methods on the datasets having imbalanced data distribution. Figure 15 shows the class data distributions and class distributions for both raw and weighted datasets according to the first three attributes of thoracic surgery dataset consisting of 16 attributes. In the two-class thoracic surgery dataset, it seems difficult to distinguish one class from another. The classification of this dataset requires both a higher computational cost and a good classifier. In the weighted thoracic surgery datasets with SBAWKMC, SBAWFCM, and SBAWMSC, the data distribution has become a more linear. In this way, the medical datasets classified with less computational cost and better performance have been obtained.

The Parkinson disease dataset consists of 22 attributes. The class distributions of the raw and weighted Parkinson disease datasets have been given to demonstrate and prove how the proposed attribute weighting methods improve the performance. Figure 16 shows the class distributions for both raw and weighted PD datasets according to the first

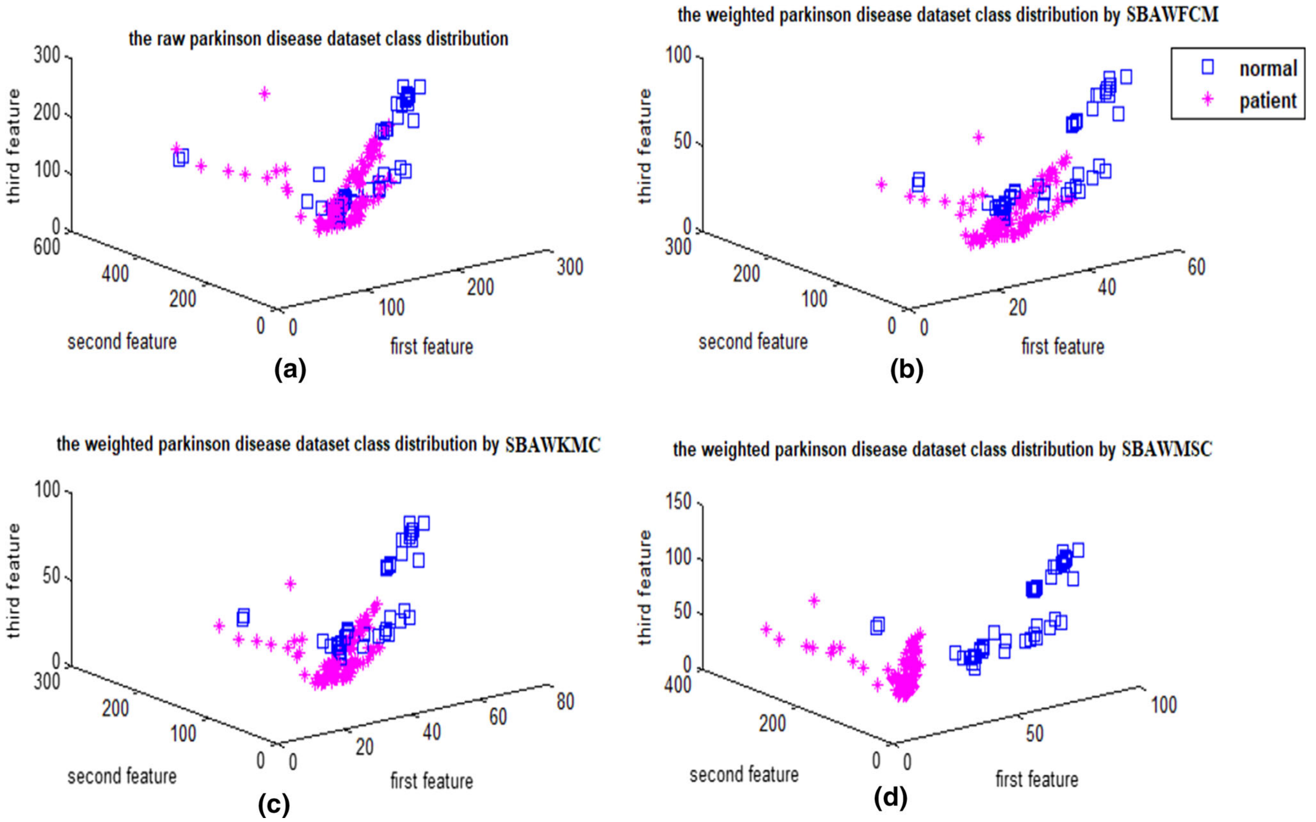

Fig. 16 Class distributions of both raw and weighted Parkinson disease datasets according to the first three attributes of the dataset. **a** The class distribution of raw Parkinson disease dataset, **b** the class distribution of weighted Parkinson disease dataset by SBAWFCM, **c** the class distribution of weighted Parkinson disease dataset by SBAWKMC, and **d** the class distribution of weighted Parkinson disease dataset by SBAWMSC

three attributes in the Parkinson disease dataset. In the two-class Parkinson disease dataset, there have been nested the classes as seen from the raw PD dataset. The classification of this PD dataset requires both a higher computational cost and a good classifier. In the case of the Parkinson disease datasets weighted with SBAWKMC, SBAWFCM, and SBAWMSC, the class distribution has become a more linear. In this way, less computational cost and better classification performance have been obtained in the classification of weighted PD dataset.

The SPECT heart disease dataset consists of 22 pixel attributes (comprising of 0 and 1). Figure 17 shows the class distributions for both raw and weighted SPECT heart datasets according to the first three attributes in the SPECT heart disease dataset. As can be seen from these graphs, the discrimination between classes in the SPECT dataset has been increased and then high classification performance has been obtained.

The hepatitis disease dataset consists of 19 numeric and categorical attributes. Figure 18 shows the class distributions for both raw and weighted hepatitis disease datasets according to the first three attributes in the hepatitis disease dataset. As can be seen from these graphs, the discrimination between classes in the hepatitis disease dataset has been increased and then high classification performance has been obtained.

The hepatitis disease dataset consists of 19 numeric and categorical attributes. Figure 18 shows the class distributions for both raw and weighted hepatitis disease datasets according to the first three attributes in the hepatitis disease dataset. As can be seen from these graphs, the discrimination between classes in the hepatitis disease dataset has been increased and then high classification performance has been obtained.

Figure 19 presents the class distributions of both raw and weighted Pima Indians disease datasets according to

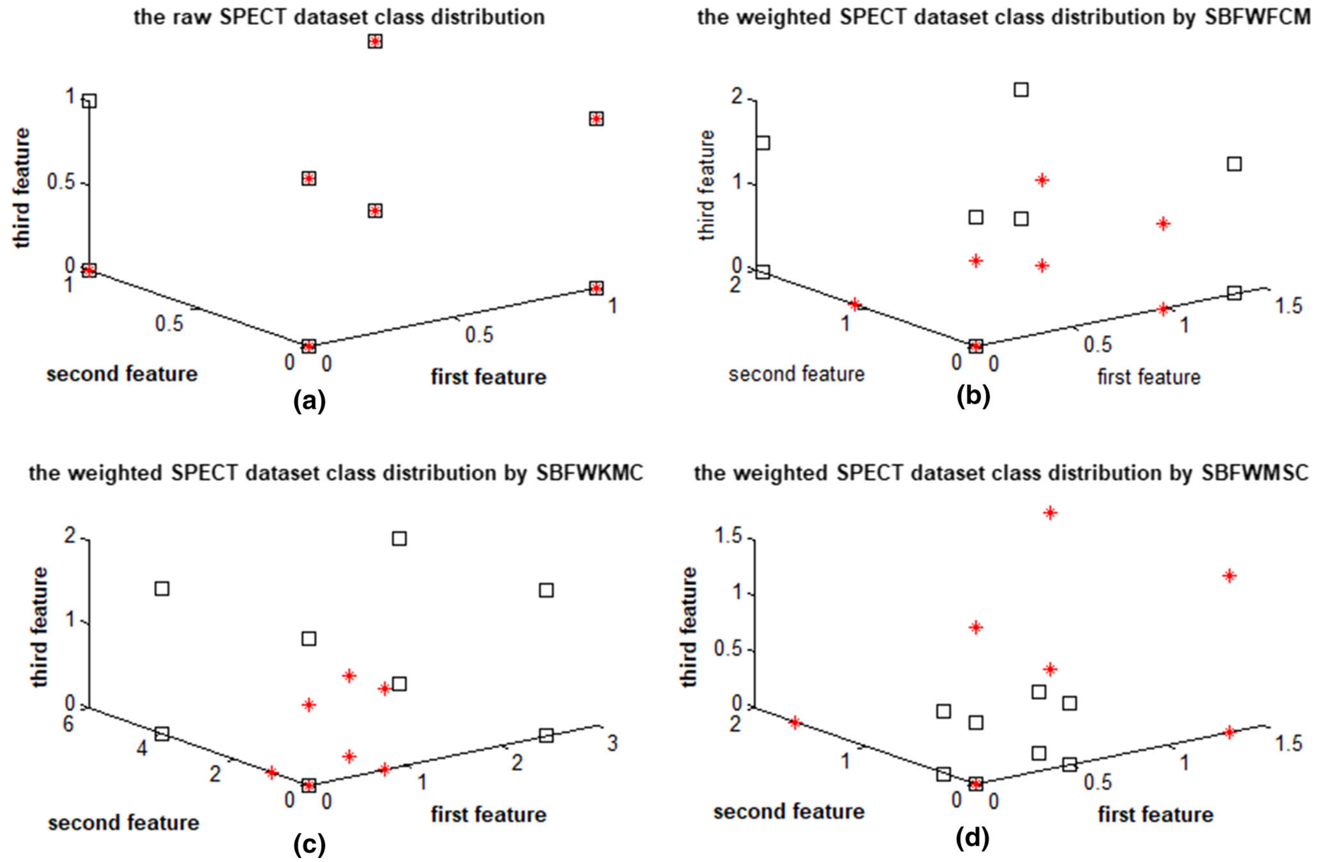

Fig. 17 Class distributions of both raw and weighted SPECT heart datasets according to the first three attributes of the dataset. **a** The class distribution of raw SPECT heart dataset, **b** the class distribution of weighted SPECT heart dataset by SBAWFCM, **c** the class distribution of weighted SPECT heart dataset by SBAWKMC, and **d** the class distribution of weighted SPECT heart dataset by SBAWMSC

the first three attributes of the dataset: (a) the class distribution of raw Pima Indians disease dataset, (b) the class distribution of weighted Pima Indians disease dataset by SBAWFCM, (c) the class distribution of weighted Pima Indians disease dataset by SBAWKMC, and d) the class distribution of weighted Pima Indians disease dataset by SBAWMSC. In this figure, the weighted Pima Indians datasets have been converted to linearly separable dataset by means of the attribute weighting methods and then the discrimination ability between two classes in the Pima Indians dataset has been increased.

As can be seen from these figures, in the original datasets, from the distribution of class distribution according to the first three attributes of the dataset, the data points appear to be scattered in many different areas. After applying the weighting method to the datasets, the class data distributions seem to be closer to each other. In this case, the distinction between the classes increases as for the distribution within the class decreases.

The results have shown that the proposed attribute weighting methods including SBAWFCM, SBAWKMC, and SBAWMSC could be used as a promising method in the classification and distinguishing medical data clusters with nonlinear separable and imbalanced distribution.

5 Conclusions and future work

In this study, attribute weighting methods based on similarity as a new preprocessing method have proposed in the classification of Parkinson, hepatitis, Pima Indians, SPECT heart, and thoracic surgery medical datasets, which have a

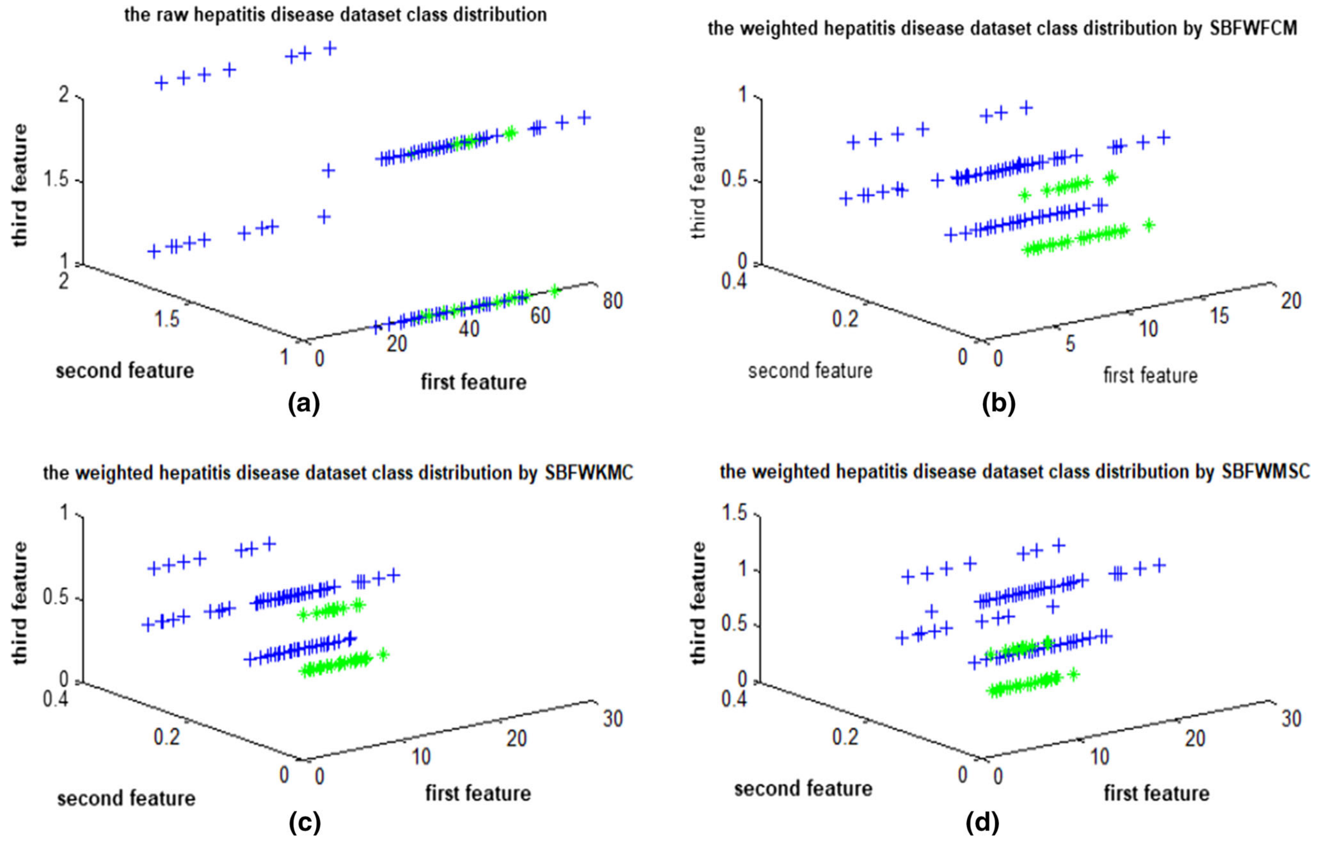

Fig. 18 Class distributions of both raw and weighted hepatitis disease datasets according to the first three attributes of the dataset. **a** The class distribution of raw hepatitis disease dataset, **b** the class distribution of weighted hepatitis disease dataset by SBAWFCM, **c** the class distribution of weighted hepatitis disease dataset by SBAWKMC, and **d** the class distribution of weighted hepatitis disease dataset by SBAWMSC

nonlinearly separable and imbalanced data distribution. *k*-means clustering, fuzzy *c*-means clustering, and mean shift clustering algorithms have been used as clustering algorithms in the proposed attribute weighting methods. The proposed method is based on the similarity between the data center cluster and the related attribute to the class in the dataset. If the similarity rate is low, the weight coefficient will be high values. If the similarity rate is high, the weight coefficient will be low values. To compare with other methods in the literature, the random subsampling has been used to handle the imbalanced dataset classification. After attribute weighting process, four classification algorithms including linear discriminant analysis (LDA), *k*-NN (*k*-nearest neighbor) classifier, support vector machine (SVM), and random forest classifier have been used to classify medical datasets. Various data preprocessing methods have been proposed to classify imbalanced datasets, but there are some problems in using these methods because they break the original structure of the dataset. The proposed attribute weighting methods could be confidently used in the signal and image classification. As the future work, the proposed methods could be applied as online in the hospital and healthcare. Also, in addition to used clustering methods, the optimization-based weighting models could be improved and then applied to this problem.

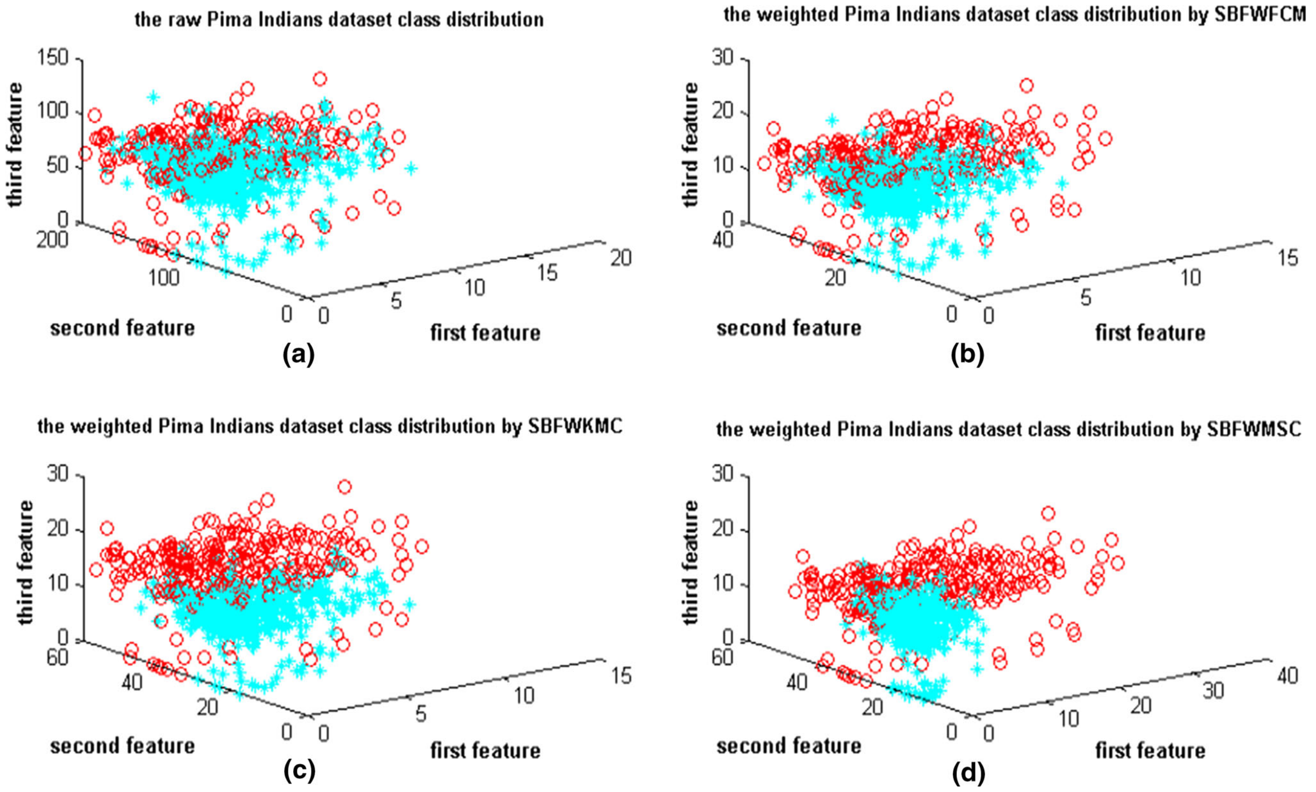

Fig. 19 Class distributions of both raw and weighted Pima Indians disease datasets according to the first three attributes of the dataset. **a** The class distribution of raw Pima Indians disease dataset, **b** the class distribution of weighted Pima Indians disease dataset by SBAWFCM, **c** the class distribution of weighted Pima Indians disease dataset by SBAWKMC, and **d** the class distribution of weighted Pima Indians disease dataset by SBAWMSC

Compliance with ethical standards

Conflict of interest We declare that there is no conflict of interest in anywhere.

References

1. Longadge R, Dongre SS, Malik L (2013) Class imbalance problem in data mining: review. Int J Comput Sci Netw (IJCSN) 2(1):83–87
2. https://classeval.wordpress.com. Accessed Feb 2018
3. http://sebastianraschka.com/Articles/2014_kernel_pca.html. Accessed Feb 2018
4. https://arxiv.org/ftp/arxiv/papers/1305/1305.1707. Accessed Feb 2018
5. Gong J, Kim H (2017) RHSBoost: improving classification performance in imbalance data. Comput Stat Data Anal 111:1–13
6. Shilaskar S, Ghato A, Chatur P (2017) Medical decision support system for extremely imbalanced datasets. Inf Sci 384:205–219
7. Zhang J, Xiao W, Li Y, Zhang S, Yang W (2017) Class-specific cost regulation extreme learning machine for imbalanced classification. Neurocomputing 261:70–82
8. UCI machine learning repository (2018) https://archive.ics.uci.edu/ml. Accessed Feb 2018
9. Short RD, Fukunaga K (1981) The optimal distance measure for nearest neighbor classification. IEEE Trans Inf Theory 27:622–627. https://doi.org/10.1109/TIT.1981.1056403
10. Weinberger KQ, Saul LK (2009) Distance metric learning for large margin nearest neighbor classification. J Mach Learn Res 10:207–244
11. Cost S, Salzberg S (1993) A weighted nearest neighbor algorithm for learning with symbolic attributes. Mach Learn 10:57–78. https://doi.org/10.1007/BF00993481
12. Breiman L (2001) Random forests. Mach Learn 45:5–32. https://doi.org/10.1023/A:1010933404324
13. Zhang Z (2014) Too much covariates in a multivariable model may cause the problem of overfitting. J Thorac Dis 6:E196–E197
14. Hastie T, Tibshirani R, Friedman JH (2003) The elements of statistical learning. Springer, New York
15. McLachlan GJ (2004) Discriminant analysis and statistical pattern recognition. Wiley, Hoboken
16. Shen R, Ghosh D, Chinnaiyan A, Meng Z (2006) Eigengene-based linear discriminant model for tumor classification using gene expression microarray data. Bioinformatics 22(21):2635–2642
17. Cortes C, Vapnik V (1995) Support-vector network. Mach Learn 20(3):273–297
18. Vapnik V (2014) Invited speaker. In: IPMU information processing and management. 15th international conference on information processing and management of uncertainty in knowledge-based systems, IPMU'2014, Montpellier, France,15–19 July 2014
19. Savitt JM, Dawson VL, Dawson TM (2006) Diagnosis and treatment of Parkinson disease: molecules to medicine. J Clin Investig 116(7):1744–1754. https://doi.org/10.1172/JCI29178
20. Levine CB, Fahrbach KR, Siderowf AD, Estok RP, Ludensky VM, Ross SD (2003) Diagnosis and treatment of Parkinson's

disease: a systematic review of the literature. Evid Rep Technol Assess 57:1–4 **(Summary)**

21. Yuvaraj R, Murugappan M, Acharya UR, Adeli H, Ibrahim NM, Mesquita E (2016) Brain functional connectivity patterns for emotional state classification in Parkinson's disease patients without dementia. Behav Brain Res 298((Pt B)):248–260. https://doi.org/10.1016/j.bbr.2015.10.036
22. http://www.frank-dieterle.de/phd/2_4_3.html. Accessed Feb 2018
23. Clark M (2015) An introduction to machine learning with applications in R, Lecture notes. University of Notre Dame, Notre Dame