J Supercomput (2016) 72:3708–3728
DOI 10.1007/s11227-015-1541-6

Improving the classification performance of biological imbalanced datasets by swarm optimization algorithms

Jinyan Li[1] · Simon Fong[1] · Sabah Mohammed[2] · Jinan Fiaidhi[2]

Published online: 16 November 2015

Abstract Classification which is a popular supervised machine learning method has many applications in computational biology, where data samples are automatically categorized into predefined labels with the aid of data mining. Often the training samples contain very few instances of interest (e.g., medical anomalies, rare disease in a population, and unusual syndromes, etc.), but many normal instances. Such imbalanced ratio of data distributions among the target labels hampers the efficacy of classification algorithms, because the induced model has not been trained with sufficient amount of instances of the interesting label(s), but overwhelmed with ordinary training records. Traditional remedies attempt to rebalance the data distributions of the target classes, by inflating the interesting instances artificially, reducing the majority of the common instances or a combination of both. Though the fundamental concept is effective, there is no clear guideline on how to strike a balance between fabricating the rare samples and reducing the norms, with the purpose of maximizing the classification accuracy. In this paper, an optimization model using different swarm strategies (Bat-inspired algorithm and PSO) is proposed for adaptively balancing the

✉ Simon Fong
ccfong@umac.mo

Jinyan Li
yb47432@umac.mo

Sabah Mohammed
sabah.mohammed@lakeheadu.ca

Jinan Fiaidhi
jfiaidhi@lakeheadu.ca

1 Department of Computer and Information Science, University of Macau, Taipa, Macau SAR

2 Department of Computer Science, Lakehead University, Taipa, Macau SAR

increase/decrease of the class distribution, depending on the properties of the biological datasets. The optimization is extended for achieving the highest possible accuracy and Kappa statistics at the same time as well. The optimization model is tested on five imbalanced medical datasets, which are sourced from lung surgery logs and virtual screening of bioassay data. Computer simulation results show that the proposed optimization model outperforms other class balancing methods in medical data classification.

Keywords Imbalanced biological data · Medical classification · Swarm algorithm · Parameter optimization

1 Introduction

In real life, the problem of imbalanced data distribution is not uncommon because the data instances of interest are often rare in quantity. Data collections of medical anomalies, rare diseases in a population, and unusual syndromes often result in highly imbalanced proportions of normal and abnormal instances. For example, the ratio of unsuccessful and successful cases in lung resections for treating primary lung cancer at the Medical University of Wroclaw in 2013 is 1:7, while the ratio of active bioactivity screens of chemical substances described in PubChem BioAssay Database is 1:70. If a classification model were to be built for automatic recognition or predicting the outcome to which a new sample should belong to, such imbalanced dataset would be used for supervised model training leading to a performance issue. The accuracy of the classification model, however, induced from the imbalanced dataset would be far from acceptable.

This phenomenon is due to the limitation of the underlying learning mechanism in the designs of the traditional classification algorithms. When training samples are loaded into the induction process, the learning method does not distinguish the target classes of the data samples. For example, in Greedy Search, which is a common directive for classification induction, the logic of the algorithm just takes whatever training data available and infers mapping relations between the data samples and the target classes. The learning proceeds without concerns of the target class distribution in the training data. As a result, the classification model tends to bias towards the majority class and lack of sufficient training to recognize the minority classes. The ill-trained classifier, often, would have a very high pseudo-accuracy in testing with the majority class samples; but when it comes to testing unknown samples from the minority class, the performance deteriorates badly. In this case, the under-training by the insufficient minority class samples is reflected by a low Kappa statistics value from the classifier, even though the accuracy may be high.

This is generally known as the ‘imbalanced dataset problem’ which received a lot of research attentions from both data mining and computational biology research communities. To tackle this problem, various computational attempts have been studied and applied to re-balance the imbalanced data distribution in the training samples over the target class and non-target classes.

2 Background and related work

Two popular solutions have been proposed and studied, respectively, in data pre-processing and the modification of algorithms at code level. The under-sampling technology [1] and over-sampling [2] technology belong to the first type of solution, data pre-processing. It works simply by changing the distribution of imbalanced data set to improve the classification performance of the subsequent classification model. The main performance issue is that it does not only need to eliminate a lot of noise information to significantly reduce the dataset's imbalanced degree, but it also tries to ensure the minimum information are retrained in the original dataset [3]. On the other hand, it is based on the cost-sensitive learning [4,5], SVM [6], Boosting [7], or classifier ensemble [8] like SMOTEBoost [9] to modify the classification method. One classical implementation is called SMOTE (synthetic minority over-sampling technique) [11] which is a commonly used over-sampling technique. The basic idea is through applying some artificial data synthesis, extra samples from the minority class are generated in order to subside the categories imbalance. But the research question is, how much artificial data from the minority class should be synthesized in order to achieve the maximum possible accuracy while at the same time some substantial Kappa value is assured? There are some key parameters whose values need to be optimized in this balancing process.

Assuming the over-sampling rate is K, the number of minority class equal to M, and each minority class can be signified as x_i $(i = 1, 2, 3 \ldots M)$ which belongs to $S_{\min}$, then every x_i searches out K neighbors of minority class from the minority class, the algorithm will random set a x_t from the K neighbors, finally it will synthesize new data, according to Eq. (1).

$$x_{new} = x_i + rand\,[0, 1]^* \,(x_t - x_i) \tag{1}$$

The function *rand* [0,1] produces a random number in the interval between 0 and 1. Once the artificial data are generated, they are added back to the training dataset, thereby updating a new version of training dataset with the class distribution modified. This synthesis process may repeat several times until certain improvement on the classification performance can be seen.

Figures 1 and 2 illustrate a sample of SMOTE operation, when the K equal to 7, and how the x_i synthesized a minority class data. It is easy to observe how to select a suitable value of K when the data dimension is low, at, e.g., 2 or 3 dimensions. However, for very high-dimensional data with many attributes, knowing the ideal proportion of extra minority class data to be synthesized is a computational challenge in using SMOTE.

Succinctly, though the concept is primitive, there are some unsolved issues associated with such balancing technique like SMOTE. It is not known about what values of the parameter variables and how many times the generation process shall repeat for yielding a classifier that produces the best performance. Please note that SMOTE runs once off; a certain amount of synthesis data is generated each time when the SMOTE function is invoked. As a research objective, we aim to find an optimal pair of S and K values which are key parameters directly influencing the data synthesis and the end

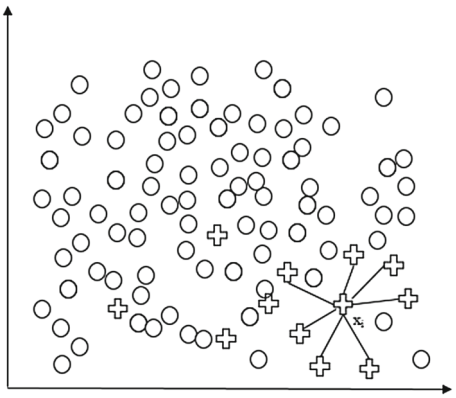

Fig. 1 SMOTE, the minority class data x_i with $K = 7$

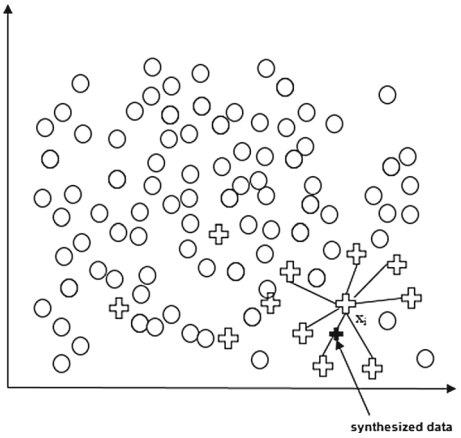

Fig. 2 Synthesized data of x_i

results—the efficacy of the classifier. To this end, swarm optimization algorithms are proposed to find a suitable pair of S and K parameter values for rebalancing the class distribution in the training data, ensuring a reliable classifier as a result.

The main advantage between using swarm optimization and computational brute-force in tuning up the S and K, is on the speed efficiency as well as making use of heuristics during reiteration. Whenever new training data are loaded, the swarm continues to move and improve again in the search space. The heuristics information help make the supervised training adaptive to the new target class distribution, whenever the underlying class distribution changes by the arrival of new data.

In the experiments which are reported in this paper, we consider using the classifier of neural network and decision tree to do the verification with two different metaheuristic optimization algorithms, namely particle swarm optimization (PSO) algorithm [11] and Bat-inspired algorithm (Bat) [12]. Neural network and decision tree are chosen because of their popularity in data mining, indeed. Another reason is that neural network represents a typical category of black-box supervised learning scheme (the non-linear relations between the attributes and the targets are numerical weights in the

Table 1 Confusion matrix

	Predicted class	
Actual class		
Label	Positive	Negative
Positive	True positive (TP)	False negative (FN)
Negative	False positive (FP)	True negative (TN)

hidden layers). The decision trees represent a non-black-box rule-induction learning scheme; decision rules that are readily human readable can be harvested from the resultant decision paths at the end. Furthermore, we use two different metaheuristics, PSO and Bat, respectively, as a comparison between the classical metaheuristic optimization algorithm and the contemporary, as well simple versus sophisticated designs. Search agents in PSO move according to two velocities, global and local; and Bat agents move in similar fashion, with the addition of acceleration in echolocation in their flight paths.

3 Experiment and datasets

As we all know that in the classification of imbalanced dataset with the original dataset, sometimes we can get a very good accuracy. But at this time the other performance index called Kappa statistics is very low, most of the time Kappa drops to almost zero. At times it may become a negative value depending on how imbalanced the data are. The confusion matrix which is defined in Table 1 shows the reason that because of the number of negative class takes much of the low proportion, the classifier misclassifies most if not all of them into wrong classes. That means if we use all negative class dataset as a testing dataset, the accuracy of the trained classification model will be extremely low, because the classifier was under-trained with the minority class data. Thus the classification result of high accuracy when it comes to classifying imbalanced dataset is meaningless.

According to confusion matrix, we can get the definition of the accuracy and Kappa as follow equations:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{P} + \text{N}} \tag{2}$$

$$\text{Kappa} = \frac{P_o + P_c}{1 - P_c} \tag{3}$$

$$P_o = \text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{P} + \text{N}} \tag{4}$$

$$P_c = \frac{(\text{TP} + \text{FP}) \times (\text{TP} + \text{FN}) + (\text{FN} + \text{TN}) \times (\text{FP} + \text{TN})}{(\text{P} + \text{N})^2} \tag{5}$$

Kappa is an alternative measure of computing classification performance in response to the consistency of testing dataset. Thus it is an important performance

indicator that tells us how to judge whether the classification accuracy is within a confidence level. Kappa is generally interpreted as the reliability of the classifier model. As the Kappa value is higher, the accuracy is more credible. The range of Kappa values (or just Kappa) [13] is between −1 and +1. Meanwhile there are three levels of Kappa that are used to estimate the credibility of classification accuracy:

1. Kappa $\geq$ 0.75: strong consistency, high credible accuracy.
2. 0.4 $\leq$ Kappa $<$ 0.75: the accuracy's confidence level is in generally.
3. Kappa $<$ 0.4: accuracy is incredible.

Our experiment is conducted over datasets that have binary classes. We use PSO and Bat algorithms to optimize the two parameters in artificially rebalancing the data distributions. Neural network and decision tree are the classifiers which we choose to measure and verify the objectives in terms of fitness in every generation with Swarm algorithms. In general, we only focus on one performance that is the accuracy to measure whether the two parameters are globally best. But here due to the specificity of imbalanced dataset we also need to ensure that the Kappa is also as large as possible so to ensure the accuracy's credibility. In the experiment of classification model testing, we use a stringent tenfold cross-validation to evaluate the classifier.

Figure 3 shows the logics of the extension of the SMOTE algorithm in flow chart where swarm algorithms are used to optimize the two key parameters values. Every time when the search agents (such as particles in PSO) or Bats move, it is hoped that through the classifiers of neural network and decision tree the new method can

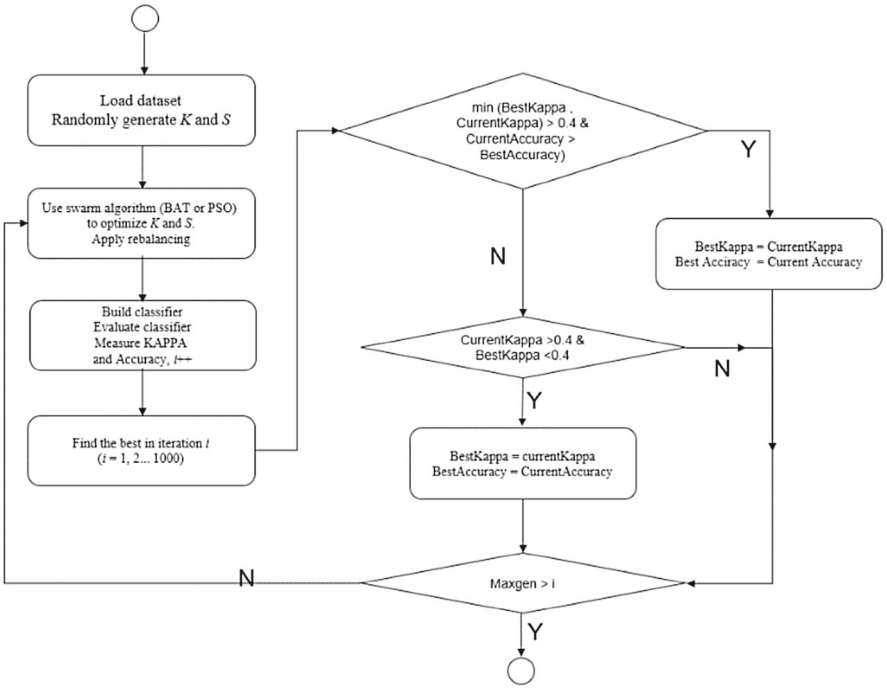

Fig. 3 Flow chart of our extended rebalancing algorithm using swarm optimization

find a local best of K and S for the sake of yielding a better performance of Kappa and accuracy. We then use the performance indicators as a measure to compare with the conditions, with the iterative processing to find the globally best K and S so to improve the values of Kappa and credibility of accuracy. In the computation, each of the parameters of S and K has its own step interval. The maximum value of S is the ratio between majority class and minority class. The minimum value of S is 1 %. The maximum and minimum values of K, respectively, are the total number of instances of the minority class and 2. For instance, if the target class of the data instances in the dataset has two labels, the number of majority class's instance is 1000, and the minority label has only 10 instances, $S\_max = 9000$ %. That means the minority class sample can grow as large as 100,000 times, and it has to increase from 20 times at least. $K\_max = 10$, and $K\_min = 2$.

Swarm-SMOTE Algorithm:

```
Specified a Meta-heuristic algorithms M (PSO/BAT/...) and a Classifier C (Neural Network/ Decision Tree/...)
Initialize the population of the M algorithm x_a (a = 1, 2, ..., n) and the other related parameters
Define the scope of K and S
//K ∈ [K_Min, K_max], S ∈ [S_Min, S_max], K is the selected Neighbor and S is the increased proportion of minority class
//data
Define the limit value of Kappa T
Load dataset
While (i < Maximum number of iteration)
    if (i = 1) // in S then
        K = Rnd (K_i)
        S = Rnd (S_i)
        // as initialize parameters of SMOTE to generate a new dataset and using C to get the Current Kappa
        //and Current Accuracy
    else
       based on the last position or solution to generate a pair of K and S
        // through the SMOTE and C to get the Current Kappa and Current Accuracy
    end if
    if (min (BestKappa, CurrentKappa) > 0.4 & CurrentAccuracy > BestAccuracy) then
        BestKappa = Current Kappa
        BestAccuracy = CurrentAccuracy
    end if
    elseif (CurrentKappa >0.4 & BestKappa <0.4 ) then
        BestKappa = Current Kappa
        BestAccuracy = CurrentAccuracy
        i = i+1
     end elseif
     back to while
end while
```

The above is the pseudo-code of swarm-SMOTE algorithm. And the principle is: in every generation of swarm optimization algorithm we set two control conditions—we

know that when Kappa equals to or greater than 0.4 (which is the minimum requirement according to the three Kappa levels), the value of accuracy is then meaningful. The two conditions for inferring a classifier with meaningful classification ability are: first of all the Kappa's value must equal or larger than 0.4, then we consider about the qualified accuracy, hence attempt to obtain the globally highest accuracy value while taking for granted that Kappa remains at least 0.4. The value 0.4 can be thought of a minimum threshold which is arbitrarily chosen. Should the user require a more robust classifier, a larger value for the minimum threshold can be used. Often, Kappa and accuracy are correlated; that means when accuracy is improved, Kappa would be likely greater than 0.4; Kappa increases while accuracy rises, and vice-versa, except in the cases of very imbalanced data being used as training dataset.

As a performance comparison benchmark, a standard class balancer algorithm is used to compare with our proposed swarm rebalancing algorithm. The experiment is tested with four imbalanced medical datasets. The same classifiers of decision tree and neural network are used throughout the experimentation. Class balancer is a traditional algorithm that turns imbalanced dataset to a completely balance dataset. It works simply by dividing the majority class data which are near the boundary of two classes into the minority class to attain the balance of the dataset quantitatively.

The software programs are coded in MatLAB version 2014a for the experimentation. The computing environment is a PC workstation with a CPU: E5-2670 V2 @2.50GHz, RAM: 128GB.

Five imbalanced medical datasets are selected from UCI [14] for our experiment and their properties are presented in Table 2. The imbalanced ratios between majority class and minority class are in the range of increasing from 2.05:1 to 70.3:1.

The surgery dataset is about the lung surgeries that were conducted over 5 years. The data were collected retrospectively at Wroclaw Thoracic Surgery Centre for patients who underwent major lung resections for primary lung cancer [15]. Out of 470 surgeries, there were only 70 cases with 1-year survival period; the rest died. The training dataset represents a typical medical case of mild imbalance where the minority class has a ratio of 14.89 % which is not uncommon in surgical treatment for tuberculosis and pulmonary diseases. The same dataset was used in prediction of the post-operative life expectancy in the lung cancer patients by using SVM [15]. The imbalance problem was tackled by embedding AdaBoost into SVM; classification accuracies were said to be improved using the new method. But Kappa was omitted from the performance consideration.

Table 2 Biological datasets used in experiment

Data name	Ratio (+:−)	# Features
Surgery	1:7.1	16
Bioassay AID362	1:70.3	144
Bioassay AID439	1:2.05	100
Bioassay AID721	1:3.47	87
Bioassay AID1284	1:5.3	103

The other dataset used in our experiment is called PubChem Bioassay. These are heavily imbalanced bioassay datasets which are obtained from different types of virtual screening using high-throughput screening (HTS) technology [17]. The use of HTS in the biopharmaceutical industry has proliferated from its basis in hit identification through the entire drug discovery and early development process. Applications of HTS approaches spanned from human to animal health, from life sciences to drug discovery, and from protein docking to virtual screening simulation. These datasets have been tested in classification using different types of machine learning methods. They include but not limited to, decision tree by Chen et al. [17], Consensus model by Liew et al. [18] and genetic algorithm-neural network by Tong and Mintram [19]. In particular, the authors in [19] embedded genetic algorithm search into a neural network activation function for doing feature selection; weights to the selection of features are assigned in favors of minority class for solving the imbalance class problem. However, extensive investigation into Kappa statistics is not done yet which is common across the previous works. In our design, we opt for an alternative approach in tackling the imbalance data in the pre-processing stage, by progressively improving so using swarm algorithm which runs iteratively along with the classification model training process.

A total of 21 bioassay datasets are generated from Pubchem and available. Both Primary and confirmatory bioassays (12 bioassays, 21 mixes) are available that could be used for training and testing for evaluating the classifiers for fitness function. Out of the many sub-datasets in bioassay datasets, which are imbalanced in nature, we randomly selected four to be used in our experiment to test our method.

4 Result analysis

Our experiment collected the performance results in terms of accuracy, Kappa (Kappa statistics), precision, recall, F-measure, ROC Area, and a new index called the imbalance ratio between minority class and majority class. These results are presented from Tables 3, 4, 5, 6 and 7 with different classification algorithms or data imbalanced processed method, respectively. Furthermore, the seven performances indices which are aforementioned are visualized in radar charts individually from Fig. 9a–g with different datasets. Meanwhile we also selected the important performance indicators such as accuracy, Kappa and imbalance ratio (Min/Maj) as the key indicators to observe the changes and results in the bar-chart diagram from Figs. 4, 5, 6, 7 and 8.

Figure 4 and Table 3 show about the classification results of Surgery dataset, the original dataset's imbalance ratio (Min/Maj) is low, the two key performance indicators of accuracy and Kappa came to two extremes, high accuracy with zero Kappa value. This extreme result means the classification power of the two classification algorithms after being trained by the imbalanced data is totally useless. After processing the original dataset by classbalancer for some but perhaps not optimal balance, we can find that no matter with which classification algorithm to be used, the effect is still poor. The performances of swarm-SMOTE approaches as proposed in this paper showed that our method manage to keep classification results in check into a reliable scope, at

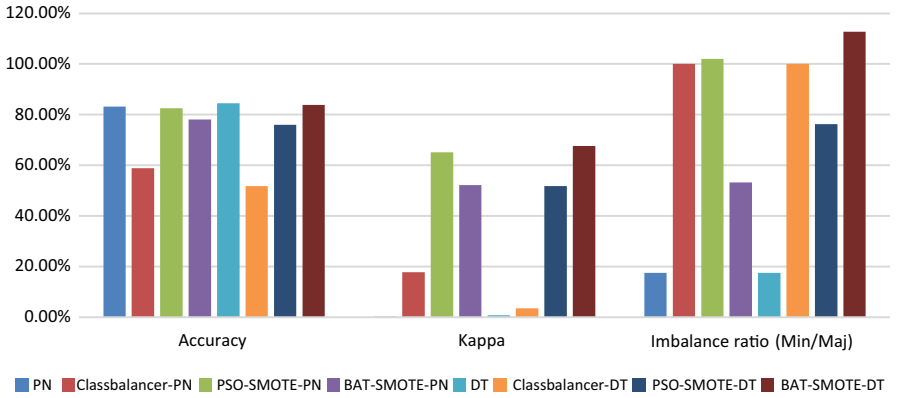

Fig. 4 Comparison of different methods in three key performances of surgery dataset

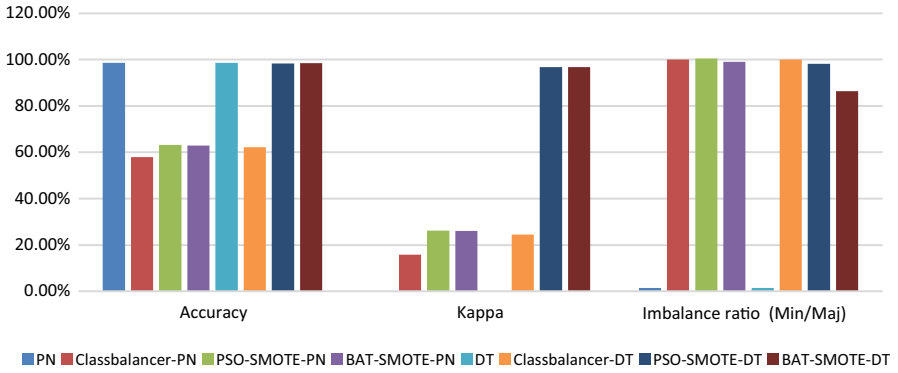

Fig. 5 Comparison of different methods in three key performances of AID362 dataset in Bioassay

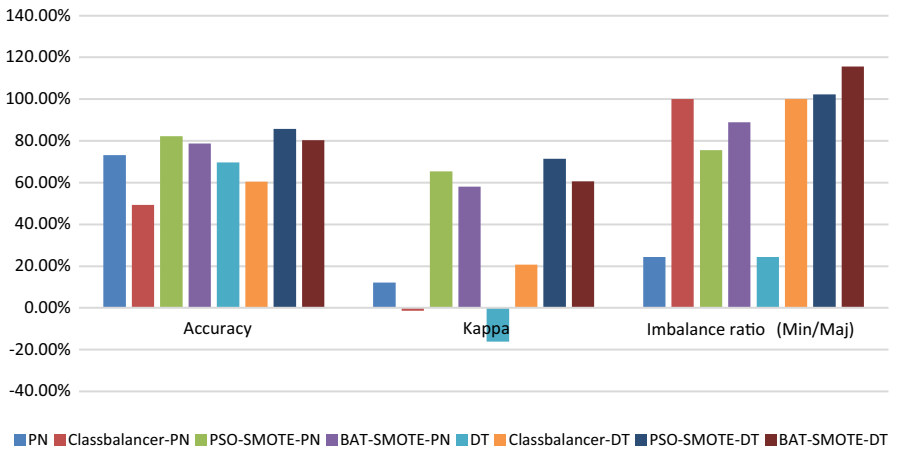

Fig. 6 Comparison of different methods in three key performances of AID439 dataset in Bioassay

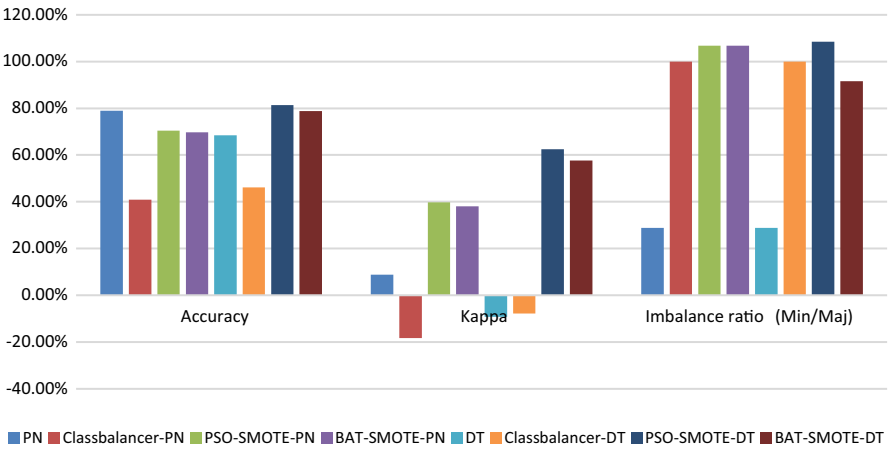

Fig. 7 Comparison of different methods in three key performances of AID721 dataset in Bioassay

the cost of some slight compromise on the accuracy. Actually the slightly lower than maximum accuracy indicates a realistic classification scenario. Through the index of imbalance ratio, we can observe the change of imbalance degree of the dataset which shows that our methods can function without needing the dataset coming to a completely balanced state (such as 50–50 % between the majority and minority data).

The other four datasets are taken from a widely diversified bioassay dataset. The results of the first Bioassay dataset which is coded AID362 with different classifiers are very extreme. It is the highest imbalanced dataset among the five. Thus there is no doubt in getting a high accuracy with a low Kappa index in the original classification. And the effect is still not very good after the original dataset processed by standard classbalancer method. But in Fig. 4 the interesting thing is that the results of our swarm-SMOTE method with neural network and decision are quite different. Under the condition of maintaining a very high accuracy, the decision tree classifier also can get a high Kappa which is very close to one; while the performance of neural network is bad whose increment was not enough to reach the credible stage, bigger or equal to 0.4. As for the third dataset of AID 439 in Bioassay, which gets a negative value of Kappa in the original classification with decision tree, it is obvious to see our method is better than the traditional method of classbalancer. In this group of results, our approach simultaneously improves the dual performance of accuracy and Kappa. And swarm-SMOTE-DT method needs to synthesize more minority samples than swarm-SMOTE-PN method. The following experiment over the dataset of AID721 in Bioassay shows that the neural network classification algorithm is worse than Decision Tree algorithm once again in Table 6 and Fig. 7. The latter classifier can get a higher accuracy and Kappa than before, whereas our method with neural network can only raise the Kappa value within a limited range with lower accuracy. It is also reflected in the results of the last dataset like before that most of the time PSO is only slightly better than Bat when it comes to finding the optimal values of the pair of parameters for best classification performance. However, it seems to achieve

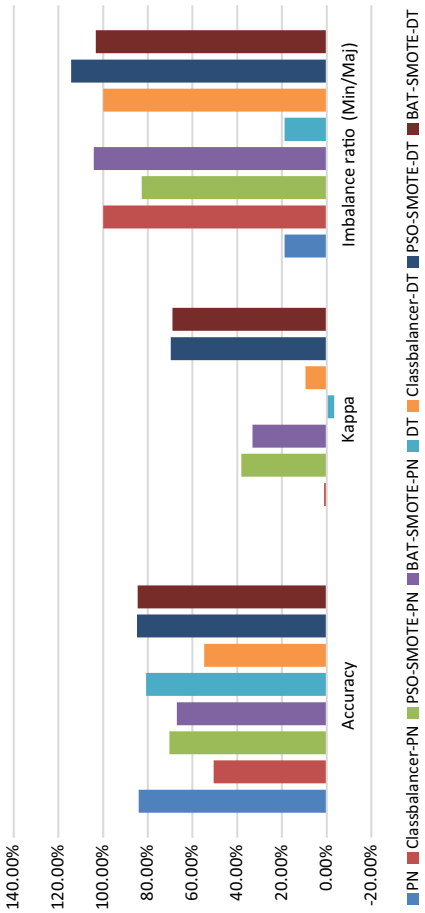

Fig. 8 Comparison of different methods in three key performances of AID1284 dataset in Bioassay

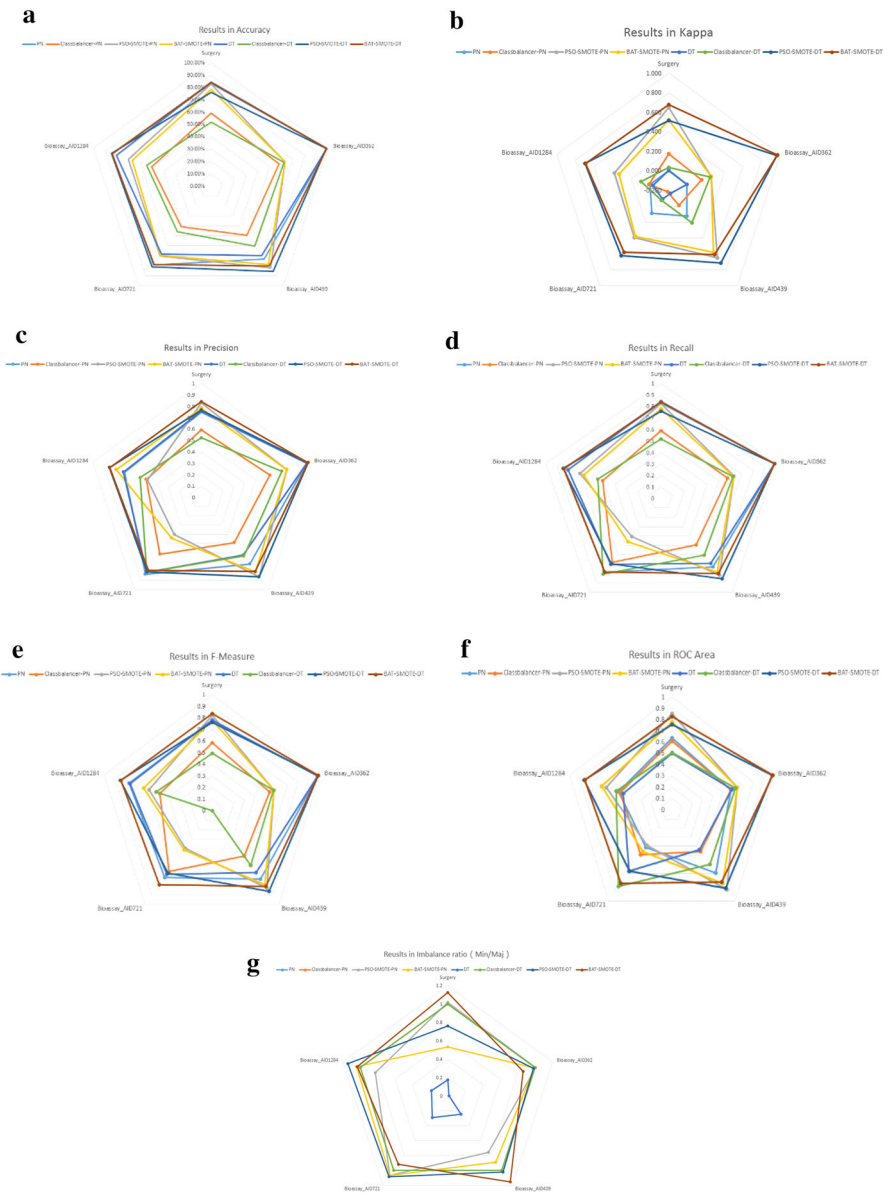

Fig. 9 **a** Radar Chart of results in accuracy, **b** Radar Chart of results in Kappa, **c** Radar Chart of results in precision, **d** Radar Chart of results in recall, **e** Radar Chart of results in *F*-measure, **f** Radar Chart of results in ROC area, **g** Radar Chart of results in imbalance ratio (Min/Maj)

this by synthesizing a lot more minority samples. Its results in Table 7 and Fig. 8 confirm the better ability to process the imbalanced dataset in classification by our method.

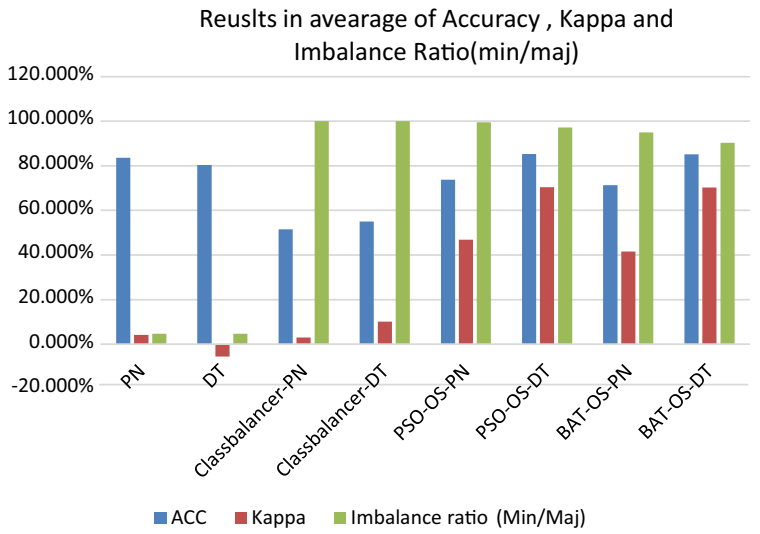

Fig. 10 Average results in different performances

Figure 9a–g, respectively, showed the different performances. It is apparent to observe so from the above analysis which is graphically displayed in the Fig. 9a–g. They, respectively, showed the different performances with different datasets. The most significant distinction is that the accuracy is very high in the original dataset classification with two different classifiers, but the value of Kappa is very low, even down to and below zero value. When this happens it means the classification accuracy is in vain. After using our method, all of the Kappa values are generally improved. But we can observe from the results that the decision tree classifier is much better than neural network. In the results of AID 721 AID 1284 with Neural Network the Kappa value struggles to go beyond 0.4, although they have risen to certain degree. It is significant to observe that by using the same swarm optimization algorithm, both the values of accuracy and Kappa have boosted. This phenomenon is most obvious in the results of AID 362 dataset. The Neural Network classification algorithms with swarm optimization algorithms performed well too. As shown in the experiment results of the last two datasets, both the two classifiers pull the Kappa value into a reliable interval, despite the fact that the decision tree is still better than neural network to keep up the high accuracy.

Figure 10 shows bar-charts of averaged values, for an overall comparison of different methods. We can see that although the class balancer method can turn the dataset to a full balance by data count, the classification performances by using it, are still very poor. There is no doubt to find in imbalanced data processed; what got balanced is the raw data count; however, the underlying mapping between the data and the target classes may remain imbalanced, nonlinearly, decision tree is much better than neural network in general.

Bat synthesis a small amount of minority class data, but it can get a good performance on par with PSO. The proposal is to recede the imbalanced degree of dataset, but the additional synthesis data may damage the structure of original dataset. Ideally, the fundamental purpose is to synthesis the least amount of data, hence minimal intervention to the original dataset, and to be able to improve the classification performance to a maximum degree by training up a highly reliable classifier.

5 Conclusion

Imbalanced data are a known challenge in classification tasks. Medical data inherently by nature are imbalanced given the small number of rare cases out of many ordinary instances. In this paper a simple and effective approach is proposed that finds an optimal values for controlling the inflation of minority class instances, called swarm-SMOTE by using stochastic swarm optimizing algorithms, such as PSO and Bat. Swarm metaheuristics are able to avoid choosing parameters blindly but to follow the optimization objective, working with constraints and progressively and heuristically refines the data subset towards the goal of balancing the two classes. This can be a good data pre-processing step suitable almost to most of the medical data classification cases. Since more synthetic samples have a greater influence on the data spatial structure, so swarm optimization algorithms do not only fix the imbalanced problem in the dataset, they help to add just the right amount of synthesis data in the minority class for a good performance results. From this aspect, in spite of the two swarm algorithms can obtain surely improved results, it is found that Bat algorithm is better than PSO for keeping the original structure of the dataset as much as possible. The proposed method is helpful for the research of biomedical research, especially in the domains involving automatic data classification. From the experimentation, our method is shown to outperform the traditional class balancing method that works solely on data counts. Moreover, decision tree can overcome the imbalanced problem, which is much better than neural network with imbalanced datasets. Our results have shown superior performance is obtained when swarm-SMOTE is coupled with classifier decision tree to solve the imbalanced dataset problem in biological medicine research. Promising classification performance is shown. It is hoped that in the future experiments we can effectively adjust the Kappa's value, instead of using a threshold, for doing multi-objective optimization, maximizing both accuracy and Kappa.

Acknowledgments The authors are thankful for the financial support from the research grant "Temporal Data Stream Mining by Using Incrementally Optimized Very Fast Decision Forest (iOVFDF)", Grant No. MYRG2015-00128-FST, offered by the University of Macau, FST, and RDAO.

Appendix

See Tables 3, 4, 5, 6 and 7.

Table 3 Results of surgery dataset

Data name				Surgery data					
Algorithms	Positive	Negative	Accuracy (%)	Kappa	Imbalance ratio (Min/Maj)	Precision	Recall	*F*-measure	ROC area
PN	70	400	83.19	0.00	0.18	0.75	0.83	0.78	0.64
Classbalancer-PN	235	235	58.89	0.18	1.00	0.59	0.59	0.59	0.61
PSO-SMOTE-PN	408	400	82.55	0.65	1.02	0.83	0.83	0.83	0.85
BAT-SMOTE-PN	213	400	78.14	0.52	0.53	0.78	0.78	0.78	0.78
DT	70	400	84.47	0.01	0.18	0.76	0.85	0.78	0.50
Classbalancer-DT	235	235	51.75	0.04	1.00	0.52	0.52	0.49	0.51
PSO-SMOTE-DT	305	400	76.03	0.52	0.76	0.77	0.76	0.76	0.76
BAT-SMOTE-DT	451	400	83.90	0.68	1.13	0.84	0.84	0.84	0.83

Table 4 Results of AID 362 in Bioassay

Data name				AID362					
Algorithms	Positive	Negative	Accuracy (%)	Kappa	Imbalance ratio (Min/Maj)	Precision	Recall	*F*-measure	ROC area
PN	48	3375	98.60	0.00	0.01	0.97	0.99	0.98	0.58
Classbalancer-PN	1711.5	1711.5	57.92	0.16	1.00	0.63	0.58	0.53	0.58
PSO-SMOTE-PN	3393	3375	63.18	0.26	1.01	0.79	0.63	0.57	0.64
BAT-SMOTE-PN	3344	3375	62.91	0.26	0.99	0.79	0.63	0.57	0.64
DT	48	3375	98.60	0.00	0.01	0.97	0.99	0.98	0.59
Classbalancer-DT	1711.5	1711.5	62.22	0.24	1.00	0.74	0.62	0.57	0.62
PSO-SMOTE-DT	3314	3375	98.36	0.97	0.98	0.98	0.98	0.98	0.98
BAT-SMOTE-DT	2916	3375	98.43	0.97	0.86	0.98	0.98	0.98	0.99

Table 5 Results of AID 439 in Bioassay

Data name				AID439					
Algorithms	Positive	Negative	Accuracy (%)	Kappa	Imbalance ratio (Min/Maj)	Precision	Recall	*F*-measure	ROC area
PN	11	45	73.21	0.12	0.24	0.72	0.73	0.73	0.69
Classbalancer-PN	28	28	49.29	−0.01	1.00	0.49	0.49	0.48	0.46
PSO-SMOTE-PN	34	45	82.28	0.65	0.76	0.86	0.82	0.82	0.87
BAT-SMOTE-PN	40	45	78.82	0.58	0.89	0.81	0.79	0.79	0.80
DT	11	45	69.64	−0.16	0.24	0.63	0.70	0.66	0.44
Classbalancer-DT	28	28	60.40	0.21	1.00	0.64	0.60	0.58	0.60
PSO-SMOTE-DT	46	45	85.71	0.71	1.02	0.86	0.86	0.86	0.86
BAT-SMOTE-DT	52	45	80.41	0.61	1.16	0.80	0.80	0.80	0.79

Table 6 Results of AID 721 in Bioassay

Data name				AID721					
Algorithms	Positive	Negative	Accuracy (%)	Kappa	Imbalance ratio (Min/Maj)	Precision	Recall	*F*-measure	ROC area
PN	17	59	78.95	0.09	0.29	0.83	0.79	0.71	0.41
Classbalancer-PN	38	38	40.88	−0.18	1.00	0.62	0.68	0.65	0.49
PSO-SMOTE-PN	63	59	70.49	0.40	1.07	0.40	0.41	0.40	0.39
BAT-SMOTE-PN	63	59	69.67	0.38	1.07	0.44	0.46	0.41	0.46
DT	17	59	68.42	−0.09	0.29	0.81	0.71	0.67	0.67
Classbalancer-DT	38	38	46.11	−0.08	1.00	0.81	0.81	0.81	0.84
PSO-SMOTE-DT	64	59	81.30	0.63	1.08	0.81	0.70	0.66	0.67
BAT-SMOTE-DT	54	59	78.76	0.58	0.92	0.79	0.79	0.79	0.80

Table 7 Results of AID 1284 in Bioassay

Data name				AID1284					
Algorithms	Positive	Negative	Accuracy (%)	Kappa	Imbalance ratio (Min/Maj)	Precision	Recall	*F*-measure	ROC area
PN	46	244	84.14	0.00	0.19	0.71	0.84	0.77	0.52
Classbalancer-PN	145	145	50.62	0.01	1.00	0.51	0.51	0.48	0.50
PSO-SMOTE-PN	202	244	70.32	0.38	0.83	0.49	0.70	0.58	0.64
BAT-SMOTE-PN	254	244	67.07	0.33	1.04	0.78	0.67	0.63	0.68
DT	46	244	80.69	−0.03	0.19	0.72	0.81	0.76	0.47
Classbalancer-DT	145	145	54.77	0.10	1.00	0.56	0.55	0.52	0.55
PSO-SMOTE-DT	279	244	84.89	0.70	1.14	0.85	0.85	0.85	0.86
BAT-SMOTE-DT	252	244	84.48	0.69	1.03	0.85	0.85	0.85	0.85

References

1. Mehta M, Agrawal R, Rissanen J (1996) SLIQ: a fast scalable classifier for data mining. In: Advances in database technology—EDBT'96. Springer, Berlin, Heidelberg, pp 18–32
2. Han J, Kamber M, Pei J (2011) Data mining: concepts and techniques: concepts and techniques. Elsevier, Amsterdam
3. Estabrooks A, Jo T, Japkowicz N (2004) A multiple resampling method for learning from imbalanced data sets. Comput Intell 20(1):18–36
4. Fan W et al (1999) AdaCost: misclassification cost-sensitive boosting. In: ICML
5. Zadrozny B, Langford J, Abe N (2003) Cost-sensitive learning by cost-proportionate example weighting. In: Third IEEE international conference on data mining, 2003. ICDM 2003. IEEE
6. Wu G, Chang EY (2005) KBA: Kernel boundary alignment considering imbalanced data distribution. Knowl Data Eng IEEE Trans 17(6):786–795
7. Joshi MV, Kumar V, Agarwal RC (2001) Evaluating boosting algorithms to classify rare classes: Comparison and improvements. In: Proceedings IEEE international conference on data mining, 2001. ICDM 2001. IEEE
8. Kotsiantis SB, Pintelas PE (2003) Mixture of expert agents for handling imbalanced data sets. Ann Math Comput Teleinform 1(1):46–55
9. Chawla NV et al (2003) SMOTEBoost: improving prediction of the minority class in boosting. In: Knowledge discovery in databases: PKDD 2003. Springer, Berlin, Heidelberg, pp 107–119
10. Chawla NV et al (2002) SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 16:321–357
11. Kennedy J (2010) Particle swarm optimization. Encyclopedia of machine learning. Springer, New York
12. Xin-She Y (2010) A new metaheuristic bat-inspired algorithm. In: Nature inspired cooperative strategies for optimization (NICSO, 2010). Springer, Berlin, Heidelberg, pp 65–74
13. Ichikawa T et al (2007) High-b value diffusion-weighted MRI for detecting pancreatic adenocarcinoma: preliminary results. Am J Roentgenol 188(2):409–414
14. Lichman M (2013) UCI Machine learning repository. University of California, School of Information and Computer Science, Irvine. http://archive.ics.uci.edu/ml. Accessed 11 Nov 2015
15. Maciej Z, Tomczak JM, Lubicz M, Witek J (2014) Boosted SVM for extracting rules from imbalanced data in application to prediction of the post-operative life expectancy in the lung cancer patients. In: Applied soft computing, vol 14, Elsevier, pp 99–108
16. Schierz AC (2009) Virtual screening of bioassay data. J Cheminform 1:1–21
17. Chen X, Wang M, Zhang H (2011) The use of classification trees for bioinformatics. Wiley Interdiscip Rev Data Min Knowl Discov 1(1):55–63
18. Ma XH, Yap CW (2010) Consensus model for identification of novel PI3K inhibitors in large chemical library. J Comput-Aided Mol Des 24(2):131–141
19. Tong DL, Mintram R (2010) Genetic algorithm-neural network (GANN): a study of neural network activation functions and depth of genetic algorithm search applied to feature selection. Int J Mach Learn Cybern 1(1–4):75–87