Pearson’s Redundancy Multi-Filtering with BAT Algorithm for Selecting High Dimensional Imbalanced Features

Ala Saleh Alluhaidan
Princess Nourah bint Abdulrahman University
Prabu P (✉ prabupphd2021@gmail.com)
Christ University
Sivakumar R
Christ University

Research Article

Keywords: Feature Selection, Imbalanced datasets, high dimensionality, BAT algorithm, Pearson’s Redundancy Algorithm.

Posted Date: November 23rd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-947703/v1

Pearson's redundancy multi-filtering with BAT algorithm for selecting high dimensional imbalanced features

Ala Saleh Alluhaidan[1],Prabu P*[2], Sivakumar R[3]

[1]Information Systems Department, College of Computer and Information Science, Princess Nourah bint Abdulrahman University, Riyadh, KSA, 84428,Email: asalluhaidan@pnu.edu.sa

[2]Department of Computer Science and Engineering, CHRIST (Deemed to be University), Bengaluru, India, Email:prabupphd2021@gmail.com

[3]Department of Computer Science and Engineering, CHRIST (Deemed to be University), Bengaluru, India

*Corresponding author:Prabu P

Abstract

Feature selection plays a vital role for every data analysis application. Feature selection aims to choose prominent set of features after removing redundant and irrelevant features from original set of features. High Dimensional dataset poses a challenging task for Machine Learning algorithms. Many state-of-art solutions were developed to handle this issue. High dimensionality in addition to imbalance ratio in the dataset becomes a tedious task. To overcome the issue, this paper introduces a novel method namely Pearson's Redundancy Based Multi Filter algorithm with improved BAT algorithm (PRBMF-iBAT) to obtain multiple feature subsets. PRBMF is implemented using multiple filters to obtain highly relevant features. iBAT algorithm uses these features to find best subset of features for classification. The results prove that PRBMF-iBAT perform better for the classifier in terms of Accuracy, Precision, Recall and F- Measure for three micro array datasets with SVM classifier. The proposed system achieves 97.99% of accuracy as highest compared to the existing rCBR-BGOA algorithm.

Keywords: Feature Selection, Imbalanced datasets, high dimensionality, BAT algorithm, Pearson's Redundancy Algorithm.

1. Introduction

The advancement in technology have paved way for the enormous growth of data in various sectors like banking, healthcare, communications, media, education, transportation, consumer trade and sports etc., This in turn has accelerated many researchers towards developing classifier model. Even though many models have been created and tested, classifying high dimensional imbalanced data still remains a leading issue in the research community. Micro array data are featured with high dimensionality, a smaller number of samples and imbalanced class distribution. Data is said to be high dimensional when it has a greater number of attributes which may be irrelevant or redundant. Many issues like high computational cost, high memory usage, difficulty in interpreting the model and decline in performance accuracy.

High dimensional dataset reduces the performance of the classifier model. To overcome this issue feature selection technique is essential. Feature selection is the most significant technique to select relevant features for developing a model that improves prediction performance thereby reducing the computational cost [1]. Feature selection reduces the number of features by selecting appropriate features for developing a model. Appropriate features do not include redundant and irrelevant attributes. Filter, wrapper and embedded methods are prominent techniques in feature selection model. Filter method uses statistical measure to evaluate and rank the features instead of using a learning algorithm. Wrapper based methods involve learning algorithm. Evaluators in embedded methods are cost function.

In addition to high dimensionality, Machine learning and data mining community leaned their interest towards imbalanced class distribution. Most of the datasets are usually imbalanced (unequal ratio of positive and negative samples) and high dimensional (more attributes in the dataset). Classes with a greater number of samples are termed as majority class and the classes with a smaller number of samples are called as minority class. The conventional classification algorithms tend favour the majority class when the classes have unequal distribution is imbalanced and degrades the performance of classifier with high dimensionality. Various solutions were provided to handle this imbalance issue including sampling techniques [2,3], cost sensitive learning methods [4,5] and ensemble techniques [6].Resampling technique includes over sampling and under sampling. The former technique increases the samples in the minority class whereas the latter technique reduces the number of samples in the majority class. The most commonly used over sampling technique is Synthetic Minority Over Sampling Technique (SMOTE) proposed by Chawla et al. [7].

In many situations exhaustive searching is not feasible so global search methods like evolutionary algorithms are used. Many researchers have hybridised other approaches to select relevant features. To handle both high dimensionality and class imbalance, main contribution of the research as follows.

1. A data pre-processing technique is proposed in which feature selection is combined with data sampling.
2. In addition, a meta heuristic algorithm is improvised to choose optimal subset of features.

The rest of the paper is organized as follows. Section 2 discusses related work. Section 3 provides pedestals of the proposed method followed by proposed methodology in Section 4. Section 5 elaborates the experimental results and finally, the conclusion and future work are summarized in Section 6.

2. Related Work

Mohan Allam et al., [8] proposed Feature Selection – Iterative Teaching Learning Based Optimization Algorithm (FS-ITLBO) model. This proposed model is an improvisation of existing TLBO algorithm.Their work involved two phases and fixed count of iterations are considered as stopping criteria. Comparison of the proposed FS-ITLBO was performedwith genetic algorithm and Binary Teaching Learning Based Optimization (BTLBO) on Wisconsin Diagnostic Breast Cancer (WDBC) and Parkinson's Disease datasets.The proposed algorithm proved better in terms of training time and error rate by selecting best optimal features.Major problem is, it does not process irrelevant high dimensional attributes.

Poolsawad et. al, [9] investigates resampling techniques to evaluate the classification algorithm's performance. The author states that resampling is one of the easiest methods to balance the minority and majority class proportions. The work proves that both of the methods has impact on sensitivity and reduces the prediction error rate for the minority class.The authors concludes that reducing the majority class samples were the suitable resampling method for clinical datasets since error rates of minority class are reduced. The major disadvantage of the proposed method is that the oversampling might duplicate the minority samples and under sampling will reduce the number of instances.It is also very important that the number of instances in each class should be significant for the classifier model to be fit.

Sotiris.K et. al.,[10] states that imbalanced dataset is one which has less number of instances in one class compared to the other class. Learning classifiers from imbalanced dataset creates problems in domains such as feature selection, information retrieval, medical diagnosis and text classification. The primary level of solution given to this problem is data level and the secondary state of solution is algorithmic level. Data level solutions are categorized into under sampling and over sampling approaches. The algorithmic solutions are based on learning techniques. This method seems to be less effective for larger datasets.

MINDEX_IB is a partition-based feature selection algorithm was proposed by Hemlata Pant,Dr. Reena Srivastava [11] for handling imbalanced datasets. This method uses micro clustering for partitioning the attribute domain and finds the relevance of attribute from the statistical measures of the microcluster.

Haoyue Liu et. al., [12] proposed weighted Gini index (FI-FSw) to deal with imbalanced classification problems. Authors revised the original Gini index using imbalanced ratio dependent weight. The main limitation in this work is the difficulty to select the first feature when multiple featuresachieve the same frequency.

Kuan-Ching et. al., [13] introduced a feature selection algorithm called Cost-sensitive Feature selection General Vector Machine (CFGVM). The pedestals of proposed method areGVM and BALO algorithm. CFGVM assigned different cost weights for the samples of different classes. More significant features were extracted by BALO algorithm that improves the performance of the classifier model. Classifier's performance on the minority class is significantly improved by the proposed methodbut declines the performance of high dimensional imbalanced small dataset.

Issue of high dimensional imbalanced dataset is handled by Garba Abdulrauf Sharifai et. al., [14]. The authors proposed Robust Correlation based Redundancy and Binary Grasshopper Optimization Algorithm.This method uses multiple filters to choose relevant features and uses Correlation based Redundancy to select the non-redundant features. Finally Binary Grasshopper Optimization Algorithm uses G-mean and AUG as fitness functions select the best feature subsets.

Considering the issuethat classifiers favour the majority class for imbalanced dataset, Zhang et al. [15] used F-measure as a performance metric for feature selection method. They introduced Support vector machine (SSVM) algorithm which selects relevant features by making use of weighted vector along with symmetric uncertainty integrated with harmony search algorithm that chose optimal feature subsets.

Zhen et al. [16] proposed WELM, to solve multi-class imbalance problems. Zhen used data level handling, algorithmic level handling and ensemble technique to upgrade the performance of the classifier. Class oriented feature selection method is used at data level. Algorithmic level usesmodified extreme learning machine (ELM) to improve the input nodes with high discrimination power.Further the author uses ensemble technique to train the performance of the model. WELM proved to be effective and outperformed other existing methods when applied on eight genetic datasets.

Multiclass imbalanced class distribution issue is solved by Du et. al., [17]. The author used improved genetic algorithm. EG-mean is used as a fitness function that favours the minority class. Du proved that his method improves the precision rate of the minority class in multi-class imbalanced datasets in the size of feature subsets.Rung-Ching Chen et. al., [18] in their work clearly showed Random Forest is the best classifier by comparing RF with different classifiers. Authors combined Random Forest (RF), Support Vector Machine(SVM), K-Nearest Neighbor(KNN) and Linear Discriminant Analysis (LDA) with different features selection method RF, RFE, and Boruta to select the best classifiers method based on the accuracy of each classifier.

3. Pedestals of the Proposed Method

3.1 Pearson's Redundancy Based Filter (PRBF)

Pearson'sχ 2test is used to measure the difference between Probability distribution of two variables. Let us take n observations which are independentof two random variables X, X' in the training data. Pearson χ 2test will be valid only if number of observations are greater than 100.

$$\chi 2\ (X, X') = \sum_{i=1}^{k} \frac{(F_i - F'_i)^2}{F'_i} \text{..........} \ (1)$$

Ifχ 2 value is large the features are not redundant.When p-value p(χ 2) > α then the two distributions are equivalent with α significance level, and thus one of the features is redundant. Cross validation techniques could be used independently for each classifier to find the best p-value.

Algorithm 1 presents the Pearson’s Redundancy Based Filter (PRBF)algorithm.Symmetrical uncertainty is used to find the relevancefurtherχ 2 test is appliedto remove redundancy.

Algorithm 1: PRBF

Input : Feature dataset

Step 1: Relevance Analysis

- Calculate SU(X, C) relevance indices and create an ordered list S of features according to the decreasing value of their relevance.

Step 2: Redundancy Analysis

- Take as X the first feature from the S list
- Find and remove all features for which X is approximately equivalent according to the Pearson χ 2 test.
- Set the next remaining feature in the list as X and repeat step 3 for all remaining features in the S list.

Output : Original features

3.2 Symmetric Uncertainty

Symmetric Uncertainity (SU) was introduced by Witten and Frank [19]was defined to measure the redundancy. features fitness is calculated by SU. Highly scored features of SU is considered as important feature.Mutual Information (MI) is extended assymmetric uncertainty (SU). MI is normalized with the entropy value of features and entropy value of features normalised withclass label. SU has been used to evaluate the effectiveness of features for classifying the data. Symmetric Uncertainity (SU) is defined as

$$SU(X,Y) = \frac{2 \times MI(X,Y)}{H(X)+H\ (Y)}\dots\dots\dots\dots\dots\dots\dots(2)$$

Here D is the high dimensional data with N is the number of instances and M is the number of features.Random features are X and Y. MI is the mutual Information.The entropies of discrete random features X and Y are H(X) and H(Y).SU normalizes the features having different values within the range of [0,1]. SU(X,Y) is1 when there is dependency between the features and SU(X,Y) is0 when the features are independent to each other[20].

3.3 Fisher Score Filter Approach

Fisher score proposed by Gu et al.,.[21]is an heuristic approach for calculatingfeaturs's score using Fisher ratio. Let µf$_i$be the mean and $\sigma_{f_i}^{k}$be the standard deviation of the k-th class and i-th feature. The Fisher score for the feature i can be calculated using the equation 3.

$$\mathrm{F(fi)} = \frac{\sum_{k=1}^{c} n_k\ (\ \mu_{f_i}^{k} - \mu^{f}i)^2}{\sum_{k=1}^{c} n_k\ (\sigma_{f_i}^{k})^2} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots..(3)$$

where n$_k$is the number of samplesin class k. Features with highest fisher score values will be selected. These selected features are suboptimal since the scores are evaluated individually. The main disadvantage of Fisher score is it fails to select the features with high aggregated discriminative power and redundant ones.

3.4 BAT Algorithm

BAT algorithm is based on echo location behaviour of micro bats. Thisalgorithm mimics the behaviour of bats while catching prey. BAT algorithm was proposed by Yang in [22]. Bats use echolocation to find its prey. When the bats are closer toits prey its pulse rate and frequency increase thereby decreases the loudness. A set of interactive parameters like position, velocity, pulse rate, loudness and frequency is assigned to each bat that affects the quality of solution and time to obtain the solution which makes the algorithm complicated compared to other metaheuristic algorithms [23].Figure 1 depicts the flowchart of BAT algorithm [24]

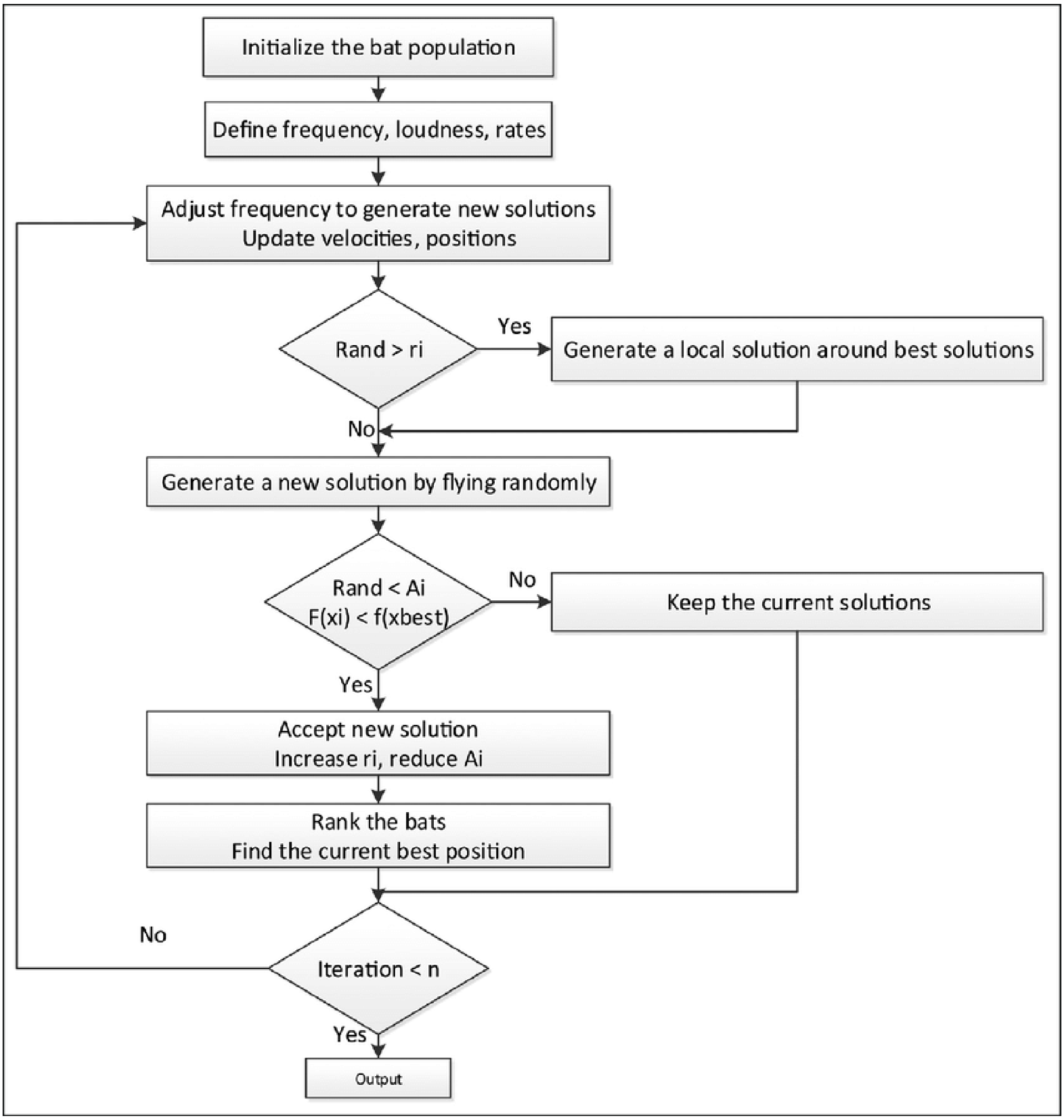
Initialize the bat population
Define frequency, loudness, rates
Adjust frequency to generate new solutions
Update velocities, positions
Rand > ri
Yes
Generate a local solution around best solutions
No
Generate a new solution by flying randomly
Rand < Ai
F(xi) < f(xbest)
No
Keep the current solutions
Yes
Accept new solution
Increase ri, reduce Ai
Rank the bats
Find the current best position
No
Iteration < n
Yes
Output

Fig 1. Flowchart of BAT Algorithm

3.4.1 Principles of BAT algorithm

A swarm of bats searches for its prey by flying randomly with velocity Vi at position Xi with fixed frequency f, varying wavelength λ and A_0 as loudness. Rate of pulse (r) emission determines the closeness of the target. r ϵ [0,1] where rate of pulse increases when the bat is closer to the target. The loudness varies from A_0 to A_{min}. Frequency and wavelength varies from $[f_{min}, f_{max}]$ and $[\lambda_{min}, \lambda_{max}]$ respectively [4.7]. The simulated bats will update their positions and velocity in a D dimensional space. The new solutions xi^t and velocity vi^t at time t is given by

$$f_i = f_{min} + (f_{max} - f_{min})\, \beta \text{ ....(4)}$$

where β ϵ [0,1] is a random vector drawn from uniform distribution.

$$v_i^t = v_i^{t-1} + (x_i^{t-1} - x^*)\, f_i \text{ ....(5)}$$

x* is the global best solution which is found after comparing all the solutions among all n bats.

$$x_i^{\,t} = x_i^{\,t-1} + v_i^{\,t} \text{.............(6)}$$

After obtaining the global best solution, a new solution for each bat is generated using a random walk with the following equation.

$$x_{new} = x_{old} + \epsilon A^t \text{ ....(7)}$$

Where Є ϵ [-1,1] is a random number and A^t is the average loudness of all the bats at time t.

As the iterations proceed, loudness and rate of pulse emission have to be updated.

$$A_i^{t+1} = \alpha A_i^t \text{.............(8)}$$

$$r_i^{t+1} = r_i^0 \text{ [1-exp(-Υt)]............(9)}$$

Where α, Ύ are constants.

4 Proposed PRBMF-iBATMethodology

The proposed method isa hybris of filter and wrapper-based feature selection technique for datasets with more features and unequal class distribution. The proposed Pearson's Redundancy Based Multi Filter algorithm with improved BAT algorithm (PRBMF-iBAT). The proposed methodworks in three phases. State of art method SMOTE is implemented on the datasets to handle imbalanced nature.Second phase enhances Pearson redundancy-based filter (PRBF) by adding an additional filter (Fisher Score) to select highly ranked features. These features are given as input to the improved BAT algorithm (iBAT) to select the optimum feature subsets in the third phase. Pearson redundancy-based filter algorithm is enhanced by including Fisher Score to choose the top ranked features. In addition to the enhanced Pearson's Redundancy Based Multi Filter algorithm, BAT algorithm is improvised to choose optimal subset of features. This method seems to be cost effective [25] since it does not check all possible combinations of features. The schematic diagram of the proposed method is given in the figure 2.

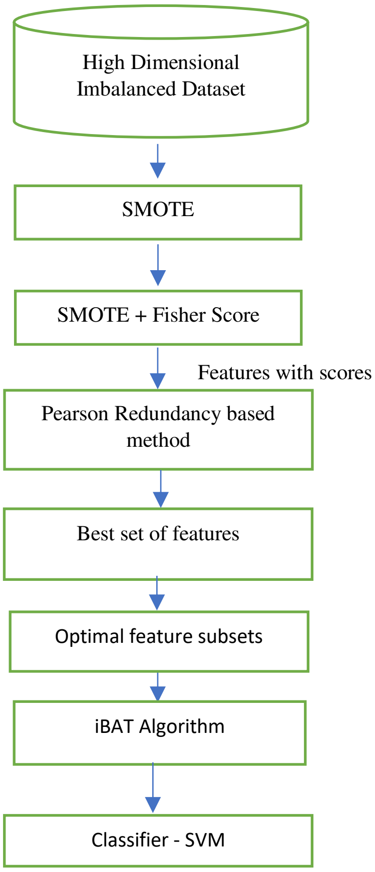

Figure 2. Schematic representation of the proposed Methodology

4.2.1 Proposed Pearson's Redundancy Based Multi Filter Algorithm

Phase 1of the proposed methodology uses Pearson's redundancy based multi filter algorithm to choose best set of features from high dimensional imbalanced dataset. The following algorithm pictures the Pearson's Redundancy Based Multi Filter Algorithm.

Algorithm 2: Pearson’s Redundancy Based Multi Filter Algorithm

```
Input:High dimensional Imbalanced Dataset
Output: relevant subset of features
Step 1: Apply SMOTE Algorithm to the high dimensional imbalanced dataset to obtain balanced class distribution
Step 2: for each feature $f_i$ in the training data and class C, Initialize S= ∅
Step 3: for i= 1 to m ( m is the set of features in the dataset)
Step 4:Calculate the fisher score value of all features in the training data.
Step 5: End for
Step 6: Sort the features with the fisher score values.
Step 7: Choose top ranked feature set S.
Step 8: Calculate feature-feature SU and feature-class SU for each feature $f_i$ in S.
Step 9: for i=i to n
Step 10: for j=i+1
Step 11: Calculate $SU_{i,c}$, $SU_{j,c}$, $SU_{i,j}$
Step 12: If $SU_{i,c}$>=$SU_{j,c}$and $SU_{i,j>=}$ $SU_{j,c}$
Step 13: remove $f_j$ else add $f_j$ to the selected set of features FS.
Step 14: End j loop and i loop
```

The proposed algorithm takes high dimensional imbalanced dataset as input. Synthetic Minority Oversampling Technique (SMOTE) is applied on the input dataset to obtain balanced dataset. SMOTE uses oversampling technique and creates synthetic samples to equalize minority and majority class instances. Further the features from the balanced dataset are scored using fisher score filter method and top ranked features are chosen and used in Pearson’s based redundancy method in which Symmetric Uncertainty is used as a relevance factor.

4.2.2 Proposed Improved BAT Algorithm

An improved BAT algorithm is applied to the output formed by the phase I. All the features from phase 1 are given as input to iBAT algorithm which results in optimal subset of features. IniBAT algorithm, a new bat position is obtained by calculating the mean between the previous position and the current best solution x*.

$v_i^t = v_i^{t-1} + \text{mean}(x_i^t, x_*) + v_i^t$ ....(10)

$x_i^t = x_i^{t-1} + v_i^t$ ....(11)

This method yields a good exploitation so naturally the solutions move towards local and global optima. The following algorithm 3 represents the iBAT method.

Algorithm 3: proposed iBAT Algorithm

```
Input: Balanced dataset with features chosen from PRBMF method
Output: Optimal subset of features
    1. For all the features in the feature set FS obtained from PRBMF algorithm
Initialize $x_i$( i=1,2,3…….n) and velocity ($v_i$) of all n bats
Define pulse frequency $f_i$ at $x_i$
Initialize $r_i$ (pulse rate) and $A_i$ (loudness)
Calculate the fitness of all bats
Find the current best position $x_i$best according to the fitness value
iter=1
while termination condition not met
iter = iter +1
For_each bat in population:
$v_i^t = v_i^t$-1 +mean ($x_i^t$-1, x*) $f_i$
```

```
$x_i^t = x_i^t-1 + v_i^t$
if rand(0,1) >$r_i^t$
select a solution among the best solutions
Generate a local solution around the best solution
Fly randomly to generate a new solution
If(rand(0,1) < $A_i$ and f($x_i$) < f(x))
Accept new solutions
Update pulsation rate and loudness
Sort the bats based on fitness value and save the current best position ($x_i$)
Choose the bats (features) with best fitness.
```

The proposed method improvises BAT algorithm to choose optimal set of features. First Initialize random values for parametersfrequency, velocity, position, loudness and pulse rate.Adjust the frequency using equation 4. In the proposed iBAT velocity is updated by adding the mean value of the previous bat position and the currentbest position. Select a solution among the best. Update pulse rate and loudness.Perform the same until the frequency is high and pulse rate is zero. When the pulse rate is zero it means the bat reached the solution.

5. Experimental Results and Discussions

The effectiveness of the proposed PRBMF-iBAT with regard to high dimensional imbalanced dataset is discussed in this section. Three microarray datasets are used to evaluate the proposed method.Implementation was carried out on an Intel Core i5 @ 2.42 GHz CPU and 16 GB RAM in Microsoft Windows 10 platform. Implementation was performed in MATLAB R 2015. The datasetsused are Lung Cancer [14], Prostrate Tumor from repository and SRBCT [14]. The details of the dataset are listed in table 5.1.

Table 1. Dataset Description

Dataset	**Size**	**No. of Classes**
Lung_Cancer	181 x 12601	2
Prostrate_Tumor	102 x 12600	2
SRBCT	83 x 2309	4

By incorporating MATLAB tool, the effectiveness of the proposed method is evaluated using Accuracy, Precision, Recall and F-measure. According to the confusion matrix the measures can be calculated using the below formula

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} * 100 \quad \ldots\ldots\ldots\ldots\ldots\ldots(12)$$

$$\text{Precision} = \frac{TP}{TP+FP} * 100 \quad \ldots\ldots\ldots\ldots\ldots(13)$$

$$\text{Recall} = \frac{TP}{TP+FN} * 100 \quad \ldots\ldots\ldots\ldots\ldots..(14)$$

$$\text{F-Measure} = 2 * \frac{(Recall*Precision)}{Recall+Precision} * 100 \quad \ldots\ldots\ldots\ldots.(15)$$

Classifiers tend to bias towards the majority class resulting high accuracy but the minority class consideration is low so accuracy alone cannot be an appropriate measure for imbalanced datasets. F-measure could be an appropriate measure to evaluate the efficiency of proposed method. Proposed PRBMF-iBAT is compared with Robust Correlation Based Redundancy and Binary Grasshopper Optimization Algorithm (rCBR-BGOA) [14]for Support Vector Machine. Fivefold cross validation is performed to compare the results.

Table 2. Comparison of proposed PRBMF-iBAT against rCBR-BGOA with regard to Accuracy, Precision, Recall and F-Measure

Dataset	Metrics	rCBR-BGOA	PRBMF-iBAT
Lung Cancer	Accuracy	87.90	91.97
	Precision	87.99	91.95
	Recall	87.89	92.04
	F-Measure	87.93	91.99

Prostrate Tumor	Accuracy	82.57	87.90
	Precision	81.61	87.93
	Recall	81.37	87.92
	F-Measure	81.49	87.93
SRBCT	Accuracy	93.22	97.99
	Precision	93.29	97.95
	Recall	93.17	98.01
	F-Measure	93.23	97.98

Table 5.2 proves that the proposed PRBMF-iBAT outperformed the rCBR-BGOA method in all the chosen metrics. PRBMF-iBAT proved to be significant in terms of accuracy, precision, recall and f-measure for SRBCT dataset comparing to Lung Cancer and Prostrate Tumor. The obtainedresults provethat multi filter method along with metaheuristic BAT algorithm is very effective in case of high dimensional imbalanced datasets. Figure 3 shows the pictorial representation of the comparative analysis of the proposed and existing approach.

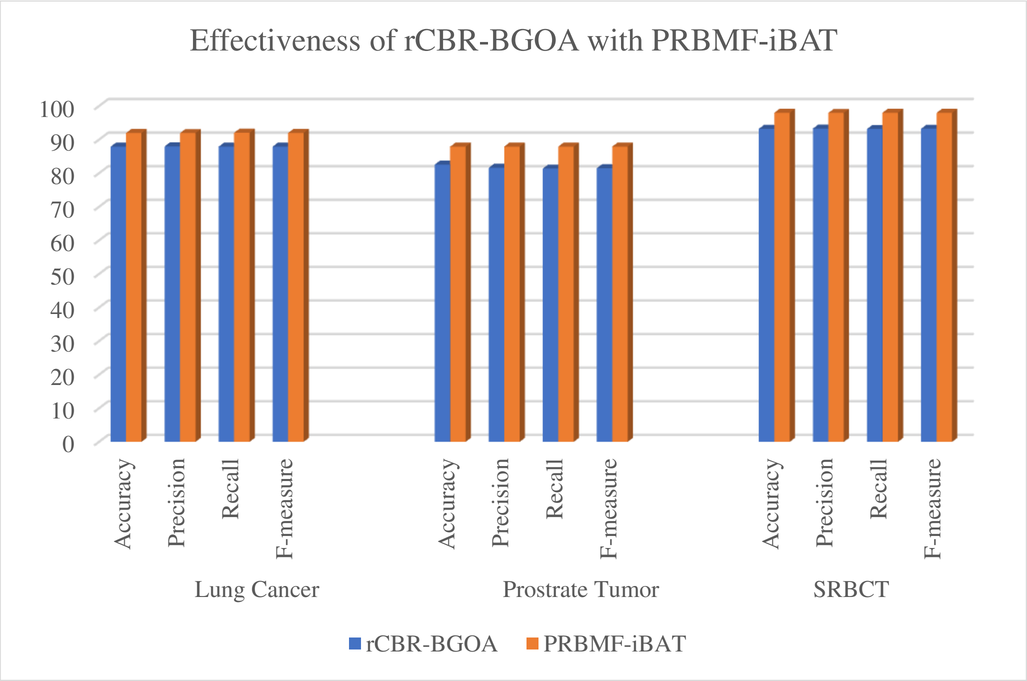

Fig 3. Comparison of proposed PRBMF-iBAT against rCBR-BGOA

The proposed method and the existing methods are also compared against the computational time. The figure 4 clearly pictures that the proposed method classifies the data in less time compared to the existing method for all three datasets.

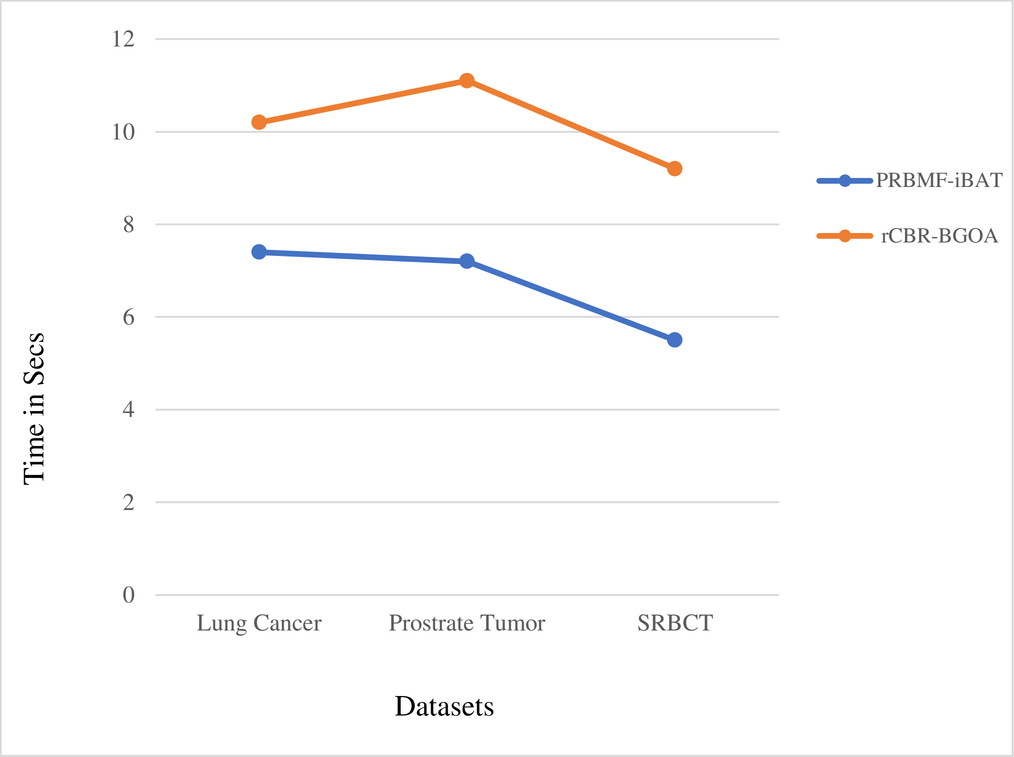

Fig 4. Comparison of proposed PRBMF-iBATandrCBR-BGOA in terms of computational time

PRBMF-iBAT algorithmis implemented for three datasets with SVM classifier. The results show that out of three datasets SRBCT shows a better result for all the four metrics. It scores 97.99% accuracy 97.95 % of precision, 98.01% of Recall and 97.98% of F-measure which is superior than the existing method and also outperforms the proposed method implemented for Lung cancer and prostrate tiumor datasets.Execution time of the algorithms also seems to be less in SRBCT dataset compared to the other datasets.The main problem noted in this method is, it increases the overlapping of classes that can introduce additional noise during balancing the imbalanced class distribution.

6. Conclusion and Future Enhancements

The objective of the proposed method is to select optimal set of features in high dimensional imbalanced datasets. A novel approach PRBMF-iBAT is proposed to overcome the issues in selecting features in high dimensional imbalanced dataset.PRBMF-iBAT involves three phases, first phase includes the implementation of SMOTE algorithm to handle imbalanced issues followed by PRBMF ensembled with improved BAT algorithm to choose optimal set of features.The results obtained in this paper proves that the proposed PRBMF-iBAT outperformed in terms of Accuracy, Precision, Recall and F-measure. The proposed method can be further improved by enhancing techniques to handle imbalance issue.

Conflict of Interest

This paper has not communicated anywhere till this moment, now only it is communicated to your esteemed journal for the publication with the knowledge of all co-authors.

Ethical approval

This article does not contain any studies with human participants or animals performed by any of the authors.

Funding

There is no funding.

Informed Consent

All authors have seen the manuscript and approved to submit it to the journal.

Author contributions

All authors have seen the manuscript and approved to submit it to the journal.

References

1. H. Liu, H. Motoda, R. Setiono, and Z. Zhao. Feature selection: An ever evolving frontier in data mining. in Proceedings of the Fourth International Workshop on Feature Selection in Data Mining, Hyderabad, India, 2010, pp. 4–13.

2. Han, H.; Wang, W.-Y.; Mao, B.-H. Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning. In Proceedings of the International Conference on Intelligent Computing, Hefei, China, 23–26 August 2005; Springer: Berlin/Heidelberg, Germany, 2005; pp. 878–887. 10.

3. Yen, S.-J.; Lee, Y.-S. Cluster-based under-sampling approaches for imbalanced data distributions. Expert Syst. Appl. 2009, 36, 5718–5727.

4. Zhou, Z.-H.; Liu, X.-Y. Training cost-sensitive neural networks with methods addressing the class imbalance problem. IEEE Trans. Knowl. Data Eng. 2006, 18, 63–77.

5. Ling, C.; Sheng, V. Cost-sensitive learning and the class imbalance problem. In Encyclopedia of Machine Learning; Springer: Berlin/Heidelberg, Germany, 2011; Volume 24.

6. Chawla, N.V.; Lazarevic, A.; Hall, L.O.; Bowyer, K.W. SMOTEBoost: Improving prediction of the minority class in boosting. In Proceedings of the European Conference on Principles of Data Mining and Knowledge Discovery, Cavtat-Dubrovnik, Croatia, 22–26 September 2003; Springer: Berlin/Heidelberg, Germany, 2003; pp. 107–119.

7. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic minority over-sampling technique. J. Artif. Intell. Res. 2002, 16, 321–357.

8. Mohan Allam and Nandhini Malaiyappan. 2020. Wrapper based Feature Selection using Integrative Teaching Learning Based Optimization Algorithm. The International Arab Journal of Information Technology, Vol. 17, No. 6, November 2020, 885-894.

9. Poolsawad, N., C. Kambhampati and J.G.F. Cleland 2014. Balancing Class for Performance of Classification with a Clinical Dataset.In the Proceedings of the World Congress on Engineering 2014 Vol I, July 2 - 4, 2014, London, U.K.

10. Sotiris Kotsiantis, Dimitris Kanellopoulos and Panayiotis pintelas. Handling Imbalanced Datasets: A review. GESTS International Transactions on Computer Science and Engineering,2006 30, 1-13.

11. Hemlata Pant, Dr. Reena Srivastava. MINDEX_IB: A Feature Selection method for Imbalanced Dataset. International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 4, April 2016.

12. Haoyue Liu, MengChu Zhou, and Qing Liu. An Embedded Feature Selection Method for Imbalanced Data Classification. IEEE/CAA Journal of AutomaticaSinica Vol. 6, No.3 May 2019.

13. Feng, Fang; Li, Kuan-Ching; Shen, Jun; Zhou, Qingguo; and Yang, Xuhui. Using cost-sensitive learning and feature selection algorithms to improve the performance of imbalanced classification (2020). Faculty of Engineering and Information Sciences - Papers: Part A. 6783.

14. Abdulrauf Sharifai G, Zainol Z. Feature Selection for High-Dimensional and Imbalanced Biomedical Data Based on Robust Correlation Based Redundancy and Binary Grasshopper Optimization Algorithm. *Genes*. 2020; 11(7):717. https://doi.org/10.3390/genes11070717.

15. Zhang, C.; Wang, G.; Zhou, Y.; Yao, L.; Jiang, Z.L.; Liao, Q.; Wang, X. Feature selection for high dimensional imbalanced class data based on F-measure optimization. In Proceedings of the 2017 International Conference on Security, Pattern Analysis, and Cybernetics (SPAC), Shenzhen, China, 15–17 December 2017; pp. 278–283.

16. Liu, Z.; Tang, D.; Cai, Y.; Wang, R.; Chen, F. A hybrid method based on ensemble WELM for handling multi class imbalance in cancer microarray data. Neurocomputing 2017, 266, 641–650.

17. Du, Lm., Xu, Y. & Zhu, H. Feature Selection for Multi-Class Imbalanced Data Sets Based on Genetic Algorithm. *Ann. Data. Sci.* **2,** 293–300 (2015). https://doi.org/10.1007/s40745-015-0060-x.

18. Bharat Singh, Nidhi Kushwaha, Om Prakash Vyas. A Feature subset selection Technique for High Dimensional Data using Symmetric Uncertainity. Journal of Data Analysis and Information Processing. Vol. 02 No.04(2014),Article ID:51490, 10 Pages.

19. Witten, I.H.; Frank, E.; Trigg, L.E.; Hall, M.A.; Holmes, G.; Cunningham, S.J. Weka: Practical Machine Learning Tools and Techniques with Java Implementations; University of Waikato: Hamilton, New Zealand, 1999.

20. Gu, Q.; Li, Z.; Han, J. Generalized Fisher Score for Feature Selection; Cornell University: Ithaca, NY, USA, 2012; arXiv preprint arXiv:1202.3725.

21. X.S Yang, A new metaheuristic bat-inspired algorithm in Nature Inspired Cooperative Strategies for Optimization (NICSO' 10), J.Gonzalez, D.Pelta, C.Cruz, G.Terrazas, and N.Krasnogor, Eds., pp.65-74, Springer, Berlin, Germany, 2010.

22. Parpinelli, R.S. and Lopes, H.S., "New inspirations in swarm intelligence: a survey", Int. J. Bio-Inspired Computation, Vol. 3, No. 1, pp.1–16, 2011.

23. Ahmed Majid Taha, Soong-Der Chen, Aida Mustapha. Natural Extensions : Bat algorithm with memory. Journal of theoretical and applied Information Technology. Vol 79, No 1, 2015.

24. https://www.researchgate.net/figure/The-bat-algorithm-flowchart_fig2_321054330

25. Yang, J.; Zhou, J.; Zhu, Z.; Ma, X.; Ji, Z. Iterative ensemble feature selection for multiclass classification of imbalanced microarray data. J. Biol. Res. Thessalon. 2016, 23, 13.