RESEARCH ARTICLE

Modified Bat Algorithm for Feature Selection with the Wisconsin Diagnosis Breast Cancer (WDBC) Dataset

Suganthi Jeyasingh[1]*, Malathi Veluchamy[2]

Abstract

Early diagnosis of breast cancer is essential to save lives of patients. Usually, medical datasets include a large variety of data that can lead to confusion during diagnosis. The Knowledge Discovery on Database (KDD) process helps to improve efficiency. It requires elimination of inappropriate and repeated data from the dataset before final diagnosis. This can be done using any of the feature selection algorithms available in data mining. Feature selection is considered as a vital step to increase the classification accuracy. This paper proposes a Modified Bat Algorithm (MBA) for feature selection to eliminate irrelevant features from an original dataset. The Bat algorithm was modified using simple random sampling to select the random instances from the dataset. Ranking was with the global best features to recognize the predominant features available in the dataset. The selected features are used to train a Random Forest (RF) classification algorithm. The MBA feature selection algorithm enhanced the classification accuracy of RF in identifying the occurrence of breast cancer. The Wisconsin Diagnosis Breast Cancer Dataset (WDBC) was used for estimating the performance analysis of the proposed MBA feature selection algorithm. The proposed algorithm achieved better performance in terms of Kappa statistic, Mathew's Correlation Coefficient, Precision, F-measure, Recall, Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Relative Absolute Error (RAE) and Root Relative Squared Error (RRSE).

Keywords: Breast cancer- Wisconsin Diagnosis Breast Cancer (WDBC) dataset- modified bat algorithm

Asian Pac J Cancer Prev, **18 (5)**, 1257-1264

Introduction

Breast cancer is the prominent cause of the cancer-related death among women aged 15-54 years (Garfinkel, Catherind, Boring, and Heath, 1997). According to the World Health Organization (WHO) report, more than 1.2 million women across the world will be diagnosed with breast cancer each year. But, the mortality rate from breast cancer has reduced in the recent years with the increased advancement in the diagnostic techniques and more effective treatments. A key factor is the early detection and accurate diagnosis of this disease (West, Mangiameli, Rampal, and West, 2005). Among other screening methods, mammography is the gold standard approach for the detection of breast cancer in the earlier stage. The radiologist should interpret a large number of mammograms on a routine basis to diagnose the cancer. However, the recent studies show that there is a failure of about 10-40 % in the detection of breast cancer during the screening stage (Beam, Layde, and Sullivan, 1996; Bird, Wallace, and Yankaskas, 1992; Elmore, Wells, Lee, Howard, and Feinstein, 1994; Giger, 2002) due to the misinterpretation of non-cancerous lesions as a cancer. The Computer-Aided Detection (CAD) can reduce the screening efforts of the radiologist (Warren Burhenne et al., 2000). Over the last decade, the CAD approaches are used for the detection of abnormalities such as micro-calcifications (Karssemeijer, 1992; Strickland and Hahn, 1996; Veldkamp, Karssemeijer, Otten, and Hendriks, 2000), masses (Kobatake, Murakami, Takeo, and Nawano, 1999; Mudigonda, Rangayyan, and Desautels, 2001; Petrick, Chan, Sahiner, and Wei, 1996), and spiculated lesions (Kegelmeyer Jr, 1993; Kegelmeyer Jr et al., 1994; Liu, Babbs, and Delp, 2001). The CAD system is classified as CADe and Computer-aided Diagnosis (CADx) systems. The CADe systems help the radiologist in detecting and locating the abnormal area in images and the CADx systems diagnose and classify the benign or malignant tissues. Generally, the CAD system involves preprocessing, initial segmentation, feature extraction, feature selection and classification of normal and abnormal tissues (Jalalian et al., 2013).

Feature selection plays a vital role in the healthcare application for the efficient classification of benign and malignant tumors (Ghazavi and Liao, 2008; Lee et al., 2003; López, Novoa, Guevara, and Silva, 2007; Soltanian-Zadeh and Rafiee-Rad, 2004; Wei et al., 2005). The feature selection approach facilitates data visualization and data understanding, requires minimum storage requirements, reduces training and utilization time and defines the curse

[1]*Department of Computer Science and Engineering, Raja College of Engineering and Technology,* [2]*Department of Electrical and Electronics Engineering, Anna University Regional Centre, Madurai, Tamilnadu, India. *For Correspondence: gksuganthi123@rediffmail.com*

of dimensionality to improve the breast cancer prediction performance (Guyon and Elisseeff, 2006). Devijver and Kittler (Devijver and Kittler, 1982) define feature selection is the process of extracting the relevant information from the raw data to improve the classification performance. Different features such as texture, morphological features, descriptor, model-based features, multi-resolution and shape features are extracted (Cheng, Shan, Ju, Guo, and Zhang, 2010; Sun, Babbs, and Delp, 2006). A feature selection method selects a subset of relevant features used for the classification purpose (Sun et al., 2006) During the last decade, the researchers applied the statistical-based approaches, machine learning algorithms and knowledge discovery techniques for the feature selection. Different from the dimensionality reduction approaches such as Principal Component Analysis (PCA) (Haka et al., 2005), the feature selection techniques select a subset of variables, without changing the original representation of the variables.

Hence, they maintain the original semantics of the variables and offer the advantage of interpretation by using a domain expert (Saeys, Inza, and Larrañaga, 2007). The feature selection techniques are categorized as wrapper, filter and embedded methods based on the combination of feature selection search and creation of the classification model. The wrapper utilizes the machine learning classifiers for scoring the feature subset according to the prediction capability. The embedded methods select the features during the training process. The filter method uses heuristics based on general data characteristics for evaluating the advantages of the features. Hence, the filter method is faster than the wrapper and embedded methods (Guyon and Elisseeff, 2003, 2006). The main drawback of the filter method is that it ignores the dependencies among the features and treats the features individually (Vanaja and Kumar, 2014). The feature selection approach reduces the number of input features in a classifier to obtain a good predictive and less computationally intensive model. The feature selection algorithm that supports the binary dataset and multiclass dataset yields high accuracy on the binary dataset and low accuracy on the multiclass dataset. The classification systems can help in the reduction of possible diagnosis errors caused by the inexperienced experts, and also enable detailed examination of the medical data within a short span of time. The classification techniques such as Support Vector Machine (SVM) (Huang, Wang, and Chen, 2006) and Artificial Neural Network (ANN) (C.-M. Chen et al., 2003; Joo, Yang, Moon, and Kim, 2004; Song et al., 2005) are used for the mass detection and classification (Jesneck, Lo, and Baker, 2007). However, ANN has some intrinsic disadvantages such as slow convergence speed, minimum generalization performance, arriving at the local minimum and over-fitting problems. The performance of the SVM classifier mainly depends on the proper choice of a kernel function among other factors.

To mitigate the existing issues, this paper proposes a combined approach of modified Bat inspired algorithm for feature selection and RF based classification for the breast cancer detection. The simple random sampling is used to modify the Bat algorithm for the selection of the random instances from the dataset. It ranks the global best features to recognize the predominant features available in the dataset. The RF algorithm ensures efficient classification of the benign and malignant cancer cells through the selection of best features. The Wisconsin Breast Cancer dataset obtained from the University of California at Irvine (UCI) Machine Learning Repository (Zwitter and Soklic) is used for evaluating the performance of the proposed work. The proposed algorithm yields maximum accuracy and minimum error rate when compared to the existing feature selection techniques.

A statistical test, namely, Mann-Whitney test is enhanced to select the features for efficient breast cancer diagnosis. The uFilter is utilized for enhancing the Mann-Whitney test by the incorporation of binary classification. An uFilter application is designed for the diagnosis of breast cancer using Computer Aided Diagnosis (CADx). Information Gain (IG), One-Rule (1Rule), Chi-square (CHI2) discretization, uFilter and Relief are applied for producing the datasets with reduced number of features. The derived datasets are trained using certain classifiers including Naive Bayes, Support Vector Machine (SVM), Linear Discriminant Analysis (LDA) and Feed-Forward back propagation neural network. The U-test method is outperformed by the uFilter method by reducing the redundant data present in the dataset (Pérez, López, Silva, and Ramos, 2015).

An intelligent classification model was proposed by combining the fuzzy-rough nearest neighbor classification and fuzzy-rough instance solution. The re-ranking algorithm was applied to evaluate the consistency of the subsets. The suggested model included three important phases, namely, instance selection, feature selection and classification. The irrelevant features are eliminated using the weak gamma evaluator and the consistency in the second phase was maintained by the re-ranking algorithm. The WDBC dataset is used for evaluation of the proposed breast cancer classification model (Onan, 2015). Various feature selection algorithms were compared by applying the relevant machine learning algorithms in the WDBC dataset. Different combination of feature selection and classification algorithms produced different results in WDBC and WPBC datasets. The IG produced better results in WDBC on comparison with the other feature selection techniques. In WPBC, the IG and Clustered Feature Selection (CFS) obtained superior results (Modi and Ghanchi, 2016). The Ant Colony Optimization (ACO) based feature selection algorithm was suggested to eliminate the redundant data and to preserve the relevant data. The accuracy of detecting the relevant data is enhanced using the new heuristic information measure. The proposed method excelled the traditional univariate and multivariate filtering methods. The ACO technique is used to rank the features present in the graph represented search space. The ranking of the features were done based on the similarity of the subsets (Tabakhi and Moradi, 2015).

A Genetic Algorithm (GA) based feature selection algorithm is proposed for dimensionality reduction to improve the diagnosis of breast cancer. The parameters are optimized using the Artificial Neural Network (ANN) techniques including Gradient descent with momentum

(GAANN_GD),resilient back-propagation (GAANN_RP), and Levenberg–Marquardt (GAANN_LM). Among the three techniques the GAANN_RP resulted in high accuracy (Ahmad, Isa, Hussain, Osman, and Sulaiman, 2015). A weighted vote based ensemble is proposed for the classification of breast cancer disease. The classifiers such as Decision tree using Gini index and information gain, naïve bayes, SVM and memory based learner were combined for heterogeneous classification. The classification accuracy is enhanced by the application of feature selection and preprocessing techniques. The features from various online breast cancer repositories are utilized for simulation (Bashir, Qamar, and Khan, 2015). The Independent Component Analysis (ICA) is proposed to reduce the features of the WDBC dataset using the decision support system. The reduced dataset is classified using ANN, K-Nearest Neighbor (KNN), SVM, and Radial Basis Function Neural Network (RBFNN). The proposed system classifies the tumors either as malignant or benign in an accurate manner (Mert, Kılıç, Bilgili, and Akan, 2015). A wrapper approach based GA feature selection is suggested for feature selection in various datasets including WDBC, Wisconsin Breast Cancer Dataset (WBC) and Wisconsin prognosis breast cancer (WPBC). The ANN, Particle Swarm (PS) and GA based classifiers are applied to identify the severity of breast cancers. On comparison with the conventional methods, the suggested model attains high accuracy, specificity and sensitivity (Aalaei, Shahraki, Rowhanimanesh, and Eslami, 2016). Ant Colony Optimization is applied for selecting the relevant subset of features from the WDBC dataset.

The features are chosen by different combinatorial optimization problem based on the behavior of ants. ACO Inspecting strategy considers the possibility of Ant Colony Optimization (ACO) to address the issue of class irregularity that happens often is proposed. The above study demonstrates that there is the need of cross breed strategy as a solitary approach is definitely not adequate or persuading for early stage location (Yu, Ni, and Zhao, 2013). The breast cancer recognition is done using the local linear wavelet neural network. The performance of the training parameters is improved using Recursive Least Square (RLS) approach. The suggested model reveals the connection weights of the neurons in the hidden and output layer. This method is found to be robust on comparison with the conventional methods (Senapati, Mohanty, Dash, and Dash, 2013).

The artificial metaplasticity algorithm is applied for classifying breast cancer. The Shannon's information theory is integrated to train the dataset and the performance of the multilayer perceptron was maximized. The results are compared with the Back propagation Algorithm and found to be better (Marcano-Cedeño, Quintanilla-Domínguez, and Andina, 2011). The datasets are classified using a novel hybrid ensemble approach to cluster the features via unsupervised learning techniques. The clustering techniques such as parallel neural-based strong clusters fusion and parallel neural network based data fusion are integrated to enhance the clustering efficiency. It utilizes Wisconsin, Pima Indian Diabetics and digital database for screening mammograms for performance analysis (Verma and Hassan, 2011). The Genetic Algorithm (GA) as applied for recognizing the cancer patterns of breast cancer datasets. The decision rules are incorporated for extracting the required patterns. The accuracy is enhanced and the simplicity is achieved using the rule extraction approach (T.-C. Chen and Hsu, 2006). A Shapely Values Embedded Genetic Algorithm (SVEGA) is proposed to reduce the dimensionality of gene data for breast cancer diagnosis. Two operators, namely, include and remove are applied to select the genes for accuracy enhancement. The classification was performed using NB, SVM,J48 and KNN classifiers (Sasikala, alias Balamurugan, and Geetha, 2015).

An ensemble based feature selection is applied to classify the types of lung cancers using machine learning methods. The suggested method helped in the accurate clinical diagnosis process (Cai et al., 2015). The performance of machine learning algorithms including SVM, C4.5, and NB are evaluated using WDBC dataset. A 10-fold cross validation technique is used to measure the accuracy of the classification algorithms. The results proved that the fusion of classifiers improve the accuracy (Gatuha and Jiang, 2015),(Venkatesan and Velmurugan, 2015).

Materials and Methods

The overall flow of the MBA feature selection and the RF classification for efficient breast cancer diagnosis are utilized. A modified Bat algorithm is proposed for feature selection from the WDBC dataset and the mined subset of features are classified using the Random Forest (RF) classification. Figure 1 shows the overall flow of the proposed MBA feature selection with RF classification for breast cancer diagnosis.

Modified Bat Algorithm

Bat algorithm is one of the bio-inspired algorithms that functions based on the echolocation characteristics of bats. This algorithm is enhanced by the incorporation of Chi-Square feature selection for selecting the best features in a random manner. Bat optimization algorithm is proposed for the selection of appropriate features from the WDBC dataset. This is one of the optimization algorithms that optimize the features of the breast cancer dataset to increase the accuracy of final results. Generally, echolocation is the practice used by microbats to identify their prey and their mates. In order to detect the obstacles in their way of travel and to detect their destination the bats generate a sound pulse in the range of 20Hz to 150 Hz. The obstacles or the prey reflects an echo on the reception of the sound pulses. According to the signal strengths of the received echo, the bat classifies whether the echo is from the prey or the obstacle. The distance and the location of the objects can also be easily computed using the strength of signals. Similarly, the essential features are selected from the dataset based on the types of attributes and the irrelevant features are eliminated by considering them as

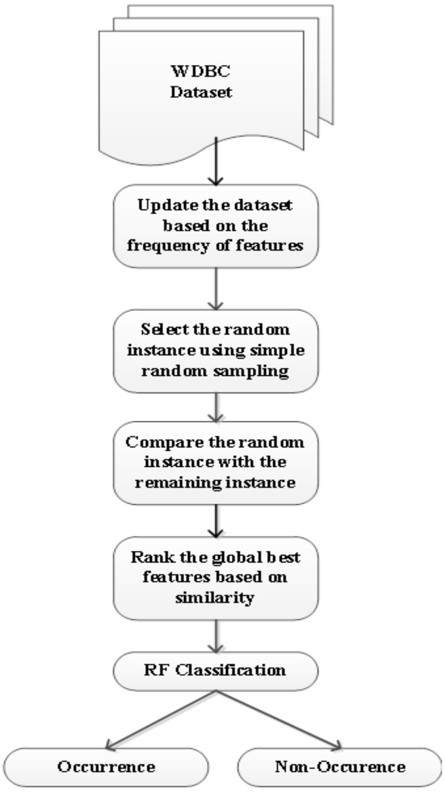

WDBC Dataset
Update the dataset based on the frequency of features
Select the random instance using simple random sampling
Compare the random instance with the remaining instance
Rank the global best features based on similarity
RF Classification
Occurrence
Non-Occurence

Figure 1. Overall Flow Diagram of the Proposed MBA Feature Selection with RF Classification

obstacles.

In the modified bat algorithm, the WDBC dataset is considered as the initial Bat population and each data has a certain frequency ω_i and velocity v_i. The frequency and velocity are estimated using the following equations and updated after each iteration.

$$\omega_i = \omega_{min} + (\omega_{max} - \omega_{min})\ \alpha$$

$$v_i^t = v_i^{t-1} + (G_i^t - G_{current})\ \omega_i$$

where, $G_{current}$ represents the present global solution and α denotes the uniform distribution function that ranges between 0 and 1.

Then, local search is performed to determine the best solutions using random walks within the dataset. Random best solutions are identified by simple random sampling technique to extract the predominant features from the entire dataset that contains a large number of features. The features are grouped together based on the similarity and ranked by resampling method. The current solution is compared with the ranked solutions and they are sorted in the best order. The solutions are updated using the following equation:

$$G_{best} = G_{current} + \epsilon\ S_i$$

where, ϵ is a random number that ranges between -1 to 1 and S_i signifies the similar features.

The proposed Bat optimization algorithm is more efficient which produce faster results on comparison with Particle Swarm Optimization (PSO) and other meta heuristic algorithms. The search area of the Bats in the datasets are utilized and supported by numerous applications. The quantitative features are essential for getting the exact results in data mining.

```
Modified Bat Algorithm for Feature Optimization
Input:   WDBC dataset
Output:  Optimized features
Initialize the dataset: $X_i$ where $1 \leq i \leq n$
Set Frequency $\omega_i$ velocity $v_i$
While t <X_n Do
   Update $\omega_i$ and $v_i$
   Estimate transfer function value
   Update $X_i$
   Generate rand
   Select random instance from $X_i$
   Compute similarity
For (rand=X_i; rand<=X_n; =X_i++)
   If (rand is highly similar than the i^th instance)
   Choose $G_{best}$ from the available solutions.
      Update $X_i$ with $G_{best}$
   End
Sort the features based on $G_{best}$
End
```

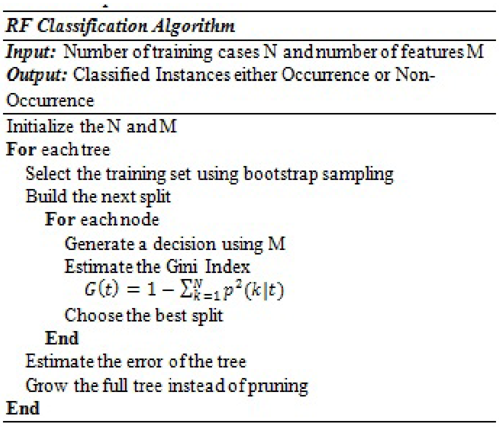
Random Forest Classification

RF is defined as a general principle of randomized ensembles of decision trees (Breiman, 2001). Generally, RF is built by the recursive partitioning of binary tree into similar nodes. The similarity of the child node is enhanced by the inheritance of data from the parent node. The original data is sampled by bootstrap sampling to generate a large number of trees to grow RF. Each tree in the RF generates a response with respect to the set of predictor values provided as input to those trees. The missing values in the dataset can be easily managed by the RF. The steps for constructing the random forest are described below:

```
RF Construction
Step 1: Draw 'n' tree bootstrap samples from the original data.
Step 2: Develop a tree for each bootstrap data set. At each node of the tree, randomly select 'm' variables for splitting the node into two child nodes. Grow the tree until the minimum node size is reached.
Step 3: Aggregate information from the 'n' trees for new data prediction such as majority voting for classification.
Step 4: Compute an Out-Of-Bag (OOB) error rate using the data that is not present in the bootstrap sample.
```

In the proposed method, the features selected via the modified Bat algorithm are provided as input to the RF to train the classifier. The accuracy of the classification is improved by the optimized features utilized for training. The training cases are split using bootstrap sampling in which the next split is selected according to the Gini index. The trees are fully grown until there is no decrease in the error. The main advantage of RF is that it can handle heterogeneous types of data and ease identification of outliers. It is not highly sensitive at the same time it has a large computational scalability (Montillo, 2009).

The RF algorithm used to train and classify the WDBC dataset is explained as follows:



```
RF Classification Algorithm
Input: Number of training cases N and number of features M
Output: Classified Instances either Occurrence or Non-Occurrence
Initialize the N and M
For each tree
    Select the training set using bootstrap sampling
    Build the next split
        For each node
            Generate a decision using M
            Estimate the Gini Index
                $G(t) = 1 - \sum_{k=1}^{N} p^2(k|t)$
            Choose the best split
        End
    Estimate the error of the tree
    Grow the full tree instead of pruning
End
```

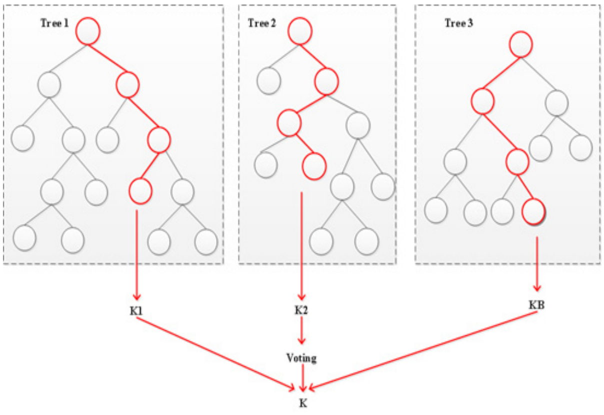


Figure 2. General Architecture of RF Algorithm

Results

Dataset description

The performance of the proposed work is evaluated by using the Wisconsin Diagnosis Breast Cancer (WDBC) dataset (Zwitter and Soklic). This dataset includes two classes having 201 instances and 85 instances that are described by nine attributes of linear and nominal type. Table I shows the attribute information of breast cancer dataset. The WEKA tool is used for simulation purpose.

Performance metrics

The metrics used for evaluating the performance of the proposed work are described below:

Correctly Classified Instances

It is the amount of cells that are correctly classified as normal or cancerous cells.

Incorrectly Classified Instances

It is the amount of normal cells that are wrongly classified as cancerous cells.

Table 1. Attribute Information of Wisconsin Breast Cancer Dataset

Attribute number	Attribute description	Attribute value
1	Class	No-recurrence and recurrence events
2	Age	10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90-99
3	Menopause	lt40, ge40, premeno
4	Tumor size	0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59
5	inv-nodes	0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26, 27-29, 30-32, 33-35, 36-39
6	Node-caps	Yes and no
7	deg-malig	1, 2, 3
8	Breast	Left and Right
9	Breast-quad	Left-up, left-low, right-up, right-low, central
10	Irradiat	Yes: No

Kappa statistic

It measures the agreement of prediction with the true class. If the kappa value is 1, there is a complete agreement with the true class.

Mean Absolute Error (MAE)

It is a measure of closeness of the detection result to the eventual results.

Root Mean Square Error (RMSE)

The MAE and RMSE are used together to diagnose the variation in the errors in a set of detection results.

Relative Absolute Error (RAE)

It is a measure of the variability of the detected results to the actual results.

Root Relative Squared Error (RRSE)

It is the measure of error rate to the variability of the actual values.

True Positive Rate (TPR) or Recall

TPR is the proportion of the benign cells that are correctly identified.

False Positive Rate (FPR)

FPR is the proportion of malignant cells that are incorrectly classified as benign.

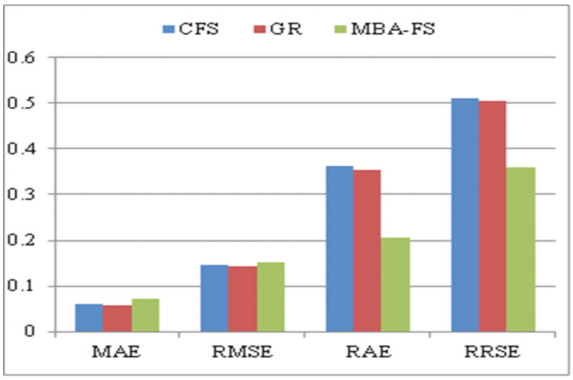


Figure 3. Comparative Analysis of Error Rate for the Proposed MBA-FS and Existing CFS and GR

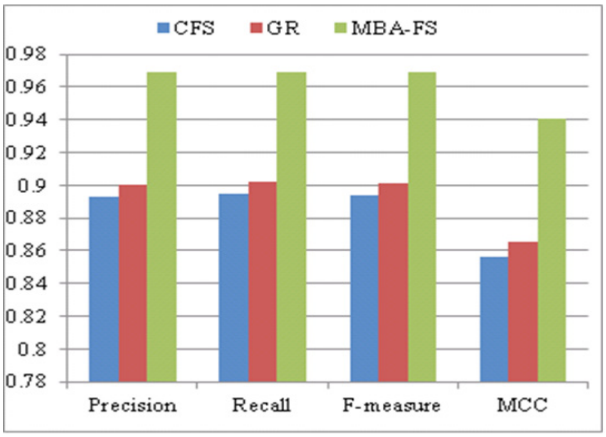


Figure 4. Accuracy Analysis of the Proposed MBA-FS and Existing CFS and GR

Precision

It is the amount of the benign cells that are correctly predicted as normal.

F-measure

It is the harmonic mean value of the precision and recall metrics.

Matthews Correlation Coefficient (MCC)

It is a correlation coefficient between the observed and predicted classification results. If MCC is +1, it represents the perfect classifications and -1 indicates the complete disagreement between the observed and predicted classification results.

Region of Curve (ROC)

It is a graphical curve that illustrates the tradeoff between the TPR and FPR of the classifier.

Precision Recall Curve (PRC) area

It is used to visualize the performance of the classifier in the balanced and imbalanced datasets.

Comparative Analysis

The proposed Modified Bat Algorithm For Feature Selection (MBA-FS) in WDBC dataset is compared with the existing Correlation-based Feature Selection (CFS) (Hall, 1999) and Gain Ratio (GR) (Karegowda, Manjunath, and Jayaram, 2010). Table II shows the comparative analysis of the CFS, GR and proposed approach. From the comparative analysis, we observe that the proposed approach yields higher Kappa statistic, precision, recall, F-measure, MCC and correctly classified rate than the CFS and GR. The RAE, RRSE and incorrectly classified rate of the proposed approach are lower than the existing CFS and GR. The MAE and RMSE of the proposed algorithm are higher than the existing CFS and GR.

Figure 3. shows the comparative analysis of the error rate for the proposed MBA-FS and existing CFS and GR. The proposed MBA-FS yields lower RAE of about 43.39% and 42.16% than the CFS and GR respectively. The RRSE of the proposed MBA-FS is about 29.246% and 28.58% lesser than the CFS and GR. The MAE of the MBA-FS is higher of about 16.94% and 18.63% than the CFS and GR. The MBA-FS provides 3.26% and 4.125% higher RMSE than the CFS and GR. Fig.4 depicts the accuracy analysis of the proposed MBA-FS and existing CFS and GR. The precision of the MBA-FS is 7.84% and 7.12% higher than the CFS and GR. The MBA-FS yields better recall of about 7.636% and 6.914% than the CFS and GR. F-measure of the MBA-FS is 7.74% and 7.02% higher than the CFS and GR. Thus MBA-FS achieves better MCC of 9.03% and 8.076% than CFS and GR.

Table 2. Comparative Analysis of Cfs, Gr and Proposed Approach

Measures	CFS	GR	MBA-FS
Kappa statistic	0.8	0.8	0.9
MAE	0.05	0.05	0.07
RMSE	0.14	0.14	0.15
RAE	36.23%	35.46%	20.51%
RRSE	50.98%	50.51%	36.07%
Precision	0.89	0.9	0.96
Recall	0.89	0.9	0.96
F-measure	0.89	0.9	0.96
MCC	0.85	0.86	0.94
Correctly Classified Instances	256	258	277
Incorrectly Classified Instances	30	28	9
Correctly classified rate	89.51%	90.21%	96.85%
Incorrectly Classified rate	10.49%	9.79%	3.15%

Discussion

In this paper, a Modified Bat Algorithm (MBA) is proposed for feature optimization in the WDBC dataset for efficient diagnosis of breast cancer. The simple random sampling technique is adopted by the Bat algorithm to choose the random instance from the dataset. Local random walk is performed via the picked random instance to identify the global best solution. The best solutions are ranked and these features are used to train the Random Forest (RF) classifier. The RF classification algorithm utilizes the bootstrap sampling to select the training sets. Gini index is estimated to recognize the best split. Finally, the unpruned tree classifies the occurrence and non-occurrence of breast cancer. The performance of the proposed MBA feature selection with RF classification is evaluated and compared with the existing feature selection algorithms such as Correlation-based Feature Selection (CFS) and Gain Ratio (GR). The MBA feature selection outperformed the CFS and GR techniques related to Kappa statistic, Mathew's Correlation Coefficient, Precision, F-measure, Recall, Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Relative Absolute Error (RAE) and Root Relative Squared Error (RRSE). Thus, the overall classification accuracy is enhanced by the MBA feature selection, which in turn makes the clinical diagnosis easier. In future, the feature selection will be improved by changing the sampling technique utilized for random instance selection.

References

Aalaei S, Shahraki H, Rowhanimanesh A, Eslami S (2016). Feature selection using genetic algorithm for breast cancer diagnosis: experiment on three different datasets. *Iran J Basic Med Sci*, **19**, 476-82.

Ahmad F, Isa NAM, Hussain Z, Osman MK, Sulaiman SN (2015). A GA-based feature selection and parameter optimization of an ANN in diagnosing breast cancer. *Pattern Anal Appl*, **18**, 861-70.

Bashir S, Qamar U, Khan FH (2015). Heterogeneous classifiers fusion for dynamic breast cancer diagnosis using weighted vote based ensemble. *Qual Quant*, **49**, 2061-76.

Beam CA, Layde PM, Sullivan DC (1996). Variability in the interpretation of screening mammograms by US radiologists: findings from a national sample. *Arch Intern Med*, **156**, 209-13.

Bird RE, Wallace TW, Yankaskas BC (1992). Analysis of cancers missed at screening mammography. *Radiology*, **184**, 613-17.

Breiman L (2001). Random forests. *Mach Learn*, **45**, 5-32.

Cai Z, Xu D, Zhang Q, et al (2015). Classification of lung cancer using ensemble-based feature selection and machine learning methods. *Mol Bio Sys*, **11**, 791-100.

Chen C-M, Chou Y-H, Han K-C, et al (2003). Breast lesions on sonograms: computer-aided diagnosis with nearly setting-independent features and artificial neural networks 1. *Radiology*, **226**, 504-14.

Chen T-C, Hsu T-C (2006). A GAs based approach for mining breast cancer pattern. *Expert Syst Appl*, **30**, 674-81.

Cheng H, Shan J, Ju W, Guo Y, Zhang L (2010). Automated breast cancer detection and classification using ultrasound images: A survey. *Pattern Recognit*, **43**, 299-17.

Devijver PA, Kittler J (1982). Pattern recognition: A statistical approach: Prentice hall. First Editon, pp1-448.

Elmore JG, Wells CK, Lee CH, Howard DH, Feinstein AR (1994). Variability in radiologists' interpretations of mammograms. *N Engl J Med*, **331**, 1493-99.

Garfinkel L, Catherind M, Boring C, Heath C (1997). Change trends: an overview of breast cancer incidence and mortality. *Cancer*, **74**, 222-27.

Gatuha G, Jiang T (2015). Evaluating diagnostic performance of machine learning algorithms on breast cancer. Paper presented at the IScIDE.

Ghazavi SN, Liao TW (2008). Medical data mining by fuzzy modeling with selected features. *Artif Intell Med*, **43**, 195-06.

Giger ML (2002). Computer-aided diagnosis in radiology. *Academic Radiol*, **9**, 1-3.

Guyon I, Elisseeff A (2003). An introduction to variable and feature selection. *J Mach Learn Res*, **3**, 1157-82.

Guyon I, Elisseeff A (2006). An introduction to feature extraction. Feature extraction, pp1-25.

Haka AS, Shafer-Peltier KE, Fitzmaurice M, et al (2005). Diagnosing breast cancer by using Raman spectroscopy. *PNAS USA*, **102**, 12371-76.

Hall MA (1999). Correlation-based feature selection for machine learning. The Univ of Waikato. Version 3, pp1-178.

Huang Y-L, Wang K-L, Chen D-R (2006). Diagnosis of breast tumors with ultrasonic texture analysis using support vector machines. *Neural Comput Appln*, **15**, 164-69.

Jalalian A, Mashohor SB, Mahmud HR, et al (2013). Computer-aided detection/diagnosis of breast cancer in mammography and ultrasound: a review. *Clin imaging*, **37**, 420-26.

Jesneck JL, Lo JY, Baker JA (2007). Breast mass lesions: computer-aided diagnosis models with mammographic and sonographic descriptors 1. *Radiology*, **244**, 390-98.

Joo S, Yang YS, Moon WK, Kim HC (2004). Computer-aided diagnosis of solid breast nodules: use of an artificial neural network based on multiple sonographic features. *IEEE Trans Med Imaging*, **23**, 1292-00.

Karegowda AG, Manjunath A, Jayaram M (2010). Comparative study of attribute selection using gain ratio and correlation based feature selection. *IJITKM*, **2**, 271-77.

Karssemeijer N (1992). Stochastic model for automated detection of calcifications in digital mammograms. *IVC*, **10**, 369-75.

Kegelmeyer Jr WP (1993). Evaluation of stellate lesion detection in a standard mammogram data set. *IJPRAI*, **7**, 1477-92.

Kegelmeyer Jr WP, Pruneda JM, Bourland PD, et al (1994). Computer-aided mammographic screening for spiculated lesions. *Radiology*, **191**, 331-37.

Kobatake H, Murakami M, Takeo H, Nawano S (1999). Computerized detection of malignant tumors on digital mammograms. *IEEE Trans Med Imag*, **18**, 369-78.

Lee, S-K, Chung P-c, Chang C-I, et al (2003). Classification of clustered microcalcifications using a Shape Cognitron neural network. *Neural Net*, **16**, 121-32.

Liu S, Babbs CF, Delp EJ (2001). Multiresolution detection of spiculated lesions in digital mammograms. I*EEE Trans on Image Proc*, **10**, 874-84.

López Y, Novoa A, Guevara MA, Silva A (2007). Breast cancer diagnosis based on a suitable combination of deformable models and artificial neural networks techniques. Paper presented at the CIARP.

Marcano-Cedeño A, Quintanilla-Domínguez J, Andina D (2011). Breast cancer classification applying artificial metaplasticity algorithm. *Neurocomputing*, **74**, 1243-50.

Mert A, Kılıç N, Bilgili E, Akan A (2015). Breast cancer detection with reduced feature set. *Comput Math Methods Med*, Article ID 265138, 1-11.

Modi N, Ghanchi K (2016). A comparative analysis of feature selection methods and associated machine learning algorithms on Wisconsin breast cancer dataset (WBCD). paper presented at the ICT4SD.

Montillo AA (2009). Random Forests. from http://www.dabi.temple.edu/~hbling/8590.002/Montillo_RandomForests_4-2-2009.pdf.

Mudigonda NR, Rangayyan RM, Desautels JL (2001). Detection of breast masses in mammograms by density slicing and texture flow-field analysis. *IEEE Trans Med Imaging*, **20**, 1215-27.

Onan A (2015). A fuzzy-rough nearest neighbor classifier combined with consistency-based subset evaluation and instance selection for automated diagnosis of breast cancer. *Expert Syst Appl*, **42**, 6844-52.

Pérez NP, López MAG, Silva A, Ramos I (2015). Improving the Mann–Whitney statistical test for feature selection: An approach in breast cancer diagnosis on mammography. *Artif Intell Med*, **63**, 19-31.

Petrick N, Chan H-P, Sahiner B, Wei D (1996). An adaptive density-weighted contrast enhancement filter for mammographic breast mass detection. *IEEE Trans Med Imaging*, **15**, 59-67.

Saeys Y, Inza I, Larrañaga P (2007). A review of feature selection techniques in bioinformatics. *bioinformatics*, **23**, 2507-17.

Sasikala S, alias Balamurugan SA, Geetha S (2015). A novel feature selection technique for improved survivability diagnosis of breast cancer. *Procedia Comp Sci*, **50**, 16-23.

Senapati MR, Mohanty AK, Dash S, Dash PK (2013). Local linear wavelet neural network for breast cancer recognition. *Neural Comput Appl*, **22**, 125-31.

Soltanian-Zadeh H, Rafiee-Rad F (2004). Comparison of multiwavelet, wavelet, Haralick, and shape features for microcalcification classification in mammograms. *Pattern Recognit*, **37**, 1973-86.

Song JH, Venkatesh SS, Conant EF, et al (2005). Artificial neural network to aid differentiation of malignant and benign breast masses by ultrasound imaging. Paper presented at the Med Imaging.

Strickland RN, Hahn HI (1996). Wavelet transforms for detecting microcalcifications in mammograms. *IEEE Trans on Med Imaging*, **15**, 218-29.

Sun Y, Babbs C, Delp E (2006). A comparison of feature selection methods for the detection of breast cancers in mammograms: adaptive sequential floating search vs. genetic algorithm. Paper presented at the IEEE-EMBS 2005.

Tabakhi S, Moradi P (2015). Relevance–redundancy feature selection based on ant colony optimization. *Pattern Recognit*, **48**, 2798-11.

Vanaja S, Kumar KR (2014). Analysis of feature selection algorithms on classification: a survey. *Int J Comput Appl*, **96**, 28-35.

Veldkamp, WJ, Karssemeijer N, Otten JD, Hendriks JH (2000). Automated classification of clustered microcalcifications into malignant and benign types. *Med Phys*, **27**, 2600-8.

Venkatesan E, Velmurugan T (2015). Performance analysis of decision tree algorithms for breast cancer classification. *Indian J Sci Technol*, **8**, 1-8.

Verma H (2011). Hybrid ensemble approach for classification. *Appl Intell*, **34**, 248-78.

Warren Burhenne LJ, Wood SA, D'Orsi CJ, et al (2000). Potential contribution of computer-aided detection to the sensitivity of screening mammography 1. *Radiology*, **215**, 554-62.

Wei J, Sahiner B, Hadjiiski LM, et al (2005). Computer-aided detection of breast masses on full field digital mammograms. *Med Phys*, **32**, 2827-38.

West D, Mangiameli P, Rampal R, West V (2005). Ensemble strategies for a medical diagnostic decision support system: A breast cancer diagnosis application. *Eur J Oper Res*, **162**, 532-51.

Yu H, Ni J, Zhao J (2013). ACOSampling: An ant colony optimization-based undersampling method for classifying imbalanced DNA microarray data. *Neurocomputing*, **101**, 309-18.

Zwitter M, Soklic M (1988). UCI Machine Learning Repository Breast cancer dataset. . from https://archive.ics.uci.edu/ml/datasets/Breast+Cancer.