ISSN (Print) : 0974-6846
Indian Journal of Science and Technology, *Vol 9(45), DOI: 10.17485/ijst/2016/v9i45/102290, December 2016* ISSN (Online) : 0974-5645

A Multi-objective Non-Dominated Sorted Artificial Bee Colony Feature Selection Algorithm for Medical Datasets

Bhuvaneswari Ragothaman and B. Sarojini

Department of Computer Science, Avinashilingam Institute for Home Science and Higher Education for Women - University, Coimbatore - 641043, Tamil Nadu, India; rbeswari@gmail.com, dr.b.sarojini@gmail.com

Abstract

Objective: There exists a huge amount of heterogeneous data in medical databases which when mined may provide valuable information for medical diagnosis. The processing of this voluminous data is tedious, owing to its high dimensionality. The presence of irrelevant and redundant data may reduce the performance of data mining algorithms. The data preprocessing technique, feature selection, removes any irrelevant or redundant features and selects discriminative features. The objective of this research work is to propose a feature selection algorithm that meets two objectives, 1. Reduce the number of features and 2. Maximize the classification accuracy. **Methods/Statistical Analysis:** The research work proposes a feature selection algorithm Multi-Objective Non-Dominated Sorted Artificial Bee Colony Algorithm (NSABC) that combines Pareto optimization and Artificial Bee Colony (ABC) algorithm for selecting the Non-Dominant optimal feature subsets of three medical datasets viz. Wisconsin Breast Cancer, Pima Indian Diabetes and Statlog Heart disease datasets. The features selected by the ABC algorithm are further optimized by applying Pareto optimization. The feature subsets selected are validated by evaluating the performance of KNN classifier in terms of the classification metrics Precision, Recall and Accuracy before and after feature selection. **Findings:** The percentage of feature reduction and the improved performance of the classifier prove that the proposed feature selection method has selected the discriminate features of the datasets. The selected feature subsets are further validated by calculating the entropy, the measure signifying the individuality and independent nature of the features in the feature subset. **Application/Improvement:** The reduced feature subset selected is used in the easy diagnosis of the disease during the medical diagnosis. This also helps in cost and time reduction for the medical diagnosis of the disease.

Keywords: Data Mining, Entropy Feature Selection, KNN Classifier, Non-Dominated Sorted Artificial Bee Colony Algorithm (NSABC), Pareto Optimization

1. Introduction

In medical domain, there is a huge volume of heterogeneous data, the patient information, drug details, electronic medical record like ECG, MRI Scan report, etc. are available that could be analyzed to discover patterns for improved diagnosis. The discovered patterns provide valuable knowledge for medical discoveries, for example identification of combinations of features that would lead to diagnosis of the disease. The correctness of the diagnostic predictive models, formed out of discovered patterns, depend on the quality of the data that is used for analysis. Many data mining techniques such as classification and clustering are proved to degrade prediction accuracy when trained on data sets containing redundant or irrelevant features[1]. Researchers realized that in order to use data mining tools on these medical databases effectively, data preprocessing is essential[2]. The application of efficient and sound data preprocessing procedures could reduce the amount of data to be analyzed without losing any critical information, improve the quality of the data, enhance the performance of the actual data mining algorithms

* *Author for correspondence*

and reduce the execution time of mining algorithms[3]. Irrelevant and redundant data are removed from the dataset by applying data preprocessing techniques. This maps the high dimensional data to low dimensional data there by reducing the storage space and search space which helps in easy analysis and visualization of the data. Removal of redundant and irrelevant data improves the performance of machine learning algorithms.

There are two preprocessing techniques: Feature Selection and Feature Extraction. Feature Extraction reduces the number of features by combining the features of the original dataset and forming a new feature subset. Feature Selection reduces the number of features by selecting relevant features that are required for the specified task. Feature Selection technique is suitable for the medical domain as it maintains the original semantics of the features which help in easy interpretability by the domain expert[4]. Though a number of feature selection methods that enhance the performance of the mining algorithm are available, still the research goes on in reducing the number of features and to identify more informative features of the datasets. The feature subset that enhances the performance of classifier is the optimal feature subset of the dataset.

In this research work, a feature selection approach, Pareto Optimization method combined with Artificial Bee Colony Algorithm is used for selecting the feasible features from the three medical datasets viz. Wisconsin Breast Cancer Dataset (WBCD), PIMA Indian Diabetics Dataset (PIMA) and Statlog Heart Disease Dataset (Heart). The selected optimal feature subsets of these datasets are validated by analyzing the classification performance of the KNN Classifier. The metrics of the classifier Accuracy, Precision and Recall are evaluated. The characteristics of the features in the selected feature subset are further analyzed by calculating the entropy of the features and feature subset.

Some of the related papers reviewed for this research work are: A two method feature selection technique using Pareto Optimization along with Particle Swam Optimization Algorithm[5]. In first algorithm the Non-Dominated Feature Subset is selected by comparison with the Pareto front value. In second algorithm the Non-Dominated Feature Subset is again sorted with comparison with the crowding distance and mutation of the features. Finally, a minified feature subset is generated and that feature subset increases the performance of the KNN classifier with an increase in the classification accuracy.

The feature selection approach proposed in[6] is a filter based feature selection that has combined the fuzzy mutual information and mutual information along with Artificial Bee Colony algorithm. In that algorithm fuzzy mutual information approach is used to find the mutual information in the datasets and the features are selected using the ABC algorithm. The selected features are used for the classification and the accuracy of the classification is increased when compared to all features in the dataset.

The feature selection approach proposed in[7] has combined Pareto optimization along with Genetic Algorithm. The algorithm overcomes the problems like computational complexity, need for parameter sharing. Those problem sorted by the feature subset selected using Pareto optimization. The feature subset selected using GE algorithm is compared with the Pareto front value and the Non-Dominated Feature Subset selected.

The feature selection method in[8] has proposed a method for restore of the distribution of system using multi objective with the artificial immune systems combined with the ant colony optimization algorithms. This maintains the population and helps in the restoring using the hyper mutation of the existing antibodies. This obtains the quick solution for the restoring of the distributed system in one or more networks.

2. Proposed Work

Multi Objective Optimization is a mathematical method of optimizing more than one objective function one after the other. The feature selection problem has two objectives that need to be solved, one is minimizing the number of features and the next one is to maximize the classification accuracy[9]. A multi-objective Non-Dominated Sorted Artificial Bee Colony (NSABC) algorithm is proposed for selecting the discriminative features of three medical datasets. The WBCD, PIMA, Heart disease datasets from UCI Repository are considered for experiments[10]. The NSBAC algorithm is implemented in Java language using NetBeans IDE. The proposed work has two main phases, 1. Optimal feature subset selection and 2. Validation of the selected feature subsets by Classification and Entropy calculation. Figure 1 shows the frame work of the Non-Dominated Sorted Artificial Bee Colony (NSABC) algorithm.

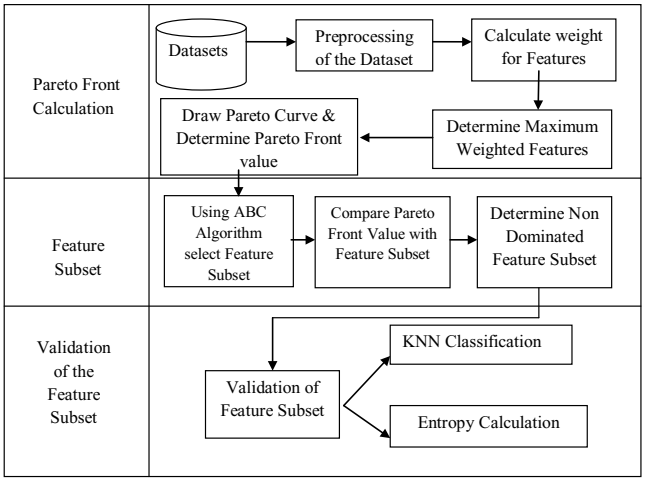

Figure 1. Framework for non-dominated feature sorted Artificial Bee Colony algorithm.

Artificial Bee Colony algorithm defined by[11] is based on the intelligent behavior of the honey bees in finding the best food source. This algorithm is used for selecting the discriminative features of the datasets. The weight or fitness value or weight of the features is calculated using the fitness function:

$$\text{Fitness function: fiti} = \begin{cases} \frac{1}{1+f_i} \; if \geq 0 \\ 1+abs(f_i) iff_i < 0 \end{cases}$$

Here, f_i is the feature in the original dataset. The features with the maximum weight when compared to the average weight of all features are considered to draw the Pareto curve. The Pareto front value, the centroid of the Pareto Curve, is calculated using the formula:

$$\text{Pareto Front Value} = \frac{f_{1+f2+\ldots+f_k}}{k}$$

The Non-Dominated Feature Subsets selected are the feasible feature subsets with minimal number of optimized features of the medical datasets. Figure 2 gives the pseudocode of the NSABC Algorithm.

In phase 2, the Non-Dominated Feature Subset Selected using NSABC is validated by analyzing the performance of K Nearest Neighbor classifier before and after feature selection is compared[12]. The KNN classifies the features using the distance. The distance between the features is calculated as:

$$\text{D (a, b)} = \sqrt{\sum_{i=1}^{n} (b_i - a_i)^2}$$

begin
Divide dataset into training and test set;
Initialize solution set X = X1;X2; :::;Xn by Eq. (1);
Evaluate two objectives of solutions;
Apply non-dominated sorting to solutions;
for cycle 1 to MCN do
foreach employed bee i do
Randomly choose a solution Xk in the neighborhood of Xi;
Add evolved solutions to X;
end
Apply non-dominated sorting on X;
Select best SN solutions based on rank and crowding distance to renew population;
Foreach onlooker bee i do
Select a food source Xi depending on probability pi;
Randomly choose a solution Xk in the neighborhood of Xi;
Add evolved solutions to X;
End
Apply non-dominated sorting on X;
Select best SN solutions based on rank and crowding distance to
renew population;
if there exits an abandoned solution then
Scout bee determines a new solution;
end
end
Calculate the classification accuracy of the feature subsets (solutions) in
the Front 1 on the test set;
Return the solutions and their classification accuracy rates;
end

Figure 2. Pseudocode of NSABC algorithm.

Where D is the distance between the features a and b. The performance is evaluated by calculating the classification metrics like Precision, Recall and Accuracy[13–15].

Precision is the number of correctly predicted positive instances divided by the sum of all correctly predicted instances. Precision is given by the formula:

$$\text{Precision} = \frac{Tp}{Fp+Tp}$$

Recall is the number of correctly predicted positive instances divided by sum of number of correctly predicted positive instance and incorrectly predicted negative instance. Recall is given by the formula:

$$\text{Recall} = \frac{Tp}{Fn+Tp}$$

Accuracy is the number of correctly predicted positive

and negative instances divided by the total number of instances present in dataset. Accuracy is given by the formula:

$$\text{Accuracy} = \frac{Tp + Tn}{Fp + Fn + Tp + Tn}$$

3. Empirical Results and Discussions

The empirical results of the Non-Dominated Sorted ABC algorithm are shown in Table 1. The results show that there is a significant reduction in the number of features selected. Out of 10 features in WBCD only 3 features Uniformity cell shape, Single Epithelial Cell Size, Bland Chromatin are selected as discriminative features. Out of 8 features in PIMA only 2 features Plasma glucose and Diastolic blood pressure are selected as discriminative features. Out of 13 features Age and Serum Cholesterol are selected as discriminative features.

Table 1. Empirical results of NSABC feature selection

Datasets	# Features	Non-Dominated Feature Subset Selected	% of Feature Reduction
Wisconsin Breast Cancer Dataset	10	3	70
PIMA Indian Diabetics Dataset	8	2	75
Statlog Heart Disease Dataset	13	2	84.62

To validate the Non-Dominated Feature Subset Selected the performance of the KNN Classifier is evaluated. Precision, Recall, Accuracy are calculated. In medical domain these metrics are very important as they tell how well the classifier behaves for the given feature subset.

The results clearly indicate a very high improvement in all three metrics as shown in Table 2. The accuracy has improved by 24% for WBCD, 30% for PIMA and 23% for Heart disease datasets. The results prove that the classification of positive and negative cases is possible with these numbers of features and the NSABC algorithm has chosen highly discriminate features of the dataset. Figures 3, 4 and 5 show the diagrammatic representation of comparison of the classification metrics of KNN Classifier before and after feature selection.

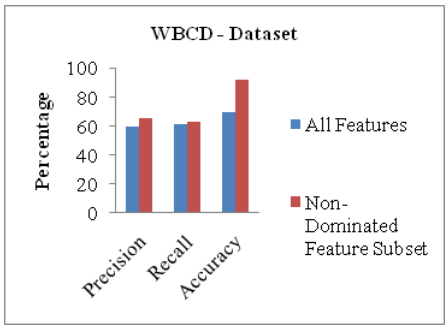

Figure 3. Comparison for WBCD dataset.

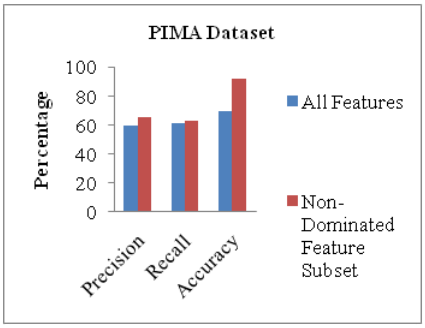

Figure 4. Comparison for PIMA dataset.

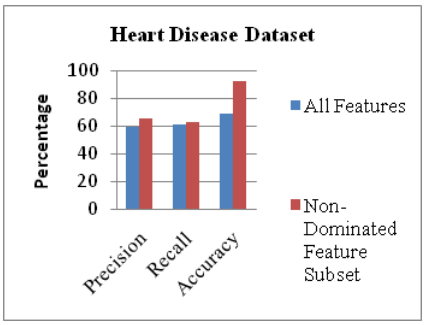

Figure 5. Comparison for heart disease dataset.

Table 2. Classification metrics for whole and Non-Dominated Feature Subset

Datasets	Whole Feature Set			Non-Dominated Feature Subset		
	Precision	Recall	Accuracy	Precision	Recall	Accuracy
Wisconsin Breast Cancer Dataset	66.481	47.889	67.618	80.817	76.988	91.201
PIMA Indian Diabetics Dataset	62.069	69.945	60.955	89.015	76.197	91.369
Statlog Heart Disease Dataset	60	61.481	69.619	65.686	64.511	92.815

4. Entropy Calculation

Entropy is used calculate the homogeneity of the features in the Non-Dominated Feature Subset selected. If the value of entropy is zero it shows that the features are fully homogeneity and if the value of entropy is one, it shows that the features are fully heterogeneity.

$$\text{Entropy} = -\sum_{i=1}^{k} p\left(Value\ i\right).\log_2\left(p\left(Value\ i\right)\right)$$

Entropy shows the independent and individuality of the features in the feature subset. The Entropy values for all three Non-Dominated Sorted ABC feature subsets datasets are shown in Table 3.

Table 3. Entropy values of the dataset

Non-Dominated Sorted ABC feature subsets	Entropy Value
Wisconsin Breast Cancer with 3 features	0.307
PIMA Indian Diabetics Dataset with 2 features	0.383
Statlog Heart Disease Dataset with 2 features	0.194

5. Conclusion

In this research work, Pareto Optimization is implemented combined with Artificial Bee Colony Algorithm to select a Non-Dominated Feature Subset. To validate the Non-Dominated Feature Subset selected the performance of the KNN Classifier is evaluated by calculating the classification metrics like Precision, Recall and Accuracy. The increase in classification accuracy for Non-Dominated Feature Subset proves that the feasible feature subset is selected. It is further validated by calculating the entropy of the feature subset, which determines the individuality and independent nature of the features in the feature subset. In future, along with classification, clustering can also be done for the selected Non-Dominated Feature Subset. Some statistics methods can also be used to validate the selected feature subset.

6. References

1. Lesh N, Mohammed J, Zaki Z, Ogihara M. Scalable feature mining for sequential data. IEEE Intelligent Systems. 2002; 15(2):1–5.
2. Cios KJ, Moore GW. Uniqueness of medical data mining. Artif Intell Med. 2002; 26(1): 1–24.
3. Balakrishnan S, Narayanasamy S. Enhancing the performance of LibSVM classifier by kernel F-score feature selection. S. Ranka et al. editors. Springer-Verlag Berlin Heidelberg: IC3 CCIS. 2009; 40:533–43.
4. Kononenko I, Bratko I, Kukar M. Application of machine learning to medical diagnosis. Machine Learning and Data Mining: Methods and Applications; 1998. p. 389–408.
5. Xue X, Zhang M, Will N, Browne B. Particle Swarm Optimization for feature selection in classification: A multi-objective approach. IEEE Transactions on Cybernetics. 2013; 43(6):1656–71.
6. Hancer E, Xuey B, Zhangy M, Karaboga D, Akay B. A multi-objective Artificial Bee Colony approach to feature selection using fuzzy mutual information. New Zealand: IEEE; 2015.
7. Deb K, Pratap A, Agarwal S, Meyarivan T. A fast and elitist multi objective Genetic Algorithm: NSGA-II. IEEE Transactions on Evolutionary Computation. 2002; 6(2);182–97.
8. Ahuja A, Das S, Pahwa A. An AIS-ACO hybrid approach for multi-objective distribution system reconfiguration. IEEE Transactions on Power Systems. 2007; 22(3):1101–11.
9. Asuncion A, Newman DJ. UCI machine learning repository. 2007.
10. Karaboga D. An idea based on honey bee swarm for numerical optimization. Technical Report TR06. Engineering Faculty, Computer Engineering Department: Erciyes University; 2005. p. 1–10.
11. Chakradeo SN, Abraham RM, Rani BA, Manjula R. Data Mining: Building social network. Indian Journal of Science and Technology. 2015 Jan; 8(S2). DOI: 10.17485/ijst/2015/v8iS2/60482.
12. Ananthapadmanabhan IKR, Parthiban G. Prediction of chances - Diabetic retinopathy using data mining classification techniques. Indian Journal of Science and Technology. 2014 Jan; 7(10):1–6.
13. Rajalakshmi V, Mala GSA. Anonymization by data relocation using sub-clustering for privacy preserving data mining. Indian Journal of Science and Technology. 2014 Jan; 7(7):976–80.
14. Mohammadi M, Nastaran M, Sahebgharani A. Sustainable spatial land use optimization through non-dominated sorting Genetic Algorithm-II (NSGA-II): (Case Study: Babol-dasht District of Isfahan). Indian Journal of Science and Technology. 2015 Feb; 8(S3):118–29.
15. Sharma M, Singh SK, Agrawal P, Madaan V. Classification of clinical dataset of cervical cancer using KNN. Indian Journal of Science and Technology. 2016 Jul; 9(28):1–5.