ISSN: 2229-6956 (ONLINE)
DOI: 10.21917/ijsc.2016.0160

R-GA: AN EFFICIENT METHOD FOR PREDICTIVE MODELING OF MEDICAL DATA USING A COMBINED APPROACH OF RANDOM FORESTS AND GENETIC ALGORITHM

S.S. Shah[1] and M.A. Pradhan[2]
Department of Computer Engineering, All India Shri Shivaji Memorial Society's College of Engineering, India
E-mail: [1]ss159862@gmail.com, [2]madhavipradhan@rediffmail.com

Abstract

Medical data mainly includes data of patients and their associated symptoms. Detecting a disease is becoming costly in terms of money and effort. Medical care will be much better if the predictions can be made with minimal efforts. Predictive modeling will help in detecting a disease early. Medical prediction methods which are computer based will help to improve diagnosis. These methods are the important components of decision support systems. This paper suggests the use of predictive modeling as a classifier. The records in dataset are used to construct classifiers using a combination of random forests and genetic algorithm. The inputs to the predictive model are records from the dataset. Genetic algorithm when used in field of computer science help to form methods that lead to a solution that is acceptable. The results of experimentation show that the random forests when used in combination with genetic algorithm gives better accuracy than random forests algorithm alone.

Keywords:

Classification, Genetic Algorithm, Predictive Modeling, Random Forests

1. INTRODUCTION

Medical care includes assessment of risks, under uncertainty. Analytical techniques are getting approval in addressing decision problems that are of classification type in fields of healthcare and medicine. Diagnosis of a disease needs the interpretation of data collected through medical tests which is recorded in datasets. The available data is evaluated by a doctor who uses case based reasoning of precedent sufferers. Disease diagnosis based on machine learning has become part of the medical environment as a supportive tool. The main focus while designing such tools is the accuracy of the overall system. Supervised learning methods have training phase [1]. In the training phase, there is input from records in form of vector or matrix, with labeled data depending on which a model is generated. Testing involves checking or testing of vector or input which is unlabeled data. There are learning algorithms which construct a set of classifiers. Then these classifiers classify test data by voting, which is collective prediction of each classifier.

This paper is based on enhancing the random forests (RF) [2] which is an ensemble algorithm, by using the genetic algorithm, which when applied in the steps of the random forests algorithm which improves the performance of the predictive model by improving accuracy. We name this approach as R-GA.

2. RELATED WORK

Guidi, Pettenati et al. [3] have made a comparison between machine learning methods of Neural networks, Support vector machine (SVM), fuzzy genetic system, classification and regression tree (CART), RF. The accuracies obtained for neural networks, SVM, fuzzy genetic, CART, RF are 77.8%, 80.3%, 69.9%, 81.8% and 83.3%, respectively. Anonymized data of Heart failure patients is used by this system for the purpose of assistance to patients.

Prinzie Anita and Dirk Van den Poel [4] have used Random MultiNomial Logit (RMNL) which involves the use of the RF. The RMNL have chosen a default of random variables as the square root of the total no of variables in the dataset. In a comparison of balanced and unbalanced RF, the unbalanced RF have shown increased Pearson Correlation Coefficient (PCC) of 24.66 with setting number of chosen variables to 252, out of 441 variables that were available, whereas, for a balanced RF with same settings, PCC was 21.67. This means that, for balancing RF, PCC must be reduced. It is stated that supervised learning is helpful in multi- class classification.

Ahmad Taher Azara et al. [5] have used RF as a classifier in the medical field, to detect lymph diseases. It has made use of classification methods based on machine learning using classifier of RF as well as genetic algorithm. The dimension reduction of lymph diseases dataset is achieved by genetic algorithm and classification is done by RF. This combination of RF and genetic algorithm gives accuracy of 0.922 as opposed to 0.812 of RF classifier without feature selection. This is very well used in the bio-medical field for suggesting the relevant medicines as per the prediction.

A. O¨zcift [6] has used RF classifier with data resampling to detect cardiac arrhythmia. The performance of RF algorithm is decided on the distribution of classes in dataset. Here, in normal distribution accuracy of RF is 76.3% and that in re sampled distribution is 90%. Prediction of individual instances has been described in [7]. A generalized approach for prediction was presented which can be used for any method of classification which gives outputs of class probabilities.

Kusiak Andrew and Verma Anoop [8] have used the RF algorithm in fields of turbines. Here also the RF is shown to give better accuracy. A comparative study of various algorithms of SVM, chi-square automatic interaction detector (CHAID) algorithm, neural network, boosting tree algorithm (BTA), RF algorithm gives accuracies of 95.8%, 96%, 97.6%, 98.8%, 99.4% respectively. Thus the RF algorithm performs better. Tripoliti E et al. [9] have suggested an alteration in the voting mechanism used in the RF algorithm and the construction of models. This paper mentions the method to use the available RF algorithm and also the modified method of RF. Accuracy is increased in the modified method, so there is an overall improvement in the performance of RF. Classical RF gives accuracy of 73.70% on breast tissue dataset. Tree selection

approaches SBS-RF and SFS-RF are explained in [9], [10]. Modifications have been done on feature selection techniques and these approaches give better accuracies.

Miao Liua et al. [11] have made a comparison of the RF algorithm has been made with SVM, neural network, boosting, back propagation and it is found that the RF algorithm is best suited for applications in pattern recognition. The average correct rates (CR) on cross validation sets of the four data sets performed by Back Propagation Neural Network (BPNN), SVM and RF were 86.68%, 66.45% and 99.07%, respectively. Structured outputs are found in problems related to predictive modeling [12]. It shows ways of learning ensemble models consisting of RF and bagging. Performance of ensembles is better than single tree models in terms of ranks, with critical distance set at a 0.05 level of significance. When structured outputs are present, ensemble method is used for predictions.

Pedergnana Mattia et al. [13] have given the methods for feature selection in attribute profiles in optimized manner by using the genetic algorithm. It aids in finding the optimal solutions iteratively. The RF algorithm gives ranking of features. This combination of RF and genetic algorithm is efficient. Training set for RF classifier is given of 100 samples from dataset of University of Pavia. Remaining samples were given for fitness function evaluation. This combination of RF and genetic algorithm is useful for feature selection. The time consumption is lowest with the use of the RF classifier in the parallel case for both the considered datasets of University of Pavia and Indian Pines. The result of applying RF on dataset of University of Pavia, directly classifying entire extended multi attribute profile (EEMAP) with the proposed approach of feature selection are more better than without the proposed approach. Overall accuracy (OA) for EEMAP with RF for principal component analysis (PCA) is 96.11% and 96.40% without the proposed approach.

A. Amirov et al. [14] have used the genetic algorithm along with the neural networks in systems making use of medical data. Training of neural network is done by using the genetic algorithm. The order in which procedure of crossover and mutation is done carried out is important. This trained network which classifies records as healthy and sick, gives value of sensitivity of 92.3% .Thus, genetic algorithm can be used in combination with other algorithms to improve accuracy.

From this literature survey, we have seen that RF algorithm is efficient. Utilization of the principles of RF is likely to further increase the accuracy of a classifier. Resampling a dataset improves accuracy thus leading to better performance. RF performs better in comparison to other methods of classification. Thus, ensemble methods are better than the single decision trees.

So using RF, which is an ensemble method, is a promising algorithm useful for designing a predictive model. Combined approach RF and genetic algorithm gives better results.

3. ALGORITHMS AND STRATEGIES USED FOR EXPERIMENTATION

In this paper, we have used a combination of random forests [2] and the genetic algorithm [15], [16].

3.1 THE RANDOM FORESTS ALGORITHM

The random forests algorithm, in machine learning, can also be thought of as an ensemble method for classification [2]. A dataset with attributes is the input to this algorithm. Random subsets of the given dataset are formed. Then, on each of the random subsets that are created, a decision tree will be formed. The resultant class of any test record is decided by the algorithm, which in this case, uses the majority vote technique.

Suppose 'x' is the input in the form of a matrix. On the matrix, there are trees formed randomly. Say there are 'b' number of trees namely $tree_1$, $tree_2$, …., $tree_b$, each of which gives decision k_1, k_2, …, k_b respectively. The majority voting method is applied. Final vote is k class, which will be decided as the class of the test record under consideration.

The random forests algorithm works on the method of randomly selecting the subsets. This means that there is no bias when the random forests algorithm is used.

3.2 GENETIC ALGORITHM

The Genetic algorithm [16], [17] is based on a methodology inspired by the biological evolution and evolutionary algorithms to find solutions. Genetic algorithm falls in the category a machine learning techniques. It optimizes a population. Operators of crossover and mutation in genetic algorithm help to generate better solutions.

4. THE STEPS FOLLOWED IN THE PROPOSED SYSTEM

The architecture proposed in this paper is described in the following figure. From the Fig.1, we see that the final prediction is given by the system, not just by direct implementation of random forests algorithm, but by using genetic algorithm along with random forests.

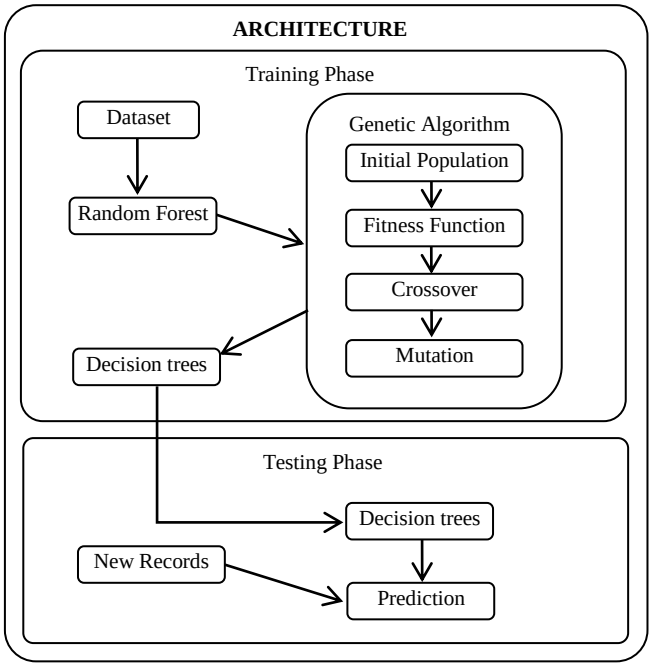

Fig.1. Steps for prediction in the proposed system

There are training and testing phases in our approach. Genetic algorithm is introduced as an additional step in the random forests.

Dataset is the input to the system. We use medical datasets of Breast Cancer, Acute inflammation, Hepatitis, MRI, Thyroid, Parkinson's, and Thoracic Surgery, all available at the UCI machine learning repository [19]. In the training phase, classifier is trained by using the combined approach of random forests and genetic algorithm. In this phase, firstly, the multiple decision trees are formed by using the random forests algorithm. Then, the genetic algorithm is applied on these trees to see if these trees can be further improved to yield better accuracy. If so, these trees are used to give the final class by making use of the majority voting technique.

In the testing phase, unlabeled records are given as input to the system. The class labels of these records are predicted based on the training done in the training phase. Thus the output is the predicted class of the unlabeled records.

4.1 HOW THE PROPOSED SYSTEM MAKES USE OF THE COMBINATION OF RANDOM FORESTS AND GENETIC ALGORITHM

The following are the steps involved in the proposed system.

4.1.1 Initialization of Population:

This step has found and given the list or set of attributes for the further three steps. The following figure, Fig.2, shows a flowchart adopted for this purpose.

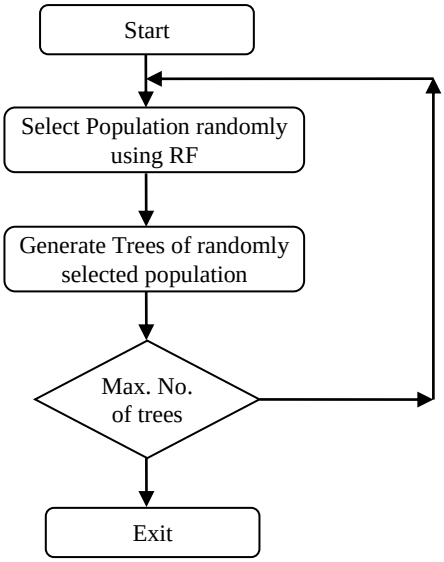

Fig.2. Flowchart for initializing the population

Firstly, the population is selected randomly by using the random forests algorithm. Then, trees for the selected population are generated. This continues until the termination criterion is met. The terminating condition is decided to be a certain variable of maximum number of trees. In our case, this variable is set to 100. So, 100 trees are generated and these trees form the initial population.

4.1.2 Fitness function and its evaluation:

This function evaluates the fitness of the attributes [17]. Measure for fitness function we use is accuracy.

4.1.3 Crossover:

Crossover involves replacing the part of the tree with another sub tree to see if the resultant tree gives better accuracy.

4.1.4 Mutation:

Mutation step involves replacing a randomly selected attribute with another attribute to find if a tree which yields better accuracy can be formed.

5. EXPERIMENTAL RESULTS AND ANALYSIS

The datasets under consideration Breast Cancer, Acute inflammations, Hepatitis, MRI, Thyroid, Parkinson's, and Thoracic Surgery, all available at the UCI machine learning repository [19].

As an example, we have considered Breast Cancer dataset, available at [20]. This dataset has 569 instances and 32 attributes, the last attribute being the class label of the record.

For our experiments, we have used performance measures of accuracy, sensitivity, specificity, precision, recall and f-measurement, all defined in [18]. The Fig.3 shows a comparison of values of performance measures on the dataset in [20]. From Fig.3, it is seen that values for all these measures obtained by using the combined approach of random forests and genetic algorithm are better than those obtained by using the random forests algorithm individually. Accuracy increases from 94.13% by using random forests algorithm to 94.65% by using R-GA. Similarly, sensitivity increases from 96.35% to 96.79%, specificity increases from 88.69% to 89.71%, precision increases from 95.43% to 95.60%, recall increases from 96.35% to 96.79%, f-measure increases from 95.88% to 96.19% by using R-GA.

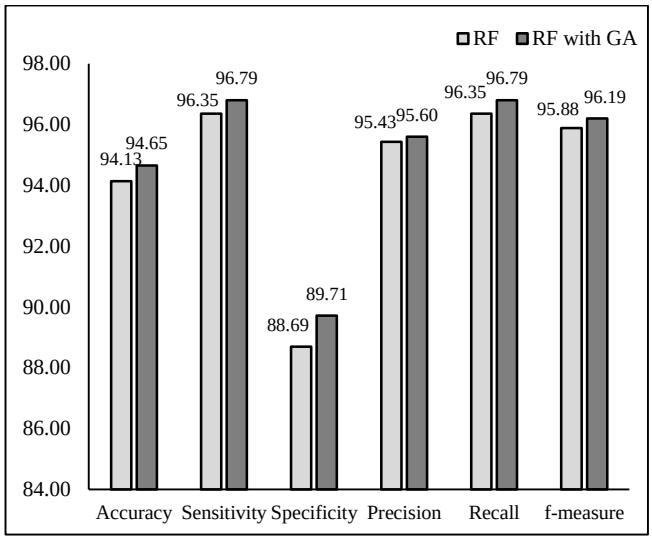

Fig.3. Comparison of performance measures of accuracy, sensitivity, specificity, precision, recall and f-measurement on Breast Cancer dataset

In the testing phase, on the datasets in the Table.1, the accuracy of the system obtained by using the combined approach of random forests and genetic algorithm RGA, is better than the accuracy obtained by using the random forests algorithm alone.

Our experimental results are shown in Table.1. It shows the comparison of the performance measure of accuracy on datasets of Breast Cancer, Acute inflammations, Hepatitis, MRI, Thyroid, Parkinson's, and Thoracic Surgery, all available at the UCI machine learning repository [19]. This table confirms that there is an increase in the accuracy by using our approach of R-GA. Similar results are obtained for sensitivity, specificity, precision, recall and f-measurement.

Table.1. Comparison of accuracy (in %)

Dataset	Accuracy obtained by using Random Forests	Accuracy obtained by using combined approach of random forests and genetic algorithm (R-GA)
Acute inflammations	94.534%	96.697%
Hepatitis	93.506%	98.376%
MRI	98.367%	99.945%
Thyroid	94.534%	96.697%
Parkinson's	92.307%	94.153%
Thoracic surgery	88.244%	89.106%
Breast Cancer	94.13%	94.65%

6. CONCLUSION

For datasets in Table.1, we compared the performances of random forests and combined approach of random forests and genetic algorithm in terms of accuracy. Genetic algorithm is proposed as an additional step in the random forests algorithm. This combined approach (R-GA) improves the performance of the classifier while preserving the randomness feature of the random forests algorithm. As the number of generations in the genetic algorithm increases, better trees are formed. So, the most optimal trees are found in the last iteration of the execution of the system. Prediction of a record whose class label is unknown is done by using these trees.

REFERENCES

[1] Jiawei Han, Micheline Kamber and Jian Pei, "*Data Mining: Concepts and Techniques*", Third Edition, Elsevier, 2012.

[2] Leo Breiman, "Random Forests", *Machine Learning*, Vol. 45, No. 1, pp. 5-32, 2001.

[3] G. Guidi, M.C. Pettenati, P. Melillo and E. Iadanza, "A Machine Learning System To Improve Heart Failure Patients Assistance", *IEEE Journal of Biomedical and Health Informatics*, Vol. 18, No. 6, pp. 1750-1756, 2013.

[4] Anita Prinzie and Dirk Van den Poel, "Random Forests for Multiclass Classification: Random MultiNomial Logit", *Expert Systems with Applications*, Vol. 34, No. 3, pp. 1721-1732, 2008.

[5] Ahmad Taher Azara, Hanaa Ismail Elshazly, Aboul Ella Hassanien and Abeer Mohamed Elkorany, "A Random Forest Classifier for Lymph Diseases", *Computer Methods and Programs in Biomedicine*, Vol. 113, No. 2, pp. 465-473, 2014.

[6] Akin Ozcift, "Random Forests Ensemble Classifier Trained with Data Resampling Strategy to Improve Cardiac Arrhythmia Diagnosis", *Computers in Biology and Medicine*, Vol. 41, No. 5, pp. 265-271, 2011.

[7] Marko Robnik- Sikonja and Igor Kononenko, "Explaining Classifications for Individual Instances", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 20, No. 5, pp. 589-600, 2007.

[8] A. Kusiak and A. Verma, "A Data-Mining Approach to Monitoring Wind Turbines", *IEEE Transactions on Sustainable Energy*, Vol. 3, No. 1, pp. 150-157, 2012.

[9] Evanthia E. Tripoliti, Dimitrios I. Fotiadis and George Manis, "Modifications of the Construction and Voting Mechanisms of the Random Forests Algorithm", *Data & Knowledge Engineering*, Vol. 87, pp. 41-65, 2013.

[10] Simon Bernard, Laurent Heutte and Sebastien Adam, "On the Selection of Decision Trees in Random Forests", *Proceedings of International Joint Conference on Neural Networks*, pp. 302-307, 2009.

[11] Miao Liu, Mingjun Wang, Jun Wang and Duo Li, "Comparison of Random Forest, Support Vector Machine and Back Propagation Neural Network for Electronic Tongue Data Classification: Application to the Recognition of Orange Beverage and Chinese Vinegar", *Sensors and Actuators B: Chemical*, Vol. 177, pp. 970-980, 2013.

[12] Dragi Kocev, Celine Vens, Jan Struyf and Sašo Džeroski, "Tree Ensembles for Predicting Structured Outputs", *Pattern Recognition*, Vol. 46, No. 3, pp. 817-833, 2013.

[13] Pedergnana Mattia, Prashanth Reddy Marpu, Mauro Dalla Mura, Jón Atli Benediktsson and Lorenzo Bruzzone, "A Novel Technique for Optimal Feature Selection in Attribute Profiles Based on Genetic Algorithms", *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 51, No. 6, pp. 3514-3528, 2013.

[14] Azamat Amirov, Olga Gerget, Dmitry Devjatyh and Arstan Gazaliev, "Medical Data Processing System based on Neural Network and Genetic Algorithm", *Procedia - Social and Behavioral Sciences*, Vol. 131, pp. 149-155, 2014.

[15] J.F. Frenzel, "Genetic algorithms", *IEEE Potentials*, Vol. 12, No. 3, pp. 21-24, 1993.

[16] Genetic algorithm, Available at https://en.wikipedia.org/wiki/Genetic_algorithm

[17] Melanie Mitchell, "*An Introduction to Genetic Algorithms*", The MIT Press, 1998.

[18] Jiawei Han, Micheline Kamber and Jian Pei, "*Data Mining: Concepts and Techniques*", Third Edition, Morgan Kaufmann Publishers, 2011.

[19] UCI Machine Learning Repository, Available at http://archive.ics.uci.edu/ml/

[20] Breast Cancer Dataset, Available at https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)