International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

Building a Classifier using Random Forests

Shah Sanika[1], Pradhan Madhavi[2]

[1,2]AISSMS CoE, Savitribai Phule Pune University, Pune, Maharashtra, India

Abstract: ***In the healthcare industry, the data is growing day by day. Many organizations store the records of patients in the form of electronic healthcare records (EHR). These records can be used by the doctor for tasks like diagnosis of diseases, etc. Also to detect a disease early, doctor may use the EHR. To assist the doctor, if an algorithm is provided, which classifies data; it becomes much easier for the doctor to identify patients who are likely to develop a disease. Thus, early detection is one advantage of building such classifiers. Another advantage is that the cost of treating any disease will be reduced. Classifier will help to predict the disease and aid in diagnosis. Here, we propose a classifier which gives predictions once the training has been done.***

Keywords: Classifier, random forests, evolutionary computing

1. Introduction

Medical diagnosis depends on the available data and also on the physician's experience. Making decisions is a process in many areas of health care and it includes diagnosis, of a patient [1]. Machine learning is used in computer introduce aided diagnosis because it has very strong ability to find hidden relationships in medical data that is very complex in nature. The medical data is very huge and it needs very strong classification methods to solve the issue of analysis of data.

Accuracy of classification methods that are used in diagnosing a disease diagnosing needs to be taken in to account. Many a times, medical data has many fields or attributes. Whenever data has many attributes, the most important features have to be selected many times, to gain information regarding the data. Some not very useful attributes have to be eliminated or not to be considered while designing a classifier.

Independent and derived attributes have to be identified. This identification plays a very vital role in the execution of the classifier, since this may affect the amount of time taken for classification purpose. Random Forests are very frequently used because they can attain a high accuracy. They can identify important an informative variables [2]. Moreover, Random Forests are utilized as a means to identify relevant and irrelevant variables in large data. Random forests are based on recursive partitioning technique. Different factors are used to decide the splitting.

2. Related Work

Several approaches for the variable selection have been proposed in literature. When variable selection is to be done, we use the Gini index, information gain, or gain ratio to determine the split variable and choose which variable to retain in the dataset [3].

The sampling methods when used will improve the accuracy of the classifier because, random forests itself uses the method of random sub sampling. Random forests were proposed by Breiman [4].

Over fitting of data becomes a problem that is usually faced by many classifier designers. Noise may be introduced while data collection. When this noise becomes evident in the output of data processing techniques, over fitting is said to occur. So, more than describing the relationships in the data attributes, noise is more prominent in output. This may cause wrong interpretations of data output and may be misleading.

Bagging is an ensemble approach [3]. This uses the method of majority vote technique. The models are developed on the basis of training. Then the testing records are provided to these trained models and then individual outputs are obtained. The final step is majority voting. Each classifier cats its vote and then the class which gets the maximum vote will be the final class of the record.

Bagging is one among the old methods for creating an ensemble of classifiers. In bagging, diversity is obtained by constructing each classifier with a different examples, obtained from the original training dataset by re-sampling. It then combines the decisions of the classifiers by the voting method.

Random forests are an approach very similar to bagging [3],[4]. The random forests also make use of ensemble technique. Randomly subsets of available datasets are selected and the trees are formed on this subset only. Several such trees are formed and then the majority voting is used to decide the class of a new testing record.

The majority vote is a method for random forest ensembles. More power can be give to a classifier when the combination or ensemble is used; these have benefits over single decision trees .Weighted majority vote is used in [5]. In this work, Piero Bonissone et al. propose and compare various combination methods and then obtain the final decision of the proposed classifier system.

In their study [6], assessment of many strategies among wrappers and filters was done so as to choose the most relevant features for cardiac arrhythmia dataset.

Torsten Hothorn, Berthold Lausen [7] have used the bagging strategy to bundle classifiers by using the already available method if bagging. This helps to achieve better accuracy, and also identify the similar types if classifiers. This similarity comparison is a very useful technique in pattern recognition.

Guvenir H.A. et al. [8], have used a voting intervals classification method and have obtained an accuracy of 68% on classification [6][8].

Gurgan F. et al. [4] used combination of support vector machines and logistic regression. Their classification accuracy with Gaussian kernel was 76.1%.

3. Techniques Used

The random forests algorithm is used. This algorithm as explained earlier uses ensemble as an approach. Evolutionary computing is also used, explained in later sections.

3.1 Random Forests

The random forest (Breiman, 2001) is an ensemble approach that is in form of predictor by neighbor method. Random forests use the divide-and-conquer methodology which increases performance. The idea for ensemble methods is that individual learners can become strong learners if the form a group or an ensemble. Machine learning comes into picture in the random forests. Individual trees are formed randomly first and the n the combination n of these trees occurs. Random forest gives accurate predictions as compared to simple classification or regression models in many cases. In these cases, there are a large number of variables and size of the sample is also big. The reason for this is that it gets the variance of many input variables and at the same time. This makes possible large number of observations to participate in the ensemble and prediction.

3.1.1 Working of the random forests

While the training s done, a set for the current tree is done by sampling, and some parts are left out of the sample for further testing. This data is used to gives unbiased estimate of the classification error. It is also used to get estimates of importance of that particular variable. Handling missing value in dataset by the random forests algorithm is computationally more costly in terms of time but gives good performance even when large amounts of data is missing in the dataset. All that it does is replace missing values and in the training phase only. It starts with only estimates, and then gradually increases the replacement of missing values with labels like unknown. Sometimes mislabeled classes may be found, and this leads to errors in the prediction. The following figure shows how a random forest algorithm would produce a combination of votes.

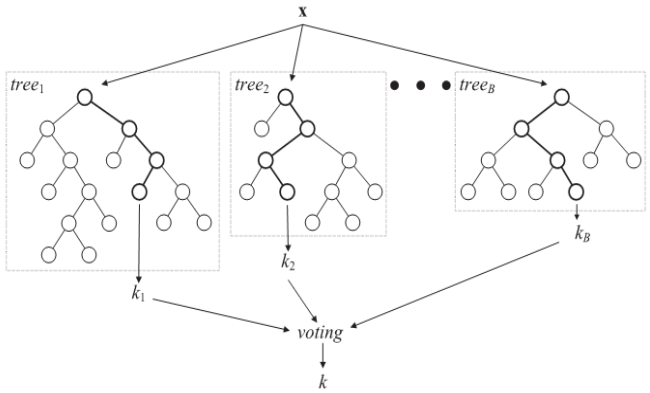

Figure 1: The figure shows the ensemble of trees decision in the random forests algorithm

3.1.2 Random forests and decision trees

Decision trees mainly are single. This means that there is a possibility that there may be biasing towards some specific attributes while considering splits. But In the case of random forests, this biasing is lessened because of the fact that the possibility of the same variable being chosen repeatedly reduces due to random selection.

3.2 Evolutionary Computing

Evolutionary computing is collection of techniques of solving problems based on principles of evolution, such as natural selection, inheritance [10]. These algorithms, called evolutionary algorithms, are based on adopting Darwinian principles. Evolutionary computing uses iterations while the progression like the development in a population would happen. The selection of the fit population takes place. These are the individuals that will go on in the generation.

The common idea for all such kinds of these techniques is stated as when there is a given population of individuals, the environmental pressure causes natural selection (survival of the fittest) and hereby the fitness of the population is growing. Such a process is very evident as an optimization process.

The figure 2 shows the steps involved in the evolutionary computing. This is a cyclic method, where all the steps are repetitive. The individuals go on evolving, in evolutionary computation.

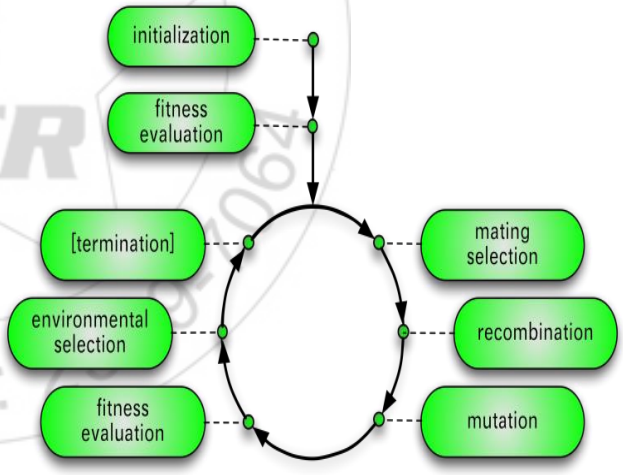

Figure 2: The cycle of evolutionary computation

From the available evolutionary computing algorithms, GA is chosen. The steps in GA are same as those followed in any evolutionary algorithm.

But some representation issues are prevalent in such evolutionary algorithms. These issues of representation are stated in literature [11].

Another issue is preservation of the diversity of the population. This issue is addressed by using certain selection methods. Evolutionary algorithms are able of optimize a collection of entities inherently. However, many components of such evolutionary algorithms are stochastic. During selection, the more fit individuals have a higher chance of

selection as compared to the lower fit ones, but the weak individuals have a possibility to either become a parent or to survive.

Evolutionary computing algorithms are robust and have the power to adapt to handle nonlinear, and complex engineering problems. They do not require any explicit knowledge of the problem beforehand or the structure or differentiability. These algorithms have the capability to providing multiple near-optimal solutions to even large, unstructured problems come into picture.

Evolutionary algorithm can be self adaptive. As a proof for the power of self-adaptation, evidence is provided in the context of changing fitness objectives. The fitness function is changing and the evolutionary process usually aims at a target. Once the objective function has changed, the current population has to be evaluated again, and it is possible for individuals to have a low fitness, because they have been used to the old fitness function.

4. Performance

The following table shows accuracy obtained by using integration of random forests and evolutionary computing on datasets from medical field.

Table 1: Performance of designed classifier on breast cancer, leukemia and ecoli datasets in terms of accuracy

Dataset	*RF*	*Integrated algorithm*
Breast cancer dataset	85.07%	87.26%
Leukemia dataset	89.67%	89.92%
E coli	91.02%	94.91%

From the table 1, we can see that the integrated algorithm gives better performance on the medical datasets of breast cancer, leukemia and Ecoli. Efficient classification is important, and our classifier does this by improving the performance.

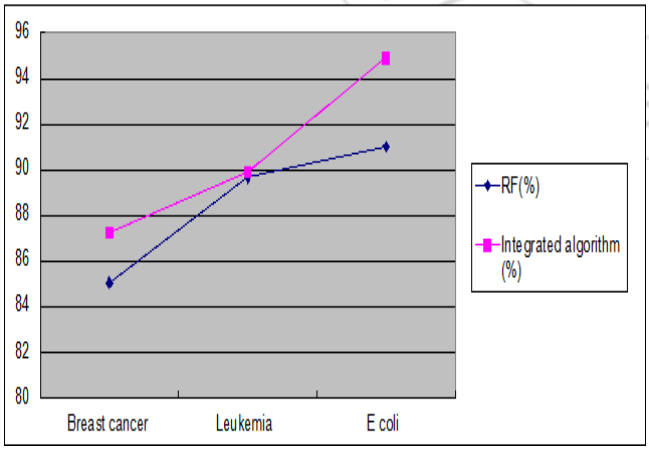

Figure 3: Graph

The figure 3 shows the comparison of accuracy on datasets breast cancer, leukemia, E coli and a graphical representation of the same.

References

[1] Ahmad Taher Azar , Hanaa Ismail Elshazly, Aboul Ella Hassanienb, Abeer Mohamed Elkorany, "A random forest classifier for lymph diseases" computer methods and programs in biomedicine 113 (2014) 465–473
[2] A. Hapfelmeier, K Ulm, "A new variable selection approach using Random Forests", Computational Statistics and Data Analysis 60 (2013) 50–69
[3] Jiawei Han, Micheline Kamber, Jian Pei, "Data Mining: Concepts and Techniques", Elsevier, 09-Jun-2011
[4] Breiman Leo, "Random forests",Machine Learning October 2001, Volume 45, Issue 1, pp 5-32
[5] Piero Bonissone , José M. Cadenas, M. Carmen Garrido, R. Andrés Díaz-Valladares, "A fuzzy random forest", International Journal of Approximate Reasoning 51 (2010) 729–747
[6] Akin O¨ zc- ift, "Random forests ensemble classifier trained with data resampling strategy to improve cardiac arrhythmia diagnosis", Computers in Biology and Medicine41(2011)265–271
[7] Torsten Hothorn, Berthold Lausen, "Bundling classifiers by bagging trees ", Computational Statistics & Data Analysis 49 (2005) 1068 – 1078
[8] H.A. Guvenir, "A supervised machine learning algorithm for arrhythmia analysis, Computers in Cardiology 24 (1997) 433–436.
[9] A.Uyar, F.Gurgen, Arrhythmia classification using serial fusion of support vector machines and logistic regression, IDAACS ,in: Proceedings of the 4th IEEE Workshop on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, 2007, pp.560–565.
[10] Evolutionarycomputation,Available: https://en.wikipedia.org/wiki/Evolutionary_computation
[11] A.E. Eiben, M. Schoenauer, "Evolutionary computing", Information Processing Letters 82 (2002) 1–6