ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 7, July 2016

New Approach for Classification and Learning Using Fuzzy Random Forest

Ms. Swati W. Bagade, Dr. Prof. Madhavi Pradhan

Department of Computer Engineering, AISSMS COE, Pune, MH, India

ABSTRACT: In machine learning system different types of approaches, machine learning strategies have applications are related sentiment analysis, classification approaches, data mining etc. Irregular Forest has huge capability of turning into a prevalent method for future classifiers in light of the fact that its execution has been observed to be practically identical with troupe strategies sacking and boosting. Thus, an inside and out investigation of existing business related to Random Forest will quicken research in the field of Machine Learning. Given work introduces a deliberate study of work done in Random Forest range. In the proposed work system classify the results base on fuzzy random forest, first its takes huge dataset from internet as well as user define directories and preprocess. Once preprocess phase done data gets in structured foam, this work has been done into first section. In second user provide the inputs as test data set from application GUI, then RF executes. In the behalf of RF first attribute selection done then calculate the weight and select the best feature set, once RF execution done system gets best set of results node. Finally fuzzy classifier will classify all the data with different probabilities and analysis phase provide the truthfulness of system how it's better than existing approaches.

KEYWORDS: Random Forests, Classification, Decision trees, Fuzzy Classifier, Learning systems.

I. INTRODUCTION

Classification has always been a challenging problem. The vast amount of information that today is available to companies and individuals further compound this problem. We have witnessed a number of methods and algorithms addressing the classification issue. In recent years, we have also seen an increase of multi-classifiers based approaches, which have shown to deliver better results than individual classifiers. Here we plan to develop we will not address the issue of how to get the best multi-classifier system. Rather, our focus will be on how to start from a multi-classifier system with performance in comparison with the best classifiers. And extend it to handle and manipulate imperfect information (missing values, linguistic labels etc.) To develop the multi-classifier, we follow the methodology of random forest. To incorporate the processing of faulty data, we construct the random forest using fuzzy trees as base classifiers. Therefore, we try to use the firmness of a tree ensemble, the power of the randomness to increase the diversity of the trees in the forest, and the ability of Fuzzy Logic and the Fuzzy Sets for Data Managing. As the data grows over, several data mining techniques become significantly harder and the computational effort for data mining applications are increases dramatically. Moreover, the increasing dimensionality of data (in the sense of many features) for various application areas (such as email classification and in silicon screening for drug discovery) often leads to severe problems for learning algorithms, since the high dimensional nature of data can have a negative influence on the classification accuracy and also increases the computational cost for both, supervised and unsupervised learning methods. High dimensionality of data in the sense of many instances also increases the computational cost for the learning algorithm but has usually a positive influence on the classification accuracy. This causes the need for effective and efficient feature selection and dimensionality reduction methods. Generally, these methods should meet the following demands: They should be computationally efficient and the resulting feature sets should allow for a compact representation of the original data. Moreover, in many application areas it is very important not to loose the interpretability of the original features. Finally, the feature reduction process should have a positive (or at least no negative) effect on the classification accuracy of the learning algorithm.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 7, July 2016

II. LITERATURE SURVEY

Random Forest (RF) is an ensemble supervised machine learning algorithm. Machine learning techniques are applied in the domain of Data Mining [8]. Random Forest [Breiman 2001] uses decision tree as base classifier. Random Forest generates multiple decision trees; the randomization is present in two ways: first random sampling of data for bootstrap samples as it is done in bagging and second random selection of input attributes for generating individual base decision trees. Strength of individual decision tree and correlation among base trees are key issues which decide generalization error of Random Forest classifier [3]. Based on accuracy measure, Random Forest classifier is at par with existing ensemble techniques like bagging [1] and boosting [6]. As per Brieman, Random Forest runs efficiently on large databases, it can handle thousands of input variables without variable deletion, it gives estimates of important variables, it generates an internal unbiased estimate of generalization error as forest growing progresses, it has effective method for estimating missing data and maintains accuracy when a large proportion of data are missing, and it has methods for balancing class error in class population unbalanced data sets [3]. The inherent parallel nature of Random Forest has led to its parallel implementations using multithreading, multi-core, and parallel architectures. Random Forest is used in many recent classification and prediction applications [13] due to above mentioned features. It has been proved theoretically and empirically that the ensemble always gives better accuracy than an individual classifier [2]. The fundamental of ensemble design is creating diversity among the base classifiers [5]. Random Forest is based on the principle of bagging. Instable base learners are good choice for bagging ensembles [1]. Decision Tree is instable in nature and hence works well as base classifier with Random Forest. As in bagging, bootstrap samples are generated for induction of each decision tree. Another source of randomization is introduced through attribute selection. Research work in the area of Random Forest aims at either improving accuracy or improving performance i.e. reducing time required for learning and classification, or both. Some work aims at experimentation with Random Forest using online continuous stream data which is very much essential today due to data streams getting generated as a result of various applications. Random Forest being ensemble technique, experiments are done with its base classifier, e.g. Fuzzy Decision Tree as base classifier of Random Forest. We have done in depth and systematic survey of current ongoing research on Random Forest and also developed "Taxonomy of Random Forest Classifier" [14]. The research work on improving accuracy is by Robnik Sikonja [7]. They have generated base trees of Random Forest using different split measures and also applied weighted voting. In [9] it is demonstrated that accuracy of random forest is improved in some domains by replacing majority voting with Dynamic Integration, which is based on local prediction performances of base decision trees. In [10], Random Forest themselves are used as base classifiers for making ensemble called Meta Random Forest, and the accuracy of this model is tested and compared with the existing Random Forest algorithm. In [4], [11], [12], it is confirmed empirically that, the generation of Random Forest should be done in such a way that the trees will be diverse as well as they retain their strength. We have performed some experimentation to achieve this and have come up with a new approach which we call as Disjoint Partitioning approach for Random Forest. With Disjoint Partitioning, we are making disjoint sets or partitions of the original dataset and using them to induce individual tree. This ensures diversity among individual trees. Another measure taken to increase diversity is use of less correlated attributes. For this, at each node, we are selecting (2/3)*m features out of total m features and then selecting $\sqrt{m}$ features randomly to decide best split at each node (which is the process for induction of Random Forest that is to be explained in the next section). Here as each individual tree is trained with less number of samples, the learning of each tree and hence learning of forest is more efficient.

III. PROBLEM STATEMENT

In the proposed research work to design and implement the a system which is provide the classification disease scenario on raw patient data using Fuzzy Random forest. First Random forest classify the instance as normal and disease and fuzzy will classify the sub category of instance.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

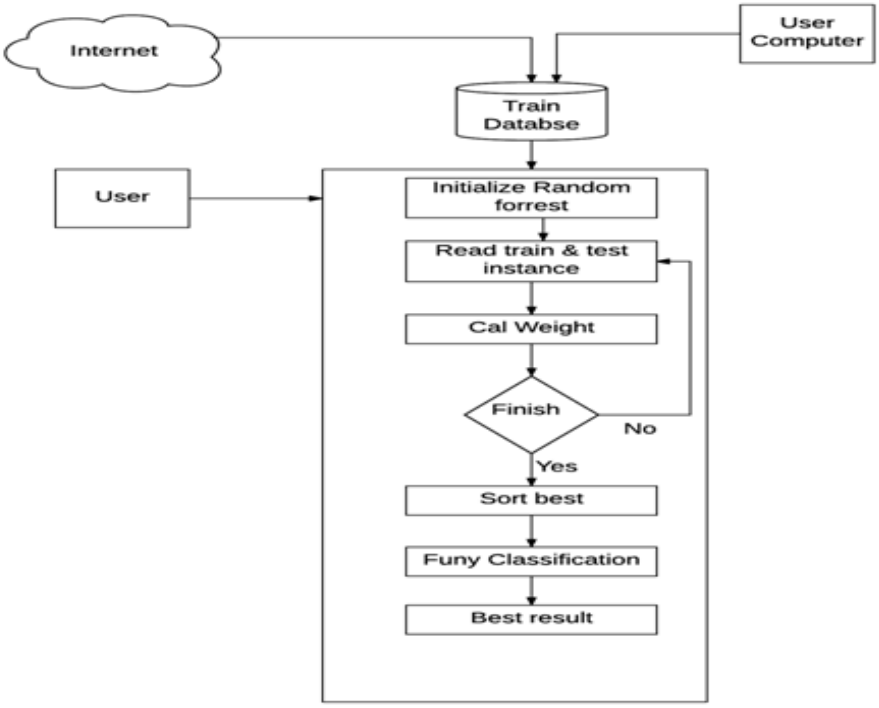
International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 7, July 2016

IV. PROPOSED SYSTEM

Classification has always been a tedious problem, As it has been commented earlier, the random forests are good categorization methods and fuzzy sets (with their estimated reasoning capability) have been introduced in the decision trees in a suitable way. Continuing these two good characteristics, in this work the multi-classifiers anticipated are a forest of fuzzy decision trees generated randomly (Fuzzy Random Forest), In this section we specify the adjustments, changes and considerations needed to construct these multi-classifiers.



a) Forest structure: at first a forest is constructed from ten trees. For that classical random forest is joint with performance measurement criteria's like Relief and numerous estimators. The forest construction is revealed below. At first, the forest started with ten trees and select a greatest fit is selected from the residual dataset and the construction is made. The similar process is continual up to the 10 trees.

b) Polynomial fitting process for feature selection: Forest construction is an iterative process. Here we have to find similarity of two Nodes $\vec{a} = (a_1, a_2, a_3, \ldots)$ and $\vec{b} = (b_1, b_2, b_3, \ldots)$, where a_n and b_n are the components of the nodes (features of the train node, or values for each features of the node) and the n is the dimension of the node:

$$\vec{a} \cdot \vec{b} = \sum_{i=1}^{n} a_i b_i = a_1 b_1 + a_2 b_2 + \cdots + a_n b_n$$

c) Fuzzy classification: the fuzzy classification executes on probability basis. The RF best nodes get input to fuzzy classifier, it will first analyze the each attribute values base on probability basis then finally classify with label.

V. RESULTS AND DISCUSSION

For the proposed system performance evaluation, we calculate matrices for accuracy. We implement the system on java 3-tier MVC architecture framework with INTEL 3.0 GHz i7 processor and 8 GB RAM. After implementation the proposed system perform the accuracy as well as classification results and just compare with all existing systems. The

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 7, July 2016

system will finally calculate the F-measure with correctness. The below table shows the experimental results with existing systems.

Fig 1: System performance proposed vs Existing

Here Fig 1 shows the overall accuracy of proposed system with some existing systems. For the proposed work system takes health care data, user can submit the query as disease name as well name of symptoms, after that system will classify the final results. Sometime system will predict the same result as wrong, that will be a false ratio. We done around 100 test queries (e.g. disease name like cancer, heart attack etc or symptoms name like flue, high BP, cough etc) with proposed system and consider the average estimated results and denoted into the final column.

VI. CONCLUSION

In this research work, we presented approaches for improving performance of Random Forest classifier using fuzzy logic in terms of accuracy, and time for learning and classification. In case of accuracy improvement, research is done using different attribute evaluation measures and combine functions. A fuzzy decision tree model along with weighted voting is suggested which improves the accuracy using sub classification . Improvement in learning time mainly concerns on reducing number of base decision trees in Random Forest so that learning and in turn, classification is faster. The approaches suggested in this direction are different partitions of training datasets to learn the base decision trees, and ranking of training bootstrap samples on the basis of diversity. Both these approaches are leading to efficient learning of Random Forest classifier. An attempt is made to find optimal subset of Random Forest classifier using Fuzzy classifier. Random Forest has inherent parallelism and can be easily parallelized for scalability and efficiency. A new parallel approach is proposed in which both, individual tree as well as entire forest is generated in parallel. The new approaches presented here are leading to effective learning and classification using Fuzzy Random Forest algorithm.

REFERENCES

[1] Leo Breiman, Bagging Predictors, Technical report No 421, September 1994.
[2] David Opitz, Richard Maclin, “Popular Ensemble Methods: An Empirical Study”, Journal of Artificial Intelligence 11, 169-198, 1999.
[3] Leo Brieman, “Random Forests”, Machine Learning, 45, 5-32, (2001).

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 7, July 2016

[4] Latinne P, Debeir O, Decastecker C, Limiting the number of trees in Random Forest, MCS, UK (2001)
[5] L Kuncheva, C Whitaker, Measures of Diversity in Classifier Ensembles and their relationship with the Ensemble Accuracy, Machine Learning, 51, 181-207, (2003).
[6] Robert E Schapire, The Boosting Approach to Machine Learning an Overview, Nonlinear Estimation and Classification, Springer, 2003.
[7] Marko Robnik, Sikonja, "Improving Random Forests", J F Boulicaut et al (eds): Machine Learning, ECML 2004 Proceedings, Springer, Berlin, 2004.
[8] Jiawei Han and Micheline Kamber, "Data Mining: Concepts and Techniques" (2nd Edition), Morgan Kaufmann Publisher, (2006).
[9] Tsymbal A, Pechenizkiy M, Cunningham P, Dynamic Integration with Random Forest, ECML, LNAI, 801-808, Springer-Verlag (2006).
[10] Boinee P, Angelis A, Foresti G, Meta Random Forest, International Journal of Computational Intelligence 2, (2006).
[11] Bernard S, Heutte L, Adam S, On the Selection of Decision Trees in Random Forest, Proceedings of International Joint Cobference on Neural Networks, Atlanta, Georgia, USA, June 14-19,302-307, (2009).
[12] Bernard S, Heutte L, Adam S, Towards a Better Understanding of Random Forests Through the Study of Strength and Correlation, ICIC Proceedings of the Intelligent Computing 5th International Conference on Emerging Intelligent Computing Technology and Applications, (2009).