International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-6, April 2020

An Empirical Methodology to Examine the Effect of Meta Classifiers in J48 and Random Tree in Weather Data

Divya R, Anju Rajan K, Deepa G

Abstract*: Weather data interpretation has become vitally important in most domains of human activity and this is because in recent years, major changes have begun to impact climate globally – peninsular India is among the regions seriously affected with this and prediction has become a particularly urgent concern. In this work to bring out a better methodology to examine the weather data using Meta classifiers, a method is postulated by formulating it with Tree classifiers – J48 and Random Tree. Implementation phase has shown distinct results for both the classifiers. Regardless, we could conclude from this work that the effect of Meta Classifiers in J48 and Random Tree algorithm shows that efficiency can be improved by applying the same.

Keywords: AdaBoost, Bagging, Data Mining J48, Random Tree.

I. INTRODUCTION

Weather prediction has always been a challenging task and in recent times, with fluctuations in Global Climate, it has become particularly difficult. Lots of new research activity has been focused on this problem. Data Mining is a broad class of techniques used to convert raw data in to useful information and many approaches to weather prediction approaches are focused on finding accurate data mining based weather prediction techniques. Data form the basis of all research and data mining helps to statistically analyze data and generate rules which help for the accurate prediction.

Dataset for this research is collected from www.kaggle.com. It is the weather data from Hungary, a country in Central Europe. Weka tool is used to further processing. Classification, Regression, Clustering etc. are available in data mining for predicting better results.

Revised Manuscript Received on April 30, 2020.
***** Correspondence Author
Divya R*, P G Student, Department of Computer Science and IT Amrita School of Arts and Science, Amrita Vishwa Vidhyapeetham, Kochi, India. Email: divyrkrishna88@gmail.com
Anju Rajan K, P G Student, Department of Computer Science and IT Amrita School of Arts and Science, Amrita Vishwa Vidhyapeetham,Kochi, India. Email: anjurajank1998@gmail.com

Lot of pre-processing steps such as Binarization, Normalization, Finding Missing values, etc. are also available.

Environmental factors trigger climatic variation. It plays an important role in determining accurate weather prediction method. The parameters we focus upon are Temperature, Humidity, Wind Speed, Cloud Cover, Pressure, etc. Several types of classifiers are available in Weka. Here, we carry out a comparison of Tree classifier algorithms - J48 and Random Tree and the Meta classifiers - boosting and bagging.

The factors used for comparison are Absolute Error and Root Mean Squared Error.

II. LITERATURE REVIEW

In [1], the authors introduced a methodology to use FP Growth algorithm to predict the weather of Bandung Regency in Indonesia. The tool they used was Weka. After the preprocessing step of Discretization and Partitioning, FP Growth algorithm is applied to the training data set. Based on the rules generated, J48 classification is applied according to which performance is measured. Precision, Recall and Accuracy are the performance measures used. [2] Deals with the Rainfall forecasting in Nagpur Station. They considered the parameters temperature, humidity and wind speed. The data set from weather department of Nagpur Station were collected and processed through the following stages namely data collection, data cleaning, data selection and data transportation. FP growth algorithm is applied to dataset to find the frequent pattern as well as to generate rules. Then they calculated Mean Absolute Error (MAE), Mean Square Error (MSE), and Standard Deviation and evaluated the next year rainfall rate in Nagpur Station. In conclusion, they state that FP Growth Algorithm shows correct monthly rainfall prediction than Neural Network. In [3] the authors provide a comparative study based on J48 algorithm and Random Forest algorithm. They study random weather data and the required features are identified. J48 and Random Forest algorithm is then applied to the data. Random forest algorithm had shown more accuracy and lesser mean absolute error as compared to the other. The future work related to this paper is to use appropriate preprocessing techniques. In [4], the comparative analysis of random forest and random tree algorithms are done. Data collected is applied to the algorithms. Attribute selection filters are applied and results are obtained.

When compared to the previous result without applying filters, random tree gives superior classification accuracy. [5] Is mainly focused on finding an improvement in j48 algorithm using the ensemble methods like bagging, boosting and stacking. SONAR dataset is used here. Stacking is applied with 2 classifiers J48 and IBK.

The paper concluded with the inference that all the ensemble methods improve the efficiency and stacking is more efficient than others.

In [7] various decision tree algorithms like J48, Decision Stump, Random Tree, Random Forest and LMT are compared using WEKA tool. They came into the conclusion that minimum time is taken by Decision Stump for the classification but the accuracy is less. In the case of J48, quite good accuracy is there but the time taken is slightly more. Maximum accuracy is given by LMT even though the time taken for building the classification model is higher than others. All other models lie between the best and worse one. In [7], Method to improve the j48 algorithm is mentioned. Pima Indian Diabetes Data set is used in this paper. WEKA tool is used as an API is the newly generated algorithm. Here arff file from WEKA is loaded into MATLAB and refined the data. Later J48 is applied and measured the accuracy. Proposed algorithm provides high accuracy and low error rate. [8] Consist of the comparative study between Random Forest and Decision Tree using R tool which is an open source R-GUI. They used weather data which consist of 256 samples for the analysis. Actually, Random Forest is made up of many decision tree.[8][9] Here they found that the error rate is less in Random Forest but the time taken for the execution is more. In [9], improvement in Radom Forest for the past 15 years is shown. In [10], Random Forest, Random Tree and LAD Tree Algorithm are taken for analysis. Tools used for the classification is WEKA. The dataset consist of News which are collected from various national and regional newspapers available in internet. It is manually classified into 7 categories and then the algorithms are applied. As a result, they reached to the conclusion that Random Tree algorithm performs well. In [11], Author aims to reduce the execution time of Random Forest by a new approach called Disjoint Partition. This helps to make diversity in the base classifier. In order to improve the diversity, different subset of the attributes were used in each stage. Then it has been found that the accuracy enhanced.

III. ALGORITHMS SELECTED FOR COMPARISON

A. J48

Weka tool incorporates a large number of classification algorithms. Classification is a process of generating a model of classes from a set of records that includes class labels. One such efficient classification algorithm is j48. Decision Tree algorithm shows how the attribute vector behaves for the number of instances. It also generates the class for the newly generated instances. J48 algorithm is the extension of the ID3 (Iterative Dichotomiser 3) algorithm, which is used to create a decision tree from the dataset. J48 algorithm is easy to comprehend .Classification and learning steps are simple and fast. In Weka, J48 is an open source implementation of C4.5 algorithm. Accounting for missing values, Decision Tree Pruning, Generating Rules etc. are the other usages of J48 algorithm.

B. Random Tree

Random Tree is a classification algorithm introduced by Leo Breiman and Adele Cutler. It creates many individual learners. It shows an idea related to 'bagging' (short for 'bootstrap aggregating') which creates a random set of data to build up decision trees. Each node in the standard tree split using the best split among all variables. Random Tree consists of a collection of tree predictors. The algorithm works as follows: It takes input vector classifier and classifies with all the trees in the forest and the class label is provided as the output based on the majority of vote it gets.

C. Adaboost

AdaBoost is the short form of Adaptive Boosting. It is an ensemble method that comes under Meta classifier in Weka Tool which is formulated by Yoav Freund and Robert Schapire. This award-winning (2003) method begins with base classifier that is created from the training data. Then another classifier is created in behind to focus on the training data instances that the first classifier identified as wrong. This process to add the classifiers continues unless and until a limit is reached in the number of models or accuracy. In order to improve the performance, it can be used in conjunction with many other types of learning algorithms.

D. Bagging

Bagging (Bootstrap Aggregating) comes under Meta classifier in Weka tool. It is a simple and powerful ensemble method which generates separate samples of the dataset that we provide and creates a classifier for each sample. Then the results from each classifier is combined either through the majority of voting or by averaging the results and produces the final result. In order to produce more accurate results than any individual model, the ensemble method technique combines the predictions from multiple ensemble method together. This technique can be used to reduce the variance of the algorithms like decision tree which have high variance.

IV. EXPERIMENTAL METHODOLOGY

In this paper, we compare the Tree classifiers and Meta classifiers. For this purpose, data collected from Kaggle.com is processed using Weka tool. Various pre-processing steps are available in Weka. The factors considered are Temperature, Visibility, Humidity, Wind speed and Summary. "Attribute Selection Filters" of Weka is utilized for selecting these features that are highly correlated to the class label.

A. Resampling

Resampling is an efficient method of taking repeated samples from the original dataset. In Weka Resampling option is visible under instances in unsupervised which come below the filters. This tool provides a pragmatic way to process the sample size percentage as follows- In SampleSizePercentage, we are able to specify the size.

In this research paper, we split 80% of the original dataset for the training data and 20% for the test data. The method select randomized samples with replacement or without replacement from the original dataset in such a manner that each number of the sample that select has a number of cases that are similar to the original dataset. The dataset should have a nominal class attribute.

B. Binarization

One of the preprocessing steps that we use is Binarisation. The dataset we got from the site www.kaggle.com was CSV format. Using Weka the dataset format is converted into ARFF. Some of the fields in this data set were nominal in type. In order to apply classification algorithms we need to binaries the data. For that first we need to apply NominalToBinary option and then the option called NumericToBinary. Before applying this method we need to set the ignore class as true.

The pre-processed data is applied to J48 algorithm and Random Tree algorithm.

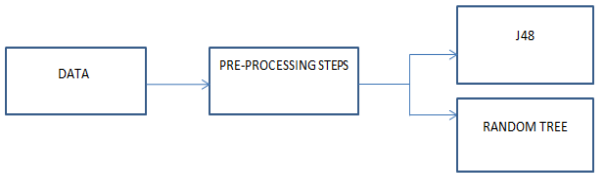

Fig.1.Preprocessed data is applied on J48 and Random Tree

The second phase of implementation starts by applying AdaBoost to J48 algorithm and Random Tree algorithm. Produced results are saved and Bagging is applied to J48 and Random Tree algorithm.

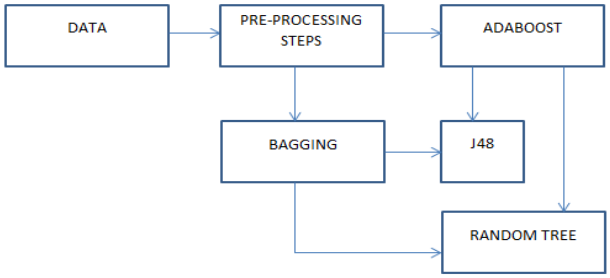

Fig.2. Preprocessed data is applied on Tree classifiers along with Meta classifiers.

V. EXPERIMENTAL RESULTS

We combined the Meta classifiers like Bagging and AdaBoost with tree classifiers, J48 and Random Tree and the results are obtained. The factors which we considered for validating the results are Mean Absolute Error and Root Mean Squared Error. Here we find the measure of Mean Absolute Error and Root Mean Squared Error for Temperature, Pressure, Humidity and Wind speed separately and then find the average of these factors to find the final result. Results when 1) Applying j48 alone 2) By applying Random Tree alone 3) Applying J48 along with AdaBoost 4) Applying J48 along with Bagging 5) By applying Random Tree along with AdaBoost and 6) By applying Random Tree along with Bagging, are shown below.

Table- I: Applying J48 alone

	Temperature	Humidity	Wind Speed	Visibility	Average
MAE	0.0084	0.0007	0.0241	0.0085	0.010425
RMSE	0.0677	0.018	0.1062	0.0643	0.06405

Table- II::Applying Random Tree alone

	Temperature	Humidity	Wind Speed	Visibility	Average
MAE	0.0083	0.0007	0.0239	0.0084	0.010325
RMSE	0.0675	0.0178	0.1056	0.0635	0.0636

Table- III: Applying J48 along with AdaBoost

	Temperature	Humidity	Wind Speed	Visibility	Average
MAE	0.0084	0.0006	0.0237	0.0082	0.010275
RMSE	0.0675	0.0179	0.1058	0.0636	0.0637

Table- IV: Applying J48 along with Bagging

	Temperature	Humidity	Wind Speed	Visibility	Average
MAE	0.0083	0.0006	0.0238	0.0083	0.01025
RMSE	0.0675	0.0178	0.1056	0.0635	0.0636

Table- V:Applying Random Tree along with AdaBoost

	Temperature	Humidity	Wind Speed	Visibility	Average
MAE	0.0084	0.0006	0.0238	0.0082	0.01025
RMSE	0.0675	0.0179	0.1058	0.0636	0.0637

Table- VI: Applying Random Tree along with Boosting

	Temperature	Humidity	Wind Speed	Visibility	Average
MAE	0.0083	0.0006	0.0238	0.0083	0.01025
RMSE	0.0675	0.0178	0.1056	0.0635	0.0636

Table- VII: Overview of the result after applying J48 and Random Tree alone

	J48	Random Tree
MAE	0.010425	0.010325
RMSE	0.06405	0.0636

Table- VIII: Overview of the result after applying Meta classifiers with Tree classifiers

	J48+AdaBoost	J48+Bagging	Random Tree+ AdaBoost	Random Tree+ Bagging
MAE	0.010275	0.01025	0.01025	0.01025
RMSE	0.0637	0.0636	0.0637	0.0636

VI. CONCLUSION

In this research, what we have tried to do is to formulate a method to find better way to reduce the error rate of J48 and Random Tree algorithm when applied to weather data. We assume that it might help for better weather prediction. For that we did an empirical methodology to examine the effect of Meta classifiers like Bagging and AdaBoost in Tree classifiers like J48 and Random Tree. Both the J48 and Random Tree algorithms are used for prediction since decades. When applied alone, Random Tree provides better results. However, Comparison of the results show that Error rate in both the Tree classifiers (J48 and Random Tree) can be reduced when applied with Meta Classifiers (AdaBoost and Bagging). The future work associated with this paper is the usage Meta Classifiers, ensemble with J48 algorithm and Random Tree algorithm in predicting weather.

REFERENCES

1. Farida Nur Khasanah, Fhira Nhita "Weather Forecasting in Bandung Regency based on FP-Growth Algorithm", International journal on ICT, December ,Volume 4,Issue 2.
2. Amruta A. Taksande , P. S. Mohod, "Applications of Data Mining in Weather Forecasting Using Frequent Pattern Growth Algorithm", International journal of Science and Research(IJSR),2013
3. S. Karthick, D. Malathi, C.Arun," Weather Prediction Analysis Using Random Forest Algorithm", International Journal of Pure and Applied Mathematics", 2018, volume 118.
4. Ajay Kumar Mishra, Bikram Kesari Ratha," Study of Random Tree and Random Forest Data Mining Algorithms for Microarray Data Analysis", International Journal on Advanced Electrical and Computer Engineering (IJAECE), 2016, Volume 3, Issue 4.
5. Aakash Tiwari, Aditya Prakash, "Improving classification of J48 algorithm using bagging, boosting and blending ensemble methods on SONAR dataset using WEKA",

International Journal of Engineering and Technical Research (IJETR), September 2014,Volume 2, Issue 9.
6. Purva Sewaiwar, Kamal Kant Verma," Comparative Study of Various Decision Tree Classification Algorithm Using WEKA ", . International Journal of Emerging Research in Management &Technology, Volume 4, Issue 10.
7. Gaganjot Kaur, Amritsar, Amit Chhabra, "Improved J48 Classification Algorithm for the Prediction of Diabetes", International Journal of Computer Applications, Volume 98, July2014
8. Prajwala T R, "A Comparative Study on Decision Tree and Random Forest using R Tool", International Journal of Advanced Research in Computer and Communication Engineering, Volume 4, Issue 1, January 2015.
9. Eesha Geol, Er.Abhilasha, " Random Forest: A Review", . International Journal of Advanced Research in Computer Science and Software Engineering, January 2017.
10. Sushikuma, Rameshpant Kalmegh , "Comparative Analyisis of WEKA Data Mining algorithm Random Forest, Random Tree and LAD Tree for classification of indigenous News data", . International Journal of Emerging Technology and Advanced Engineering, January 2015, Volume 5, Issue 1.
11. Vrushali Y Kulkarni, Pradeep K Sinha," Effective Learning and Classification using Random Frorest Algorithm", International Journal of Engineering and Innovative Technology(IJEIT), Volume 4, Issue 9, May 2014.

AUTHOR'S PROFILE

Divya R, PG Student, Department of computer science and IT, Amrita School of Arts and Sciences, Kochi, Amrita Vishwa Vidhyapeetham, India.

Anju Rajan K, PG Student, Department of Computer Science and IT, Amrita School of Arts and Sciences, Kochi, Amrita Vishwa Vidhyapeetham, India.

Deepa G, Assistant Proffessor, Department of Computer Science and IT, Amrita School of Arts and Sciences, Kochi, Amrita Vishwa Vidhyapeetham, India.