PERFORMANCE OF CLASSIFICATION ALGORITHMS WITH WEKA AND SPARK TOOLS

[1]Kasarapu Ramani
[1]Professor and Head
[1]Department of IT
[1]Sree Vidyanikethan Engg College (Autonomous), Tirupati, India

Abstract— Because of the extensive number of impacting factors, it is hard to foresee the natural disasters such as earthquake. Analysts are working seriously on earthquake forecast. Death toll and property can be limited with earthquake prediction. In this paper, the performance of the methods such as the Decision Tree and Naïve Bayes has been compared to find the algorithm that best fits the earthquake prediction. Also the performance of these classification algorithms is tested on two different tools such as Weka and Spark. The performance is expressed in terms of parameters correctly classified instances, incorrectly classified instances, errorrate and precision. The Decision Tree algorithm has given improved precision than the Naïve Bayes by approximately 3-4 percentage. It can be concluded by the analysis that the performance of both the algorithms was high when performed in the Spark tool as compared to the Weka tool.

Index Terms — Classification, Decision Tree, Naïve Bayes, Earthquake Prediction.

I. INTRODUCTION

The main objective of paper is to study the impact of Decision Tree and Naïve Bayes classification algorithms on the Seismic bumps dataset in Weka and Spark tools. The parameters for judging the algorithms are correctly classified instances, incorrectly classified instances, error rate and precision. These are helpful when training data is used instead of testing data and comparing them to know the correctly classified instances, incorrectly classified instances, error rate and precision of the particular algorithm. This paper is categorized as follows. Section II inclines the related work. Section III gives the procedure and discusses the characteristics of the classification algorithms and the dataset. Section IV gives analysis of the generated by the algorithms. Section V concludes the paper.

II. RELATED WORK

The results of [1] proved that the Random Forest Algorithm gives better results on large datasets keeping the same number of attributes while Decision Tree is a finest and easy method for smaller datasets with less number of instances.[2] proposed a system for earthquake prediction by investigating seismic bump data. 94.11% accuracy was achieved in the study through the k nearest neighbour algorithm. In [3] proposed a method to enable system for earthquake prediction and it achieves 91% classification accuracy through Support Vector Machine classification algorithm.

III. METHODOLOGY

The following are the steps included in the classification process carried out in this work:

- Seismic-bumps dataset is chosen for the classification process.
- Two different classifiers namely-Decision Tree and Naive Bayes are chosen.
- Two different tools Weka and Spark are used to perform the classification by each of the classifier.
- The correctly classified instances, incorrectly classified instances, error rate and precision of each classifier are calculated.
- Finally the results are analysed and the best suited algorithm for the chosen dataset is found. The performance of both the tools is also analysed.

III.I Dataset

The dataset considered in this work is seismic-bumps dataset from UCI machine learning repository. The data describe the problem of high energy (higher than 10^4 J) seismic bumps forecasting in a coal mine. The dataset is composed of 19 attributes with one attribute for the class label. The dataset has a total of 2584 instances. The class distribution of the dataset is as follows:

"hazardious state" (class 1) : 170 instances (6.6%)
"non-hazardious state" (class 0): 2414 instances (93.4%)

III.II Classifiers

III.II.I Decision Tree

A decision tree classifier is a classifier that classifies the given input model into one of its possible classes. Decision tree classifier is a tree structured classifier that classifies by extracting knowledge through making decision rules from the huge amount data. A decision tree classifier is a simple form of classification which is briefly stored and can powerfully classify new data. The advantages of decision tree classifier are its ability to handle different types of input data such as textual,numerical and nominal. Its ability to handle missing values and errors in the datasets. Its availability across various platforms in different packages.

III.II.II Naive Bayes

A Naive Bayes classifier assumes that the incidence of a particular feature in a class is not related to the incidence of any other feature. Naive Bayes classifier is a simple classifier that is based on the Bayes Theorem of conditional probability along with strong independent assumptions. This classifier emphasizes on measure of probability that whether the document belong to a particular class or not. It is an independent feature model. It is based on the assumption that the occurrence or non-occurrence of a specific attribute is unrelated to the occurrence or non-occurrence of a specific attribute. The major benefit of Bayesian classifier is that it needs only a small training data set for classification. It is efficient, easier for implementation and fast to classify. It is non-sensitive to extraneous features.

III.III Tools

III.III.I WEKA

The full form of WEKA is Waikato Environment for Knowledge Learning. Data pre-processing, classification, clustering, association, regression and feature selection are the standard data mining tasks supported by Weka tool. It is an open source application available. In Weka datasets should be structured to the ARFF format. Weka Explorer provides the classification tasks through the classify tab. Weka uses a variety of classifiers such as Bayes, function, tree etc.

III.III.II Spark

Apache Spark is a general purpose cluster computing engine which is very fast and reliable. This system provides Application programing interfaces in various programing languages such as Java, Python, Scala. Spark tool is specialized at making data analysis faster. The in-memory processing capability of spark makes it much faster than any traditional data processing engine. Spark also provides enormous impressive high level tools such as machine learning tool M Lib, structured data processing, Spark SQL, graph processing took Graph X, stream processing engine called Spark Streaming, and Shark for fast interactive question device. The classification algorithms supported by Spark are part of the Spark machine learning tool mlib.

IV. RESULTS

The experimental setup used includes Windows 10 Operating System, intel core i5 processor, 8GB RAM, Weka tool version 3.8.1 and Spark tool version 1.6.1. The Results of following analysis on the seismic-bumps dataset are clearly given by the tables 1, 2 and 3. Tables 1 and 2 have given the positive and negative instances correctly classified with total number of training and testing instances in the dataset using Decision Tree and Naïve Bayes classifiers in Weka and Spark tools respectively. Table 3 listed the error rate and precision measures to analyse the classifiers in both Weka and Spark.

Comparing the Decision Tree and Naïve Bayes Classification Algorithms in both Weka and Spark tools, it can be concluded that the performance of the Decision Tree Classifier is better on the considered seismic-bumps dataset. The Decision Tree classifier is 3-4 % more accurate than the Naïve Bayes classifier. Also an improved performance of nearly 3.5 % on an average is achieved through the Spark tool. The pictorial representation of this analysis is provided through Fig. 1 and 2.

Table 1 Comparing Decision Tree and Naïve Bayes Classification Algorithms in Weka

WEKA Classification Algorithm	**No of Training instances**	**No of testing instances**	**No of positive instances correctly identified**	**No of negative instances correctly identified**	**No of correctly identified instances**
J48(Decision Tree)	1809	775	5	711	716
NaiveBayes	1809	775	21	653	674

Table 2 Comparing Decision Tree and Naïve Bayes Classification Algorithms in Spark

Spark Classification Algorithm	**No of Training instances**	**No of testing instances**	**No of positive instances correctly identified**	**No of negative instances correctly identified**	**No of correctly identified instances**
Decision Tree	1755	829	25	756	781
NaiveBayes	1793	791	13	727	740

Table 3 Comparing the performance of Classification Algorithms in Weka and Spark Tools

	Error rate	**Precision**
Weka Decision Tree	0.0761	0.9238
WekaNaiveBayes	0.1303	0.8696
Spark Decision Tree	0.058	0.942
Spark NaiveBayes	0.0644	0.9355

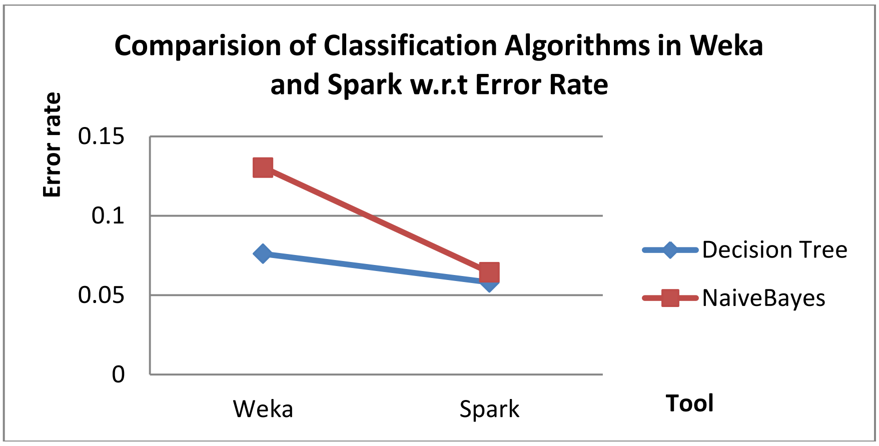

Fig. 1 Comparing Decision Tree and Naïve Bayes with its Error rate in Weka and Spark

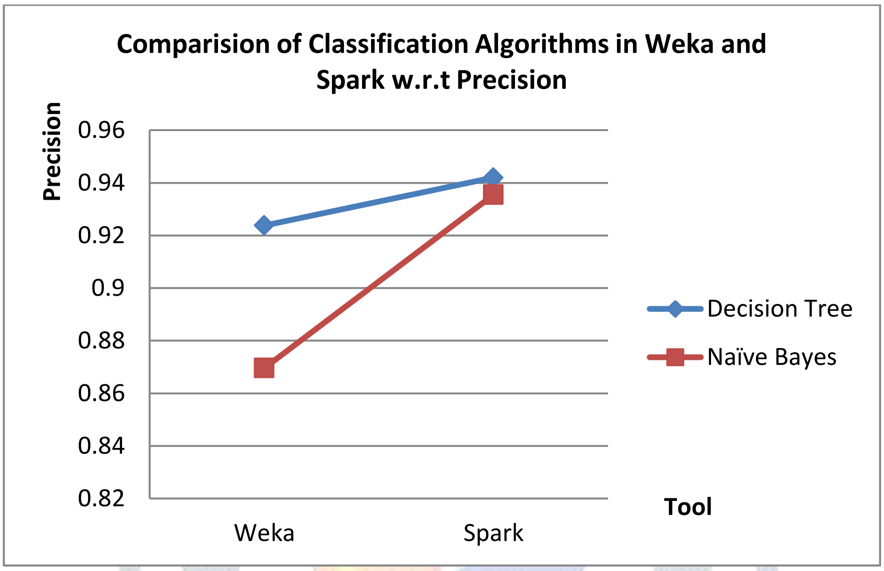

Fig. 2 Comparing Decision Tree and Naïve Bayes with its Precision in Weka and Spark

V. CONCLUSION AND FUTURE WORK

In this paper we have compared the performance of Decision Tree and Naïve Bayes classifiers in both Weka and Spark tools. Seismic-bumps dataset is used for experimentation from the UCI machine learning repository. It is concluded that the performance of Decision Tree classification technique was better on the considered data set. Also higher performance was achieved through Spark tool. Our future work will focus on improvement of the classification Technique thus improving the effectiveness of classification in reduced time.

REFERENCES

[1] Jehad Ali, Rehanullah Khan, Nasir Ahmad, Imran Maqsood," Random Forests and Decision Trees", IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 5, No 3, September 2012.

[2] M. Bilen, A.H. Işık, T. Yiğit, “Seismic hazard prediction with classification of seismic pulses”, International Burdur Earthquake & Environment Symposium (IBEES2015), Burdur, Turkey, 2015, 41-48.

[3] E. Celik, M. Atalay, and H. Bayer, “Earthquake prediction using seismic bumps with Artificial Neural Networks and Support Vector Machines”, Signal Processing and Communications Applications Conference, Trabzon, Turkey, 730-733, 2014.

[4] V. Vaithiyanathan, K. Rajeswari, KapilTajane and Rahul Pitale, “Comparison of Different Classification Techniques Using Different Datasets”, International Journal of Advances in Engineering & Technology,ISSN: 2231-1963, May 2013.

[5] Ananthi S and G. Thailambal, “Comparison of Classification Algorithms in Text Mining”, International Journal of Pure and Applied Mathematics, Volume 116 No. 22, 425-433, 2017.

[6] Musa Peker, “Seismic Hazard Prediction Using Seismic Bumps: A Data Mining Approach”, American Journal of Engineering Research, Volume-5, Issue-4, pp-106-111, 2016.