Comparison of Classification Techniques Using Mushroom Datasets

Ameer Rashed Khan [1], Rajeswari [2], Dr.S.Shajun Nisha
[1, 2] Dept of Computer Science
[3] Asst.Prof & Head, Dept of Computer Science
[1, 2, 3] Sadakathullah Appa College, Tirunelveli, India

Abstract- *Data mining is a process of extracting information from datasets and it converts into useful information to make knowledge it extracts the hidden data that enhances the knowledge. It also referred as KDD (Knowledge Discovery in Databases) various techniques have been used in data mining like Classification, Cluster, and Association etc. Classification is one of the major data mining techniques known as supervised learning used to extract important data into class labels. This paper focuses on different classification techniques ZeroR, Bayes Net and J48 for mushroom dataset and compares the results using WEKA tool. WEKA stands for Waikato Environment for Knowledge Analysis developed in Java. Finally the results show the best techniques for mushroom datasets by its Accuracy, Mean Absolute Error and Kappa Statistic.*

Keywords- ZeroR, Bayes Net, J48, KDD, WEKA.

I. INTRODUCTION

Data mining is a very crucial research domain in recent research world and it is an interdisciplinary subject used in various areas. Data mining is a process of extracting information from datasets and it converts into useful information to make knowledge it extracts the hidden data that enhances the knowledge. It derives its basics from statistics, artificial intelligence and machine learning. Supervised and Un-supervised are two different types of learning methods in the data mining. In Data mining, classification is major technique widely used in various fields where it assigns data into a group or class labels. Agriculture is the most significant function area particularly in the developing counties like India. The yield of agriculture is very low as compare to land which farmers have. Use of information technology in agriculture can change the position of decision making and farmers can yield in better way. Mushroom cultivation is gradually becoming popular outcome all over the world, it not only the nutritional food, medicinal purposes and additionally adds to the income, particularly for the growers with poor land. It becomes hobby for the aged persons as well as homemakers those who can grow mushrooms in small areas and it faces only less difficulties for cultivations. Cultivation of mushroom has been in trend for almost 200 years. Although India has started mushroom farming recently. It is an excellent source of vitamins, minerals, protein and is a good source of iron for anemic patients. There are many characters of mushrooms to analyze suitably with data mining tools.

A. Related Work

When western countries are taking help of advance computer technologies in determining when, what, where and how of agriculture and animal husbandry, we have not even completely digitized our agriculture data [07]. The advances in computing and information storage have provided vast amounts of data. The challenge has been to extract knowledge from this raw data that has lead to new methods and techniques such as data mining that can bridge the knowledge gap [10]. Data Mining involves the various data analysis tools for identifying previously unknown, valid patterns and relationships in huge data set. There are different data mining techniques like classification association, preprocessing, transformation, clustering, and pattern evaluation. Classification and Association are the popular techniques used to predict user interest and relationship between those data items which has been used by users [02]. For classification, a small amount of training data is expected to predict the class labels [05]. The demand of food is increasing now a day; so the researchers, farmers, agricultural scientists and government are trying to put extra effort and techniques for more production. And as a result, the agricultural data increases day by day. Still today, a very few farmers are actually using the new methods, tools and technique of farming for better production. Data mining can be used for predicting the future trends of agricultural processes [09]. Data mining plays a crucial role for decision making on several problems related to agriculture field [06].Enhancement of food technology is today's need. India is living in the era of huge population wherein the ratio and proportion of food and humans has no toning, resulting in high rates of inflation so that we keep getting high quality and good quantity crop productions. Therefore the researchers are always thinking, about how to increase productivity of crops. The more the production of food material, the cheaper will be the cost of

food products [12]. Raw data is useless without techniques to extract information from it. According to I. H. Witten and E. Frank [04].Selection of a learning technique is a difficult task that depends on the database and the types of desired results. Schlimmer found optimal rules for the Mushroom database using back propagation methods in 1987 to 95% accuracy [04]. Classification of data is one of the major steps towards extracting useful information in data mining [7]. It is the process of findings a model or function that describes and distinguishes data classes and concepts. Classification enables us to predict the classes of object whose class label is not known [11]. Zero Rule or ZeroR is the simplest classification method which relies on the target and ignores all predictors [13]. In Bayes Net classifier conditional probability of each node is calculated first and then Bayesian Network is formed. Bayesian Networks is a directed acyclic graph [02]. Bayes Net can compute the probabilities of another subset of variable called query variable [03]. J48 is a program that creates a decision tree based on a set of labeled input data. This decision tree can then be tested against unseen labeled test data to quantify how well it generalizes The decision trees generated by C4.5 can be used for classification, and for this reason, C4.5 is often referred to as a statistical classifier [08]. To make a decision, the attribute with the highest normalized information gain is used [02]. WEKA is a collection of machine learning algorithms for data mining tasks. It is a comprehensive collection of data preprocessing and modeling techniques. It is fully implemented in the Java programming language and thus runs on almost any modern computing platform [13].

B. Motivation and Justification

Data mining is an important part of knowledge discovery process that we can analyze an enormous set of data and get hidden and useful knowledge. Data mining in agriculture is a very recent research topic. Recent technologies are now a day's able to provide a lot of information on agricultural-related activities, which can be analyzed in order to find important information for agriculture. Mushroom cultivation is gradually becoming popular outcome all over the world, it not only the nutritional food, medicinal purposes and additionally adds to the income, particularly for the growers with poor land. Classification is used to find out in which group each data instance is related within a given dataset. It is used for classifying data into different classes according to some constrains. Several major kinds of classification algorithms including J48, Bayes Net, ZeroR, KNN classifier, Naive Bayes, SVM, and ANN are used for classification. Bayes Net classifier is easy to implement in which it requires a small amount of training data to estimate the parameters and provides good results in most of the cases. The J48 algorithm is based on C4.5 algorithm which can be easily interpreted and easy to implement. This motivates to work in WEKA tool because it has been mostly used for agricultural purposes.

II. METHODOLOGY

A. Outliine of The Work

Our objective of data mining techniques is to separate eatable mushrooms from poisonous ones using different classification algorithms such as ZeroR, Bayes Net and J48 in WEKA. The performance of these classification algorithms are investigated using Accuracy, Mean absolute Error, Time and Kappa Statistic for the given mushroom data set.

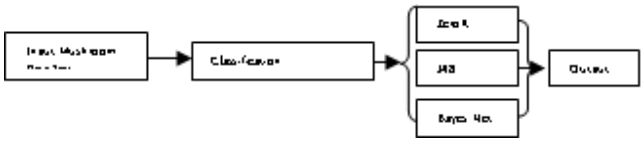

Fig.0 Flow Diagram

Dataset

The Mushroom dataset has been collected from the UCI Repository. It consists of 8124 instances and 22 attributes in .CSV file format.

Classification

Classification maps data into predefined different groups or different classes by their similarities it gives class labels to orders the objects in the data collection it also known as supervised learning. Classification is done using two stages learning and testing stage. Various algorithms are used under classification techniques they are ZeroR, Bayes Net and J48.

ZeroR

ZeroR is the simplest classification method which relies on the target and ignores all predictors. It simply predicts the majority category (class). Although there is no predictability power in ZeroR it is useful for determining a baseline performance as a benchmark for other classification methods. Algorithm Construct a frequency table for the target and select it's most frequent value. Predictors Contribution There is nothing to be said about the predictor's contribution to the model because ZeroR does not use any of them. Model Evaluation the ZeroR only predicts the majority class correctly. As mentioned before, ZeroR is only useful for determining a baseline performance for other classification methods [14].

Bayes Net

Bayesian networks are directed acyclic graphs whose nodes represent random variables in the Bayesian sense. Edges represent conditional dependencies; nodes which are not connected represent variables which are conditionally independent of each other. Based on Bayesian networks, these classifiers have much strength, like model interpretability and accommodation to complex data and classification problem settings [02].

J48(C4.5)

Decision trees are used in the data mining process. Using the algorithm C4.5 the decision trees are generated. Ross Quinlan developed the C4.5 algorithm. The decision trees are generated using set of labeled input data. In the data mining tool WEKA, the C4.5 algorithm is implemented and termed as J48 by using JAVA. J48 is a free classifier and accepts nominal classes only which use both continues and discrete attributes. The output of the J48 is in the form of Decision tree [03].

III. EXPERIMENTAL RESULTS

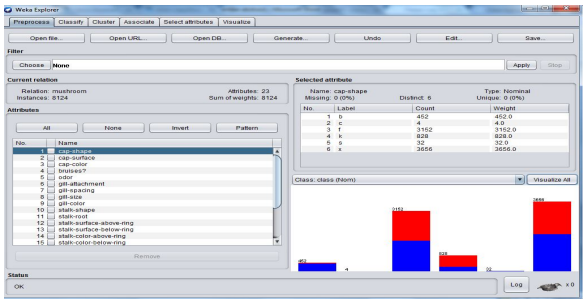

Figure 1. Pre Processing

By using WEKA, we visualize all the 22 attributes with histograms. This helps us to examine the distribution and think about what it's telling us. The histograms are listed below.

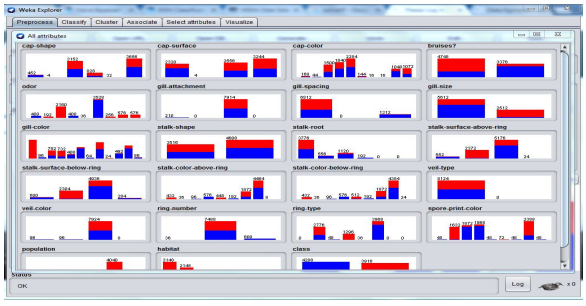

Fig.2. Attributes of Mushroom

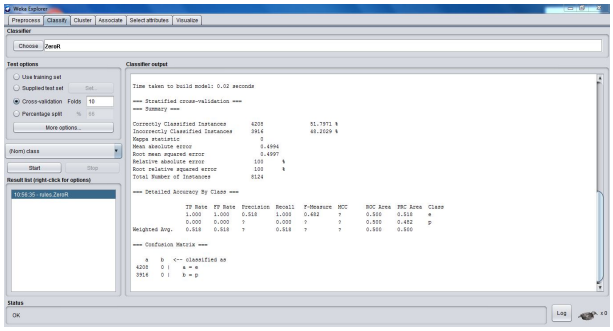

Fig.3. Result for ZeroR Algorithm

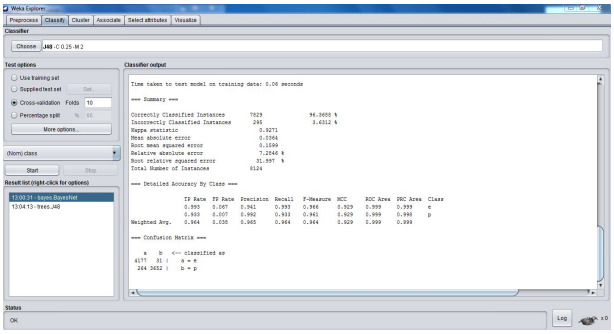

Fig.4. Result for Bayes Net Algorithm

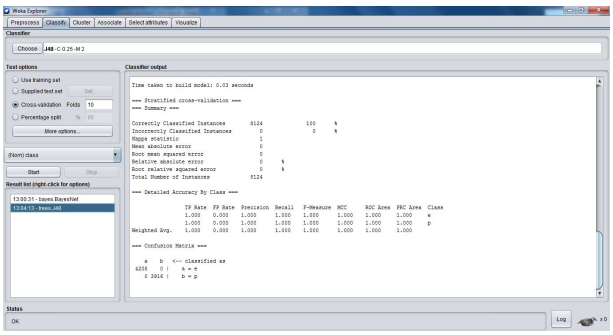

Fig.5. Result for J48 Algorithm

Performance of Various Classification Algorithms for Mushroom Data Sets

S. No.	Name Of The Algorithm	Accuracy	Time	Kappa Statistic	Mean Absolute Error	Correctly Classified Instances	In Correctly Classified Instances
1	ZeroR	51.7971%	.01 sec	0	0.4994	4208	3916
2	Bayes	96.36	.06	0.92	0.03	7829	295

	Net	88%	sec	71	64		
3	J48	100%	.03 sec	1	0	8124	0

Table.1. Comparison of Accuracy, Time, Kappa Static, Mean Absolute Error of Algorithms

Accuracy is defined as the number of instances classified correctly. The Accuracy of a classifier on a given set is the percentage of test set tupels that are correctly classified by the classifier. Accuracy = TP + TN / P + N [02].Kappa Statistic: Inter observer variation can be measured in any situation in which two or more independent observers are evaluating the same thing. Mean Absolute Error: The mean absolute error is a quantity used to measure how close forecasts or predictions are to the eventual outcomes. Correctly Classified Instance: If the problem is a multi-class one (i.e. more than two classes) then AUC is calculated for each class in turn by treating all other classes as the negative class. It is possible to achieve a high AUC on one class while the overall classification accuracy is somewhat lower. Incorrectly Classified Instance: Incorrectly classified instances refer to the case where the instances are used as test data and again are the most important statistics here for our purposes. Time is based on the processing to build a model for a given data sets.

IV. CONCLUSION

Our goal is to evaluate and compare the different classification techniques as ZeroR, Bayes Net, and J48 among them J48 gives the best result for mushroom data set. This result is given after analysis on Accuracy, Mean Absolute Error, Correctly Classified, Incorrectly Classified Instances, Kappa Statistics and Time. In future I have been implemented in IoT datasets for Agriculture, Mushrooms and so on

REFERENCES

[1] Aditya Shastry, Sanjay, Kavya “A Novel Data Mining Approach For Soil Classification”
[2] Deepali Kharche, K. Rajeswari, Deepa Abin “Comparison of Different Datasets Using Various Classification Techniques with WEKA”
[3] Hemageetha, Nasira “Analysis of Soil Condition Based On Ph Value Using Classification Techniques “K. Elissa, “Title of paper if known,” unpublished.
[4] Hemendra Pal Singh, Seth Gyaniram “Data Mining: The Mushroom Database”
[5] E.Manjula, S.Djodiltachoumy “Data Mining Technique to Analyze Soil Nutrients Based On Hybrid Classification"
[6] B. Murugesakumar, Dr. K.Anandakumar, Dr. A.Bharathi “A Survey on Soil Classification Methods Using Data Mining Techniques”
[7] Nikhita Awasthi, Abhay Bansal “Application of Data Mining Classification Techniques on Soil Data Using R”
[8] [8] Prediction Jay Gholap, Anurag Ingole, Jayesh Gohil, Shailesh Gargade, Vahida Attar “Soil Data Analysis Using Classification Techniques And Soil Attribute”
[9] [9] Rajshekhar Borate, Rahul Ombale, Sagar Ahire, Manoj Dhawade, Mrs. Prof. R. P. Karande “Applying Data Mining Techniques To Predict Annual Yield Of Major Crops And Recommend Planting Different Crops In Different Districts In India”
[10] Singaraju, Jyothi, Peyakunta, Bhargavi “Applying Naive Bayes Data Mining Technique for Classification of Agricultural Land Soils”
[11] Sunita Beniwal, Bishan “Mushroom Classification Using Data Mining Techniques”
[12] Vrushali Bhuyar “Comparative Analysis of Classification Techniques on Soil Data to Predict Fertility Rate for Aurangabad District”
[13] Wikipedia
[14] Zero Maps