Int J Pharm Bio Sci 2015 Jan; 6(1): (B) 1170 - 1176

Research Article **Allied sciences**

International Journal of Pharma and Bio Sciences **ISSN 0975-6299**

MUSHROOM CLASSIFICATION USING DATA MINING TECHNIQUES

SUNITA BENIWAL[1*] AND BISHAN DAS[2]

[1]Department of Computer Science and Engineering, Guru Jambeshwar University of Science and Technology. Hisar125001, Haryana
[2]Department of Computer Science and Engineering, Maharishi Markandeshwar University, Mullana, Ambala, Haryana 133207, India

ABSTRACT

This paper focuses on the use of classification techniques for analyzing mushroom data set. Mushroom dataset is composed of records of different types of mushrooms, which are edible or non- edible. WEKA (Waikato Environment for Knowledge Analysis) is used for implementation of the classification techniques. Different classification techniques like naïve bayes, bayes net, and ZeroR are used to categorize different mushrooms and the performance of the classification techniques is evaluated using accuracy, mean absolute error, kappa statistic. After analyzing it was found that bayes net outperformed the other techniques with highest accuracy, lowest mean absolute error and naïve bayes is the second best performer. It was also found that accuracy increased with the increase in size of the training set.

KEYWORDS : **Bayes net, Naïve bayes, KDD, Accuracy.**

*Corresponding author

SUNITA BENIWAL
Department of Computer Science and Engineering, Guru Jambeshwar University of Science and Technology. Hisar125001, Haryana

INTRODUCTION

Agriculture data can be analysed for taking various decisions like to increase the productivity, to classify soil type, to categorize different types of crops etc. Large amounts of data of different fields are available. Manually analyzing the large amounts of data is cumbersome task; data mining techniques can be applied for the same. The term KDD refers to the automated process of knowledge discovery from databases. KDD has many steps for analysis, namely data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation and knowledge representation. Data mining is a step in the whole process of knowledge discovery which can be applied to a dataset to extract any previously unknown, valid, novel, useful and understandable patterns if present in the dataset. Data mining is the non-trivial process that automatically collects the useful hidden information from the data and is taken on as forms of rule, concept, pattern and so on[1]. The mining techniques can be broadly divided into four categories: class/concept description, association analysis, classification or prediction and clustering analysis[2]. Data pre-processing is required before any data mining technique can be applied. Data pre-processing can be done using steps like data integration, data cleaning, discretization, and attribute selection[3].This paper presents the use of different classification techniques on mushroom data to classify various types of mushrooms as edible or non-edible. Section I provides an overview of classification techniques used in this paper and different parameters used for measuring the performance of different techniques. Section II provides a brief review of weka and the dataset used. In Section III results of classification techniques on mushroom data is given.

RELATED WORK

Classification is the process of finding a model or a function that describes and distinguishes data classes and concepts. Classification enables us to predict the classes of objects whose class label is not known. Classification is done using two steps learning and testing. The data set is divided into two before doing classification: training set and test set. In learning step, classifier is built describing a predetermined set of classes or concepts by analyzing the training set made up of database tuples and their associated labels. In the second step model is used for classification by estimating the predictive accuracy of classifier built during the first step using the test data. The accuracy of classifier on a given test set is percentage of tuples that are correctly classified by the classifier. If the accuracy is above some acceptable level, the classifier can be used to predict future tuples whose class label is not known[4]. Different types of classifiers are available like rule based classifiers, decision trees, support vector machines, bayes classifiers, genetic algorithms, neural networks etc.[3]. The techniques which have been used for mushroom classification are ZeroR, Bayes Network and Naïve Bayes.ZeroR is a trivial classifier, but it gives a lower bound on the performance of a given dataset which should be significantly improved by more complex classifiers. The ZeroR is the simplest method which relies on the frequency of target. A ZeroR classifier simply assigns every value to the most common class after examining the training data.A Bayesian network consists of a directed, acyclic graph and a probability distribution for each node in that graph given its immediate predecessors[5]. A Bayes Network Classifier is based on a bayesian network which represents a joint probability distribution over a set of categorical attributes. It consists of two parts, the directed acyclic graph G consisting of nodes and arcs and the conditional probability tables. Assuming discrete variables, the strength of the relationship between variables is quantified by conditional probability distributions associated with each node. The nodes represent attributes, whereas the arcs indicate direct dependencies. The density of the arcs in a network is one measure of its complexity. Sparse BNs can represent simple probabilistic models like naïve Bayes models, whereas dense BNs can capture highly complex models[6].A Naive Bayes classifier is a simple Bayes net probabilistic classifier with strong independence assumptions. The probability

model is such that a naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. It is an efficient approach to supervised classification with extremely high accuracy because of the precise nature of the probability model.The measures used for performance analysis of various techniques are accuracy, mean absolute error, Kappa statistic. Classification accuracy refers to the ability of the model to correctly predict the class label of new or previously unseen data. Mean Absolute Error or MAE as the name suggests, is the mean of the absolute errors. The absolute error is the absolute value of the difference between the forecasted value and the actual value. MAE tells us how big of an error, we can expect from the prediction. Kappa statistic is used to measure the concordance level between categorical data during prediction.P. Bhargavi, Dr. S. Jyothi in their research used data mining algorithms like Genetic algorithm, Fuzzy classification and Fuzzy clustering for classifying soil data. Both Supervised and Unsupervised techniques are used for mining. GA and Fuzzy Classification rules were used for supervised learning where Fuzzy rules performed better than GA as reported. Fuzzy C-Means algorithm was used for unsupervised learning[7]. P. Revathi, Dr. M. Hemalatha used J48, multilayer perceptron, Naive Bayes etc. for classification of cotton seeds on the germination stages and to find whether meaningful patterns exist among seed germination and seedling. reported that J48 performed better with more accuracy[8].Clara Eusebi et al used various data mining techniques to analyze Mushroom Database and used a human-machine interface to increase the accuracy of machine learning using data mining tool Weka where J48 performed best amongst all used[9]. Sally Jo Cunningham and Geoffrey Holmes used image data of mushrooms to categorize mushrooms into quality grades and achieving accuracy similar to that attained by human inspectors. They reported that visually-based attributes which can be automatically extracted from digitized images are sufficient for good separation of mushrooms[10].

METHODOLOGY

The mushroom data set was retrieved from the UCI repository[11] and it evaluates samples of mushrooms from the Agaricus and Lepiota Family and then they are classified as definitely edible, definitely poisonous or of unknown edibility and not recommended. This latter class is combined with the poisonous one. The mushroom dataset chosen has a large number of instances (8124), large number of attributes(22) , 2 possible class labels and a very even class distribution (out of 2 possible class values 52.8% are edible and 48.2% are poisonous).Weka is used for classification of the mushroom database[12]. Weka (Waikato Environment for Knowledge Analysis) is a popular suite of machine learning software written in Java, developed at the University of Waikato, New Zealand. The Weka workbench contains a collection of visualization tools and algorithms for data analysis and predictive modelling, together with graphical user interfaces for easy access to this functionality. ARFF file format is used for storing the data. Attribute Relationship File Format (ARFF) is the text format file used by Weka to store data in a database. For the purpose of training and testing, dataset is split into different percentages from the whole dataset. The performance on the selected classification algorithms, namely Baysnet, Naïve Bayes, and ZeroR classifier are investigated using Accuracy, mean absolute error, kappa statistic on the given dataset. Different size of training data is used for this purpose, i.e. 40, 50, 60, 70, 80 (%) for measuring the performance.

RESULTS AND DISCUSSION

Bayes Net Classifier Technique

Table 1 shows the values of different parameters like total no. instances, correctly classified instances and Kappa Statistic with the respective training dataset size for bayes net classification technique. It is clear from the table that the accuracy is high when size of training dataset is large as compared to when dataset is small and mean absolute error decreases as size of dataset increases. The values of Kappa Statistic also increase with the increase in size of training set.

Table1
Simulation result of algorithm Bayes Net

S. No.	Training Size (%)	No. of Instances used for training (8124)	Correctly Classified Instances %(value)	Incorrectly Classified Instances %(value)	Mean Absolute Error	Kappa Statistic
1	40	3250	95.0462	4.9538	0.0477	0.9009
2	50	4062	96.0364	3.9636	0.0402	0.9202
3	60	4874	95.9921	4.0008	0.0413	0.9196
4	70	5687	97.2217	2.7783	0.0289	0.9402
5	80	6499	97.1842	2.8155	0.03	0.9428

Naïve Bayes Classifier Technique
Table 2 shows the performance for the Naïve Bayes classifier techniques on the mushroom dataset using different parameters like accuracy, mean absolute error and Kappa Statistic on different training dataset sizes. The table shows that highest accuracy is when 70 % tuples are used for training set. The other training set yields an average accuracy of around 95%. Mean absolute error decreases with the increase in size of training set. When training dataset is small the value of Kappa Statistic is low as compared to larger size of training set.

Table2
Simulation result of algorithm Naïve Bayes

S. No.	Training Size (%)	No. of Instances used for training (8124)	Correctly Classified Instances %(value)	Incorrectly Classified Instances %(value)	Mean Absolute Error	Kappa Statistic
1	40	3250	94.4615	5.5385	0.0523	0.8892
2	50	4062	95.5441	4.4559	0.0443	0.9103
3	60	4874	95.5478	4.4522	0.0452	0.9165
4	70	5687	96.8173	3.1827	0.0333	0.9316
5	80	6499	96.738	3.262	0.0338	0.9338

ZeroR Classifier Technique:
Table 3 shows the performance for the ZeroR classifier technique on the mushroom dataset using different parameters like accuracy, mean absolute error and Kappa Statistic on different training dataset sizes.

Table 3
Simulation result of algorithm ZeroR

S.No.	Training Size (%)	No. of Instances used for training (8124)	Correctly Classified Instances %(value)	Incorrectly Classified Instances %(value)	Mean Absolute Error	Kappa Statistic
1	40	3250	50.1846	49.8154	0.5	0
2	50	4062	52.6342	47.3658	0.4986	0
3	60	4874	52.2158	47.7842	0.499	0
4	70	5687	64.2518	35.7482	0.4594	0
5	80	6499	57.2857	42.7143	0.4894	0

It is clearly evident from the table that the accuracy increases with increase in dataset size and it is maximum for training set of 70 % of the whole data set. Mean Absolute Error decreases gradually from 40% to 70% of training data set size. Kappa statistic has constant zero value.

Comparison of different classification techniques

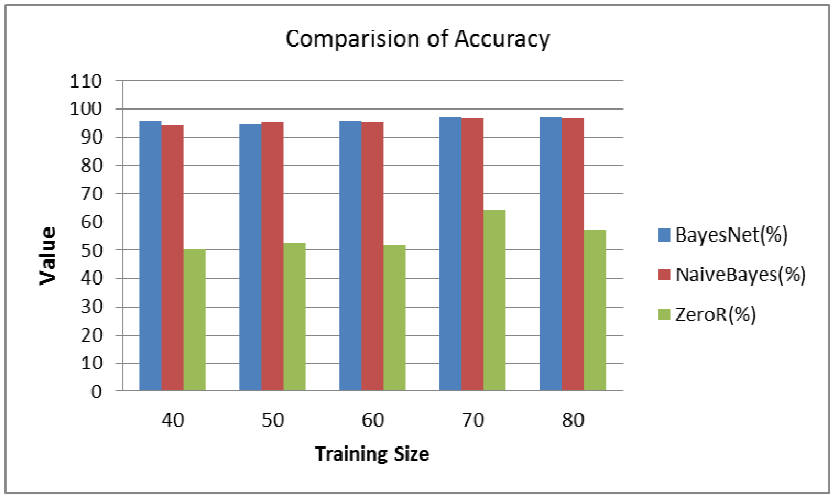

Figure 1
Comparison based on Accuracy

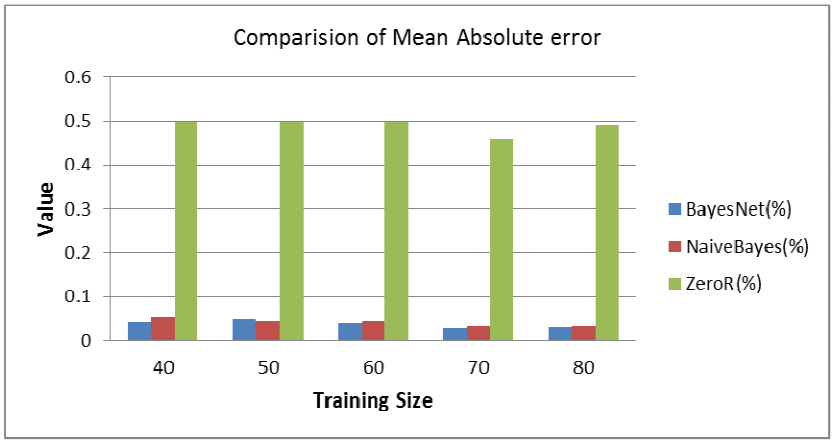

Figure 2
Comparison based on Mean Absolute Error

Figure1 clearly depicts that the accuracy of Bayes Net classifier is the best among these three classifier techniques.Figure 2 shows that the highest Mean absolute error is 0.0477% and lowest Mean absolute error is 0.0289% in the Bayes Net classifiers. And the highest Mean absolute error is 0.0523% and lowest Mean absolute error is 0.0333% in the Naïve Bayes classifiers and the highest Mean absolute error is 0.5% and lowest Mean absolute error is 0.4594% in the ZeroR classifiers. From the above graph we can clearly see that the Mean absolute error rate of ZeroR classifier is the highest among these three classifier techniques.

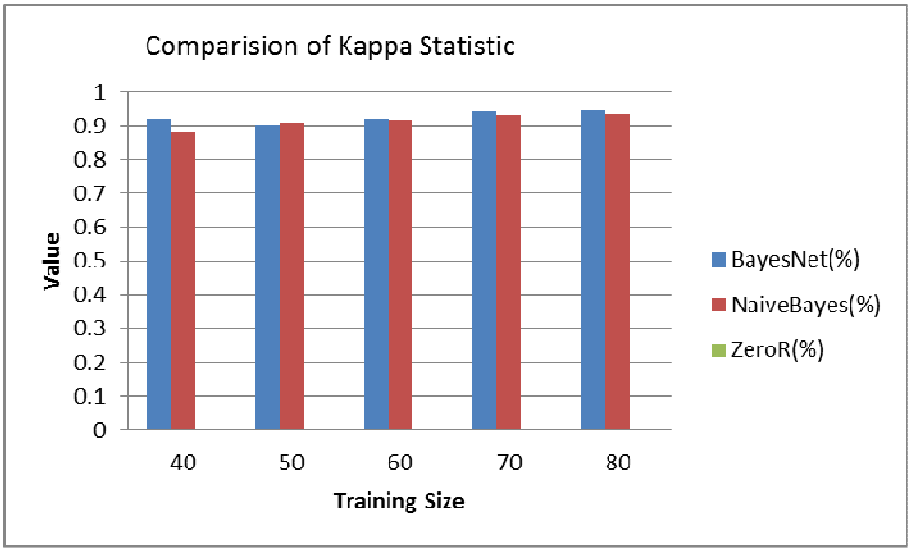

Figure 3
Comparison based on Kappa Statistic

Figure 3 shows that the Kappa Statistic is zero in the Zeros classifiers for all training size. From the above graph we can clearly see that the Kappa Statistic rate of Bayes Net classifier is the highest among these three classifier techniques.

CONCLUSION

It can be seen from the above results that the bayes net classification technique performs best among the three techniques used for classification. It is also seen that performance of all the techniques is low when dataset size is small and the performance improves with increase in size of training set up to when training set is 70% of the whole dataset. So it is very clear that size of training set as well as selection of classification technique depending on the data to be analysed is very important for mining of patterns efficiently.

REFERENCES

1. Shan TJ., Wei H., Yan Q. Application of genetic algorithm in data mining. First International Workshop on Education Technology and Computer Science, 2, 353- 356. (2009)
2. Shi ZZ. Knowledge discovery, Beijing, Tsinghua University Press. (2001)
3. Kumar D., Beniwal S. Genetic algorithm and programming based classification: a survey. Journal of Theoretical and Applied Information Technology, 54(1), 48-58,(2013).
4. Han J., Kamber M. Data mining concepts and techniques. San Francisco, Morgan Kaufmann, 285-289(2006).
5. Darwiche A. Modeling and Reasoning with Bayesian Networks. Cambridge University Press,(2009).
6. Beniwal S., Arora J. Classification and Feature Selection Techniques in Data Mining. International Journal of Engineering Research & Technology, 1(6), 1-6 ,(2012).
7. Bhargav P., Jyothi S. Soil Classification Using Data Mining Techniques: A Comparative Study. International Journal of Engineering Tends and Technology, July to Aug Issue, 55-59. (2011).
8. Revathi P., Hemalatha M. Categorize the Quality of Cotton Seeds Based on the Different Germination of the Cotton Using Machine Knowledge Approach.

International Journal of Advanced Science and Technology, 36, 9-14. (2011)
9. Eusebi C., Gliga C., John D., Maisonave A. Data Mining on a Mushroom Database. Proceedings of Student-Faculty Research Day, CSIS Pace University B2.1-B2.9 (2008).
10. Cunningham SJ., Holmes G. Developing innovative applications in agriculture using data mining. Proceedings of the Southeast Asia Regional Computer Confederation conference, Singapore, 1-12(1999).
11. Bache K., Lichman M. UCI Machine Learning repository. [http://archive.ics.uci.edu/ml]", Irvine, CA: University of California, School of Information and Computer Science, (2013).
12. Hall M., Frank E., Holmes G., Pfahringer B., Reutemann P., Witten IH. The WEKA Data Mining Software: An Update. SIGKDD Explorations, 11(1), 10-18, (2009)