

A Comparative Study of Classification Techniques On Adult Data Set

S.Deepajothi¹, Dr.S.Selvarajan²

¹Assistant Professor of department of CSE , Chettinad college of Engineering and Technology ,TamilNadu,India

²Direof the Muthayammal T Ccollege of Engineering and Technology, Tamilnadu, India

Abstract:

Data mining is the extraction of hidden information from large database. Classification is a data mining task of predicting the value of a categorical variable by building a model based on one or more numerical and/or categorical variables (predictors or attributes). Classification mining function is used to gain a deeper understanding of the database structure. There are various classification techniques like decision tree induction, Bayesian networks, lazy classifier and rule based classifier. In this paper, we present a comparative study of the classification accuracy provided by different classification algorithms like Naïve Bayesian, Random forest, Zero R, K Star on census dataset and provide a comprehensive review of the above algorithms on the dataset.

Keywords — Data Mining, Bayesian, classification technique.

1.INTRODUCTION:

Data mining involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data set. These tools can include statistical models, mathematical algorithm and machine learning methods. Consequently, data mining consists of more than collection and managing data, it also includes analysis and prediction [1]. Data mining applications can use a variety of parameters to examine the data. They include association, sequence or path analysis, classification, clustering, and forecasting .Classification technique is capable of processing a wider variety of data and is growing in popularity. The various classification techniques are Bayesian network, tree classifiers, rule based classifiers, lazy classifiers, Fuzzy set approaches, rough set approach etc.

Our next section discusses about Classification. Section 3 describes about Bayesian Network whereas decision tree classifier is described in section 4. Section 5 describes about rule based classifier Section 6 describes about lazy classifier .Section 7and 8 describes about the adult dataset and Experimentation and results. Finally, Section 9 concludes this work the last section discusses about future work.

2. CLASSIFICATION:

Classification predicts categorical class labels. It classifies the data based on the training set and the values in classifying the attributes and uses it in classifying the new data. Data classification is a two step process consisting of model construction and model usage. Model construction is used for describing predetermined classes. Model usage is used for classifying future or unknown objects. There are various preprocessing steps that may be applied to the data which helps to improve the accuracy, efficiency, and scalability of the classification process. They are data cleaning, relevance analysis, data transformation and reduction. The various classification techniques are discussed in the next sections.

3. BAYESIAN CLASSIFIERS

Bayesian classifiers are statistical classifiers. They can predict class membership probabilities, such as the probabilities, such as the probability that a given tuple belongs to particular class. Bayesian classification is based on Bayes Theorem. Bayesian classifiers exhibit high accuracy and speed when applied to large database. It consists of Naïve Bayesian Classifiers and Bayesian Belief Networks. Naive Bayesian Classifiers assume that the effect of an attribute value on a given class is independent of the values of the other attribute while Bayesian Belief Networks are graphical methods which allow the representation of dependencies among subsets of attributes. In this paper for comparative study of classification algorithms we have taken Naïve Bayesian Classification. The Naïve Bayesian classification is a simple and well-known method for performing supervised learning of a classification problem. It makes the assumption of class conditional independence, i.e, given the class label of a tuple, the values of the attributes are assumed to be conditionally independent of one another.

4.DECISION TREE INDUCTION

Decision tree induction is the learning of the decision trees from class-labeled training tuples. A decision tree is a flow chart-like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node holds a class label. The topmost node in a tree is the root node. Internal nodes are denoted by rectangles, and leaf nodes are denoted by ovals. Some decision tree algorithms produce only binary trees, whereas others can produce non binary trees. The construction of decision tree classifiers does not require any domain knowledge or parameter setting, and therefore is appropriate for exploratory knowledge discovery. They can handle high dimensional data and have good accuracy. The decision tree induction algorithm applied on the dataset for study is Random Forest. **Random forest** (or **random forests**) is an ensemble classifier that consists of many decision trees and outputs the class that is the mode of the classes output by individual trees. It runs efficiently on large data bases and can handle thousands of input variables without variable deletion. Generated forests can be saved for future use on other data.

5. RULE BASED CLASSIFIERS

Rule-based reasoning methods group methods providing explicit knowledge model, which can be expressed by formal rules or not, to be applied for further prediction. Provide a set of classification rules that can be used later to evaluate a new case and classify in a predefined set of classes. They classify records by using a collection of “if...then...” rules where a rule is represented as **Rule: (Condition) \rightarrow y**. Here condition is a conjunction of attributes and y is the class label. The left hand side denotes rule antecedent or condition and right hand side denotes rule consequent. Zero R classifier is applied on the adult dataset for comparative study. The idea behind the ZeroR classifier is to identify the most common class value in the training set. It always returns that value when evaluating an instance. It is frequently used as a baseline for evaluating other machine learning algorithms.

6. LAZY CLASSIFIERS

Lazy classifiers store all of the training samples and do not build a classifier until a new sample needs to be classified. It differs from eager classifiers, such as decision tree induction, which build a general model (such as a decision tree) before receiving new samples. In lazy classifiers, no general

model is built until a new sample needs to be classified. Lazy classifiers are simple and effective. However, it's slow at predicating time since all computation is delayed to that time. In this paper KStar classifier is applied on the dataset for comparative study with other classification algorithms. KStar is an instance-based classifier that is the class of a test instance is based upon the class of those training instances similar to it, as determined by some similarity function.

7.DATASET

Adult dataset available on UCI Machine Learning Repository and has a size of 3,755KB. The adult dataset consists of 32561 records and 15 attributes.

A. DATA PREPROCESSING

Data preprocessing is a type of processing on raw data to make it easier and effective for further processing. It is an important step in data mining process. The product of data preprocessing is the final training set. Kotsiantis et al. (2006) present a well known algorithm for each step of data pre-processing [8].

The data preprocessing techniques are

- Data Cleaning
- Data Integration

- Data Transformation
- Data Reduction

These data preprocessing techniques are not mutually exclusive; they may work together. Data processing techniques, when applied before mining, can substantially improve the overall quality of patterns mined and the time required for actual mining. Data preprocessing techniques can improve the quality of the data, accuracy and efficiency of the mining process.

B.PREPROCESSING OF ADULT DATASET

In order to improve the quality of the data, accuracy and efficiency of the mining process the adult dataset undergoes a preprocessing step. The less sensitive attributes like final weight, capital gain, capital loss, hours per week are removed since they are not considered as relevant attribute for privacy preservation in data mining. So the number of attributes is reduced to 10. The first 100 instances of the dataset is taken and then the instances with missing values are removed resulting in a dataset of 91 attributes.

8.EXPERIMENTATION AND RESULTS

Here we are going to implement the above described classification algorithms in the sections 3, 4, 5, 6,7 on the preprocessed adult dataset and tabulate the results. Here we have used Weka 3.5.7 for our experimentation. Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes. Table 1 shows the classification accuracy of each classification algorithms.

ID	WORK CLASS	EDUCATION	MARITAL STATUS	OCCUPATION	RELATIONSHIP	RACE	SEX	NATIVE COUNTRY	INCOME
1	Self-emp-not-inc	Bachelors	Married-civ-spouse	Exec-managerial	Husband	White	Male	United-States	<=50K
2	Private	HS-grad	Divorced	Handlers-cleaners	Not-in-family	White	Male	United-States	<=50K
3	Private	11th	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	United-States	<=50K
4	Private	Bachelors	Married-civ-spouse	Prof-specialty	Wife	Black	Female	Cuba	<=50K
5	Private	Masters	Married-civ-spouse	Exec-managerial	Wife	White	Female	United-States	<=50K
6	Private	9th	Married-spouse-absent	Other-service	Not-in-family	Black	Female	Jamaica	<=50K
7	Self-emp-not-inc	HS-grad	Married-civ-spouse	Exec-managerial	Husband	White	Male	United-States	>50K
8	Private	Masters	Never-married	Prof-specialty	Not-in-family	White	Female	United-States	>50K
9	Private	Bachelors	Married-civ-spouse	Exec-managerial	Husband	White	Male	United-States	>50K
10	Private	Some-college	Married-civ-spouse	Exec-managerial	Husband	Black	Male	United-States	>50K
11	State-gov	Bachelors	Married-civ-spouse	Prof-specialty	Husband	Asian-Pac	Male	India	>50K
12	Private	Bachelors	Never-married	Adm-clerical	Own-child	White	Female	United-States	<=50K
13	Private	Assoc-acdm	Never-married	Sales	Not-in-family	Black	Male	United-States	<=50K
14	Private	7th-8th	Married-civ-spouse	Transport-moving	Husband	Amer-ind	Male	Mexico	<=50K
15	Self-emp-not-inc	HS-grad	Never-married	Farming-fishing	Own-child	White	Male	United-States	<=50K
16	Private	HS-grad	Never-married	Machine-op-inspct	Unmarried	White	Male	United-States	<=50K
17	Private	11th	Married-civ-spouse	Sales	Husband	White	Male	United-States	<=50K
18	Self-emp-not-inc	Masters	Divorced	Exec-managerial	Unmarried	White	Female	United-States	>50K
19	Private	Doctorate	Married-civ-spouse	Prof-specialty	Husband	White	Male	United-States	>50K
20	Private	HS-grad	Separated	Other-service	Unmarried	Black	Female	United-States	<=50K
21	Federal-gov	9th	Married-civ-spouse	Farming-fishing	Husband	Black	Male	United-States	<=50K
22	Private	11th	Married-civ-spouse	Transport-moving	Husband	White	Male	United-States	<=50K
23	Private	HS-grad	Divorced	Tech-support	Unmarried	White	Female	United-States	<=50K
24	Local-gov	Bachelors	Married-civ-spouse	Tech-support	Husband	White	Male	United-States	>50K
25	Local-gov	HS-grad	Never-married	Craft-repair	Own-child	White	Male	United-States	<=50K

Figure1:Preprocessed adult dataset with 91 instances and 10 attributes

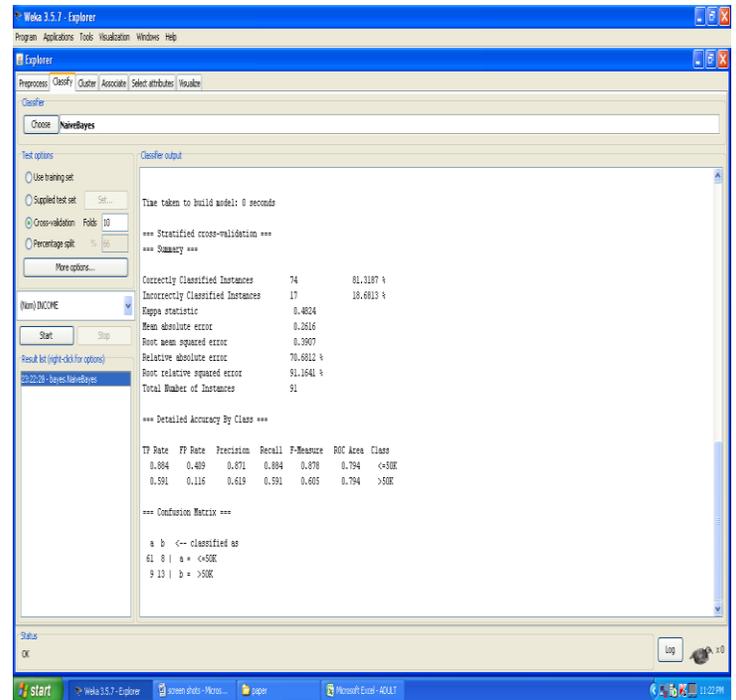


Figure2:Naïve Bayesian implemented on adult dataset

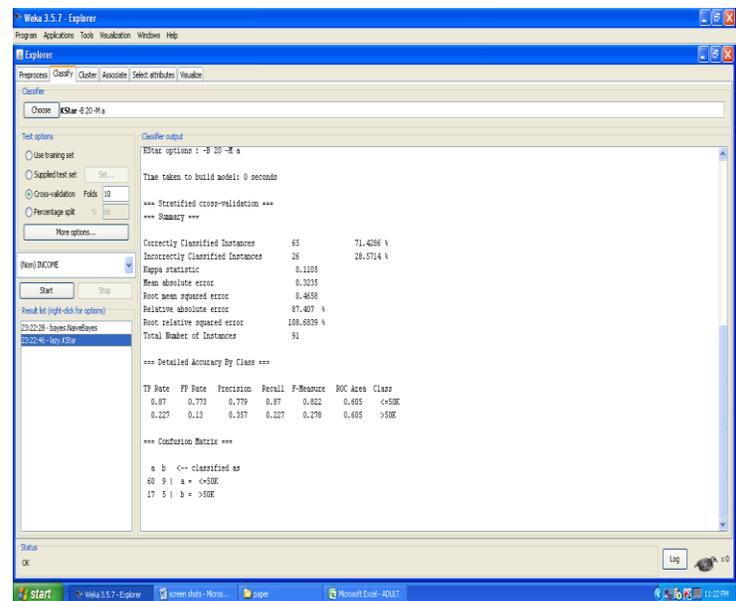


Figure3:KStar implemented on adult dataset

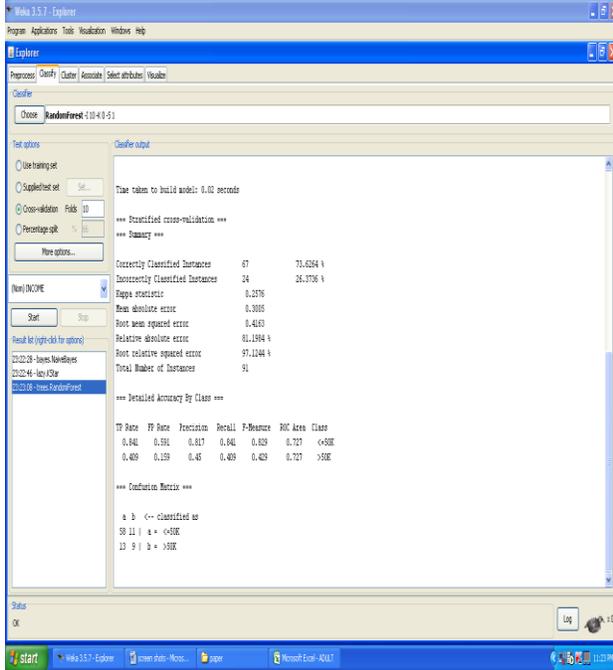


Figure2:Random Forest implemented on adult dataset

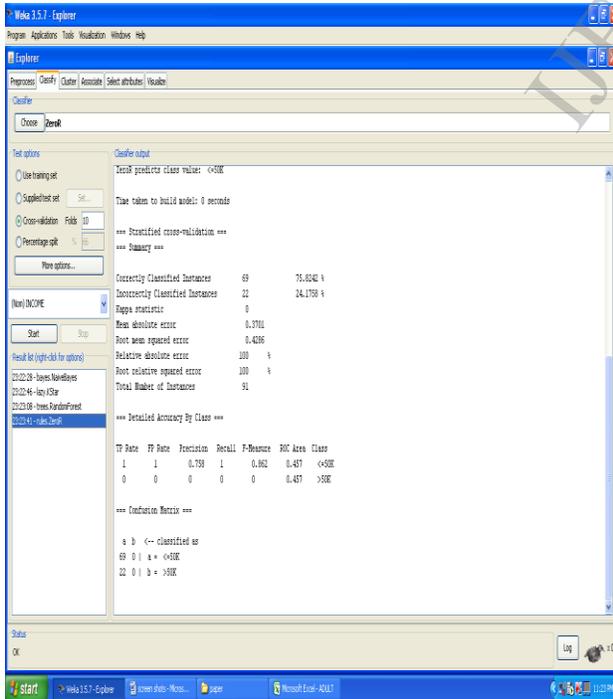
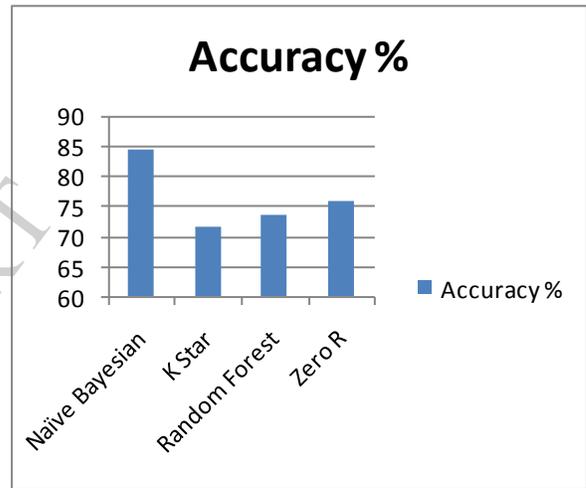


Figure2:Zero R implemented on adult dataset

Classification algorithm	Accuracy %
Naïve Bayesian	84.3187
K Star	71.4286
Random Forest	73.6264
Zero R	75.8242

Table 1:Classification accuracy of the algorithms.



3. CONCLUSION:

The above experimentation of various classification algorithm on adult data set shows that Naïve Bayesian is the best. Though Naïve Bayesian is followed by Zero R then Random Forest and KStar , these three algorithms remain in the same range. So the Naive Bayes classifier is simple and fast and they also exhibit higher accuracy rate than the algorithms discussed above.

4. FUTURE WORK

Our work can be extended to other data mining techniques like clustering, association etc. It can also be extended for other classification algorithms. We have implemented the classification technique and found the accuracy for a dataset with just 91 instances. This study can be carried forward by implementing the same algorithms on larger data sets.

REFERENCES

- [1] *Survey of Classification Techniques in Data Mining*, Thair Nu Phyu, *Proceedings of the International MultiConference of Engineers and Computer Scientists 2009 Vol I, IMECS 2009, March 18-20, 2009, Hong Kong.*
- [2] Man Leung Wong, Member, IEEE, and Kwong Sak Leung, Senior Member, IEEE *An Efficient Data Mining Method for Learning Bayesian Networks Using an Evolutionary Algorithm-Based Hybrid Approach* " , *IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION*, VOL. 8, NO. 4, pp. 378 -404, AUGUST 2004.
- [3] Mr. V. K. Pachghare Parag Kulkarni , "Pattern Based Network Security using Decision Trees and Support Vector Machine " , pp. 254 - 257 , 2011 IEEE.
- [4] Rodrigo C. Barros, Ricardo Cerri, Pablo A. Jaskowiak and Andr´ C. P. L. F. de Carvalho, "A Bottom-Up Oblique Decision Tree Induction Algorithm", pp. 450- 456 2011 IEEE 2011 11th International Conference on Intelligent Systems Design and Applications.
- [5] Wang Tongwen and Guan Lin , " A Data Mining Technique Based on Pattern Discovery and k-Nearest Neighbor Classifier for Transient Stability Assessment" 2007 RPS pp-118 - 123 *The 8th International Power Engineering Conference (IPEC 2007)*.
- [6] Nie Qianwen, Wang Youyuan , "An Augmented Naive Bayesian Power Network Fault Diagnosis Method based on Data Mining" 978-1-4244-6255-1/11/\$26.00 ©2011 IEEE
- [7] Jiabing WANG1, Pei ZHANG1 Guihua WEN1, 2, Jia WEI1, " *Classifying Categorical Data by Rule-based Neighbors* 2011 11th IEEE International Conference on Data Mining" , pp. 1248 - 1253 2011 IEEE
- [8] "Data Preprocessing for Supervised Learning" by S. Kotsiantis, D.Kanellopoulos , P.Pintelas.
- [9] "UCI Repository of Machine Learning Databases" by D.J. Newman, S. Hettich, C.L. Blake, and C.J. Merz, Available at www.ics.uci.edu/~learn/MLRepository.html, University of California, Irvine, 1998.
- [10] *Choosing the Right Data Mining Technique: Classification of Methods and Intelligent Recommendation* Karina Giberta, Miquel Sànchez-Marrèa, Víctor Codinaa, *International Environmental Modelling and Software Society (iEMSs)* , 2010 International Congress on Environmental Modelling and Software ,Modelling for Environment's Sake, Fifth Biennial Meeting, Ottawa, Canada ,David A. Swayne, Wanhong Yang, A. A. Voinov, A. Rizzoli, T. Filatova (Eds.) .
- [11] *A Generic Framework for Rule-Based Classification*, Arnaud Giacometti, Eynollah Khanjari Miyaneh Patrick Marcel, Arnaud Soulet, L.I. Universite Francois Rabelais de Tours, 41000 Blois, France.
- [12] *Lazy Classifiers Using P-trees* , William Perrizo, QinDing Anne Denton. William Perrizo, Department of Computer Science North Dakota State University Fargo, ND 58015, QinDing and Anne Denton. QinDing and Anne Denton, Department of Computer Science, Penn State Harrisburg, Middletown, PA 17057.
- [13] John G. Cleary, Leonard E. Trigg: *K*: An Instance-based Learner Using an Entropic Distance Measure*. In: *12th International Conference on Machine Learning*, 108-114, 1995.
- [14] R. Cheng, D. Kalashnikov, and S. Prabhakar, "Evaluating probabilistic queries over imprecise data," in *ACM SIGMOD International Conference on Management of Data*, 2003, pp. 551–562.
- [15] C. Chui, B. Kao, and E. Hung, "Mining frequent itemsets from uncertain data," in *Proc. of the*

Methodologies for Knowledge Discovery and Data Mining, Pacific-Asia Conference (PAKDD) 2007, pp. 47–58.

IJERT