J48 and JRIP Rules for E-Governance Data

Anil Rajput *drar1234@yahoo.com*
Principal, Bhabha Engineering Research Institute-MCA, Bhopal-26, India

Ramesh Prasad Aharwal *ramesh_ahirwal_neetu@yahoo.com*
Asstt. Prof., Department of Mathematics and Computer Science, Govt. P.G. College Bareli (M.P.), 464668, India

Meghna Dubey
Asstt. Prof., Department of Computer science, SCOP College, Bhopal (M.P.) - India.

S.P. Saxena
HOD, T.I.T. Engineering college-MCA, Bhopal, India

Manmohan Raghuvanshi
Asstt. Prof. BIST Bhopal (M.P.) India

Abstract

Data are any facts, numbers, or text that can be processed by a computer. Data Mining is an analytic process which designed to explore data usually large amounts of data. Data Mining is often considered to be "a blend of statistics. In this paper we have used two data mining techniques for discovering classification rules and generating a decision tree. These techniques are J48 and JRIP. Data mining tools WEKA is used in this paper.

Keywords: Data Mining, Jrip, J48, WEKA, Classification.

1. INTRODUCTION

Data mining is an interdisciplinary research area such as machine learning, intelligent information systems, database systems, statistics, and expert systems. Data mining has evolved into an important because of theoretical challenges and practical applications associated with the problem of extracting interesting and previously unknown knowledge from huge real-world databases. Indeed, data mining has become a new paradigm for decision making, with applications ranging from E-commerce to fraud detection, credit scoring, even auditing data before storing it in a database. The fundamental reason for data mining is that there is a lot of money hidden in the data. Without data mining all we have are opinions, we need to understand the data and translate it into useful information for decision making. According to Han and Kamber (2001), the term 'Data Mining' is a misnomer.

2. CLASSIFICATION

Classification problems aim to identify the characteristics that indicate the group to which each case belongs. This pattern can be used both to understand the existing data and to predict how New instances will behave. Data mining creates classification models by examining already Classified data (cases) and inductively finding a predictive pattern. These existing cases may come from historical database. They may come from an experiment in which a sample of the entire database is tested in the real world and the results used to create a classifier. Sometimes an expert classifies a sample of the database, and this classification is then used to create the model which will be applied to the entire database (TCC, 1999). A number of data mining algorithms have been introduced to the community that perform summarization of the data, classification of data with respect to a target attribute, deviation detection, and other forms of

data characterization and interpretation. One popular summarization and pattern extraction algorithm is the association rule algorithm, which identifies correlations between items in transactional databases.

In data mining tasks, classification and prediction is among the popular task for knowledge discovery and future plan. The classification process is known as supervised learning, where the class level or classification target is already known. There are many techniques used for classification in data mining such as Decision Tree, Bayesian, Fuzzy Logic and Support Vector Machine (SVM). In fact, there are many techniques from the decision tree family such as C4.5, NBTree, SimpleCart, REPTree, BFTree and others. The C4.5 classification algorithm is easy to understand as the derived rules have a very straightforward interpretation. Due to these reasons, this study is aimed to use this classification algorithm to handle issue on E-governance data.

2.1 Decision Tree

Decision tree can produce a model with rules that are human-readable and interpretable. Accourding to Hamidah Jantan et al 2010 (H. Jantan et a), the classification task using decision tree technique can be performed without complicated computations and the technique can be used for both continuous and categorical variables. This technique is suitable for predicting categorical outcomes (H. Jantan et a). Decision tree classifiers are quite popular techniques because the construction of tree does not require any domain expert knowledge or parameter setting, and is proper for exploratory knowledge discovery. In present, there are many research that in use decision tree techniques such as in electricity energy consumption (G. K. F. Tso and K. K. W. Yau), prediction of breast cancer (D. Delen), accident frequency (L. Y. Chang). It is stated that, the decision tree is among the powerful classification algorithms some of decision tree classifiers are C4.5, C5.0, J4.8, NBTree, SimpleCart, REPTree and others (H. Jantan et a).

2.2 Decision Tree Classifier

The C4.5 technique is one of the decision tree families that can produce both decision tree and rule-sets; and construct a tree. Besides that, C4.5 models are easy to understand as the rules that are derived from the technique have a very straightforward interpretation. J48 classifier is among the most popular and powerful decision tree classifiers. C5.0 and J48 are the improved versions of C4.5 algorithms. WEKA toolkit package has its own version known as J48. J48 is an optimized implementation of C4.5.

3. DATA SOURCES AND DESCRIPTION

Data is taken from questioners which is fillip from individuals. A questionnaire contains ten questions and demographic information. These Questionnaires have fill up from Bhopal which is a capital of Madhya Pradesh. After the initial data collection, new database was created in Ms Excel format. Ms Excel was used for preparing the dataset into a form acceptable by the selected data mining software, and Knowledge studio Weka. The database table is for E-governance data with 15 columns and 397 rows.

Question No.	Descriptions	Values
Q1	Whether are you have T. V.	True/false
Q2	Purpose of T. V.	True/false
Q3	How many mobiles are you have in your home?	True/false
Q4	Whether are you having Computer in your home?	True/false
Q5	Purpose of Computer at home	True/false
Q6	Whether you have a internet at your home	True/false
Q7	Whether you use a internet	True/false
Q8	Whether are you know about E-Governance	True/false
Q9	Whether are you know about Common Service Centre	True/false
Q10	Whether are you gain information from E-Governance	True/false

TABLE 1: Descriptions of each questions

4. EXPERIMENT

This study has three phases; the first phase is the data collection process which involved the data cleaning and data preprocessing. The second phase is to generate the classification rules using j48 classifier for the training dataset. In this case, we use all the selected attributes defined in Table 1. The J48 classifier produced the analysis of the training dataset and the classification rules. In the experiment, the third phase of experiment is the evaluation and interpretation of the classification rules using the unseen data. In the experiment we have use 66% instances of the database as a training datasets and remaining instances for tested dataset. The analyses were performed using WEKA environment. Inside the Weka system, there exist many classification algorithms which can be classified into two types; rule induction and decision-tree algorithms (N. Ulutagdemir and Ö. Daglı). Rule induction algorithms generate a model as a set of rules. The rules are in the standard form of IFTHEN rules. Meanwhile, decision-tree algorithms generate a model by constructing a decision tree where each internal node is a feature or attribute.

4.1 Data Mining Tool Selection

Data mining tool selection is normally initiated after the definition of problem to be solved and the related data mining goals. However, more appropriate tools and techniques can also be selected at the model selection and building phase. Selection of appropriate data mining tools and techniques depends on the main task of the data mining process. In this paper we have used WEKA software for extracting rules and built a decision tree. The selected software should be able to provide the required data mining functions and methodologies. The data mining software selected for this research is WEKA (to find interesting patterns in the selected dataset). The suitable data format for Weka data mining software are MS Excel and arff formats respectively. Scalability-Maximum number of columns and rows the software can efficiently handle. However, in the selected data set, the number of columns and the number of records were reduced. Weka is developed at the University of Waikato in New Zealand. “Weka” stands for the Waikato Environment of Knowledge Analysis. The system is written in Java, an object-oriented programming language that is widely available for all major computer platforms, and Weka has been tested under Linux, Windows, and Macintosh operating systems. Java allows us to provide a uniform interface to many different learning algorithms, along with methods for pre and post processing and for evaluating the result of learning schemes on any given dataset. Weka expects the data to be fed into to be in ARFF format (WEKA website).

4.2 Screen Shots Which is Generated During Experiment

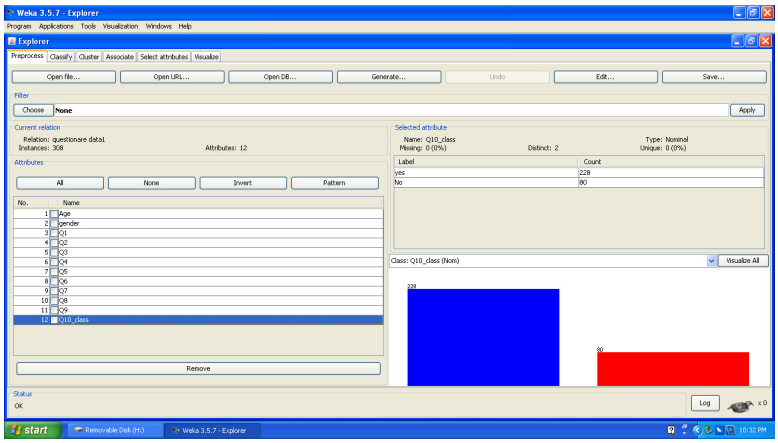

FIGURE: 1 WEKA explorer

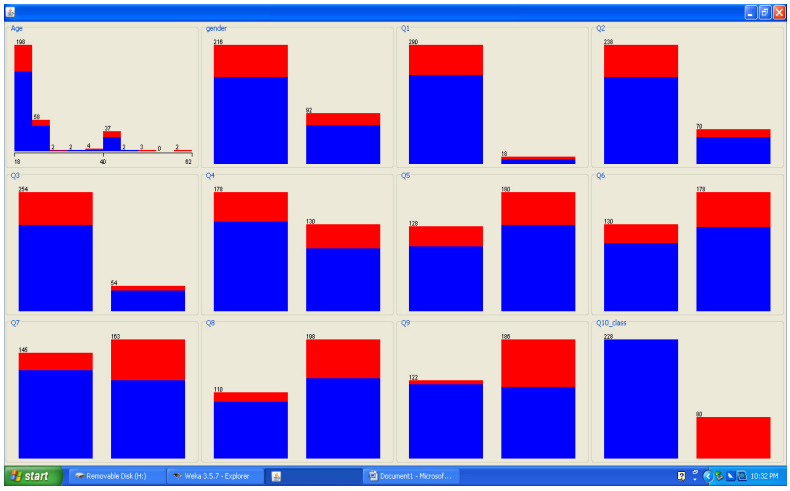

FIGURE. 2: Visualization of each attributes of experimental data

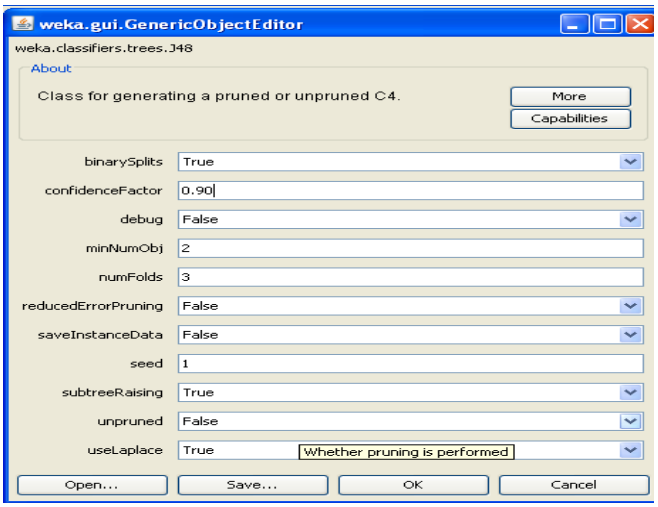

FIGURE 3: parameter setting

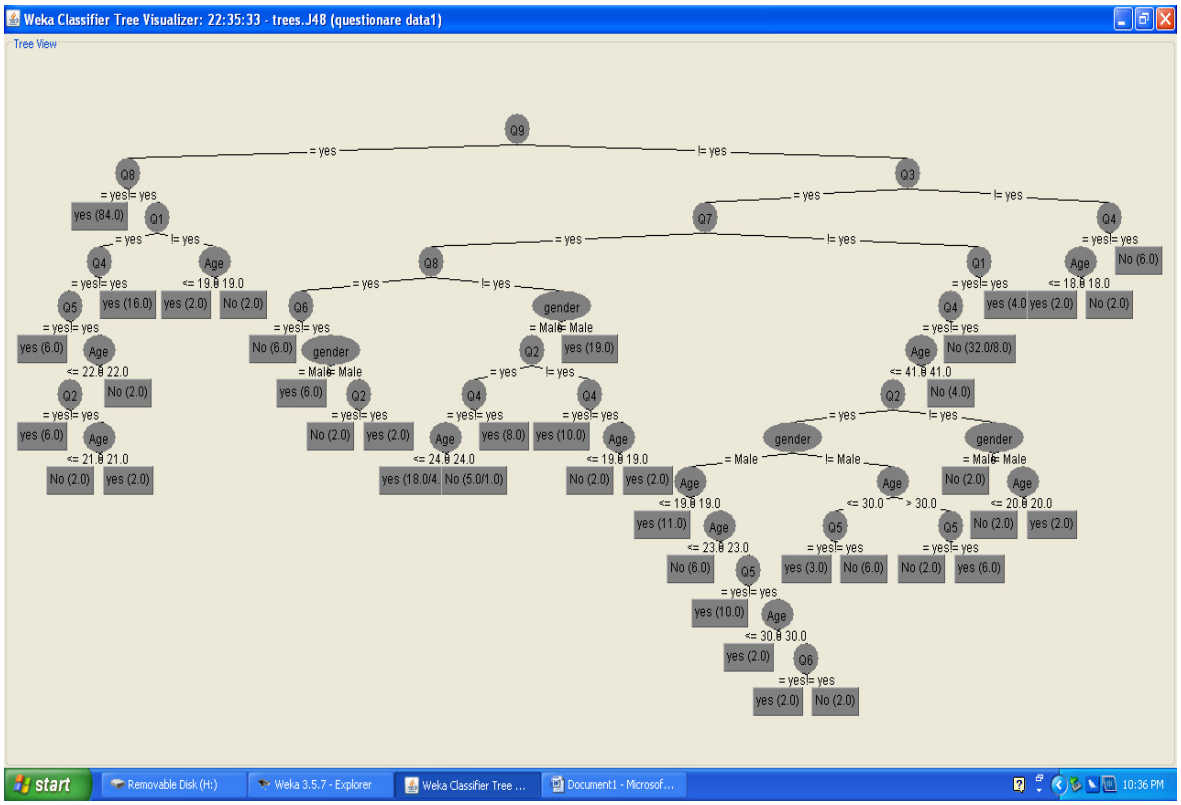

FIGURE 4: Decision tree generated from WEKA

Experimental Result

=== Run information ===

Scheme: weka.classifiers.trees.J48 -C 0.9 -B -M 2 -A
Relation: questionare data1
Instances: 308
Attributes: 12
Age, gender, Q1, Q2, Q3, Q4, Q5, Q6, Q7, Q8, Q9, Q10_class
Test mode: split 66% train, remainder test
=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances	90	85.7143 %
Incorrectly Classified Instances	15	14.2857 %

=== Confusion Matrix ===

a b <-- classified as

68 8 | a = yes

7 22 | b = No

=== Predictions on test split ===

5. JRIP RULES CLASSIFIERS

JRip (RIPPER) is one of the basic and most popular algorithms. Classes are examined in increasing size and an initial set of rules for the class is generated using incremental reduced error JRip (RIPPER) proceeds by treating all the examples of a particular judgment in the training data as a class, and finding a set of rules that cover all the members of that class. Thereafter it proceeds to the next class and does the same, repeating this until all classes have been covered.

=== Run information ===
Scheme: weka.classifiers.rules.JRip -F 5 -N 2.0 -O 2 -S 1 -D
Relation: questionnaire data1
Instances: 308
Attributes: 12
Age,gender, Q1, Q2, Q3, Q4, Q5, Q6,Q7,Q8, Q9, Q10_class
Test mode: split 66% train, remainder test
JRIP rules:

==========

(Q9 = No) and (Q7 = No) and (Q4 = No) => Q10_class=No (42.0/12.0)

(Q9 = No) and (Q8 = yes) and (gender = Female) => Q10_class=No (10.0/2.0)

(Q9 = No) and (Q7 = No) and (Age >= 20) and (Age <= 22) => Q10_class=No (10.0/0.0)

(Age >= 30) and (Age >= 49) => Q10_class=No (5.0/0.0)

(Q8 = No) and (Q3 = No) and (Age >= 19) => Q10_class=No (4.0/0.0)
=> Q10_class=yes (237.0/23.0)

5.1 Interpretations of Rules

In this section we try to interpret Jrip rule.

Rule first interpreted as If a person do not know about common service center and do not use a computer and also he does not have computer at his home then he do not gain information from E-governance.

Rule second interpreted as If a person does not know about common service center and he know about E-governance and he is a female then he does not gain information from E-governance.

Third rule is interpreted as If a person do not know about common service center and he does not use a computer and age is above 19 and below 23 then he do not gain information from E-governance.

Similarly last rule is interpreted as, if a person do not know about E-governance and he have not mobile at home and age is greater and equal 19 years then he gain information from E-governance

Same way other rules can be interpreted as above

6. CONCLUSION

This paper focuses on the use of decision tree and JRIP classifiers for E-governance data. Decision tree classifier generates the decision tree. From generated decision tree useful and

meaningful rules can be extracted. JRIP classifier generates some useful rules which are interpreted in above section.

7. REFERENCES

[1] L. Y. Chang and W. C. Chen, *"Data mining of tree-based models to analyze freeway accident frequency*," *Journal of Safety Research,*36(1): 365-375, 2005.

[2] D. Delen, G. Walker, and A. Kadam, "*Predicting breast cancer survivability: A comparison of three data mining methods," Artificial Intelligent in Medicine,* 34:113- 127, 2005.

[3] H. Jantan et al. *"Classification for Prediction",* International Journal on Computer Science and Engineering, 2(8): 2526-2534, 2010.

[4] J. Han and M. Kamber, "*Data Mining: Concept and Techniques*". Morgan Kaufmann (2006).

[5] G. K. F. Tso and K. K. W. Yau, "*Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks*," *Energy,* 32 : 1761-1768, 2007.

[6] N. Ulutagdemir and Ö. Daglı, "*Evaluation of risk of death in hepatitis by rule induction algorithms*", Scientific Research and Essays, 5(20): 3059-3062, 2010, ISSN 1992-2248.

[7] Weka website: Data Mining Software in Java, http://www.cs.waikato.ac.nz/ml/weka/