International Journal of Advance Engineering and Research Development (IJAERD)
Volume 1,Issue 6,June 2014, e-ISSN: 2348 - 4470 , print-ISSN:2348-6406

The Empirical comparison of decision tree approaches with classification task for finding graduate employment status

Bangsuk Jantawan[1], Cheng-Fa Tsai[2]

[1]*Department of Tropical Agriculture and International Cooperation, National Pingtung University of Science and Technology Pingtung, jantawan4@hotmail.co.th*
[2]*Department of Management Information Systems, National Pingtung University of Science and Technology Pingtung, cftsai2000@yahoo.com.tw*

Abstract—Classification task is one of the most useful techniques in data mining, which is to find out meaningful and useful pattern in large volumes of data. For classification task, Decision tree approaches are the most commonly used because of its facility of implementation and easy to understand. The study purpose various approaches available for classification like J48, Simple Cart, Random Forest, LAD Tree, REP Tree, Decision Stump etc. The research we introduce eight algorithms from them. We take graduates dataset in academic year of 2011 Maejo University Thailand. Implement theses algorithms in graduate dataset and compare TP-rate, Fp-rate, Precision, Recall and ROC Curve parameter.

Keywords-classification task; data mining technique; decision tree; graduate employment; precision

I. INTRODUCTION

Data mining is the procedure of analyzing data based on various perspectives and summarizing it into valuable information. Data mining technique which conducts the assigning of objects into related classes (output attribute) are called classifiers. [1]. Classification task is one of the most useful techniques in data mining, which is to find out meaningful and useful pattern in large volumes o f data. The main phased of classification task are divided into two parts as follows: the first phase is used to find a model for the class attribute as a function of other variables of the datasets, and the second phase is applied previously designed model on unseen datasets for defining the involved class of each record [1]. There are various methods for the classification task such as Decision Trees, Support Vector Machine, Naïve Bayes, Linear Regression, Logistic Regression etc.

Decision tree approaches are the most commonly used because of its facility of implementation and easy to understand [2]. It is a flow-chart-like tree structure, in these tree structures, leaves node is represented by class labels and the branches nodes are represented by conjunction of features that lead to those class labels.

The advantage of decision tree approach not only easy to understand and facility of implementation but also from the following [5]:

- Decision trees are easily converted to a suite of production rules;
- They also can classify both numerical and categorical data, but the output attribute must be categorical;
- There are no a priori assumptions about the nature of the data.

In this paper we describe and apply difference decision tree approaches for predicting graduate employability status from graduate historical database. Moreover, we also compared TP-rate, Fp-rat, Precision, Recall and ROC Curve parameter of each decision tree approach. The comparison of various decision trees approaches (BF Tree, Decision Stump, FT, J48, J48graft, LAD Tree, LMT, NB Tree, Random Forest, Random Tree, REP Tee, Simple Cart) we used "graduate.arff" dataset the basic information of graduate historical in higher education for academic year 2011, Maejo University Thailand. The outcome (Class lebels) of the decision tree predicted the number of graduate who are only employed (JOB), employed and studies (JOBandStu), unemployed (NOJOB), and only studied (OnlyStdu).

The rest of the research is organized as follows: in section 2, preliminary work or previous works involved with this area and the motivations have been presented. Section 3 describes about dataset for decision tree approach. The procedure of this experiment is presented in section 4. Section 5 evaluates the results and assesses the efficacy of the classifiers. Finally, section 6 concludes the research and describes future works.

II. PRELIMINARY

The goal of this research is to select the best approaches for graduate employment dataset which can be integrated in our WEKA tool, we have to search among those that can support categorical data, handle with duplicate data and incomplete data offer a natural interpretation to instructors and be accurate working with samples. Therefore, we analysis six of the most common tree approaches techniques, namely BF Tree, Decision Stump, FT, J48, J48graft, LAD Tree, LMT, NB Tree, Random Forest, Random Tree, REPT, and Simple Cart.

2.1. Decision tree

Decision tree are used to classify an instance to a predetermined set of class labels (graduate status) based on their attributes. It is a flowchart such as tree structure [3]. The decision tree is consisting of three components, which are leaf node, internal node, and root node. Top of tree denotes the root node. Leaf node of tree denotes the terminal element of the structure and the nodes in between is called the internal node. Each internal node is represented test on an attribute, each branch represents an outcome of the test, and each leaf node holds a class label (graduate status) [3]. The approach is created based on "Divide and Conquer" [4]. That is the tree is constructed by framing rules which will branch out from the nodes and sub-nodes until the decision is made. There are different approaches of forming the decision rules for Decision Trees. Nodes of tree are chosen from the top level based on quality attributes such as Gain Ratio, Gini Index, and Information Gain and so on.

The following gives the short introduction of six decision tree approaches.

- The C4.5 or J48: this algorithm uses Gain Ratio to build the tree, the component with highest gain ratio is taken as the root node and the dataset is split based on the root component values. Again the information gain is calculated for all the sub-nodes individually and the step is duplicated until the prediction is completed [3].
- REP Tree: this algorithm is a fast decision tree learner which constructs a decision or regression tree using information gain as the splitting criteria, and prunes it using receded error pruning. It only arranges values for numeric attributes once. Missing value carry out with using C4.5's method of using faction instances [5].
- Decision Stump: a decision stump is basically a one level decision tree where the split at the root level is based on a specific attribute [5].

- Random Forest: this algorithm is a set of regression trees or unpruned classification, occurred from bootstrap samples of the training data, using random attribute selection in the tree induction procedure. Prediction is made by aggregating (most of vote for classification or averaging for regression) the predictions of the ensemble [5].
- NB Tree: this algorithm is combine both of naive Bayesian classification and decision tree learning. In NB Tree, a local naive Bayes is applied on each leaf of a traditional decision tree, and an instance is classified, using the local naive Bayes on the leaf into which it falls. The algorithm for learning an NB Tree is analogous to to C4.5. After a tree is grown, a naive Bayes is created for each leaf using the data involved with that leaf. An NB Tree classifies an example by sorting it to a leaf and applying the naive Bayes in that leaf to assign a class label to it. NB Tree frequently achieves higher accuracy than either a naive Bayesian classifier or a decision tree learner [5].
- Random Tree: this algorithm is a tree drawn at random from a suite of possible trees. In this context "at random" denotes that each tree in the suite of trees has an equal chance of being sampled [5].

2.2. WEKA Tool

The Waikato Environment for Knowledge Analysis (WEKA) come about through the perceived need for a unified workbench that would allow researcher easy access to state-of-the-art techniques in machine learning. At the time of the project's inception in 1992, learning algorithms were available in various languages, for use on different platforms, and operated on a variety of data formats. The task of collection together learning schemes of a comparative study on a collection of data sets was daunting at best. It was envisioned that WEKA would not only provide a toolbox of learning algorithms, but also a framework inside which researchers could implement new algorithms without having to be concerned with supporting infrastructure for data manipulation and scheme evaluation [6].

The data file normally use by WEKA is in ARFF file format, which consists of special tags to indicate different things in the data file. The main interface in WEKA is explorer, It has a set of panel, each of which can be used to perform a certain task. Once a dataset has been loaded, one of the other panels in the Explorer can be used to perform further analysis [1].

III. DATASET

In our experiment for comparing all decision tree approaches we used "graduate.arff" dataset obtained from historical database on academic year of 2011 at Maejo University, Thailand. Involved Information:

Table 1.

Variable	Values
Gender (GEND)	Male (M), Female (F)
Domicile (Domicile)	76 provinces in Thailand such as Suratthani province (Suratthani), Chiang Mai province (ChingMai) and etc.
Degree (DEG)	Doctorate (Doctoral), Master degree (Master), Bachelor degree (Bachelor)
Work province (WorkPro)	76 provinces in Thailand such as Suratthani province (Suratthani), Chiang Mai province (ChingMai) and etc.
Educational background (ED)	Bachelor of Science (BSc), Bachelor of Landscape Architecture

	(BLA), Bachelor of Engineering (BEng), BA Economics (BEcon), Bachelor of Business Administration (BBA), Doctor of Philosophy (PhD), Bachelor of Arts (BA),Bachelor of Agricultural Technology (BSPlantScience), Bachelor of Accountancy (BAcc), Bachelor of Political Science (BPolSc), Master of Science (MSc), Master of Business Administration (MBA), Master of Engineering (MEng), Doctor of Arts (PhDArts), Master of Arts (MArts), Bachelor of Technology (BTech)
Faculty (ISCED)	Faculty of Agricultural Production (FaOAgPr), Administration (FaOAd), Faculty of Science (FaOSc), Faculty of Engineering and Agro-Industrial (FEnA), High School- Phrae Honor (HSPhr),High School – Chumphon (HScChu), College of Management Sciences (CoMaSc), The School of Renewable Energy (ScReEn), Faculty of Tourism Development (FOToDe), Faculty of Liberal Arts (FOLiBr), Faculty of Economics (FaOEc), Faculty of Fisheries an Aquatic Resources (FOFiAq), Faculty of Information and Communication (FaOInACo), Faculty of Architecture and Environmental Design (FaOArAEnDe), Faculty of Animal Science and Technology (FaOAnScATe)
Grade Point Average (GPA)	Numerical data (1.00, 2.50, …)
Talent (TAL)	Foreign languages (ForeL), Computer (Comp), The recreational activities (TReAc), Arts (ART), Sports (Spor), The Performing Arts / Music chorus (PAMC), Have Not (NO), Other (OT)
Position (JOB)	Officials / authorities of the state/State enterprise (OASE), Companies / organizations/ private businesses(COPB), Independent Business / Owner (IBO), Employees bodies / international (EBI), Other (OT), Not specified(No)
Length of time for finding job (TFJ)	Unknown (Nu), Get a job immediately after graduation (IMME), 1-3 months (M13), 4-6 months (M46), 7-9 months (M79), 10-12 months(M1012), Over 1 year (Over1Y), Work before and during the study (BNS)
The matching between field of graduate and job (MAF)	Match (Yes), Not Match (No), Unknown(Un)
Required for studying (ReSt)	Demand (Yes), Needless (No)

Note: Class: Employed, Unemployed, Employed and study, Only Study

IV. IMPLEMENTATION

First of all open WEKA Tool and select dataset in our experiment we used graduate dataset and it's in built of WEKA as figure 1.

In Figure 1, it shows the basic information of dataset such as the name of attribute, attribute type, number of instance, number of attribute, and various values. We can also change various values on this window.

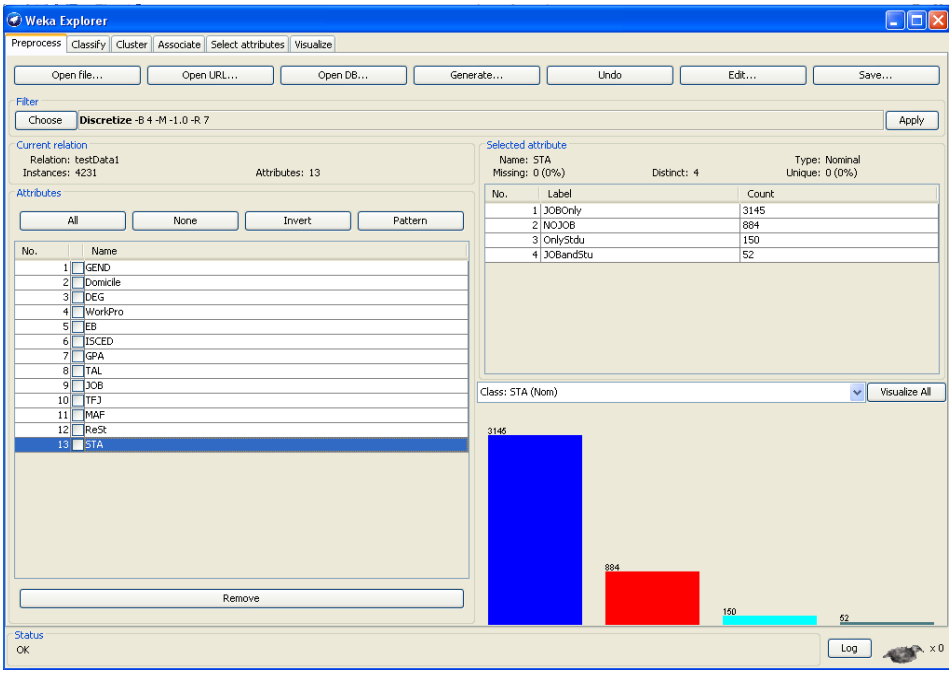

Figure 1. the preprocess window

In Figure 2, it provides the value of graduate dataset of every attribute. After preprocessing we directly perform for classification select classifier based on classification approach and click on start button. Finally, Figure 3 shows output regarding all parameter like error rate and confusion matrix. In this paper we implemented eight algorithms under Decision tree approach.

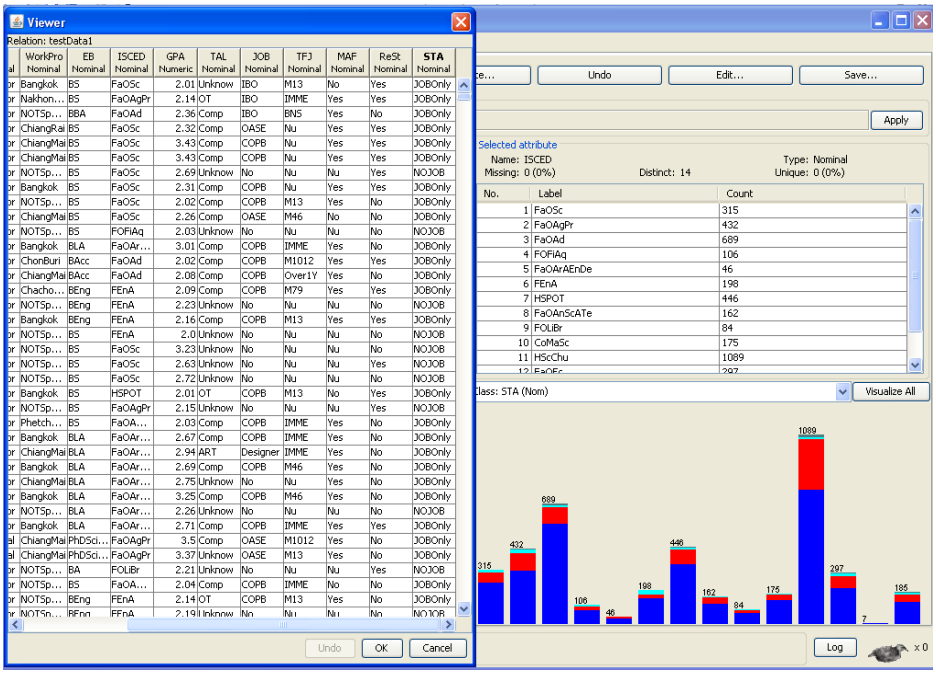

Figure 2. the window of editing attributes

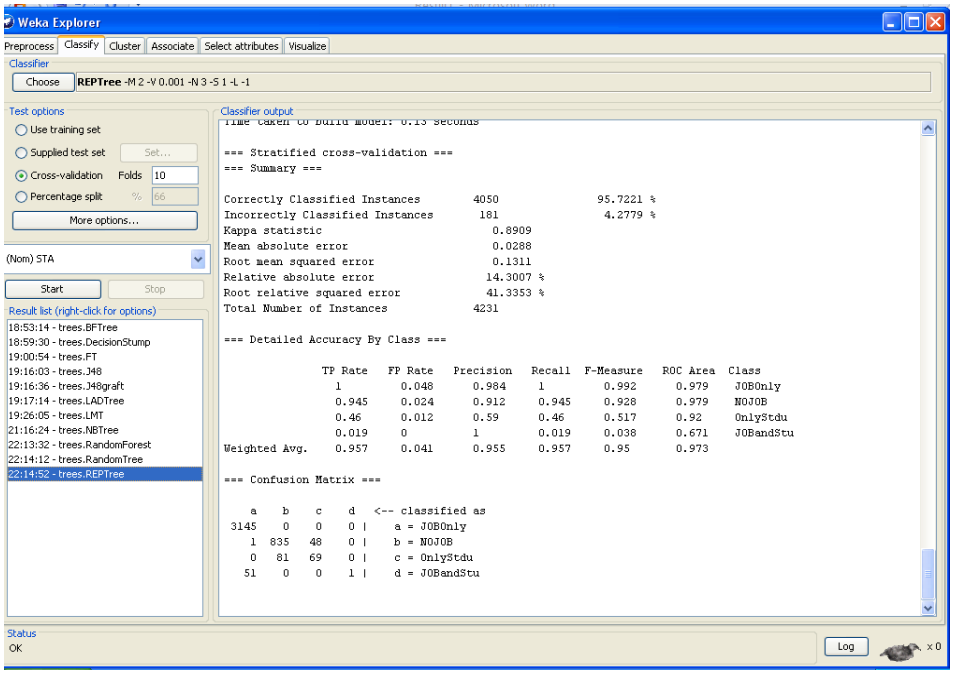

Figure 3. the result of REP Tree algorithm

V. EXPERIMENTAL RESULT

We conducted eight algorithms using different parameter setting and different number of folds for cross validation, in order to discover whether they have a great affection the result Finally, we set the algorithms with default parameters and used 10 fold cross validation w show the classification accuracy and rates obtained with the four algorithms for the graduate dataset.

Table 2 Comparison of all six algorithms

Algorithms	TP-rate	FP-Rate	Precision	Recall	F-Measure	ROC Area
Decision Stump	0.952	0.046	0.910	0.952	0.930	0.968
J48	0.954	0.042	0.941	0.954	0.947	0.974
NB Tree	0.955	0.041	0.944	0.955	0.949	0.984
Random Forest	0.958	0.041	0.949	0.958	0.950	0.981
Random Tree	0.942	0.045	0.938	0.942	0.940	0.955
REP Tree	0.957	0.041	0.955	0.957	0.950	0.973

Table 3 Comparison of all six algorithms under the correct, incorrect classified, and time for analyzing

Algorithms	Correctly classified instances	Incorrectly classified instances	Time for analyzing (second)
Decision Stump	95.202%	4.798%	0.02
FT	94.564%	5.436%	77.22
J48	95.415%	4.585%	0.08
Random Forest	95.793%	4.207%	0.25
Random Tree	94.162%	5.838%	0.02
REP Tree	95.722%	4.278%	0.13

VI. CONCLUSION

In this research, we compare six algorithms on graduate dataset with some parameter. In graduate dataset have a simple and class attribute. FT tree will implement on graduate dataset then it is less efficient than all other algorithms. In algorithm of J48 is more accurate and efficient all parameters like TP-rate, FP-rate, Precision, Recall and ROC area.

REFERENCES

[1] P. Kanu, V. Jay, and P. Jaymit, "Comparison of various classification algorithms on iris datasets using WEKA", International journal of Advance Engineering and Research Development vol. 1, pp. 1-7, February 2014.

[2] R. Anju, and M. R. prakash, "Survey on Decision Tree Classification algorithms for the Evaluation of Student Performance, International journal of computer & technology, vol. 4, pp.244-247, March-April, 2013.

[3] P. Nikita, and U. Saurabh, "Study of various decision tree pruning methods with their empirical comparison in WEKA", International journal of computer applications, vol. 60, pp.20-25, December, 2012.

[4] S. Despa, "What is data mining" Cscu.cornell.edu, 2002 [Online] December 2002 Retrieved from http://www.cscu.cornell.edu/news/statnews/stnews55.pdf [Accessed on June 14, 2014].

[5] Z. Yongheng, and Z. Yanxia, "Comparison of decision tree methods for finding active objects", Advances of Space Research, v.1, pp.1-10, August, 2007.

[6] H. Mark, F. Eibe, H. Geoffrey, and P. Bernhard, "The WEKA Data mining Software: An Update", SIGKDD Explorations, vol. 11, pp.10-18, July, 2009.

