Comparison of Classification Techniques in Education Marketing

Sheila A. Abaya, Bobby D. Gerardo, and BartolomeT. Tanguilig III

***Abstract*— One of the predicaments of Higher Educational Institutions (HEIs) is to identify the potential schools for enrollment. Most HEIs conduct School-To-School-Promotion (STSP) to several secondary schools to sustain, if not, increases the enrollment rate. The classification techniques in data mining were used to classify feasible secondary institutions as target markets for promotion. This technique may also eliminate, if not, alleviate the expenses of HEIs by filtering which among the visited secondary schools do not contribute to the enrollment rate. Experimentation on J48, and Bayesian Network classification techniques were implemented using WEKA 3.6.0 [2] [4]. These techniques were also identified based on the accuracy of classifying the data set. C4.5 outperformed other classifying technique. The output of this research is beneficial in identifying the best classifying technique to be used in the given data set of determining which among the probable secondary schools are qualified for enrollment in the HEI.**

***Index Terms*— Bayesian Network, C4.5, classification technique, WEKA**

I. Introduction

CLASSIFICATION technique is one of the data analyses in data mining where it can be used to create models in determining the target market [1]. This technique identifies the probability of the schools to produce potential enrollees. Several classification methods were implemented and certain technique outperformed the others. This research aims to identify which among the following techniques: J48 (C4.5) and Bayesian Network works best as a classifier in the given training set of students who enrolled in the higher educational institution. Moreover, it also establishes the preciseness of the aforementioned techniques in terms of classifying instances whether the school provides enrollees or not and to determine which classifier is more accurate. The training set was used to check the correctness of classifier. Pruning was also implemented using the test set to avoid over fitting.

Manuscript submitted January 02, 2014; revised January 27, 2014. This work was supported and financed by the University of the East, Manila, Philippines.

S. A. Abaya is a doctoral student under the program of Doctor in Information Technology of the Technological Institute of the Philippines and currently a faculty of the Department of Computer Studies and Systems of the University of the East, Caloocan Philippines, phone: +63-9063025931; (e-mail: sheila_abaya@yahoo.com.ph).

B. D. Gerardo was with West Visayas State University in Iloilo City, Philippines. He is now the Vice President for Admin and Finance of the same institution (e-mail: bgerardo@wsu.edu.ph).

B. T. Tanguilig is currently the Dean of the Graduate School of the Technological Institute of the Philippines, Quezon City Philippines (e-mail: bttanguilig_3@yahoo.com).

A. Related Studies

Several studies have been conducted to compare different classification techniques.

Sharma, *et al* [3] worked on the comparative analysis of J48, ID3, ADTree, and SimpleCART classification techniques for spam emails. The research focused on the data analysis of email to identify whether the message is a spam email or not. The experiment was done using WEKA by WEKA Machine Learning Project of the University of Waikato in New Zealand. There were 4,601 instances with 1,831 spam category and 58 attributes from which 57 are continuous and 1 is nominal. The result of the experiment proved that J48 (C4.5) has the highest classification accuracy of 92.7624% where 4,268 instances were classified correctly and 333 instances were classified otherwise.

Grossman, *et al* [10] labored on the comparison of Bayesian Network Classifier (BNC) with other algorithms of classification such as C4.5, Naïve Bayes (NB), Tree-Augmented Naïve Bayes (TAN) by Friedman *et al* (1997), original Bayesian network structure search algorithm (HGC) by Heckerman *et al* (1995), Maximum Likelihood Learners using the MDL score (ML-MDL) and two-parent nodes (ML-2P) and NB-ELR and TAN-ELR, NB and TAN with parameters optimized for conditional log likelihood of Greiner and Zhou (2002). Based on the result, BNC can be learned by maximizing conditional likelihood and thus provide a better classification probability among the other methods.

II. METHODOLOGY

A. Preparation of Data

To identify the classification accuracy of these techniques, a training set was provided and cleaned by removing invalid data and supplying them with missing value to make sure that it provides a reliable result. The training set is actually the historical data of students who took the entrance exam and chose to enroll in a particular institution. The data considered five attributes: General Weighted Average (GWA) of secondary school students; Radial Distance, the proximity of the secondary school from the tertiary institution; School Ownership, the type of school whether publicly or privately owned; the Income Bracket, the salary range of the parents of a particular student; and the Class, it identifies whether the student enrolled or not in the organization. These criteria were categorized based on the possible value presented in Table 1. The data were stored in

ISBN: 978-988-19252-5-1
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

Excel and saved as Comma Separated Value (CSV) format. The CSV file was imported in Notepad and was converted into an Attribute Relation File Format (ARFF) file. An ARFF file has three components: @relation which gives the name of the data set, @attribute which identifies the elements of the tables with the corresponding value and @data which lists all the records as shown in Fig.1. There were 1,970 instances in the training set.

TABLE I
ATTRIBUTE VALUES FOR RELATION ENROLL

Attribute	Category	Lowerbound - Upperbound
GWA	GWA1	95 – 100
	GWA2	90 – 94
	GWA3	85 – 89
	GWA4	80 – 84
	GWA5	75 - 79
RadialDistance	DistanceA	0 – 9 km
	DistanceB	10 – 20 km
	DistanceC	21 – 99999 km
SchoolOwnership	Private	Private
	Public	Public
IncomeBracket	IncomeA	51000 – 10000000
	IncomeB	21000 – 50999
	IncomeC	10000 – 20999
Final	Enrolled	Enrolled
	DidNotEnroll	DidNotEnroll

```
@relation enroll

@attribute schoolownership{PRIVATE,PUBLIC}
@attribute Radialdistance
{DistanceA,DistanceB,DistanceC}
@attribute IncomeBracket
{IncomeA,IncomeB,IncomeC}
@attribute GWA{GWA1,GWA2,GWA3,GWA4,GWA5}
@attribute final{Enrolled,DidNotEnroll}

@data
PRIVATE,DistanceA,IncomeB,GWA3,DidNotEnroll
PRIVATE,DistanceA,IncomeB,GWA3,DidNotEnroll
PRIVATE,DistanceC,IncomeB,GWA2,DidNotEnroll
PRIVATE,DistanceA,IncomeB,GWA2,Enrolled
PRIVATE,DistanceB,IncomeB,GWA3,Enrolled
PRIVATE,DistanceA,IncomeC,GWA5,DidNotEnroll
PRIVATE,DistanceB,IncomeB,GWA2,DidNotEnroll
PRIVATE,DistanceA,IncomeA,GWA3,DidNotEnroll
PRIVATE,DistanceB,IncomeB,GWA3,DidNotEnroll
PRIVATE,DistanceC,IncomeB,GWA2,Enrolled
PRIVATE,DistanceC,IncomeB,GWA3,DidNotEnroll
PRIVATE,DistanceC,IncomeB,GWA2,DidNotEnroll
PRIVATE,DistanceB,IncomeB,GWA2,DidNotEnroll
PUBLIC,DistanceB,IncomeB,GWA4,Enrolled
PRIVATE,DistanceC,IncomeB,GWA4,Enrolled
PUBLIC,DistanceA,IncomeB,GWA4,DidNotEnroll
PRIVATE,DistanceA,IncomeB,GWA3,DidNotEnroll
PRIVATE,DistanceB,IncomeB,GWA3,DidNotEnroll
PRIVATE,DistanceB,IncomeC,GWA4,Enrolled
PRIVATE,DistanceC,IncomeB,GWA4,Enrolled
PRIVATE,DistanceB,IncomeB,GWA4,DidNotEnroll
PRIVATE,DistanceB,IncomeB,GWA3,DidNotEnroll
PRIVATE,DistanceC,IncomeB,GWA3,Enrolled
PUBLIC,DistanceC,IncomeB,GWA1,Enrolled
PRIVATE,DistanceB,IncomeB,GWA2,Enrolled
PRIVATE,DistanceB,IncomeB,GWA2,Enrolled
```

Fig. 1. An Excerpt of the Enroll1 ARFF File. Generating this ARFF file with the name of the relation, relevant attributes and the extracted instances from the HistoricalData will trigger the techniques to produce the classified data from the training set.

B. Applying WEKA

Launching the application of WEKA which is available in http://www.cs.waikato.ac.nz/ml/weka will show the opening screen as shown on Fig. 2.

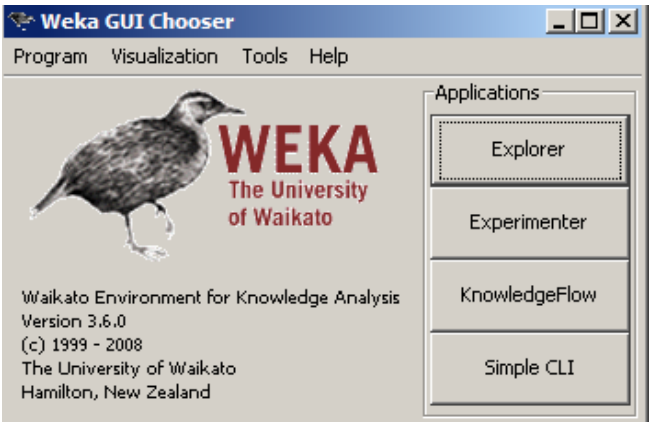

Fig. 2. Initial Interface of WEKA. The window displays the possible processes that can be done with WEKA.

The application has four (4) options: 1.) Explorer, it allows to preprocess the file, select attributes, and visualize the training data set; 2.) Experimenter, it compares different machine learning algorithms and identifies which method works best in a particular problem set; 3.) Knowledge Flow, it provides a visual representation of the Knowledge Discovery and Data Mining (KDD) process; and 4.) Simple Command Line Interface (SimpleCLI), is a window for typing commands. Most users of WEKA preferred to launch Explorer first to preprocess the data.

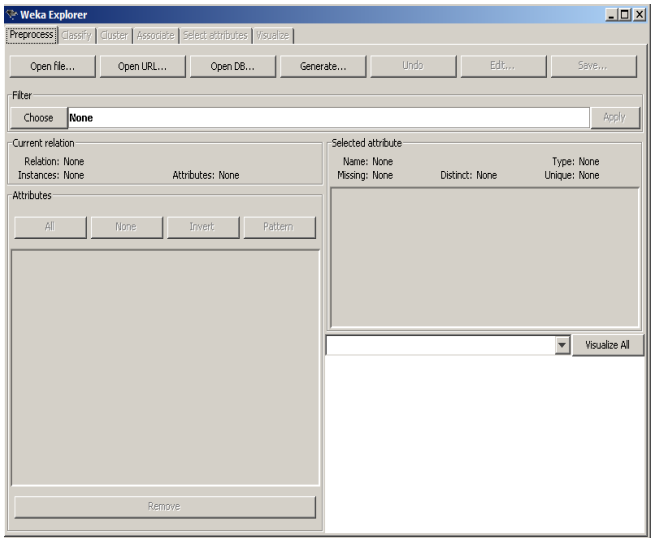

Fig. 3. Explorer Interface of WEKA. It allows selecting a training data to be used in classifying.

Initially, the window does not contain any information or data on the attribute, selected attribute, and visualize panes since there is/are no file/s selected yet. The preprocess tab enables to open the data as the training set. Clicking the Open file button names the ARFF file as "Enroll1.arff."

ISBN: 978-988-19252-5-1
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

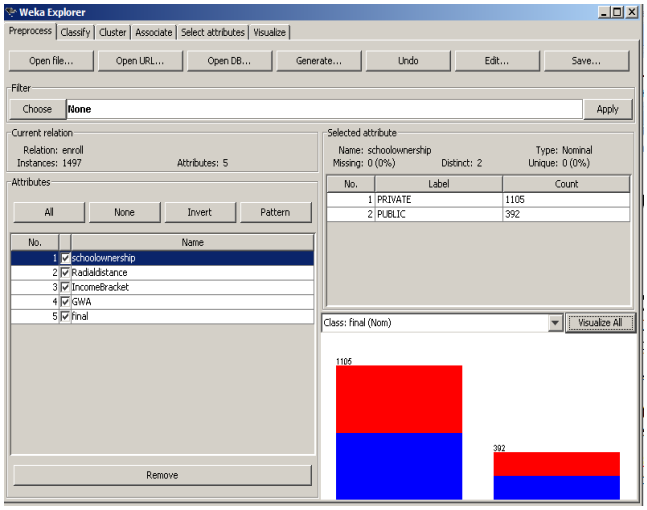

Fig. 4. Enroll1.ARFF Information. It displays the name of the relation, the number of instances, the attributes available in the relation, the values for the chosen attribute and the visualize form of the chosen attribute.

In Fig. 4, the attributes pane displays the available variables in the data set. Enroll1.arff has five (5) attributes: School Ownership, Radial Distance, Income Bracket, GWA, and Final. Upon checking the attribute, the information is displayed on the selected attribute pane, which identifies the element, type, and the missing and corresponding value. The Visualize pane displays the histogram of the selected attribute, which happens to be the School ownership attribute with values Private and Public. The histogram of all attributes can be displayed by clicking the Visualize All button as shown in Fig. 5. Every bar in the histogram represents the values of corresponding attribute.

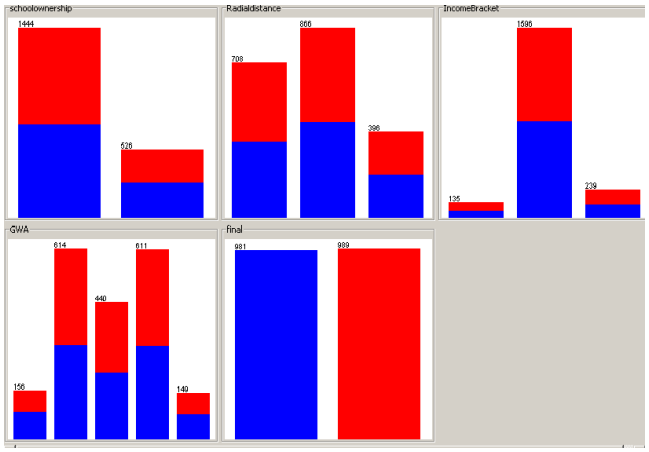

Fig. 5. Histogram Window. This is the visual representation of all the attributes selected.

Choosing the **Classify tab** as seen in Fig. 6 opens the option of selecting the classifiers J48 (C4.5) and Bayesian Net. For the Test Options, Use Training Set was chosen as "Enroll1.arff" to create the model with 1,970 instances. Once the initial model is created, it validates the accuracy result of the model. Furthermore, another test was done using the option Supplied Test Set named "Enroll1a.arff" with 27 instances. Records were not included in "Enroll1.arff". Clicking the Start button determines the functionality of the model.

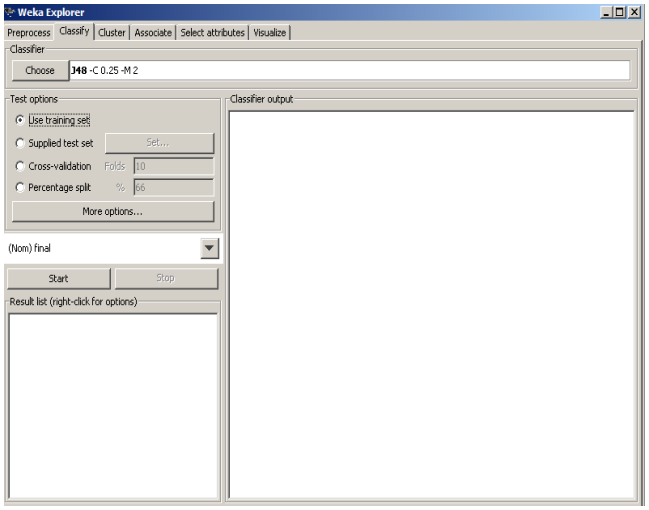

Fig. 6. Classify Tab Interface. This option opens the gate of choosing the classification technique to be used in the training set described in the preprocess data.

III. CLASSIFICATION TECHNIQUES

This section gives an overview of the different techniques used in this paper.

A. *J48/C4.5*

A decision tree learner that implements C4.5 and the successor of ID3 works best in dealing with numeric attributes, missing data, noisy data, and generating rules from the tree. The algorithm works in heuristic based reasoning where the candidate cuts off a smallest number of instances on the numeric attributes. Based on the heuristic observation of Quinlan (1986), if there is an S candidate on a certain numeric attribute at the node, it is considered splitting $\log_2(S)/N$ is subtracted from the information gain where N is the number of instances at the node which prevents over fitting.

B. *Bayesian Network/Bayes Net*

Bayesian Networks (BNs) is known as belief networks where the model is believed to be true but with some uncertainties and it is a graphical model of subjective probability [7]. "The probability of a model M after observing data D is proportional to the likelihood of the data D assuming that M is true, times the prior probability of M. (Bayes)". This structure works on probability theory, graph theory and statistics. It shows the probabilistic relationship among different variables which can be used for data analysis. This model can handle missing values, predict consequences of intervention, ideal for combining prior knowledge, and avoids over fitting [8][9].

IV. EXPERIMENTAL RESULTS

This section presents the result of the classifiers J48 (C4.5) and Bayes Net using WEKA before and after pruning. The training set "Enroll1.arff" has 1,970 instances with 5 attributes while the test set "Enroll1a.arff" has 27 instances

ISBN: 978-988-19252-5-1
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

with 5 attributes and consists of records which do not belong to the training set.

TABLE II
RESULTS OF WEKA

Classifier	After Pruning		Before Pruning	
	CCI	ICI	CCI	ICI
J48/C4.5	54.5178%	45.4822%	51.8519%	48.1481%
Bayes Net	51.1168%	48.8832%	70.3704%	29.6296%

The training set was used to evaluate the accuracy of the technique. In Table II, C4.5 or J48 in WEKA identified 55% (54.5178%) of CCI and 45% (45.4822%) ICI while BayesNet has 51% (51.1168%) CCI and ICI of 49% (48.8832%). After running the same classifiers with supplied test set, it shows that J48 (C4.5) has 52% (51.8519%) of CCI and 48% (48.1481%) of ICI. Bayes Net classified 70% (70.3704%) of CCI and 30% (29.6296%) of ICI. The CCI of J48 (C4.5) is 55% from the training set and the CCI of 52% from the test set indicates that the model is still accurate and can handle unknown data or any changes that may be applied to it in the future. In the case of Bayes Net, comparing the CCI of 51% from the training set and the CCI of 70% from the test set is an indicator that the model is vulnerable to handle unknown data or future data that can be applied to it. After pruning, the accuracy of data set using Bayes Net will be at risk since there is a big difference in CCI of 19%.

V. CONCLUSION/RECOMMENDATION

Considering the experimental results using WEKA, it is therefore concluded that since J48/C4.5 is the technique that obtained the highest percentage result, then it outperforms Bayes Net. However, since the result is only 56% accurate, based on the initial analysis, it is not really a recommended classifier in generating the model given the set of training data.

It is recommended for future work to try on other classification techniques that produce a better model for identifying the potential market for HEIs.

ACKNOWLEDGMENT

S. A. Abaya would like to thank RBA and PLORS for the continued support in this endeavor. This project will not be completed without your continuous motivation.

REFERENCES

[1] J. Han and M. Kamber, "Data mining concepts and technique," 2nd ed., 2006, p. 285.

[2] W. Wang, *A tutorial in WEKA*, Data Mining & Statistics within the Health Services, University of East Anglia, 2010.

[3] A. K. Sharma and S. Sahni, "A Comparative Study of Classification Algorithms for Spam Emails Data Analysis," *International Journal of Computer Science and Engineering*, vol. 3, no. 5, pp. 1890 – 1895, May 2011.

[4] I. Witten, E. Frank and M. Hall, "Data mining practical machine tools and techniques," 3rd ed., Elsevier Inc., 2011.

[5] T. Jyothirmayi and S. Reddy, "An Algorithm for Better Decision Tree," *International Journal of Computer Science and Engineering*, vol.2, no. 9, pp. 2827 – 2830, 2010.

[6] F. Ruggeri, F. Faltin and R. Kenett, "Bayesian networks encyclopedia of statistics in quality & reliability," Wiley and Sons, 2007.

[7] D. Heckerman, "A tutorial in learning bayesian networks," Microsoft Research Advanced Technology Division, Microsoft Corporation Redmond, WA98052, March 1995.

[8] P. Myllymaki, "On probabilistic modeling and bayesian network," Complex Systems Computation Group (CoSCo), Helsinki Institute for Information Technology (HIIT), Finland, 2002.

[9] D. Grossman and P. Domingos, "Learning bayesian network classifiers by maximizing conditional likelihood," in *Proc. 21st International Conference on Machine Learning*, Banff Canada, 2004.