Journal of Scientific Research & Reports
7(7): 494-500, 2015; Article no.JSRR.2015.230
ISSN: 2320-0227

SCIENCEDOMAIN *international*
www.sciencedomain.org

Implementing Classification Techniques of Data Mining in Creating Model for Predicting Academic Marketing

Sheila A. Abaya[1*], Bobby D. Gerardo[2] and Bartolome T. Tanguilig[3]

[1]Information Technology, Technological Institute of the Philippines, Currently a Faculty of the Department of Computer Studies and Systems, University of the East, Caloocan, Philippines.
[2]Administration and Finance, West Visayas State University, Iloilo City, Philippines.
[3]Academic Affairs, Dean of CITE and Graduate Programs Department, Technological Institute of the Philippines, Quezon City, Philippines.

Authors' contributions

The work was carried out in collaboration between all authors. Author SAA wrote the first draft of the manuscript. Author BDG suggested the classification techniques possible for comparison as well as some literature searches. Author BTT recommended improvement on data scaling. Author SAA managed the experimental procedures and identified the best classification technique for predicting probable academic market for tertiary education.

Article Information

DOI: 10.9734/JSRR/2015/16940
Editor(s):
(1) Saad Mohamed Saad Darwish, Department of Information Technology, Institute of Graduate Studies and Research (IGSR), University of Alexandria, Egypt.
(2) James P. Concannon, Associate Professor of Education, Westminster College, Fulton, Missouri, USA.
(3) Luigi Rodino, Professor of Mathematical Analysis, Dipartimento di Matematica, Università di Torino, Italy.
Reviewers:
(1) Anonymous, Stefan cel Mare University Suceava, Romania.
(2) M. Bhanu Sridhar, Dept. of CSE, GVP College of Engg. For Women, Vizag, India.
(3) Anonymous, Tribhuvan University, Nepal.
(4) G. Y. Sheu, Accounting Information Systems, Chang-Jung Christian University, Tainan, Taiwan.
(5) Anonymous, The University of Silesia, Poland.
(6) Anonymous, University of Tampere, Finland.
Complete Peer review History: http://www.sciencedomain.org/review-history.php?iid=1131&id=22&aid=9563

Original Research Article

Received 19th February 2015
Accepted 7th May 2015
Published 3rd June 2015

ABSTRACT

The education domain is one of the business areas with abundant data. Nowadays, most of tertiary educational institutions have dilemmas in identifying probable secondary schools which are considered as feeders for enrollment. The data mining technique of classification has been used in

**Corresponding author: Email: sheila_abaya@yahoo.com.ph;*

this research to easily identify the target secondary schools for enrollment. With these techniques, higher educational institutions may lessen the marketing cost by filtering which of these secondary schools are considered enrollment contributors. The techniques of ID3, C4.5, BayesNet and Naïve Bayes were used in this research implemented on WEKA 3.6.0 toolkit [1]. Based on the experimental results, C4.5 outperformed ID3, BayesNet and Naïve Bayes in determining the best classification technique to identify the targeted secondary schools qualified for enrollment in tertiary level. The model created can aid in education management's decision making process in terms of student recruitment.

Keywords: C4.5; J48; Id3; bayes net; naïve bayes.

1. INTRODUCTION

Data Mining (DM) has been treated as the forefront of business technologies [2]. With the overwhelmed increasing size of data in every business, patterns can be identified, validated and can be used for prediction. DM has several functionalities or tasks [3] that identify what kind of data patterns can be mined. One of which is the classification technique, [4] that can be used in predicting the target secondary feeder schools for enrollment in higher education. The methods of ID3, C4.5, BayesNet and Naïve Bayes were used in this research to identify which classifier works best in producing the model that identifies and determines the probable secondary schools which can be considered as the target schools for enrollment in higher educational institutions.

2. RELATED LITERATURES

Several studies have been conducted to compare different classification techniques.

Sharma, et al. [5] worked on the comparative analysis of J48, ID3, ADTree, and SimpleCART classification techniques for spam emails. The research focused on 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 categories and 58 attributes from which 57 are continuous and 1 is nominal. The results 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.

Bresfelean, [6] research focused on the application of classification technique in predicting probable student's choice in continuing their education with post university studies and their preference in certain fields of study and the data mining technique of clustering in grouping students with dissimilar behavior. Based on this paper, J48 (an implementation of C4.5 in WEKA) is known to be the most used WEKA classification algorithm that is noted to provide stability on precision, speed and interpretability of results simply because of the use of decision tree.

Grossman, [7] 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, original Bayesian network structure search algorithm (HGC) by Heckerman, 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.

Heckerman, [8] technical report on learning Bayesian network discusses the advantages of using BayesNet in classification and prediction. BayesNet can handle missing data entries; used for causal relationship that understands problem domains and predict the outcome of intervention; used ideally for representing prior knowledge; and avoids over fitting of data.

Naenudorn, [9] research compares the classification techniques of ID3, J48, Naïve Bayes and OneR in predicting the features of students who are likely to undergo the process of student admission. The data set of student used is from 2009 – 2011 with 6 attributes and 2,148 instances. The results of the experiment identified J48 (C4.5 implementation in WEKA) as the algorithm that provides the highest accuracy model and can be used to predict the future outcome of the pattern of students willing to enroll in the university.

Adhatrao, [10] works on applying the classification techniques of ID3 and C4.5 in predicting student's performance. Classification techniques were used rather than clustering because the former is suitable for prediction which is the subject matter of the research while clustering works on unknown class and are discovered from data. The tool was developed using PHP to interpret the decision trees of ID3 and C4.5 after data processing. It was found out from the results that both ID3 and C4.5 achieved an accuracy result of approximately 75.275%.

Abaya, [11] compares the classification algorithms of C4.5 and Bayes Net using 1,970 instances with 4 attributes and the final class is defined as "Enrolled" and "DidNotEnroll". A test set was also used to check for the accuracy of the model with 27 instances. The algorithms were implemented in WEKA. Based on the experimental results, the accuracy is close to 56% in favor of C4.5 algorithm in identifying potential market for enrollment.

3. BACKGROUND KNOWLEDGE

3.1 ID3 and C4.5

ID3 (Iterative Dichotomiser) is a decision tree algorithm developed by Ross Quinlan in the late 1970s and early 1980s and a predecessor of C4.5 were originally intended for classification. These methods follow the greedy approach in constructing decision trees. Trees are constructed in a top down approach in a divide-and-conquer manner. ID3 uses information gain in selecting relevant attributes while C4.5 uses the extension of information gain which is known as the gain ratio [4].

ID3/C4.5 Pseudocode [12]:

If the set of remaining non-class attributes is empty or if all of the instances in D are in the same class

return an empty tree
else {
compute the class entropy of each attribute over the dataset D
let a* be an attribute with minimum class entropy
create a root node for a tree T; label it with a*
for each value b of attribute a* {
let T(a*=b) be the tree computed recursively by ID3 on input (D\|a*=b, A-a*, C),
where:D\|a*=b contains all instances of D for which a* has the value b A-a* consists of all attributes of A except a*
attach T(a*=b) to the root of T as a subtree
}
return the resulting decision tree T }

3.2 BayesNet (BN) and NaiveBayes

BN is a graphical model for probability relationships among a set of variable features [13]. The model is believed to be true but with uncertainties and considered as subjective probability [14]. A Naïve Bayes is a simple BN classifier that produces a simple structure node which serves as the parent node of all nodes and no other connections are allowed [15].

This technique uses the algorithm of K2 by Cooper (1992).

K2 Pseudocode:

Procedure K2
For i:=1 to n do
$\pi_i = \phi$;
$P_{old} = g(i, \pi_i)$;
OKToProceed := true
while OKToProceed and $|\pi_i| < u$ do
let z be the node in Pred(x_i)- π_i that maximizes g(i, $\pi_i \cup \{z\}$);
$P_{new} = g(i, \pi_i \cup \{z\})$;
if $P_{new} > P_{old}$ then
$P_{old} := P_{new}$;
$\pi_i := \pi_i \cup \{z\}$;
else OKToProceed := false;
end {while}
write("Node:", "parents of this nodes :", π_i);
end {for}
end {K2}

4. METHODOLOGY

4.1 Data Preparation

The data preparation structure presented in Fig. 1 illustrates how the training data set as well as the test data sets was derived.

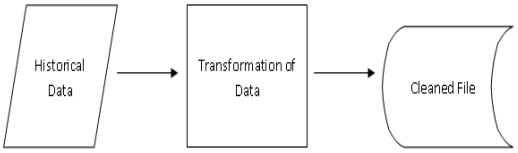

Fig. 1. Composition of data preparation

This historical data consist of the students' demographic information who took the College Entrance Test. This database goes into transformation where it was preprocessed and cleaned to achieve reliable results. Preprocessing and cleaning happen when the historical file continuous values are converted into discrete values. When these data are transformed, the attributes [11] are identified as Students General Weighted Average (Average); Parent's Income Bracket (Salary); School's Distance (Distance); School Ownership (Ownership) and Class. These attributes define the segmentation of the university's target market. Table 1 defines the attributes used in the data set while Table 2 defines the attribute values for Average, Distance, Ownership, Salary and Class.

Table 1. Definition of relevant attributes

Attribute	Definition
Average	The average grade of student before entering the higher education institution
Distance	Refers to the proximity of (target) secondary schools to the tertiary institution
Ownership	It identifies the type of management the university has.
Salary	Defines the parent's salary range of the prospective secondary students
Class	It defines prospective students whether they will "enroll" or "willnotenroll" in the higher educational institution.
	It refers to the final predictive value of :Enrolled" or "did not Enroll"

4.2 Experimental Results

The four classifiers (ID3, J48 an implementation of C4.5, BayesNet and Naïve Bayes) were used in the WEKA toolkit. The training dataset has 6,409 instances with 5 attributes while the test dataset has 1,015 instances with 5 attributes. Fig. 2 (a) shows the decision tree derived from C4.5, Fig. 2 (b) presents the pruned tree interpreted in the IF_THEN_RULES while Fig.3 shows the visualized graph of BayesNet where C represents the Class (Enrolled or DidNotEnroll), A is for Average, S for Salary, D represents Distance and O is identified as Ownership.

Table 2. Attribute values

Attributes	Alias	Values
Average	A	A1{75-79), A2(80-84},A3{85-89},A4{90-94},A5{95-100}
Distance	D	D1{1-9KM},D2{10-20KM},D3{=>21KM}
Ownership	O	PRI{Private}, PUB{Public}
Salary	S	S1{500-61999}, S2{62000-192999}, S3{193000-603999}, S4{604000-9999999}
Class	C	Enrolled or DidNotEnroll

An excerpt of the pruned tree is interpreted in the IF-THEN-RULES as follows:

If Distance = D1 and

If Ownership = PRI then 3,216 instances are classified as "DidNotEnroll"

If Ownership = PUB and Average = A1 and Salary = S1 then 73 instances are classified as "Enrolled"

If Distance = D2 and

If Ownership = PRI then 473 instances are classified as "Enrolled"

If Distance = D3 then There are 480 instances are classified as "DidNotEnroll"

Fig. 3 is interpreted as:

The occurrence of Average (A), Salary (S), Distance (D) and Ownership (O) is caused by the attribute Class (C).

Table 3 presents the accuracy result after running the classifiers in WEKA. Using the training data set test option ID3 and C4.5 identified 63% of Correctly Classified Instances (CCI) and 37% of Incorrectly Classified Instances (ICI), while BayesNet and Naïve Bayes identified 62% CCI with 38% ICI. Using the Supplied Test Data Set option, ID3 has the CCI of 71% which is not far from the CCI of C4.5 which is 72% and ICI of 29% and 28% respectively, while BayesNet and Naïve Bayes both identified a CCI of 52% and an ICI of 47%.

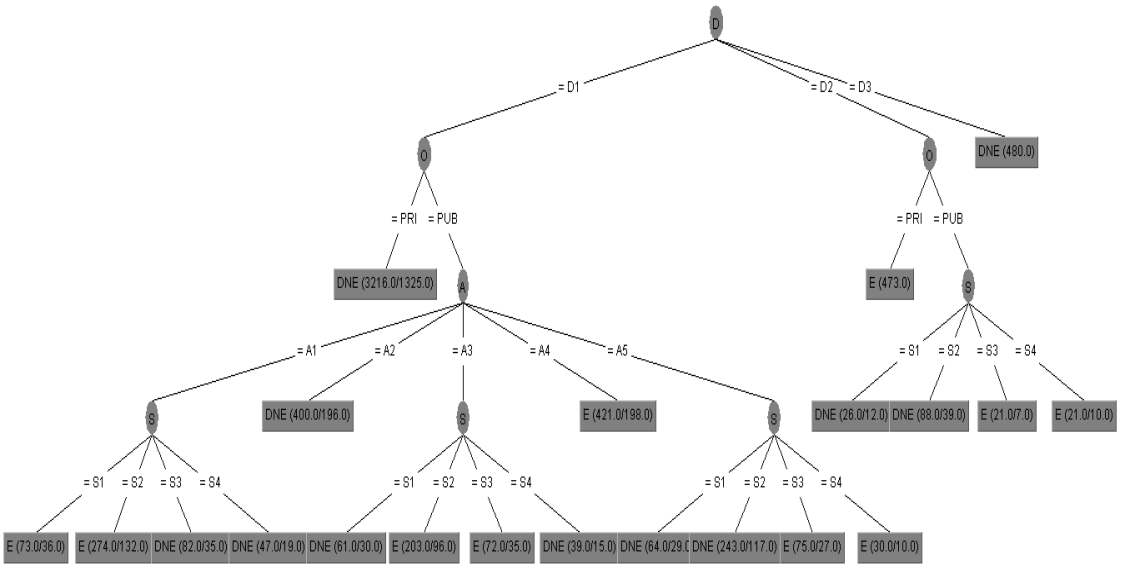

Fig. 2(a). J48 an example of visualize decision tree

```
D = D1
|   O = PRI: DNE (3216.0/1325.0)
|   O = PUB
|   |   A = A1
|   |   |   S = S1: E (73.0/36.0)
|   |   |   S = S2: E (274.0/132.0)
|   |   |   S = S3: DNE (82.0/35.0)
|   |   |   S = S4: DNE (47.0/19.0)
|   |   A = A2: DNE (400.0/196.0)
|   |   A = A3
|   |   |   S = S1: DNE (61.0/30.0)
|   |   |   S = S2: E (203.0/96.0)
|   |   |   S = S3: E (72.0/35.0)
|   |   |   S = S4: DNE (39.0/15.0)
|   |   A = A4: E (421.0/198.0)
|   |   A = A5
|   |   |   S = S1: DNE (64.0/29.0)
|   |   |   S = S2: DNE (243.0/117.0)
|   |   |   S = S3: E (75.0/27.0)
|   |   |   S = S4: E (30.0/10.0)
D = D2
|   O = PRI: E (473.0)
|   O = PUB
|   |   S = S1: DNE (26.0/12.0)
|   |   S = S2: DNE (88.0/39.0)
|   |   S = S3: E (21.0/7.0)
|   |   S = S4: E (21.0/10.0)
D = D3: DNE (480.0)
```

Fig. 2 (b). An example of J48 pruned tree

Table 3. Comparison of accuracy result

Classifier	Training dataset		Supplied Test dataset	
	CCI (%)	ICI (%)	CCI (%)	ICI (%)
ID3	63.25	36.75	71.43	28.57
J48/C4.5	63.05	36.95	72.02	27.98
BayesNet	61.59	38.41	52.32	47.68
NaiveBayes	61.59	38.41	52.32	47.68

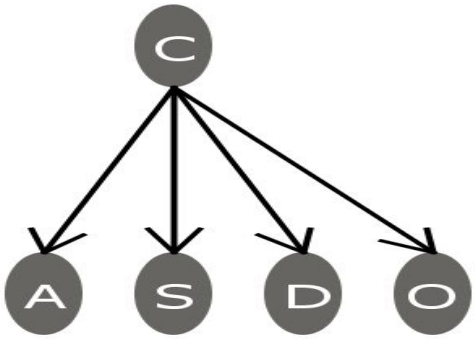

Fig. 3. An example of BayesNet graph

5. CONCLUSION / RECOMMENDATION

The relevant attributes that were used in creating a model for predicting the probable secondary school feeders for higher educational institutions are as follows: Average (which defines the general weighted average of a student before entering the university), Salary (which identifies the income bracket of the student's parents), Distance (which defines the proximity of secondary school to the prospective university/college institution), Ownership (which defines the type of school management whether it is privately owned or publicly operated), and Class (Final outcome of the prediction whether the student will enroll or will not enroll in the tertiary institution). Values for these attributes were also presented in the paper which was based from the actual values used in schools. With the results derived after running the four classifiers in WEKA toolkit, since ID3 and C4.5 correctly classifies instances of 71% and 72% respectively vis-à-vis with the result of BayesNet and Naïve Bayes wherein both classifiers correctly classified 52% of instances. Based from the generated results, it is therefore concluded that decision tree classifiers outperformed graphs in creating models for prediction. Moreover, the model that was created using the decision tree classifiers can be used in predicting the qualified secondary schools for academic recruitment.

ACKNOWLEDGMENTS

The author expresses gratitude to the University of the East, Philippines.

DISCLAIMER

This manuscript was presented in the conference "Proceedings of the International Multi Conference of Engineers and Computer Scientists 2014" available link IS " HTTP://WWW.IAENG.ORG/PUBLICATION/IMECS2014/ IMECS2014_PP342-345.PDF " DATE: MARCH 12 - 14, 2014,VOL I .

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

1. Wang W. A tutorial in WEKA. Data Mining & Statistics within the Health Services, University of East Anglia; 2010.
2. Witten I, Frank E, Hall M. Data mining practical machine tools and techniques. 2nd ed., Elsevier Inc; 2005.
3. Calders T, Pechenizkiy M. Introduction to the special section on educational data mining. SIGKKD Explorations. 2012;13(2).
4. Han J, Kamber M. Data mining concepts and technique. 2nd ed; 2006.
5. Sharma AK, Sahni S. A comparative study of classification algorithms for spam emails data analysis. International Journal of Computer Science and Engineering. 2011; 3(5):1890–1895.
6. Bresfelean VP. Data mining applications in higher education and academic intelligence management. Theory and Novel Applications of Machine Learning, I-Tech, Vienna, Austria; 2009.
7. Grossman D, Domingos P. Learning bayesian network classifiers by maximizing conditional likelihood. In Proc. 21st International Conference on Machine Learning, Banff Canada; 2004.
8. Heckerman D. A tutorial in learning bayesian networks. Microsoft Research Advanced Technology Division, Microsoft Corporation Redmond, WA98052; 1995.
9. Naenudorn E, et al. Classification model induction for student recruiting. Latest Advances in Educational Technologies. 2012;117–122.
10. Adhatrao K, et al. Predicting student's performance using ID3 and C4.5 classification algorithms. International Journal of Data Mining & Knowledge Management Process. 2013;3(5).
11. Abaya S, et al. Comparison of classification techniques in education marketing. Proceedings of the International

Multi Conference of Engineers and Computer Scientists, Vol. 1, IMECS 2014; 2014.

12. Available:http://www.cs.bc.edu/~alvarez/ML/id3.html (Retrieved March 9, 2015)
13. Al-Nabi D, Ahmed S. Survey on classification algorithms for data mining: (Comparison and Evaluation). Computer Engineering and Intelligent Systems. 2013; 4(8).
14. Ruggeri F, Faltin F, Kenett R. Bayesian networks encyclopedia of statistics in quality & reliability. Wiley and Sons; 2007.
15. Cheng J, Greiner R. Comparing Bayesian Network classifiers. Department of Computing Science, University of Alberta, Canada. Available:http://arxiv.org/ftp/arxiv/papers/1301/1301.6684.pdf (Retrieved March 10, 2015)

Peer-review history:
The peer review history for this paper can be accessed here:
http://www.sciencedomain.org/review-history.php?iid=1131&id=22&aid=9563