ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 6, June 2014

A Classification Model on Graduate Employability Using Bayesian Approaches: A Comparison

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

Department of Tropical Agriculture and International Cooperation, National Pingtung University of Science and Technology, Pingtung 91201, Taiwan[1]

Department of Management Information Systems, National Pingtung University of Science and Technology Pingtung 91201, Taiwan[2]

ABSTRACT: The aim of study presents a graduate employability model that uses Bayesian methods to search the most important factor of graduate employability, and to compare the accuracy of each algorithm under Bayesian methods including Naïve Bayesian Simple, Naïve Bayesian, Averaged One-Dependence Estimators, Averaged One-Dependence Estimators with subsumption resolution, Bayesian networks, and Naïve Bayesian Updateable. The results show that 3 factors with a direct effect on employability are the work province, occupation type, and times find work.

KEYWORDS: Bayesian methods, Classification techniques, Data-mining, Graduate employability

I. INTRODUCTION

According to the report by the Impact of Economic Crisis on Higher Education in 2012 indicated that The Thai higher education system is facing serious problems of graduate unemployment crisis. There were 320,815 graduates with bachelor's degrees and above in 2006. In 2007, the number of graduates increased to 371,982. About 75.02 per cent of graduates without those from open universities were employed in 2006 and 18 percent of graduates were unemployed. After that, the proportion of employed graduates decreased to 68.65 percent in 2008 and unemployment increased to 28.98 per cent. Effecting to graduates employment, graduates are likely to have more difficulties in finding jobs. This crisis is explicitly affecting to community, social and the nation. For these reason, it is necessary to require in-depth studies that enable identify the factors underlying graduates employability to be accurately identified so that more effective measures and standards can be implemented [1].

Classification approach is one of the most important data-mining especially in the area of the predicting. The capability of classification approach not only handles with huge volume of data to discover hidden patterns and relationships helpful in decision making, but also reduce their flexibility the data-generation structure, irrespective of complexity, great predictive and in some case, interpretative potential.

Therefore, the aim of this study is to compare the accuracy of classification model under Bayesian methods for identifying the factors underlying graduate employability. This useful information may provide in-depth data valuable for the ministry of education to monitor and improve various aspects of an administrative system in the institutions of higher education.

II. RELATED WORK

The Economist Intelligence Unit (2012) reveals that Thai universities are lacking of manufacturing graduates that possess good language skills, technical and information technology skills. Moreover, both of employer and employees reveals that the gaps in generic behavior skills consisting communication, leadership, social skills, time management, teamwork, and adaptability [2].

Yilmaz (2010) reveals six major considerations for hiring firms in Thailand, which are interpersonal skills, technical skills, Educational level, ethnic quotas, loyalty, and experience, whereas more than one fifth of all survey participants Thai firms noted technical skills as a major consideration in their hiring decisions [3]. Komintarachat (2012) reveals four criteria in term of BE concerned by BE graduates of Assumption University batch thirty-eight, graduate users of BE graduates in Thai employment market, and the Commission of Higher Education's five domains of learning form the Thailand Qualifications Framework for Thailand's higher education, these are degree of learning, usefulness of skills, continuance and relevance respectively [4].

The research by the Council for Industry and Higher Education of United Kingdom reveals that employers observe six competencies in people who can change the organization and add value to their careers [5]. These six competencies are personal capabilities, cognitive skills or brainpower, personal capabilities, generic competencies, technical ability, business or organization awareness, and practical elements. These competencies cover a set of achievements that include the skills, understandings, and personal attributes that make graduates more likely to gain employment and become successful in their selected occupations. This advantages the graduates, the community, and also the economy.

Moreover, data mining techniques have been used in the education domain for predicting and classifying. Research by Minaei-Bidgoli *et al* have classified students to predict their final grade using six common classifiers: 1-nearest neighbor (1-NN), the Quadratic Bayesian classifier, the Quadratic Bayesian classifier, k-nearest neighbor (k-NN), Multilayer Perceptron, Parzen-window, and Decision tree methods [6].

Guruler *et al* [7] identified that individual student characteristics are associated with their success according to the Grade Point Average (GPA) by using a Microsoft Decision Tree classification technique [8]. These studies have revealed some applications of classification in data mining technique of educational domain that extracts useful information from huge data sets. Data mining and analytical tools can help users' access current information for the decision-making process.

This study we are therefore applied classification techniques to construct graduate employability for identifying the factors underlying graduate employability from historical database and compare the accuracy of each models under Bayesian methods, in order to search the real factor and relationship are accrued.

III. METHODOLOGY

This study developed the research methodology based on three phases of the data-mining techniques including data preprocessing, classification task, and interpretation and evaluation.

A. *Data Preprocessing*

We collected raw data from the database of graduate historical in Khon Kaen University, Thailand for the academic year of 2009. The data set consists of 3,090 instances and 17 attributes. We created the classifiers by using the Waikato Environment for Knowledge Analysis (WEKA) program. This software was developed at the University of Waikato, New Zealand, and it can be easily applied to the data set. The default file type for data analysis in WEKA is the Attribute-Relation File Format file type, moreover, the data can also be imported in various formats such as CSV: Comma-Separated Values (text file), CSV file, and so on.

The data preprocessing phase includes two steps to prepare the data set for the classification task. The first step cleans and eliminates data with missing values in significant attributes, removes duplicate data, and identifies outliers. Then, we discrete the values of attributes into intervals with categorical or nominal attributes to prepare the data set for the classification task. These discretized values can be described as follows:

- GENDER is transformed into a nominal value from its previous value as a code (1 or 2).
- GPA is transformed into a grade range from its previous value as a continuous number.
- MATCH_EDU is transformed into a nominal value from a previous value as a code (1 or 2).
- ADDPROGRAM 1, 2, 3, 4, and 5 values are transformed into nominal values from their previous values as a code (1 or 2).
- WORK STATUS is transformed into a nominal value from its previous value as a code (1, 2, and 3).

B. *Classification Task*

In this part, we used classification technique which called Classification task. The classification task in this paper is to construct graduate employability model and predict the employment status (working, not working, or other) for graduate profiles. There are two stages in classification task consisting training and testing. The testing data set is used to estimate the predictive accuracy. The classify phase in WEKA includes four test modes as testing options: the training set, supplied test set, cross-validation, and percentage split [9].

- Training set: If we use this option as test, the test data will be sourced from training data, therefore, this option will decrease reliable measuring of the true error rate.
- Supplied test set: this option, we can use test data which been prepared separately from the training data.
- Cross-validation: this option is appropriate for limited dataset wherewith the number of fold can be determined by user. 10-fold class validation is widely used to get the best for measuring error. It has been widely assay on numerous datasets with different learning techniques.
- Percentage split: this option is evaluated on how well it predicts a certain percentage of the data that are held for testing. The amount of data held depends on the value entered in the % field.

We selected hold-out validation method with 70-30 percentage split in order to avoid over fitting of data, whereby 70% out of the 2,327 instances is used for training while the remaining instances are prepared for testing.

1) Bayesian Methods

The classification task in these methods comprises classifying a class variable based on a set of attribute variables. This is a type of statistical analysis in which the prior distribution is estimated from the data before any new data are observed; thus, every parameter is assigned a prior probability distribution [10]. The Naïve Bayesian algorithm works as follows: Let D be a training set of tuples and their associated class labels. As usual, each tuple is represented by an n-dimensional attribute vector, X = (x1, x2, …, xn), depicting n measurements made on the tuple from n attributes, respectively, $A1$, $A2$, … , An. Suppose that there are m classes, $C1$, $C2$, …, Cm. Given a tuple, X, the classifier predicts that X belongs to the class with the highest posterior probability, conditioned on X; that is, the Naïve Bayesian algorithm predicts that tuple X belongs to the class Ci if and only if $P(Ci|X) > P(Cj|X)$ for $1 \leq j \leq m; j \neq I$. Thus, we maximize $P(Ci|X)$. The class Ci for which $P(Ci|X)$ is maximized is called the maximum posteriori hypothesis.

We performed this experiment with six algorithms under the Bayesian methods in WEKA: the Averaged One-Dependence Estimators (AODE), Averaged One-Dependence Estimators with subsumption resolution (AODEsr), Bayesian Network, Naïve Bayesian, Naïve Bayesian Simple, and Naïve Bayesian Updateable algorithms.

The AODE and Naïve Bayesian algorithms were also used by Affendey et al. [11], and the remaining algorithms were selected to compare the results of the Bayesian algorithm experiment using the same data set. The AODE algorithm achieved the highest accuracy percentage in averaging all smaller searching-space in alternative Naïve Bayes-like models, which have weaker, and hence, less detrimental independence assumptions than the Naïve Bayesian algorithm. The resulting algorithm is computationally efficient and achieves highly accurate classification in many learning tasks. The AODEsr algorithm complements the AODE algorithm with Subsumption Resolution, which is capable of detecting specializations between two attribute values at classification time, and deletes the generalization attribute value. Bayesian Network learning uses various search algorithms and quality measures. In the Naïve Bayesian algorithm, numeric estimator precision values are chosen based on an analysis of the training data. The Naïve Bayesian Updateable algorithm uses a default precision of 0.10 for numeric attributes when build classifier is called with zero training instances. The Naïve Bayesian Simple modeled numeric attributes by a normal distribution.

C. *Interpretation and Evaluation*

In this part, we compared the performance under Bayesian methods. The AODEsr algorithm achieved the highest accuracy of 98.3% using the graduate data set. The second highest accuracy was achieved using AODE algorithm with an accuracy of 96.1%.

IV. RESULTS

Table 1 shows the classification accuracies for various algorithms under the Bayesian method. This table provides comparative results for the kappa statistics mean absolute error, root mean squared error, relative absolute error, and root relative squared error of the 699 testing instances. The OADEer algorithm achieved a higher accuracy percentage than other algorithms.

In Fig. 1, it shows a comparison of the accuracy and root relative squared error from each algorithm under Bayesian methods. This knowledge can be used to gain insights into the employment trend of graduates from local institutions of higher learning. A chart display of the accuracy and root relative squared error all algorithms under Bayesian methods reveal the accuracy of these approaches. The highest accuracy mean that the results in a better forecast.

Table 1. Classification accuracy used various algorithms under the Bayesian method

Algorithms	Accuracy (%)	Error Rate (%)	Kappa Statistic	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error (%)	Root Relative Squared Error (%)
Naïve Bayes	92.4	7.6	0.539	0.041	0.178	61.1	94.3
AODE	96.1	3.9	0.657	0.02	0.126	27.73	66.7
AODEsr	98.3	1.7	0.88	0.0128	0.075	17.76	39.8
Naïve Bayes Simple	92.4	7.66	0.53	0.044	0.178	61.1	94.2
BayesNet	92.4	7.6	0.550	0.044	0.179	60.92	94.6
Naïve Bayes Updateable	92.4	7.6	0.539	0.441	0.178	61.1	94.2

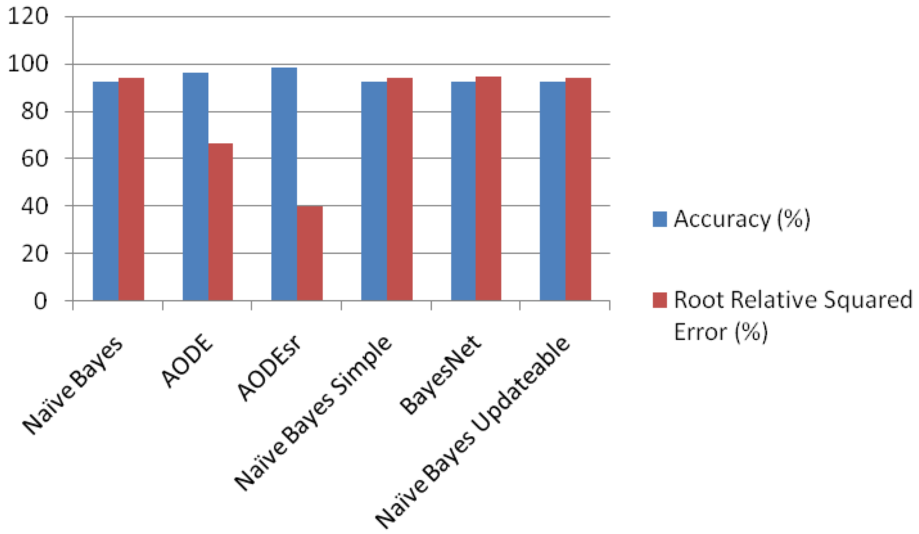

Fig.1. A comparison of the accuracy and root relative squared error under Bayesian methods

V. CONCLUSION

In this study, we compare six algorithms under Bayesian methods on graduate dataset with some parameter. In graduate dataset have a simple and class attribute. The results show that the AODEsr algorithm, achieved the highest accuracy of 98.3%. The second highest accuracy was achieved using AODE algorithm with an accuracy of 96.1%. In addition, the experiment show that 3 factors with a direct effect on employability are the work province, occupation type, and times find work.

ACKNOWLEDGMENTS

B. Jantawan would like to express thanks Dr. Cheng-Fa Tsai, professor in the Management Information Systems Department, National Pingtung University of Science and Technology, Taiwan for supporting the outstanding scholarship, and highly appreciates to the Khon Kaen University, Thailand for giving the information of this study.

REFERENCES

1. United Nations Educational Scientific and Cultural Organization, "The Impact of Economic Crisis on Higher Education", UNESCO Bangkok, Asia and Pacific Regional Bureau, Bangkok: Thailand, 2012.
2. Economist Intelligence Unit, "Skilled Labor Shortfalls in Indonesia, the Philippines, Thailand, and Vietnam: A custom research report for the British Council", Retrieved November 29, 2013, from http://www.eiu.com
3. Yilmaz, Y., "Higher Education Institutions in Thailand and Malaysia–can they deliver?, Retrieved November 29, 2013, from http://siteresources.worldbank.org/INTEASTASIAPACIFIC/Resources/Thailand-Malaysia-HEInstitutions.pdf
4. Komintarachat, H., "Development of a model for effective business English curriculum in an international university in Thailand", Journal of Assumption University Thailand, Vol. 4, pp. 84-92, 2012.
5. Rees, C., Forbes, P., and Kubler, B., "Introduction. Student Employability Profiles: A Guide for Higher Education Practitioners" (2nd ed., pp. 3), United Kingdom: The Higher Education Academy, 2006.
6. Minaei-Bidgoli, B., Kashy, D. A., Kortemeyer, G., and Punch, W. F., "Predicting student performance: An application of data mining methods with an educational Web-based system", 33rd Frontiers in Education Conference, pp. 13-18, 2003.
7. Guruler , H., Istanbullu, A., and Karahasan, M., "A new student performance analysing system using knowledge discovery in higher educational databases", Computers & Education, Vol. 55, pp. 247-254, 2010.
8. Kumar, V. and Chadha, A., "An Empirical Study of the Applications of Data Mining Techniques in Higher Education", International Journal of Advanced Computer Science and Applications, Vol. 2, pp. 80-84, 2011.
9. Kirkby, R. and Frank, E., "WEKA Explorer User Guide for Version 3-4-3", University of Waikato, 2004.
10. Jaynes, E. T., "Probability Theory: The Logic of Science", United Kingdom: Cambridge University Press, 2003.
11. Affendey, L. S., Paris, I. H. M., Mustapha, N., Sulaiman, M. N., and Muda, Z, "Ranking of influencing factors in predicting student academic performance", Information Technology Journal, Vol. 9, No. 4, pp. 832-837, 2010.

BIOGRAPHY

Bangsuk Jantawan is a Ph.D. student in the Department of Management Information System at the National Pingtung University of Science and Technology, Taiwan. Her research is the Application of Data Mining to Build Classification Model for Predicting Graduate Employment. Ms. Jantawan present research interests include the data mining, education system development, decision making, and machine learning.

Cheng-Fa Tsai is a full professor in the Department of Management Information Systems (MIS), National Pingtung University of Science and Technology, Taiwan. He has published over 160 well-known journal papers and conference papers and several books in the MIS. He holds, or has applied for, nine U.S. patents and thirty ROC patents in his research areas. Prof. Dr. Tsai research interests are in the areas of education, data mining and knowledge management, database systems, mobile communication and intelligent systems, with emphasis on efficient data analysis and rapid prototyping.