Journal of Engineering and Applied Sciences 12 (16): 4202-4207, 2017
ISSN: 1816-949X
© Medwell Journals, 2017

Application of Data Mining in Forecasting Graduates Employment

[1]Mohd Tajul Rizal and [2]Yuhanis Yusof
[1]Kolej Professional Mara Indera Mahkota, Kuantan, Pahang, Malaysia
[2]School of Computing, Universiti Utara Malaysia, 06010 Sintok, Kedah, Malaysia

Abstract: Obtaining information on graduate employability is crucial to every higher education institution. This is because such data would provide insight on the effectiveness of the institution curriculum in preparing human capital for the market needs. To date, the MARA Professional College (KPM) in Malaysia relies on graduates to manually provide data on their employment. Such an approach is not reliable as not all graduates provide the information to the institution. This study presents the application of data mining techniques in forecasting the KPM graduates employment type. In data mining, there exist three main tasks; classification, clustering, and association mining. The aim of this study is to forecast whether a particular graduate will be "employed", "unemployed" or "further study" 6 months after the completion of his study. The undertaken experiments include the utilization of five data mining techniques, namely, the Naive Bayes, Logistic regression, multilayer perceptron, K-nearest neighbor and decision tree. Furthermore, the experimental setup-up is based on three types of data proportion (training-testing) 70-30, 80-20 and 90-10. Based on the obtained result, it is learned that the Logistic regression is the best classifier for the in-hand dataset. In particular, the classifier is at its best when the 80-20 proportion is adopted. The produced classification model will benefit the management of the college as it provides insight to the quality of graduates that they produce and how their curriculum can be improved to cater the needs from the industry.

Key words: Data mining, graduates employment, Naive Bayes, logistic regression, multilayer perceptron, K-nearest neighbor, decision tree

INTRODUCTION

Graduates employment is one of the issue in Malaysia as there are as many as 53 higher education institutions that includes public and private university. The large number of institution will then produce a large number of human resources hence obtaining information on the whereabouts of the graduates can contribute to the strategic planning of the particular institution. Even though Malaysian education institutions are reported to produce more then 180,000 graduates each year, unfortunately there isn't any complete statistics on all of these graduates.

According to MOHE (MHEM, 2012), employment can be defined as the potential to secure a job at workplace and employability can be defined as the potential to maintain, secure and grow in a particular job at workplace. Based on Robinson (2000) in order to make sure that the graduates are employed, it depends on the employability skills such as model of understandings, achievements and personal attitude to get employment and successful in career. Another researcher, Buck and Barrick (1987) defined graduate employability as an attitude based on the personal value, decision making skills, problem solving, communication skills, relations with other people and commitment to get a job. Today, most companies hire employee not only based on the academic result and ability to write, listen and communicate but the employers are also concern about the creative thinking, problem solving, decision making and reasoning of the potential employee.

One of the research undertaken by Shafie and Nayan (2010) revealed that the highest employability, given 100 graduates is from Universiti Teknologi MARA (i.e., 77). This is followed by Universiti Sains Malaysia (USM) with 74, Universiti Islam Antarabangsa (UIA) with 71, Universiti Malaya (UM) with 63, Universiti Putra Malaysia (UPM) with 61, Universiti Kebangsaan Malaysia (UKM) with 38, Universiti Teknologi Malaysia (UTM) with 35 and Universiti Malaysia Sarawak (UNIMAS) with 34.

MARA Professional College (KPM) was formerly known as MARA Institute of Commerce (IPM) which was established on May 1977. KPM offers accredited diploma program from Malaysian Qualification Agency (MQA). To

Corresponding Author: Mohd Tajul Rizal, Kolej Professional Mara Indera Mahkota, Kuantan, Pahang, Malaysia

date, KPM have 6 branches which includes Seri Iskandar (KPMSI) Perak, Beranang (KPMB) Selangor, Bandar Melaka (KPMBM) and Ayer Molek (KPMAM) Melaka, Bandar Penawar (KPMBP) Johor and Indera Mahkota (KPMIM) Pahang. For this study, the study is focusing on KPMIM which offers 4 diploma programs such as Diploma in Accountancy (DIA) Diploma in English Communication (DEC) Diploma in Computer Networking (DCN) and Diploma in Business Digital Media Creative (DDC). At the moment, there are 1500 students and every year, a total of 300 graduates are produced by KPMIM. In practice, information on employment of KPMIM graduates is obtained directly from the graduates. It is the responsible of the alumni unit to acquire the employment status from every graduate and to date, the process is done manually. Such a time consuming process is not efficient as the number of KPMIM graduates are increasing and there exist graduates who do not respond to the request from the alumni unit. As the graduate employment status is important in determining the quality of the program (i.e., how relevant is the program to the market) offered by an education institution, there is a need to automatically capture the statistics.

To date, various approaches have been used to study graduate employability and this includes data mining. Data Mining (DM) is a part of artificial intelligence technology where it analyzes raw data and transforms it into knowledge. DM has been applied in various applications such as web mining and decision support system (Sheng *et al*., 2010). Data mining offers 4 main tasks; classification, clustering, regression and association mining. These tasks can be achieved using various methods such as Decision Tree (DT) linear regression, logistic regression, Naive Bayes, Neural Network (NN) K-nearest Neighbor (K-NN) Relevance Vector Machine (RVM) and Support Vector Machine (SVM).

In this study, data mining (i.e., classification task) is performed on KPMIM data in order to produce a classification model that best suits KPMIM. It is hoped that the obtained model can be used by the KPMIM management to forecast its graduate employment. Upon completing a program, KPMIM will know in advance whether a graduate will be employed in public or private sector, unemployed or continue study.

Literature review: Graduates employment has been defined in various forms. In general, graduates employment means that a graduate is able to get a job once he has completed his training or education program. In higher education, it is important that graduates are able to find job that is relevant to their qualification. Today, the number of graduates is increasing vastly as compared to the demands from the market. One of the issues of why graduates fail to secure a job is personal attributes of the graduate which includes communication skills, confidence level, teamwork, professional and decision making skills. These skills are consider competent in application and practice. Other than that is understanding for scientific and technologies (Life Long Learning).

The Malaysian Government conducted a survey on Malaysian graduates and it was discovered that about 60,000 Malaysian graduates were unemployed due to a lack of experience, poor in English, poor communication skills and because they had pursued studies irrelevant to the market (Ting and Ying, 2012). The research further mentioned that the typical unemployed graduate was female, mainly from the Malay ethnic group and from the lower income group. Most unemployed graduates had majored in business studies or information technology. A total of 81% of the unemployed graduates had attended public universities where the medium of instruction in many courses was the Malay language. The Ministry of Human Resource recently reported that a large number of graduates are still jobless. According to the report, 70% graduates of from public universities and institutions of higher learning are still unemployed. This is in contrast with 26% from private institutions of higher learning and 34% who are foreign graduates (Suresh, 2006).

Peter Knight from the Institute for Educational Technology at the Open University is quoted in the Hobsons Directory 2005 (www.get.hobsons.co.uk) for graduate-level vacancies, discussing skills looked on favorably amongst employers: “When hiring, employers generally value good evidence of ability to cope with uncertainty, ability to work under pressure, action-planning skills, communication skills, IT skills, proficiency in networking and team working, readiness to explore and create opportunities, self-confidence, self-management skills and willingness to learn”.

A study reported in the Conference Board of Canada, divided graduate employability into three layers of fundamental skills, personal management and teamwork skills. Reich (1991) made a list of necessary skills consists of abstraction, system thinking, experimentation and collaboration. A domestic scholar divided graduate employability into social compatibility, pre-professional image and characters for job. For deaf college student can conclude in positive perception IT will improved their initiative of study and conversation, social network, get the job information and negative is daily used are not suitable because get the harmful information.

Besides of studies that determine attributes to contribute in determining employability, there also exist studies on the automatic classification of the graduates. In Jantawan and Tsai (2013) they reported on graduate students of Maejo University in Thailand where they determine if graduates are employed within 12 months after graduation. The research also identifies attributes for skilled graduates. The utilized DM methods includes Bayesian Network and decision tree. They concluded that WAODE algorithm in Bayesian Network is a better classifier because it obtained 99.77% accuracy as compared to J48 algorithm that produces 98.31% accuracy.

A study by Abu Tair and El-Halees (2009) employs other classifiers such as neural network and K-nearest neighbour to classify graduate employment and their work was based on the Khanyounis College of Science and Technology graduates. The study includes 4 stages; association, classification, clustering and outlier detection. In the association stage, they associate the student's grades as either being "excellent", "very good", "good" and "average". In classification their used rule induction and Naive Bayesian. Based on analysis researcher suggest used the Naive Bayesian because this method can predict on time average student. In the clustering, they apply single value decomposition to cluster plot average from 0-3. Finally, in the outlier detection stage, researchers fou nd two main approaches, it is distance based approach with means student excellent result at the matriculation can be excellent result in the college. Researchers identify association rules and used the classification to predict the graduate employment.

Prior to that in a study performed by Wook *et al.* (2009) try to combine Artificial Neural Network (ANN) and decision tree. The proposed a classifier was evaluated on students from the Computer Science Department, Faculty of Science and Defence Technology, Universiti Pertahanan Malaysia. There are six stages involved; study understanding, data collection, data preparation, modelling, evaluation and deployment. The research employs 85 students of semester 1 2008/2009 and focuses on 2 types of information (i.e., personnel and academic) to predict the success of the student. The utilized ANN architecture operates based on sigmoid function and the decision tree was employed to represent and understand each cluster. Nevertheless, no data on classification accuracy was reported (So-In *et al.*, 2014).

Based on the a forementioned studies, it is noted that data mining helps to automate the classification of graduates. The success of data mining can also be seen in other domains such as in customer analysis (Zhang *et al.*, 2008). The employed algorithms includes the C5.0 clementine decision tree model and Classification and Regression Trees (CART). Based on these classifications the highest accuracy algorithm model is C5.0 is 82.53% and CART is 82.24%.

In year 2010, a survey was done on Mobile Telecom Market to develop customer classification model using data mining approach (Ahn *et al.*, 2008). The researcher divided 2 sections of DM techniques. Section 1 includes several DM classifications such as Logistic Regression (LR) Decision Tree (DT) and Artificial Neural Network (ANN). Section 2 employs the Genetic Algorithm classification to provide probability. The utilized techniques includes Simple Averaging of LR, DT and ANN (SAVG) Majority Voting of LR, DT and ANN (MVOTE) and Weighted Averaging of LR, DT and ANN (GAOW). Experiments results showed that DT produced the highest accuracy with 62.87%. In addition, in the research researchers use data mining to investigate intrusion detection. Based on this model he analyzed and find different result among classes with is normal versus attack, type of attack and individual attack. For this case, the best classification model was the K-nearest neighbor as requires the lowest computational complexity compared to other classification models.

Another area that benefits from data mining is the image processing (Wagle *et al.*, 2013). The researchers wanted to improve medical image classification model using DM techniques. Their used the K-Nearest Neighbor (KNN) and compare with other techniques such as Naive Bayes Classifier (NB), Neural Network (NN) and Support Vector Machine (SVM). Based on the undertaken experiment they concluded that KNN classifier is much easier to use and implemented on normal or severe category of medical images.

MATERIALS AND METHODS

The main objective of this study is to find the best classification model for KPMIM graduates employment. The aim is to use the model to forecast the types of employment for future graduates. Hence, 4 research phase were undertaken; data collection, data pre-processing, data mining and model evaluation as shown in Fig. 1.

Data collection: The data for this study is obtained from 3 resources (units) examination, alumni and the

Table 1: Attributes for the classification of graduates employment

Attributes	Values	Description
Gender	Male, female	Gender of graduate
Program	Diploma in English communication, diploma in accountancy, diploma in computer networking, HND bit, diploma in business digital media creative	Program enrolled in KPMIM
Semester	Number of semesters	Number of semesters taken by a student to complete his program
Co-curriculum	Active, very_active or not_active	Student involvement in clubs and societies
Academic	Excellent, good, satisfactory, pass and fail	CGPA
Employment	Private, public, unemployed and further_study	Types of graduate employment

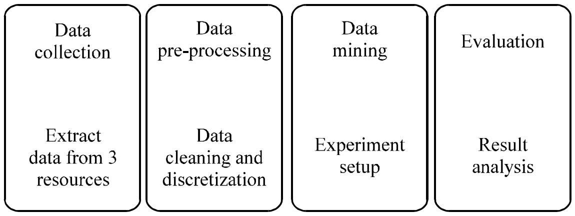

Fig. 1: Research phases

curriculum. The first unit provides academic information that includes program, CGPA and number of semester. An example for the first attribute includes Diploma in English Communication while the values 3.88 and 6 are examples for the second and third attributes.

Data obtained from the alumni unit includes gender and status of employment. The second attribute represents the types of employment. In general, the study is to classify a future graduate into either "employed", "unemployed" or "further_study". However, to provide the institution with a detail information, this study further classifies the "employed" category into "private" or "public". Hence, the overall class label for the KPMIM dataset includes 6 label; "private", "public", "unemployed" and "further_study".

The third data source which is the curriculum unit, provides the data on student involvement in co-curriculum activities such as clubs and societies. The involvement is represented as either "active", "very_active" or "not_active". A student is indicated as "active" if he joins a minimum of 2 clubs or societies and "very_active" if he holds any managing position in the society.

Data pre-processing: The second phase of the study includes 2 activities; data cleaning and data discretization. The first activity involves replacing any missing values and removing the duplicate attributes. On the other hand, there are several attributes that need to be discretized where data is represented in categories form. This study discretizes the "CGPA" data into five categories. The first category is represented as "excellent" that includes grade points of 4.00-3.67 while the points of 3.65-3.00 is included in the "Good" category. The "Satisfactory" and "Pass" bin includes the points of 2.99-2.33 and 2.32-2.00, respectively. The final category (i.e., "Fail") then includes the CGPA that is <2.00. The outcome of phase 2 is presented in Table 1.

Data mining: In determining the best forecasting model for the in-hand dataset, this study executes the pre-processed dataset on 5 data mining techniques offered by weka; Naive Bayes, logistic regression, multilayer perceptron, K-nearest neighbor and decision Tree J48. In addition, different data proportion for the test option is also employed. This includes the 70-30, 80-20 and 90-10 proportion. The larger portion is used for training while the smaller ones indicate the percentage of data used for testing purposes.

RESULTS AND DISCUSSION

Evaluation: Analysis on the accuracy is undertaken to determine which classifier is best suited for KPMIM dataset. Table 2 includes the obtained accuracy rate while using the different classifiers and test options (i.e., training, cross validation and testing). The findings showed that the highest accuracy classification is produced by logistic regression which is at 92.47% while using data proportion of 80-20. The second best classifier for the same test option is the Naive Bayes with 79.40%. On the other hand, the Bayes classifier performed the worst when the 70-30 data proportion is employed, obtaining only 62.40%.

The multilayer perceptron classifier obtained its highest accuracy (i.e., 69.93%) when testing option 90-10 is utilized. The K-nearest neighbor shows similar capability as it produces 69.45%. Nevertheless, the J48 decision tree was at its best when training set was used as the test option. A more detail result is shown in Table 3 for the 80-20 data proportion. Data includes the various error rate obtained based on the accuracy.

Table 2: Accuracy for types of graduates employment

	Accuracy					
		Cross validation		Testing		
Classifier	Training	5	10	70	80	90
Naive Bayes	65.34	63.60	63.76	62.40	79.40	63.95
Logistic regression	66.77	65.03	64.55	61.65	92.47	64.17
Multilayer perceptron	68.98	63.44	62.65	63.90	68.91	69.23
K-nearest neighbor	69.93	63.92	63.76	64.66	70.09	69.45
J48	67.24	64.24	63.60	63.15	66.53	65.93

Table 3: Detail error rate

Classifier	Error	MAE	RMSE	RAE	RRSE
Naive Bayes	20.59	0.15	0.24	51.99	65.41
Logistic regression	7.520	0.03	0.17	10.61	45.65
Multilayer perceptron	31.08	0.17	0.30	60.16	78.58
K-nearest neighbour	29.90	0.16	0.29	58.41	76.35
J48	33.46	0.19	0.31	68.05	82.54

CONCLUSION

Every year, the number of graduates produced by higher education institutes is increasing. The optimal scenario is that the number matches job opportunities in the market. However, in reality, this is hard to achieve. Hence the goal of this study is to assist KPM management in estimating their graduate employability. This is achieved by creating a classification model that is able to automatically forecast whether a graduate will be employed in public sector, private sector, continue study or even unemployed. Based on the experiments, it is learned that the Logistic Regression Model is the best classifier for the KPMIM dataset. The model produces as high as 92.5% accuracy and has outperformed the other 4 classifiers.

SUGGESTIONS

As for future work, there is a need to include more attributes in the model such as the grades for major courses in the program and the types of co-curriculum involvement (being a president, secretary or treasurer). Inclusion of these attributes would provide better understanding on the employment pattern of KPIM graduates. Furthermore, the new set of attributes need to be tested on other machine learning classifiers such as Support Vector Machine and Least Suqares Support Vector Machine.

ACKNOWLEDGEMENT

The researcher would like to thank the management of Kolej Professional MARA in particular from the Indera Mahkota branch in providing the data and financial support for the study.

REFERENCES

Abu Tair, M.M. and A.M. El-Halees, 2012. Mining educational data to improve students performance: A case study. Int. J. Inf. Communi. Technol. Res., 2: 140-146.

Ahn, H., C. Song, J.J. Ahn, H.Y. Lee and T.Y. Kim et al., 2010. Using hybrid data mining techniques for facilitating cross-selling of a mobile telecom market to develop customer classification model. Proceedings of the 43rd Hawaii International Conference on System Sciences (HICSS), January 5-8, 2010, IEEE, Honolulu, Hawaii, ISBN:978-1-4244-5509-6, pp: 1-10.

Buck, L.L. and R.K. Barrick, 1987. They're trained, but are they employable?. Vocational Educ. J., 62: 29-31.

Jantawan, B. and C. Tsai, 2013. The application of data mining to build classification model for predicting graduate employment. Intl. J. Comput. Sci. Inf. Secur., 11: 1-8.

MHEM., 2012. The National Graduate Employability Blueprint 2012-2017. Ministry of Higher Education Malaysia, Putrajaya, Malaysia, ISBN: 978-967-0334-43-1, Pages: 62.

Reich, R., 1991. The Work of Nations: Preparing Ourselves for 21st Century Capitalism. Alfred A. Knopf, New York, USA., Pages: 331.

Robinson, J.P., 2000. What are employability skills. Workplace, 1: 1-3.

Shafie, L.A. and S. Nayan, 2010. Employability awareness among Malaysian undergraduates. Int. J. Bus. Manage., 5: 119-123.

Sheng, Q., Z.L. Shan and Y.Y. Xiang, 2010. A web-based distributed group decision support system for railway construction organization. Proceedings of the 2nd International Asia Conference on Informatics in Control, Automation and Robotics (CAR) Vol. 2, March 6-7, 2010, IEEE, Wuhan, China, ISBN:978-1-4244-5192-0, pp: 362-365.

So-In, C., N. Mongkonchai, P. Aimtongkham, K. Wijitsopon and K. Rujirakul, 2014. An evaluation of data mining classification models for network intrusion detection. Proceedings of the 4th International Conference on Digital Information and Communication Technology and it's Applications (DICTAP), May 6-8, 2014, IEEE, Bangkok, Thailand, ISBN:978-1-4799-3724-0, pp: 90-94.

Suresh, 2006. 70% of grads from public institutions jobless. The Daily Sun, Bangladesh.

Ting, S.K.T. and C.Y. Ying, 2012. Business graduates competencies in the eyes of employers: An exploratory study in Malaysia. World Rev. Bus. Res., 2: 176-190.

Wagle, S., J.A. Mangai and V.S. Kumar, 2013. An improved medical image classification model using data mining techniques. Proceedings of the 7th IEEE Conference and Exhibition (GCC), November 17-20, 2013, IEEE, Doha, Qatar, ISBN:978-1-4799-0722-9, pp: 114-118.

Wook, M., Y.H. Yahaya, N. Wahab, M.R.M. Isa and N.F. Awang et al., 2009. Predicting NDUM student's academic performance using data mining techniques. Proceedings of the 2nd International Conference on Computer and Electrical Engineering ICCEE'09 Vol. 2, December 28-30, 2009, IEEE, Dubai, UAE., ISBN:978-1-4244-5365-8, pp: 357-361.

Zhang, L., Y. Chen, Y. Liang and N. Li, 2008. Application of data mining classification algorithms in customer membership card classification model. Proceedings of the International Conference on Information Management, Innovation Management and Industrial Engineering ICIII'08. Vol. 1, December 19-21, 2008, IEEE, Taipei, Taiwan, ISBN: 978-0-76 95-3435-0, pp: 211-215.