Jurnal Teknologi

Full Paper

DATA MINING APPROACHES IN BUSINESS INTELLIGENCE: POSTGRADUATE DATA ANALYTIC

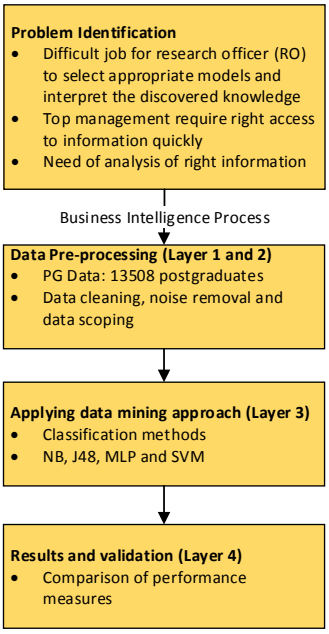
Shamini Raja Kumaran, Mohd Shahizan Othman*, Lizawati Mi Yusuf

Department of Computer Science, Faculty of Computing, Universiti Teknologi Malaysia, 81310 UTM Johor Bahru, Johor, Malaysia.

Article history
Received
20 November 2015
Received in revised form
9 May 2016
Accepted
16 March 2016

*Corresponding author
shahizan@utm.my

Graphical abstract



Abstract

Over recent years, there has been tremendous growth of interest in business intelligence (BI) for higher education. BI analysis solutions are operated to extract useful information from a multi-dimensional datasets. However, higher education-based business intelligence is complex to build, maintain and it faces the knowledge constraints. Therefore, data mining techniques provide an effective computational methods for higher education-based business intelligence. The main purpose of using data mining approaches in business intelligence is to provide decision making solution to higher education management. This paper presents the implementation of data mining approaches in business intelligence using a total of 13508 postgraduates (PG) data. These PG data are to allow the research to identify the postgraduates who Graduate On Time (GOT) via business intelligence process integrating data mining approaches. There are four layers will be discussed in this paper: data source layer (Layer 1), data integration layer (Layer 2), logic layer (Layer 3), and reporting layer (Layer 4). The main scope of this paper is to identify suitable data mining which is to allow decision making on GOT so as to an appropriate analysis to education management on GOT. The results show that Support Vector Machine (SVM) classifier is with better accuracy of 99%. Hence, the contribution of data mining in business intelligence allows an accurate decision making in higher education.
Keywords: Business intelligence (BI), data mining, postgraduate (PG) data, higher education, decision making

Abstrak

Sejak kebelakangan ini, terdapat pertumbuhan pesat bagi kepentingan dalam perisikan perniagaan (BI) berdasarkan pengajian tinggi. Penyelesaian analisis BI dikendalikan untuk mendapatkan maklumat yang berguna dari set data yang multi-dimensi. Walau bagaimanapun, BI berasaskan pengajian tinggi adalah kompleks untuk dibina, atau untuk dikekalkan dan ia menghadapi kekangan pengetahuan. Oleh itu, teknik-teknik perlombongan data merupakan satu kaedah pengiraan yang berkesan untuk BI berasaskan pengajian tinggi. Tujuan utama menggunakan pendekatan perlombongan data dalam BI adalah untuk menyediakan penyelesaian pengurusan bagi membuat keputusan di sektor pengajian tinggi. Kertas ini membentangkan pelaksanaan pendekatan perlombongan data dalam Risikan Perniagaan menggunakan 13508 data pasca siswazah (PG). Data PG ini membolehkan pihak pengurusan untuk mengenal pasti pasca-siswazah yang memperolehi *Graduate On Time* (GOT) melalui proses BI dengan implementasi perlombongan data. Terdapat empat lapisan akan dibincangkan dalam kertas ini: lapisan sumber data (*Layer 1*), lapisan data integrasi (*Layer 2*), lapisan logik (*Layer 3*) dan lapisan laporan (*Layer 4*). Fokus utama kertas kerja ini adalah lapisan logik dengan pendekatan perlombongan data. Keputusan menunjukkan bahawa klasifikasi *Support Vector Machine* (SVM) mendapat ketepatan 99%. Oleh yang demikian, sumbangan perlombongan data dalam proses BI membolehkan membuat keputusan yang lebih baik dalam bidang pengajian tinggi.
Kata kunci: Perisikan perniagaan (BI), perlombongan data, data pasca siswazah (PG), pengajian tinggi, membuat keputusan

1.0 INTRODUCTION

Higher education systems all over the world nowadays are challenged by the requirement of gaining information from data [1]. Gartner [2] stated that in recent years, business intelligence (BI) has emerged as one of the top ten spending priorities for many Chief Information Officers. This clearly shows that business intelligence becoming a demand nowadays because of its capability and ability to extract and present accurate information in a real-time mannered for decision makers to produce a good business decision [3]. Business intelligence is a process for extracting, transforming, managing and analysing large data results to gain information and knowledge to help make decisions in the complex environment [4].

Here, data mining approach playing the main role in BI to analysis the data. In addition, this analysis process via the BI based system to deliver the lists of data which are ideally placed to see the progress of higher studies activities grow on daily basis. Despite this, the importance of usage of data mining as a tool for higher institutions are due to having a large amount of data. Alnoukari [5] stated that decision makers depend on detailed and accurate information when they required to make decisions. Hence, BI can offer decision makers with such accurate information by using appropriate BI framework tools for data analysis along with proper data mining method. Under this article, the case study presents decision-making on Graduate on Time (GOT) for higher education which will be discussed in detail at Section 2.0.

This paper is arranged as follows: current section gives an overview to what the research comprises about with Section 2, BI process with its GOT case study attributes information. In Section 3, discusses on the layers in business intelligence processes. Next, Section 4 reveals detailed explanation on data mining (Layer 3) in business intelligence and continued Section 5 with results using data mining approaches. Finally, Section 6 draws a conclusion from the results in previous section.

2.0 RESEARCH CASE STUDY

To begin the research, three main problems are identified via the interviews with research officers in higher education sector. The first problem is related to the rapid increase of data causes barrier toward successful decision-making within department. The second problem is due to long working process beginning with extraction of data, manual data cleaning, manual data analysis and interpretation of analyzed data to the discovered knowledge. Lastly, the third problem is related to the requirement of ad-hoc reporting by the top management. Hence, the solutions are via the application of data mining in business intelligence process. All BI processes under each layers play the important role to achieve good decision-making.

The layers involved in the business intelligence framework is data source layer, data integration layer, logic layer and reporting layer. Data source layer (Layer 1) comprises internal and external data source that provide data to undergo data pre-processing. Data integration layer (Layer 2) consists of data warehouse and data marts and involves data pre-processing before the data injected into logic layer (Layer 3). Using data mining techniques under logic layer, many kinds of knowledge from data can be discovered and it operates on large volumes of data which are helpful in decision-making [6]. The last layer is the reporting layer (Layer 4) where it visualizes the data in the form of reporting for the view of decision makers. The most important part in BI process is the Layer 3 in which the processed raw data will transform into knowledge and decisions. These decisions improve classification accuracy compared to when using raw data directly for decision-making.

The research case study discussed under this article is to identify suitable data mining which is to allow decision making on GOT for an appropriate analysis to education management. Graduate on Time (GOT) under the context of higher education refers to students who completed their study within specified duration. Basically, the measurement of GOT under this case study for PhD postgraduates to complete their studies are maximum 6 semesters (based on university measurement).

Hence, the obtained attributes reveals a major role in predicting postgraduates who GOT in the future. In addition, this will assist the higher education management to identify postgraduates who are at risk and able to provide better additional boost to postgraduates to achieve GOT. This paper presents the implementation of data mining approaches in business intelligence using a total of 13508 postgraduates (PG) data. The attributes that are used under this case study are as shown in Table 1. Table 1 depicts 13 attributes is to identify suitable data mining which is to allow decision making on GOT.

Table 1 GOT case study attributes

Attributes	Sample Data
Student ID	XX103367
Faculty	Faculty of Science
CGPA Bachelor	3.66
CGPA Masters	3.70
Study Mode	Full Time
Discipline	Science and Technology
Sponsorship	myBrain PhD
Gender	Female
Citizenship	Malaysia
Enrolment Year	2010
No. of Semester	5
Intake CGPA>=3.00	Yes

These attributes obtained from the education department which is in charge in generating monthly report Graduate on Time (GOT) postgraduates for top management. Further discussion on layers in Section 3.0.

3.0 PROCESSES IN BUSINESS INTELLIGENCE (BI) LAYERS

Data source layer (Layer 1) is the utmost important layer in BI process and consists of external and internal data in the form of databases, Spreadsheets, or Microsoft Excel. Under this case study, the data extracted from the databases and injected into data integration layer (Layer 2). However, it is hard to get clean data from the databases. There are many types of data constraints that effect the BI process. Some possibilities that faced under this research such as data redundancy, and data defects (for example, manually entered data with falsified data, missing data, and correct data in the wrong fields) [7]. To fix this defective data, BI process allow extract, transform and load process before the integration of data from different databases into data warehouse and dispersed into data marts. This process will result in accurate, complete and non-obsolete data into logic layer (Layer 3) which the data mining process will be conducted.

Under this case study, the initial data was 13508 postgraduates (graduated year 2014), after initializing the scope of the case study which was extracting PhD enrolment from the year 2011 till 2012 who was graduated year 2014, the total was compressed to 273 postgraduates. As stated earlier, there are data which are falsified, correct data in wrong fields, data duplicated, and some data were outdated. These data then narrowed to 121 postgraduates (PG) by going through severe cleaning process. Total of 121 PG data then injected into logic layer to undergo data mining process mainly is to identify suitable data mining which is to allow decision making on GOT.

According to [8], university able to predict the students demographic and can use data mining techniques to identify the intellectual features. The main task of data mining is to examine the large amount of data in order to extract the patterns. Data mining as BI components represents the computational data process from its techniques perspectives, with the goal to extract, trends and converts information into data [9]. This will assist to BI process to discover, detect and explain the GOT phenomena. Figure 1 shows the cycle of application of data mining in BI system for higher education. Figure 1 describes the data mining process in Layer 3 as proposed under this research.

The definition of data mining is the extraction of previously unknown data and analyzing data to discover meaningful patterns and rules [10]. The basic understanding of data mining study is generally begun with the need for new knowledge. This data mining will be embedded along with BI to maintain the quality of decision making in higher education. Therefore, the quality of decision making in any higher education institutions is most likely to obligate the requirements of top managements. In addition, Figure 2 shows the layers involved in BI processes for better understanding for description that explained under this Section 3.0. In the upcoming Section 4.0, Layer 3 has been focused and 121 PG data has been used is to identify suitable data mining which is to allow decision making on GOT.

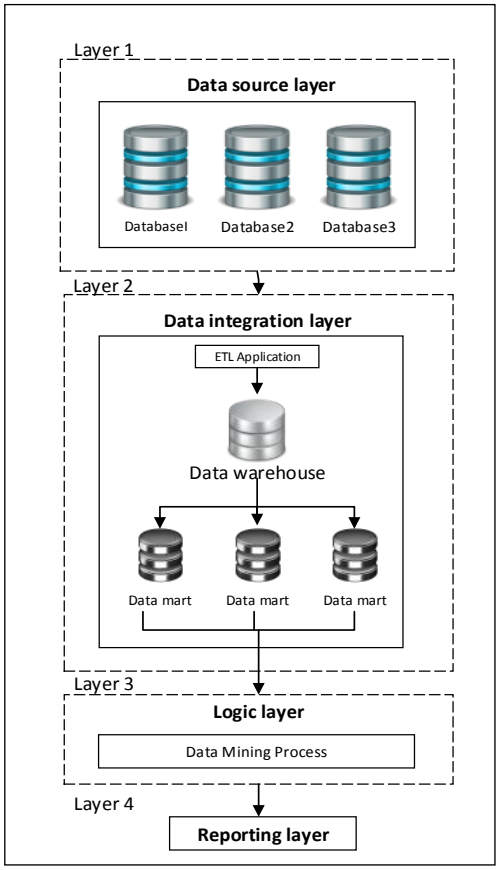


Figure 2 Business intelligence process accordance to layers

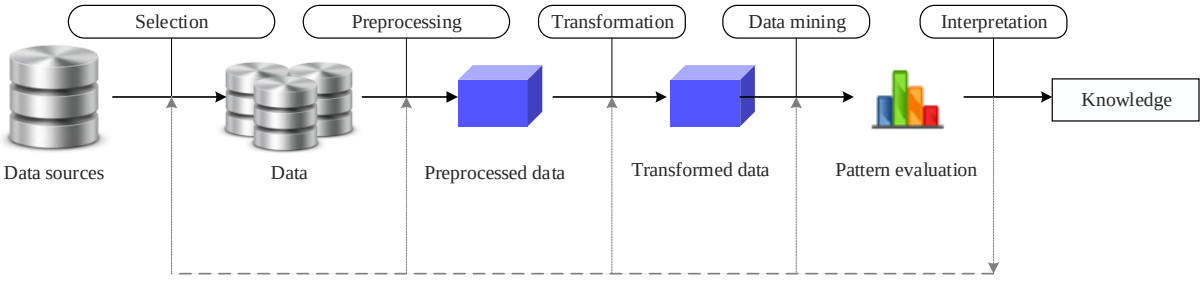


Figure 1 Steps to create knowledge using data mining [8]

4.0 DATA MINING PROCESS

Basically, data mining technology uses data to automatically develop a BI model specialized for the organization [11]. Data mining is to identify nuggets of information or decision making knowledge in bodies of data, and extracting it in a way can be put into decision support, prediction, forecasting, and estimation [12-14]. Data mining method, the algorithms are divided into two basic groups:

(1) Unsupervised algorithms: Clustering and association rules
(2) Supervised algorithms: Classification (For example, decision trees, induction rules, Bayesian networks, and neural networks)

To conclude the context of unsupervised learning is mainly to discover the patterns of data without involving any supervision of data. The main vision of this algorithm is to uncover the patterns of dataset of the input attributes. Looking upon supervised algorithms, it is suitable to be used as the basis of to classify to which the data set will belong [13]. Hence, classification is the best method suits under this case study. The main goal of classification is classify the data based on given data based on specific characterizes. The selected classification techniques for this article are used to discover the best way is to identify suitable data mining which is to allow decision making on GOT.

4.1 Classification Method

Different classifiers have been employed under this case study to discover the best way is to identify suitable data mining which is to allow decision making on GOT whom achieve GOT and Not GOT. To obtain accuracy rate, the calculation of accuracy and sensitivity, Equation 1 and 2 as follow [15]:

$$Accuracy = \frac{True\ Positive + False\ Negative}{Number\ of\ PG} \quad (1)$$

$$Sensitivity = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (2)$$

As stated by Cristina [15], true positive defined as positive cases that has been predicted, false negative defined as positive cases that has been incorrectly classified. Receiver operator curve (ROC) used in evaluating the prediction results and the datasets classified correctly [15]. Brief explanation of classifiers as follows:

4.1.1 J48 Decision Tree (DT)

J48 classifier is a C4.5 decision tree for classification. It is a decision tree approach. J48, basically, ignores the missing values and its basic idea to divide data into range based on the attributes [16]. In this context of PG dataset, J48 maximize true positive rate rather than achieving higher classification accuracy.

4.1.2 Support Vector Machine (SVM)

Support Vector Machine (SVM) is an approach that offer to separate ambiguous data from real data [17]. The performance of the classification using SVM classifiers produce high accuracy rate.

4.1.3 Naïve Bayes (NB)

Naïve Bayes (NB) is a probabilistic classifier that rarely holds true in real world application [16]. NB also simplifies problems relying on assumptions of familiar classification and discover probabilistic knowledge.

4.1.4 Multilayer Perceptron (MLP)

Multilayer Perceptron (MLP) is a part of neural network classifier and it is suitable for approximating a classification function. MLP is a classifier that uses back propagation to classify instances [18].

4.2 Dataset of Case Study

In this section, the section express the list of data sets that are used in case study. As per the input to the model 13 attributes are used. Table 2 shows the list of attributes, type of data and missing rate of data on each attribute. The attributes comprise on numeric and nominal data only.

Table 2 Description of attributes

Attributes	Type	Missing value (%)
Student ID	Numeric	0
Faculty	Nominal	0
CGPA Bachelor	Numeric	0
CGPA Masters	Numeric	0
Study Mode	Nominal	0
Discipline	Nominal	0
Sponsorship	Nominal	0
Gender	Nominal	0
Citizenship	Nominal	0
Enrolment Year	Numeric	0
No. of Semester	Numeric	0
Intake CGPA>=3.00 (Yes or No)	Nominal	0

5.0 EXPERIMENTAL RESULTS

The classifiers J48, NB, MLP and SVM have been applied on the PG datasets. This is mainly to identify suitable data mining which is to allow decision making on GOT.

Comparing with work from the research officers via interviews, approximately will take 3-5 days to undergo pre-processing data and process the data into identify GOT postgraduates. In addition, they have highlighted, there are possibilities of errors and required to re-do the work process to obtain accurate decision making. In Table 3 shows the evaluated datasets with correctly classified GOT postgraduates and time taken to identify the correct GOT students. Comparing with other

classifier, SVM correctly classified higher number of GOT postgraduates.

Table 3 GOT classification by classifiers

GOT Classification	NB	MLP	J48	SVM
Correctly classified	115	106	119	120
Incorrectly classified	6	15	2	1

A very good result was achieved as NB, 95%, J48 DT 98.3% and SVM, 99.6% as highlighted in Table 4. Table 4 represent performance measures by each data mining (classification) approach.

Table 4 Performance measures

Classifiers	Accuracy	Sensitivity	ROC
NB	0.95	0.95	0.98
J48 DT	0.98	0.99	0.97
SVM	0.99	0.99	0.99
MLP	0.87	0.876	0.885

Hence, by implementing data mining in business intelligence will assist in within seconds and real-time processing allow high accurate decision making. The question arises here which classifier suitable to be implemented. The proposed best classifier method to be in business intelligence framework is Support Vector Machine (SVM) as proven with 99% accuracy and low number of incorrect classified of GOT postgraduates.

6.0 CONCLUSIONS AND FUTURE WORKS

This paper presents data mining approach in business intelligence that uses different type of classification method at Layer 3, the logic layer. Based on our experiment, SVM shows the best accuracy (99%) comparing to NB (95%), J48 (98%), and MLP (87%) in obtaining postgraduates who completed their study within specified duration. For future study, the research will be extended by embedding feature selection into BI system for obtaining the best features that influence GOT results in higher education sector. The main importance of pursuing this research to next step is to visualize the GOT postgraduates and analyze the performance of postgraduates for being graduates on time. SVM classifier proved to provide much broader perspective toward obtaining the accurate prediction.

Acknowledgement

This research is fully supported by Research University Grant (RUG) under the Vote No. 02G87 and Fundamental Research Grant Scheme (FRGS) under the Vote No. 4F783. The authors fully acknowledged the Ministry of Higher Education (MOHE) and Universiti Teknologi Malaysia (UTM) for the approved fund which makes this important research viable and effective.

References

[1] Shade, O., K., Goga, P., Awodele, P., Okolie, D. 2013. Framework of Intelligent Recommendation System for a Private Tertiary Institution in Nigeria. *Framework of Intelligent Recommendation System for a Private Tertiary Institution in Nigeria, Volume 3*(4): 1-9.

[2] Gartner. 2009. Gartner EXP Worldwide Survey of More than 1,500 CIOs Shows IT Spending to Be Flat in 2009. Gartner

[3] Gartner. 2011. Gartner Says Worldwide Business Intelligence, Analytics and Performance Management Software Market Surpassed the $10 Billion Mark in 2010. Gartner

[4] Fitriana, R., Eriyatno, Djatna, T. 2011. Progress in Business Intelligence System research: A literature Review. *International Journal of Basic and Applied Sciences IJBAS-IJENS.* 11(03): 96-105.

[5] Alnoukari, M. 2009. Arab International University Case Study. *Using Business Intelligence Solutions for Achieving Organization's Strategy.* 1(2): 11-14.

[6] Baradwaj, B., & Pal, S. 2012. Mining Educational Data To Analyze Student's Performance. *International Journal On Advanced Computer Science and Applications.* 2(6): 63–69.

[7] Kumari, N. 2013. Business Intelligence In A Nutshell, 1(4): 969–975.

[8] Guster, D. and Brown, C. G. 2012. The Application Of Business Intelligence Higher Education: Technical And Managerial Perspectives. *Journal of Information Technology*. 23: 42-62.

[9] Beckett, T. and McComb, B. E. 2012. Increase Enrollment, Retention and Student Sucess: Best Practices for Information Delivery and Strategic Alignment. We*Focus, Ed., New York: Information Builders*. 1-34.

[10] Shah, K. N., Patel, M. R., Trivedi, N. V, Gadariya, P. N., Shah, R. H., Adhvaryu, M. N., & Review, A. L. 2015. Study of Data Mining in Higher Education-A Review. 6(1): 455–458.

[11] Buyetendjik, F., and Tepanier, L. 2010. Predictive Analytics : Bringing The Tools To The Data, (September). 1 –14.

[12] Aquila, C. D., Tria, F. D. I., Lefons, E., Tangorra, F., Informatica, D., and Bari, U. 2008. Evaluating Business Intelligence Platforms : A Case Study. *Integration The Vlsi Journal.* 558–564.

[13] Azma, F., and Mostafapour, M. A. 2012. Business intelligence As A Key Strategy For Development Organizations. Procedia Technology, 1: 102–106. doi:10.1016/j.protcy.2012.02.020.

[14] Khan, R., & Quadri, S. 2012. Business Intelligence: An Integrated Approach. *Business Intelligence Journal.* 5(1): 64–70.

[15] Oprea, C., & Ti, P. 2014. Performance Evaluation of The Data and Mining Classification Methods. 249–253.

[16] Patil, T. R., & Sherekar, S. S. 2013. Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification. *International Journal Of Computer Science And Applications*. ISSN: 0974-1011, 6(2): 256–261.

[17] Shaviata, Walia, A. 2014. *International Journal of Advanced Research in Computer Science and Software Engineering*. 4(5): 442–458.

[18] Devasena, L. C. 2014. Efficiency Comparison of Multilayer Perceptron and SMO Classifier for Credit Risk Prediction. 3(4): 6155–6162.