International Journal of Research Studies in Computer Science and Engineering (IJRSCSE)
Volume 1, Issue 1, May 2014, PP 29-35
www.arcjournals.org

Holistic Prediction of Student Attrition in Higher Learning Institutions in Malaysia Using Support Vector Machine Model

Anbuselvan Sangodiah
Faculty of Information and Communication Technology,
Department of Information System
Universiti Tunku Abdul Rahman
Kampar, Malaysia
anbuselvan@utar.edu.my

Balamuralithara Balakrishnan
Faculty of Art, Computing and Creative Industry
Universiti Pendidikan Sultan Idris
Tanjung Malim, Malaysia
balab@fskik.upsi.edu.my

Abstract: *Attrition or better known as student dismissal or drop out from completing courses in higher learning institutions is prevalent in higher learning institutions in Malaysia and abroad.There are several reasons attributed to the attrition in the context of student in higher learning institutions. The degree of attrition varies from one institution to another and it is cause for concern as there will be a lot of wastage of resources of academic and administrative besides the adverse effect on the social aspect. In view of this, minimizing the attrition rate is of paramount importance in institutions.There have been numerous non technical approaches to address the issue, but they have not been effective to predict at early stage the likelihood of students dropping out from higher learning institutions. Technical approach such as data mining has been used in predicting student attritionby some researchers in their past research work.However, not all prediction data mining techniques and other relevant and significant factors attributed to student attrition have been fully explored to address the issue. As of result this, this paper will focus on using support vector machine model to predict student attrition. Itwillalso examinerelevant and other factors that contribute to the attrition among students in Malaysia. With all these in place, a model with high accuracy in predicting student attrition is expected to achieve.*

Keywords: *Attrition, Data Mining, Support Vector Machine, Classification model*

1. Introduction

Attrition becomes a normal scenario in many higher education institutions (HEIs) all over the world. The factors that lead to attrition among undergraduates and postgraduates in HEIs are varied from one to another taking into the consideration of geographical factors, ethnicity, education system of a country and etc*[1]*. Problem of attrition has been studied since 1980's where *[2]* and *[3]* have expressed the concern on attrition rate among students in HEIs. There are many studies have been carried out to find out the reasons which lead to attrition and also many investigations have been conducted focusing on overcome the attrition problem through various approaches including intervention strategies *[4]*,*[5]*.

The attrition rate among Malaysian students in both public and private HEIs is contributed by many factors. In private HEIs, most of attritions cases are due to financial issues since the tuition fee is very expensive, lack of facilities provided by the management and quality of teaching *[6]*. While the majority of students leave public HEIs because of two major reasons; (i) not able to continue their study in the programme that they enrolled due to lack of interest and (ii) failed in the examination *[6]*. These factors, which related to attrition, need to be handled carefully in order to solve the issue of attrition. Many strategies have been put forward to encounter this issue which mainly focuses on non-technical approaches that did not have much impact to solve the problems that lead to attrition. Therefore, an efficient mechanism using a technical approach needs to be identified and put forward that can guarantee to reduce the attrition rate and improve retention rate. Thus, in this study, the support vector machine model would be used to predictattrition rates. The reason that the model is chosen is due to its support for high dimensional space and sparse *[7]*. Besides that, it eliminates the need for feature selection hence making, it is easier to use. Also it is known for its superior performance compared to other data mining models in solving pattern recognition problem *[8]*.

Every university and colleges have a strong aim to provide students different backgrounds with a conducive learning experience that could lead students to achieve success completing their study earning certificates, diploma, degree or postgraduate degrees. However, the aim could not reach every student where some of them might fail to complete their study due to many reasons. Attrition is disadvantageous for both students and HEIs. For students, fail to earn a degree will hamper their efforts to improve the socioeconomic status while for HEIs, attrition creates a major financial problem due to loss of income through tuition fees from students.

In Malaysia, the situation of students dropped out in their tertiary level education is alarming where, according to the latest statistics; out of 168000 college students who pursue their studies for certificates and diploma, 30000 would not graduate while , out of 100000 students who went for their degree program me, only 83000 able to finish the program me*[9]*. It means that 17.5% of total students who enrolled in tertiary education have dropped out in Malaysia. *[10]* Reported that a private university in Malaysia has an attrition rate exceeding 14% in just 6 months in the year 2012.

Universities outside Malaysia are also facing the attrition problem among the students. Griffith University in UK has an attrition rate of 21.2% in year 2011 *[11]* while*[12]* reported that in the year 2007, out of 32 Australian universities there were about average of 10.5% dropped out in each university *[12]*.The rate of attrition varies according to the field of study, in the US, students who enrolled in *[13]* related programs either change to another program or leaving colleges/universities without completing their study (48% - bachelor degree – and 69% - postgraduate level) between 2003 and 2009. In non STEM programmes, both undergraduate and postgraduate suffered with an attrition rate of 56% between 2003-2009 *[13]*.*[14]*found that about 57% of doctoral students across disciplines have to leave HEIs without completing their degree. *[15]* Cited that in the US, the attrition rate among PhD students is 50%.

HEIs all over the world are facing challenges in retaining number of students in their respective institutions where there is a need to identify factors that influencing students to leave higher education institutions before graduating. There are many factors from different aspects contribute to attrition; institutional environment, personal and financial problems are among the major reasons for students' attrition. *[11]* Indicated the following reasons for students' attrition which can be seen as universal reasons.

i. Personal difficulties – the most commonly given explanation for attrition, relating tohealth, finances, family, work, and difficulty fitting in or making friends.

ii. Academic difficulties – lack of academic preparedness, weak academic knowledge orspecific study skills required to tackle the demands of the program; weak academic entryscores and low GPAs in first semester are all associated with greater attrition.

iii. Full time vs part-time status – part-time students are significantly less likely to continueinto second year compared to full-time students.

iv. Making an uncertain or the wrong subject/program/university choice is linked toattrition. In some cases, this may reflect poor information provided prior to enrolment,or inadequate consideration of educational and career goals

v. Not being the University of First Choice – a proportion of students leave one university totake up a more attractive opportunity at another institution if they are able to

vi. Loss of interest in the program or subject area

vii. Inability to manage time and workload demands and in consequence falling behind

viii. Dissatisfaction with the university experience, quality of curriculum or teaching." (pg.2)

Those numbers and reasons pertaining to the attrition rate among the students in HEIs showed that there is a need for a good mechanism that need to be developed and adopted by HEIs to overcome the problem of attrition in which this mechanism could hinder attrition from every angle of factors that contribute to the rise of number of dropped out. Thus, in this investigation; a holistic prediction of student attrition in HEIs in Malaysia using support vector machine model has been proposed in which it focuses on the feasibility of the usage of the system in mitigating attrition.

2. Literature Review

Data mining (DM) is a computer-based information system (CBIS) *[16]* devoted to scan huge data repositories, generate information, and discover knowledge. The meaning of the traditional mining term biases the DM grounds. But, instead of searching natural minerals, the target is knowledge.DM pursues to find out data patterns, organize information of hidden relationships, structure association rules, estimate unknown items' values to classify objects, compose clusters of homogenous objects, and unveil many kinds of findings that are not easily produced by a classic CBIS. Thereby, DM outcomes represent a valuable support for decision-making.

Concerning education, it is a novel DM application target for knowledge discovery, decisions-making, and recommendation *[17]*. Nowadays, the use of DM in the education arena is incipient and gives birth to the educational data mining (EDM) research field *[18]*.

There are several data mining techniques or methods, including generalization, characterization, classification; clustering, association, evolution, pattern matching, data visualization and meta-rule guided mining *[19]*. As this study focuses on prediction through classification method, generally there are several models that are popular such as decision tree, neural network, support vector machine, Naïve Bayes and decision rule *[20]*. However in this study, support vector machine is used over other models to predict student attrition and the justification for using it in this study has been mentioned earlier.

There have been some past research work in predicting student attrition. A case study has been performed to predict electrical engineering students drop out in the Department of Electrical Engineering Eindhoven University of Technology. The experimental results show that rather simple and intuitive classifiers (decision trees) give a useful result with accuracies between 75 and 80%. *[21]*. certainly the research work is focused on decision tree model.

Similar research work has also been conducted at University of Science and Technology in Iran *[22]* where its focused classification models were regression and decision tree. An accuracy of 88.5% has been recorded considering selected academic and nonacademic factors in predicting student attrition.

The work by *[23]* uses three classification models which are logistic, decision tree and neural network in accordance to SEMMA methodology. Despite this research work primarily focused on comparing the accuracy between models in the context of student attrition, support vector machine model was not explored. Feature selection to determine contributing factors to student attrition was focused on non academic factors.

The importance of student dropout prediction using classification model is also evident in the research work by *[24]*. The work uses decision tree model to identify most contributing factors to student attrition and subsequently use those factors for prediction. An accuracy of 38.10% was recorded.

Other related work uses *[25]* prediction model particularly, artificial neural networks to determine the pass and fail cases on different educational majors at the end of the high school. Besides that, classification models have also been used to identify demographic, educational, and economic factors associated with atypically long time between doctoral program admission and degree completion *[26]*.

The use of classification models such as neural network, decision tree and Naïve Bayes has also been popular in the education field to predict students' behavior as evidenced in these research work *[27]*.In a nutshell, the objective of this study is to focus on using support vector machine model with high number of vectors or predictors to observe the accuracy of student attrition.

3. Methodology

As data mining is a process of discovering various models, summaries, and derived values from a given collection of data *[28]*. It is only appropriate that data that needs to be analysed using data mining models should be accurate and reliable. The degree of accuracy of results obtained from data mining models is directly dependent on the accuracy of data, thus a data mining process or methodology is necessary. There are several well known data mining methodologies and for this study Knowledge Discovery from Data (KDD) methodology is chosen. KDD is a process that provides, in such way, this knowledge. KDD *[29],[30]*is an interactive and iterative process *[31]*. It is a multi stages process. According to *[29]* KDD process has 8 phases: problem formulation, data retrieval, data selection, data cleaning, data transformation, data mining, patterns evaluation, and knowledge integration and use as shown in Figure 1.

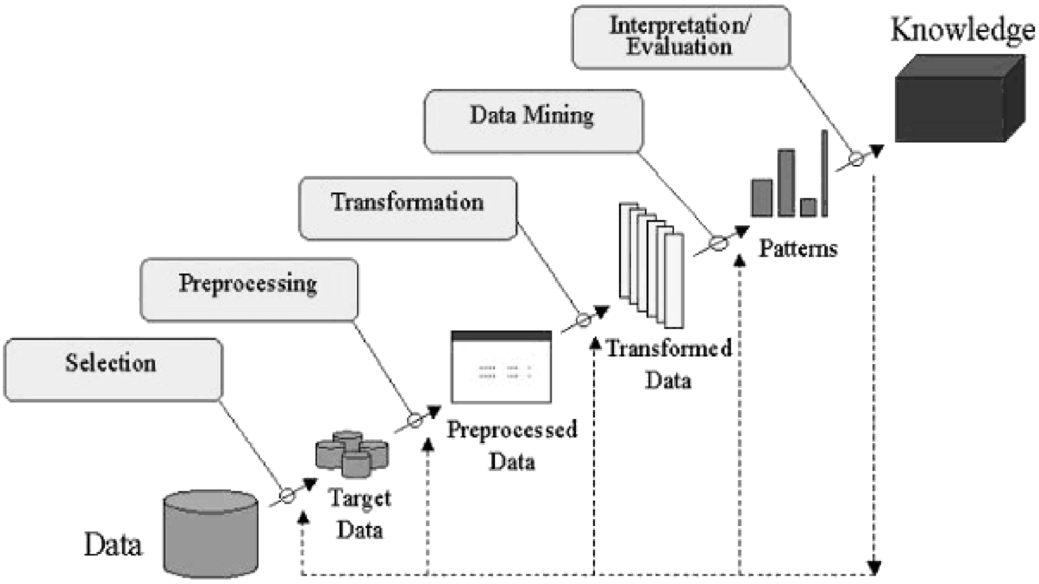


Fig 1

In this proposed study, selection phase will involve in selecting relevant and important factors or attributes attributed to the student attrition. Certainly a high number of vectors or attributes will be tested in this study. The selected attributes are expected to be non academic and the following proposed attributes will be considered.

Proposed attributes: gender, age, location/hometown, parent’s occupation, parent’s income, health status, media social involvement, choice of programme, status of study, type of prior knowledge and other attributes.

Later in the preprocessing and transforming stages, there are inherent challengesin this study as the support vector model only supports input or data in the form of numerical and binary and most of the selected attributes are categorical based data.These attributes whose values are category based must be converted into numeric or binary forms before they can be analysed using support vector machine.

A simple technique can be deployed to convert a categorical attribute into binary form*[32]*. If there are m categorical values, then each original value can be uniquely assignmend to an integer n he interval$[0,\quad m-1]$. Thereafter, the each m integers can be converted to a binary number using n =$\log_2(m)$.

Example: parent’s occupation attribute; assume that there are three possible category values which are professional, labour, others.

Categorical value	Integer value	x_1	x_2
Professional	0	0	0
Labour	1	0	1
others	2	1	0

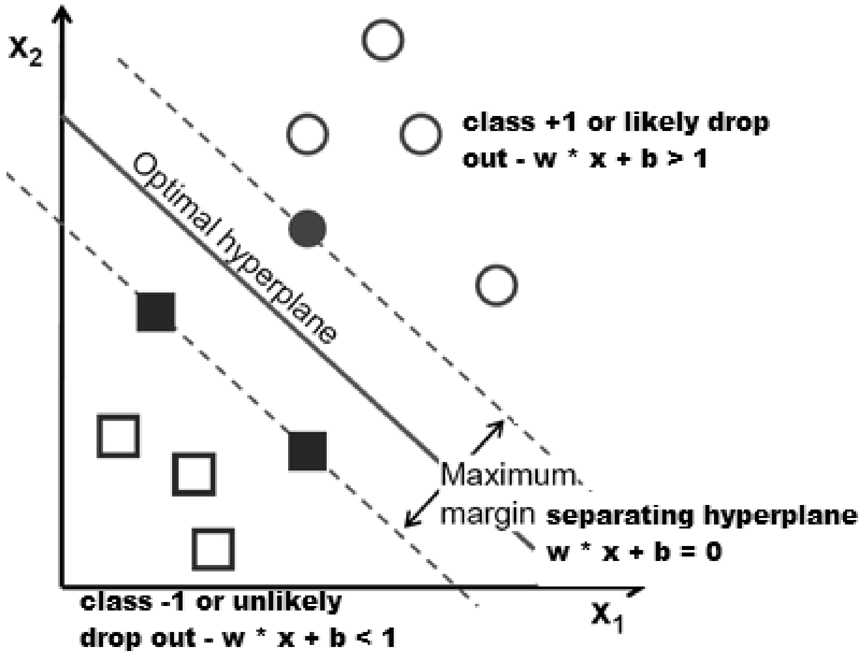
Besides binarization process as shown above, some attributes particularly income attribute though its value is expressed in numerical form but it has the potential to influence the final outcome as large values will take precedence over other attributes. Therefore, such attribute, its value will be normalized or scaled to prevent bias over other attributes.

Example: income attribute; assume the value is 20,000 and it will be normalized to a value between 0 to 1.

The next phase is about selecting a classification model. In this study, support vcctor machine model is chosen. Support Vector Machines, which were developed by *[33]*,performbinaryclassification,i.e.,separateasetoftrainingvectorsfortwo different classes$(x_1,y_1),(x_2,y_2),\ldots,(x_m,y_m)$, where $x_i \in R^d$denotes vectors in a d-dimensional feature space and $y_i \in \{-1,+1\}$is a class label. In this study, there will be two class labels which are likely drop out and unlikely drop out. The strength of SVM lies in decision boundary of a linear classifier which separate class labels and it can be written in the following form:

$W\cdot x + b = 0$ where w and b are parameters of the model.

After putting some training data through SVM, the following graph will be observed as shown in Figure 2.


X2
class +1 or likely drop out - w * x + b > 1
Optimal hyperplane
Maximum margin
separating hyperplane w * x + b = 0
class -1 or unlikely drop out - w * x + b < 1
X1

Once the parameters of the decision boundary are found, a test instance z is classified as follows:

$f(\mathbf{z}) = sign\ (\mathrm{w} \cdot \mathbf{z} + \mathrm{b}) = sign\ (\sum_{i=1}^{N} \lambda_i\ \ y_i x_i \cdot \mathbf{z} + \mathrm{b})$

If $f(\mathbf{z}) > 1$, then the test instance is classified as a positive class or likely drop out class label otherwise, it is classified as a negative class or unlikely drop out class label.

4. Conclusion and Future Work

Apparently, this paper identifies the potential use of classification model which is support vector machine to predict student attorition in a holistic and accurate manner.

Future work would involve in coming up training and test data to measure the accuracy of this model in predicating student attrition compared to other classification models to which past research work in the context of student attrition have used. Measuring the accuracy in predicting student attrition is vital before it can be put into practice in the real world context.

References

[1] Veenstra C. P., A strategy for improving freshman college retention. Journal for Quality & Participation, 31, 19-23,(2009).Retrieved from http://asq.org/pub/jqp/

[2] Oscar T. L., Philip E. B., Ken S., Retention and attrition: Evidence for action and research, Boulder, Colorado: NCHEMS, (1980).

[3] Stadtman V. A., Academic adaptations: Higher education prepares for the 1980s and 1990s. San Francisco, California: Jossey-Bass, Inc, (1980).

[4] Traci S., and Stefanie K., The best laid Plans: Examining the Conditions Under which a planning Intervention Improves learning and reduces attrition. Journal of AppliedPsychology, doi: 10.1037/a0027977, (April 2012).

[5] Wilson K., (2009). The impact of institutional, programmatic and personal interventions on an effective and sustainable first-year student experience. In 12th First Year in Higher Education Conference 2009, Townsville. (2009). Retrieved from http://www.fyhe.com.au/pat_papers/ papers09/ ppts/Keithia_Wilson_paper.pdf

[6] Lesley W., Julie C., and Sally J., Beyond the first-year experiences: the impact of attrition ofstudent experience throughout undergraduate degree studies in six diverse universities. Innovations in Education and Teaching International, 36(3), 331-352. (2011).

[7] Joachims T., Text categorization with support vector machines: Learning with many relevant features. In Proceedings of the Tenth European Conference on Machine Learning (ECML '98), Lecture Notes in Computer Science, Number 1398, 137–142, (1998).

[8] Lewis D. L., Yang Y., Rose T. G., and Li F., RCV1: A new benchmark collection for text categorization research. Journal of Machine Learning Research, 5: 361-397, (2004).

[9] Borneo Post, Staggering dropout rate before SPM (September 26, 2012). Retrieved from http://www.theborneopost.com/2012/09/26/staggering-dropout-rate-before-spm/

[10] Govindarajo N. S., and Kumar D., How to Combat Attrition? Case Study on a Malaysian Educational Institution, (2012). Retrieved from http://cprenet.com/uploads/archive/IJBBS_12-1167.pdf

[11] Griffth University, Operation Student Success: Griffith's Student Retention Strategy 2012 – 2014,(2012).Retrievedfromhttp://www.griffith.edu.au/__data/assets/pdf_file/0006/419469/Student-Retention-Strategy.pdf

[12] AUIDF, Retention and Attrition in Australian Universities (2007), Retrieved from http://www.spre.com.au/download/AUIDFRetentionResultsFindings.pdf

[13] STEM Attrition: College Students' Paths Into and Out of STEM Fields (2013). Retrieved from http://nces.ed.gov/pubs2014/2014001rev.pdf

[14] Gardner, S. K., Student and faculty attributions of attrition in high and low-completing doctoral programs in the United States. Higher Education, 58, 97-112. doi:10.1007/s10734-008-9184-7, (2009).

[15] Petroff L., The politics of graduate school. National Social Science Journal, 35, 133-138.(2011).Retrieved from http://www.nssa.us/journals.htm

[16] Vlahos G. E., Ferratt T. W., and Knoepfle G., The use of computer-based information systems by German managers to support decision making. Journal of Information & Management, 41(6), 763–779, (2004).

[17] Vialardi-Sacin C., Bravo-Agapito J., Shafti L., and Ortigosa A., Recommendation in higher education using data mining techniques. In Proceedings of the 2nd international conference on educational data mining (pp. 190–199), (2009).

[18] Anjewierden A., Kolloffel B., and Hulshof C., Towards educational data mining: using data mining methods for automated chat analysis to understand and support inquiry learning processes. In Proceedings of the international workshop on applying data mining in e-Learning (pp. 23–32), (2007).

[19] Liao S. H., Chu P. H., Hsiao P. Y., Data mining techniques and applications – A decade review from 2000 to 2011.Journal of Expert Systems with Applications 2:11304, (2012).

[20] Alejandro P. A., Educational data mining: A survey and a data mining-based analysis of recent works. Journal of Expert Systems with Applications pg 1432, (2013).

[21] Gerben W. D., Mykola P., Jan M.V., Predicting Students Drop Out: A Case Study. 2ndInternational Conference on Educational Data Mining Proceedings Cordoba, Spain, (2009).

[22] Mahdi N., Behrouz M., Fereydoon V., Predicting GPA and Academic Dismissal in LMS Using Educational Data Mining: A Case Mining. 6th National and 3rd International conference of e-Learning and e-Teaching (ICELET2012), IEEE (2012).

[23] Mario J., Željko G., Maja M., Student Dropout Analysis with Application of Data Mining Methods. Management, Vol. 15, 1, pp. 31-46, (2010).

[24] Mohammad N. M., Linkon C., Md S. K. Students Dropout Prediction for Intelligent System

from Tertiary Level in Developing Country. IEEE/OSAIIAPR International Conference on Informatics, Electronics & Vision, IEEE, (2012).

[25] Sayyed M. M., Sayyede A. A., Prediction of Success or Fail of Students on Different Educational Majors at the End of the High School with Artificial Neural Networks Methods. International Journal of Innovation, Management and Technology, Vol. 4, No. 5, (2013).

[26] Gayle, M., Data Mining Analysis of the Effect of Educational, Demographic, and Economic Factors on Time from Doctoral Program Entry to Degree Completion in Education, Ph.D Thesis The Florida State University College Of Education, (2006).

[27] Vasile P. B., Analysis and Predictions on Students' Behavior Using Decision Trees in Weka Environment. Proceedings of the ITI 2007 29th Int. Conf. on Information Technology Interfaces, IEEE, (2007).

[28] Mehmed, K. Data Mining: Concepts, Models, Methods and Algorithms Second Edition. Wiley, (2011).

[29] Fayyad U., Piatetsky-Shapiro G., Smyth P., From Data Mining to Knowledge Discovery Databases. In: Advances in knowledge discovery and data mining. American Association for Artificial Intelligence, ISBN: 0-262-56097-6, pp. 1 – 34, (1996).

[30] CRISP-DM, Cross Industry Standard Process for Data Mining, http://www.crisp-dm.org, 2003.

[31] BenAyed M., Ltifi H., Kolski C., Alimi A. M., A User-centered Approach for the Design and Implementation of KDD-based DSS.DSS: A case Study in the Healthcare Domain, Decision Support Systems, 2010, Vol. 50, pp.64–78. (2010).

[32] Pang-Ning T., Michael S., Vipin K., Introduction to Data mining. Pearson, (2006).

[33] Vapnik V. N., Statistical learning theory. New York: John Wiley and Sons, (1998).

AUTHORS' BIOGRAPHY

Mr. Anbuselvan Sangodiah has done Bsc (Hons) from UPM. He has done Msc in Information Technology from UPM in the year of 2000. He has vast experience of almost 15 years in teaching in various renowned universities in Malaysia. He is also has been a trainer for 6 years in the area of Information Technology where he has conducted various IT training courses for various renowned companies and universities. He is a certified professional and expert in Java and also certified of Microsoft Server. Besides that, he is also a certified trainer of Ministry of Human Resource. He is a member of MNCC (Malaysia National Computer Confederation) and ACM (Association for Computing Machinery). He is also actively involved in research work where he has published a few conference papers in the area of information system. His areas of interest are data mining, text mining, software development and database. At present he is working as lecturer in FICT in UTAR.

Balamuralithara Balakrishnan has received his PhD in year 2011 in the area of Creative Multimedia. Currently, he is attached with Universiti Pendidikan Sultan Idris as Senior Lecturer. His research interests are in Educational Research and Engineering Education.