Investigation of Factors for Students' Decisions to Study at Phetchabun Rajabhat University by Using Data Mining Techniques

การศึกษาปัจจัยที่มีผลต่อการตัดสินใจเข้าศึกษาที่มหาวิทยาลัยราชภัฏเพชรบูรณ์ด้วยเทคนิคเหมืองข้อมูล

Jetsadaporn Pakamwang[1], Kan Khoomsab[1]and Kriengkri Timsorn[1]
เจษฎาพร ปาคำวัง[1], กาญจน์ คุ้มทรัพย์[1] และ เกรียงไกร ทิมศร[1]
[1]Faculty of Science and Technology, Phetchabun Rajabhat University
[1] คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏเพชรบูรณ์
Received: January 14, 2019
Revised: April 20, 2019
Accepted: April 26, 2019

Abstract

This study presents the use of data mining for investigation of important factors that affect students' decisions for studying at Phetchabun Rajabhat University. 400 students, including 200 freshmen of Phetchabun Rajabhat University and 200 other freshmen from the other universities, were studied based on questionnaires that had 14 attributes. Score values were collected from the attributes and input data were first analyzed for data classification by ANN and decision tree techniques. 10-fold cross validation and RMSE values were also employed to evaluate classification accuracy. Then, factor identification based on filter ranking method was investigated. Its results were explained by decision tree visualization and IF-THEN rules. The results showed that 400 students were correctly classified into two groups (1) freshmen group from Phetchabun Rajabhat University and (2) freshmen group from the other universities. The classification accuracy of ANN and decision tree was 93.00% and 88.25%, respectively. RMSE (root-mean-square deviation) values were 0.2574 and 0.3031, respectively. Based on filter ranking method, the identified factors affecting students' decisions were ranked as university brand recognition, family income, and number of majors, university location, parents' careers and modern curriculum, respectively. From these results, data mining techniques with ANN and decision tree are useful for data processing and they provide meaningful information for university admission system analysis as well as other applications.

Keywords: data mining, artificial neural network, decision tree, attributes selection, data science

บทคัดย่อ

งานวิจัยนี้ศึกษาปัจจัยที่มีผลต่อการตัดสินใจเข้าศึกษาต่อที่มหาวิทยาลัยราชภัฏเพชรบูรณ์ของนักศึกษาด้วยเทคนิคเหมืองข้อมูล โดยได้ทำการเก็บข้อมูลจากแบบสอบถาม 14 คุณลักษณะจากนักศึกษาจำนวน 400 คน แบ่งเป็นนักศึกษาใหม่ของมหาวิทยาลัยราชภัฏเพชรบูรณ์ จำนวน 200 คน และอีก 200 คน เป็นนักศึกษาใหม่จากสถาบันอื่น ข้อมูลที่ได้จะถูกนำมาวิเคราะห์ความแตกต่างของคุณลักษณะเบื้องต้น โดยใช้เทคนิคโครงข่ายประสาทเทียม (Artificial Neural Network--ANN) และเทคนิคต้นไม้ตัดสินใจ (decision tree) เพื่อจำแนกข้อมูล ซึ่งมีวิธีการวัดความแม่นยำในการจำแนกจากค่าประสิทธิภาพของโมเดลด้วยการตรวจสอบไขว้จากการแบ่งข้อมูลออกเป็น 10 ส่วน (10-fold cross validation) และค่า RMSE (Root Mean Squared Error--RMSE) หลังจากได้ผลการวิเคราะห์เบื้องต้นแล้ว ในงานวิจัยนี้จะใช้เทคนิคการคัดเลือกคุณลักษณะของข้อมูลด้วย filter ranker method แสดงผลและอธิบายด้วยแผนภาพต้นไม้ตัดสินใจและกฎ IF-THEN ผลการวิจัยพบว่า การจำแนกนักศึกษาออกเป็น 2 กลุ่ม คือ (1) กลุ่มนักศึกษาใหม่ของมหาวิทยาลัยราชภัฏเพชรบูรณ์ (2) กลุ่มนักศึกษาใหม่จากสถาบันอื่น ด้วยเทคนิคโครงข่ายประสาทเทียมและเทคนิคต้นไม้ตัดสินใจ มีค่าความแม่นยำเป็น 93.00% และ 88.25% ตามลำดับ และมีค่า RMSE เป็น 0.2574 และ 0.3031 ตามลำดับ คุณลักษณะที่เป็นปัจจัยสำคัญที่มีผลต่อการตัดสินใจของนักศึกษาได้แก่ ชื่อเสียงของมหาวิทยาลัย รายได้ของผู้ปกครอง จำนวนหลักสูตร ที่ตั้งของมหาวิทยาลัย อาชีพของผู้ปกครอง และความทันสมัยของหลักสูตร ตามลำดับ ผลการวิจัยนี้แสดงให้เห็นว่าเทคนิคเหมืองข้อมูลมีประโยชน์ในการจัดการข้อมูลและให้ข้อมูลที่มีความหมายที่สามารถใช้เป็นข้อมูลในการพิจารณาหาวิธีการเพิ่มจำนวนนักศึกษาของมหาวิทยาลัยได้

คำสำคัญ: เหมืองข้อมูล, โครงข่ายประสาทเทียม, ต้นไม้ตัดสินใจ, การเลือกคุณลักษณะ, วิทยาศาสตร์ข้อมูล

Introduction

Nowadays, one of the major problems for a group of Rajabhat University is mainly related to a decrease in a number of freshmen (Ippoodom, 2017). It needs to quickly solve the problem for Rajabhat universities. Different factors that affect high school student's decision to study in universities such as interesting and various curriculums, university locations, facilities as well as student's family background should be investigated for understanding of student's decision. Knowing these factors is very useful for a group of Rajabhat University to improve admission system and increase a number of freshmen, especially for high school students in locals. Data mining has increased great attention in data analysis which includes classification, prediction, clustering, data visualization and pattern recognition (Liao et al., 2012). It is the process of analyzing and discovering the patterns of large data sets that contain correlated variables and transforms them into meaningful information (Angeli et al., 2017; Kaur et al., 2015; Natek & Zwilling, 2014; Sen & Ucar, 2012; Ahmed et al., 2017). The methods used for data mining involve machine learning and statistics, e.g., Artificial Neural Networks--ANN, Decision Tree--DT, Principal Component Analysis--PCA and Support Vector Machine--SVM (Jones et al., 2016; Jothi et al., 2015)

These methods have been successfully studied and applied to different applications including education (Kaur et al., 2015; Sen & Ucar, 2012; Sen et al., 2012; Rodrigues et al., 2018), sensor technology (Timsorn et al., 2016; Timsorn et al., 2017) and marketing (Hsu, 2009; Ozyirmidokuz et al., 2015; Packianather et al., 2017) etc. For ANN and decision tree, they show unique data process compared with the other techniques. ANN well performs for a complexity system with many correlated variables and provides a high degree of accuracy (Ahmad, 2017: Schmitz et al., 1999; Deng et al., 2008; Han & Kamber, 2006). Decision tree is very useful for interpretation of relationship between input and output data (Pu et al., 2018; Schmitz et al., 1999; Panto & Theantong, 2014). Sittichat (Sittichat, 2017) successfully studied educational attributes to estimate student's achievement in Calculus I for Engineer course using data mining with ANN and decision tree. Tsai et al. (Tsai et al., 2017) applied statistical analysis and conjoint analysis to study factors which affect international students' decision for selecting universities in Taiwan.

In this study, we propose data mining technique for investigation of important factors that affect student's decision for studying at Phetchabun Rajabhat University. ANN and decision tree methods were used for data classification based on score values of 14 attributes corresponding to information of Phetchabun Rajabhat university and student's family background, which were collected from 400 freshmen from Phetchabun Rajabhat university and other universities. It should be noted that these 14 selected attributes are based on three main factors (Tsai et al., 2017; Phang, 2012) including; (1) Institutional Image and Environment (university fame, university activity, graduate quality, university location, facility, university environment, lecturer quality and university administration), (2) Desired program/course (modern curriculum, curriculum number, employment and student moral) and (3) Influences and recommendation from family (parent career and family income). The important factors were identified using filter ranker method with decision tree.

Research Objective

The objective of this study is to use data mining technique for investigation of important factors that affect student's decision for studying at Phetchabun Rajabhat University.

Materials and Methods

Research plan

The research diagram of this study is presented in Fig. 1. It consists of five steps; (1) Data collection and preparation, (2) Model building, (3) Testing and evaluation, (4) Attribute identification and (5) Results and discussion.

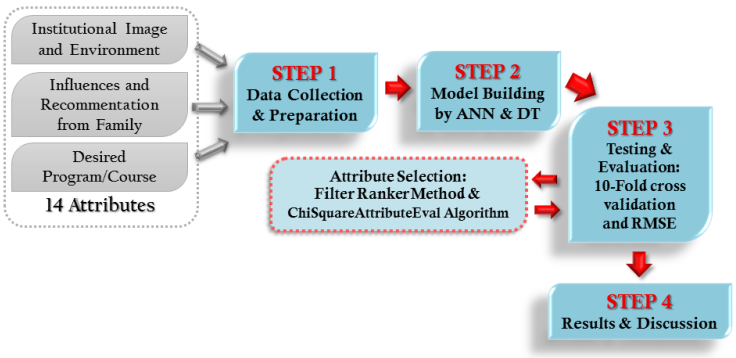

Figure 1 Investigation diagram used in this work.

Data Collection and Preparation

In this step, score values of 14 attributes on questionnaires were collected from 400 undergraduate students in the first semester of 2018, 200 freshmen of Phetchabun Rajabhat University and 200 freshmen from the other universities. 14 attribute details with their score values are shown in Table 1. To cleaning data, the missing data were replaced with mean values of data by using replace missing values option in WEKA. Then, data were normalized and arranged in a M×N matrix format. M and N represent rows of data number (400 student data) and columns of attributes (14 attributes), respectively.

Classification Model Building

For the step, ANN and decision tree methods were employed to build classification models using 3.9.2 version of WEKA program. It should be noted that ANN is a machine learning method based on concept of neural network system of human brain. The ANN purpose is to convert inputs into meaningful outputs. The ANN model for this study was a multilayer perceptron--MLP as shown in Fig. 2. It consists of three main layers; input layer, hidden layer and output layer. Nodes in the input layer receive input data and distribute them into nodes in the hidden layer. Nodes in the hidden layer weight their inputs with the strengths of the respective connections from the input layer and sum them. The outputs of nodes in the hidden layer can be calculated as follows (Deng et al., 2008; Lou & Nakai, 2001):

$$y = \sum_{i=1}^{n} w_i x_i \qquad (1)$$

Where y is an output of each node in the hidden layer, xi is input data and w i is the strengths of respective connections.

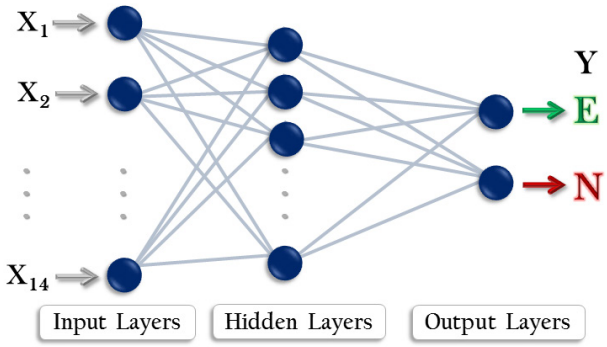

Figure 2 ANN network used for data classification

Note, Adapted from "Evaluation of bacterial population on chicken meats using a briefcase electronic nose". by Timsorn, K., Thoopboochagorn, T., Lertwattanasakul, N., & Wongchoosuk, C., 2016, *Biosystems Engineering, 151*(1), 116-125. ©2016 by IAgrE. Elsevier.

In this study, the number of nodes in the input layer was 14 attributes and the number of nodes in the output layer was 2 classes; E (enroll at Phetchabun Rajabhat university) and N (not enroll).

Decision tree is a classification algorithm that divides a data set into smaller subdivision on the basis of a set of tests for decision making (Pu et al., 2018). Its network resembles a tree including root nodes, internal nodes (branch nodes) and leaf nodes (Pu et al., 2018; Lui et al., 2018). The root nodes include test data samples and the leaf nodes present decision results. The objective of decision tree is to build a model that can classify data samples into the same category with optimal partition attributes (Pu et al., 2018). Moreover, tree representative is easy to understand its classification principles based on IF-THEN rules (Pu et al., 2018; Lui et al., 2018).

Model Evaluation

To evaluate the classification model, 10-fold cross validation and Root Mean Square Error (RMSE) were performed. The 10-fold cross validation is the method that divides the data into 10 equal sets. Nine sets are used as a training set and the other set is used for a testing set (Sittichat, 2017). Then, they are alternatively repeated to cover a series of 10 data sets. RMSE is error measurement between the actual data and the classified data.

Attribute Identification

The factors that affect student's decision for studying at Phetchabun Rajabhat University were identified by feature selection of attributes. There are two techniques widely used for feature selection of attributes namely wrappers and filters (Ramaswami & Bhaskaran, 2009). In this study, filter ranker method with information gain values was employed. It should be noted that ranking of features (attributes) determines the importance of each attribute depending on general characteristics of data (Ramaswami & Bhaskaran, 2009). The identification process was performed using decision tree and the result was interpreted by decision tree visualization and IF-THEN rule.

Table 1

Attributes and their score values

Attribute	Scores
Parent Career	1: Farmer
	2: Sale
	3: Employee
	4: Private Business (such as barber, tailor, mechanic)
	5: Government
Family Income	1: < 10,000
	2: 10,000–20,000
	3: > 20,000
University Fame	1: Strongly Disagree
University Activity	2: Disagree
Graduate Quality	3: Neither Agree/ Nor Disagree
University Location	4: Agree
Facility	5: Strongly Agree
University Environment	
Modern Curriculum	
Employment	
Curriculum Number	
Student Moral	
Lecturer Quality	
University Administrations	

Results and Discussion

1. Classification results

Table 2 shows classification results of data based on 14 attributes collected from 400 freshmen using ANN and decision tree models. From both methods, it is obviously seen that data were classified into two groups; E and N, corresponding to freshmen from Phetchabun Rajabhat University and the other universities. The overall classification accuracy of ANN and decision tree models was 93.00% and 88.25%, respectively. RMSE values of ANN and decision tree models were 0.2574 and 0.3031, respectively. For ANN model, in the E class, 182 freshmen were correctly classified and 190 freshmen were correctly classified for the N class. In case of decision tree model, a correctly classified number of freshmen in both E and N classes were smaller than that of ANN model. The result shows that the ANN model displays better performance than that of decision tree model for classifying two groups of data samples. Unfortunately, the ANN

is difficult to interpret its result in the meaning (Schmitz et al., 1999). To identify the factors that affect student's decision, decision tree is suitable for this purpose and its result will be discussed in the next section.

Table 2

Classification results in confusion matrix

	Enroll(E)	Not-Enroll(N)	Accuracy(%)
Artificial Neural Network (ANN)			
Enroll (E)	182	18	91.00
Not-Enroll (N)	10	190	95.00
		Overall	93.00
Decision Tree (DT)			
Enroll (E)	174	26	87.00
Not-Enroll (N)	21	179	89.50
		Overall	88.25

2. Attribute identification result

The six important attributes obtained from filter ranker method were university fame, family income, curriculum number, university location, parent career and modern curriculum. They were arranged in decision tree visualization as illustrated in Fig. 3. From the visualization, the university fame was taken as the root node while family income, curriculum number, university location, parent career and modern curriculum were internal nodes. The leaf nodes were E and N classes. The prioritization of these attributes can be described by IF-THEN rules as presented in Fig. 4. For example, if the score value of university fame is lower than or equal 3 value, then the family income is next considered. If its score is lower than or equal 2 value, then the data samples are classified into the E class. If family income score is more than 2 value, then curriculum number is next considered, and so on. Based on filter ranker method, the decision tree visualization and IF-THEN rules, it was found that the best important factor was university fame. This indicates that students first consider the university fame for their decision. Family income and university location are the second factors for decision and then curriculum number, parent career and modern curriculum, respectively.

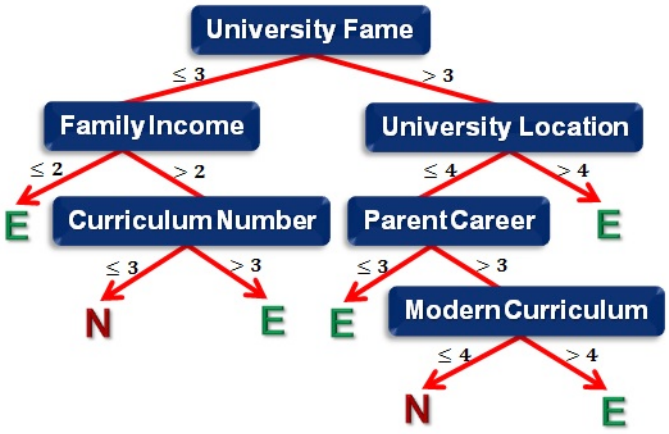

Figure 3 Generated visualization of decision tree.

1) IF UniversityFame > "3" AND UniversityLocation > "4" THEN Result = "E"

2) IF UniversityFame > "3" AND UniversityLocation ≤ "4" AND ParentCareer ≤ "3" THEN Result = "E"

3) IF UniversityFame > "3" AND UniversityLocation ≤ "4" AND ParentCareer > "3" AND ModernCurriculum > "4" THEN Result = "E"

4) IF UniversityFame ≤ "3" AND FamilyIncome ≤ "2" THEN Result = "E"

5) IF UniversityFame ≤ "3" AND FamilyIncome > "2" AND CurriculumNumber > "3" THEN Result = "E"

6) IF UniversityFame ≤ "3" AND FamilyIncome > "2" AND CurriculumNumber ≤ "3" THEN Result = "N"

IF UniversityFame > "3" AND UniversityLocation ≤ "4" AND ParentCareer > "3" AND ModernCurriculum ≤ "4" THEN Result = "N"

Figure 4 IF-THEN rule diagram.

Conclusions

This study investigated the important factors that affect student's decision for studying at Phetchabun Rajabhat University by using data mining with ANN and decision tree methods. 400 students, 200 freshmen from Phetchabun Rajabhat University and 200 other freshmen from other universities were studied. The main data used were score values of 14 attributes on questionnaires. To evaluate the collected data difference based on classification models, the results show that ANN classification accuracy was higher than that of decision tree. In case of factor identification, decision tree was used. From filter ranker method for attribute selection and decision tree visualization and IF-THEN rule for interpretation, the important factors affecting student's decision are university fame, family

income, university location, curriculum number, parent career and modern curriculum, respectively. The results reveal that data mining techniques can be used to identify the factors affecting student's decision for choosing universities. Moreover, it is useful for improving admission system of a group of Rajabhat University to increase a number of freshmen.

Acknowledgements

This work was financially supported by Research and Development Institute Phetchabun Rajabhat University.

References

Ahmad, M.W., Mourshed, M., & Rezgui, Y. (2017). Trees vs neurons: Comparison between random forest and ANN for high-resolution prediction of building energy consumption. *Energy and Buildings, 147*(1), 77-89.

Ahmed, A.M., Rizaner, A., & Ulusoy, A.H. (2016). Using data mining to predict instructor performance. *Procedia Computer Science, 102*(1), 137-142.

Angeli, C., Howard, S.K., Ma, J., & Yang, J. (2017). Data mining in educational technology classroom research: Can it make a contribution? *Computer & Education, 113*(1), 226-242.

Deng, W.J., Chen, W.C., & Pei, W. (2008). Back-propagation neural network based importance-performance analysis for determining critical service attributes. *Expert Systems with Applications, 34*(2), 1115-1125.

Han, J., & Kamber, M. (2006). *Data mining: Concepts and techniques*. San Francisco: Morgan Kaufmann Publishers.

Hsu, C.H. (2009). Data mining improve industrial standards and enhance production and marketing: An empirical study in apparel industry. *Expert Systems with Applications, 36*(3), 4185-4191.

Ippoodom, T. (2017). *Thai University crisis when the educational institutions wage a war for students to survive*. Retrieved from https://thematter.co/pulse/war-of-thai-university/25611.

Jones, D.E., Ghandehari, H., & Facelli, J.C. (2016). A review of the applications of data mining and machine learning for the prediction of biomedical properties of nanoparticles. *Computer Methods and Programs in Biomedicine, 132*(1), 93-103.

Jothi, N., Rashid, N.A.A., & Husain, W. (2015). Data mining in Healthcare–A Review. *Procedia Computer Science, 72*(1), 306-313.

Kaur, P., Singh, M., & Josan, G.S. (2015). Classification and prediction based data mining algorithms to predict slow learners in education sector. *Procedia Computer Science, 57*(1), 500-508.

Liao, S.H., Chu, P.H., & Hsiao, P.Y. (2012). Data mining techniques and applications: A-decade review from 2000 to 2011. *Expert Systems with Applications, 39*(12), 11303-11311.

Lou, W., & Nakai, S. (2001). Application of artificial neural networks for predicting the thermal inactivation of bacteria: A combined effect of temperature, pH and water activity. *Food Research International, 34*(7), 573-579.

Lui, X., Li, Q., Li, T., & Chen, D. (2018). Differentially private classification with decision tree ensemble. *Applied Soft Computing, 62*, 807-816.

Natek, S., & Zwilling, M. (2014). Student data mining solution-knowledge management system related to Higher Education Institutions. *Expert Systems with Applications, 41*(14), 6400-6407.

Ozyirmidokuz, E.K., Uyar, K., & Ozyirmidokuz, M.H. (2015). A data mining based approach to a firm's marketing channel. *Procedia Economics and Finance, 27*, 77-84.

Packianather, M.S., Davies, A., Harraden, S., Soman, S., & White, J. (2017). Data mining techniques applied to a manufacturing SME. *Procedia CIRP, 62*(1), 123-128.

Panto, O., & Theantong, M. (2014). A comparative efficiency of classification of VARK learning style using data mining techniques. *Journal of Industrial Technology Ubon Ratchathani Rajabhat University, 4*(1), 1-11. (in Thai)

Phang, S.L. (2012). *Factors influencing international students' study destination decision abroad*. Master of Communication Thesis, University of Gothenburg.

Pu, Y., Apel, D.B., & Lingga, B. (2018). Rockburst prediction in kimberlite using decision tree with incomplete data. *Journal of Sustainable Mining, 17*(3), 158-165.

Rodrigues, M.W., Isotani, S., & Zarate, L.E. (2018). Educational data mining: A review of evaluation process in the e-learning. *Telematics and Informatics, 35*(6), 1701-1717.

Schmitz, GJ., Aldrich, C., & Gouws, F.S. (1999). ANN-DT: An algorithm for extraction of decision trees from artificial neural networks. *IEEE Transactions on Neural Networks, 10*(6), 1392-401.

Sen, B., & Ucar, E. (2012). Evaluating the achievements of computer engineering department of distance education students with data mining methods. *Procedia Technology, 1*(1), 262-267.

Sen, B., Ucar, E., & Delen, D. (2012). Predicting and analyzing secondary education placement-test scores: A data mining approach. *Expert Systems with Applications, 39*(10), 9468-9476.

Sittichat, S. (2017). Study of educational attributes using data mining technique. *Information Technology Journal, 13*(2), 20-28. (in Thai)

Ramaswami, M., & Bhaskaran, R. (2009). A study on feature selection techniques in educational data mining. *Journal of Computing, 1*(1), 7-11.

Timsorn, K., Lorjaroenphon, Y., & Wongchoosuk, C. (2017). Identification of adulteration in uncooked Jasmine rice by a portable low-cost artificial olfactory system. *Measurement, 108*(1), 67-76.

Timsorn, K., Thoopboochagorn, T., Lertwattanasakul, N., & Wongchoosuk, C. (2016). Evaluation of bacterial population on chicken meats using a briefcase electronic nose. *Biosystems Engineering, 151*(1), 116-125.

Tsai, Y.C., Trang, L.T., & Kobori, K. (2017). Factors influencing international students to study at Universities in Taiwan. *International Journal for Innovation Education and Research, 5*(1), 1-11.