Educational Data Mining for Supporting Students' Courses Selection

Thi Yen Tran[†], Ba Lam To[†]

Ho Chi Minh City University of Transport, Viet Nam[†]

Summary
At the beginning of each semester, students must choose courses in the course list to study. Many students are confused about choosing the most suitable courses. Some students want to choose courses to improve academic achievement and improve grades. Some other students want to choose courses to avoid being academic warning, academic probation, academic suspension or academic dismissal. Academic advisors and students' knowledge also partly support the selection of courses for students. However, this support depends on the experience of the academic advisor and the students' knowledge. This paper aims to build a model to support students' courses selection using educational data mining. The proposed model is experimented with educational data from Faculty of Civil Engineering, Ho Chi Minh City University of Transport, Vietnam during the period of 2013-2016. Experimental results bring many positive results for supporting courses selection.

Key words:
Educational Data Mining, Apriori, J48, K-Means

1. Introduction

Selecting the most suitable courses to improve academic achievement or avoid being academic warning is very difficult for students. Data Mining is the data analyzing technique from different perspectives also summarizing the useful information results. The data mining process uses many principles as machine learning, statistics and visualization techniques to discover and present knowledge in an easily comprehensible form. Data mining is a process to create knowledge from transactional database by using statistic procedure and machine learning and training set to get the exact information for predicted decision.

Educational data mining (EDM), is an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings, and using those methods to better understand students, and the settings which they learn in [1-2]. Scholars in educational data mining have used many data mining techniques such as Decision Trees, Support Vector Machines, Neural Networks, Naïve Bayes, K-Nearest neighbor, among others to discover many kinds of knowledge such as association rules, classifications and clustering [3-6].

Data mining is "Knowledge Discovery in Databases (KDD)" and includes a combination of advancements in the fields of machine learning, pattern recognition, database, statistics, artificial intelligence, knowledge acquisition and data visualization. This process of extraction is conducted following six steps as shown in Figure 1 [7-9].

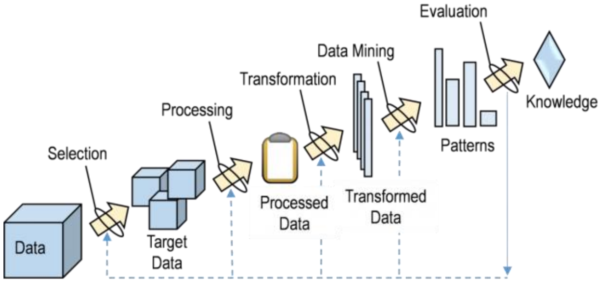


Fig. 1 Data Mining Model

This paper proposed a data mining based solution to extract knowledge from educational data and support students selecting the most suitable courses from extracted knowledge. The main contribution of this paper focuses on the following sections:

- Pre-processing educational data to make clean data.
- Extracting knowledge using data mining algorithms via Weka.
- Supporting for students to select the most suitable courses based on the extracted knowledge above.

This paper is organized as follows. The proposed model is described in detail of Section 2. Section 3 presents the experimental data, the experimental process and the experimental results. Section 4 gives conclusions and outlines future research directions.

2. Proposed Model

The proposed model including 2 modules is shown in Figure 2.

Manuscript received July 5, 2019
Manuscript revised July 20, 2019

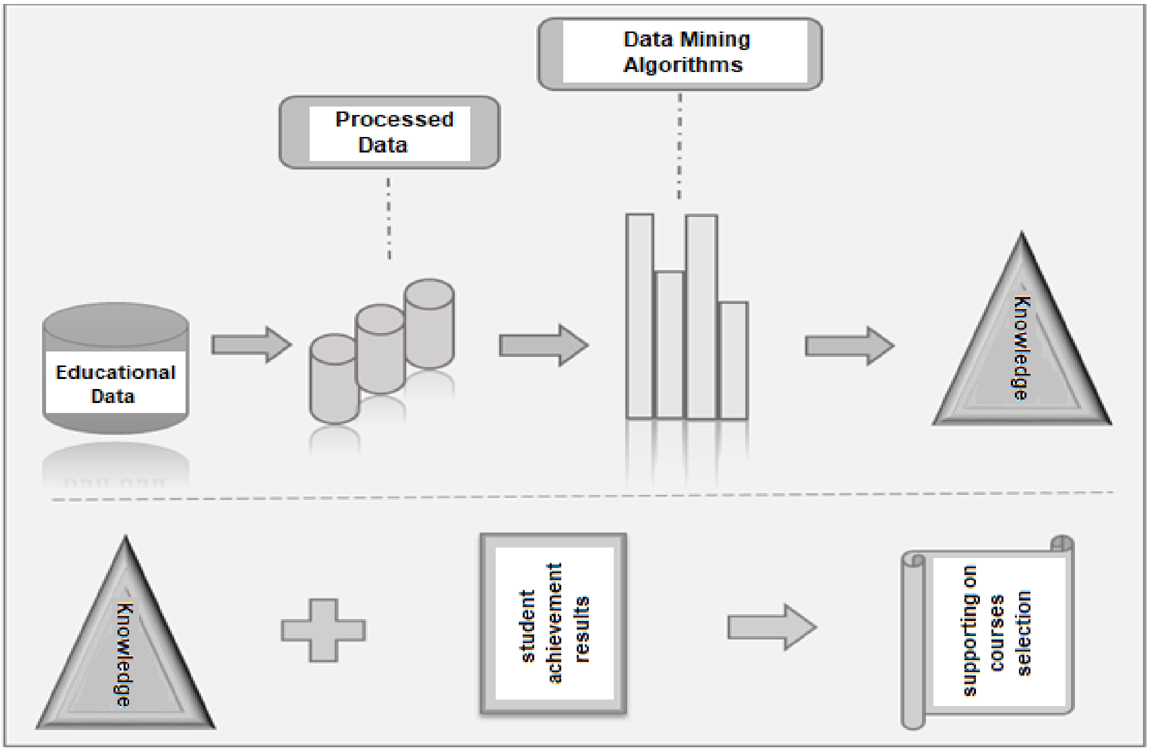


Fig. 2 Proposed Model

- The first module is the "Extracting Knowledge" module which extracts knowledge from educational data based on data mining algorithms.
- The second module is the "Supporting on courses selection" module which supports students selecting the most suitable courses based on extracted knowledge from the first module.

2.1 "Extracting Knowledge" Module

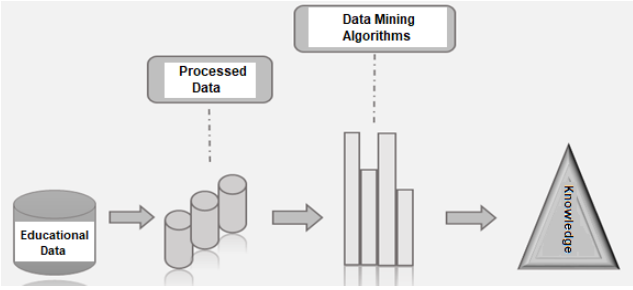


Fig. 3 "Extracting Knowledge" Module

This module is described in Figure 3. This module uses data mining algorithms to extract knowledge from educational data. There are many techniques in data mining that can be applied to educational data, such as classification, clustering, and association rules. These techniques will help extract hidden knowledge and useful information.

Association rule mining is one of the most well studied data mining tasks. It discovers relationships among attributes in databases, producing if-then statements concerning attribute values [10]. Apriori algorithm is the first and best-known algorithm for association rules mining. Apriori was proposed by R.Agrawal and R.Srikant [11] in 1994. Association rules help uncover all such relationships between items from huge databases. Association Rule Mining approach which uses Apriori algorithm to analyze the relationship among courses.

A decision tree is one of the most well-known classification approaches that are commonly used to examine data and induce the tree in order to make predictions [12]. The purpose of the decision tree is to classify data into distinct groups or branches that generate the strongest separation in the values of the dependent variable [13]. J48 is an open source Java implementation of the C4.5 algorithm under WEKA data mining platform. J48 uses gain ratio to classify the decision tree [14]. The aim of classification is to predict the future output based on the available data. Hence, in order to predict the future output of students based on their available previous and current students data, which make classification one of the techniques better suited for educational analysis. Analyzing all the courses that are required in the study plan will identify the list of courses that have a huge impact on final GPAs.

Clustering is finding groups of objects such that the objects in one group will be similar to one another and different from the objects in another group. Clustering can be considered the most important unsupervised learning technique. Simple K-means algorithm is a type of unsupervised algorithm in which items are moved among the set of cluster until required set is reached. Clustering to group student according to their features. Clustering by majors, clustering by learning results, clustering by course's score, etc. Knowledge extracted from clustering will support students to choose the most suitable courses.

2.2 "Supporting on courses selection" Module

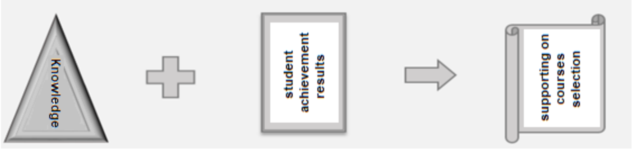


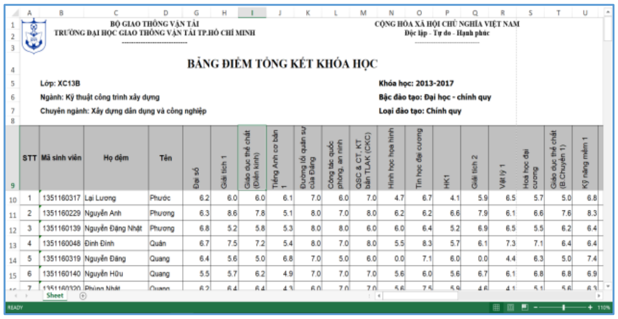
Fig. 4 "Supporting on courses selection" Module

This module is shown in Figure 4. This module uses the extracted knowledge from "Extracting knowledge" module to support students selecting the most suitable courses.

3. Experimental Setup

3.1 Experimental Data

The data of this experiment was collected from Faculty of Civil Engineering, Ho Chi Minh City University of Transport, Vietnam during the period of 2013-2016. There are 16 scoring files corresponding to 16 classes of the Faculty of Civil Engineering during the period 2013 to 2016. Each semester's score corresponds to one student per row, each column is one course code. The data is pre-processed, and transformed to be appropriated format so as to apply data mining techniques to extract knowledge. Figure 5 shows the score of students in 2013.


BỘ GIAO THÔNG VẬN TẢI
TRƯỜNG ĐẠI HỌC GIAO THÔNG VẬN TẢI TP.HỒ CHÍ MINH

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập - Tự do - Hạnh phúc

BẢNG ĐIỂM TỔNG KẾT KHÓA HỌC

Lớp: XC13B
Ngành: Kỹ thuật công trình xây dựng
Chuyên ngành: Xây dựng dân dụng và công nghiệp

Khóa học: 2013-2017
Bậc đào tạo: Đại học - chính quy
Loại đào tạo: Chính quy

STT	Mã sinh viên	Họ đệm	Tên	Đại số	Giải tích 1	Giáo dục thể chất (Điền kinh)	Tiếng Anh cơ bản 1	Đường lối quân sự của Đảng	Công tác quốc phòng, an ninh	QSC & CT, KT bắn TLAK (CKC)	Hình học họa hình	Tin học đại cương	HK1	Giải tích 2	Vật lý 1	Hoá học đại cương	Giáo dục thể chất (B.Chuyền 1)	Kỹ năng mềm 1
1	1351160317	Lại Lương	Phước	6.2	6.0	6.0	6.1	7.0	6.0	7.0	4.7	6.7	4.1	5.9	6.5	5.7	5.0	6.8
2	1351160229	Nguyễn Anh	Phương	6.3	8.6	7.8	5.1	8.0	7.0	8.0	6.2	6.2	6.6	7.9	6.1	6.6	7.6	8.3
3	1351160139	Nguyễn Đặng Nhật	Phương	6.8	5.2	5.8	5.3	8.0	8.0	6.0	6.0	6.4	5.2	6.9	6.5	5.5	6.2	6.4
4	1351160048	Đinh Đinh	Quân	6.7	7.5	7.2	5.4	8.0	7.0	8.0	5.5	8.3	5.7	6.1	7.3	7.1	6.4	6.4
5	1351160319	Nguyễn Đăng	Quang	6.4	5.6	5.0	6.8	7.0	5.0	6.0	0.0	7.1	6.0	0.0	4.4	6.3	5.0	7.4
6	1351160140	Nguyễn Hữu	Quang	5.5	5.7	6.2	4.9	7.0	7.0	7.0	5.6	6.0	5.6	6.7	6.1	6.8	6.8	6.9
7	1351160320	Phùng Nhất	Quang	6.2	6.4	6.4	4.3	6.0	7.0	7.0	5.6	7.5	5.9	4.6	4.1	5.1	5.6	6.3

Fig. 5 The score of students in 2013

3.2 Data Pre-Processing

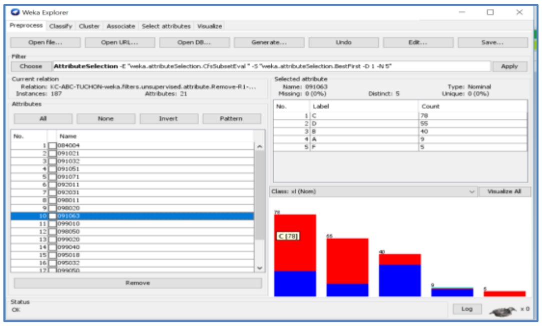
Initially the datasets were collected in Ms Excel sheet and initial preprocessing was done manually by filling the missing values in the data set. Some irrelevant attributes were removed. Feature selection was used as a method to select relevant attributes from the full set of attributes as a measure of dimensionality reduction. The objective of feature selection was to choose a subset of input variables by eliminating features, which are irrelevant or of no predictive information [15].

This study used data mining software to investigate the most important variables. The open source software WEKA [16], offering a wide range of machine learning algorithms for Data Mining tasks, was used as a data mining tool for the research implementation. The selected attributes were transformed into a form acceptable to WEKA data mining software. The data file was saved in Comma Separated Value (CSV) file format in Microsoft excel and later was converted to Attribute Relation File Format (ARFF) file inside WEKA for ease of analysis.

3.3 Extracting Knowledge

3.3.1 Apriori Algorithm

After processing educational data, ARFF data files are put into Weka to process. The data features in Weka are shown in Figure 6 and Figure 7.



Filter: AttributeSelection -E "weka.attributeSelection.CfsSubsetEval " -S "weka.attributeSelection.BestFirst -D 1 -N 5"

Current relation
Relation: KC-ABC-TUCHON-weka.filters.unsupervised.attribute.Remove-R1-...
Instances: 187
Attributes: 21

Selected attribute
Name: 091063
Type: Nominal
Missing: 0 (0%)
Distinct: 5
Unique: 0 (0%)

No.	Label	Count
1	C	78
2	D	55
3	B	40
4	A	9
5	F	5

Attributes

No.	Name
1	084004
2	091021
3	091032
4	091051
5	091071
6	092011
7	092031
8	098011
9	098020
10	091063
11	099010
12	098050
13	099020
14	099040
15	095018
16	095032
17	099050

Class: xl (Nom)

Status
OK

Fig. 6 Data Attributes in Weka

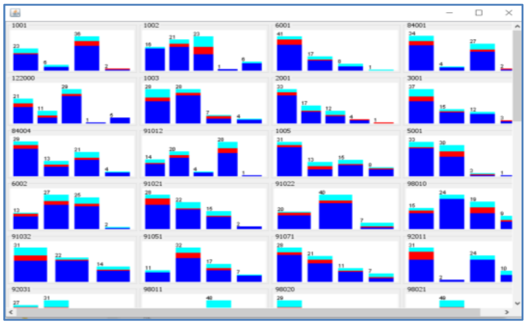
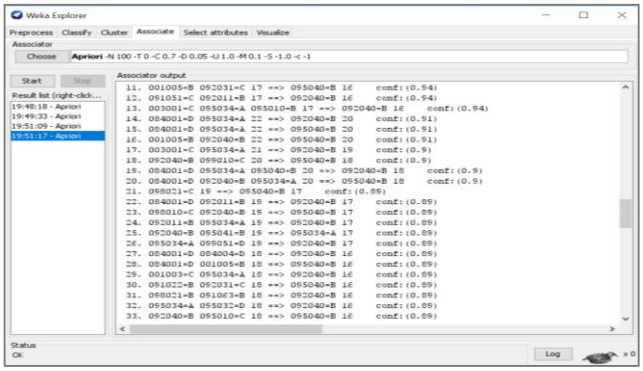


Fig. 7 Data Attributes in Weka

The "Association rule" knowledge generated by the Apriori algorithm via Weka is shown in Figure 8. Extracting meaningful rules to support courses selection.



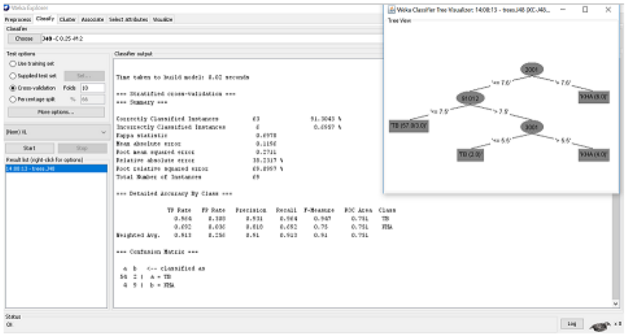
Apriori -N 100 -T 0 -C 0.7 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1

```
 11. 001005=B 092031=C 17 ==> 095040=B 16    conf:(0.94)
 12. 091051=C 092011=B 17 ==> 092040=B 16    conf:(0.94)
 13. 003001=C 095034=A 095010=B 17 ==> 092040=B 16    conf:(0.94)
 14. 084001=D 095034=A 22 ==> 092040=B 20    conf:(0.91)
 15. 084001=D 095034=A 22 ==> 095040=B 20    conf:(0.91)
 16. 001005=B 092040=B 22 ==> 095040=B 20    conf:(0.91)
 17. 003001=C 095034=A 21 ==> 092040=B 19    conf:(0.9)
 18. 092040=B 098010=C 20 ==> 095040=B 18    conf:(0.9)
 19. 084001=D 095034=A 095040=B 20 ==> 092040=B 18    conf:(0.9)
 20. 084001=D 092040=B 095034=A 20 ==> 095040=B 18    conf:(0.9)
 21. 098021=C 19 ==> 095040=B 17    conf:(0.89)
 22. 084001=D 092011=B 19 ==> 092040=B 17    conf:(0.89)
 23. 098010=C 092040=B 19 ==> 095040=B 17    conf:(0.89)
 24. 092011=B 095034=A 19 ==> 092040=B 17    conf:(0.89)
 25. 092040=B 095041=B 19 ==> 095034=A 17    conf:(0.89)
 26. 095034=A 098051=D 19 ==> 092040=B 17    conf:(0.89)
 27. 084001=D 084004=D 18 ==> 092040=B 16    conf:(0.89)
 28. 084001=D 001005=B 18 ==> 095040=B 16    conf:(0.89)
 29. 001003=C 095034=A 18 ==> 092040=B 16    conf:(0.89)
 30. 091022=B 092031=C 18 ==> 095040=B 16    conf:(0.89)
 31. 098021=B 091063=B 18 ==> 092040=B 16    conf:(0.89)
 32. 095034=A 095032=D 18 ==> 092040=B 16    conf:(0.89)
 33. 092040=B 095010=C 18 ==> 095040=B 16    conf:(0.89)
```

Fig. 8 "Association Rule" Knowledge

3.3.2 J48 Decision Tree

After processing educational data, ARFF data files are put into Weka to process. The knowledge generated by the J48 decision tree algorithm via Weka is shown in Figure 9. From "decision tree" knowledge, extracting meaningful rules to support courses selection.



J48 -C 0.25 -M 2

Folds 10 % 66

```
Time taken to build model: 0.02 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          63               91.3043 %
Incorrectly Classified Instances         6                8.6957 %
Kappa statistic                          0.6978
Mean absolute error                      0.1156
Root mean squared error                  0.2711
Relative absolute error                 38.2317 %
Root relative squared error             69.8957 %
Total Number of Instances               69

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall   F-Measure   ROC Area   Class
                 0.964     0.308     0.931      0.964     0.947       0.751     TB
                 0.692     0.036     0.818      0.692     0.75        0.751     KHA
Weighted Avg.    0.913     0.256     0.91       0.913     0.91        0.751

=== Confusion Matrix ===

  a  b   <-- classified as
 54  2 |  a = TB
  4  9 |  b = KHA
```

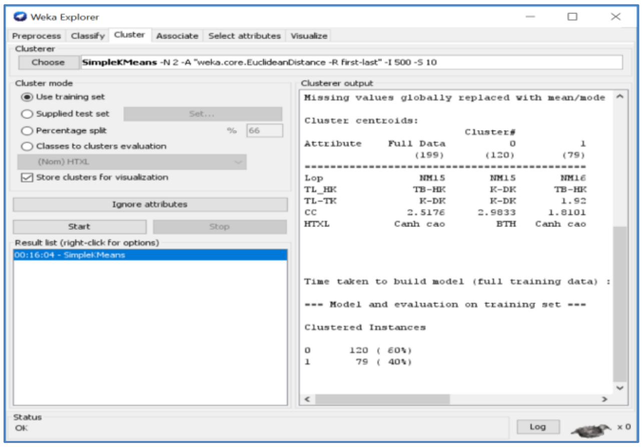
Fig. 9 "Decision Tree" Knowledge

3.3.3 K-Means Clustering

After processing educational data, ARFF data files are put into Weka to process. The clustering knowledge generated by the K-Means algorithm via Weka is shown in Figure 10 and Figure 11. From clustering knowledge, extracting meaningful knowledge to support courses selection.

3.4 Supporting Courses Selection

The following is an illustration of support for courses selection for 2 students. A student who wants to improve academic achievement, another student being academic warning wants to escape this situation.



SimpleKMeans -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -S 10

% 66

☑ Store clusters for visualization

```
Missing values globally replaced with mean/mode

Cluster centroids:
                                   Cluster#
Attribute       Full Data              0             1
                  (199)            (120)          (79)
=======================================================
Lop                NM15             NM15          NM16
TL_HK             TB-HK             K-DK         TB-HK
TL-TK              K-DK             K-DK          1.92
CC               2.5176           2.9833        1.8101
HTXL           Canh cao              BTH      Canh cao




Time taken to build model (full training data) :

=== Model and evaluation on training set ===

Clustered Instances

0       120 ( 60%)
1        79 ( 40%)
```

Fig. 10 Clustering Results

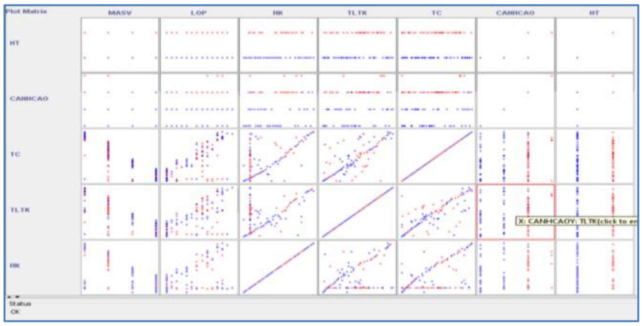


Fig. 11 Distribution of Attribute Data

3.4.1 The First Case

A student of year 2016 is studying civil and industrial construction. Student's GPA is 1.96, the student wants to improve his GPA in the next semester. Based on the extracted knowledge above, supporting the student to select the following courses:

- Theoretical mechanics (3 credits)
- Structural mechanics Part 1 (3 credits)
- Structural mechanics Part 2 (2 credits)
- Reinforced concrete structure Part 1 (3 credits)
- Soil mechanics (3 credits)
- Architectural projects (1 credits)
- Testing construction materials (1 credits)
- Construction Engineering (3 credits)

The student registers a total of 19 credits (24 credits maximum of a semester). Selecting such courses ensures a balance for students' schedule, ensures the completion of failed courses, reinforces the knowledge of some courses, and improves knowledge of new courses.

3.4.2 The Second Case

A student of year 2016 is studying civil and industrial construction. Student's GPA is less than 1.0, the student is being academic warning. Based on the extracted knowledge above, supporting the student to select the following courses:
- Selecting failed courses to complete these courses.
- Selecting courses with low scores to improve GPA.
- Registering 23 or 24 credits to ensure academic progress.

4. Conclusion

This paper proposed a solution using data mining technique to extract knowledge from education data. Then, using the knowledge to support students to select the most suitable courses in the next semester. The solution is experimented with the data collected from Faculty of Civil Engineering, Ho Chi Minh City University of Transport, Vietnam during the period of 2013-2016. Experimental results bring many positive results. In the future, the proposed solution will be improved to support better.

Acknowledgments

The authors thank reviewers for their reading of our manuscript and their insightful comments and suggestions.

References

[1] Karim, M., & Rahman, R. M. (2013, April). Decision Tree and Naïve Bayes Algorithm for Classification and Generation of Actionable Knowledge for Direct Marketing. Journal of Software Engineering and Applications, 6, 196-206.

[2] Kabakchieva, D., Stefanova, K., & Kisimov, V. (2011). Analyzing University Data for Determining Student Profiles and Predicting Performance. Conference Proceedings of the 4th International Conference on Educational Data Mining (EDM 2011), 6- 8 July 2011 (pp. 347-348).

[3] Sahay, A., & Mehta, K. (2010). Assisting higher education in assessing, predicting, and managing issues related to student success: A web-based software using Data Mining and Quality Function Deployment. Academic and Business Research Institute Conference, (pp. 1-12). Las Vegas.

[4] Fong, S., Yain-Whar, S., Robert, P., & Aghai, B. (2009). Applying a Hybrid Model of Neural Network and Decision Tree Classifier for Predicting University Admission.

[5] Padmapriya, A. (2012, November). Prediction of Higher Education Admissibility using Classification Algorithms. International Journal of Advanced Research in Computer Science and Software Engineering, 2(11), 330- 336.

[6] Siraj, F., & Abdoulha, M. (2009). Uncovering hidden information within university's student enrolment data using data mining. MASAUM Journal of Computing, 1(2), 337-342.

[7] D'Oca S, Hong T. Occupancy schedules learning process through a data mining framework. Energy and Buildings, 2015.

[8] D'Oca S, Corgnati S, Hong T. Data mining of occupant Behavior in office buildings, 6th International Building Physics Conference, IBPA, 2015.

[9] D'Oca S, Hong T. A data-mining approach to discover patterns of window opening and closing behavior in offices. Building and Environment, 2014.

[10] Agrawal, R., Imielinski, T. and Swami, A.N., Mining Association Rules between Sets of Items in Large Databases. In Proceedings of SIGMOD, 207-16, 1993.

[11] Agrawal R., Srikant R.,"Fast algorithms for mining association rules", In Proceedings 20th International Conference on VeryLarge Data Bases (VLDB' 94), pp. 487-499, 1994.

[12] H. Edelstein, "Introduction to Data Mining and Knowledge Discovery", Third Edition. Two Crows Corporation, Potomac, MD, USA, 1999.

[13] Parr Rud, O. "Data Mining Cookbook. Modeling Data for Marketing, Risk, and Customer Relationship Management". John Wiley & Sons, Inc.; 2001.

[14] A. Kumar Sharma and S. Suruchi, "A Comparative Study of Classification Algorithms for Spam Email Data Analysis", International Journal on Computer Science and Engineering (IJCSE), Vol. 3 No. 5, pp. 1891-1895. May 2011

[15] D. Koller and M. Sahami, "Toward optimal feature selection," In Proceedings of the Thirteenth International Conference on Ma-chine Learning, pp. 284–292, 1996.

[16] L. Breiman, J. Friedman, R. Olshen, C. Stone.: Classification and Regression Trees. Chapman & Hall (1984).

Thi Yen Tran received B.Sc. degree in Computer Science. Currently, She is preparing to get a M.Sc degree in Computer Science. Her current research interests include Data Mining, Machine Learning and Artificial Intelligence.

Ba Lam To is a lecturer in the Faculty of Information Technology, Ho Chi Minh City University of Transport, Vietnam. He received his B.Eng. degree in Information Technology from HCM City University of Technology in 2006 and received Master and Ph.D degree in 2008 and 2012 respectively from Université Pierre et Marie Curie (UPMC -Paris 6). His current research interests include Cloud Computing, Routing in Cognitive Radio Network, Network Security, Intelligent transportation system.