International Journal of Artificial Intelligence and Neural Networks – IJAINN
Volume 4 : Issue 4 [ISSN 2250 – 3749]
Publication Date : 27 December, 2014

The Knowledge Representation Of The Program Analysis Using Decision Tree Data Mining Technique

Sasitorn Kaewman, and Chatklaw Jareanpon

***Abstract*— The Education Data Mining (EDM) is the data mining for analyzing the data from and for education. This paper proposed the framework to analysis the undergraduate program. The selected data mining technique is the well-known Decision Tree Data Mining Technique. This result of this proposed is able to analysis the major subjects that will effect with the student and the registered planning. Moreover, the course management is able to plan the relationship and the program. The experimental result tested from computer simulation from real registration data and grade of department of Informatics, Mahasarakham University shows that the knowledge representation of the program analysis using Decision Tree Data Mining Technique is very well. The average accuracy from k-fold cross validation is 74.05%.**

***Keywords*— Knowledge Representation, Program Analysis, Decision Tree, Data Mining Technique.**

I. Introduction

Education Data Mining (EDM) is the research field concerned with the application of data mining to information generated from education such as university and intelligent tutoring systems. EDM has contributed to theories of learning investigated by researchers in educational psychology, learning science, and education management. This field consists of the four main users which are learners, educators, researchers, and administrators. The main objective is trying to understand learners, and how they interact with learning environments.

The EDM is interested by many researcher in various topics. Nawal et al. proposed the web usage mining analysis on moodle [1]. This research proposed the new static variables using SCORM (Shareable Content Object Reference Model) content tree and new visualization technique. It is automatic recommendations for teachers/tutors and learners. Agathe and Kalina proposed the analysis tool called Logic-ITA that is able to help the teacher as well as the learner [2].

Sasitorn Kaewman, and Chatklaw Jareanpon
Polar lab, Faculty of Informatics, Mahasarakham University
THAILAND

The Logic-ITA based on Association of mistakes destined to the student, along with tools dedicated to teacher for managing teaching configuration setting and material as well as for collecting and analyzing data. Apart from web analysis intelligence, Brijesh and Saurabh proposed the Decision tree for evaluating the student's performance of MCA (Master of Computer Applications) course [3]. This research will help the students and the teachers to improve the division of the student. Abeer and Ibrahim proposed the classification task used to predict the final grade student using decision tree [4]. Vatun Kumar and Anuoama Chadha proposed the association rule mining technique to enhancing the quality of student's performances at Post Graduation level [5]. Kavita and Rajanish investigated the arrangement and rearrangement of the subjects in two postgraduate academic programmed using k-mean clustering [6]. The idea of classification of student's performance of each subject is will help student and lecturer to improve each course.

All of the researches are interested in the subject or student. However, in the education curriculum still found the problem that how to improve the course or the program that the graduated student get the exact major job. That knowledge is possible to plan and help to manage the program. This research proposed the program analysis for finding the subject rule that effect to get the exact job by major.

This paper is now organized as follows. Section 2 will cover the concept of Education Data Mining Process. Section 3 will propose the design and method. Section 4 is the experimental and result, Section 5 is the Knowledge Representation, and Section 6 is conclusion and future work.

II. Educational Data Mining Process

The goal of the EDM is to use the large-scale educational data sets to better understand learning and to provide information about the learning process. Variety researchers are interested in this topic, it is able to classify into 3 groups which are 1) developing computational tools and techniques that work on extending and understanding the necessary tool foundation to EDM, 2) understanding and determining data that works on determining what are new, exciting questions to ask the data which is necessary for EDM to grow, and 3) expanding the list of stakeholders for whom who can provide information, and where this information is received.

The EDM process is following this step:

Domain: The goal for using EDM.

Data: Where is collected data?, and what's size of data?

Objective: The objective of data mining in EDM is used for analysis the learning method, students, or helping the class management.

Technique: The majority of traditional data mining techniques including but not limited to classification, clustering, and association analysis techniques have been already applied successfully in the educational domain.

Data representatives: The output will analysis and represent to the stakeholders.

III. Program analyzed Data mining using Decision tree

This process consisted of 4 steps as shown in Figure 1.

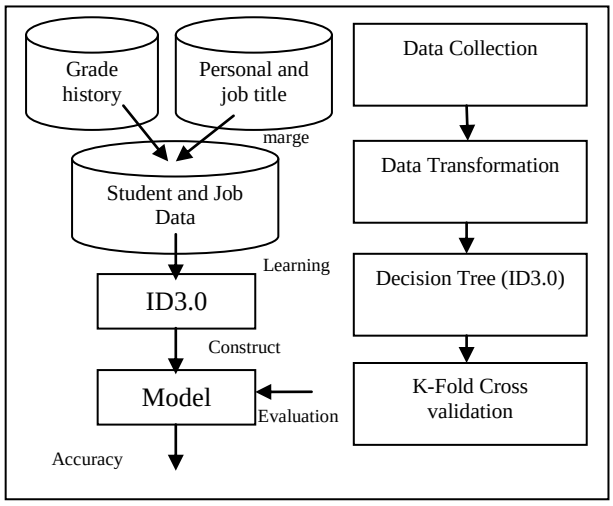


Figure 1. Program analyzed Data mining Using Decision tree process and diagram.

A. Data collection

The data set used in this research was obtained from a student's Faculty of Informatics Mahasarakham University. This faculty consisted of 6 programs that are Computer Science, Information technology, Information Science, Mass Communication, Geo-informatics, and Computer game and animation [7]. The data is divided into 2 parts that are subject enrolment and grade history, and job title collected from 2004-2008.

B. Data Selection and transformation

In this step, the unnecessary data will remove and the necessary data is transform within suitable format. From the collected data, the Geo-informatics program has still not found the graduated student. Thus, five remained programs will use in this research with 1,621 student's records and 153,198 enrolments major subjects of all students as shown in Table 1. The data will transform and show in Table 2 and 3.

TABLE I. RAW DATA OF EACH PROGRAM ENROLMENT

Programs	Records	Enrolment Subjects
Computer Science (CS)	242	33,154
Information Technology (ICT)	406	19,975
Computer game and animation (NMD)	273	20,613
Information Science (IS)	315	51,532
Mass Comminication (MC)	404	27,924

TABLE II. RAW DATA OF JOB OF EACH STUDENT

Attributes	Example Value
Year	2010
Code of Program	CS
StudentID	50011210205
Job title	System Engineering
Exact job by major	(Yes=1, No = 0)

TABLE III. ENROLLMENT DATA

Attributes	Example Value
Enrolment Year	2012
Subject Code	1204304
StudentID	50011210205
Grade	(A=4, B+ = 3.5, … , F=0)

C. Decision Tree (ID3.0)

Decision trees are popular structure for supervised learning. Many researches are successfully applying the decision tree models to real-world problem. In this research, the ID3 is used to classify the exact job by major. The ID3 algorithm is to construct the decision tree by employing a top-down, greedy search though the given sets each attribute at every tree node.

The step of ID3 is shown as follow:

1. Let P be the set of training instance.
2. Choose an attribute that best differentiates the instances contained in T calculated by Equation 1 and splitting criteria calculated by Equation 2.
3. Create a tree node that value is the chosen attribute. Create child links from this node where each link represents a unique value for the chosen attribute. If the subclass does not satisfy the predefined criteria and there is at least one attribute to further subdivide the path of the tree, let T be the current set of subclass instance and return to step 2.

The programming structure that is suitable for creating the tree is recursive programming or dynamic programming.

$$Entropy = \sum_{j}^{r} - P_j \log_2 P_j \quad (1)$$

$$Gain(S, A) = \quad (2)$$

$$Entropy(s) - \sum_{v \in Value(A)} \left| \frac{S_v}{S} \right| Entropy(S_v)$$

D. *K-fold Cross Validation, Precision and Recall*

In k-fold cross-validation [8], the original sample is randomly partitioned into k equal size subsamples. Of the k subsamples, a single subsample is retained as the validation data for testing the model, and the remaining k−1 subsamples are used as training data. The cross-validation process is then repeated k times (the folds), with each of the k subsamples used exactly once as the validation data. The k results from the folds then can be averaged (or otherwise combined) to produce a single estimation as shown in Figure 2.

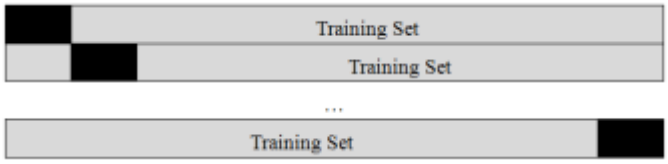


Figure 2. 10-fold Cross Validation (Black is the Test set and gray is the Training set)

The performance of the algorithm is able to calculate from the Precision and Recall calculated from Equation 3 and 4.

$$P_j = \frac{TP_j}{TP_j + FP_j} \quad (3)$$

$$R_j = \frac{TP_j}{TP_j + FN_j} \quad (4)$$

TABLE IV. PRESITION AND RECALL VARIABLE

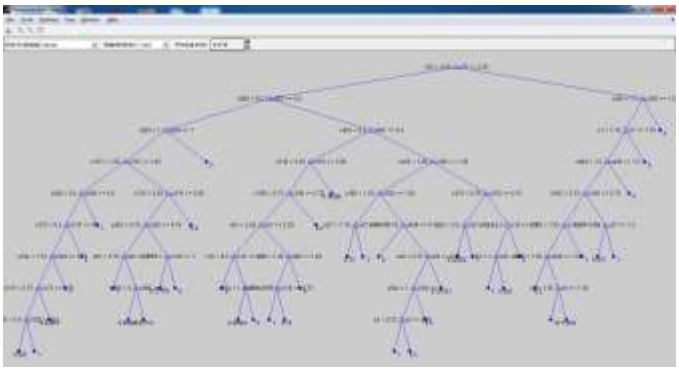
	Actual Class	
Predicted class	$TP_j(TruePositive)$	$FP_j(FalsePositive)$
	$FN_j(FalseNegative)$	$TN_j(TrueNegative)$

IV. Experimental and result

In this research, the ID3 is used to analysis each program. Then the five tree structure will create as shown in Figure 3. The rule for each program is represented in Table 4. The average accuracy from the k-fold cross validation is 74.05%.

TABLE V. THE RULE OF EACH PROGRAM

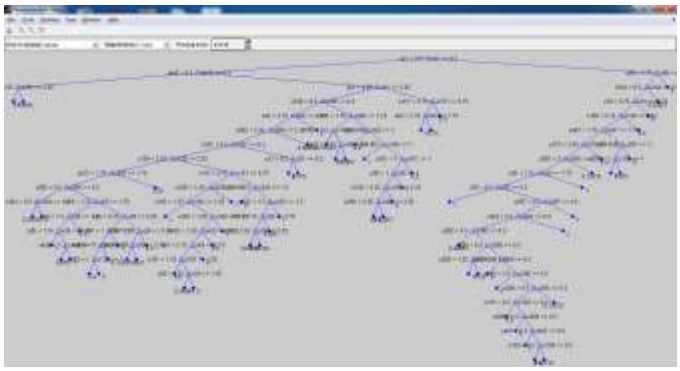
Programs	Rule number
Computer Science (CS)	80
Information Technology (ICT)	164
Computer game and animation (NMD)	102
Information Science (IS)	138
Mass Comminication (MC)	138



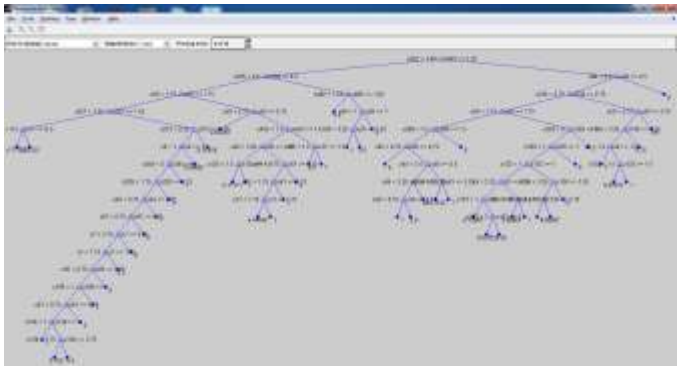
(a) Computer Science program rule

(b) Information Technology (ICT) program rule

(c) Information Science program rule



(d) Mass Comminication (MC) program rule



(e) Computer animation and game program rule

Figure 3. Analysis rule of each program using ID3

V. Knowledge Representation

This section will expend the rule from the tree that will show the knowledge from this research. For example, the Computer science program analysis, the rules of the subject that will effect the exact job by major are five rules as follows:

1. if Computer and Network Security>C+ and Fundamentals of UNIX > C+ and Application Program Development >= D+ this student will get the exact job by major;
2. if Computer and Network Security>C+ and Fundamentals of UNIX > C+ and Application Program Development >= D this student will get the exact job by major;

Then from the example rule, the lecturers, program management, and students must be concerned and interested in the Computer and Network Security subject. Moreover, that rule is able to represent the job in the market.

VI. Conclusion and Future work

This research tries to analysis the program with the exact job by major constraint. The contribution is valuable for the administrator to manage the program, for the learner to aware the enrolment subject. This framework is easy, simple and able to apply in other program. The accuracy is 74.05%.

For the future work, other algorithm such as association rule, neural network will be able to apply with this data. Moreover, some rule is a subset of some rule, then the mathematics concept such as set theory will use to pruning the tree. It will be reduced the computational time of seeking the rule.

References

[1] Nawak Sael, Abekaziz, and Hicham Behja,: Web Usage Mining Data Preprocessing and Multi Level analysis on Moodle, IJCSU International Journal of Advanced Computer Science, Vol. 10, Issue 2, No 2, 347-354 (2013).

[2] Agathe Merceron and Kalina Yacef,: Educational Data Mining: a Case Study, Proceedings of the 12th international Conference on Artificial Intelligence in Education AIED,467-474 (2005).

[3] Brijesh Kumar Baradwaj, and Saurabh Pal,: Mining Educational Data to Analyze Student Performance, IJCSU International Journal of Advanced Computer Science, Vol. 2, No 6, 69-69 (2011).

[4] Abeer Badr EI Din Ahmed, and Ibrahim Elaraby,: Data Mining: A prediction for Student's Performance using Classification Method, World Journal of Computer Application and Technology, Vol. 2, No 2, 43-47 (2014).

[5] Varun Kumar and Anupama Chadha,: Mining Association Rules in Student's Assessment Data, IJCSU International Journal of Advanced Computer Science, Vol. 9, Issue 15, No 3, 211-216 (2012).

[6] Kavita Oza and Rajanish Manat,: Applying Data Mining for Framing of Computer Science Curriculum, Proceedings of the ICTEC'13 Conference, (2013).

[7] Faculty of Informatics, Mahasakham University, http://it.msu.ac.th/eng/.

[8] Geisser Seymour, Predictive Inference. New York, NY: Chapman and Hall, (1993).

About Author (s):

Asst. Prof. Dr. Chatklaw Jareonpon is interested in Robotics and Data Mining Applications.

Asst. Prof. Sasitorn Keawmun interested in Software Development.