The application of Association Rule in data mining and building an assessment and classification system of students in Intermediate Professional Education

Nguyen Thi Thuy Trang

Abstract—**Data mining is defined as a process used to extract usable data from a larger set of any raw data. It implies analysing data patterns in large batches of data using one or more software. Data mining has applications in multiple fields, like science and research. Data mining involves effective data collection and warehousing as well as computer processing. For segmenting the data and evaluating the probability of future events, data mining uses sophisticated mathematical algorithms. Data mining is also known as Knowledge Discovery in Data (KDD).**

In this article, the author attempts to present a brief description of Data Mining and build applications using the Apriori-Tid algorithms to devise association rules to evaluate and classify students based on their reliability and participation.

***Index Terms*—Data Mining, Association Rule.**

I. INTRODUCTION

For the Education and Training sector, developing appropriate schemes is essential and extremely urgent especially during the current times. Therefore, there needs to be specific, accurate, and convincing information with scientific basis, from which to make timely adjustments to avoid errors. Exploring association rule in the evaluation and classification of students in Education and Training is immensely meaningful, and shall offer data based on reliable scientific foundations, which are also valuable to supporting educational management.

In actuality, the author has studied and developed a program to assess and classify the quality of intermediate professional students at the University of Technology and Education - the University of DaNang.

II. RELATED WORD

Data mining is the process of extracting useful information. Basically it is the process of discovering hidden patterns and information from the existing data. In data mining, one needs to primarily concentrate on cleansing the data so as to make it feasible for further processing. The process of cleansing the data is also called as noise elimination or noise reduction or feature elimination [1].

This can be done by using various tools available supporting various techniques. The important consideration in data mining is whether the data to be handled static or dynamic. In general, static data is easy to handle as it is known earlier and stored. Dynamic data refers to high voluminous and continuously changing information which is not stored earlier for analyzing and processing like static data. It is difficult to maintain dynamic data as it changes with time. Many algorithms are used to analyze the data of interest. Data can be sequential, audio signal, video signal, spatio -temporal, temporal, time series etc.

Data mining is a part of a bigger framework, referred to as knowledge discovery in databases (KDD) that covers a complex process from data preparation to knowledge modeling [2]. Main data mining task is classification which has main work to assign each record of a database to one of the predefined classes. The next is clustering which works in the way that it finds groups of records instead of only one record that are close to each other according to metrics defined by user. The next task is association which defines implication rules on the basis of that subset of record attributes can be defined. Data mining is the main important step to reach the knowledge discovery. Normally for data preprocessing it goes through various process such as data cleaning, data integration, data selection and data transformation and after these it is prepared for mining task. Its main contribution is in the fields of traditional sciences as astronomy, biology, high engineering physics, medicine and investigations. Various algorithms and tools can be used according to the application as given by Soni and Ganatra [3].

Data mining encompasses a range of techniques for uncovering valuable information hidden in large data sets (Databases). In essence, data mining involves analyzing data and using different techniques to find regular patterns in the data set.

Knowledge discovery in database (KDD) refers to the entire process of discovering useful knowledge from large data sets, in which data mining is a particularly important step using special algorithms for extracting patterns or models from data.

At a certain level of abstraction, data mining can be defined as a process of finding and discovering new, hidden, useful knowledge in large databases.

Knowledge discovery in database (KDD) is the main goal of data mining; as a result, these two concepts are considered

***Nguyen Thi Thuy Trang**, The University of Danang – University of Technology and Education, trangntt07@gmail.com, Danang, Vietnam.*

two equivalent fields. If the two concepts are to be separated, data mining is a core step in the KDD process.

KDD-based activities are mainly about extracting knowledge from data. In this case, knowledge is the relationship and common patterns between data elements. What counts as knowledge must be something new which is difficult to perceive by sensory and usable. In the KDD process, data mining is used separately at the knowledge discovery stage. The KDD environment covers many areas: Machine learning, statistics, expert system.

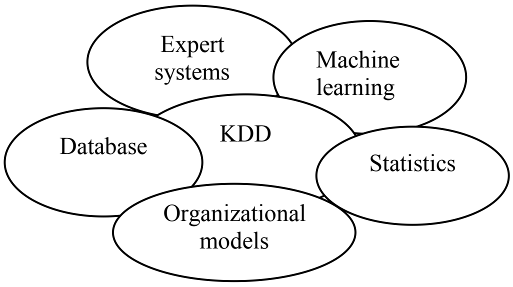


Figure 1. Data mining model

Data mining uses a combination of an explicit knowledge base, sophisticated analytical skills, and domain knowledge touncover hidden trends and patterns. These trends and patterns form the basis of predictive models that enable analysts to produce new observations from existing data. Data mining models and algorithms: Models house the steps, modules, and resources of the data mining process. Some data mining models include the entireprocess for a particular purpose, be it to cluster or predict. A model is, however, different from an algorithm [6].

III. DATA MINING AND ALGORITHMS

Data mining: This is the most important and time-consuming step of the knowledge discovery process, applying mining techniques (most are machine learning techniques) to exploit and extract patterns of information and special relationships in the data.

Knowledge representation & evaluation: Using data display techniques to present information samples (Knowledge) and special relationships found in the data, exploited in the previous step in a form that is familiar to users such as graphs, trees, tables, rules, etc. At the same time, this step also assesses the knowledge discovered according to certain criteria.

During the data mining phase, user interaction may be required to tune and extract the required knowledge. The knowledge received can also be saved and reused.

We have I = {I1, I2, ..., Im}, a set with m number of items (or attributes) $X \subseteq I$ is called an itemset

$T = \{T_1, T_2, \ldots, T_n\}$: a set with m number of transactions (or records), each transaction is defined by TID (Transaction Indentification) [8].

R is a binary relation above I and T (or $R \subseteq IxT$). If transaction T includes item I, we can write $(I,T) \in R$. A database D, in terms of form, is a binary relation R as above. In terms of meaning, a database is a set of transactions, each transaction T is an itemset, $T \in 2^1$ (2^1 is a subset of I).

Example of database:

I = {a, b, c, d, e}, T = {1, 2, 3, 4, 5, 6}

Info on transactions given in table 1.

Transaction Indentification (TID)	Itemset
1	a b d e
2	b e
3	e b d e
4	a b c e
5	a b c d e
6	b c d

Table 1: Example of a transactional database - D

We have an itemset $X \subseteq I$.

The notation s(X) being the support of itemset X - the percentage of transactions in database D containing X on the total number of transactions in database D.

s(X) = Card (X) / Card (D)%.

3.1. Frequent set

We have an itemset $X \subseteq I$ and Minimum Support threshold minsup $\in$ (0,1) defined by user. An itemset X is called a frequent itemset on minsupp threshold if and only if its support is greater than or equal to a minsupp threshold. $s(X) \geq$ minsupp.

The notation FX (T, I, R, minsupp) is a set of frequent set on the minsupp threshold, meaning

FX(T, I, R, minsupp) = $\{X \subseteq I \mid s(X) \geq \text{minsupp}\}$

For (T, I, R) in table 1 database and minsupp threshold value = 50%, all frequent itemsets will be listed as in table 2:

Frequent itemsets	Corresponding support
b	100% (6/6)
e, be	83% (5/6)
a, c, d, ab, ae, bc, bd, abe	67% (4/6)
ad, ce, de, abd, ade, bce, bde, abde	50% (3/6)

Table 2. Frequent sets from table 1 database with minsupp = 50%

The support level s of association rule $X \rightarrow Y$ is the percentage of transactions in D containing X and Y

$s(X \rightarrow Y) = Card(X \cup Y) / Card(D)\%$.

association rule in $X \xrightarrow{c} Y$ form in which:

X and Y are itemsets satisfying $X \cap Y = \emptyset$

c being the rule's reliability.

$c = s(X \cup Y) / s(X)\%$ ($c = Card(X \cup Y) / Card(X)\%$): is the percentage of transactions in D containing X that contains Y. In terms of probability, reliability c of a rule is the probability (with conditions) that results in Y given that X has occurred.

3.2. Reliable association rule

A rule is considered reliable if the reliability c is greater than or equal to a minimum confidence threshold minconf ∈ (0,1].c ≥ minconf). The minconf threshold reflects the likelihood of the given Y.

The association rule to look for is one that matches the given minsupp and minconf. (Only pay attention to rules with a support greater than the minimum support and reliability greater than the minimum confidence).

The problem of solving association rule (in simple form) is as follows: We have a database D, with minimum support minsupp, minimum confidence minconf. Find all association rules in the form of $X \rightarrow Y$ satisfying support $s(X \cup Y) \geq$ minsupp, with confidence $c(X \rightarrow Y) = s(X \cup Y)/ s(X)$, $c \geq$ minconf.

Most proposed algorithms for mining association rules are divided into two steps:

Step 1: Finding all frequent itemsets from the database, meaning finding all itemsets X satisfying $s(X) \geq$minsupp.

Step 2: Generate reliable rules from frequent itemsets found in step 1.

If X is a frequent itemset, the association rule generated from X has the form:

$X' \xrightarrow{c} X \setminus X'$, in which:

X' is another null subset of X.

X \ X' is the difference between two sets X and X'.

c being the rule's credibility satisfying c ≥ minconf.

Example: For frequent sets as shown in Table 2, the minimum support minsupp = 50% and the minimum confidence minconf = 70%. Considering the abe frequent itemset with 67% support, it is possible to generate association rules from the abe frequent itemsets as in table 3.

Association Rule	Credibility c ≥ minconf
$a \xrightarrow{100\%} be$ (c = s(abe)/s(a) = 100%	Yes
$a \xrightarrow{67\%} ae$	No
$e \xrightarrow{80\%} ab$	Yes
$ab \xrightarrow{100\%} e$	Yes
$be \xrightarrow{80\%} a$	Yes
$ae \xrightarrow{100\%} b$	Yes

Table 3: Association rule generated from abe frequent set

3.3. Max pattern:

We have M ∈ FX (T, I, R, minsupp), M being called a max pattern if there is no X ∈ FX (T, I, R, minsupp), M ≠ X, M ⊂X.

IV. CRITERIA FOR ASSESSING AND CLASSIFYING STUDENTS

Data mining is a powerful tool for academic intervention. Through data mining, a university could, for example, predict with 85 percent accuracy which students will or will not graduate. The university could use this information to concentrate academic assistance on those students most at risk.

In order to understand how and why data mining works, it's important to understand a few fundamental concepts. First, data mining relies on four essential methods: Classification, categorization, estimation, and visualization. Classification identifies associations and clusters, and separates subjects under study. Categorization uses rule induction algorithms to handle categorical outcomes, such as "persist" or "dropout," and "transfer" or "stay." Estimation includes predictive functions or likelihood and deals with continuous outcome variables, such as GPA and salary level. Visualization uses interactive graphs to demonstrate mathematically induced rules and scores, and is far more sophisticated than pie or bar charts. Visualization is used primarily to depict three-dimensional geographic locations of mathematical coordinates.

Higher education institutions can use classification, for example, for a comprehensive analysis of student characteristics, or use estimation to predict the likelihood of a variety of outcomes, such as transferability, persistence, retention, and course success [5].

Throughout the learning process, students' learning and training results are influenced by various factors in the environment, themselves, their families, society, etc.

I would like to offer some key elements to assess and classify students more comprehensively through the 5 following criteria:

- All-year academic performance: Excellent, Good, Average, Weak.
- All-year conduct ratings: Excellent, Good, Average, Weak.
- Community activities (school relationships, youth union participation, etc.): Excellent, Good, Average, Weak.
- Discipline (Number of absence, willingness to study, misconducts, etc.) Excellent, Good, Average, Weak.
- Conditions of learning: Excellent, Good, Average.

The above criteria are obtained through examining the learning and training process at the school, circulars of the Ministry of Education and Training, student surveys, etc.

Thereby, we can assess students' ability and classifying them as satisfactory or unsatisfactory.

4.1. Data warehouse: a container of data relating to the number of students enrolled, graduating, academic results, training results, related documents, etc. After the pre-processing stage, the data is written to text file as follows:

```
081350411101  2  3  4  4  3
081350411104  1  2  3  4  2
081350411107  2  3  4  4  3
081350411108  3  4  2  4  2
081350411110  3  4  4  4  3
081350411111  2  4  3  4  4
```

In it, each line is the result of a student's rating. Ratings are coded as numerical as good, good corresponding to number 4, fairly corresponding to number 3, average to number 2, weak to 1.

The hypothetical data set is based on the collection of a number of results according to the evaluation and classification criteria given by high school students at the Technical Pedagogical University, the student survey online and online Internet [2].

4.2. System modeling stage

Applying Apriori algorithm to the processed data.

Building association rules.

Boolean operators:

OR: generates plural, with either conditions

AND: narrow the range to search for rules.

4.3. Some conventions in the program:

Rating academic performance: performance excellent, performance good, performance average, performance weak correspond to transactions a, b, c, d.

Rating conducts (practice): practice excellent, practice good, practice average, practice weak correspond to transactions g, h, i, j.

Rating community adaptability (via questionnaires for community activities): adaptability excellent, adaptability good, adaptability average corresponds to transactions k, m, n.

Rating discipline: discipline excellent, discipline good, discipline average, discipline weak correspond to transactions q, r, s, t.

Rating conditions of learning (via questionnaires): condition excellent, condition average, condition weak correspond to transactions u, v, x, y, z.

4.4. Results tallying and assessment stage:

Achievements of the system:

Calculate the support of each transaction, for example transaction b's support level was 43, c's was 53, etc.

We enter min. support and min. confidence, and after execution, the algorithm generates rules.

There are clear advantages when using this system to assess compared to traditional methods in that it is possible to predict certain outcomes, for instance students with good academic performance may have excellent conduct, or students with good academic performance and excellent community adaptability will have excellent conduct rating. As a result, as soon as students are enrolled, the school can determine appropriate instructions and teaching methods for each of them, and the students will have a chance to develop in every respect.

4.5. Building the program:

After the preprocessing stage, the database is stored as texts.

Rules are in the following form: X → Y (X being the premise, Y the conclusion), X and Y are attribute sets. The association rule shall have supp and conf values.

V. METHODS

Apriori is a solvable algorithm first proposed by Rakesh Agrawal, Tomasz Imielinski and Arun Swami in 1993. The algorithm finds transaction t with a level of support and reliability greater than a certain threshold value.

The algorithm trims candidate sets with infrequent subsets before calculating the support level.

L_k: Collection of k sets-frequent items (with a certain minsup)

C_k: Collection of k sets-candidate entries (potential frequent itemsets)

5.1. The Apriori algorithm:

Input: Database D, minsup.

Output: Set of popular item sets.

```
$L_1$ = {1 - itemset popular};
k=2;
While ( $L_{k-1}$! =∅)
{$C_k$ = apriori_gen($L_{k-1}$, minsup5. for( ∀deal t∈ D)
{$C_t$=Subset ($C_k$,t);t
   for ( ∀candidates c ∈ $C_t$)
 c.count ++;
}
$L_k$={ c ∈ $C_k$ /c.count ≥minsup}
 k++;
 }
 Return L= $\cup_k L_{k'}$;
// new candidate (**)
Void apriori_gen($L_{k-1}$, minsup )
{ for (∀itemset $l_1$∈ $L_{k-1}$)
for (∀itemset $l_2$∈ $L_{k-1}$)
if (($L_1$(1) == $L_2$(1)&&$L_1$(2) == $L_2$(2)&&...&& $L_1$(k-2 == $L_2$(k-2)) &&$L_1$(k-1) == $L_2$(k-1))
        {c= $L_1$ connect $L_2$;
        if( has_inrequent_subset(c, $L_{k-1}$)) delete c;
        else add c to $C_k$;
}
return $C_k$
}
Boolean  has_infrequent_subset(c, $L_{k-1}$)
{ for (∀(k-1)-subset s∈ c)
   If (s ∉ $L_{k-1}$) return TRUE;
   else return FALSE ;
}
```

5.2. Data structure:

The Apriori algorithm is highly effective by gradually reducing the size of candidate sets. However, in case the database exploited has many frequent itemsets, the itemsets are large in size or have small minsupp, the Apriori algorithm incurs two important costs:

The cost of handling a large number of candidate sets.

The cost of repeatedly approving the database and examining a large number of candidates by matching patterns.

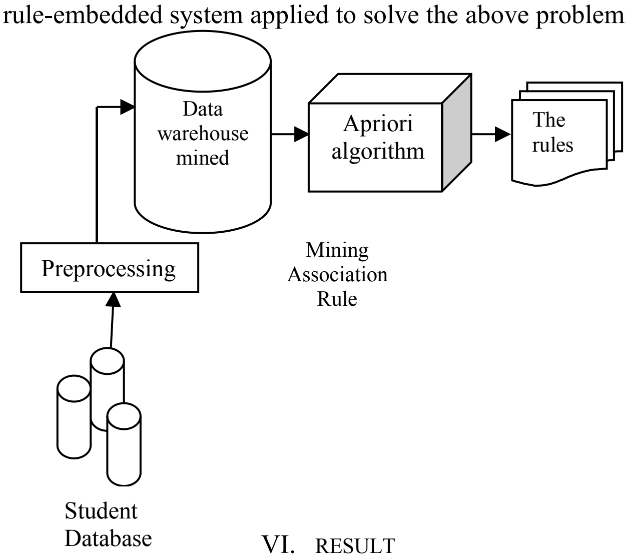
Thus, the problem of huge cost for Apriori algorithm in discovering frequent itemsets is caused by the occurrence and reviewing of candidate sets. If we try to limit the generation of candidate sets, the problem above is basically settled, as a result we can build a Apriori algorithm faster.

In order to achieve convincing assessments and classifications, we need to have effective data mining methods, from which to draw scientific conclusions and bring about practical significance.

On that basis, educational institutes must adopt future plans to improve their quality and meet the demands of today's society such as: improving teachers' proficiency, equip facilities that meet social requirements, employ both theory and practice, increase the number of scholarships for students with good and excellent academic performance, etc.

Acquiring the above information at such high demand is impossible if we use traditional tallying methods to render it. Therefore, it is necessary to use data mining - association rules.

Stages of implementation for the association rule-embedded system applied to solve the above problem:



VI. RESULT

I used Apriori-TID algorithm and built the program on Visual Studio 2008; database was collected at the Training Department at the University of Technology and Education in Excel format. After the preprocessing stage, the data is stored as texts.

Entering minimum support value (minsup): 20%

Entering minimum confidence value (mincof): 70%

The program results executed has resulted in the following rule interface:

Frequency of occurrence	
Attribute	Support
b	39
c	54
d	24
h	34
i	71
m	61
Association rule	

Rule	Credibility	Elaborating generated association rule
b -> h	79.49%	Performance_Good -> Practice_Good
c -> i	100.00%	Performance_Good -> Practice_Excellent
cn->i	100.00%	Excellent Performance_Excellent conditions -> Practice_Excellent
cs ->i	100.00%	Performance_Specialized practice_excellent -> Practice_Excellent

There are 2 parts to the results:

Part 1: The frequency of occurrence includes attributes and support

Part 2: Generated association rule

- If academic performance is average, discipline is good, conf: 79.49%

- If academic performance is good, discipline is excellent, conf: 100%

- If academic performance is good and adaptability is excellent, discipline is excellent, conf: 100%

- If academic performance is good and specialized practice is good, discipline is excellent, conf: 100%

Therefore, with minsup value of: 20% and minconf: 70%, the rules above are generated and the students are all qualified graduates in terms of academic performance, discipline and specialized practice.

VII. CONCLUSION

In this article, I have presented the general idea of data mining and Apriori-TID algorithm, including knowledge discovery, the approach and research of data mining to build a multi-dimensional data warehouse in which exploring association rules to build a comprehensive assessment and classification system for students is an important method of knowledge discovery in Data Mining, which is also the focus of the article.

REFERENCES

[1] PhridviRaj MSB., GuruRao CV (2013) *Data mining – past, present and future – a typical survey on data streams.* INTER-ENG Procedia Technology 12:255 – 263

[2] Srivastava S (2014) Weka: *A Tool for Data preprocessing, Classification, Ensemble, Clustering and Association Rule Mining*. International Journal of Computer Applications (0975 – 8887) 88:.10 [3] Soni N, Ganatra A (2012)

[3] Jing Luan, PhD. *Data Mining Applications in Higher Education, Chief Planning and Research Officer, Cabrillo College Founder, Knowledge Discovery Laboratories.*

[4] Chengqi Zhang, Shichao Zhang. (2002). *Association rule mining - models and practice.*

[5] Goulbourne, G., Coenen, F. and Leng, P. (2000). "*Algorithms for Computing Association Rules Using a Partial-Support Tree*", Journal of Knowledge-Based Systems, pp141-149.

[6] [6] Han, J., Pei, J. and Yiwen, Y. (2000). "*Mining Frequent Patterns Without Candidate Generation*". Proceedings ACM-SIGMOD International Conference on Management of Data, ACM Press, pp1-12.

[7] Coenen, F., Goulbourne, G. and Leng, P., (2003). *Tree Structures for Mining Association Rules.* Journal of Data Mining and Knowledge Discovery, Vol 8, No 1, pp25-51.

[8] Han, Jiawei, Micheline Kamber, and Jian Pei. "Data mining: concepts and techniques. 2001." San Francisco: Morgan Kauffman (2006).