Association Analysis for an Online Education System

Behrouz Minaei-Bidgoli[1], Gerd Kortemeyer[2], and William Punch[1]
[1]*Computer Science Department, Michigan State University, East Lansing, MI, 48824, USA*
{minaeibi, punch}@cse.msu.edu
[2]*College of Natural Science, LITE lab, Michigan State University, East Lansing, MI 48824, USA*
korte@lite.msu.edu

Abstract

An important goal of data mining is to discover the unobvious relationships among the objects in a data set. Web-based educational systems collect vast amounts of data on user patterns, and data mining methods can be applied to these databases to discover interesting associations between student attributes, problem attributes, and solution strategies. In this paper, we propose a framework for the discovery of interesting association rules within a web-based educational system. A hybrid measure of subjective and objective measure for rule interestingness is proposed which is called contrasting rules. Contrasting association rule is one in which a conjunction of attributes is compared for complementary subsections of a data set. We provide a new algorithm for mining contrasting rules that can improve these systems for both teachers and students – allowing for greater learner improvement and more effective evaluation of the learning process. A larger advantage of developing this approach is its wide application in any other data mining application.

1. Motivation

The growth of the world wide web has a great impact on the education arena. Recently developed online education systems allow researchers to study how students learn (descriptive studies) and which learning strategies are most effective (causal/predictive studies). Michigan State University (MSU) has pioneered systems to provide an infrastructure for online instruction. The research presented here was performed on a part of the latest online educational system developed at MSU, the *Learning Online Network with Computer-Assisted Personalized Approach* (*LON-CAPA*) [1-3]. LON-CAPA involves three types of large data sets: 1) educational resources such as web pages, demonstrations, simulations, and individualized problems designed for use on homework assignments, quizzes, and examinations; 2) information about users who create, modify, assess, or use these resources; and 3) activity log databases which log actions taken by students in solving homework and exam problems. In other words, we have three ever-growing pools of data. This paper investigates methods for extracting useful and interesting patterns from these large databases using online educational resources and their recorded paths within the system. Our research is guided and inspired by the following questions:

Can we find any associative rules between the attributes of students, problems within their courses, and the methods they use to solve them? How do contrasting groups differ in a particular course? Which attributes are associated with course success? For example, is there a relationship between gender and course success? How can we compare data from the same course between sections or over multiple semesters, in addition to examining attributes across courses, at a global scale?

Based on the current state of the student in their learning sequence, as well as other student attributes, the system could then make suggestions for improving student performance and course design. As more and more students enter the online learning environment, databases concerning student access and performance will grow – yielding greater pattern clarity. We develop such techniques in order to provide information that can be usefully applied by instructors to increase student learning.

2. Background

This section states a formal definition of association analysis and contrasting rules. It also provides a descriptive model of different data attributes in order to supply a formal statement of the problem.

2.1. Association rules

The task of discovering association rules was initiated in [4]. With its main use in business environments, association rule mining is focused on market "basket data" which stores items purchased on a per-transaction basis – similar to the "shopping cart" from an online store. A typical example of an association rule regarding market "basket data" is that 68% of a book store's customers who purchase a book on HTML also purchase a book on XML. Finding association rules is a valuable source for many applications in the business arena as well as medical diagnosis and remote sensing. It is precisely the lack of association rule mining in the field of education that led to this study.

The time computational complexity is mainly determined by the first step, which is the generation of frequent itemsets. According to [4], the basic formal definition of the association rule is as follows: Let $I = \{i_1, i_2, \ldots, i_m\}$ be the set of all items and $T = \{t_1, t_2, \ldots, t_N\}$ the set of all transactions where m is the number of items and N is the number of transactions. Each transaction t_j is a set of items such that $t_j \subseteq I$. Each transaction has a unique identifier, is referred to as TID. An *association rule* is an implication statement of the form $X \Rightarrow Y$, where $X \subset I$, $Y \subset I$, and X and Y are disjoint, that is, $X \cap Y = \varnothing$. X is called the antecedent while Y is called the consequence of the rule.

There are two basic measurements for each rule, support and confidence. The rule $X \Rightarrow Y$ has support, s, in the transaction set, T, if s% of transactions in T contains $X \cup Y$. The rule has *confidence*, c, if c% of transactions in T that contain X also contains Y. Support indicates how frequently the pattern occurs, while confidence indicates the strength of the rule [5].

In other words, support measures the fraction of transactions that contain all items belonging to the set $X \cup Y$. Confidence measures the fraction of times the itemset Y is present in transactions that contain X. Formally, these measurements are defined as follows: support, $s(X \Rightarrow Y) = \frac{s(X \cup Y)}{N}$ and confidence, $c(X \Rightarrow Y) = \frac{s(X \cup Y)}{s(X)}$ where N is the total number of transactions [5].

2.2. Rule interestingness

The techniques which mine association rules often generate too many rules; while most of the rules are useless to the user, manual inspection of the rules' interestingness is usually a time-consuming, difficult task. This problem sometimes is called post-mining rule analysis [6]. In literature different measures are proposed to discover the interestingness of a rule. *Rule templates* [7-8] is a technique that separates only those rules that match the template. *Actionability* [9-10] is a subjective measure of the benefit/advantage obtained by applying a rule. *Unexpectedness* [6, 11] is interpreted either in the probabilistic sense or in regards to the user's beliefs. *Neighborhood-based interestingness* [12] defines interestingness within a set of rules in terms of their density and relative confidences. Though all of these methods are valid within certain regimes, we suggest a new method of determining a rule's interestingness that will expand the current repertoire and be applicable in all data mining applications.

Bay and Pazzani [13] presented the definition of contrast sets as a conjunction of attributes and values that differ meaningfully in their distribution across groups. They developed the STUCCO (Search and Testing for Understandable Consistent Contrast) algorithm to find Significant Contrast Sets. They use a chi-square test for testing the null hypothesis that contrast-set support is equal across all groups. The goal in this work is to find such surprisingly contrasting sets. We extend the idea of contrasting sets to the discovery of contrasting rules, introducing new measures for finding the significant differences between the groups of rules for contrasting elements.

3. Problem Statement

Let D be a data set of N transactions with d-dimensional attributes. The data model for contrasting rules is a generalization of the association data model for the grouped categorical attributes. Let B be an attribute with k mutually exclusive elements. Let A be an attribute or any conjunction of the attributes such that $A \cap B = \varnothing$. We define the contrasting rule as follows:

$$\begin{cases} A \Rightarrow B_1 \\ A \Rightarrow B_2 \\ \ldots \\ A \Rightarrow B_k \end{cases}$$

where A is a possible conjunction of at most d-1 attributes; B and $\overline{B}$ are the elements of a of the target class and $1 \le i, j \le k$ and $\forall i \neq j : B_i \cap B_j = \varnothing$.

Suppose $A \Rightarrow B_i$ and $A \Rightarrow B_j$ are two rules, and let Ω be a ranking measure for rules (explained in 2.3).

$A \Rightarrow B_i$ and $A \Rightarrow B_j$ are contrasting rules if and only if $| \Omega(A \Rightarrow B_i) - \Omega(A \Rightarrow B_j) | \ge \sigma$, where σ is a user defined threshold, which implies that there is high gap between both support and confidence of these two rules.

In the most basic case, when k=2, the contrast rules would be: $\begin{cases} A \Rightarrow B \\ A \Rightarrow \overline{B} \end{cases}$

As an example, consider gender as the attribute of B, while A can be the conjunction of a student's grade point average (GPA) and success within a particular course (pass/fail). If an instructor is interested in the difference between males and females who pass his or her course and have a particular GPA, then the above contrasting rule is of value.

3.2. Criteria for Ranking the Rules

Let D is a data set N transactions, and let $D_1 \subset D$ be the subset of D which includes B_i, $N_1 = |D_1|$, and $D_2 \subset D$ be the subset of transactions which includes B_j, $N_2 = |D_2|$; n_1 is the number of transactions of D_1 that includes $A \cup B_i$ and n_2 is the number of transactions of D_2 that includes $A \cup B_j$ as it shown in figure 1.

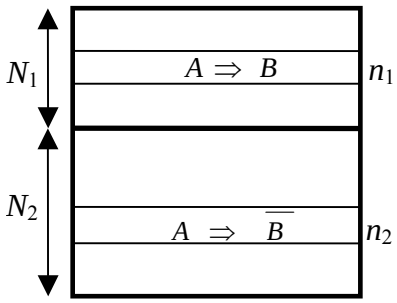

Figure 1. An illustration of operationalized contrasting rules

We propose some functions to measure the rules and rank them. For example, in the case that the contrasting group is gender; we divide the transactions into male and female subsets. Methodological detail will be explained in the experimental study, but first a discussion of rule ranking criteria is necessary.

1. *Difference of proportions:*

The difference shown in Eq. (1) expresses the coverage of the rule between B_i and B_j – since it is the difference in ratios of the "affected" to the whole population for each.

$$\left| \frac{n_1}{N_1} - \frac{n_2}{N_2} \right| \tag{1}$$

2. *Log Odds Ratio:*

Odds ratio shows how two proportions differ. Let $p=n_1/N_1$, $q=n_2/N_2$ and "odds ratio"$=(p/(1-p))/(q/(1-q))$. When p and q are equal, then their odds ratio is equal to 1 (not significant); when it is not equal to 1 it shows that the proportion of the one element of a contrasting group is greater than another. The value range for this ratio is between zero and infinity, with a value of one implying a balance between diagonals in the contingency table. Unfortunately, as though any two comparable odds ratio values will lie in ranges of differential size (zero to one versus one to infinity).

$$|log\ (p/(1-p))/(q/(1-q))| \quad \text{where } p = n_1/N_1 \text{ and } q = n_2/N_2 \tag{2}$$

The logs odds ratio (Eq. 2) is in the interval $(-\infty, +\infty)$, when p and q are equal it will be equal to zero. The advantage for these criteria is the improvement over the simple odds ratio – the scale of output is smoother. Yet, it is not without problems. As a result of improving the output scale, the speed of convergence for more significant rules is low.

3. *Chi-Square value:*

$$\chi^2 = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{3}$$

One of the most common tests for significance is shown in Eq. (3) – the chi-square test – where i is the number of rows and j is the number of columns in the contingency table. We implemented all these measures in this paper and the results and the comparison will be discussed in experimental study section.

4. Methodology

Association analysis is not an easy task for many applications of data mining. For example, using an Apriori algorithm with very low support to find patterns is quite difficult. From one side, if we put the minimum-support high enough, we lose many interesting, but low-support patterns. On the other hand if we choose a very low minimum-support the Apriori algorithm will find too many rules and finding the interesting rules becomes a very hard task.

We propose a new method to discover hidden patterns, even those with low support. An automatic rule miner finds the common rules amongst the contrast elements. We develop an algorithm, Mining Contrasting Rules (MCR) to discover the association rules for the contrast elements. Here, we describe the framework (see Figure 3) in which this algorithm can work for the purpose of discovering association rules while mining for contrasting groups:

- Selecting data from course and students databases
- Preprocessing; cleansing data
- Attribute subset extraction/selection
- Discretizing the continuous attributes
- Pruning the values of attribute with very high support
- Select an interesting contrast group
- Applying the MCR algorithm given a contrasting attribute
- Post-processing to identify the rule interestingness
- Select another measure or contrast group and repeat the procedure

Figure 2. Framework to Mine Contrasting Rules

Detail of this framework will be explained in the experimental study. Now we explain the MCR algorithm as shown in Figure 4.

As it is explained in the algorithm we divide data set D, into M disjoint subsets. In the case of that the contrasting group is gender; we divide the transactions into male and female subsets. We use the Apriori[1] algorithm in order to find the closed frequent itemset in

[1] We used the C. Borgelt's implementation of Apriori version 4.19.

each subset. The advantage of using closed frequent itemsets is stated in [5].

Input:
D – Input set of *N* transactions of students per problem
A – Interested attribute includes contrast groups
σ – Minimum (very) low support
Ω – A measure for ranking the rules
k – Number of the most interesting rules
M – Number of contrasting elements to be compared
Divide data set *D* based on contrasting elements into *M* spaces
for *j* = 1 to *M*
  Find the close frequent itemsets for *D*(*j*) given σ (Apriori)
  Generate possible rules for *D*(*j*) based on the frequent itemsets
end
Find common rules among the *M* contrast groups
Rank the common rules with respect to the Ω
Sort the rules with respect to their rank; Select *k-top* rules;
Validate selected rules, *R*, as a candidate set of interesting rules (optional)
return *R*

Figure 4. Mining Contrasting Rules (MCR) algorithm for discovering the interesting rules candidates

We choose a very low minimum support because we need to obtain as many frequent itemsets as is possible. Using perl scripts, we find the common rules between two contrasting subsets. Finally, we rank the common rules with all of the measures explained in the previous section, and then the *k*-top rules of the sorted ranked-rules are chosen as a candidate set for interesting rules.

Therefore an important parameter for this algorithm is minimum support σ; the lower the σ, the larger number of common rules. If the user selects a specific ranking measure Ω, then the algorithm will rank the rules with respect to that measure.

5. Experiments

In this section first we provide a general model for data attributes, data sets and their selected attributes, and then, we discuss the results and experimental issues.

5.1. Data model

In order to better understand the interactions between students and the online educational system, a model is required to analyze the data. Ideally, this model would be both descriptive and predictive in nature. 1

As shown in Figure 5 (Appendix A), each student is characterized by a set of attributes which are static for any particular analysis (GPA, gender, ethnicity, etc.) and can be easily quantized. The *u*-tuple ($S_i^{(1)}, S_i^{(2)}, \ldots, S_i^{(u)}$) describes the characteristics of the *i*-th student. The set of problems is determined by the scope of the analysis – at this time, single courses over individual terms, but with future possibilities for multi-term analysis – and characterized by a set of attributes, some of which are fixed (Bloom's taxonomic categorization, content type, simulation-dependent, etc.). The *v*-tuple ($P_j^{(1)}, P_j^{(2)}, \ldots, P_j^{(v)}$) describes the characteristics of the *j*-th problem.

The interaction of these two sets becomes a third space where larger questions can be asked. The *k*-tuple ($SP_{ij}^{(1)}, SP_{ij}^{(2)}, \ldots, SP_{ij}^{(k)}$) describes the characteristics of the *i*-th student linking to the *j*-th problem. LON-CAPA records and dynamically organizes a vast amount of information on students' interactions with and understanding of these materials. We can extract from these logged data sets, many features which belong to this third space of attributes.

5.2. Selected attributes

We have extracted the following attributes per student per problem from the activity log:

- Total number of attempts before correct answer is derived
- Success on the problem
- Total time from first attempt until the correct answer
- GPA
- Major
- Ethnicity
- Gender
- Level Transferred (LT)GPA (i.e. High School GPA)
- Student's Age
- Student's Grade

An aggregation of "grade" attributes must be added to the total attribute list. Besides the 9 possible labels for grade (a 4.0 scale with 0.5 increments), we can group the students regarding their final grades into the 2-Classes (Failed, Passed): Categorize students with one of two class labels: "*Passed*" for grades above 2.0, and "*Failed*" for grades less than or equal to 2.0.

5.3. Data sets

For this paper we selected two student/course data sets of LON-CAPA courses, which were held at MSU in fall semester 2003 as shown in Table 1: LBS271 is a Physics course with 200 students integrated 174 online homework problems, used LON-CAPA. This course has an activity log with approximately 152 MB. However it is much smaller than CEM141, general chemistry I, which 2048 student enrolled for this course and activity log of this course exceeds 750MB and includes more than 190k transactions of per student per problem records.

For this paper we selected two contrast groups, gender and 2-Classes, in order to find the contrasting rules for the elements for these two sets. The count and percentage of these elements for these three courses are shown in table 2. The proportion of "female" in all three courses is

greater than the "male". The proportion of "passed" is greater than the "failed" in LBS271.

Table 1. Characteristics of two of MSU courses which used LON-CAPA in fall semester 2003

Data set	Course Title	# of Students	# of Problem	Size of Activity log	# of Transactions
LBS 271	Physics_I	200	174	152.1 MB	32,394
CEM 141	General Chemistry_I	2048	114	754.8 MB	190,859

Table 2. Characteristics of three of MSU courses which held by LON-CAPA in fall semester 2003

Data set	Female	Male	Passed	Failed
LBS 271	20,468 63.6%	11,696 36.4%	29,412 91.4%	2,752 8.6%
CEM 141	106,296 55.7%	84563 44.3%	121,540 63.7%	69,319 36.3%

5.5. Results[2]

The question that arises here is that how we can determine whether an unexpected rule is a candidate for "interesting" status or not. We need criteria to discover the surprising rules, methods for finding the greatest difference between two contrasting elements. We can divide the set of discovered rules into three categories:

1. *Expected and previously known*: This type of rule confirms user beliefs, and can be used to validate our approach. Though perhaps already known, many of these rules are still useful for the user as a form of empirical verification of expectations. For our specific situation (education) this approach provides opportunity for rigorous justification of many long-held beliefs.
2. *Unexpected*: This type of rule contradicts user beliefs. This group of unanticipated correlations can supply interesting rules, yet their interestingness and possible actionability still requires further investigation.
3. *Unknown*: This type of rule does not clearly belong to any category, and should be categorized by domain-specific experts. For our situations, classifying these complicated rules would involve consultation with not only the course instructors and coordinators, but also educational researchers and psychologists.

Figure 5 provides seven examples of obtained rules running the MCR algorithm, the outputs of which are in italics. The "coverage" of a rule over a related subset shown in brackets represents the fraction of transactions that hold true for the left-hand side of the rule [14]. "Support" and "confidence" of the rule are denoted in parentheses by the values of *s* and *c*, and are both evaluation of rule quality.

The examples in Fig. 5(a) suggest that a student with level-transfer-GPA between 3.0 and 3.5 is more likely to pass the course. Thus, these rules likely belong to the first category, since there is a well-established correlation between high grades and course success. Rules in Fig. 5(b) suggest that a student with GPA between 3.0 and 3.5 who attempts a problem more than 10 times is more likely to pass the course. Since successful students might be assumed to succeed on problems more quickly, these rules might belong to the second (unexpected) category. The remaining rules in Fig. % are completely open for interpretation, and therefore are placed into the third category. It is interesting to note that rules generated by the difference of proportion criterion tend to have significantly higher coverage than those of the chi-square and log odds ratio criteria.

(a) CEM141 data, using difference of proportion

(Lt_GPA=[3,3.5)) ==> Passed
[44187 (36.4)%] (s=23.2%, c=87.6%)
(Lt_GPA=[3,3.5)) ==> Failed
[6283 (9.1)%] (s=3.3%, c=12.4%)

(b) CEM141 data, using log odds ratio

(GPA=[3,3.5) & Tries>=10) ==> Passed
[1272 (1.0)%] (s=0.7%, c=83.8%)
(GPA=[3,3.5) & Tries>=10) ==> Failed
[245 (0.4)%] (s=0.1%, c=16.2%)

(c) CEM141 data, using chi-square value

(Ethnicity=Asian & GPA=[3,3.5) & Sex=Male) ==> Passed
[1236 (1.0)%] (s=0.6%, c=85.7%)
(Ethnicity=Asian & GPA=[3,3.5) & Sex=Male) ==> Failed
[206 (0.3)%] (s=0.1%, c=14.3%)

(d) CEM141 data, using log odds ratio

(GPA=[3,3.5) & Sex=Male & Time=1_20_hours) ==> Passed
[1156 (1.0)%] (s=0.6%, c=92.2%)
(GPA=[3,3.5) & Sex=Male & Time=1_20_hours) ==> Failed
[98 (0.1)%] (s=0.1%, c=7.8%)

(e) LBS271 data, using chi-square value

(Major=PREDENTAL & Time=1_5_minutes) ==> Passed
[122 (0.1)%] (s=0.1%, c=63.5%)
(Major=PREDENTAL & Time=1_5_minutes) ==> Failed
[70 (0.1)%] (s=0.0%, c=36.5%)

(f) LBS271 data, using difference of proportion

(Age=20 & GPA=[3.5,4] & Tries=1) ==> Male
[934 (8.0)%] (s=2.9%, c=20.7%)
(Age=20 & GPA=[3.5,4] & Tries=1) ==> Female
[3586 (17.5)%] (s=11.1%, c=79.3%)

(g) LBS271 data, using log odds ratio

(Age=20 & Lt_GPA=[3,3.5) & Time>20_hours) ==> Passed
[1216 (1.0)%] (s=0.6%, c=85.4%)
(Age=20 & Lt_GPA=[3,3.5) & Time>20_hours) ==> Failed
[208 (0.3)%] (s=0.1%, c=14.6%)

Figure 4. Examples of obtained binary rules using different criteria

[2] Experiments were conducted on a 1.7 GHz Pentium 4 PC running RedHat Linux 7.3 kernel x-2.4.20-19 with 1GB RAM

In conclusion, LON-CAPA servers are tracking students' activities in large logs. We developed an algorithm to discover a set of surprising contrasting rules. This tool can help instructors to design courses more effectively, detect anomalies, and help students use resources more efficiently.

6. Acknowledgment

This work was partially supported by the National Science Foundation under ITR 0085921.

7. References

[1] Kortemeyer, G., Bauer, W., Kashy, D.A., Kashy, E., and Speier, C., "The LearningOnline Network with CAPA Initiative", *IEEE Frontiers in Education Conference Proceedings*, vol. 31,(2001) p. 1003. See also www.loncapa.org.

[2] Minaei-Bidgoli, B., Punch, W.F., "Using Genetic Algorithms for Data Mining Optimization in an Educational Web-based System", In Proc. of *the Genetic and Evolutionary Computation Conference GECCO* 2003, pp. 2252-2263.

[3] Minaei-Bidgoli, B., Kashy, D.A., Kortemeyer, G., Punch, W.F., "Predicting Student Performance: An Application of Data Mining Methods with an educational Web-based System", *(IEEE/ASEE) FIE 2003 Frontier In Education*, Nov. 2003 Boulder

[4] Agrawal, R., Imielinski, T.; Swami A., "Mining Associations between Sets of Items in Massive Databases", *Proc. of the ACM-SIGMOD 1993 Int'l Conference on Management of Data*, Washington D.C., May 1993.

[5] Agrawal, R., Srikant, R. "Fast Algorithms for Mining Association Rules", *Proceeding of the 20th International Conference on Very Large Databases*, Santiago, Chile, September 1994.

[6] Liu B., and Hsu. W., Post-analysis of learned rules. In *Proceedings of AAAI*, pages 828-834, 1996.

[7] Fu Y. and Han, J., "Meta-rule-guided mining of association rules in relational databases". *Proc. 1995 Int'l Workshop. on Knowledge Discovery and Deductive and Object-Oriented Databases (KDOOD'95)*, Singapore, December 1995, pp. 39-46.

[8] Klemettinen, M., Mannila, H., Ronkainen, P., Toivonen, H., and Verkamo, A. I. "Finding interesting rules from large sets of discovered association rules". In *Proceedings of the Third InternationalConference on Information and Knowledge Management*, pages 401-407, Gaithersburg, Maryland, 1994.

[9] Piatetsky-Shapiro G. and Matheus. C. J., "The interestingness of deviations". In *Proceedings of the AAAI-94 Workshop on Knowledge Discovery in Databases*, pages 25-36, 1994.

[10] Silberschatz, A. and Tuzhilin, A., "On subjective measures of interestingness in knowledge discovery". In *Proceedings of the First International Conference on Knowledge Discovery and Data Mining*, pages 275-281, Montreal, Canada, August, 1995.

[11] A. Silberschatz and A. Tuzhilin. What makes patterns interesting in Knowledge discoverysystems. *IEEE Transactions on Knowledge and Data Engineering*, 8(6):970-974, December, 1996.

[12] Dong, G., Li, J., "Interestingness of Discovered Association Rules in terms of Neighborhood-Based Unexpectedness", *Proceedings of Pacific Asia Conference on Knowledge Discovery in Databases (PAKDD)*, pages 72-86. Melbourne, 1998.

[13] Bay, S. D. and Pazzani, M. J., "Detecting Group Differences: Mining Contrast Sets. *Data Mining and Knowledge Discovery*, 2001, Vol 5, No 3 213-246.

[14] Minaei-Bidgoli, B., Tan, P-N., Punch, W.F., "Mining Interesting Contrast Rules for a Web-based Educational System", *(IICMLA 2004)* in press.

Appendix A

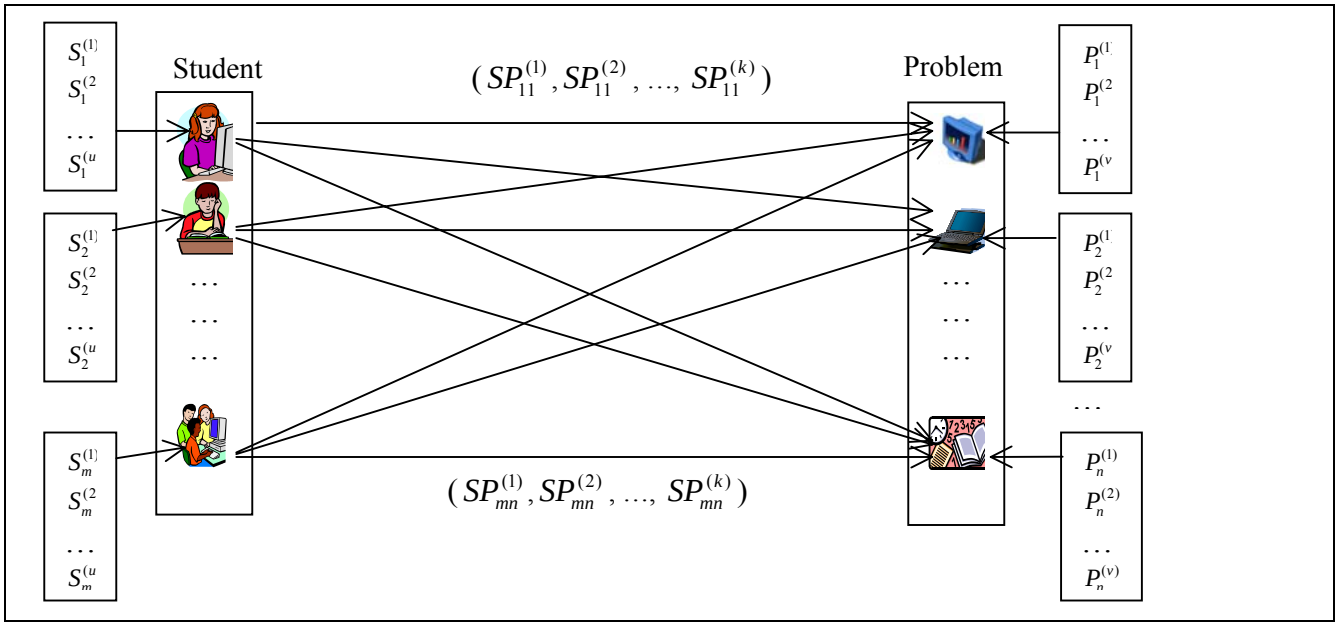

Figure 4. Attribute mining model, Fixed students' attributes, Problem attributes, and Linking attributes between students and problem