# Discovering Prerequisite Structure of Skills through Probabilistic Association Rules Mining

Yang Chen, Pierre-Henri Wuillemin, Jean-Marc Labat
Sorbonne Universites, UPMC, Univ. Paris 6, UMR 7606, LIP6, Paris, France
CNRS, UMR 7606, CNRS, Paris, France
4 Place Jussieu, 75005 Paris, France
{yang.chen, pierre-henri.wuillemin, jean-marc.labat}@lip6.fr

## ABSTRACT
Estimating the prerequisite structure of skills is a crucial issue in domain modeling. Students usually learn skills in sequence since the preliminary skills need to be learned prior to the complex skills. The prerequisite relations between skills underlie the design of learning sequence and adaptation strategies for tutoring systems. The prerequisite structures of skills are usually studied by human experts, but they are seldom tested empirically. Due to plenty of educational data available, in this paper, we intend to discover the prerequisite structure of skills from student performance data. However, it is a challenging task since skills are latent variables. Uncertainty exists in inferring student knowledge of skills from performance data. Probabilistic Association Rules Mining proposed by Sun et al. (2010) is a novel technique to discover association rules from uncertain data. In this paper, we preprocess student performance data by an evidence model. Then the probabilistic knowledge states of students estimated by the evidence model are used by the probabilistic association rules mining to discover the prerequisite structure of skills. We adapt our method to the testing data and the log data with different evidence models. One simulated data set and two real data sets are used to validate our method. The discovered prerequisite structures can be provided to assist human experts in domain modeling or to validate the prerequisite structures of skills from human expertise.

## Keywords
Probabilistic association rules mining, Skill structure, Prerequisite, DINA, BKT

## 1. INTRODUCTION
In most Intelligent Tutoring Systems (ITSs) and other educational environments, learning sequence is an important issue investigated by many educators and researchers. It is widely believed that students should be capable of solving the easier problems before the difficult ones are presented to them, and likewise, some preliminary skills should be learned prior to the learning of the complex skills. The prerequisite relations between problems and between skills underlie the adaptation strategies for tutoring and assessments. Furthermore, improving the accuracy of a student model with the prerequisite structure of skills has been exemplified by [1, 2]. The prerequisite structures of problems and skills are in accordance with the Knowledge Space Theory [3] and Competence-based Knowledge Space Theory [4]. A student's knowledge state should comply with the prerequisite structure of skills. If a skill is mastered by a student, all the prerequisites of the skill should also be mastered by the student. If any prerequisite of a skill is not mastered by a student, it seems difficult for the student to learn the skill. Therefore, according to the knowledge states of students, we can uncover the prerequisite structure of skills. Most prerequisite structures of skills reported in the student modeling literature are studied by domain or cognition experts. It is a tough and time-consuming task since it is quite likely that the prerequisite structures from different experts on the same set of skills are difficult to come to an agreement. Moreover, the prerequisite structures from domain experts are seldom tested empirically. Nowadays, some prevalent data mining and machine learning techniques have been applied in cognition models, benefiting from large educational data available through online educational systems. Deriving the prerequisite structures of observable variables (e.g. problems) from data has been investigated by some researchers. However, discovering prerequisite structures of skills is still challenging since a student's knowledge of a skill is a latent variable. Uncertainty exists in inferring student knowledge of skills from performance data. This paper aims to discover the prerequisite structures of skills from student performance data.

## 2. RELATED WORK
With the emerging educational data mining techniques, many works have investigated the discovery of the prerequisite structures within domain models from data. The Partial Order Knowledge Structures (POKS) learning algorithm is proposed by Desmarais and his colleagues [5] to learn the item to item knowledge structures (i.e. the prerequisite structure of problems) which are solely composed of the observable nodes, like answers to test questions. The results from the experiments over their three data sets show that the POKS algorithm outperforms the classic BN structure learning algorithms [6] on the predictive ability and the computational efficiency. Pavlik Jr. et al. [7] used the POKS algorithm to analyze the relationships between the observable item-type skills, and the results were used for the hierarchical agglomerative clustering to improve the skill model. Vuong et al. [8] proposed a method to determine the dependency relationships between units in a curriculum with the student performance data that are observed at the unit level (i.e. graduating from a unit or not). They used the statistic binominal test to look for a significant difference between the performance of students who used the potential prerequisite unit and the performance of students who did not. If a significant difference is found, the prerequisite relation is deemed to exist. All these methods above are proposed

to discover prerequisite structures of the observable variables. Tseng et al. [9] proposed to use the frequent association rules mining to discover concept maps. They constructed concept maps by mining frequent association rules on the data of the fuzzy grades from students' testing. They used a deterministic method to transfer frequent association rules on questions to the prerequisite relations between concepts, without considering the uncertainty in the process of transferring students' performance to their knowledge. Deriving the prerequisite structure of skills from noisy observations of student knowledge is considered in the approach of Brunskill [10]. In this approach, the log likelihood is computed for the precondition model and the flat model (skills are independent) on each skill pair to estimate which model better fits the observed student data. Scheines et al. [11] extended causal discovery algorithms to discover the prerequisite structure of skills by performing statistical tests on latent variables. In this paper, we propose to apply a data mining technique, namely the probabilistic association rules mining, to discover prerequisite structures of skills from student performance data.

# 3. METHOD

Association rules mining [12] is a well-known data mining technique for discovering the interesting association rules in a database. Let $I=\{i_1,i_2,\cdots,i_m\}$ be a set of attributes (called items) and $D=\{r_1,r_2,\cdots,r_n\}$ be a set of records (or transactions), i.e. a database. Each record contains the values for all the attributes in $I$. A pattern (called itemset) contains the values for some of the attributes in $I$. The support count of pattern $X$ is the number of records in $D$ that contain $X$, denoted by $\sigma(X)$. An association rule is an implication of the form $X\Rightarrow Y$, where $X$ and $Y$ are related to the disjoint sets of attributes. Two measures are commonly used to discover the strong or interesting association rules: the support of rule $X\Rightarrow Y$ denoted by $Sup(X\Rightarrow Y)$, which is the percentage of records in $D$ that contain $X\cup Y$, i.e. $P(X\cup Y)$; the confidence denoted by $Conf(X\Rightarrow Y)$, which is the percentage of records in $D$ containing $X$ that also contains $Y$, i.e. $P(Y|X)$. The rule $X\Rightarrow Y$ is considered strong or interesting if it satisfies the following condition:

$$(Sup(X \Rightarrow Y) \geq minsup) \wedge (Conf(X \Rightarrow Y) \geq minconf) \tag{1}$$

where *minsup* and *minconf* denote the minimum support threshold and the minimum confidence threshold. The support threshold is used to discover frequent patterns in a database, and the confidence threshold is used to discover the association rules within the frequent patterns. The support condition makes sure the coverage of the rule, that is, there are adequate records in the database to which the rule applies. The confidence condition guarantees the accuracy of applying the rule. The rules which do not satisfy the support threshold or the confidence threshold are discarded in consideration of the reliability. Consequently, the strong association rules could be selected by the two thresholds.

To discover the skill structure, a database of students' knowledge states is required. The knowledge state of a student is a record in the database and the mastery of a skill is a binary attribute with the values mastered (1) and non-mastered (0). If skill $Si$ is a prerequisite of skill $Sj$, it is most likely that $Si$ is mastered given that $Sj$ is mastered, and that skill $Sj$ is not mastered given that $Si$ is not mastered. Thus this prerequisite relation corresponds with the two association rules: $Sj=1\Rightarrow Si=1$ and $Si=0\Rightarrow Sj=0$. If both the association rules exist in a database, $Si$ is deemed a prerequisite of $Sj$. To examine if both the association rules exist in a database, according to condition (1), the following conditions could be used:

$$(Sup(Sj = 1 \Rightarrow Si = 1) \geq minsup) \wedge (Conf(Sj = 1 \Rightarrow Si = 1) \geq minconf) \tag{2}$$

$$(Sup(Si = 0 \Rightarrow Sj = 0) \geq minsup) \wedge (Conf(Si = 0 \Rightarrow Sj = 0) \geq minconf) \tag{3}$$

When condition (2) is satisfied, the association rule $Sj=1\Rightarrow Si=1$ is deemed to exist in the database, and when the condition (3) is satisfied, the association rule $Si=0\Rightarrow Sj=0$ is deemed to exist in the database. Theoretically, if skill $Si$ is a prerequisite of $Sj$, all the records in the database should comply with the two association rules. To be exact, the knowledge state $\{Si=0, Sj=1\}$ should be impossible, thereby $\sigma(Si=0, Sj=1)$ should be 0. According to the equations (4) and (5), the confidences of the rules in the equations should be 1.0. Since noise always exists in real situations, when the confidence of an association rule is greater than a threshold, the rule is considered to exist if the support condition is also satisfied. We cannot conclude that the prerequisite relation exists if one rule exists but the other not. For instance, the high confidence of the rule $Sj=1\Rightarrow Si=1$ might be caused by the high proportion $P(Si=1)$ in the data.

$$\begin{aligned} &Conf(Sj = 1 \Rightarrow Si = 1) = P(Si = 1 | Sj = 1) \\ &= \frac{\sigma(Si = 1, Sj = 1)}{\sigma(Si = 1, Sj = 1) + \sigma(Si = 0, Sj = 1)} \to 1 \end{aligned} \tag{4}$$

$$\begin{aligned} &Conf(Si = 0 \Rightarrow Sj = 0) = P(Sj = 0 | Si = 0) \\ &= \frac{\sigma(Si = 0, Sj = 0)}{\sigma(Si = 0, Sj = 0) + \sigma(Si = 0, Sj = 1)} \to 1 \end{aligned} \tag{5}$$

The discovery of the association rules within a database depends on the support and confidence thresholds. When the support threshold is given a relatively low value, more skill pairs will be considered as frequent patterns. When the confidence threshold is given a relatively low value, the weak association rules within frequent patterns will be deemed to exist. As a result, the weak prerequisite relations will be discovered. It is reasonable that the confidence threshold should be higher than 0.5. The selection of the two thresholds requires human expertise. Given the data about the knowledge states of a sample of students, the frequent association rules mining can be used to discover the prerequisite relations between skills.

However, a student's knowledge state cannot be directly obtained since student knowledge of a skill is a latent variable. In common scenarios, we collect the performance data of students in assessments or tutoring systems and estimate their knowledge states according to the observed data. The evidence models that transfer the performance data of students to their knowledge states in consideration of the noise have been investigated for several decades. The psychometric models DINA (Deterministic Input Noisy AND) and NIDA (Noisy Input Deterministic AND) [13] have been used to infer the knowledge states of students from their response data on the multi-skill test items. The well-known Bayesian Knowledge Tracing (BKT) model [14] is a Hidden Markov model that has been used to update students' knowledge states according to the log files of their learning in a tutoring system. A Q-matrix which represents the items to skills mapping is required in these models. The Q-matrix is usually created by domain experts, but recently some researchers [15, 16, 17] investigated to extract an optimal Q-matrix from data. Our method assumes that an accurate Q-matrix is known, like the method in [11]. Since the noise (e.g. slipping and guessing) is considered in the evidence models, the likelihood that a skill is mastered by a student can be estimated. The estimated knowledge state of a student is probabilistic, which incorporates the probability of each skill mastered by the student. Table 1 shows an example of the database consisting of probabilistic knowledge states. For example, the probabilities that skills *S1*, *S2* and *S3* are mastered by student "st1" are 0.9, 0.8 and 0.9 respectively.

We discover the prerequisite relations between skills from the probabilistic knowledge states of students that are estimated by an evidence model. The frequent association rules mining can no longer be used to discover the prerequisite relations between skills from a probabilistic database. Because any attribute value in a probabilistic database is associated with a probability. A probabilistic database can be interpreted as a set of deterministic instances (named possible worlds) [18], each of which is associated with a probability. We assume that the noise (e.g. slipping, guessing) causing the uncertainty for different skills is mutually independent. In addition, we assume that the knowledge states of different students are observed independently. Under these assumptions, the probability of a possible world in our database is the product of the probabilities of the attribute values over all the records in the possible world [18, 19, 20]. For example, a possible world for the database in Table 1 is that both the knowledge states of the students "st1" and "st2" are {*S1*=1, *S2*=0, *S3*=1}, whose probability is about 0.0233 (i.e. 0.9×0.2×0.9×0.2×0.9×0.8). The support count of a pattern in a probabilistic database should be computed with all the possible worlds. Thus the support count is no longer a deterministic number but a discrete random variable. Figure 1 depicts the probability mass function (*pmf*) of the support count of pattern {*S1*=1, *S2*=1} in the database of Table 1. For instance, the probability of $\sigma$(*S1*=1, *S2*=1)=1 is about 0.7112, which is the sum of the probabilities of all the possible worlds in which only one record contains the pattern {*S1*=1, *S2*=1}. Since there are an exponential number of possible worlds in a probabilistic database (e.g. $2^6$ possible worlds in the database of Table 1), computing the support count of a pattern is expensive. The Dynamic-Programming algorithm proposed by Sun et al. [20] is used to efficiently compute the support count *pmf* of a pattern.

**Table 1. A database of probabilistic knowledge states**

| Student ID | Probabilistic Knowledge State |
|---|---|
| st1 | {S1: 0.9, S2: 0.8, S3: 0.9} |
| st2 | {S1: 0.2, S2: 0.1, S3: 0.8} |

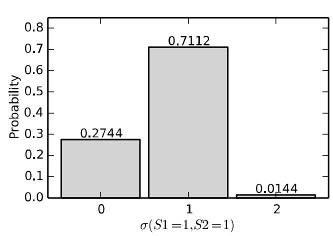


**Figure 1. The support count *pmf* of the pattern {*S1*=1, *S2*=1} in the database of Table 1**

To discover the prerequisite relations between skills from the probabilistic knowledge states of students, the probabilistic association rules mining technique [20] is used in this paper, which is an extension of the frequent association rules mining to discover association rules from uncertain data. Since the support count of a pattern in a probabilistic database is a random variable, the conditions (2) and (3) are satisfied with a probability. Hence the association rules derived from a probabilistic database are also probabilistic. We use the formula proposed by [20] to compute the probability of an association rule satisfying the two thresholds. It can be also interpreted as the probability of a rule existing in a probabilistic database. For instance, the probability of the association rule *Sj*=1⇒*Si*=1 existing in a probabilistic database is the probability that the condition (2) is satisfied in the database:

$$\begin{aligned} & P(Sj=1 \Rightarrow Si=1) \\ & = P\big((Sup(Sj=1 \Rightarrow Si=1) \geq minsup) \wedge (Conf(Sj=1 \Rightarrow Si=1) \geq minconf)\big) \\ & = \sum_{n=minsup \times N}^{N} f_{Si=1,Sj=1}[n] \sum_{m=0}^{\frac{(1-minconf)n}{minconf}} f_{Si=0,Sj=1}[m] \end{aligned} \tag{6}$$

where $N$ is the number of records in the database and $f_X$ denotes the support count *pmf* of pattern $X$, and $f_X[k]=P(\sigma(X)=k)$.

The probability of the rule related to condition (3) is computed similarly. According to formula (6), the probability of an association rule changes with the support and confidence thresholds. Given the two thresholds, the probability of an association rule existing in a probabilistic database can be computed. And if the probability is very close to 1.0, the association rule is considered to exist in the database. If both the association rules related to a prerequisite relation are considered to exist, the prerequisite relation is considered to exist. We can use another threshold, the minimum probability threshold denoted by *minprob*, to select the most possible association rules. Thus, if both P(*Sj*=1⇒*Si*=1)≥*minprob* and P(*Si*=0⇒*Sj*=0)≥*minprob* are satisfied, *Si* is deemed a prerequisite of *Sj*. When a pair of skills are estimated to be the prerequisite of each other, the relation between them are symmetric. It means that the two skills are mastered or not mastered simultaneously. The skill models might be improved by merging the two skills with the symmetric relation between them.

# 4. EVALUATION

We use one simulated data set and two real data sets to validate our method. The prerequisite structure derived from the simulated data is compared with the presupposed structure that is used to generate the data, while the prerequisite structure derived from the real data is compared with the structure investigated by another research on the same dataset or the structure from human expertise. Moreover, we adapt our method to the testing data and the log data. Different evidence models are used to preprocess the two types of data to get the probabilistic knowledge states of students. The DINA model is used for the testing data, whereas the BKT model is used for the log data.

## 4.1 Simulated Testing Data

**Data set.** We use the data simulation tool available via the R package CDM [21] to generate the dichotomous response data according to a cognitive diagnosis model (the DINA model used here). The prerequisite structure of the four skills is presupposed as Figure 3(a). According to this structure, the knowledge space decreases to be composed of six knowledge states, that is ∅, {*S1*}, {*S1*, *S2*}, {*S1*, *S3*}, {*S1*, *S2*, *S3*}, {*S1*, *S2*, *S3*, *S4*}. The reduced knowledge space implies the prerequisite structure of the skills. The knowledge states of 1200 students are randomly generated from the reduced knowledge space restricting every knowledge state type in the same proportion (i.e. 200 students per type). The simulated knowledge states are used as the input of the data simulation tool. There are 10 simulated testing questions, each of which requires one or two of the skills for the correct response. The slip and guess parameters for each question are restricted to be randomly selected in the range of 0.05 and 0.3. According to the DINA model with these specified parameters, the data simulation tool generates the response data. Using the simulated response data as the input of a flat DINA model, the slip and guess parameters of each question in the model are estimated and the probability of each student's knowledge on each skill is computed. The tool for the parameter estimation of DINA model is also available through the R package CDM [21], which is performed by the Expectation Maximization algorithm to maximize the marginal likelihood of data.

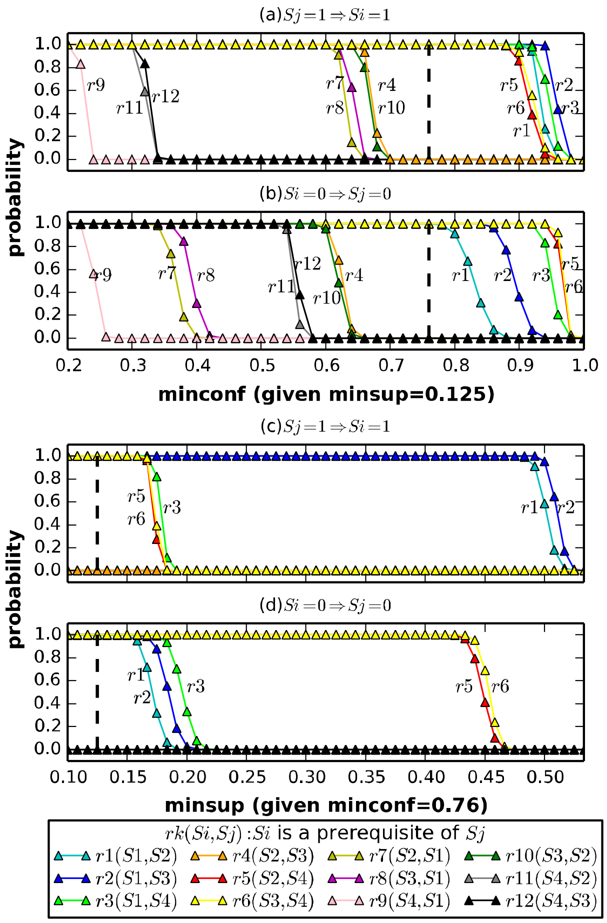


**Figure 2. The probabilities of the association rules in the simulated data given different confidence or support thresholds**

**Result.** The estimated probabilistic knowledge states of the simulated students are used as the input data to discover the prerequisite relations between skills. For each skill pair, there are two prerequisite relation candidates. For each prerequisite relation candidate, we examine if the two corresponding association rules *Sj*=1⇒*Si*=1and *Si*=0⇒*Sj*=0 exist in the database. The probability of an association rule existing in the database is computed according to formula (6), which is jointly affected by the selected support and confidence thresholds. For the sake of clarity, we look into the effect of one threshold leaving the other one unchanged. The joint effect of the two thresholds will be discussed in section 4.4. Giving a small constant to one threshold that all the association rules satisfy (perhaps several trials are needed or simply assign 0.0), we can observe how the probabilities of the association rules change with different values of the other threshold.

Figure 2 (a) and (b) describe how the probabilities of the corresponding association rules in the simulated data change with different confidence thresholds, where the support threshold is given as a constant (0.125 here). When the probability of a rule is close to 1.0, the rule is deemed to satisfy the thresholds. All the association rules satisfy the support threshold since their probabilities are almost 1.0 at first. The rules in the two figures corresponding to the same prerequisite relation candidate are depicted in the same color. In the figures, when the confidence threshold varies from 0.2 to 1.0, the probabilities of the different rules decrease from 1.0 to 0.0 in different intervals of threshold value. When we choose different threshold values, different sets of rules will be discovered. In each figure, there are five rules that can satisfy the significantly higher threshold. Given minconf=0.78, the probabilities of these rules are almost 1.0 whereas others are almost 0.0. These rules are very likely to exist. Moreover, the discovered rules in the two figures correspond to the same set of prerequisite relation candidates. Accordingly, these prerequisite relations are very likely to exist. To make sure the coverage of the association rules satisfying the high confidence threshold, it is necessary to know the support distributions of these rules. Figure 2 (c) and (d) illustrate how the probabilities of the corresponding association rules change with different support thresholds. The confidence threshold is given as a constant 0.76, and five association rules in each figure satisfy this threshold. Only on these rules, the effect of different support thresholds can be observed. In each figure, the rules gather in two intervals of threshold value. For example, in Figure 2 (c), to select the rules corresponding to *r3*, *r5* and *r6*, the highest value for the support threshold is roughly 0.17, while for the other two rules, it is 0.49. If both the confidence threshold and the support threshold are appropriately selected, the most possible association rules will be distinguished from others. As a result, the five prerequisite relations can be discovered in this experiment.

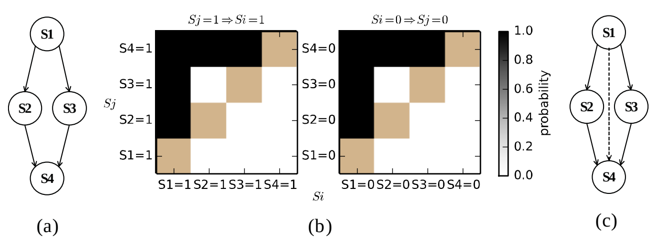


**Figure 3. (a) Presupposed prerequisite structure of the skills in the simulated data; (b) Probabilities of the association rules in the simulated data given *minconf*=0.76 and *minsup*=0.125, brown squares denoting impossible rules; (c) Discovered prerequisite structure**

Figure 3 (b) illustrates the probabilities of the corresponding association rules in the simulated data given *minconf*=0.76 and *minsup*=0.125. A square's color indicates the probability of the corresponding rule. Five association rules in each of the figures whose probabilities are almost 1.0 are deemed to exist. And the prerequisite relations corresponding to the discovered rules are deemed to exist. To qualitatively construct the prerequisite structure of skills, every discovered prerequisite relation is represented by an arc. It should be noted that the arc representing the relation that *S1* is a prerequisite of *S4* is not present in Figure 3 (a) due to the transitivity of prerequisite relation. Consequently, the prerequisite structure discovered by our method which is shown in Figure 3 (c), is completely in accordance with the presupposed structure shown in Figure 3 (a).

## 4.2 Real Testing Data

**Data set.** The ECPE (Examination for the Certification of Proficiency in English) data set is available through the R package CDM [21], which comes from a test developed and scored by the English Language Institute of the University of Michigan [22]. A sample of 2933 examinees is tested by 28 items on 3 skills, i.e. Morphosyntactic rules (*S1*), Cohesive rules (*S2*), and Lexical rules (*S3*). The parameter estimation tool in the R package CDM [21] for DINA model is also used in this experiment to estimate the slip and guess parameters of items according to the student response data. And with the estimated slip and guess parameters, the probabilistic knowledge states of students are assessed according to the DINA model, which are the input data for discovering the prerequisite structure of skills.

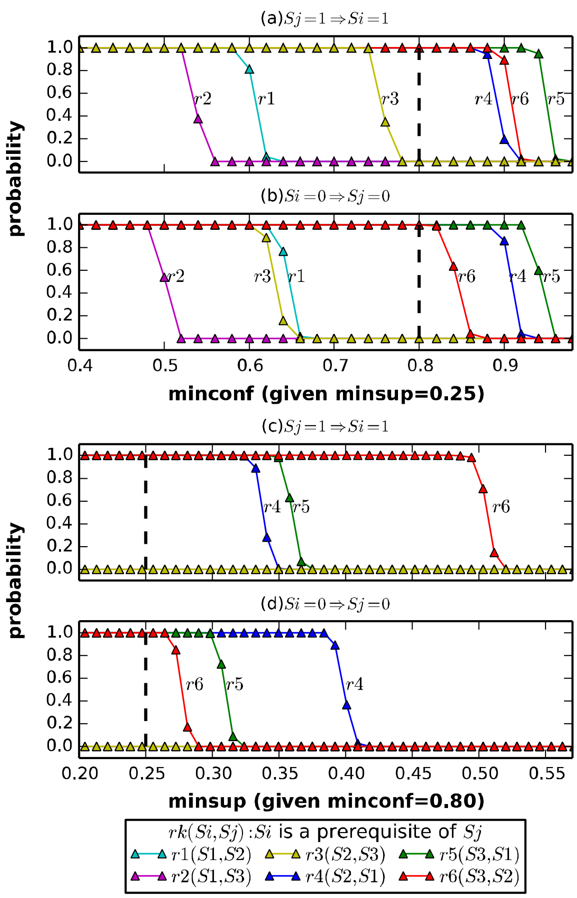


**Figure 4. The probabilities of the association rules in the ECPE data given different confidence or support thresholds**

**Result.** The effect of different confidence thresholds on the association rules in the ECPE data is depicted in Figure 4 (a) and (b) given the support threshold as a constant (0.25 here). In each figure, there are three association rules that can satisfy a significantly higher confidence threshold than others. The maximum value of the confidence threshold for them is roughly 0.82. And these rules in the two figures correspond to the same set of prerequisite relation candidates, that is, *r4*, *r5* and *r6*. Thus these candidates are most likely to exist. It can be noticed that in Figure 4 (a) the rule *S3*=1⇒*S2*=1 can satisfy a relatively high confidence threshold. The maximum threshold value that it can satisfy is roughly 0.74. However, its counterpart in Fig 4 (b), i.e. the rule *S2*=0⇒*S3*=0, cannot satisfy a confidence threshold higher than 0.6. When a strong prerequisite relation is required, the relation corresponding to the two rules cannot be selected. Only when both the two types of rules can satisfy a high confidence, the corresponding prerequisite relation is considered strong. Likewise, the effect of different support thresholds is shown in Figure 4 (c) and (d), where the confidence threshold is given as 0.80. And in each figure, only the three association rules which satisfy the confidence threshold are sensitive to different support thresholds. It can also be found that these rules are supported by a considerable proportion of the sample. Even when *minsup*=0.27, all the three rules in each figure satisfy it. According to the figures, when the support and confidence thresholds are appropriately selected, these rules can be distinguished from others. Consequently, the strong prerequisite relations can be discovered.

Given the confidence and support thresholds as 0.80 and 0.25 respectively, for instance, the probabilities of the corresponding association rules are illustrated in Figure 5 (b). The rules that satisfy the two thresholds (with a probability of almost 1.0) are deemed to exist, which are evidently distinguished from the rules that do not (with a probability of almost 0.0). Three prerequisite relations shown in Figure 5 (c) are found in terms of the discovered association rules. To validate the result, we compare it with the findings of another research on the same data set. The attribute hierarchy, namely the prerequisite structure of skills, in ECPE data has been investigated by Templin and Bradshaw [22] as Figure 5 (a). Our discovered prerequisite structure totally agrees with their findings.

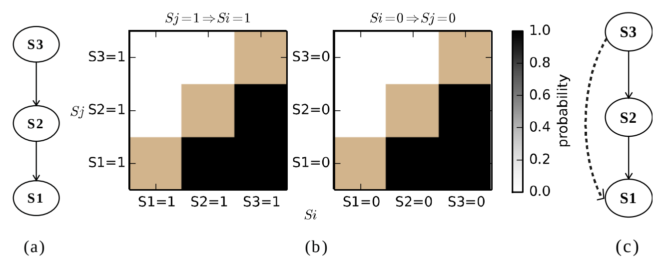


**Figure 5. (a) Prerequisite structure of the skills in the ECPE data discovered by Templin and Bradshaw [22]; (b) Probabilities of the association rules in the ECPE data given *minconf*=0.80 and *minsup*=0.25, brown squares denoting impossible rules; (c) Discovered prerequisite structure**

## 4.3 Real Log Data

**Data set.** We use the 2006-2007 school year data of the curriculum "Bridge to Algebra" [23] which incorporates the log files of 1146 students collected by Cognitive Tutor, an ITS for mathematics learning. The units in this curriculum involve distinct mathematical topics, while the sections in each unit involve distinct skills on the unit topic. A set of word problems is provided for each section skill. We use the sections in the units "equivalent fractions" and "fraction operations" as the skills (see Table 2). There are 560 students in the data set performing to learn one or several of the item-type skills in these units. The five skills discussed in our experiment are instructed in the given order in Table 2. A student's knowledge of the prior skills has the potential to affect his learning of the new skill. Hence, it makes sense to estimate whether a skill trained prior to the new skill is a prerequisite of it. If the prior skill *Si* is a prerequisite of skill *Sj*, students who have mastered skill *Sj* quite likely have previously mastered skill *Si*, and students not mastering the skill *Si* quite likely learn the skill *Sj* with great difficulty. Thus if both the rules *Sj*=1⇒*Si*=1 and *Si*=0⇒*Sj*=0 exist in the data, the prior skill *Si* is deemed a prerequisite of skill *Sj*.

**Table 2. Skills in the curriculum “Bridge to Algebra”**

| Skill | Example |
|---|---|
| S1: Writing equivalent fractions | Fill in the blank: $\frac{2}{3}=\frac{\square}{6}$. |
| S2: Simplifying fractions | Write the fraction in simplest form: $\frac{24}{30}=\frac{\ }{\ }$. |
| S3: Comparing and ordering fractions | Compare the fractions $\frac{3}{4}$ and $\frac{5}{6}$. |
| S4: Adding and subtracting fractions with like denominators | $\frac{2}{10}+\frac{3}{10}=$ |
| S5: Adding and subtracting fractions with unlike denominators | $\frac{2}{3}-\frac{1}{4}=$ |

To discover the prerequisite relations between skills, firstly we need to estimate the outcomes of student learning according to the log data. A student learns a skill by solving a set of problems that requires applying that skill. At each opportunity, student knowledge of a skill probably transitions from the unlearned to learned state. Thus their knowledge should be updated each time they go through a problem. The BKT model has been widely used to track the dynamic knowledge states of students according to their activities on ITSs. In the standard BKT, four parameters are specified for each skill [14]: $P(L_0)$ denoting the initial probability of knowing the skill a priori, $P(T)$ denoting the probability of student’s knowledge of the skill transitioning from the unlearned to the learned state, $P(S)$ and $P(G)$ denoting the probabilities of slipping and guessing when applying the skill. We implemented the BKT model by using the Bayes Net Toolbox for Student modeling [24]. The parameter $P(L_0)$ is initialized to 0.5 while the other three parameters are initialized to 0.1. The four parameters are estimated according to the log data of students, and the probability of a skill to be mastered by a student is estimated each time the student performs to solve a problem on that skill. In the log data, students learned the section skills one by one and no student relearned a prior section skill. If a prior skill *Si* is a prerequisite of skill *Sj*, the knowledge state of *Si* after the last opportunity of learning it has an impact on learning *Sj*. We use the probabilities about students’ final knowledge state of *Si* and *Sj* to analyze whether a prerequisite relation exists between them. Thus students’ final knowledge states on each skill are used as the input data of our method.

**Result.** The probabilities of the association rules in the log data changing with different confidence thresholds are illustrated in Figure 6 (a) and (b) given the support threshold as a small constant (0.05 here). In Figure 6 (a), compared with the rules *S4*=1⇒*S3*=1 and *S5*=1⇒*S3*=1, all the other association rules can satisfy a significantly higher confidence, while in Figure 6 (b) if given *minconf*=0.6, only three rules satisfy it. The effect of different support thresholds on the probabilities of the association rules is depicted in Figure 6 (c) and (d) given the confidence threshold as a constant (0.3 here). All the association rules satisfy the confidence threshold as the probabilities of the rules are almost 1.0 at first. In Figure 6 (c), there are six rules that can satisfy a relatively higher support threshold (e.g. *minsup*=0.2). But in Figure 6 (d), even given *minsup*=0.14, only the rule *S4*=0⇒*S5*=0 satisfy it, and the maximum value for the support threshold that all the rules can satisfy is roughly 0.07.

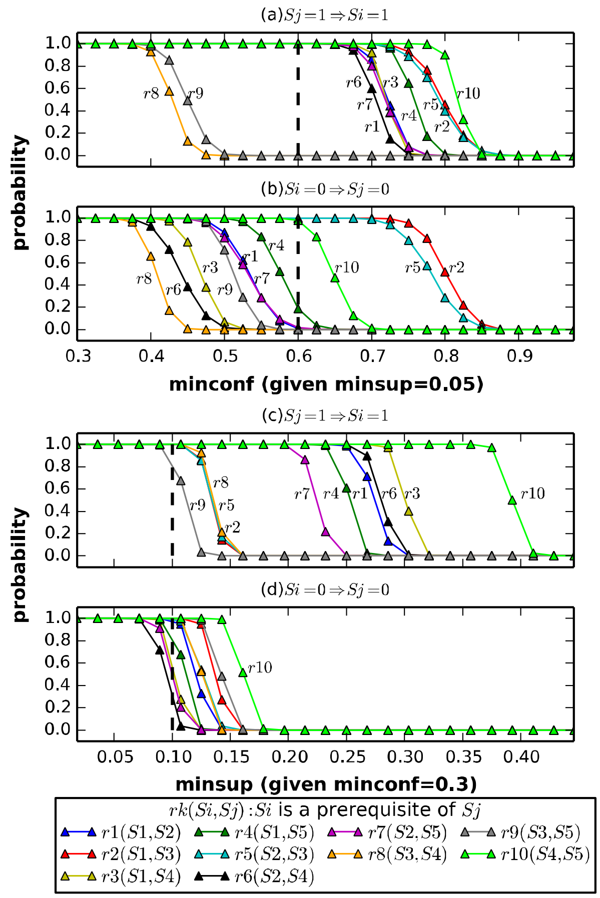


**Figure 6. The Probabilities of the association rules in the “Bridge to Algebra 2006-2007” data given different confidence or support thresholds**

Given the confidence and support thresholds as 0.6 and 0.1 respectively, the probabilities of the association rules in the log data are depicted in Figure 7 (b). There are eight of the rules in the form of *Sj*=1⇒*Si*=1 (left) and three of the rules in the form of *Si*=0⇒*Sj*=0 (right) discovered, whose probabilities to satisfy the thresholds are almost 1.0. According to the result, only the three prerequisite relations shown in Figure 7 (c), whose corresponding rules both are discovered, are deemed to exist. Figure 7 (a) shows the prerequisite structure of the five skills from the human experts’ opinions. It makes sense that the skills *S1* and *S2* rather than skill *S3* are required for learning the skills *S4* and *S5*. This is supported by the chapter warm-up content in the student textbook of the course [25]. The discovered rules in the form of *Sj*=1⇒*Si*=1 completely agree with the structure from human expertise. But the discovered rules in the form of *Si*=0⇒*Sj*=0 is inconsistent with it. The counterparts of a large part of the discovered rules *Sj*=1⇒*Si*=1 do not satisfy the confidence threshold. Even reducing the confidence threshold to the lowest value, i.e. 0.5, the rules *S1*=0⇒*S4*=0 and *S2*=0⇒*S4*=0 still do not satisfy it (see Figure 6 (b)). It seems that the rules *Sj*=1⇒*Si*=1 are more reliable than

*Si*=0⇒*Sj*=0 since most of the former can satisfy a higher support threshold than the latter (see Figure 6 (c) and (d)). In addition, the log data is very likely to contain much noise. It is possible that some skills could be learned if students take sufficient training, even though some prerequisites are not previously mastered. In this case, the support count $\sigma(Si{=}0, Sj{=}1)$ would increase. Or perhaps students learned the prerequisite skills by solving the scaffolding questions in the process of learning new skills, even though they performed not mastering the prerequisite skills before. In this case, the observed values of $\sigma(Si{=}0, Sj{=}1)$ would be higher than the real values. According to the equations (4) and (5), if $\sigma(Si{=}0, Sj{=}1)$ increases, the confidence of the rules will decrease. And when the noise appears in the data, the confidences of the association rules which are supported by a small proportion of sample will be affected much more than those supported by a large proportion of sample.

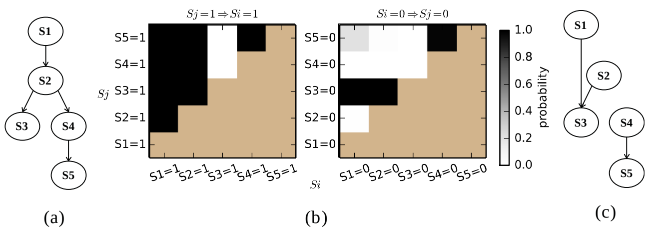


**Figure 7. (a) Prerequisite structure from human expertise; (b) Probabilities of the association rules in the "Bridge to Algebra 2006-2007" data given *minconf*=0.6 and *minsup*=0.1, brown squares denoting impossible rules; (c) Discovered prerequisite structure**

## 4.4 Joint Effect of thresholds

We have discussed the effect of one threshold on the probability of association rules while eliminating the effect of the other one in the three experiments. To determine the values for the thresholds, we investigate how the two thresholds simultaneously affect the probability of an association rule. Figure 8 depicts how the probabilities of the association rules for the skill pair *S2* and *S3* in the ECPE data change with different support and confidence thresholds, where (a) and (c) involve one relation candidate while (b) and (d) involve the other one. The figures demonstrate that the probability of a rule decreases almost from 1.0 to 0.0 when the confidence and support thresholds vary from low to high. It can be found that the rules in the left figures can satisfy an evidently higher confidence threshold than those in the right figures, and have the same support distributions with them. If we set *minconf*=0.8 and *minsup*=0.25, only the rules in the left figures satisfy them. Suppose that a rule satisfy the thresholds if its probability is higher than 0.95, i.e. *minprob*=0.95. When we change the values of the confidence and support thresholds from 0.0 to 1.0, for each rule, we can find a point whose coordinates consist of the maximum values of the confidence and support thresholds that the rule can satisfy. Finding the optimal point is hard and there are probably several feasible points. To simplify the computation, the thresholds are given by a sequence of discrete values from 0.0 to 1.0. We find the maximum value for each threshold when only one threshold affects the probability of the rule given the other as 0.0. And for each threshold, *minprob* is given as 0.97, roughly the square root of the original value. The found maximum values for the two thresholds are the coordinates of the point. The found point is actually an approximately optimal point. For convenience, the point is named maximum threshold point in this paper. The points for all the rules in the three data sets are found by our method as well as plotted in Figure 9 (some points overlap). When we set certain values to the thresholds, the points located in the upper right area satisfy them and the related rules are deemed to exist. For one prerequisite relation, a couple of related points should be verified. Only when both of them are located in the upper right area, they are considered eligible to uncover the prerequisite relation. The eligible points in Figure 8 and Figure 9 are indicated given the thresholds.

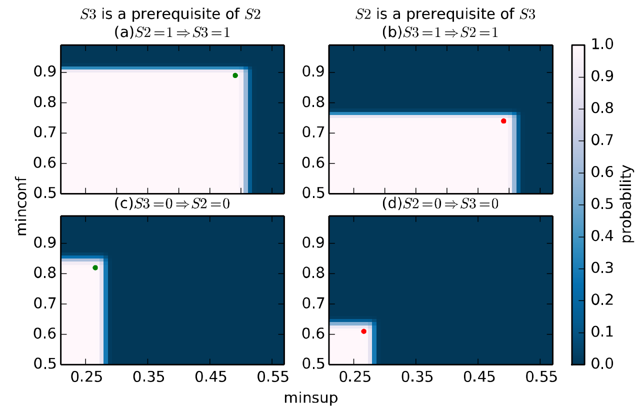


**Figure 8. Probabilities of the association rules within the skill pair *S2* and *S3* in the ECPE data given different confidence and support thresholds, and their maximum threshold points which are eligible (green) or not (red) given *minconf*=0.8 and *minsup*=0.25**

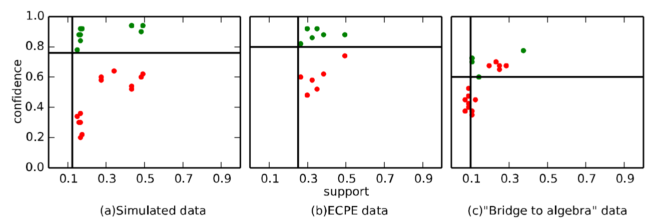


**Figure 9. Maximum threshold points for the association rules in our three experiments, where eligible points are indicated in green given the thresholds**

# 5. CONCLUSION AND DISCUSSION

Discovering the prerequisite structure of skills from data is challenging in domain modeling since skills are the latent variables. In this paper, we propose to apply the probabilistic association rules mining technique to discover the prerequisite structure of skills from student performance data. Student performance data is preprocessed by an evidence model. And then the probabilistic knowledge states of students estimated by the evidence model are used as the input data of probabilistic association rules mining. Prerequisite relations between skills are discovered by estimating the corresponding association rules in the probabilistic database. The confidence condition of an association rule in our method is similar to the statistical hypotheses used in the POKS algorithm for determining the prerequisite relations between observable variables (see the details in [5]). But our method targets on the challenge of discovering the prerequisite relations between latent variables from the noisy observable data. In addition, our method takes the coverage into account (i.e. the support condition), which could strengthen the reliability of the discovered prerequisite relations. Determining the appropriate confidence and support thresholds is a crucial issue in our method. The effect of a single threshold and the joint effect of two thresholds on the probabilities of the rules are discussed. The maximum threshold points of the probabilistic association rules are proposed for determining the thresholds. We adapt our method to two common types of data, the testing data and the log data, which are preprocessed by different evidence models, the DINA model and the BKT model. An accurate Q-matrix is required for the evidence models, which is a limitation of our method. According to the results of the experiments in this paper, our method performs well to discover the prerequisite structures from a simulated testing data set and a real testing data set. However, applying our method in the log data still needs to be improved. Since much noise exist in the log data, the strategies to reduce the noise need to be applied. The prerequisite structures of skills discovered by our method can be applied to assist human experts in skill modeling or to validate the prerequisite structures of skills from human expertise.

## 6. REFERENCES


[1] Käser, T., Klinger, S., Schwing, G., Gross, M.: Beyond Knowledge Tracing: Modeling Skill Topologies with Bayesian Networks. In *Proceedings of the 12th International Conference on Intelligent Tutoring Systems*, Honolulu, USA, 188-198, 2014

[2] Chen, Y., Wuillemin, P.H., Labat, J.M.: Bayesian Student Modeling Improved by Diagnostic Items. In *Proceedings of the 12th International Conference on Intelligent Tutoring Systems*, Honolulu, USA, 144-149, 2014

[3] Falmagne, J.C., Cosyn, E., Doignon, J.P., Thiéry, N.: The Assessment of Knowledge, in Theory and in Practice. In *Proceedings of the 4th International Conference on Formal Concept Analysis*, Dresden, Germany, 61-79, 2006

[4] Heller, J., Steiner, C., Hockemeyer, C., Albert, D.: Competence-based Knowledge Structures for Personalized Learning. *International Journal on E-learning*, 5, 75-88, 2006

[5] Desmarais, M.C., Meshkinfam, P., Gagnon, M.: Learned Student Models with Item to Item Knowledge Structures. *User Modeling and User-adapted Interaction*, 16(5), 403-434, 2006

[6] Spirtes, P., Glymour, C., Scheines, R.: *Causation, Prediction, and Search*. 2nd Edition, The MIT Press, Cambridge, 2000

[7] Pavlik Jr., P.I., Cen, H., Wu, L., Koedinger, K.R.: Using Item-type Performance Covariance to Improve the Skill Model of an Existing Tutor. In *Proceedings of the 1st International Conference on Educational Data Mining*, Montreal, Canada, 77-86, 2008

[8] Vuong, A., Nixon, T., Towle, B.: A Method for Finding Prerequisites within a Curriculum. In *Proceedings of the 4th International Conference on Educational Data Mining*, Eindhoven, Netherlands, 211-216, 2011

[9] Tseng, S.S., Sue, P.C., Su, J.M., Weng, J.F., Tsai, W.N.: A New Approach for Constructing the Concept Map. *Computers & Education*, 49(3), 691-707, 2007

[10] Brunskill, E.: Estimating Prerequisite Structure from Noisy Data. In *Proceedings of the 4th International Conference on Educational Data Mining*, Eindhoven, Netherlands, 217-222, 2011

[11] Scheines, R., Silver, E., Goldin, I.: Discovering Prerequisite Relationships among Knowledge Components. In *Proceedings of the 7th International Conference on Educational Data Mining*, London, UK, 355-356, 2014

[12] Agrawal, R., Srikant, R.: Fast Algorithm for Mining Association Rules. In *Proceedings of the 20th International Conference on Very Large Data Bases*, San Francisco, USA, 487-499, 1994

[13] Roussos, L.A., Templin, J.L., Henson, R.A.: Skills Diagnosis Using IRT-based Latent Class Models. *Journal of Educational Measurement*, 44(4), 293-311, 2007

[14] Corbett, A.T., Anderson, J.R.: Knowledge Tracing: Modeling the Acquisition of Procedural Knowledge. *User Modeling and User-Adapted Interaction*, 4(4), 253-278, 1995

[15] Barnes, T.: The Q-matrix Method: Mining Student Response Data for Knowledge. *American Association for Artificial Intelligence 2005 Educational Data Mining Workshop*, 2005

[16] Desmarais, M.C., Naceur, R.: A Matrix Factorization Method for Mapping Items to Skills and for Enhancing Expert-Based Q-Matrices. In *Proceedings of the 16th International Conference on Artificial Intelligence in Education*, Memphis, USA, 441-450, 2013

[17] González-Brenes, J.P.: Modeling Skill Acquisition Over Time with Sequence and Topic Modeling. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, San Diego, USA, 296-305, 2015

[18] Bernecker, T., Kriegel, H.P., Renz, M., Verhein, F., Zuefle, A.: Probabilistic Frequent Itemset Mining in Uncertain Databases. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Paris, France, 119-128, 2009

[19] Chui, C.K., Kao, B., Hung, E.: Mining Frequent Itemsets from Uncertain Data. In *Proceedings of the 11th PAKDD Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, Nanjing, China, 47-58, 2007

[20] Sun, L., Cheng, R., Cheung, D.W., Cheng, J.: Mining Uncertain Data with Probabilistic Guarantees. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, USA, 273-282, 2010

[21] Robitzsch, A., Kiefer, T., George, A.C., Uenlue, A.: Package CDM (Version 3.4-21, 2014), http://cran.r-project.org/web/packages/CDM/index.html

[22] Templin, J., Bradshaw, L.: Hierarchical Diagnostic Classification Models: A Family of Models for Estimating and Testing Attribute Hierarchies. *Psychometrika*, 79, 317-339, 2014

[23] Stamper, J., Niculescu-Mizil, A., Ritter, S., Gordon, G.J., & Koedinger, K.R.: Bridge to Algebra 2006-2007. Development data set from KDD Cup 2010 Educational Data Mining Challenge. Find it at http://pslcdatashop.web.cmu.edu/KDDCup/downloads.jsp

[24] Chang, K., Beck, J., Mostow, J., & Corbett, A.: A Bayes Net Toolkit for Student Modeling in Intelligent Tutoring Systems. In *Proceedings of the 8th International Conference on Intelligent Tutoring Systems*, Jhongli, Taiwan, 104-113, 2006

[25] Hadley, W.S., Raith, M.L.: *Bridge to Algebra Student Text*. Carnegie Learning. 2008