*Liet. matem. rink.*, **44**, *spec. nr.*, 2004, 339–344

# Mining the applications to study at Lithuania's institutions of higher education using association rules

Ieva MITAŠIŪNAITĖ (VU), Gediminas ADOMAVIČIUS (University of Minnesota)
e-mail: ieva.mitasiunaite@maf.vu.lt, gedas@umn.edu.lt

**Abstract.** The secondary outcome of implementing the computerized application process for studies at Lithuania's institutions of higher education is the availability of the collected data. An important problem is how to use this data effectively, e.g., how to extract some useful knowledge from it. Discovered patterns and relationships can facilitate a better understanding of the applicants' preferences, popularity of various study programs, and, more generally, the overall application process. This paper uses the association rule mining technique and focuses on the relationships between the gender of the applicants and their preferred study programs. The data mining procedure, which can also be considered as a framework for similar data mining tasks, is presented through several illustrative examples.

*Keywords:* data mining, association rules.

## 1. Introduction

The objective of this research project is to extract useful, interesting and comprehensible knowledge from the applications submitted in 2003 to study at Lithuania's institutions of higher education that were collected by the Computer Centre of Vilnius University. More specifically, the collected data includes the following information about each applicant: some demographic information (e.g., date of birth, gender, country of citizenship), the institutions he/she has graduated from along with the corresponding grading information, his/her application form indicating the preferred study programs, and the admission results. Each applicant is authorized to specify a preference list of no more than twenty different study programs, stated in the order of descending priority. A study program is defined by the name of the subject (e.g., psychology), study form (e.g., full time studies), and the respective institution of higher education. The following institutions participated in this centralized application process: General Jonas Žemaitis Military Academy of Lithuania (GJŽMAL), International School of Management (ISM), Kaunas University of Medicine (KUM), Kaunas University of Technology (KUT), Klaipėda University (KU), Law University of Lithuania (LUL), Lithuanian Academy of Music (LAM), Lithuanian Academy of Physical Education (LAPE), Lithuanian University of Agriculture (LUA), Lithuanian Veterinary Academy (LVA), Vilnius Academy of Fine Arts (VAFA), Vilnius Gediminas Technical University (VGTU), Vilnius Pedagogical University (VPU), Vilnius University (VU), Vytautas Didysis University (VDU), Šiauliai University (ŠU).

The association rule discovery procedure employed in this project is illustrated using one specific data mining task from this project, i.e., the analysis of the relationships

between the gender of an applicant and his/her preferred study programs. As a baseline gender distribution, it is useful to know that in 2003 women made up 57% and men made up 43% of the total of 32269 applicants.

## 2. Data mining and association rules

The progress in the use of computer technologies in numerous areas of human activity has resulted in increasing volume of the collected data. The emerging challenge is to use that data efficiently. Collected data needs to be converted into information and knowledge to become useful [1]. Data mining discipline can be defined as follows: "data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner" [2]. The term "observational data" is used as an opposite to "experimental data" in order to emphasize the fact that data mining deals with the data that was collected without an explicit intention to use it for some research analysis, while statistics often deals with "experimental data" that is collected to answer specific questions [2]. The project described in this paper considers the mining of interesting relationships from the data that was collected (although not explicitly for this reason) by the Computer Centre of Vilnius University.

We consider the problem of finding interesting relationships in this data, namely, in the preference lists of study programs specified by each applicant. The preference list specified by an applicant can be considered as an itemset (i.e., a set of items), where an item, or element, is a study program. Such an itemset, associated with a given applicant, can be easily extended with additional elements that are specific to the applicant, such as his/her gender. Then the problem of finding relationships among the preferred study programs and the gender can still be considered as the same problem of finding relationships among the elements of itemsets. Consequently, since we can regard data as a collection of itemsets, the association rule discovery technique [3] is highly appropriate; moreover, association rules are easy to understand and constitute a natural and intuitive way to express relationships and patterns.

More specifically, association rule mining algorithms find elements that commonly occur together in an itemset. Formally, the association rule model can be stated as follows [3]: let $I = \{i_1, i_2, \ldots, i_m\}$ be a set of literals, called *items*. Let $D$ be a set of transactions (i.e., a database), where each transaction $T$ is a set of items (itemset) such that $T \subseteq I$. An association rule is an implication of the form $X \rightarrow Y$, where $X \subseteq I$, $Y \subseteq I$, and $X \cap Y = \varnothing$. Given itemset $X$, its frequency $fr(X)$ is the number of transactions in $D$ that contain $X$. The rule $X \rightarrow Y$ holds in the transaction set $D$ with *confidence* $c$, if $c$ percent of the transactions in $D$, containing $X$, also contain $Y$, i.e., $c = fr(X \cup Y) \times 100/fr(X)$. The rule $X \rightarrow Y$ has *support* $s$, if $s$ percent of the transactions in $D$ contain $X \cup Y$, i.e., $s = fr(X \cup Y) \times 100/|D|$, where $|D|$ is a total number of transactions in $D$. Finally, given a set of transactions $D$, the problem of association rule mining is to discover all association rules that have support and confidence greater or equal to the user-specified support threshold (called *minSupp*) and confidence threshold (called *minConf*) respectively. Many algorithms for fast discovery of association rules have been proposed in data mining literature [3, 4, 5], with the Apriori algorithm [3] being the most popular and widely used.

## 3. Mining the relationships between the gender of applicants and their preferred study programs

Before proceeding with the association rule generation, a number of necessary initial stages that are typical to any data mining process were performed, namely, the preliminary data analysis and data preparation, including data cleaning, reorganization, and aggregation. To generate the association rules, an open source software package [6] implementing the Apriori algorithm [3] was used.

We consider the itemsets of the form: $\{G, P_1, \ldots, P_k\}$, where $G$ stands for the applicant's gender, and $P_i$, where $i = 1, \ldots, k$, represent the study programs specified by that applicant. $G$ can have the values F and M, where F denotes a female applicant and M denotes a male applicant. The specific values of $P_i$ (where $i = 1, \ldots, k$) and the value of $k$ are experiment-specific and will be explicitly defined in each subsequent subsection. One itemset $\{G, P_1, \ldots, P_k\}$ corresponds to one applicant. The problem is to find relationships among the elements $G$ and $P_i$.

### *3.1. Mining the relationships where program priority numbers are included*

We consider itemsets $\{G, P_1, \ldots, P_k\}$, where $\max(k) = 20$ and each $P_i$ takes the form of $[pr, p, sf, u]$, where $pr$ denotes the priority number of a study program in the applicant's preference list, $p$ denotes the name of the study program, $sf$ denotes the form of studies, and $u$ denotes the institution of higher education. The priority number $pr$ can take the values from 1 to 20, where 1 represents the highest priority, and 20 – the lowest priority. Study form $sf$ takes the values FTM, EVE and EXT: FTM denotes full time studies, EVE – evening studies, and EXT – extramural studies. In this example, we consider rules of the form $P_i \rightarrow G$. The following rule is one example of the relationships that were mined having set $minConf = 85\%$ and $minSupp = 0.5\%$:

$$[1, \text{Computer Science, FTM, VU}] \rightarrow \text{M } (1.5\%/482, 91.1\%) \quad (1)$$

The information represented by association rule (1) can be interpreted as follows: 91.1% of applicants who have specified the full time Computer Science studies at VU as their number one priority were men. Altogether there were 482 such men, i.e., who have specified this study program with the highest priority, or 1.5% of all applicants. Speaking in association rule terms, the support of the above rule is 1.5% and the confidence of this rule is 91.1%. All other association rules presented in this paper should be interpreted in the same manner. The following are some additional examples of association rules discovered for this data mining task:

$$[1, \text{Software Engineering, FTM, VU}] \rightarrow \text{M } (0.7\%/238, 94.1\%) \quad (2)$$

$$[1, \text{Computer Science, FTM, KUT}] \rightarrow \text{M } (1.1\%/354, 89.4\%) \quad (3)$$

$$[1, \text{Psychology, FTM, VU}] \rightarrow \text{F } (1.2\%/392, 88.5\%) \quad (4)$$

$$[1, \text{Pharmaceutics, FTM, KUM}] \rightarrow \text{F } (0.7\%/214, 86.3\%) \quad (5)$$

The emerging interpretation would be that there were almost only men who have specified the Computer Science and Software Engineering studies while there were almost

only women who have specified the Psychology and Pharmaceutics studies at the corresponding universities with the highest priority.

### 3.2. *Mining the relationships where priority numbers are discarded*

Here we consider itemsets $\{G, P_1, \ldots, P_k\}$, where $\max(k) = 20$ and each $P_i$ takes the form of $[p, sf, u]$. Here $p, sf$ and $u$ have the same definitions as in Section 3.1. The priority number is discarded in order to have a more general understanding about the presence or the absence of the study program in the preference list, regardless of the priority with which this program is specified. Again, we consider the rules of the form $P_i \rightarrow G$. The following are some examples of rules that were discovered for this data mining task, having set $minConf = 80\%$ and $minSupp = 2.5\%$:

[Computer Science, FTM, VU] → M (6.6%/2116, 80.2%) (6)

[Software Engineering, FTM, VU] → M (4.5%/1458, 83.3%) (7)

[Computer Science, FTM, KUT] → M (5.1%/1642, 83.1%) (8)

[Psychology, FTM, VU] → F (10.7%/3466, 82.1%) (9)

[Pharmaceutics, FTM, KUM] → F (2.5%/793, 81.3%) (10)

Note the parallels between rules (6–10) and the rules (1–5) presented in Section 3.1. The support of rules (6–10) is obviously higher than the support of their counterparts from the previous section, since in this data mining task the study programs within a rule are no longer restricted to one priority number. Interestingly, the confidence of rules (6–10) has decreased by 5–10%, compared to rules (1–5). For example, the percentage of men among applicants who specified the full time Computer Science studies at VU as their number one priority was 91.1% (according to rule 1). On the other hand, the percentage of men among applicants who specified the same study program anywhere on the list (i.e., regardless of priority) was 80.2% (rule 6).

### 3.3. *Mining the relationships based on high-priority study programs*

For this task, we consider itemsets $\{G, P_1, \ldots, P_k\}$, where $\max(k) = q$ and each $P_i$ takes the form of $[p, sf, u]$. Here $p, sf$ and $u$ have the same definitions as in Section 3.1. It is reasonable to assume that the study programs that appear higher on the applicant's preference list represent the applicant's true interests better than the programs that appear lower on this list. Therefore, it seems useful to mine relationships that involve only high-priority study programs and, for this reason, in this task we discarded the low-priority study programs. More specifically, we ran this data mining task twice, setting the value of $q$ to 3 and 5, i.e., only first $q$ study programs from each preference list were used to form the itemsets. Again, we searched for the rules of the form $P_i \rightarrow G$, where $minConf = 80\%$ and $minSupp = 1\%$. The following are some examples ($q$ value is specified for each rule):

[Software Engineering, FTM, VU] → M (2.9%/928, 88.3%); $q = 5$ (11)

[Software Engineering, FTM, VU] → M (2.1%/666, 90.9%); $q = 3$ (12)

$$[\text{Pharmaceutics, FTM, KUM}] \rightarrow \text{F}\ (1.7\%/550, 82.6\%);\ q = 5 \tag{13}$$

$$[\text{Pharmaceutics, FTM, KUM}] \rightarrow \text{F}\ (1.3\%/421, 83.9\%);\ q = 3 \tag{14}$$

One interesting observation about Software Engineering and related rules (7), (11), (12), and (2) is that the confidence is consistently increasing when this study program is specified with higher priority. The same observation can also be made about Pharmaceutics and rules (10), (13), (14), and (5).

### *3.4. Mining the relationships among more than one study program*

In previous sections we presented association rule examples that had only one study program on their left-hand sides (i.e., in their antecedents). However, the left-hand side of the rule can consist of more than one element. In this task, we search for the rules of the form $P_i$, $P_j \rightarrow G$. Again, we consider itemsets $\{G, P_1, \ldots, P_k\}$, where each $P_i$ and $P_j$ is of the previously defined form $[p, sf, u]$. Here are some examples of the discovered rules ($minConf = 85\%$, $minSupp = 0.5\%$):

$$[\text{Computer Physics, FTM, VU}], [\text{Computer Science, FTM, VU}] \rightarrow \text{M}\ (1.7\%/540, 92.3\%) \tag{15}$$

$$[\text{Software Engineering, FTM, VU}], [\text{Computer Science, FTM, VU}] \rightarrow \text{M}\ (3.8\%/1234, 84.8\%) \tag{16}$$

The comparison of rules (6) and (15) leads to the following observation: the appearance of the Computer Physics studies in addition to the Computer Science studies at VU on the applicant's preference list raises the confidence of the inference that the applicant is male by 12.1%. A similar observation can be made about the rules (6) and (16) as well: the combination of the Computer Science studies with the Software Engineering studies at VU raise the confidence of the inference that the applicant is male by 4.6%. Based on the support parameter values of rules (15) and (16), it is also interesting to note that the combination of the Computer Science and Software Engineering studies at VU was 2.3 times more frequent than the combination of the Computer Science and Computer Physics studies at VU on the preference lists specified by male applicants.

## 4. Conclusions

This paper demonstrates that the data mining approach can be successfully applied to discover patterns and relationships existing in raw data sets such as the data collected to perform the computerized and centralized application process for the institutions of higher education. We have shown that the problem of mining the relationships among different attributes can be formulated in terms of association rules and illustrated the corresponding data mining procedure using a specific knowledge discovery task, i.e., the analysis of the relationships between the gender of an applicant and his/her preferred study programs. Moreover, the data mining procedure described in this paper

can be easily generalized to mine other possible relationships and patterns (e.g., between the applicant's geographical location and the preferred institution of higher education) by switching between various aggregation levels and/or extending itemsets with additional items.

## References

1. M. Kantardzic, *Data Mining: Concepts, Models, Methods, and Algorithms*, IEEE Press (2003).
2. D. Hand, H. Mannila, P. Smyth, *Principles of Data Mining*, The MIT Press (2001).
3. R. Agrawal, R. Srikant, Fast algorithms for mining association rules, in: *Proceedings of the 20th International Conference on Very Large Data Bases*, Santiago, Chile (1994).
4. M. Zaki, S. Parthasarathy, M. Ogihara, W. Li, New algorithms for fast discovery of association rules, in: *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining*, Newport, California, USA (1997).
5. J. Han, J. Pei, Y. Yin, Mining frequent patterns without candidate generation, in: *Proceedings of the ACM SIGMOD International Conference on Management of Data* (2000).
6. C. Borgelt, *Finding Association Rules/Hyperedges with the Apriori Algorithm.* `http://fuzzy.cs.uni-magdeburg.de/borgelt/doc/apriori/`

REZIUMĖ

***I. Mitašiūnaitė, G. Adomavičius. Stojimo į Lietuvos aukštąsias mokyklas duomenų tyrimas panaudojant asociacijų taisykles***

Kompiuterizuoto bendrojo priėmimo į Lietuvos aukštąsias mokyklas antrinis rezultatas yra sukaupti duomenų įrašai. Svarbus uždavinys yra šiuos duomenis efektyviai panaudoti, tai yra, iš duomenų gauti vertingos informacijos. Duomenyse atrasti dėsningumai bei sąryšiai gali padėti geriau suprasti stojančiųjų prioritetus, įvairių studijų programų populiarumą ir patį priėmimo į aukštąsias mokyklas procesą. Panaudojant asociacijų taisyklių duomenų tyrimo metodą, šiame straipsnyje nagrinėjami sąryšiai tarp stojančiųjų lyties ir jų norimų studijuoti studijų programų. Keletu iliustracinių pavyzdžių aprašoma duomenų tyrimo procedūra, kuri gali būti naudojama ir kitiems panašiems uždaviniams spręsti.