STUDIA UNIV. BABEŞ–BOLYAI, INFORMATICA, Volume **LXIII**, Number 2, 2018
DOI: 10.24193/subbi.2018.2.07

INCREMENTAL RELATIONAL ASSOCIATION RULE MINING OF EDUCATIONAL DATA SETS

LIANA MARIA CRIVEI

Abstract. *Educational Data Mining* is an attractive research field in which the underlying idea is that of bringing the *data mining* perspective into *educational environments*. The main focus is to better understand the educational related phenomena by extracting, through data mining techniques, meaningful hidden patterns from educational data sets. *Incremental Relational Association Rule Mining* ($IRARM$) has been introduced as an effective *online data mining* method for dynamically mining interesting *relational association rules* (RARs) in a dynamic data set which is extended with new data instances. The study conducted in this paper is aimed to emphasize the effectiveness of both RAR and $IRARM$ mining methods in *educational data mining* settings. Experiments performed on various academic data sets highlight the potential of using *relational association rules* for uncovering relevant knowledge from educational related data.

1. Introduction

Data mining (DM) techniques are extensively applied nowadays in various domains including medicine, bioinformatics, software engineering, to discover relevant patterns in large databases, especially due to their potential of uncovering hidden information from data.

Applying DM techniques in education [5] has attracted researchers from both DM and educational research and thus a new interdisciplinary research discipline known as *educational data mining* (EDM) emerged. The main focus in EDM is to develop methods for extracting knowledge from data that come from various educational information systems and educational environments.

Received by the editors: April 15, 2018.

2010 *Mathematics Subject Classification.* 68T05, 68P15.

1998 *CR Categories and Descriptors.* H.2.8[**Database management**]: Database Applications – *Data Mining*; I.2.6[**Computing Methodologies**]: Artificial Intelligence – *Learning*.

Key words and phrases. Data mining, Educational data mining, Relational association rule, Incremental algorithm.

Through mining educational data sets, EDM's purpose is to better understand the students' learning process and thus to offer additional insights into educational related phenomena.

Within the DM domain, *association rule* (AR) mining represents an important data analysis and mining technique [9] applied in various *supervised* and *unsupervised* learning scenarios for extracting rule based patterns from data sets. *Ordinal association rules* (OARs) [7] were proposed as a particular class of ARs which express ordinal relationships between the attributes characterizing a data set. Afterwards, *relational association rules* (RARs) [6, 11] have been introduced as an extension of OARs capable to capture various type of non-ordinal relations between data attributes.

We are approaching in this paper the problem of *incremental relational association rule mining* (*IRARM*) in the context of EDM. The process of incremental RAR mining is appropriate specifically for *online* DM scenarios, where the data set to be mined is dynamic and thus continuously extended with real-time arriving data streams. In such situations, *IRARM* approach aims to progressively adapt the interesting RARs identified in a data set, when it is enlarged with new instances. Since the learning processes within educational environments are by nature online processes, the idea of investigating the *IRARM* perspective in EDM comes naturally. The EDM literature also reveals that DM is very useful in the educational field particularly when exploring the online learning environment [18].

The contribution of the paper is summarized as follows. First, we are emphasizing the relevance of RAR mining in the field of *educational data mining* (EDM) with the goal of uncovering meaningful patterns within educational data sets. Secondly, we extend the experimental evaluation of our previously proposed *incremental relational association rule* mining approach (*IRARM*) [17] on several EDM case studies. The effectiveness of *IRARM* is emphasized through the reduction in mining time achieved when using *IRARM* against RAR mining from scratch when a data set is extended with new instances. The study conducted in this paper is novel in the EDM literature, since neither the classical nor the incremental RAR mining approaches have been applied on academic data sets, so far.

The rest of the paper is structured as follows. Section 2 introduces the EDM domain and emphasizes its relevance within the larger DM field. A background on RAR mining and its incremental extension *IRARM* is presented in Section 3, together with an example of RAR mining in EDM. Section 4 describes the experiments performed for highlighting the performance of *IRARM* on four academic data sets and discusses upon the obtained experimental results. The conclusions of the paper and directions for future improvements are highlighted in Section 5.

2. Educational Data Mining

EDM is an attractive research field in which the underlying idea is that of bringing the *data mining* perspective into *educational environments*. The main focus is to better understand the educational related phenomena by extracting, through data mining techniques, meaningful hidden patterns from educational data sets.

Extracting relevant patterns from the educational processes would also be useful for understanding students and how they learn, as well as improving the educational outcomes (e.g. learning outcomes). EDM has received lately considerable attention from the research community since extracting hidden knowledge from educational data is of particular interest for the academic institutions and also useful for improving their teaching methodologies and learning processes [18].

Various applications using data mining techniques have been developed, so far, in the EDM field. *Machine learning* methods are intensively investigated, both from a *supervised* and *unsupervised* perspective, as data mining techniques for building course planning systems, detecting what type of learners are the students, grouping students according to their similarity, predicting the students' performance for courses, assisting instructors in the educational process [15].

We briefly review, in the following, several recent approaches which have been developed for assessing the performance of students in educational environments.

Ayers et al. applied in [3] several clustering algorithms such as hierarchical agglomerative clustering, K-means and model based clustering for grouping students according to their skill sets.

Bharadwaj and Pal conducted in [4] a study towards identifying features which are strongly correlated with the academic students' performances. The authors found out that characteristics such as the living location, medium of teaching, mother's qualification, the family annual income, and student's family status highly influence the performance of the prediction task. Pal and Pal conducted in [21] a study using decision tree based classification algorithms to identify the students needing special advising and counseling from the teachers.

Supervised classification models such as *Naive Bayes*, *decision trees*, *neural networks* have been applied in [15] together with *Synthetic Minority Over-Sampling* (SMOTE) method to improve the accuracy of a machine learning model for predicting the students' final grade for a particular course. An analysis of the performance of the previous mentioned machine learning models, including *support vector* classification was performed by Shahiri et al. in [23]. Additionally, a study was conducted upon the effectiveness of the attributes involved in the classification process.

Ahmed et al. focused in [2] on predicting the performance of instructors and analyzed the factors that affect students' academic achievements, with the purpose of improving the quality of the educational system. Several classifiers such as J48 Decision Tree, Multilayer Perceptron, Naïve Bayes, and Sequential Minimal Optimization were applied and compared to identify the best performing classification algorithm. Among all considered classifiers, J48 provided the best classification accuracy of 84.8%.

The problem of predicting the students performance (PSP) has been considered in [24] as regression problem and a hybrid method combining a collaborative filtering-based system and a regression-based one has been proposed.

3. Background on relational association rules

In Section 3 the fundamental concepts related to *relational association rule* (RAR) mining [11] are reviewed. Then, the relevance and importance of RAR mining in the context of EDM is emphasized through an example on an educational data set. Section 3.2 briefly presents the *incremental relational association rule* mining (*IRARM*) approach [17].

3.1. Relational association rule mining.

Association rule (AR) mining represents an important data analysis and mining technique [9] useful in multiple *machine learning* tasks for uncovering meaningful rule based patterns in data sets. *Ordinal association rules* (OARs) [7] were proposed as a particular class of ARs which express ordinal relationships between the attributes characterizing a data set. RARs [6, 11] have been introduced as an extension of OARs able to express different type of non-ordinal relations between data attributes.

The *Relational Association Rules* (*RARs*) notion is defined in the following paragraphs.

We consider $\mathcal{D} = \{d_1, d_2, \ldots, d_n\}$ a set of *instances* or *records*. Let $\Omega = (a_1, \ldots, a_m)$ be a sequence of m attributes characterizing each instance from the data set $\mathcal{D}$. Each attribute a_i takes values from a non-empty and non-fuzzy domain Δ_i, which also contains a *null* (*empty*) value. We denote by $\Psi(d_j, a_i)$ the value of attribute a_i for an instance d_j.

We denote by $\mathcal{T}$ the set of all possible relations that are not necessarily ordinal which can be defined between two domains Δ_i and Δ_j.

Definition 3.1. *A* relational association rule [11] *is an expression*

$$(a_{i_1}, a_{i_2}, a_{i_3}, \ldots, a_{i_h}) \Rightarrow (a_{i_1}\tau_1 a_{i_2}\tau_2 a_{i_3} \ldots \tau_{h-1} a_{i_h}),$$

where $\{a_{i_1}, a_{i_2}, a_{i_3}, \ldots, a_{i_h}\} \subseteq \Omega$, $a_{i_k} \neq a_{i_p}$, $k, p = 1, \ldots, h$, $k \neq p$ *and* $\tau_k \in \mathcal{T}$ *is a relation over* $\Delta_{i_k} \times \Delta_{i_{k+1}}$, Δ_{i_k} *being the domain of the attribute* a_{i_k}.

a) *If* $a_{i_1}, a_{i_2}, \ldots, a_{i_h}$ *are non-missing in* m *instances from the data set then we call* $s = \frac{m}{n}$ *the* support *of the rule.*

b) *If we denote by* $D' \subseteq \mathcal{D}$ *the set of instances where* $a_{i_1}, a_{i_2}, a_{i_3}, \ldots, a_{i_h}$ *are non-missing and all relations* $\Psi(d_j, a_{i_1})\tau_1\Psi(d_j, a_{i_2})$, $\Psi(d_j, a_{i_2})\tau_2$ $\Psi(d_j, a_{i_3})$, $\ldots$, $\Psi(d_j, a_{i_{h-1}})\tau_{h-1}\Psi(d_j, a_{i_h})$ *hold for each instance* d *from* D' *then we call* $c = \frac{|D'|}{n}$ *the* confidence *of the rule.*

Interesting RAR's were defined in [11] as those rules which have both their *support* and *confidence* greater than or equal to specified minimum thresholds. For mining interesting RARs an Apriori-like algorithm named *DRAR* (Discovery of Relational Association Rules) was proposed in [12] as an extension of the *DOAR* algorithm introduced in [7] for uncovering OARs.

3.1.1. *Example.* For a better understanding of the concept of RAR, an example on an EDM related data set is considered. The aim is to highlight the relevance of applying RAR mining in the context of educational data sets.

The data set used in our example is a real data set, containing the grades obtained by students at a Computer Science undergraduate course offered at Babeş-Bolyai University in a time frame of four academic years (2014-2018). The complete data set is available at [1]. There are a total of 867 instances characterized by 6 attributes, denoted by $a_1, a_2, \ldots, a_6$. These attributes represent the following: written exam score (a_1), seminar score (a_2), laboratory score (a_3), first practical test score (a_4), second practical test score (a_5) and final grade (a_6). Considering the minimum support threshold at $s_{min} = 1$ and the minimum confidence threshold at $c_{min} = 0.6$, we applied *DRAR* mining algorithm. Since all the attributes in our experiment have integer values, two possible binary relations between integer valued attributes were used: $\leq$ and $>$. The discovered maximal interesting RARs are illustrated in Table 1.

Each line from Table 1 describes a RAR of a certain length (depicted in the first column), which has the confidence illustrated in the third column. For example, the first line in Table 1 refers to the RAR $a_1 \leq a_3$ of length **2** (i.e. the rule contains two attributes) having a confidence of **0.739**. This rule has the following interpretation: the value of the attribute a_1 is less or equal than the value of the attribute a_3 in 73.9% of instances from the analyzed data.

Analyzing the last rule depicted in Table 1 one observes that for 61.3 % of the students the grade for the written exam is less or equal than the first test

Length	Rule	Confidence
2	$a_1 \leq a_3$	0.739
2	$a_1 \leq a_5$	0.751
2	$a_1 \leq a_6$	0.825
2	$a_2 \leq a_3$	0.751
2	$a_2 \leq a_4$	0.828
2	$a_2 \leq a_5$	0.819
2	$a_2 \leq a_6$	0.669
2	$a_3 \leq a_4$	0.722
2	$a_3 \leq a_5$	0.711
2	$a_3 > a_6$	0.605
2	$a_4 \leq a_5$	0.713
2	$a_5 > a_6$	0.604
3	$a_1 \leq a_4 > a_6$	0.613

TABLE 1. Interesting maximal relational association rules mined for $s_{min} = 1$ and $c_{min} = 0.6$.

grade, which is greater than the final grade. This suggests that the grades for the practical test are greater than those for the written exam, which could be considered typical because the written exam requires wider knowledge. Analyzing other interesting rules depicted in the table 1: $a_3 \leq a_4$ and $a_3 \leq a_5$ we observe that the grade obtained for the laboratory is less than the both practical test scores. This is an indication that some of the laboratory assignments are more difficult or complex than the actual practical test. The complexity of the practical test could be increased. We also observe that $a_3 > a_6$ meaning the laboratory score is less than the final grade score.

The RARs mined from the academic data may be relevant for the professor and can provide indications about the complexity of the laboratory assignments or the written exams.

3.2. Incremental relational association rule mining.

We have previously introduced in [17] an *incremental relational association rule mining* approach, called *IRARM*, which is useful when a data set to be mined is extended with new objects. In such situations, for uncovering the interesting RARs from the extended data set, *IRARM* will efficiently adapt the RARs discovered in the data set before the extension. This incremental process will be more effective than running *DRAR* from scratch on the extended data set.

We consider in the following that the data set $\mathcal{D}$ to be mined is dynamic, being extended at a certain time with a non-empty set of instances $\{d_{n+1}, d_{n+2},$

$\ldots, d_s\}$. We denote the enlarged set of instances by $\mathcal{D}^{ext} = \{d_1, d_2, \ldots, d_s\}$, while the set of newly added instances is $\mathcal{D}^{new} = \mathcal{D}^{ext} \setminus \mathcal{D}$. For pre-specified minimum support (s_{min}) and confidence thresholds (c_{min}), we analyze the problem of incrementally identifying all *interesting* RARs in the extended data set $\mathcal{D}^{ext}$ by adapting the set of interesting RARs mined in $\mathcal{D}$ before its extension. Through the *IRARM* method we aim to reduce the running time required to mine the set $\mathcal{R}ules^{ext}$ of interesting RARs from $\mathcal{D}^{ext}$.

Certainly, new interesting RARs could be produced by the newly added instances, but also RARs which were interesting enough in the data set before extension may become uninteresting on the extended data set. The set $\mathcal{R}ules^{ext}$ of all interesting RARs may be discovered by applying the *DRAR* method from scratch, on the extended set of objects. But this process can be computationally expensive. That is why our goal is to replace it by a more efficient algorithm *IRARM* (*Incremental Relational Association Rule Mining*), which preserves the completeness of the RARs generation procedure. Considering the newly added instances, *IRARM* adjusts the set $\mathcal{R}ules$ of all interesting RARs in the initial data set $\mathcal{D}$ to produce the set of all interesting RARs in the extended data set $\mathcal{D}^{ext}$.

The idea behind determining the set $\mathcal{R}ules^{ext}$ will be further described. Two main *stages* characterize the *IRARM* method. The **first stage** is **filtering** the set $\mathcal{R}ules$ of interesting RARs from the initial data set $\mathcal{D}$ in order to maintain only the rules which are interesting in the extended data set $\mathcal{D}^{ext}$ as well. The **second stage** consists of **extending** the subset previously obtained with new rules which were not interesting in $\mathcal{D}$, but become interesting on the extended data set $\mathcal{D}^{ext}$. After the second phase is completed the set $\mathcal{R}ules^{ext}$ of interesting RARs from the extended data set $\mathcal{D}^{ext}$ will have been mined.

More details about the description of *IRARM* algorithm can be found in [17].

The educational process is essentially dynamic therefore the results of the evaluation for new students are available in an incremental manner. While new information is accessible the existent academic data set is continuously updated. In such situations the discovery of interesting relational association rules in academic data sets is an incremental process. Consequently it is more efficient from a computational viewpoint to apply the *IRARM* incremental method (by adapting the set of rules identified before updating the data) rather than applying *DRAR* from the scratch on the entire data set.

4. Results and discussion

We provide in the following an experimental evaluation of *IRARM* on two academic data sets, as well as a discussion upon the obtained results.

4.1. **Case studies and data sets.**

In order to evaluate the performance of *IRARM*, two case studies will be conducted on four educational data sets.

The first case study is performed starting from a real academic data set collected from the Babeş-Bolyai University.

4.1.1. *First case study.* The *first data set* used in our study is the real data set described in Section 3.1.1 and available at [1].

The *second data set* considered in our evaluation is synthetically generated from the first data set and is available at [1]. The number of instances from this data set is 867, as in our first data set. Since our first data set contained a relatively small number of attributes, namely 6, we extended the set of attributes to 10 attributes, $a_1, a_2, \ldots, a_{10}$. The first 5 attributes from this data set have the same meaning as described in Section 3.1.1. Attributes a_6, a_7, a_8 and a_9 represent the scores for four additional practical tests, while the last attribute a_{10} represents the final grade. We mention that the values for the added attributes a_6–a_9 were randomly generated, using a uniform distribution, within the interval determined by the attributes a_2, a_3, a_4, a_5.

4.1.2. *Second case study.* The second case study used for evaluating *IRARM* contains the *Turkiye Student Evaluation* data sets publicly available at [14]. There are two data sets (*Turkey Student Evaluation Generic* and *Turkey Student Evaluation Specific*) each containing a total of 5820 evaluation scores provided by students from Gazi University in Ankara (Turkey). Each data set contains a total of 28 course specific questions and additional 5 specific attributes. Details about the attributes can be found at [14].

4.2. **Experimental results.**

Let us denote by s the number of instances from the analyzed data set. For both data sets from our first case study $s = 867$, while for the data sets from the second case study $s = 5820$.

In the performed experiments, for all data sets considered for evaluation, the following experimental methodology was applied. We have started with n instances in the data set (for various values for n) and afterwards the data set was extended with $s - n$ entities. Different values were used for the minimum confidence threshold c_{min}, while s_{min} was set to 1 since our data sets do not contain missing values.

For each experiment, the set of interesting RARs on the extended data set containing s instances was obtained in two ways:

(1) by adapting using *IRARM* the set of RARs obtained on the data set before its extension;
(2) by applying the *DRAR* method from scratch on the extended data set.

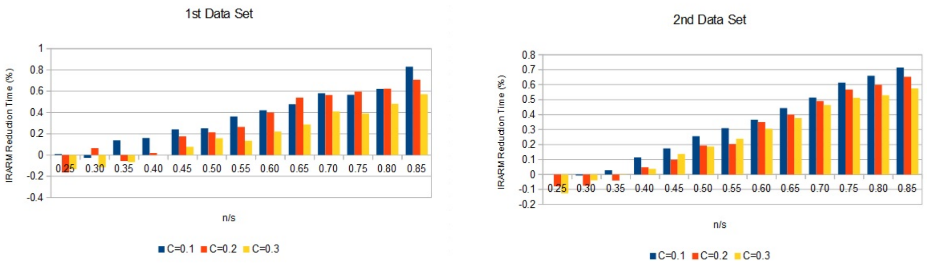
We mention that, using method (1) or (2), the set of interesting RARs discovered in data is the same, but we expect the total mining time for $IRARM$ to be lower than the total mining time of $DRAR$ applied from scratch. The experiments presented in this section were performed on a PC with an Intel Core i7 Processor at 2.30 GHz, with 4 GB of RAM.

In the mining process, we used the following binary relations between the integer valued attributes: $>, <, =$.

For all four data sets from our case studies, we have repeatedly run $DRAR$ and $IRARM$ for different values for c_{min} and different values for $\frac{n}{s}$. Tables 2 and 3 present the results obtained when considering $c_{min} = 0.1$ and varying $\frac{n}{s}$ from 0.25 to 0.85 with a step size of 0.05. For a certain combination of parameters (n, s, c_{min}), the mining method ($DRAR$ or $IRARM$) was executed 20 times and the results were averaged over these executions. The fifth column from the tables gives the reduction in total running time achieved by $IRARM$ computed as $\frac{DRAR\,time - IRARM\,time}{DRAR\,time}$.

From Tables 2 and 3 we observe that, when the percentage of initial instances $\frac{n}{s}$ is larger than 0.35, the running time of $IRARM$ is increasingly reduced with respect to the running time of $DRAR$, as the number of instances added to the data set decreases. The maximum reduction in mining time obtained by $IRARM$ is achieved when $\frac{n}{s}$ is 0.85 and is higher than 70%.

Figure 1 depicts, for the data sets from our first case study, how the percentage of $IRARM$'s running time reduction increases when increasing $\frac{n}{s}$, for three different minimum confidence thresholds: 0.1, 0.2 and 0.3.



(A) *First data set.*

(B) *Second data set.*

FIGURE 1. $IRARM$'s reduction in total mining time for the data sets from our first case study, using different minimum confidence thresholds.

Figures 2 and 3 illustrate, for the data sets from the first case study, how the running time for the main operations of $IRARM$ algorithm (*Filter* function, candidates generation process, *Select* function, support and confidence

Data set	**n**	**s-n**	**Time DRAR** (ms)	**Time** *IRARM* (ms)	*IRARM* **time** reduction (%)
First data set	217	650	5.6	5.55	**0.0089**
	260	607	5.65	5.8	**-0.0265**
	303	564	5.8	5	**0.1379**
	347	520	5.65	4.75	**0.1593**
	390	477	5.8	4.4	**0.2414**
	433	434	5.8	4.35	**0.25**
	477	390	5.8	3.7	**0.3621**
	520	347	5.7	3.3	**0.4211**
	564	303	5.45	2.85	**0.4771**
	607	260	5.6	2.35	**0.5804**
	650	217	5.3	2.3	**0.5660**
	694	173	5.55	2.1	**0.6216**
	737	130	5.85	1	**0.8291**
Second data set	217	650	15.05	15.05	**0**
	260	607	14.9	15	**-0.0067**
	303	564	14.8	14.4	**0.0270**
	347	520	15.05	13.35	**0.1130**
	390	477	14.95	12.35	**0.1739**
	433	434	15.05	11.2	**0.2558**
	477	390	15	10.35	**0.31**
	520	347	15.15	9.6	**0.3663**
	564	303	15	8.35	**0.4433**
	607	260	14.9	7.25	**0.5134**
	650	217	15.15	5.85	**0.6139**
	694	173	15.15	5.15	**0.6601**
	737	130	15.1	4.3	**0.7152**

Table 2. Experimental results on the data sets from our first case study for $s_{min} = 1$ and $c_{min} = 0.1$.

computation) evolved when varying $\frac{n}{s}$ for $c_{min} = 0.1$. From the figures we observe that running times for the *Filter* and *Select* operations decrease while $\frac{n}{s}$ increases.

Figure 4 illustrates for the data sets from the second case study, how the percentage of *IRARM*'s running time reduction increases when increasing $\frac{n}{s}$, for two different minimum confidence thresholds: 0.8 and 0.85.

The experimental results presented in this section highlighted the effectiveness of *IRARM* method, which reduces the mining time against the time achieved by applying DRAR mining from scratch when a data set is extended with new instances.

4.3. **Comparison to related work.** The *incremental relational association rule mining* approach previously introduced in [17] and applied in this paper on educational data sets is new both in the DM and EDM literature. Existing incremental approaches from the DM literature handle only *non-relational*

Data set	**n**	**s-n**	**Time DRAR** (ms)	**Time** *IRARM* (ms)	*IRARM* **time** reduction (%)
Turkiye Student Evaluation Generic data set [14]	1455	4365	814	875.2	**-0.0751**
	1746	4074	552.2	564.75	**-0.0227**
	2037	3783	782.7	703.25	**0.1015**
	2328	3492	794.35	664.55	**0.1634**
	2619	3201	834.65	678.35	**0.1873**
	2910	2910	765	564.95	**0.2615**
	3201	2619	816.85	525.15	**0.3571**
	3492	2328	692.4	386.1	**0.4424**
	3783	2037	739.95	358.7	**0.5152**
	4074	1746	570	238.25	**0.5820**
	4365	1455	809.65	312.35	**0.6142**
	4656	1164	628.55	228.3	**0.6368**
	4947	873	501.65	156.45	**0.6881**
Turkiye Student Evaluation Specific data set [14]	1455	4365	686.7	751.65	**-0.0946**
	1746	4074	684.35	695.65	**-0.0165**
	2037	3783	742.4	676.55	**0.0887**
	2328	3492	845.95	704.05	**0.1677**
	2619	3201	967.55	779.25	**0.1946**
	2910	2910	1010.95	637.9	**0.3690**
	3201	2619	980.75	540.4	**0.4490**
	3492	2328	936.45	475.7	**0.4920**
	3783	2037	730.25	383.85	**0.4744**
	4074	1746	565.1	252.7	**0.5528**
	4365	1455	748.35	291.85	**0.6100**
	4656	1164	952.9	360.8	**0.6214**
	4947	873	961.85	288.2	**0.7004**

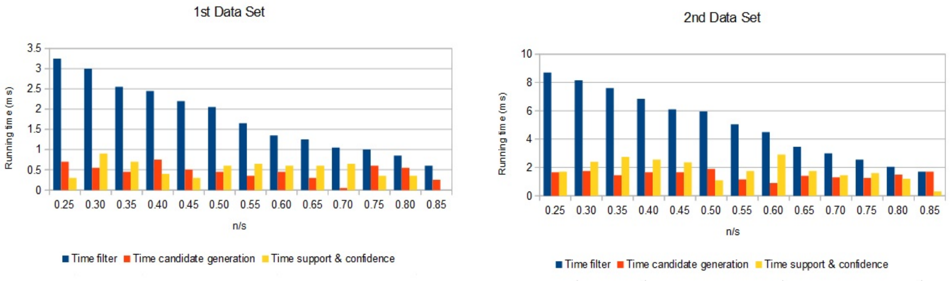
TABLE 3. Experimental results on the data sets from our second case study for $s_{min} = 1$ and $c_{min} = 0.8$.

association rules. In the EDM literature we have not found, so far, approaches using *relational association rule mining* or *incremental relational association rule mining* on EDM scenarios.

We present in the following several data mining methods which deal with the *incremental mining* perspective on *non-relational* association rules.

Sarda and Srinivas introduced in [22] an algorithm for incremental association rule mining, in which the data set is extended with new instances. The proposed adaptive algorithm was able to identify new rules for the updated database, avoiding multiple scans of it. Yafi *et al.* proposed in [25] an incremental association rules mining algorithm named YAMI based on the Apriori model on evolving databases. The authors also introduced the concept of *shocking interesting rule*, as a rule which surpass all user's expectations.

The incremental association rule mining on dynamic transactional databases was investigated by Chandraker and Sao in [8]. Nath *et al.* present in [19] a



(A) *First data set.*

(B) *Second data set.*

FIGURE 2. Running time (ms) for the main operations of $IRARM$, for both data sets from the first case study, for $c_{min} = 0.1$.

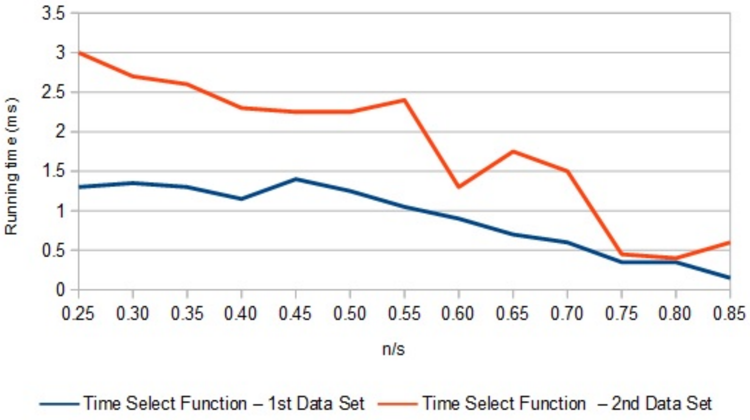
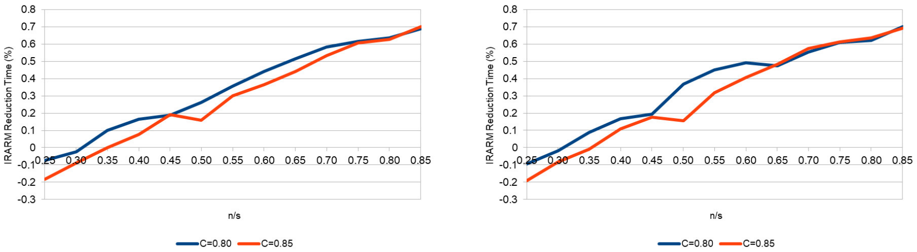


FIGURE 3. Running time (ms) for the *Select* function, for both data sets from the first case study, for $c_{min} = 0.1$.

survey on incremental association rule mining. They review frequent itemset generation techniques, rule generation techniques and incremental association rule mining techniques. The authors emphasize several research issues and challenges, such as the incremental behaviour of the data set, the number of data set scans and the number of generated candidate itemsets. Dhanabhakyam and Punithavalli [13] propose An Adaptive Association Rule Mining with Faster Rule Generation Algorithm (FRG-AARM) with the intent of acquiring a more efficient Market Basket Analysis.



(A) *Turkiye Student Evaluation Generic* data set.

(B) *Turkiye Student Evaluation Specific* data set.

FIGURE 4. *IRARM*'s reduction in total mining time for the data sets from the second case study, using different minimum confidence thresholds.

Ogunde *et al.* [20] introduced an Adaptive Incremental Mining Algorithm (AIMA) aimed to adapt to the trend of constant data updates in distributed databases.

An incremental association rule mining algorithm has been proposed by Yu-Dong *et al.* [26]. This was named VSIFP-Growth (Improved FP-Growth) and used together with parallel computing techniques with the purpose of developing the PVSIFP-Growth algorithm for frequent itemsets generation. Li et al. [16] developed *TDUP*, a *three-way decision update pattern approach* together with a synchronization mechanism in order to reduce the number of scans of the initial data set.

The above presented methods deal with *incremental* AR mining, but from a *non-relational* perspective. Unlike the classical *association rules*, RARs are capable to express relationships between data attributes. Thus, RARs may be more powerful than classical ARs in various machine learning scenarios, including those related to EDM tasks.

5. Conclusions and future work

We investigated in this paper the application of classical and incremental RAR mining for knowledge discovery in data sets from educational environments, with the goal of uncovering meaningful patterns within educational data sets. The relevance of uncovering RARs in academic data sets has been emphasized in the context of the students' learning process, as offering additional insights into educational related phenomena. Additionally, the effectiveness of *incremental* RAR mining in online EDM scenarios was highlighted through several case studies.

Future work will be done in order to extend the experimental evaluation of *IRARM* on other EDM tasks, to further test its performance. An *incremental adaptive* RAR mining will be also investigated for academic data sets, when both new instances and new features are added to the data set. Furthermore, we plan to apply RAR, gradual RARs [10] and *IRARM* mining algorithm in supervised learning EDM scenarios, such as predicting student's academic performance.

Acknowledgements

The author acknowledges the assistance received by using the UCI Machine Learning Repository.

References

[1] Academic data set, 2018. http://www.cs.ubbcluj.ro/∼liana.crivei/AcademicDataSets.

[2] Ahmed Mohamed Ahmed, Ahmet Rizaner, and Ali Hakan Ulusoy. Using data mining to predict instructor performance. *Procedia Computer Science*, 102:137 – 142, 2016. 12th International Conference on Application of Fuzzy Systems and Soft Computing, ICAFS 2016, 29-30 August 2016, Vienna, Austria.

[3] Elizabeth Ayers, Rebecca Nugent, and Nema Dean. A comparison of student skill knowledge estimates. In *Educational Data Mining - EDM 2009, Cordoba, Spain, July 1-3, 2009. Proceedings of the 2nd International Conference on Educational Data Mining.*, pages 1–10, 2009.

[4] Brijesh Kumar Baradwaj and Saurabh Pal. Mining educational data to analyze students' performance. *CoRR*, abs/1201.3417, 2012.

[5] Alejandro Bogarín, Rebeca Cerezo, and Cristóbal Romero. A survey on educational process mining. *Wiley Interdisc. Rew.: Data Mining and Knowledge Discovery*, 8(1), 2018.

[6] Alina Câmpan, Gabriela Şerban, and Andrian Marcus. Relational association rules and error detection. *Studia Universitatis Babes-Bolyai Informatica*, LI(1):31–36, 2006.

[7] Alina Campan, Gabriela Şerban, Traian Marius Truta, and Andrian Marcus. An algorithm for the discovery of arbitrary length ordinal association rules. In *DMIN*, pages 107–113, 2006.

[8] Toshi Chandraker and Neelabh Sao. Incremental mining on association rules. *International Jurnal of Engineering and Science*, 1(11):31–33, 2012.

[9] H. Y. Chang, J. C. Lin, M. L. Cheng, and S. C. Huang. A novel incremental data mining algorithm based on fp-growth for big data. In *2016 International Conference on Networking and Network Applications (NaNA)*, pages 375–378, July 2016.

[10] I. G. Czibula, G. Czibula, D.-L. Miholca. Enhancing relational association rules with gradualness. *International Journal of Innovative Computing, Information & Control*, 13(1):289-305, 2017.

[11] Gabriela Şerban, Alina Câmpan, and Istvan Gergely Czibula. A programming interface for finding relational association rules. *International Journal of Computers, Communications & Control*, I(S.):439–444, June 2006.

[12] Gabriela Czibula, Maria-Iuliana Bocicor, and Istvan Gergely Czibula. Promoter sequences prediction using relational association rule mining. *Evolutionary Bioinformatics*, 8:181–196, 04 2012.

[13] M. Dhanabhakyam and M. Punithavalli. An efficient market basket analysis based on adaptive association rule mining with faster rule generation algorithm. *The Standard International Journals on Computer Science Engineering and its Applications (CSEA)*, 1(3):105–110, 2013.

[14] N. Gunduz and E. Fokoue. UCI machine learning repository, 2013.

[15] Syed Tanveer Jishan, Raisul Islam Rashu, Naheena Haque, and Rashedur M. Rahman. Improving accuracy of students' final grade prediction model using optimal equal width binning and synthetic minority over-sampling technique. *Decision Analytics*, 2(1):1, Mar 2015.

[16] Yao Li, Zhi-Heng Zhang, Wen-Bin Chen, and Fan Min. Tdup: an approach to incremental mining of frequent itemsets with three-way-decision pattern updating. *International Journal of Machine Learning and Cybernetics*, 8(2):441–453, Apr 2017.

[17] Diana-Lucia Miholca, Gabriela Czibula, and Liana Maria Crivei. A new incremental relational association rules mining approach. In *22nd International Conference on Knowledge-Based and Intelligent Information & Engineering Systems*, KES2018, page to be published. Procedia Computer Science, 2018.

[18] Siti Khadijah Mohamad and Zaidatun Tasir. Educational data mining: A review. *Procedia - Social and Behavioral Sciences*, 97:320 – 324, 2013. The 9th International Conference on Cognitive Science.

[19] B. Nath, D. K. Bhattacharyya, and A. Ghosh. Incremental association rule mining: A survey. *Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 3(3):157–169, 2013.

[20] Adewale O. Ogunde, Olusegun Folorunso, and Adesina S. Sodiya. The design of an adaptive incremental association rule mining system. In *Proceedings of the World Congress on Engineering 2015 - Volume I*, London, UK, 2015.

[21] Kumar Ajay Pal and Saurabh Pal. Analysis and mining ofeducational data forpredictingthe performance of students. *International Journal of ElectronicsCommunication and Computer Engineering*, 4(5):278—-4209, 2013.

[22] N. L. Sarda and N. V. Srinivas. An adaptive algorithm for incremental mining of association rules. In *Proceedings of the 9th International Workshop on Database and Expert Systems Applications*, DEXA '98, pages 240–, Washington, DC, USA, 1998. IEEE Computer Society.

[23] Amirah Mohamed Shahiri, Wahidah Husain, and Nur'aini Abdul Rashid. A review on predicting student's performance using data mining techniques. *Procedia Computer Science*, 72:414 – 422, 2015. The Third Information Systems International Conference 2015.

[24] Thi-Oanh Tran, Hai-Trieu Dang, Viet-Thuong Dinh, Thi-Minh-Ngoc Truong, Thi-Phuong-Thao Vuong, and Xuan-Hieu Phan. Performance prediction for students: A multi-strategy approach. *CYBERNETICS AND INFORMATION TECHNOLOGIES*, 17(2):164 – 182, 2017.

[25] Eiad Yafi, Ahmed Al-Hegami, Afshar Alam, and Ranjit Biswas. YAMI: Incremental mining of interesting association patterns. *The International Arab Jurnal of Information Technology*, 9(6):504–510, 2012.

[26] Guo Yu-Dong, Li Sheng-Lin, Li Yong-Zhi, Wang Zhao-Xia, and Zeng Li. Large-scale dataset incremental association rules mining model and optimization algorithm. *International Journal of Database Theory and Application*, 9(4):195–208, 2016.

Department of Computer Science,, Faculty of Mathematics and Computer Science,, Babeş-Bolyai University, Kogălniceanu 1, Cluj-Napoca, 400084, Romania

Email address: liana.crivei@cs.ubbcluj.ro