Mining Rare Association Rules from e-Learning Data

Cristóbal Romero, José Raúl Romero, Jose María Luna, Sebastián Ventura
cromero@uco.es, jrromero@uco.es, i32luarj@uco.es, sventura@uco.es
Dept. of Computer Science, University of Córdoba, Spain

Abstract. Rare association rules are those that only appear infrequently even though they are highly associated with very specific data. In consequence, these rules can be very appropriate for using with educational datasets since they are usually imbalanced. In this paper, we explore the extraction of rare association rules when gathering student usage data from a Moodle system. This type of rule is more difficult to find when applying traditional data mining algorithms. Thus we show some relevant results obtained when comparing several frequent and rare association rule mining algorithms. We also offer some illustrative examples of the rules discovered in order to demonstrate both their performance and their usefulness in educational environments.

1 Introduction

Nowadays most research on Association Rule Mining (ARM) has been focused on discovering common patterns and rules in large datasets. In fact, ARM is widely and successfully used in many different areas, such as telecommunication networks, market and risk management, inventory control, mobile mining, graph mining, educational mining, etc. The patterns and rules discovered are based on the majority of commonly repeated items in the dataset, though some of these data can be either obvious or irrelevant [5]. Unfortunately, not enough attention has been paid to the extraction process of rare association rules, also known as non-frequent, unusual, exceptional or sporadic rules, which provide valuable knowledge about non-frequent patterns. The aim of Rare Association Rule Mining (RARM) is to discover rare and low-rank itemsets to generate meaningful rules from these items. Notice that this specific type of rule cannot be revealed easily using traditional association mining algorithms.

In previous works, other authors have applied ARM to e-learning systems extensively to discover frequent student-behavior patterns [13], [7]. However, RARM has been hardly applied to educational data, despite the fact that infrequent associations can be of great interest since they are related to rare but crucial cases. For instance, they might allow the instructor to verify a set of rules concerning certain infrequent/abnormal learning problems that should be taken into account when dealing with students with special needs. Thus, this information could help the instructor to discover a minority of students who may need specific support with their learning process. From the perspective of knowledge discovery, the greatest reason for applying RARM in the field of education is the imbalanced nature of data in education, as in other real-world tasks, i.e., some classes have many more instances than others. Furthermore, in applications like education, the minor parts of an attribute can be more interesting than the major parts; for example, students who fail or drop out are usually less frequent than those students who fare well. In the field of association rule mining, the rare item problem [6] is essentially considered to be a data imbalance problem which may, on either side of the association rule, give rise to severe problems. The problem of imbalance has only been dealt with in one educational data mining study [9]. However, in this work, data was firstly modified/preprocessed to solve the problem of imbalance and then several

different classification algorithms were applied instead of specific association rule algorithms.

In this paper, we explore the application of RARM to student data stored in a large Moodle repository to discover information about infrequent student behavior. This paper is organized as follows. Section 2 presents some background on frequent and infrequent association rule mining while Section 3 describes the experiments carried out and the analysis of the most relevant results obtained, as well as including a description of the most accurate rules mined by applying both ARM and RARM to a Moodle dataset containing real information. Finally, conclusions are outlined in Section 4.

2 Background

Association Rule Mining is one of the most popular and well-known data mining methods for discovering interesting relationships between variables in transaction databases or other data repositories [2]. An association rule is an implication $X \Rightarrow Y$, where X and Y are disjoint itemsets (i.e., sets with no items in common). The intuitive meaning of such a rule is that when X appears, Y also tends to appear. The two traditional measures for evaluating association rules are support and confidence. The confidence of an association rule $X \Rightarrow Y$ is the proportion of the transactions containing X which also contain Y. The support of the rule is the fraction of the database that contains both X and Y. The problem of association rule mining is usually broken down into two subtasks. The first one is to discover those itemsets whose occurrences exceed a predefined support threshold, and which are called frequent itemsets. A second task is to generate association rules from those large items constrained by minimal confidence. Nowadays, the problem of frequent itemset mining has been studied widely and many algorithms have already been proposed [3], mainly variations or improvements of the Apriori algorithm [2] which is the first, simplest and most common ARM algorithm. Most research in the area of ARM is focused on the sub-problem of efficient frequent rule generation. However in some data mining applications relatively infrequent associations are likely to be of great interest, too. Though these algorithms are theoretically expected to be capable of finding rare association rules, they actually become intractable if the minimum level of support is set low enough to find rare rules [5].

The problem of discovering rare items has recently captured the interest of the data mining community [1]. As previously explained, rare itemsets are those that only appear together in very few transactions or some very small percentage of transactions in the database [5]. Rare association rules have low support and high confidence in contrast to general association rules which are determined by high support and a high confidence level. Figure 1 illustrates how the support measure behaves in relation to the two types of rules.

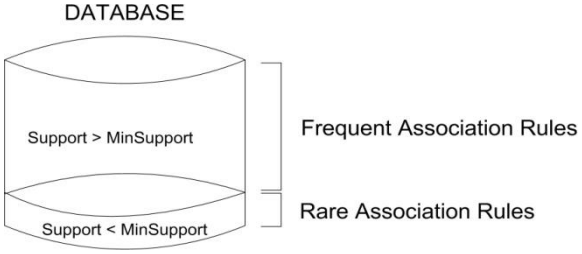

Figure 1. Rules in a database.

There are several different approaches to discover rare association rules. The simplest way is to directly apply the Apriori algorithm [2] by simply setting the minimum support threshold to a low value. However, this leads to a combinatorial explosion, which could produce a huge number of patterns, most of them frequent with only a small number of them actually rare. A different proposal, known as Apriori-Infrequent, involves the modification of the Apriori algorithm to use only the above-mentioned infrequent itemsets during rule generation. This simple change makes use of the maximum support measure, instead of the usual minimum support, to generate candidate itemsets, i.e., only items with a lower support than a given threshold are considered. Next, rules are yielded as generated by the Apriori algorithm. A totally different perspective consists of developing a new algorithm to tackle these new challenges. A first proposal is Apriori-Inverse [4], which can be seen as a more intricate variation of the traditional Apriori algorithm. It also uses the maximum support but proposes three different kinds of additions: fixed threshold, adaptive threshold and hill climbing. The main idea is that given a user-specified maximum support threshold, *MaxSup*, and a derived *MinAbsSup* value, a rule *X* is rare if $Sup(X) < MaxSup$ and $Sup(X) > MinAbsSup$. A second proposal is the Apriori-Rare algorithm [11], also known as Arima, which is another variation of the Apriori approach. Arima is actually composed of two different algorithms: a naïve one, which relies on Apriori and hence enumerates all frequent itemsets; and MRG-Exp, which limits the considerations to frequent itemsets generators only. Finally, please notice that the first two approaches (Apriori-Frequent and Apriori-Infrequent) are taken to ensure that rare items are also considered during itemset generation, although the two latter approaches (Apriori-Inverse and Apriori-Rare) try to encourage low-support items to take part in candidate rule generation by imposing structural constraints.

The algorithms aforementioned are the most important RARM proposals. Next, we will explore how these approaches can be applied over educational data in such a way that their usefulness in this research area is shown.

3 Experimentation and Results

In order to test the performance and usefulness of applying RARM to e-learning data, we have used student data gathered from the Moodle system to compare several ARM and RARM algorithms and show examples of discovered rules.

3.1 Experimentation

The experiments were performed using data from 230 students in 5 Moodle courses on computer science at the University of Córdoba. Moodle (http://moodle.org) is one of the most frequently used free Learning Content Management Systems (LCMS) and keeps detailed logs of all activities that students perform (e.g., assignments, forums and quizzes). This student usage data has been preprocessed in order to be transformed into a suitable format to be used by our data mining algorithms [10]. First, a summary table (see Table 1) has been created, which integrates the most important information about the activities and the final marks obtained by students in the courses. Notice that we have transformed all the continuous attributes into discrete attributes that can be treated

as categorical attributes. Discretization allows the numerical data to be divided into categorical classes that are easier for the instructor to understand.

Name	Description	Values
course	Identification number of the course.	C218, C94, C110, C111, C46
n_assigment	Number of assignments done.	ZERO, LOW, MEDIUM, HIGH
n_quiz	Number of quizzes taken.	ZERO, LOW, MEDIUM, HIGH
n_quiz_a	Number of quizzes passed.	ZERO, LOW, MEDIUM, HIGH
n_quiz_s	Number of quizzes failed.	ZERO, LOW, MEDIUM, HIGH
n_posts	Number of messages sent to the forum.	ZERO, LOW, MEDIUM, HIGH
n_read	Number or messages read on the forum.	ZERO, LOW, MEDIUM, HIGH
total_time_assignment	Total time spent on assignments.	ZERO, LOW, MEDIUM, HIGH
total_time_quiz	Total time spent on quizzes.	ZERO, LOW, MEDIUM, HIGH
total_time_forum	Total time spent on forum.	ZERO, LOW, MEDIUM, HIGH
mark	Final mark obtained by the student in the course.	ABSENT, FAIL, PASS, EXCELLENT

Table 1. Attributes used for each student instance

Due to the way their values are distributed, the course and mark attributes are clearly imbalanced, i.e., they have one or many values with a very low percentage of appearance:

- **Course:** From a total of 230 students, 80 took course 218 (34.78%), 66 students did course 94 (28.69%), 62 students did 110 (26.95%), 13 students took course 111 (5.65%) and 9 students took course 46 (3.91%). Thus, there are three predominant courses (C218, C94 and C110) and two minority courses (C111 and C46).

- **Mark**: From among 230 students, 116 students PASS the final exam with a normal/medium score (50.43%), 87 students FAIL the exam (38.82%), 15 students obtain an EXCELLENT or very good/high score in the exam (6.52%) and 12 students were ABSENT from the exam (5.21%). So, there are two majority marks (PASS and FAIL) and two minority marks (EXCELLENT and ABSENT).

A better view of such imbalanced value distribution for these two attributes (mark and course) can be seen graphically in Figure 2.

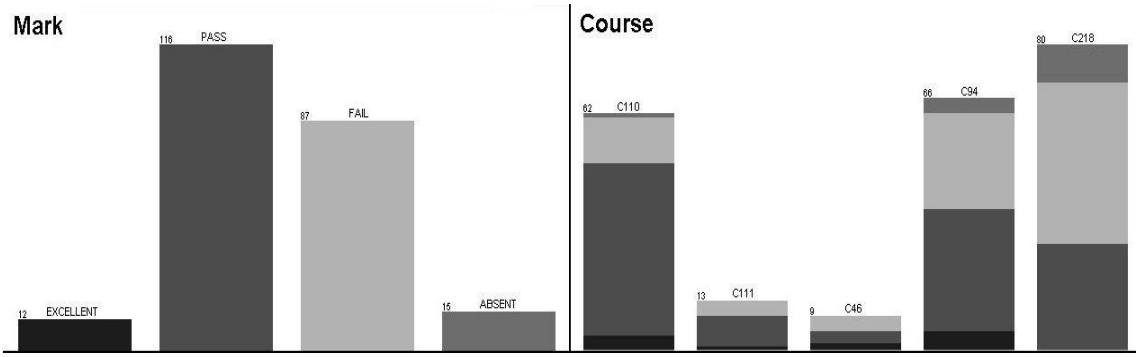

Figure 2. Value distribution for the attributes Mark and Course. The different colours on the right image correspond to the different Marks.

We performed a comparison between ARM and different RARM algorithms to discover rare class association rules [12] from the aforementioned data. A class association rule is a special subset of association rules with the consequent of the rule limited to a target class label (only one predefined item), whereas the left-hand may contain one or more attributes. It is represented as $A \rightarrow C$, where A is the antecedent (in our case, the course and activity attributes) and C is the class (in our case, the mark attribute). This type of rule is more easily understood than general association rules, since it only comprises one element in the consequent and usually represents discovered knowledge at a high level of abstraction, and so can be used directly in the decision making process [8]. In the context of EDM, class association rules can be very useful for educational purposes, since they show any existing relationships between the activities that students perform using Moodle and their final exam marks. To obtain class association rules we need to filter the resulting rules from the ARM or RARM algorithms, so we only select those rules that have a single attribute (i.e., the mark attribute) in their consequent.

We evaluated the four different Apriori proposals following the configuration parameters stated below:

- Apriori-Frequent [2], setting the minimum support threshold at a very low value (0.05);
- Apriori-Infrequent, setting the maximum support at 0.1;
- Apriori-Inverse and Apriori-Rare, using the same support threshold set at 0.1.

We also assigned the value 0.7 as the confidence threshold for all these algorithms.

3.2 Evaluation of results

Table 2 summarizes the results obtained from the four Apriori proposals, and shows the number of frequent and infrequent itemsets mined, the number of rules discovered, and their average support and confidence.

Algorithm	# Freq. Itemsets	# UnFreq. Itemsets	# Rules	Avg Support/ ± Std Deviation	Avg Confidence/ ± Std Deviation
Apriori-Frequent	11562	--	788	0.162±0.090	0.717±0.211
Apriori-Infrequent	--	1067	388	0.058±0.060	0.863±0.226
Apriori-Inverse	--	3491	46	0.056±0.070	0.883±0.120
Apriori-Rare	--	5750	44	0.050±0.080	0.885±0.108

Table 2. Comparison of ARM and RARM proposals.

Notice that the Apriori-Frequent is the only algorithm that uses frequent itemsets. Therefore, it discovers the greatest number of rules (both frequent and rare) with the highest average support but not the highest confidence. This means that the instructor needs to search manually for the rare rules. On the other hand, Apriori-Infrequent mines the smallest number of infrequent itemsets. Though it discovers a great number of rare rules, most of them are redundant. Finally, Apriori-Inverse and Apriori-Rare behave in a very similar fashion and are the best at discovering rare association rules, since they use a higher number of infrequent items than Apriori-Infrequent and discover a lower number of rare rules. A lower number of rules is easier than a higher number of rules for the instructor to use and understand. Furthermore, the standard deviation shows that both Apriori-Inverse and Apriori-Rare tend to be very close to the average, so one expects to obtain rules that will not vary much from these values.

3.3 Examples of discovered rules

Due to the imbalanced nature of the data source, different versions of the conditional support were defined. Conditional support is a well-known measure for the processing of imbalanced data using class association rules [12]. Thus, three different measures are considered to evaluate rule support, as defined in continuation:

- The traditional support of a rule $A \rightarrow C$, with A as the antecedent and C as the consequent, is defined as $Sup(A \rightarrow C) = \frac{n(A \cap C)}{N}$, where $n(A \cap C)$ is the number of instances that matches both the antecedent and consequent, and N is the total number of instances.

  However, the support of rules that contain course and mark attributes (imbalanced attributes) must be defined as follows:

  - The conditional support with respect to the mark of a class association rule $A \rightarrow Mark$, where *Mark* stands for the imbalanced attribute mark, and is defined as $SupM(A \rightarrow Mark) = \frac{n(A \cap Mark)}{n(Mark)}$, where $n(A \cap Mark)$ is the number of instances that matches both the antecedent and consequent *and n(Mark)* is the number of instances that matches the ”mark” attribute.

- The conditional support with respect to the course of a class association rule $A \cap Course \rightarrow Mark$, where *Course* stands for the imbalanced attribute course and *Mark* for the class attribute, is defined as $SupC(A \cap Course \rightarrow Mark) = \frac{n(A \cap Course \cap Mark)}{n(Course)}$, where *n(Course)* is the number of instances that matches the "course" attribute.

Next, there are some examples of rules that were mined using both the ARM and the different RARM algorithms. For each rule, we show the antecedent and the consequent constructed, as well as the support and confidence measures. Firstly, Table 3 shows some representative association rules mined using the Apriori-Frequent algorithm. A further description is detailed below.

Rule	Antecedent	Consequent	Sup	SupC/SupM	Conf
1	total_time_forum=HIGH	mark=PASS	0.24	--/0.47	0.82
2	n_posts=MEDIUM AND n_read=MEDIUM AND n_quiz_a=MEDIUM	mark=PASS	0.13	--/0.25	0.71
3	course=C110 AND n_assignment=HIGH	mark=PASS	0.14	0.52/0.27	0.89
4	total_time_quiz=LOW	mark=FAIL	0.21	--/0.55	0.78
5	n_assignment=LOW	mark=FAIL	0.23	--/0.60	0.70
6	n_quiz_a=LOW AND course=C218	mark=FAIL	0.18	0.51/0.47	0.83

Table 3. Rules extracted using the Apriori-Frequent algorithm.

As can be seen, all the rules discovered (not only the 6 rules shown in Table 3 but also the other 788 rules discovered) contain only frequent itemsets, such as mark=PASS (students who passed the exam), mark=FAIL (students who failed), course = 119 (students who took the course 119), course=218 and course=94. Secondly, we can see that these rules have low support (but not very low), a medium value in the two conditional supports and are of high confidence (but not very high). Finally, to explain the usefulness of these rules for the instructor, we are going to describe their meaning. Rule 1 shows that if students spend a lot of time in the forum (a high value) then they pass the final exam. It provides information to the instructor about how the forum has been a good activity for students with a confidence of 0.82. Rule 2 shows that if students have submitted and read messages to/from the forum, and they have passed quizzes, then they have passed the exam. The information provided is similar to the previous data but adds the quizzes as another determining factor in the final mark (as is logical). Rule 3 shows that students in course 110 who sent in many assignments then passed the final exam (rule 5 is the opposite version but for any course). So, the number of assignments is directly related to the final mark. Rule 4 and 5 show that if the total time in quizzes is low or the number of passed quizzes is low (and the course is 218),

then students obtain a bad mark. So, quizzes are also directly related to the mark and can be used to detect in time students at risk of failing the final exam.

Next, Table 4 shows some representative rare association rules obtained using the Apriori-Rare algorithm. Please notice that due to the Apriori-Inverse approach obtains almost the same set of rules, so we don't present another analysis similar to the following table. A detailed description of this rule set is presented below.

Rule	Antecedent	Consequent	Sup	SupC/SupM	Conf
1	n_quiz=HIGH AND n_quiz_a=HIGH	mark=EXCELLENT	0.045	--/0.69	0.86
2	total_time_assignment=HIGH	mark=EXCELLENT	0.045	--/0.69	0.86
3	n_posts=HIGH AND course=C46	mark=EXCELLENT	0.045	1.00/0.69	1.00
4	total_time_assignment=ZERO AND total_time_forum=ZERO AND total_time_quiz=ZERO]	mark=ABSENT	0.050	--/0.76	0.78
5	n_posts=ZERO AND n_read=ZERO	mark=ABSENT	0.050	--/0.76	0.78
6	n_quiz=ZERO AND course=C111	mark=ABSENT	0.050	0.88/0.76	1.00

Table 4. Rules extracted using the Apriori-Rare algorithm.

As can be seen, all the rules that are discovered (not only the 6 rules shown in Table 4 but also the other 44 rules discovered) contain only infrequent itemsets, such as mark=EXCELLENT (students who passed the exam with a very high score), mark=ABSENT (students who did not take the exam), course = 46 and course = 111 (students who did courses 46 and 111 respectively). We can see that these rules have a very low support, a very high confidence level (the maximum value) and also a high value for the conditional supports; indicating that they are rare/infrequent rules and their data is imbalanced with respect to the course and mark attributes. To explain the usefulness of these rules for the instructor, we are going to describe their meaning. Rule 1 shows that if students execute all the quizzes and pass them, then they obtain an excellent score in the final exam. It could be an expected rule that shows the instructor that quizzes can be used in order to predict very good student results. Rule 2 shows that if students spend a lot of time on assignments, they obtain an excellent score. This is the opposite of rule 5 in Table 3 and so it proves again that the number of assignments is directly related to the final mark. Rule 3 shows that if students in course 46 send a lot of messages to the forum, they obtain an excellent score. The instructor can use this information to detect very good students in course 46 depending on the number of messages they send to the forum. The last three rules are about students who have been absent for the exam. They show the instructor that if students do not spend time on assignments, forum participation and quizzes, then they do not take the exam. The instructor can detect this type of student in time to help him/her to take part in course activities and also do the final exam.

Finally, we have also compared the values of the evaluation measures, shown in Table 3 and Table 4. Firstly, we can see that confidence values (*Conf*) are normally higher in Table 4 (rare association rules) than in Table 3 (frequent association rules). Secondly, the support values (*Sup*) of rare association rules are much lower in Table 4 than the support of frequent association rules in Table 3. Then, we can see that relative support values (*SupC* and *SupM*) of rare association rules are higher in Table 4 than the relative support of frequent ones in Table 3. It proves that rare association rules have high confidence levels, and although they have very low values of support with respect to all the data, these support values are high with respect to imbalanced attributes (as show the relative support measure).

4 Concluding remarks and future work

In this paper we have explored the use of RARM over educational data gathered from the Moodle system installed at the University of Córdoba. The use of this approach has shown to be an interesting research line in the context of EDM, since most real-world data are usually imbalanced. Rare-association rules are more difficult to mine using traditional data mining algorithms, since they do not usually consider class-imbalance and tend to be overwhelmed by the major class, leaving the minor class to be ignored. In fact, we have shown that the regular Apriori algorithm [2] (known as Apriori-Frequent) discovers a huge number of rules with frequent items. Hence we explored how some specific algorithms, such as Apriori-Inverse and Apriori-Rare, are better at discovering rare-association rules than other non-specific algorithms, such as Apriori-Frequent and Apriori-Infrequent. In fact, the set of rules discovered by Apriori-Rare are included into the set of rules discovered by Apriori-Inverse but they are included neither into the set of rules discovered by Apriori-Infrequent nor Apriori.

Finally, we have shown how the rules discovered by RARM algorithms can help the instructor to detect infrequent student behavior/activities in an e-learning environment such as Moodle. In fact, we have evaluated the relation/influence between the on-line activities and the final mark obtained by the students.

In the future, we would like to develop a new algorithm specifically to discover RARM using evolutionary algorithms, and to compare its performance and usefulness in e-learning data versus the previous algorithms. We also plan to explore the use of other different rule evaluation measures for rare association rule mining.

Acknowledgment

This research is supported by projects of the Regional Government of Andalucia and the Ministry of Science and Technology, P08-TIC-3720 and TIN2008-06681-C06-03 respectively, and FEDER funds.

References

[1] Adda, M., Wu, L, Feng, Y. Rare itemset mining. In *Sixth Conference on Machine Learning and Applications*. 2007, Cincinnati, Ohio, USA, pp-73-80.

[2] Agrawal, R., Imielinski, T., Swami, A. Mining association rules between sets of items in large databases. *Proceedings of the ACM SIGMOD International Conference on Management of Data*, Washington, DC, 1993, pp. 207–216.

[3] Kotsiantis, S., Kanellopoulus, D. Association rules mining: A recent overview. *International Transactions on Computer Science and Engineering Journal*, 2006. 32, 1, pp. 71-82.

[4] Koh, Y., Rountree, N. Finding sporadic rules using Apriori-Inverse. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining.* 2005. Berlin, pp. 97-106.

[5] Koh, Y., Rountree, N. Rare association rule mining and knowledge discovery. 2009. *Information Science Reference.*

[6] Liu, B., Hsu, W., Ma, Y. Mining association rules with multiple minimum supports. In *Proceedings of fifth ACMSIGKDD International Conference on Knowledge Discovery and Data Mining*. 1999, New York, USA, pp. 337-341.

[7] Merceron, A., and Yacef, K, Interestingness Measures for Association Rules in Educational Data. In *International Conference on Educational Data Mining*. 2008. Montreal, Canada, pp. 57-66.

[8] Romero, C., Ventura, S., Bra, P. D. Knowledge discovery with genetic programming for providing feedback to courseware author. *User Modeling and User-adapted Interaction: The Journal of Personalization Research*, 2004, 14 (5), pp. 425–464.

[9] Romero, C., Ventura, S., Espejo, P., Hervás, C. Data Mining Algorithms to Classify Students. In *International Conference on Educational Data Mining*. 2008, Montreal, Canada, pp. 8-17.

[10] Romero, C., Ventura, S., Salcines, E. Data mining in course management systems: Moodle case study and tutorial. *Computer & Education*, 2008, 51(1), pp. 368-384.

[11] Szathmary, L., Napoli, A., Valtchev, P. Towards rare itmeset mining. In *International Conference on Tools with Artificial Intelligence*, Washington, DC. 2007, pp. 305-312.

[12] Zhang, H., Zhao, Y., Cao, L., Zhang, C., Bohoscheid, H. Rare class association rule mining with multiple imbalanced attributes. *Rare Association Rule Mining and Knowledge Discovery*. 2009. Information Science Reference.

[13] Zaïane, O., Building a Recommender Agent for e-Learning Systems. *Proceedings of the International Conference in Education.* 2002, New Zealand, pp. 55-59.