An evolutionary algorithm for the discovery of rare class association rules in learning management systems

Article *in* Applied Intelligence · April 2014

DOI: 10.1007/s10489-014-0603-4

CITATIONS
37

READS
566

4 authors:

José María Luna
University of Cordoba (Spain)
82 PUBLICATIONS **1,399** CITATIONS
SEE PROFILE

Cristóbal Romero
University of Cordoba (Spain)
127 PUBLICATIONS **9,310** CITATIONS
SEE PROFILE

José Raúl Romero
University of Cordoba (Spain)
78 PUBLICATIONS **1,269** CITATIONS
SEE PROFILE

Sebastian Ventura
University of Cordoba (Spain)
336 PUBLICATIONS **11,960** CITATIONS
SEE PROFILE

Noname manuscript No.
(will be inserted by the editor)

An Evolutionary Algorithm for the Discovery of Rare Class Association Rules in Learning Management Systems

J.M. Luna · C. Romero · J.R. Romero · S. Ventura

Received: / Revised: / Accepted:

Abstract Association rule mining, an important data mining technique, has been widely focused on the extraction of frequent patterns. Nevertheless, in some application domains it is interesting to discover patterns that do not frequently occur, even when they are strongly related. More specifically, this type of relation can be very appropriate in e-learning domains due to its intrinsic imbalanced nature. In these domains, the aim is to discover a small but interesting and useful set of rules that could barely be extracted by traditional algorithms founded in exhaustive search-based techniques. In this paper, we propose an evolutionary algorithm for mining rare class association rules when gathering student usage data from a Moodle system. We analyse how the use of different parameters of the algorithm determine the rule characteristics, and provides some illustrative examples of them to show their interpretability and usefulness in e-learning environments. We also compare our approach to other existing algorithms for mining both rare and frequent association rules. Finally, an analysis of the rules mined is presented, which allows information about students' unusual behaviour regarding the achievement of bad or good marks to be discovered.

Keywords Rare Association Rules · Grammar Guided Genetic Programming · Evolutionary Computation · Educational Data Mining

1 Introduction

E-learning systems provide a great variety of information that is very important for analysing students' behaviour and which could create a gold mine of educational data [28,30]. However, due to the large amounts of data that these systems can generate on a daily basis, it is very difficult to do a manual inspection. To overcome this problem, some data mining (DM) techniques were used in recent years, association rule mining (ARM) being one of the most well studied in this sense [2,37].

ARM was conceived as an unsupervised learning task for finding close relationships among patterns in large databases. An association rule is an if-then statement concerning attribute-value pairs [10]. Most current approaches and tools for the discovery of association rules are based on models that mine frequent patterns [16,26,34]. ARM proposals were originally designed for market basket analysis, obtaining patterns that exceed a minimum frequency and then, using the patterns to form reliable association rules. However, there are situations where it is interesting to determine abnormal or unusual behaviour in data, mining

J.M. Luna · C. Romero · J.R. Romero · S. Ventura
Dept. of Computer Science and Numerical Analysis, University of Cordoba, Rabanales Campus, 14071 Cordoba, Spain. Dr. Ventura also belongs to Department of Computer Science, King Abdulaziz University, Saudi Arabia Kingdom.
Tel.: +34957212218
Fax: +34957218630
E-mail: {jmluna,cromero, jrromero, sventura}@uco.es

relationships that do not follow the trend of the others [1]. Banking fraud detection [31], or the recognition of patients who suffer a particular rare disease [21] are some examples where rare association rules mining (RARM) plays an important role.

Even so, not enough attention has been paid to the extraction of rare and reliable association rules [13, 17], especially in educational tasks where infrequent associations can be of great interest [25]. For instance, infrequent associations might allow the instructor to verify a set of rules concerning certain unusual learning problems, for instance dealing with students with special needs. Thus, this information could help the instructor to discover a minority of students who may need specific support in their learning process. The idea is to find infrequent relations that show unusual behaviours in the form of **IF** *a student spends high time doing task* **THEN** *the student fails or does not obtain a good mark in the course*. This could be a rare behaviour since students that spend high time doing tasks probably will obtain an excellent mark. In this sense, those that do not achieve an excellent mark should be analysed to determine why they do not achieve the aim and which specific needs they require. In addition, it should be noticed that infrequent attributes are usually more interesting than those that appear frequently, e.g. students who drop out or absent of a course/subject are usually more infrequent than those students who fail or fare well.

In early studies, the problem of finding infrequent patterns was originally addressed by using algorithms for mining frequent patterns. Then, specific algorithms for mining infrequent association rules were proposed [12]. A major drawback of these proposals is that they perform an exhaustive search among the dataset patterns, so their execution over huge datasets with a large number of attributes is computationally hard. Furthermore, most RARM algorithms discover a huge number of rules barely understandable by the instructor, who actually only requires a small set of interesting and reliable rules for easier detection of those learning needs. Different evolutionary algorithms and especially genetic programming (GP) [4] algorithms have been proposed to extract frequent association rules [16]. These algorithms can be used where optimization is needed, i.e., to find the best solution to a problem where there are many solutions. Moreover, these kinds of algorithms perform well, e.g., in terms of scalability. In GP, individuals are represented with variable-length hierarchical structures usually in a tree-form, where the shape, size and structural complexity of the solution are not constrained a priori. In some application domains it is sometimes possible to know the syntax form of the desired solution, in which case it is useful to constrain the GP process by searching for solutions with different syntax forms. Methods to implement such restrictions consist in using some form of constrained-syntax to enforce syntactic and semantic constraints to GP trees, which is known as grammar guided genetic programming (G3P) [10]. G3P is an extension of GP where each individual is a derivation tree that generates and represents a solution using the language defined by the grammar.

Motivated by these problems, as well as the great interest of RARM in educational tasks and the promising performance of using evolutionary methodologies in ARM, we propose the application of a new evolutionary algorithm for mining rare association rules in e-learning datasets. Moreover, the mining of class association rules in these domains requires a previous knowledge of the syntax form of the solutions, so the use of G3P in this context is fully justified. For the sake of analysing the effectiveness of this proposal, we compare it to other exhaustive search ARM and RARM algorithms. We make use of a real student usage dataset in order to discover information about infrequent students' behaviour concerning their resulting course marks.

This paper is organised as follows. Section 2 introduces some related works. Section 3 describes the proposed algorithm. Results obtained from the use of a Moodle dataset are presented and discussed in Section 4. Finally, some concluding remarks are provided in Section 5.

2 Background

ARM is one of the most popular and well-known DM techniques for extracting interesting and close relationships between patterns in transaction databases or other data repositories [2]. An association rule is an implication of the form $A \rightarrow C$, where A and C are disjointed item-sets, i.e. sets with no items in common. A and C being the antecedent and consequent, respectively. The intuitive meaning of such a rule is that when A appears, C also tends to appear. A special type of association rule is known as class association rule (CAR) [38], which describes an implicative co-occurring relationship between a set

of items and a predefined class in its consequent. While ARM discovers all frequent and reliable rules without any predetermined target, in the CAR mining process there is one and only one predetermined target, i.e. the class. CAR is considered as a type of target-constrained association rule.

The process of evaluating association rules is a major issue because of the large number of them that could be extracted from a specific problem. Some objective measures for evaluating the interest of these rules have been proposed by different researchers [33]. Two of the most important and widely used measures for evaluating association rules are support and confidence. The support measure is defined in Equation 1 as the proportion of the number of transactions T including the antecedent A and the consequent C in a dataset D. The confidence measure is detailed in Equation 2 as the probability of finding C in transactions under the condition that these transactions also contain A, i.e. the proportion of the number of transactions which include A and C among all the transactions that include A.

$$support(A \rightarrow C) = \frac{|\{A \cup C \subseteq T, T \in D\}|}{|D|} \tag{1}$$

$$confidence(A \rightarrow C) = \frac{support(A \rightarrow C)}{support(A)} \tag{2}$$

Despite the fact that most proposals in ARM are based on a support–confidence framework, these measures are not sufficient to select interesting associations between patterns [3] over some application domains. Sometimes, it is required that the occurrence of the antecedent does not imply an increment in the occurrence of the consequent. Lift (see Equation 3) was defined to solve this problem, establishing how many times the antecedent and the consequent occur together more often than would be expected if they were statistically independent. An association rule is defined as of interest if its confidence value is higher than the value of the support of its consequent. On the other hand, if the confidence of the rule is equal to the support of its consequent, then both antecedent and consequent are independent.

$$lift(A \rightarrow C) = \frac{confidence(A \rightarrow C)}{support(C)} \tag{3}$$

The ARM process is usually divided into two steps. The first one discovers those patterns whose occurrences exceed a predefined support threshold. The second step generates reliable association rules from those frequent patterns previously extracted. Additionally, some methodologies have been proposed to reduce the number of rules to be discovered, mining the top k association rules [5]. Nowadays, the problem of mining frequent patterns has been studied in depth and many algorithms have already been proposed for this purpose [2, 10, 16]. Nevertheless, there is also an increasing interest in the extraction of unusual patterns, which are extremely important in many diverse domains (banking, health, education, etc.) and therefore the discovery of this type of rules has recently captured the interest of the DM community [1].

In the ARM field, the rare item problem is essentially considered as a data imbalanced problem. The imbalanced learning problem is concerned with the performance of learning algorithms in the presence of under-represented data, that is, the number of instances in one attribute is much smaller than the number of instances in other attributes [7]. Generally, the mining of association rules is related to the discovery of highly frequent and reliable relationships between set of items. Nevertheless, when we deal with the discovery of infrequent relationships, attributes should be under-represented in data. The number of imbalanced attributes is even higher in the educational data mining field, where students with special needs could represent an under-represented group within data.

As mentioned above, infrequent patterns are those that rarely appear in the database [13]. Rare association rules have low support and high confidence in contrast to frequent association rules, which are determined by their high support and confidence level. The process of mining rare patterns was originally addressed by using algorithms for mining frequent patterns. Though these algorithms are theoretically expected to be capable of finding rare association rules, they actually become intractable when using large datasets since they require a low support threshold value to widen the search space constructed by the exhaustive search methods, and therefore, requiring a significantly higher computational time. The use of Apriori [2], which is one of the most well-known ARM algorithms, appears to be the simplest way to discover association rules. However, it should be noted that low support values imply a combinatorial

explosion and an increment of the runtime [8]. A different proposal, known as Apriori-Infrequent [23], involves the modification of the Apriori algorithm to use only infrequent patterns during the rule generation. This simple alteration provides the use of a maximum support threshold, instead of the usual minimum support, to generate candidate patterns. Similarly to Apriori, reliable rules are obtained by using the candidate patterns previously discovered.

A different perspective was considered with the Apriori-Inverse algorithm [12], which is considered as a variation of the traditional Apriori algorithm. During its execution, Apriori-Inverse keeps those patterns with a support value greater than a minimum threshold value but less than a maximum value. Finally, a set of association rules is obtained by using a minimum confidence threshold.

Apriori-Rare, also known as Arima, is another important RARM proposal [32]. This algorithm works by obtaining the minimal rare patterns, i.e. those infrequent patterns that comprise frequent subsets. In subsequent steps, Arima discovers all the supersets that can be obtained from the minimal rare patterns and whose support is distinct from zero. Finally, a set of rare association rules is obtained by using the rare patterns previously mined.

To sum up, it should be noted that the first two approaches (Apriori and Apriori-Infrequent) are taken to ensure that rare patterns are also considered during pattern mining. In contrast, the two latter approaches (Apriori-Inverse and Apriori-Rare) try to encourage low-support patterns to take part in candidate rule generation by imposing structural constraints.

Finally, there are a large number of works that have extensively applied ARM to educational domains [18], finding close relationships between patterns and applying them to the following tasks: automatic guidance of the learner's activities and the intelligent generation and recommendation of learning materials [14]; identification of attributes characterising patterns of performance disparity between various groups of students [20]; discovery of interesting relationships from student's usage information in order to provide feedback to the course author [27]; enquiry of relationships between each pattern of a learner's behaviour [35]; discovery of students' common mistakes [19]; guidance of the search for the best fitting transfer model of student learning [6]; optimisation of the content of an e-learning portal by determining the content of most interest to the user [24]; and extraction of patterns to help educators and web masters evaluate and interpret on-line course activities [36]. All these previous studies were focused on the frequent pattern mining problem. However, educational real world datasets comprise infrequent data, which can be very useful for instructors to discover students that may need extra help in their learning process.

To our knowledge, there is only one previous work that deals with educational problems from a RARM perspective [25]. This is an initial work for comparing different well-known RARM algorithms using data from diverse learning management systems. Unlike this previous work, this paper presents a new evolutionary approach for discovering infrequent CARs. This approach is compared to ARM and RARM algorithms by using real student usage data gathered from Moodle learning management systems [25]. CARs are easier to be understood than traditional association rules. Remember that CARs only comprise one attribute in their consequent, so in this work, they show the relationships between the activities that students perform by using Moodle and their final exam marks.

3 Proposed Evolutionary Algorithm

Evolutionary algorithms are based on the Darwin's theory of evolution, where each individual codifies a solution (a single rule or a rule set) that evolves to a new individual by means of genetic operators (mutation and crossover). A technique based on evolutionary algorithms is GP [4], its main feature being the individual representation. As mentioned above, individuals in GP are encoded as variable-length hierarchical structures, usually in a tree-form, where the shape, size and structural complexity of the solutions are not constrained a priori. Sometimes, it is useful to restrict the GP process by searching for solutions that present different syntax forms. Therefore, some researchers have focused their studies on methods that implement such restrictions by using some form of constrained syntax, or by building solutions conformant to a predefined grammar [22]. The use of grammars in GP is known as grammar guided genetic programming (G3P) [10]. In G3P, each individual is represented as a tree that generates and represents a solution by means of the language defined by the grammar, which allows of defining some

$G = (\Sigma_N, \Sigma_T, P, S)$ with:
S = {Rule}
Σ_N = {Rule, Antecedent, Consequent, Comparison}
Σ_T = {'AND', '=', 'name', 'value'}
P = {Rule = Antecedent, Consequent ;
Antecedent = Comparison | 'AND', Comparison, Antecedent ;
Consequent = Comparison ;
Comparison = '=', 'name', 'value' ;}

Fig. 1 Context-free grammar for encoding the rules

syntax constraints, restricting the search space and obtaining expressive solutions in different attribute domains.

Our algorithm only considers individuals conformant to the context-free grammar G, defined as a four-tuple $(\Sigma_N, \Sigma_T, P, S)$ where $\Sigma_N \cap \Sigma_T = \emptyset$, Σ_N is the alphabet of non-terminal symbols, Σ_T is the alphabet of terminal symbols or tokens, P is the set of production rules and S stands for the start symbol. Productions have the format $A \rightarrow \alpha$ where $A \in \Sigma_N$, and $\alpha \in \{\Sigma_T \cup \Sigma_N\}^*$. So, as defined in Figure 1, in order to obtain individuals a derivation process is carried out starting from the start symbol of the grammar. This grammar generates traditional IF-THEN rules with categorical values. Nevertheless, the grammar could easily be changed to deal with other different types of condition. In our problem, we have preferred to use categorical values due to the intrinsic nature of data and its readability and comprehensibility for instructors. Besides, it should be noted that a shortcoming when using grammars is the possibility of obtaining excessively deep trees. To avoid this issue, the number of production rules to be applied through the derivation process is predefined. Therefore, it is not possible to obtain individuals having a size greater than a maximum in order to reduce bloating.

In the proposed approach, each individual is determined by its genotype, which denotes a derivation syntax tree, and its phenotype, which represents the entire rare association rule comprising an antecedent and a consequent (see Figure 2).

Once each individual is built conformant to the grammar defined above, an evaluation process is performed to calculate its fitness value. As already mentioned in Section 2, two of the most important and widely used measures are support and confidence. In our proposal, the fitness function (see Equation 4) is defined as the support measure, previously described in Equation 1. Similarly to most existing RARM proposals, a maximum support threshold is required. However, the main difference is the use of a minimum

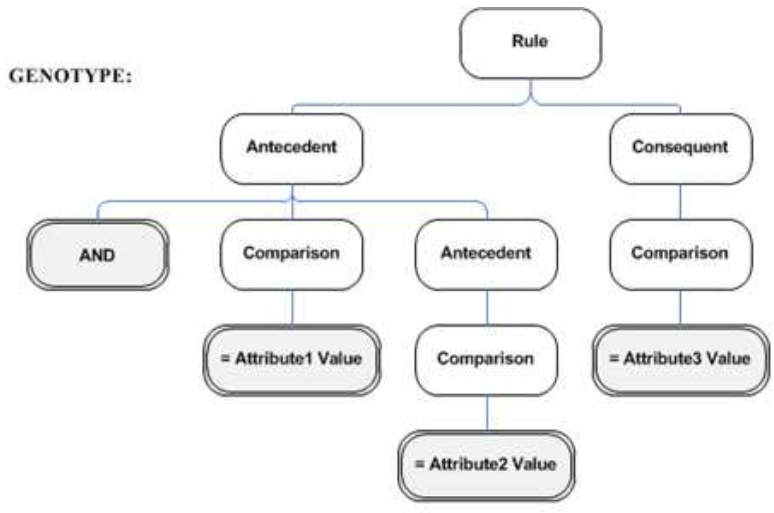

Fig. 2 Genotype and phenotype of a sample individual

support value, which discriminates against rules that have an extremely low frequency. This minimum support threshold is also very interesting in education, since instructors search for groups of students that somehow require extra help in their learning, not for those specific or particular cases that could be obtained if the support was extremely low, e.g. by counting only one student.

$$fitness = support(A \rightarrow C) = \frac{|\{A \cup C \subseteq T, T \in D\}|}{|D|} \tag{4}$$

The proposed algorithm (see Figure 3) follows a generational schema. This proposal uses an elite population to maintain the best individuals during the evolutionary process. In each generation, this pool is updated with those individuals that exceed at least the following quality criteria: (1) the fitness function value must be greater than zero, and (2) the confidence must be greater than the minimum confidence threshold. Next, the execution of two genetic operators, i.e. crossover and mutation, allows new and diverse individuals to be obtained in every generation. The algorithm uses the well-known tournament selector to obtain parents to be crossed and mutated. After selecting two individuals that act as parents, the crossover operator swaps the highest support condition within one individual for the lowest support condition within the other parent. Alternatively, the mutation operator selects the highest support condition within one parent only, and changes it with a new random condition. Both genetic operators allow of obtaining

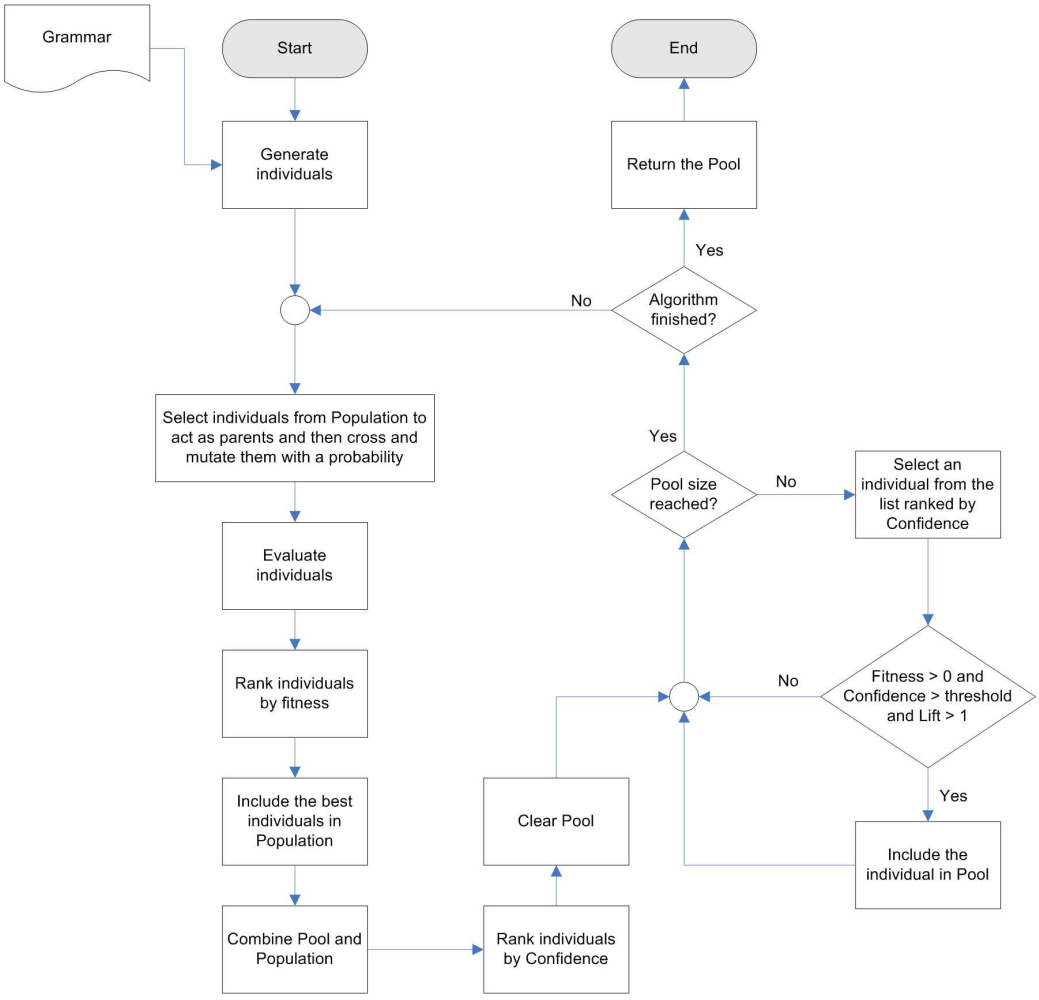

Fig. 3 Overview of the G3P algorithm for mining RARs

Listing 1 Proposed algorithm for mining rare association rules in educational environments

Input: $max\_generations, num\_individuals, max\_Pool\_size, confidenceThreshold$
Output: $Pool$
1: $Pop \leftarrow$generateIndividuals($num\_individuals$)
2: $Pool \leftarrow \emptyset$
3: $Aux \leftarrow \emptyset$
4: $num\_generations \leftarrow 0$
5: **while** $num\_generations < max\_generations$ **do**
6: $Parents \leftarrow$ selectParents (Pop)
7: $Aux \leftarrow$ geneticOperators ($Parents$)
8: Evaluate (Aux)
9: $Pop \leftarrow$ rankIndividualFitness ($Aux \cup Pop$)
10: $Aux \leftarrow$ rankIndividualConfidence ($Pop \cup Pool$)
11: $Pool \leftarrow$ updatePool (Aux, max_Pool_size, $confidenceTreshold$)
12: $Aux \leftarrow \emptyset$
13: $num\_generations++$
14: **end while**
15: **return** $Pool$

procedure rankIndividualConfidence
Input: Aux, max_Pool_size, $confidenceTreshold$
Output: Aux'
1: $Aux' \leftarrow \emptyset$
2: $i \leftarrow 0$
3: **for all** $individuals \in Aux$ **do**
4: **if** $individual_i^{Aux}$ is not in A' **then**
5: **if** getFitness($individual_i^{Aux}$) > 0 **then**
6: **if** getConfidence($individual_i^{Aux}$) $> confidenceTreshold$ **then**
7: **if** getLift($individual_i^{Aux}$) > 1 **then**
8: $Aux' \leftarrow (Aux' \cup individual_i^{Aux})$
9: **end if**
10: **end if**
11: **end if**
12: **end if**
13: $i++$
14: **if** getSize(Aux') $= max\_Pool\_size$ **then**
15: **return** Aux'
16: **end if**
17: **end for**
18: **return** Aux' **end procedure**

new individuals having the same size to the parents since they only swap sub-trees starting with the non-terminal symbol comparison, so the problem of bloating could not appear.

The algorithm proposed (see Figure 3 for a general sketch and Listing 1 for the pseudocode) for the extraction of rare association rules follows a generational schema, which starts by generating a set of new individuals conformant to the specified grammar (see line 1 in the pseudocode). Several steps are performed for each generation: (1) a set of individuals are selected to act as parents (line 6) from the general population and genetic operators are applied over them immediately afterwards with a certain probability (see line 7). Next, (2) these new individuals are evaluated (line 8). In the following step, (3) a new population is obtained (see line 9) by ranking the individuals by fitness from the new set obtained (line 7) and the population obtained in the previous generation. This new updated population comprises the individuals with the highest fitness value. Following, (4) an auxiliary population is formed by ranking individuals (see line 10) by their confidence value from the new population and the pool set. This auxiliary population is used to get the individuals that will comprise the new pool set (see procedure $rankIndividualConfidence$). Only those individuals having a fitness value greater than zero, a confidence value greater than the minimum confidence threshold, and a lift value greater than unity are considered prompting the discovery of infrequent, reliable and interesting association rules.

4 Experimental study

In order to test the performance and usefulness of our evolutionary algorithm in e-learning domains, we have used student usage data gathered from the Moodle system. Next, we compare the results obtained with other ARM and RARM algorithms, and show some examples of the rules discovered.

4.1 Description of the data

The experiments were performed using data collected from 230 students on 5 Moodle courses on computer science at the University of Cordoba. Moodle keeps detailed logs of all the activities performed by these students (e.g. assignments, forums [11], or quizzes). All this information was properly preprocessed, so it was transformed into a suitable format to carry out data mining [29]. This preprocessing includes the transformation of every continuous attribute into a discrete domain, so they can be treated as categorical attributes. Discretization allows the numerical data to be divided into categorical classes, making it easier for the instructor to understand. This discretization step was carried out by experts in the domain, considering educator from the degree of computer sciend at the University of Cordoba. The following list of attributes summarises the most important information about the activities monitored by Moodle from students during the life of the course:

- course: identifies the course. Its available values are: C218, C94, C110, C111 and C46.
- n_assigment: determines the number of assignments done. Its available values are: ZERO, LOW, MEDIUM, HIGH.
- n_quiz: establishes the number of quizzes taken. Its available values are: ZERO, LOW, MEDIUM, HIGH.
- n_quiz_pass: determines the number of quizzes passed. Its available values are: ZERO, LOW, MEDIUM, HIGH.
- n_quiz_fail: identifies the number of quizzes failed. Its available values are: ZERO, LOW, MEDIUM, HIGH.
- n_posts: determines the number of messages sent to the forum. Its available values are: ZERO, LOW, MEDIUM, HIGH.
- n_read: identifies the number of messages read on the forum. Its available values are: ZERO, LOW, MEDIUM, HIGH.
- total_time_assignment: establishes the total time spent on assignments. Its available values are: ZERO, LOW, MEDIUM, HIGH.
- total_time_quiz: determines the total time spent on quizzes. Its available values are: ZERO, LOW, MEDIUM, HIGH.
- total_time_forum: determines the total time spent on the forum. Its available values are: ZERO, LOW, MEDIUM, HIGH.
- mark: establishes the final mark obtained by the student on the course. Its available values are: FAIL, ABSENT, PASS, EXCELLENT.

It is worth mentioning that the values of two of these attributes (course and mark) are clearly distributed in an imbalanced way. So, as shown in Figure 4, from 230 students, 116 students obtained a PASS in the final exam with a normal/medium score, 87 students obtained a FAIL in the exam, 15 students obtained an EXCELLENT or a very good/high score in the exam and 12 students were ABSENT from the exam. Thus, there are two predominant marks (PASS and FAIL) and two minority marks (EXCELLENT and ABSENT).

On the other hand, concerning the course attribute (see Figure 5), from a total of 230 students, 80 took course 218, 66 students did course 94, 62 students did course 110, 13 students took course 111 and 9 students took course 46. Thus, there are three predominant courses (C218, C94 and C110) and two minority courses (C111 and C46).

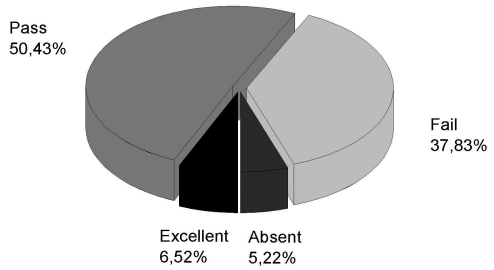

Fig. 4 Value distribution for the *mark* attribute

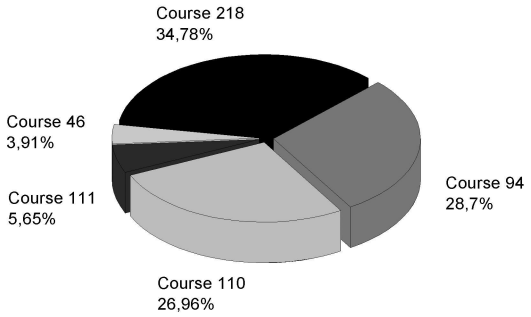

Fig. 5 Value distribution for the *course* attribute

4.2 Comparison and analysis of results

In this section, we compare our approach to five other well-known ARM and RARM algorithms: Apriori-Frequent, setting the minimum support threshold at a very low value (0.05); Apriori-Infrequent, setting the maximum support at 0.1. Furthermore, in order to obtain the optimal parameters that allow us to obtain the best results using the evolutionary proposal, a series of experiments were carried out. The best results were obtained with maximum and minimum support thresholds set to 0.1 and 0.01, respectively. Similarly, Apriori-Inverse and Apriori-Rare algorithms were tuning using the same values, i.e., 0.1 and 0.01 as maximum and minimum support threshold values, respectively. The same value for the confidence threshold value, 0.7, is set for all algorithms. Additionally, it should be noted that the G3P-Rare algorithm allows to obtain the N most reliable rules, those having the highest confidence values. The value of this parameter could be adapted to each specific problem or domain. For example, in specific educational problem, it is desired to provide a low number of reliable rules to the instructor, so only the best 30 rules are obtained by the G3P-Rare algorithm.

Table 1 summarises the results obtained in the experimentation, showing the number of frequent (#Frequent) and infrequent (#Infrequent) patterns mined, the number of rules discovered (#Rules), the runtime in seconds, their average support (Avg Support) and confidence (Avg Confidence) values. Table also shows the standard deviation for the set of rules discovered, considering the standard deviation for support and confidence quality measures. Analysing the aforementioned table, it is shown that the Apriori-Frequent algorithm discovers the greatest number of rules (both frequent and rare) with the highest average support, but not the highest confidence. This means that instructors must manually discriminate between the rare rules and the others. On the other hand, Apriori-Infrequent mines the smallest number of infrequent item-sets, and whilst it discovers a great number of rare rules, most of them are redundant.

Table 1 Comparison of ARM and RARM proposals

Algorithm	**#Frequent**	**#Infrequent**	**#Rules**
Apriori-Frequent	11562	-	788
Apriori-Infrequent	-	1067	388
Apriori-Inverse	-	3491	46
Apriori-Rare	-	5750	44
G3P-Rare	-	-	30

Algorirthm	**Runtime (seconds)**	**Avg Support ± Std Deviation**	**Avg Confidence ± Std Deviation**
Apriori-Frequent	52	0.162 ± 0.090	0.717 ± 0.211
Apriori-Infrequent	24	0.058 ± 0.060	0.863 ± 0.226
Apriori-Inverse	3	0.056 ± 0.070	0.883 ± 0.120
Apriori-Rare	2	0.050 ± 0.080	0.885 ± 0.108
G3P-Rare	2	0.031 ± 0.011	1.000 ± 0.000

Finally, Apriori-Inverse, Apriori-Rare, and our proposal behave in very similar fashions, and are the best at discovering rare association rules. Both Apriori-Inverse and Apriori-Rare use a higher number of infrequent items than Apriori-Infrequent and discover a slightly smaller number of rare rules.

Focusing on our approach, it does not mine any frequent or infrequent items since it discovers directly rules through the use of a grammar without requiring a previous step for the mining process. Furthermore, our approach allows the discovery of the N most reliable association rules within a range of support values. This freedom to be adapted to each specific domain makes the G3P-Rare proposal an interesting algorithm, especially in educational domains. For instance, our proposal discover the best 30 rules having a support in the range [0.01, 0.1], which is easier for the instructor to understand.

Finally, a runtime study is carried to determine the performance of the proposed approach. The results for the data under study determine that Apriori-Rare and G3P-Rare behave similarly in runtime. On the contrary, there are huge differences with regard to the execution time of Apriori-Frequent and Apriori-Infrequent. It demonstrate that the performance of the model used to discover rare class association rules in Learning Management Systems is really good. It should be noted that, since the final aim of this work is to solve a specific problem, that is, mining rare and interesting relations between attributes in Learning Management Systems, no additional datasets are considered in this study.

4.3 Examples of discovered rules

Next, we illustrate how the information is provided to the instructor after the mining task execution. Some rules discovered by RARM algorithms are shown and described. This analysis allows a demonstration of their usefulness in making decisions about how to detect in time successful and failed students starting on their activities in the Moodle environment. For every rule, the antecedent and the consequent, as well as their support and confidence values, and two different versions of the conditional support are shown. Conditional support is a well-known measure for the processing of imbalanced data using class association rules [36]:

The conditional support with respect to the mark of a class association rule $A \rightarrow Mark$, where Mark stands for the imbalanced attribute mark, is defined as shown in Equation 5. Notice that $n(A \cap Mark)$ is the number of instances that matches both the antecedent and consequent over the number of instances that matches the mark attribute. This quality measure has been also used as a generality measure in subgroup discovery tasks [9,15]. This metric quantifies the quality of the subgroups according to the patterns covered, and it is known as support on the basis of examples of the class.

$$supportM(A \rightarrow Mark) = \frac{n(A \cap Mark)}{n(Mark)} \tag{5}$$

On the contrary, the conditional support with respect to the course of a class association rule, where Course stands for the course imbalanced attribute and Mark for this class attribute, is defined as depicted in Equation 6. Notice that $n(Course)$ is the number of instances that matches the course attribute.

$$supportC(A \cap Course \rightarrow Mark) = \frac{n(A \cap Course \cap Mark)}{n(Course)} \quad (6)$$

Table 2 shows some representative association rules mined using the Apriori-Frequent algorithm. All the rules discovered (not only the 5 rules shown but also the 788 rules mined) only contain frequent item-sets, such as mark=PASS (students who passed the exam), mark=FAIL (students who failed), course=110 (students who took course 110), course=218 and course=94. Note that these rules have a high support value, a medium value in conditional support and a not very high confidence.

Next, we explain how these rules should be interpreted to illustrate their usefulness to the instructor. Rule 1 shows that if students spend a lot of time in the forum (a HIGH value), then they will pass the final exam. It provides information to the instructor about how beneficial the forum is for students with a confidence of 0.82. Rule 2 shows that students present in course 110 who submitted many assignments passed the final exam (rule 4 is the directly opposite version but for any course). So, the number of assignments is directly related to the final mark. Rules 3 and 5 show that, if the total time in quizzes is low or the number of passed quizzes is low (only for course 218), a failed mark should be expected.

Table 3 enumerates some representative rare association rules discovered using the Apriori-Rare and Apriori-Inverse approaches. Notice that these algorithms obtain almost the same set of rules, with very low support but high conditional support and confidence. Analysing the results, all the rules discovered (not only the 6 rules shown above but also the entire set of rules) only contain infrequent item-sets, such as mark=EXCELLENT (students who passed the exam with an outstanding score), mark=ABSENT (students who did not take the exam), course=46 and course=111 (students who attended courses 46 and 111 respectively). This type of rule provides information about rare or exceptional patterns that can also be useful in education for providing help to those students.

A more in-depth explanation about these rules is now discussed. Rule 1 shows that if students complete all the quizzes and they pass these quizzes, then an excellent score is expected in their final exam. This rule confirms that quizzes can be useful in predicting 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 4 in Table 2,

Table 2 Rules extracted by executing the Apriori-Frequent algorithm

Rule	Antecedent	Consequent	Support	SupportC/ SupportM	Confidence
1	total_time_forum=HIGH	mark=PASS	0.24	–/0.47	0.82
2	course=C110 AND n_assignment=HIGH	mark=PASS	0.14	0.52/0.27	0.89
3	total_time_quiz=LOW	mark=FAIL	0.21	–/0.55	0.78
4	n_assignment=LOW	mark=FAIL	0.23	–/0.60	0.70
5	n_quiz_pass=LOW AND course=C218	mark=FAIL	0.18	0.51/0.47	0.83

Table 3 Rules extracted by executing the Apriori-Rare algorithm

Rule	Antecedent	Consequent	Support	SupportC/ SupportM	Confidence
1	n_quiz=HIGH AND n_quiz_pass=HIGH	mark=EXCELLENT	0.045	–/0.69	0.86
2	total_time_assignment=HIGH AND	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 AND	mark=ABSENT	0.050	–/0.76	0.78
5	n_posts=ZERO AND n_read=ZERO AND	mark=ABSENT	0.050	–/0.76	0.78
6	n_quiz=ZERO AND course=C111 AND	mark=ABSENT	0.050	0.88/0.76	1.00

which proves again that the number of assignments submitted 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 their active participation on the forum. The last three rules (4 – 6) 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 will not take the exam. Note that the support, the confidence and the conditional supports provide additional valuable information to the instructor, whose main objective is to detect in time those students that really need encouragement. Finally, Table 4 shows some examples of representative rare association rules obtained using the RARM algorithm proposed in this paper.

Focusing on Table 4, the algorithm proposed in this paper discovers rules with any type of mark and course, i.e. it discovers rare rules that contain not only infrequent but also frequent patterns, favouring the diversity of study cases. Furthermore, the rules mined have both the lowest support values and highest confidence values. Nevertheless, the conditional support varies depending on whether the rules contain frequent itemsets and thus low conditional supports are obtained, or whether the rules have infrequent item-sets, so higher conditional supports are obtained.

In Table 4, Rule 1 shows that if students do not spend any time on quizzes and the number of assignments is low, then they fail the final exam. So, it is an expected rule that shows the instructor the importance of using quizzes and assignments to pass the exam. Rule 2 is a very similar rule. This rule states that if students do not fail any quizzes but their number of assignments is low, then these students also fail the final exam. This rule is very interesting because it shows how the fact of passing the quizzes may not be condition enough for passing the final exam. Rule 3 is an interesting rule, since it shows that students that only spend a short time on quizzes could also pass the exam if they read a lot in the forum. So, this rule states that reading messages in the forum could significantly help student pass the exam. Rule 4 is very similar to Rule 3. Here, students that fail very few quizzes and read a lot in the forum pass the exam. Rule 5 shows an interesting rule. It states that students that do not spend any time on the forum and do not take any quizzes do not take the final exam. Similarly to Rule 5, Rule 6 shows that if the students do not do assignments, then they will be absent from the exam. Finally, Rules 7 and 8 identify students with an excellent mark. More specifically, these rules show that if students have a high number of assignments and spend a lot of time on them, or they submit a large number of assignments and they are subscribed to course 94, then these students obtain an excellent mark. Observe that this kind of rules help the instructor predicts the final performance of the students (both pass, fail, absent or excellent) before the exam.

Table 4 Rules extracted by executing the proposed evolutionary algorithm

Rule	Antecedent	Consequent	Support	SupportC/ SupportM	Confidence
1	total_time_quiz=ZERO AND n_assignment=LOW	mark=FAIL	0.021	–/0.40	1.00
2	n_quiz_fail=ZERO AND n_assignment=LOW	mark=FAIL	0.021	–/0.40	1.00
3	n_read=HIGH AND total_time_quiz=LOW	mark=PASS	0.039	–/0.48	1.00
4	n_quiz_fail=LOW AND n_read=HIGH	mark=PASS	0.052	–/0.51	1.00
5	total_time_forum=ZERO AND n_quiz_pass=ZERO	mark=ABSENT	0.047	–/0.72	1.00
6	total_time_assignment=ZERO AND n_assignment=ZERO	mark=ABSENT	0.047	–/0.72	1.00
7	total_time_assignment=HIGH AND n_assignment=HIGH	mark=EXCELLENT	0.017	–/0.83	1.00
8	course=94 AND total_time_assignment=HIGH	mark=EXCELLENT	0.013	0.65/0.25	1.00

5 Concluding Remarks and Future Works

In this paper we have explored the use of a GP algorithm for discovering rare association rules in an educational dataset. The application of this approach has shown to be an interesting research line in the context of educational data mining, where most real-world data are usually imbalanced. Rare-association rules are more difficult to mine using traditional ARM algorithms, since they do not usually consider class-imbalance and tend to be overwhelmed by the major class, whilst ignoring the minor class. In fact, we have shown that the Apriori algorithm discovers a huge number of rules with frequent items.

On the other hand, RARM, 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. However, these algorithms only use infrequent item-sets for discovering rare rules. Besides which, all these algorithms have strong restrictions, e.g. they only handle categorical attributes, and a slight variation of thresholds may cause a combinatorial explosion, requiring inappropriate runtime. In order to solve these drawbacks we have proposed a new evolutionary algorithm that uses grammars for generating the rare rules. We have compared our algorithm to existing RARM algorithms using a real Moodle dataset, which has shown that our algorithm discovers a lower number of the best rare rules, comprising not only of infrequent but also frequent item-sets. Furthermore, the rules mined have the lowest support values and the highest confidence values. We have also shown how the rules discovered by the ARM and RARM algorithms can help the instructor detect infrequent students' behaviours/activities in an e-learning environment, such as Moodle. As a proof of concept, we have evaluated the relationship between the on-line activities performed by the students and their final mark.

In future works, we plan to combine jumping emerging patterns with rare association rules, providing interesting association rules that comprise items whose frequency changes signicantly from one dataset to another.

Acknowledgements This research was supported by the Regional Government of Andalusia, project P08-TIC-3720, by the Spanish Ministry of Science and Technology, project TIN-2011-22408, and by FEDER funds. This research was also supported by the Spanish Ministry of Education under FPU grant AP2010-0041.

References

1. M. Adda, L. Wu, and Y. Feng. Rare itemset mining. In *Proceedings of the 6th International Conference on Machine Learning and Applications*, ICMLA '07, pages 73–80, Cincinnati, Ohio, 2007.
2. R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases*, VLDB '94, pages 487–499, San Francisco, CA, USA, 1994. Morgan Kaufmann Publishers Inc.
3. F. Berzal, I. Blanco, D. Sánchez, and M. A. Vila. Measuring the accuracy and interest of association rules: A new framework. *Intelligent Data Analysis*, 6(3):221–235, 2002.
4. P. G. Espejo, S. Ventura, and F. Herrera. A survey on the application of genetic programming to classification. *IEEE Transactions on Systems, Man and Cybernetics: Part C*, 40(2):121–144, 2010.
5. P. Fournier-Viger, C. Wu, and V.S. Tseng. Mining top-k association rules. In *Proceedings of the 25th Canadian Conference on Advances in Artificial Intelligence*, Canadian AI'12, pages 61–73, 2012.
6. J. Freyberger, N. T. Heffernan, and C. Ruiz. Using association rules to guide a search for best fitting transfer models of student learning. In *Workshop on Analyzing Student-Tutor Interaction Logs to Improve Educational Outcomes*, ICITS 2004, pages 1–10, 2004.
7. Q. Gu, Z. Cai, L. Zhu, and B. Huang. Data mining on imbalanced data sets. In *Proceedings of the International Conference on Advanced Computer Theory and Engineering*, pages 1020–1024, 2008.
8. H. Ha, D. Hwang, B. Ryu, and K. H. Yun. Mining association rules on significant rare data using relative support. *Journal of Systems and Software*, 67(3):181–191, 2003.
9. F. Herrera, C. J. Carmona, P. González, and M. J. del Jesus. An overview on subgroup discovery: Foundations and applications. *Knowledge and Information Systems*, 29(3):495–525, 2011.
10. R. I. Hoai, N. X. Whigham, P. A. Shan, Y. O'neill, and M. McKay. Grammar-based genetic programming: A survey. *Genetic Programming and Evolvable Machines*, 11(3-4):365–396, 2010.
11. J. Kim and J. Kang. Towards identifying unresolved discussions in student online forums. *Applied Intelligence*, 40(4):601–612, 2014.
12. Y. S. Koh and N. Rountree. Finding sporadic rules using apriori-inverse. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 3518:97–106, 2005".

13. Y. S. Koh and N. Rountree. *Rare Association Rule Mining and Knowledge Discovery: Technologies for Infrequent and Critical Event Detection*. Information Science Reference, Hershey, New York, 2010.
14. J. Lu. Personalized e-learning material recommender system. In *Proceedings of the International Conference on Information Technology for Application*, pages 374–379, 2004.
15. J. M. Luna, J. R. Romero, C. Romero, and S. Ventura. Discovering subgroups by means of genetic programming. In *Proceedings of the 16th European Conference*, EuroGP 2013, pages 121–132, Vienna, Austria, 2013.
16. J. M. Luna, J. R. Romero, and S. Ventura. Design and behavior study of a grammar-guided genetic programming algorithm for mining association rules. *Knowledge and Information Systems*, 32(1):53–76, 2012.
17. J.M. Luna, J.R. Romero, and S. Ventura. On the adaptability of g3parm to the extraction of rare association rules. *Knowledge and Information Systems*, 38(2):391–418, 2014.
18. A. Merceron and K. Yacef. Interestingness measures for association rules in educational data. *Educational Data Mining*, 2008.
19. Agathe Merceron and Kalina Yacef. Mining student data captured from a web-based tutoring tool: Initial exploration and results. *Journal of Interactive Learning Research*, 15(4):319–346, October 2004.
20. B. Minaei-Bidgoli, P. N. Tan, and W. F. Punch. Mining interesting contrast rules for a web-based educational system. In *Proceedings of the International Conference on Machine Learning Applications*, ICMLA, pages 320–327. IEEE Computer Society, 2004.
21. C. Ordoñez, N. Ezquerra, and C. Santana. Constraining and summarizing association rules in medical data. *Knowledge and Information Systems*, 9(3):259–283, 2006.
22. L. Raedt, T. Guns, and S. Nijssen. Constraint programming for data mining and machine learning. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, SIGKDD, pages 204–212, 2008.
23. A. Rahman, C. I. Ezeife, and A. K. Aggarwal. Wifi miner: An online apriori-infrequent based wireless intrusion system. In *Proceedings of the 2nd International Workshop in Knowledge Discovery from Sensor Data*, SensorKDD '08, pages 76–93, Las Vegas, USA, 2008.
24. A. A. Ramli. Web usage mining using apriori algorithm: Uum learning care portal case. In *Proceedings of the International Conference on Knowledge Management*, pages 1–19, Malaysia, 2005.
25. C. Romero, J. M. Luna, J. R. Romero, and S. Ventura. Mining Rare Association Rules from e-Learning Data. In *Proceedings of the 3rd International Conference on Educational Data Mining, EDM 2010*, pages 171–180, 2010.
26. C. Romero, J. M. Luna, J. R. Romero, and S. Ventura. Rm-tool: A framework for discovering and evaluating association rules. *Advances in Engineering Software*, 42(8):566–576, 2011.
27. C. Romero, C. Ventura, and P. De Bra. Knowledge discovery with genetic programming for providing feedback to courseware author. user modeling and user-adapted interaction. *Journal of Personalization Research*, 5(14):425–464, 2004.
28. C. Romero and S. Ventura. Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, 40(6):601–618, 2010.
29. C. Romero, S. Ventura, and E. García. Data mining in course management systems: Moodle case study and tutorial. *Computers & Education*, 51(1):368–384, 2008.
30. C. Romero, S. Ventura, M. Pechenizky, and R. Baker. *Handbook of Educational Data Mining*. SChapman and Hall/CRC Press, 2010.
31. D. Sánchez, J. M. Serrano, L. Cerda, and M. A. Vila. Association rules applied to credit card fraud detection. *Expert systems with applications*, (36):3630–3640, 2008.
32. L. Szathmary, A. Napoli, and P. Valtchev. Towards rare itemset mining. In *Proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence*, ICTAI '07, pages 305–312, Patras, Greece, 2007.
33. P. Tan and V. Kumar. Interestingness measures for association patterns: A perspective. In *Proceedings of the Workshop on Postprocessing in Machine Learning and Data Mining*, KDD '00, New York, USA.
34. S. Yen, Y. Lee, and C. Wang. An efficient algorithm for incrementally mining frequent closed itemsets. *Applied Intelligence*, 40(4):649–668, 2014.
35. P. Yu, C. Own, and L. Lin. On learning behavior analysis of web based interactive environment. In *Proceedings of the International Conference on Implementing Curricular Change in Engineering Education*, ICCEE, Oslo, Norway.
36. O. R. Zaíane. Building a recommender agent for e-learning systems. In *Proceedings of the International Conference on Computers in Education*, ICCE '02, Washington, DC, USA, 2002. IEEE Computer Society.
37. C. Zhang and S. Zhang. *Association rule mining: models and algorithms*. Springer-Verlag, Berlin, Heidelberg, 2002.
38. H. Zhang, Y. Zhao, L. Cao, and C. Zhang. Class association rule mining with multiple imbalanced attributes. In MehmetA. Orgun and John Thornton, editors, *AI 2007: Advances in Artificial Intelligence*, volume 4830 of *Lecture Notes in Computer Science*, pages 827–831. Springer Berlin Heidelberg, 2007.