Evolutionary algorithms for subgroup discovery applied to e-learning data

C.J. Carmona, P. González, M.J. del Jesus
Department of Computer Science
University of Jaen
Jaen, Spain
{ccarmona, pglez, mjjesus}@ujaen.es

C. Romero, S. Ventura
Department of Computer Science and Numerical Analysis
University of Cordoba
Cordoba, Spain
{cromero, sventura}@uco.es

Abstract—**This work presents the application of subgroup discovery techniques to e-learning data from learning management systems (LMS) of andalusian universities. The objective is to extract rules describing relationships between the use of the different activities and modules available in the e-learning platform and the final mark obtained by the students. For this purpose, the results of different classical and evolutionary subgroup discovery algorithms are compared, showing the adequacy of the evolutionary algorithms to solve this problem. Some of the rules obtained are analyzed with the aim of extract knowledge allowing the teachers to take actions to improve the performance of their students.**

Subgroup discovery; educational data mining; e-learning systems; evolutionary algorithms; fuzzy rules

I. INTRODUCTION

For almost as long as LMSs exist, researchers have been interested in study how the usage logs of these systems can be used to improve the learning process. The more used approach to exploit this data uses automated evaluation of system logs and databases [1] using data mining techniques to provide additional information for teaching staff about the quality of the student experience. In this sense, data mining techniques can be applied to analyze student's usage data in order to identify useful patterns and to evaluate web activity to get more objective feedback for instruction and more knowledge about how the students learn on the LMS [2].

Educational data mining is an emerging interdisciplinary research area that deals with the development of methods to explore data from an educational context [2]. It is concerned with the development of mining methods to explore the unique types of data in educational settings and, using these methods, to better understand students and learning settings. A data mining algorithm can discover knowledge using different representation models and techniques from two different perspectives: predictive induction, whose objective is the discovery of knowledge for classification or prediction [3]; or descriptive induction, whose main objective is the extraction of interesting knowledge from data. In this area, attention can be drawn to the discovery of association rules following an unsupervised learning model [4], subgroup discovery [5] and other approaches to non-classificatory induction.

Association rule mining [4] is one of the better-studied descriptive data mining methods whose objective is to discover descriptive rules about relations between attributes of a set of data. It has been applied to LMS in order to reveal which contents students tend to access together, or which combination of tools they use [5].

The extraction of association rules has been successfully applied in e-learning systems to discover relationships or associations between the different visited web pages, activities performed, marks obtained, etc. A pioneering work in applying web mining techniques to e-learning systems is [6] which proposes the use of agents [7] to recommend online learning activities or shortcuts in a web course based on access records of students, thereby enhancing the online learning process. Another work that uses association rule mining techniques and collaborative filtering is [8] with the aim of discover helpful navigation patterns and propose a navigation model. The use of methods such as linear regression in combination with association rules to obtain learning transferring patterns of students from the log files in intelligent tutoring systems is proposed in [9]. In [10] is described the use of fuzzy association rules to discover relationships between patterns of student behavior, including access time, number of pages read, questions answered, or read and sent messages. Evolutionary algorithms are used in [11] as a technique for discovering useful information for the authors of such courses, in order to make improvements of the contents, the structure or the adaptation of the courses. Another work using evolutionary algorithms is [12], where an association analysis is performed to predict student performance. Association rules can be adapted to subgroup discovery, a new and very interesting task in educational environments. In fact, this paper proposes the application of data mining techniques for subgroup discovering over educational data.

Subgroup discovery (SD) is a descriptive inductive learning area in which, given a set of data and a property of interest to the user, an attempt is made to locate subgroups which are statistically "most interesting" for the user [13]. A subgroup is interesting if it has an unusual statistical distribution with respect to the property of interest. The objective is to discover interesting properties of subgroups by obtaining simple rules, which are highly significant and with high support.

978-1-4244-6571-2/10/$26.00 ©2010 IEEE

A rule describing a subgroup has the form:

Condition $\rightarrow$ *Class_label*

where the property of interest for subgroup discovery is the value of the variable (*Class_label*) which appears in the consequent of the rule, and the antecedent of the rule (*Condition*) is a conjunction of features (attribute-value pairs) selected between the features describing the instances of the data set.

Genetic Algorithms (GAs) are beginning to be used to solve SD problems [14], [15] because they offer a set of advantages for knowledge extraction and specifically for rule induction processes. A fuzzy approach in a SD algorithm, which considers descriptive fuzzy rules, allows us to obtain knowledge in a similar way to human reasoning, and so the obtaining of more interpretable and actionable solutions in the field of SD, and in general in the analysis of data in order to establish relationships and identify patterns.

In this work we apply the evolutionary subgroup discovery algorithms SDIGA [14] and MESDIF [16] to obtain fuzzy rules which describe relationships between the student's usage of the different resources provided by the e-learning systems and the final marks obtained. The objective is to characterize subgroups of students whose final marks are significantly different from those of all students, and use this knowledge to improve the learning process. The results obtained by these algorithms are compared with the ones obtained by classical subgroup discovery algorithms, showing the suitability of the evolutionary approach to this problem.

The work is arranged in the following way: Section 2 introduces the subgroup discovery and describes the evolutionary rule induction algorithms used, SDIGA and MESDIF, and the classical subgroup discovery algorithms Apriori-SD and CN2-SD. Section 3 describes the e-learning case study using the Moodle e-learning system implanted in the University of Cordoba, the experimentation carried out, the analysis of the results and of some of the rules obtained. Finally, the conclusions and further research are outlined.

II. SUBGROUP DISCOVERY: CLASSICAL APPROACHES AND EVOLUTIONARY ALGORITHMS

In the literature, several different proposals for the extraction of rules in the area of subgroup discovery can be found. Two of the most used classical algorithms for subgroup discovery are adaptations of the well-known algorithms Apriori (used for the extraction of association rules) and CN2 (used for the extraction of rule bases for classification), which are named Apriori-SD and CN2-SD respectively.

There are also evolutionary proposals for the SD task. SDIGA [14] is an evolutionary algorithm for the induction of fuzzy rules which uses linguistic rules as description language for the specification of the subgroups, and adaptations of the measures used in the association rule induction algorithms as quality measures for the subgroup discovery task. MESDIF [16] is a multiobjective evolutionary algorithm using the same quality measures as SDIGA.

A. Classical subgroup discovery algorithms

The most widely used state-of-art SD algorithms are described below:

- Apriori-SD [17] uses the weighted relative accuracy as quality measure for the induced rules, using support and significance of each individual rule for the evaluation. It is an adaptation for subgroup discovery of the classification rule learning algorithm Apriori-C [18], based in the original Apriori association rule learning algorithm [4].
- CN2-SD [19] is a modified version of the CN2 classification rule algorithm [20] which induces subgroups in the form of rules using as quality measure the relation between true positives and false positives. As Apriori-SD, it uses a modified weighted relative accuracy measure for the selection of the rules.

B. Evolutionary rule induction

In any data mining process there are different tasks or problems which can be approached and solved as optimization and search problems. GAs are general purpose search algorithms which use principles inspired by natural genetics to evolve solutions to problems [21].

GAs have several advantages as a rule induction method: they tend to cope well with attribute interaction (because they usually evaluate a rule as a whole), have ability to scour a search space thoroughly, or allow to find complex interactions due to the implicit backtracking in its search of the rule space. This makes GAs particularly suited for the SD task.

The genetic representation of solutions is the most determinant aspect of any rule induction GA. In this sense, the proposals in the specialized literature follow two approaches in order to encode rules within a population of individuals [22]:

- The *"Chromosome = Rule"* approach, in which each individual codifies a single rule.
- The *"Chromosome = Set of rules"*, also called the Pittsburgh approach, in which each individual represents a set of rules.

Within the "*Chromosome = Rule*" approach, three generic proposals can be found: the *Michigan* approach, the *Iterative Rule Learning* (IRL) approach and the *cooperative-competitive* approach. A complete description of the different proposals for the evolutionary induction of rules can be found in [22].

In processes aimed to the extraction of rules for the subgroup discovery task, the "*Chromosome = Rule*" approach is more suited because the objective is to find a reduced set of rules in which the quality of each rule is evaluated independently from the rest, and it is not necessary to evaluate the set of rules jointly.

Among the different proposals of GAs for the description of subgroups, below are described two focused in this approach.

C. Evolutionary rule induction algorithm SDIGA

SDIGA (Subgroup Discovery Iterative Genetic Algorithm) is an evolutionary model for the extraction of fuzzy rules for the subgroup discovery task. This algorithm is described in [14], including here a brief summary of its key features.

In the subgroup discovery task we must distinguish between a set of descriptive variables and a single target variable which describes the subgroups. As the objective is to obtain a set of rules describing subgroups for all the values of the target variable, the GA of this proposal discovers fuzzy rules with the consequent prefixed to one of the possible values of the target variable. In this way, each run of SDIGA obtains a set of rules corresponding to the value specified for the target variable, and the algorithm must be run for each one of the possible values of the target variable.

In this proposal, each candidate solution is coded according to the *"Chromosome = Rule"* approach representing only the antecedent of the rule in the chromosome (since all the individuals of the population are associated with the same value of the target variable). The antecedent of a rule is composed of a conjunction of variable-value pairs. The information related to each rule is stored in a fixed length chromosome using an integer representation model (in which the i-th position indicates the value of the i-th variable).

The core of SDIGA is a GA which uses a post-processing step based on a local search (a hill-climbing procedure). The hybrid GA extracts one simple and interpretable fuzzy rule. The post-processing step is applied in order to increase the generality of the extracted rule. The optimized rule will substitute the original one only if it overcomes minimum confidence.

The GA uses a modified steady-state reproduction model, with the aim of increasing the diversity of the population, in which the original population is modified through the substitution of the worst individuals by individuals resulting from crossover and mutation. A two-point crossover operator is applied to the two best individuals of the population, obtaining two new individuals, who will substitute the two worst individuals in the population. Mutation is carried out by means of a biased random mutation operator applied to the gene selected according to the mutation probability. The mutation can eliminate the variable selected or assign a random value.

This hybrid GA is included in an iterative process for the extraction of a set of rules describing different parts (not necessarily apart) of the search space. A set of solutions generated in successive runs of the GA is obtained, corresponding with one value of the target variable. This is accomplished by checking the instances covered by the obtained rule, preventing a new rule to cover exactly the same examples. Thus different fuzzy rules are obtained, although they may be overlapping.

The model uses fuzzy rules, which provide better interpretability for the rules extracted by means of the use of a knowledge representation near to the expert, also allowing the use of numerical variables without a prior discretization. The fuzzy sets for the linguistic labels defined by the corresponding membership functions can be specified by the user or defined by a uniform partition if there is no available expert knowledge (using uniform partitions with triangular membership functions).

The evaluation function of the GA combines, according to the following expression, three factors: confidence (*Conf*), support (*Supp*) and unusualness (*Unus*):

$$fitness(c) = \frac{\omega_1 \cdot Conf(c) + \omega_2 \cdot Supp(c) + \omega_3 \cdot Unus(c)}{\omega_1 + \omega_2 + \omega_3} \quad .1$$

in which w_i are the weights assigned to each measure, allowing to establish the importance of each one of the measures.

These measures are computed in the following way:

- *Confidence*. Determines the relative frequency of examples satisfying the complete rule among those satisfying only the antecedent. In this paper we use an adaptation of Quinlan's accuracy expression [23] in order to generate fuzzy rules [24]: the sum of the degree of membership of the examples of this class to the zone determined by the antecedent, divided the sum of the degree of membership of all the examples that verifies the antecedent part of this rule (irrespective of their class) to the same zone.
- *Support*. It is a measure of the degree of coverage than the rule offers to the examples of the class. This measure aims to promote the extraction of different rules in successive runs of the hybrid GA. This, for the computing of the support we only consider the examples not covered by rules obtained in previous runs of the GA. This way, support is defined as the quotient between the number of new examples covered by the rule and the number of not covered examples of the data set.
- *Unusualness*. It is also known as the weighted relative accuracy. It measures the balance between the coverage of the rule and its accuracy gain and is normally used as a measure of the interest of the rule.

The overall objective of the evaluation function is to guide the search towards rules that maximize the accuracy and the interest of the rule, also minimizing the number of negative and not covered examples.

D. Multiobjective evolutionary algorithn MESDIF

This multiobjective evolutionary proposal [16] extracts rules using the same representation as the SDIGA algorithm, and its objective is the extraction of a variable number of different rules for each value of the target variable. The algorithm allows the extraction of fuzzy and/or crisp rules, for problems with continuous and/or categorical variables.

The multiobjective GA follows the SPEA2 approach [25], and so applies the concepts of elitism in the rule selection (using a secondary population) and the search of optimal solutions in the Pareto front (the individuals of the population are ordered considering if each one of them is or not dominated

by others using the concept of Pareto optimal). Any multiobjective GA must be designed aimed to reach two objectives: obtaining of good approximations to the Pareto front and maintaining the diversity of the solutions in order to adequately sample the space of the solutions and not to converge to a single solution or to a bounded section of the Pareto front. To preserve diversity at a phenotypic level the algorithm uses a niche technique that considers how close the values of the objectives are. Table I shows the operation scheme of the proposed model.

In this subgroup discovery process the objective is the extraction highly descriptive capacity, comprehensible and interesting rules. So, three objectives have been defined in this multiobjective proposal: support, confidence and unusualness. Confidence and unusualness are defined in the same way as in SDIGA, but a different definition of support is used because the algorithm obtains sets of different rules without using an iterative model. In his case, support is defined as the quotient between the number of examples of the class described by the rule and the total number of examples of the class.

This algorithm uses an elite population with a fixed size, and so it is necessary to define functions for truncation and filling. The truncation function allows the elimination of non-dominated solutions from the elite population if it exceeds the defined size. For this purpose it is used a niche schema defined around the density measured by the distance to its k-th nearest neighbour, in which, in an iterative process, in each iteration it is eliminated from the elite population the individual that is nearest of others respect of the values of the objectives. The fill function allows adding dominated individuals from the population and the elite population until the exact size of the set is reached (ordering the individuals according to their fitness values).

The algorithm uses the following reproduction model:

- Join the original population with the elite population obtaining then the non-dominated individuals of the joining of these populations.
- Apply a binary tournament selection on the non-dominated individuals.
- Apply recombination to the resulting population by a two point cross operator and a biased uniform mutation operator in which half the mutations carried out have the effect of eliminating the corresponding variable, in order to increase the generality of the rules.

III. E-LEARNING CASE STUDY: USAGE DATA OF THE MOODLE E-LEARNING SYSTEM OF THE UNIVERSITY OF CORDOBA

In this section we examine the e-learning case study. First, we describe the problem and then the experimental results obtained in the execution of the different subgroup discovery algorithms are shown. Finally, an analysis from the point of view of the teacher of several of the rules obtained is performed with the aim of improving the e-learning courses.

TABLE I. OPERATION SCHEME OF THE PROPOSED EVOLUTIONARY ALGORITHM

Initialization: Generate an initial population P_0 and create an empty elite population $P'_0 = \emptyset$.
Repetir
  Fitness assignment: Compute fitness values for the join of the individuals in population P_t and in elite population P'_t.
  Environmental selection: Copy al non-dominated individuals in population P_t and in elite population P'_t in the elite population P'_{t+1}. If the size of P'_{t+1} is bigger than the number of individuals to store, reduce P'_{t+1} by a truncation operator; otherwise, if the size of P'_{t+1} is lower than the number of individuals, fill with dominated individuals of P_t and P'_t.
  Mating selection: Perform binary tournament selection with replacement on elite population P'_{t+1} applying later crossover and mutation operator to fill the matting pool. The result is the population P_{t+1}.
While stip criteria is not verified.
Return the non-dominated individuals in elite population P'_{t+1}.

As we have mentioned previously, we have used student's usage data of the Moodle system [26], one of the most used web-based e-learning systems. Moodle is an alternative to proprietary commercial online learning solutions, is distributed free under open source licensing and has been installed at universities and institutions all over the world.

The aim of the use of subgroup discovery in this problem is to analyze the possible relation between the usage of complementary activities of a course and the final marks obtained by the students. The final mark is used as the variable to characterize, using the different marks to divide the data into classes and codifying them as values of the consequent of the rules.

We have run different subgroup discovery algorithms in order to compare the results and show what type of algorithm discovers more useful knowledge for the course teacher. The objective is to present the results to the teacher in the form of rules in order to allow the use of this knowledge in the decision making concerning the complementary activities of the course. For example, the teacher can decide to promote the use of some type of activities to obtain a high mark, or on the contrary eliminate some activities because they are associated with low marks.

The Moodle system contains a great deal of detailed information on course content, users, usage, etc., stored in a relational data base keeping detailed logs of all the activities performed by the students. We can use these logs in order to determine which students have been active in the course, what they did, when, or if everyone has done a certain task or spent a required amount of time online within certain activities [27].

We have available information corresponding to 192 different courses of the University of Cordoba. Among all these courses, we have chosen 5 courses, corresponding to the subjects with the highest usage of the activities and resources, involving a total of 293 students. Although our approach can be applied to just one course we have selected courses with high student participation in order to generalize the results. However, there is no a minimum amount of students to obtain any rule.

We have applied a pre-processing step to the information, obtaining a summary table with the most important information related to our objective. Table II shows this summary table, including the activities completed and the mark obtained by each student in an e-learning course. We have discretized the marks into classes (fail, pass, good and excellent) in order to codify them as the values of the rule consequent. The information obtained has been exported to a text file using the structure of the KEEL platform files [28] because the subgroup discovery algorithms used are implemented for the KEEL data mining platform.

TABLE II. ATTRIBUTES USED FOR EACH STUDENT

Name	Description
course	Identification of the course
n_assigment	Number of assignments completed
n_assigment_a	Number of assignments passed
n_assigment_s	Number of assignments failed
n_quiz	Number of quizzes completed
n_quiz_a	Number of quizzes passed
n_quiz_s	Number of quizzes failed
n_messages	Number of messages sent to the chat
n_messages_ap	Number of messages sent to the teacher
n_posts	Number of messages sent to the forum
n_read	Number of forum messages read
mark	Discretized student's mark

A. Experimental results of the application of subgroup discovery algorithms

The experiments had been performed using four different algorithms in order to compare their results and determine the most appropriated approach: the classical subgroup discovery algorithms Apriori-SD and CN2-SD and the evolutionary algorithms SDIGA and MESDIF.

We have performed several runs of the different algorithms in order to obtain the average values of the measures used to evaluate the quality of the rules.

For the classic deterministic algorithms Apriori-SD and CN2-SD we have performed a set of runs, varying one of their parameters each time. In the case of Apriori-SD, we have used 4 minimum confidence values (0.6, 0.7, 0.8 and 0.9) varying the minimum support value (0.03, 0.1, 0.2, 0.3, and 0.4). In the case of CN2-SD, we have used the γ parameter (0.9, 0.7, 0.5 and additive) varying the star size (1, 2, 3, 4, 5).

For the evolutionary algorithms, we have performed 5 different runs for each of the values of the target variable (fail, pass, good and excellent) using a population of 100 individuals, a maximum number of evaluations of individuals in each GA run of 10000, a crossover probability of 0.6, a mutation probability of 0.01, and 5 linguistic labels for the continuous variables (very high, high, medium, low, very low).

In addition, the weights used for the fitness function in the SDIGA algorithm are: 3 for accuracy, 1 for coverage and 4 for significance. This set of weights has been chosen according to the results obtained in an experimental study. For the SDGA algorithm, different values of the minimum confidence (0.6, 0.7, 0.8 and 0.9) are used. For the MESDIF algorithm, different elite population sizes are used (3. 4, 5 and 10).

Table III shows the results obtained by the classic algorithms with their different parameter values and the averages of the 5 runs of the evolutionary algorithms with each value of minimum confidence. The table shows the total number of rules obtained, the number of attributes in the antecedents of the rules and the values of their quality measures (support, confidence, unusualness and significance).

In order to analyze de results with respect to the number of rules and the number of variables of the rules, we have to take into account that a key aspect in our problem is the interpretability of the results. In this sense, we are interested in the extraction of a reduced set of rules with a low number of attributes. This will facilitate the understandability of these rules for the teacher.

TABLE III. EXPERIMENTAL RESULTS OF THE ALGORITHMS

Algorithm	*Number of rules*	*Number of attributes*	*Support*	*Confidence*	*Unusualness*	*Significance*
SDIGA MinCnf 0.6	5.4	3.1600	0.7418	0.7709	0.0340	20.0671
SDIGA MinCnf 0.7	4.8	2.8000	0.8096	0.5716	0.0378	23.1872
SDIGA MinCnf 0.8	5.4	3.2583	0.6534	0.6805	0.0319	21.4395
SDIGA MinCnf 0.9	6.2	3.1821	0.3815	0.7769	0.0200	14.4846
MESDIF Elite 3	12.0	3.2000	0.9795	0.4466	0.0173	21.7596
MESDIF Elite 4	16.0	3.4125	0.9890	0.4242	0.0234	24.7158
MESDIF Elite 5	20.0	3.6500	0.9884	0.4517	0.0241	23.6884
MESDIF Elite 10	40.0	3.6950	1.0000	0.4732	0.0284	23.8734
Apriori-SD MinCnf 0.6	9.8	1.0400	0.5924	0.6157	0.0183	27.3901
Apriori-SD MinCnf 0.7	10.4	1.3238	0.5513	0.6301	0.0176	31.4304
Apriori-SD MinCnf 0.8	5.0	0.8294	0.3734	0.3842	0.0205	21.2968
Apriori-SD MinCnf 0.9	4.6	1.1692	0.2089	0.3787	0.0192	21.0734
CN2-SD (γ=0.5)	15.6	5.6406	0.9342	0.7143	0.0264	45.8554
CN2-SD (γ=0.7)	18.4	5.6857	0.9876	0.7177	0.0349	47.0058
CN2-SD (γ=0.9)	25.2	5.7177	0.9890	0.7184	0.0315	47.2862
CN2-SD (add)	31.5	5.7741	1.0000	0.7129	0.0273	54.7134

Thus, in this aspect the results of SDIGA are better than the results of the classical subgroup discovery algorithms.

With respect to the quality measures, it can be seen that:

- For the support measure, MESDIF and CN2-SD obtain the best results, obtaining rules with a high generality, so covering most of the students. MESDIF obtains the same results as CN2-SD but with a lower number of attributes. SDIGA also obtains high values in this measure.
- Related with the confidence or precision of the rule, which indicates the number of students covered by the antecedent of the rule and which correspond to the class associated to the same, the best results are obtained by SDIGA and CN2-SD, which obtain rules with a good level of precision.
- For the unusualness, the best results are obtained by the SDIGA algorithm, but all the algorithms obtain similar values in this quality measure.
- Finally, significance measures the relevance and interest of the rule. CN2-SD is the algorithm that obtains the best results in this measure, and both evolutionary algorithms obtain results very similar to that of Apriori-SD.

The most desirable algorithm with regard to the values of these quality measures would be an algorithm that simultaneously shows the highest values for all the measures. As we have seen there is not a single algorithm which achieves this. Then we have to consider the interpretability of the results for the teachers. In this sense, the evolutionary algorithms offer knowledge with a reduced number of rules and variables. In addition, the use of fuzzy rules in SDIGA and MESDIF contributes to the interpretability of the extracted rules due to the use of a knowledge representation nearest to the expert, also allowing the use of continuous features without a previous discretization.

B. Analysis of the comprehensibility of the rules obtained by evolutionary algorithms

According to the results shown in the previous section, the evolutionary algorithms for subgroup discovery take advantage over the classic algorithms in relation to the comprehensibility of the rules extracted when using them in the decision making of the teacher of a course. This is due to the use of attributes of the form "LABEL = VALUE", where VALUE are linguistic labels provided by the expert, which allow an easier interpretation for the teacher.

On one hand, rule induction algorithms are normally also considered to produce comprehensible models because they discover a set of IF-THEN classification rules that are a high-level knowledge representation and can be used directly for decision making. Some examples of rules obtained by a rule induction algorithm are:

IF n_assignment < 6 THEN mark = FAIL

IF n_assignment > 10 AND n_quiz_a > 9 THEN mark = GOOD

IF course = 29 AND n_quiz_a = 0 THEN mark = FAIL

IF course = 110 AND n_quiz_a > 7 THEN mark = GOOD

On the other hand, fuzzy rule algorithms obtain IF-THEN rules that use linguistic terms that make them more comprehensible/interpretable by humans. So, this type of rule is very intuitive and easily understood by problem-domain experts like teachers.

Below are shown a few examples of rules discovered by the evolutionary algorithms, analyzing their meaning a applied to the improvement of the courses.

IF course = 110 AND n_assignment = High AND n_posts = High
THEN mark = Good
Support: 0.7045 Confidence: 0.7231

For the course 110, students who have done many activities and have sent many messages to the forum have obtained a high mark. The teacher of this course should continue promoting such follow activities as they have been proven effective in the final mark obtained by the students who carry them out.

IF course = 88 AND n_messages = Very High
THEN mark = Fail
Support: 0.1930 Confidence: 0.9444

For the course 88, students who have sent a lot of messages to the chat have failed. The teacher of this course should eliminate the chat because it has no provided any benefit to the students, but on the contrary it can be seen as a source of distraction.

IF n_read = High AND n_messages_ap = Very low
THEN mark = Fail
Support: 0.1176 Confidence: 0.7500

For any course, if the number of messages read from the forum is high but the number of messages sent to the teacher is low, then the final mark is fail. This rule provides information on a small group of students who tend to fail. The teacher can then pay more attention to these students trying to motivate them in time to pass the course.

Finally, in our educational problem the final objective is to show the instructor interesting information about student marks depending on the usage of Moodle courses. Then, the instructor can use the discovered knowledge for decision making about activities. For example, some of the rules discovered show that the number of quizzes passed in Moodle was the main determinant of the final marks, but there are some others that could help the teacher to decide whether to promote the use of some activities to obtain higher marks, or on the contrary, to decide to eliminate some activities because they are related to low marks. It could also be possible for the teacher to detect new students with learning problems in time to remedy (students predicted as *Fail*).

IV. CONCLUSIONS

In this work the application of subgroup discovery techniques to data from a real world problem of knowledge extraction in e-learning systems.

After testing the proposed algorithms by comparing them with other classical subgroup discovery algorithms, it appears that evolutionary algorithms are well suited for solving the proposed problem. They obtain a reduced set of understandable rules (due both to their small size and to use of linguistic labels) that make them more interpretable to the teacher, in addition to obtaining similar values in the other quality measures. Based on the rules obtained, teachers can make decisions about course activities to improve the performance of their students.

ACKNOWLEDGMENT

This work was supported by the Spanish Ministry of Education, Social Policy and Sports under projects TIN-2008-06681-C06-02 and TIN-2008-06681-C06-03, and by the Andalusian Research Plan under projects TIC-3928 and TIC-3720.

REFERENCES

[1] J. Hung, and K. Zhang, "Data mining applications to online learning.", in Proceedings of World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education, Chesapeake, VA, pp. 2014-2021, 2006.

[2] C. Romero and S. Ventura, "Educational data mining: a survey from 1995 to 2005", Expert Systems with Applications, vol 3(1), pp. 135-146, 2007.

[3] D. Michie, D.J. Spiegelhalter and C.C. Taylor, Machine learning, neural and statistical classification, Ellis Horwood, 1994.

[4] R. Agrawal, T. Imielinski and A. Swami, "Mining association rules between sets of items in large databases", ACM SIGMOD Conference on Management of Data, pp. 207-216, 1993.

[5] F. Wang, "On using data-mining technology for browsing log file analysis in asynchronous learning environment", in Proceedings of the Conference on educational multimedia, hypermedia and telecommunication, Denver, Colorado, pp. 2005– 2006, 2002.

[6] O.R. Zaïane, "Web Usage Mining for a Better Web-Based Learning Environment", Conference on Advanced Technology for Education, Alberta, pp 60-64. 2001.

[7] O.R. Zaïane, "Building a Recommender Agent for e-Learning Systems", International Conference on Computers in Education, New Zealand. pp 55-59. 2002.

[8] F. Wang, "On Analysis and Modeling of Student Browsing Behavior in Web-Based Asynchronous Learning Environments". International Conference on Web-based Learning. pp. 69-80. 2002.

[9] J. Freyberger, N.T. Heffernan and C. Ruiz, "Using association rules to guide a search for best fitting transfer models of student learning", International Conference on Intelligent Tutoring Systems, pp 1-10. 2004.

[10] P. Yu, C. Own and L. Lin, "On the learning behavior analysis of web based interactive environmnent", ICCE, pp. 1-8, 2001.

[11] C. Romero, S. Ventura and P. de Bra, "Knowledge discovery with genetic programming for providing feedback to courseware author", User Modeling and User-Adapted Interaction, vol. 14(5), pp. 425-464, 2004.

[12] B. Minaei-Bidgoli and W.F. Punch, Predicting student performance: an application of data mining methods with the educational web-based system LON-CAPA, IEEE Frontiers in Education, pp 1-6, 2003.

[13] W. Klösgen, "Explora: A multipattern and multistrategy discovery assistant", in Advances in Knowledge Discovery and Data Mining, Menlo Park, California, AAAI Press, pp. 249–271, 1996.

[14] M.J. del Jesus, P. González, F. Herrera and M. Mesonero, "Evolutionary fuzzy rule induction process for subgroup discovery: A case study in marketing", IEEE Transactions on Fuzzy Systems, vol. 12(3), pp. 296-308, 2007.

[15] C. Romero, P. González, S. Ventura, M.J. del Jesus and F. Herrera, "Evolutionary algorithm for subgroup discovery in e-learning: A practical application using Moodle data", Expert Systems with Applications, vol. 36 , pp. 1632-1644, 2009.

[16] F. Berlanga, M.J. del Jesus, P. González, F. Herrera and M. Mesonero, "Multiobjective evolutionary induction of SD fuzzy rules: A case study in marketing", LNCS, vol. 4065, pp. 337-349, 2006.

[17] B. Kavsek, N. Lavrac and V. Jovanoski, "APRIORI-SD: Adapting association rule learning to subgroup discovery", Advances in Intelligent Data Analysis, vol. V, pp. 230-241, 2003.

[18] L. Bing, H. Wynne and M. Yiming, "Integrating Classification and Association Rule Mining", Fourth International Conference on Knowledge Discovery and Data Mining, pp. 80-86, 1998.

[19] N. Lavrac, B. Kavsec, P. Flach and L. Todorovski, "Subgroup discovery with CN2-SD", Journal of Machine Learning Research, vol. 5, pp. 153-188, 2004.

[20] P. Clark and T. Niblett, "The CN2 induction algorithm", Machine Learning, vol. 3(4), pp. 261–283, 1989.

[21] D.E. Goldberg, Genetic algorithms in search, optimization and machine learning, Addison-Wesley, 1989.

[22] O. Cordón, F. Herrera, F. Hoffmann and L. Magdalena, Genetic fuzzy systems: evolutionary tuning and learning of fuzzy knowledge bases, World Scientific, 2001.

[23] J.R. Quinlan, "Generating production rules from decision trees", International Joint Conference on Artificial Intelligence, Milan, pp. 304-307, 1987.

[24] O. Cordón, M.J. del Jesus and F. Herrera, "Genetic learning of fuzzy rule-based classification systems co-operating with fuzzy reasoning methods", International Journal of Intelligent Systems, vol. 13(10/11), pp. 1025-1053, 1998.

[25] E. Zitzler, M. Laumanns and L. Thiele, "SPEA2: Improving the strength pareto evolutionary algorithm for multiobjective optimisation", Evolutionary methods for design, optimisation and control, CIMNE, pp. 95-100, 2002.

[26] M. Flate, Online education and learning management systems. Global e-learning in a Scandinavian perspective, NKI Gorlaget, Oslo, 2003.

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

[28] J. Alcalá, M.J. del Jesús, J.M. Garrell, F. Herrera, C. Hervás and L. Sánchez, "Proyecto KEEL: Desarrollo de una Herramienta para el Análisis e Implementación de Algoritmos de Extracción de Conocimiento Evolutivos", Tendencias de la Minería de Datos en España, pp. 413-423, 2004.