Expert Systems with Applications 38 (2011) 15020–15031

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Multiple instance learning for classifying students in learning management systems

Amelia Zafra *, Cristóbal Romero, Sebastián Ventura

Department of Computer Science and Numerical Analysis, University of Cordoba, Spain

ARTICLE INFO

Keywords:
Educational data mining
Multiple instance learning
Traditional supervised learning

ABSTRACT

In this paper, a new approach based on multiple instance learning is proposed to predict student's performance and to improve the obtained results using a classical single instance learning. Multiple instance learning provides a more suitable and optimized representation that is adapted to available information of each student and course eliminating the missing values that make difficult to find efficient solutions when traditional supervised learning is used. To check the efficiency of the new proposed representation, the most popular techniques of traditional supervised learning based on single instances are compared to those based on multiple instance learning. Computational experiments show that when the problem is regarded as a multiple instance one, performance is significantly better and the weaknesses of single-instance representation are overcome.

1. Introduction

Academic failure or success on the part of university students has been the subject of many debates. Many educational psychologists have tried to understand this issue and then explain it, and many statisticians have tried to predict outcomes (Busato, Prins, Elshout, & Hamaker, 2000; Chidolue, 2001; Minnaert & Janssen, 1999). Nowadays, advances in technology and the impact of Internet in the last few years have promoted the appearance of the virtual learning environment (VLE) or e-learning platforms (Nagi & Suesawaluk, 2008) that routinely collect vast amounts of data on student activity providing an alternative way to deal with the same questions. E-learning platforms generate log files that collect all available information which then gives us a chance to apply data mining methods to discover hidden patterns, associations, and anomalies present in this educational data and use this knowledge to improve decision-making processes in e-learning systems.

A large number of automatic tools that work with vast quantities of data have been developed in the last few years. These systems incorporate educational background knowledge that helps to avoid problems during the learning process and improve student performance (Kotsiantis & Pintelas, 2005; Superby, Vandamme, & Meskens, 2006). The main property shared to date by all previous studies is the use of the traditional supervised learning perspective that uses single instances. However, the essential factor when facing this problem is that the information is incomplete because each course has different types and numbers of activities, and each student carries out different numbers of activities, dedicating more or less time to resolve them according to his/her interest and motivation. This peculiarity means that the problem contains disperse information that hinders our ability to foresee student performance. An innovative learning framework, called multiple instance learning, has become quite popular over the last few years. This type of learning provides a more flexible representation that can be adapted to the diverse information in each example. Its strong point is that each example or pattern, called bag in this learning process can be represented by a different number of instances. Thus, general information about each pattern can be stored by means of bag attributes, and specific information about the student's work in each pattern by means of a variable number of instances.

This paper proposes a new representation based on multiple instance learning (MIL) in the context of virtual learning systems to improve the efficiency and effectiveness of classical representation in order to predict student performance. The main idea in this research study is to determine whether data mining technology can solve this problem more efficient using representation based on multiple instances rather than classical representation based on single instances. With this aim in mind, the paper presents both traditional supervised learning representation and a novel proposal to work in a MIL scenario. Algorithms of the most representative paradigms in the two learning frameworks are compared and experimental results show how our representation based on MIL is more effective and obtains more accurate models as well as a more optimized representation, thereby exceeding the shortcomings of classical representation.

The paper is organized as follows. A review of related studies and an introduction to multi-instance learning are covered in

* Corresponding author. Address: Department of Computer Science and Numerical Analysis, University of Cordoba, Campus Universitario Rabanales, Edificio Einstein, Tercera Planta, 14071 Cordoba, Spain. Tel.: +34 957212031; fax: +34 957218360.

E-mail addresses: azafra@uco.es (A. Zafra), cromero@uco.es (C. Romero), sventura@uco.es (S. Ventura).

0957-4174/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2011.05.044

Section 2. Section 3 introduces the problem of predicting student performance and presents two different representations, one based on single-instance learning and the other on multi-instance learning. Section 4 describes the data in the study carried out. Section 5 reports on and compares experimental results for all algorithms tested. Finally, Section 6 summarizes the main contributions of this paper and raises some future research issues.

2. Overview

2.1. Data mining in the learning environment

The most important challenge that higher education faces is to facilitate more efficient, effective and accurate educational processes in the universities. Data mining is the process of extracting patterns from data. As more data is gathered, with the amount of data doubling every three years, data mining is becoming an increasingly important tool to transform this data into information. As a newer sub-field of data mining, educational data mining is considered to be the most suitable technology to resolve its own unique range of research questions and approaches. Educational data mining provides additional insight into the behavior of lecturers, students, managers, and other educational staff and acts as an active automated assistant that helps them to take better decisions about their educational activities (Chen, Hsieh, & Hsu, 2007).

The impressive proliferation of web-based educational systems or e-learning platforms and the great amount of information that generate have originated a considerable number of studies researching the prospect of using data mining in e-learning decision making. Delavari, Beikzadeh, and Shirazi (2004) propose a model with different types of education-related questions and the data-mining techniques appropriate for dealing with large amounts of data collected in academic institutions. An example of a specific case study of clustering students with similar characteristics (such as *self starters* and *high interaction*) is given in Luan, Zhao, and Hayek (2004). Anjewierden, Kollnöffel, and Hulshof (2007) investigate the application of data mining methods to provide learners with real-time adaptive feedback about the nature and patterns of their on-line communication accrued while learning collaboratively. Lazcorreta, Botella, and Fernández-Caballero (2008) propose a new learning method towards automatic personalized recommendation based on the behavior of a single user in accordance with all other users in web-based information systems. Chanchary, Haque, and Khalid (2008) analyze student logs pertaining to a learning management system with data mining and statistical tools in search of relationships between students' access behavior and overall performances. Hwang, Kuo, Yin, and Chuang (2010) study an optimization problem that models objectives and criteria for determining personalized, context-aware, ubiquitous learning paths to maximize learning efficacy for individual students by taking into account the meaningfulness of the learning paths and the number of simultaneous visitors to each learning object. Finally, studies about the prediction of students' marks or their academic success also can be found in this area. Fausett and Elwasif (1994) predict students' grades (classified in five classes: A, B, C, D and E or F) from test scores using neural networks. Martnínez (2001) predicts student academic success (classes that are successful or not) using discriminant function analysis. Minaei-Bidgoli and Punch (2003) classify students by using genetic algorithms to predict their final grade. Superby et al. (2006) predict a student's academic success (classified in low, medium and high risk classes) using different data mining methods. Kotsiantis and Pintelas (2005) predict a student's marks (pass and fail classes) using regression techniques in Hellenic Open University data and Romero, Espejo, Zafra, Romero, and Ventura (2011) show how web usage mining can be applied in e-learning systems in order to predict the marks that university students will obtain in the final exam of a course. They compare the performance of different data mining techniques to solve the problem.

Further information can be found in a survey compiled by Romero and Ventura (2010) which provides a good complete review of the main research studies using data mining techniques, grouped by task, in the e-learning environment.

2.2. Multiple instance learning

MIL is designed to learn a concept that correctly classifies training data and generalizes on unseen data. Although the actual learning process is quite similar to traditional supervised learning, the two approaches differ in the class labels provided from which they learn. In MIL the training patterns are given as bags of instances. Different bags can contain different number of instances and each instance represents a different view of the training pattern attached to it. Following the Dietterich, Lathrop, and Lozano-Perez (1997)'s work, Fig. 1 shows the difference between the two learning frameworks. Both approaches are designed to learn a good approximation to function $f(m)$ that given an input object returns the class to which that object belongs, but the two approaches differ in the class labels provided from which they learn. In a traditional machine learning setting, the input object m is represented by a single feature vector. However, in the multiple instance setting, each input object m may have i various instances denoted $m_1, m_2, \ldots, m_i$ and each one of these instances will be represented by a distinct feature vector.

According to the Dietterich hypothesis is assumed that if the result observed is *positive*, then at least one of the variant instances must have produced that positive result. Furthermore, if the observed result is *negative*, then none of the variant instances could have produced a positive result. The key challenge in MIL is to cope with the ambiguity of not knowing which of the instances in a positive bag are the actual positive examples and which ones are not. In this sense, the multiple instance learning problem can be regarded as a special kind of supervised learning problem where the labeling information is incomplete.

Learning with multi-instances has flourished enormously in the last few years due to the great number of applications that have found a more appropriate form of representation in this learning than in traditional learning. Thus we can find proposals for text categorization (Andrews, Tsochantaridis, & Hofmann, 2002), content-based image retrieval (Herman, Ye, Xu, & Zhang, 2008; Pao, Chuang, Xu, & Fu, 2008), image annotation (Qi & Han, 2007; Yang, Dong, & Fotouhi, 2005), drug activity prediction (Maron & Lozano-Pérez, 1997; Zhou & Zhang, 2007), web index page recommendation (Zafra, Ventura, Romero, & Herrera-Viedma, 2009), semantic video retrieval (Chen & Chen, 2009), video concept detection (Gao & Sun, 2008; Gu, Mei, Tang, Wu, & Hua, 2008) and pedestrian detection (Pang, Huang, & Jiang, 2008). In all cases MIL provides a more natural form of representation that achieves better the results than those obtained by traditional supervised learning.

In order to solve these problems, a great number of new methods os multi-instance learning have been designed. A glance at the literature shows specifically developed algorithms for solving MIL problems, such as Axes Parallel Rectangle (APR) algorithms (Dietterich et al., 1997), Diverse Density (DD) (Maron & Lozano-Pérez, 1997), Diverse Density with Expectation Maximization (EM-DD) (Zhang & Goldman, 2001) and more recently the proposal made by Pao et al. (Pao et al., 2008). Also, it can be found algorithms that are adaptations of popular machine learning paradigms, such as, multi-instance lazy learning algorithms (Wang & Zucker, 2000), multi-instance tree learners (Blockeel, Page, & Srinivasan, 2005; Ruffo,

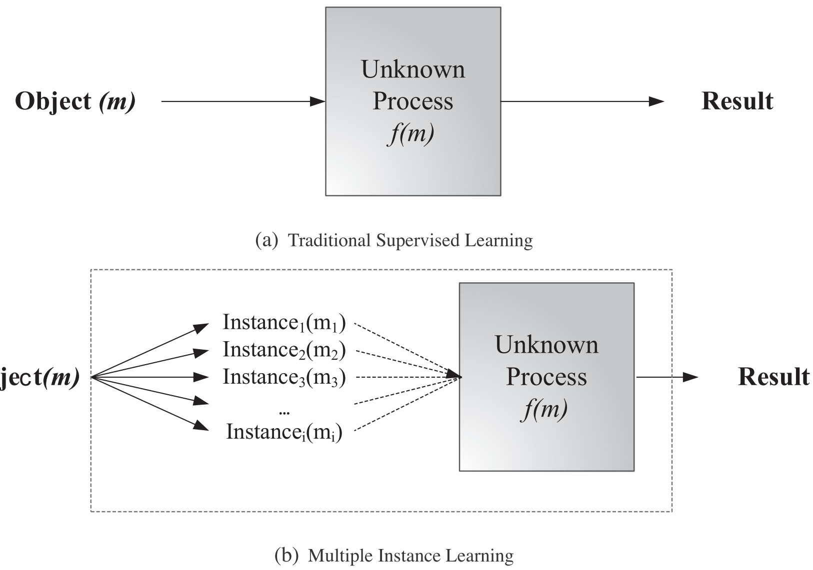


Fig. 1. Differences between traditional supervised learning (with single instances) and multiple instance learning.

2000; Chevaleyre & Zucker, 2001), multi-instance rule inducers (Chevaleyre & Zucker, 2001), multi-instance logistic regression methods (Xu & Frank, 2004), multi-instance neural networks (Chai & Yang, 2007; Zhang & Zhou, 2004, 2006; Zhang & Zhou, 2005), multi-instance kernel methods (Andrews et al., 2002; Chen & Wang, 2004; Chen, Bi, & Wang, 2006; Gartner, Flach, Kowalczyk, & Smola, 2002; Gu et al., 2008; Mangasarian & Wild, 2008), multi-instance ensembles (Zhang & Zhou, 2005; Zhou & Zhang, 2007) and evolutionary algorithms (Zafra et al., 2009; Zafra, Gibaja, & Ventura, in press).

3. Predicting students' performance based on the virtual learning platform

Predicting student performance based on work on the VLE is an issue of growing interest in the learning environment. This problem allows interesting relationships to be obtained that can suggest activities and resources to students and educators that favour and improve both learning and the effective learning process. Thus, it can be determined if all the additional material provided to the students (web-based homework) helps them to better assimilate the concepts and subjects developed in the classroom or if some activities are more worthwhile than others for improving final results.

The problem could be formulated in the following way. A student can do different activities in a course to enable him to assimilate and strengthen the concepts acquired in class. Later, at the end of the course, students face a final exam. A student with a mark over a fixed threshold passes a module, while a student with a mark lower than that threshold fails that lesson or module. With this premise, the problem consists of predicting if the student will pass or fail the module considering the number, time and type of activities that he/she has undertaken during the course. In continuation, the information available is set out in detail and both traditional supervised learning representation and multiple instance learning representation will be described.

3.1. Activities considered on the virtual learning platform

There is an enormous variety of e-learning platforms and most of them have common features and services. Nowadays, one of the most commonly used is Moodle (Modular Object Oriented Developmental Learning Environment), a free learning management system enabling the creation of powerful, flexible and engaging online courses and experiences (Rice, 2006). This system stores a great deal of detailed information about course content, users and usage in a relational database. The present study is based on the information stored about three activities: quizzes, assignments and forums.

- The quizzes are a useful tool for students to test their level of knowledge and review each of the subjects studied. They are a great tool for giving students rapid feedback on their performance and for gauging their comprehension of materials. Feedback on performance is a critical part of a learning environment. Quizzes show what students do and do not understand. A well-designed test can give us critical information about student performance and help students to gauge their own performance and be more successful.
- The assignments are a tool for collecting student work. The assignment module provides an easy way for students to upload digital content to be graded. They can be asked to submit essays, spreadsheets, presentations, web pages, photographs, or small audio or video clips. This activity allows us to keep all the students' work and to verify when the work was submitted.
- Finally, forums are a powerful communication tool. They allow educators and students to communicate with each other at any time, from anywhere with an Internet connection. Students do not have to be logged on at the same time to communicate with the teacher or their classmates. Thus, students can take their time composing replies, which can lead to more thoughtful discussions. Forums create many opportunities to replicate the conversations held in class, to formulate project discussions between groups of students or to bring the best ideas and questions from the forum into the classroom.

There are lots of strategies and studies for the effective use of these activities to improve the learning process. Chiu and Hsiao (2010) explored the differences among online elementary school student groups based on their communication features. Ventouras, Triantis, Tsiakas, and Stergiopoulos (2010) compared the use of multiple-choice questions as an examination method to examinations based on constructed-response questions. Dringus and Ellis (2010) examined the overarching issue of how temporal transitions, specifically duration of message flow, affects momentum or the quality of discussion in an asynchronous forum.

A summary of the information considered for each activity in our study is shown in Table 1.

3.2. Representation of information for working with data mining algorithms

The main idea in this research study is to determine whether data mining technology can solve this problem more efficient using representation based on multiple instances rather than classical representation based on single instances. Table 1 contemplates the available information, developing two representations of the problem. One of them is a classical representation to solve the problem with traditional supervised learning algorithms and the other with multiple instance learning algorithms.

The classical representation considers one student per pattern. Each pattern is represented by only one instance which contains information on all the possible activities that a student can carry out or a course can contain. Thus, the information in each instance represents all activities that the student might do, whether he does them or not. A summary of the attributes (features) that belong to the instances are presented in Fig. 2. In this problem, each student can execute a different number of activities: a hard-working student may do all the activities available and, on the other hand, there can be students who have not done any activities. Moreover, there are some courses that include only a few activities while others consider an enormous variety and number of activities. With this representation, in spite of different information about each student and course, all instances share the same information. This means that most examples have missing attribute values either because the student did not do an activity of a certain type or because that course did not have an available activity of that type.

The representation based on multiple instances that we propose considers each student registered in each course by pattern. Each student is regarded as a bag which represents the work carried out and is composed of one or several instances where each instance represents the different types of work that the student has done. Therefore, each pattern/bag will have as many instances as different types of activities carried out by the student. This representation allows new activities to be added without affecting previous patterns that not considered that activity. Its essential peculiarity lies in the fact that general information about students is stored as bag attributes while the information about each student and course are instance attributes being the number of instances per student variable. This representation fits the problem perfectly because activities that are not very common in the courses could be studied without increasing the general information about each pattern. A summary of the attributes that belong to the bag and the information that corresponds to the instances are presented in Fig. 3. Each instance is divided into 3 attributes: type of activity, number of exercises in that activity and the time devoted to completing that activity. Eight activity types are considered (therefore, a pattern will have a maximum number of eight instances). They are, *ASSIGNMENT_S*, number of assignments that the student has submitted, *ASSIGNMENT* referring to the number of times the student has visited the activity without finally submitting any file. *QUIZ_P*, number of quizzes passed by the student, *QUIZ_F* number of quizzes failed by the student, *QUIZ* referring to the number of times the student has visited a survey without actually answering, *FORUM_POST* number of messages that the student has submitted, *FORUM_READ* number of messages that the student has read and *FORUM* that refers to the number of times the student has seen different forums without entering them. In addition, the bag contains three attributes, that are: student identification, course identification and the final mark obtained by the student in that course.

Fig. 4 shows information available about two students using both representations described. Fig. 4(a) shows the information according to traditional supervised learning; each student is a pattern which contains all the information considered, even though this student may not have actually done any type of activity. Thus, it can be seen that student 1 had many fields empty of content in this representation. Fig. 4(b) and Fig. 4(c) show the information according to MIL representation. Fig. 4(b) shows the representation of student 1, this student has carried out only one activity and we can see the information that belongs to the bag or to the instance in each case. This representation considers the information about the tasks carried out by the student without including any empty fields for activities left undone, since these are simply not instances belonging to this student. This student only did one activity therefore he/she only contains one instance in his/her representation. Fig. 4(c) shows student 2 with respect to the bag and instances. This student has carried out a great number of activities, therefore he/she has a higher number of instances than student 1.

4. A study case in the university of Cordoba

This study employs the students' information from the virtual learning environment at Cordoba University using the Moodle platform (Rice, 2006). This platform stores the tasks carried out by the students during an academic year from September to June, just before the Final Examinations. In order to collect information, each user in the system is assigned an identifier and every time he/she logs onto the framework, all movements within a particular course site are stored with respect to his/her access to content and tools (e.g., calendar, discussion forum, email archive, chat, quizzing, and assignments). In our work only is considered the information about quizzes, forums and assignments. This information is preprocessed to work with data mining algorithms using the two types of representations described in Section 3.2.

The details about the 7 e-Learning courses which have 419 registered students are given in Table 2. We only consider the

Table 1
Information about activities considered in our study.

Activity	Attribute	Description
ASSIGNMENT	NumberAssignment	Number of practices/tasks done by the user in the course.
	TimeAssignment	Total time in seconds that the user has taken in the assignment section.
FORUM	NumberPosts	Number of messages sent by the user in the forum.
	NumberRead	Number of messages read by the user in the forum.
	TimeForum	Total time in seconds that the user has taken in the forum section.
QUIZ	NumberQuiz	Number of quizzes seen by the user.
	NumberQuiz_a	Number of quizzes passed by the user.
	NumberQuiz_s	Number of quizzes failed by the user.
	TimeQuiz	Total time in seconds that the user has taken in the quiz section.

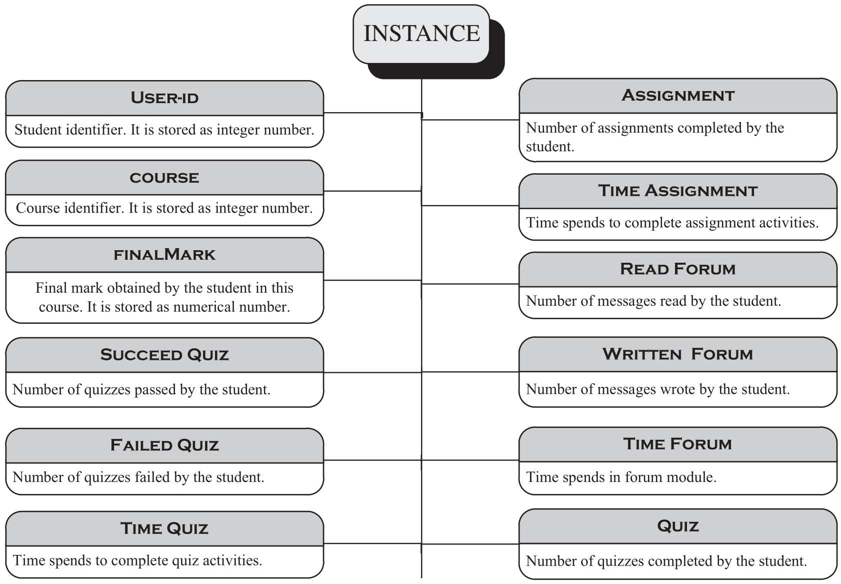


Fig. 2. Information about instances in single instance learning.

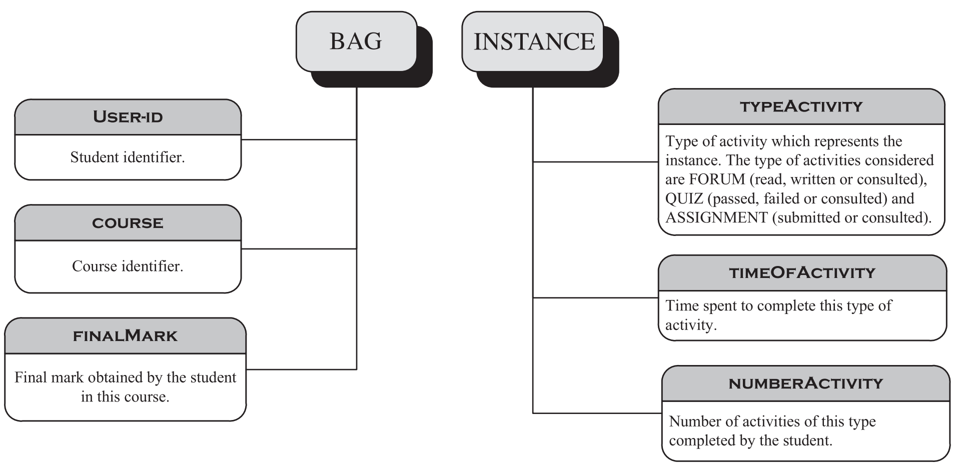


Fig. 3. Information about bags and information about instances in multiple instance learning.

identifier of each student, his/her activity with respect to forums, assignments and quizzes and the final mark obtained in the course.

An analysis of the information collected is specified in the next section together with the format of the representation used.

4.1. Analysis of data

The courses considered have a total of 419 registered students: 244 of them failed the course while 175 passed. In continuation, there is information about each activity considering, on one hand, the number of students who carried out the activity independently of the final mark achieved by the student and, on the other hand, the number of students that did that activity failed or passed, which is considered separately.

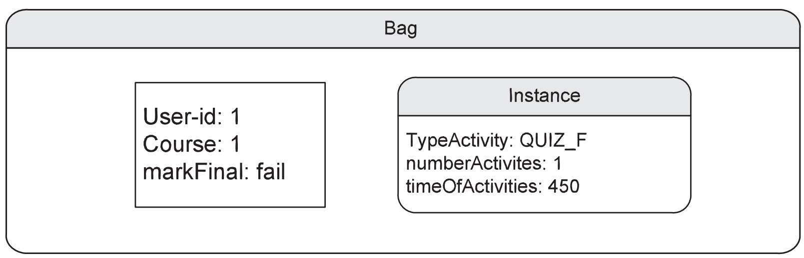
The first activity considered is the assignments. Table 3 shows information about them, the number of tasks is grouped to describe the information. Thus, the categories considered are: not doing any activities, doing between one and five activities, doing between six and ten activities or finally doing between eleven and fifteen activities. It can be seen that most students submit between one and ten tasks. Concretely, students that did not do any activities usually failed, students with more than ten activities submitted usually passed the exam while that students inside of group that did the rank of activities between six and ten, did not show a clearly predictable result.

Fig. 5 represents the percentages of students that carrying out different numbers of activities passed or failed the course. Fig. 5(a) considers students who carried out no assignments, Fig. 5(b) students who carried out from 1 to 5 assignments, Fig. 5(c) students who carried out from 6 to 10 assignments and Fig. 5(d) students who did between 11 and 15 assignments. It is shown that the percentage of students that did not do any assignments and still passed the course is much lower than that of students that failed. Concretely, out of 22% of students that did no assignment, 20% passed versus 80% that failed. On the other hand, students that carried out more than 10 tasks usually passed the course with a higher percentage.

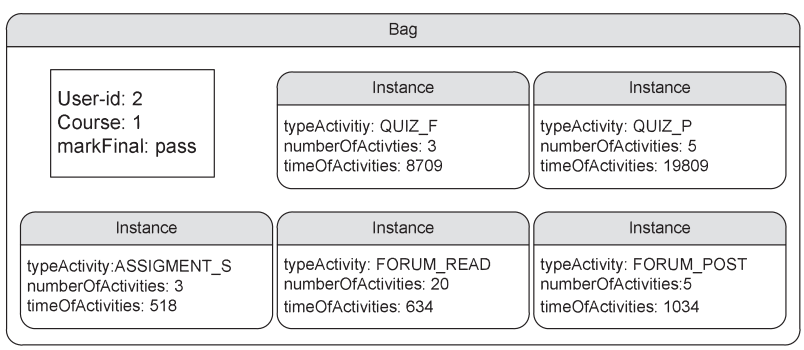
Similar information about quizzes is shown in Table 4 and Fig. 6. Again, students that did not do any quizzes have a higher probability of failing the course. Numerical information shows that out of 55% of students that did not do any activity, 73.36% failed the course and only 26.64% passed it. Whereas if we consider students that did more than six quizzes, 75% passed the course and on observing

Attributes	Student1	Student2
User_id	1	2
Course	1	1
n_assigment	0	3
total_time_assigment	0	8709
n_posts	0	5
n_read	0	20
total_time_forum	0	1034
n_quiz	1	8
n_quiz_a	0	5
n_quiz_s	1	3
total_time_quiz	450	19809
Final_mark	Fail	Pass

(a) Available Information for student1 and student2 to traditional supervised learning



(b) Information about bag and instances for student 1 to MIL



(c) Information about bag and instances for student 2 to MIL

Fig. 4. Information about two students.

Table 2
General information about data sets.

Course identifier	Number of students	Number of assignments	Number of forums	Number of quizzes
ICT-29	118	11	2	0
ICT-46	9	0	3	6
ICT-88	72	12	2	0
ICT-94	66	2	3	31
ICT-110	62	7	9	12
ICT-111	13	19	4	0
ICT-218	79	4	5	30

students with more than 13 quizzes, we can see that only 19% failed the course. However, the probability of students passing or failing a course was similar when they did between one and six quizzes.

Finally, with respect to forum activity, Table 5 shows information about students who read messages on the forum and Table 6, information about the messages written by students. In the first place, we can see that this activity is carried out by a lower number of students, being a priori the lowest representative activity, 349 students do not do any reading in the forums. Figs. 7 and 8 indicate the percentage of students who did this activity. These figures

Table 3
Information about assignments.

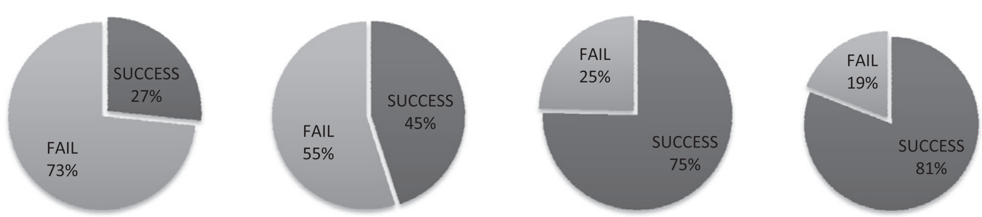
	All students	Percentage (%)	Succeeded student	Failed student
Students do no assignments	92	22	14	78
Students do between 1 and 5 assignments	127	30	52	75
Students do between 6 and 10 assignments	137	33	69	68
Students do between 11 and 15 assignments	63	15	40	23

(a) 0 assignments (b) From 1 to 5 assignments (c) From 6 to 10 assignments (d) From 11 to 15 assignments

Fig. 5. Assignments carried out by students.

Table 4
Information about quizzes.

	All students	Percentage (%)	Succeeded student	Failed student
Students do no quizzes	229	55	61	168
Students do between 1 and 6 quizzes	100	24	45	55
Students do between 7 and 12 quizzes	69	16	52	17
Students do between 13 and 18 quizzes	21	5	17	4



(a) 0 quizzes (b) From 1 to 6 quizzes (c) From 7 to 12 quizzes (d) From 13 to 18 quizzes

Fig. 6. Quizzes carried out by students.

Table 5
Information about reading forums.

	All students	Percentage (%)	Succeeded student	Failed student
Students do not read any messages	349	83	138	211
Students read between 1 and 5 messages	66	16	36	30
Students read between 6 and 10 messages	4	1	1	3

Table 6
Information about writing forums.

	All students	Percentage (%)	Succeeded student	Failed student
Students do not write any messages	296	71	105	191
Students write between 1 and 5 messages	109	26	61	48
Students write between 6 and 10 messages	14	3	9	5

Fig. 7. Messages read by students.

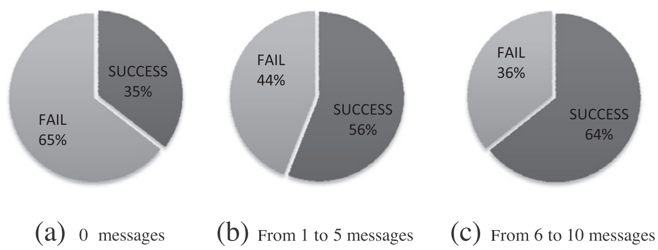


Fig. 8. Messages write by students.

show that 83% of them did not read any messages and 71% of them did not write any messages. Moreover, the fact that some students wrote more messages in the forums was not related to a higher probability of passing the course.

In general, it is difficult to obtain decisive conclusions to predict if a student has a high probability of being successful in a course or not. First, there is no information about the marks achieved by each student in the activities (only about quizzes we have information if the quiz was passed or failed). Thus, students who carried out many activities could not be linked with a high probability of passing the course. Moreover, we only have commented in this section information about activities recorded separately, it would be interesting to study them together. Thus, a student that did not do any quizzes and passed the course could have done a lot of assignments. Also, in order to make this preliminary study easier, there has been a grouping of different numbers of tasks done by students; however, it would be interesting not to summarize the information and to work with the specific number of tasks carried out by each student. Finally, there are courses that do not have any types of activities and therefore no students in these courses can do this type of activity; in these cases this activity is not so relevant as in others.

For this reason, data mining technology becomes more relevant by presenting it as a necessary tool to uncover interesting relationships about students according to the work that each individual has done during the course and the final performance obtained. These tools allow us to obtain patterns from available information and apply them to new students to detect any problems in the learning process and improve the efficiency of the courses.

4.2. Format of the representation

All the information of each student for both representations is exported to a text file using Weka ARFF format (Witten & Frank, 2005). This format has two different sections: the first section contains the header information (the name of the data collection, a list of attributes and their types); the second contains data information and depends on the representation used. Fig. 9 shows an example that represents the two students of Fig. 4 in this format. Fig. 9(a) shows a single-instance representation, each student represents a row where there is a column for each attribute value. Fig. 9(b) shows a multi-instance representation, each student is a row where a student can have several instances separated by $\backslash n$ and only the information available is represented.

5. Experimental section

The purpose of the experimentation is to show that multi-instance representation improves the efficiency and effectiveness of the classical representation to predict students' performance. Thus, first a comparative study is carried out between the most relevant algorithms using a single instance representation and then the different multi-instance proposals are evaluated for solving the same problem. Finally, a comparison is carried out between single instance and multi-instance proposals, and a statistical test is used to determine if there are significant differences between the results achieved by each of the two representations. All experiments are performed using 10-fold stratified cross validation (Wiens, Dale, Boyce, & Kershaw, 2008) and the average values of accuracy, sensitivity and specificity are reported in this section.

5.1. Comparison with supervised learning algorithms

This study has considered different and representative paradigms used in traditional supervised learning that have shown good results in other applications. The main paradigms considered are methods based on trees, rules, neural networks and support vector machines. They are briefly summarised below; more in-depth information can be consulted on the WEKA workbench (Witten & Frank, 2005) where they were designed.

- *Methods based on trees:* decision tree classifiers are used successfully in many diverse areas, these methods generate decision trees that classify instances by sorting them based on feature values. Each node in a decision tree represents a feature in an instance to be classified, and each branch represents a value that the node can assume. Instances are classified starting at the root node and sorted based on their feature values. The algorithms considered in this paradigm are: DecisionStump (Witten & Frank, 2005), RandomForest (Breiman, 2001), RandomTree (Witten & Frank, 2005) and C4.5 (J48) (Quinlan, 1993).
- *Methods based on rules:* these methods generate decision rules which can be understood by the user. Classification rules are composed of an antecedent and a consequent. The consequent represents the class and the antecedent represents a number of disjunctions and conjunctions of conditions which evaluate the coverage of the set of instances. The algorithms considered in this paradigm are: NNge (Martin, 1995), Ridor (Gaines & Compton, 1995), ZeroR (Witten & Frank, 2005) and OneR (Holte, 1993).
- *Methods based on neural networks:* neural networks have emerged as an important tool for classification. The recent numerous research activities in neural classification have established that neural networks are a promising alternative for various conventional classification methods. The advantage of neural networks lies in the following theoretical aspects. First, neural networks are data driven self-adaptive methods in that they can adjust themselves to the data without any explicit functional or distributional specification for the underlying model. Second, they are universal functional approximators in that neural networks can approximate any function with arbitrary accuracy. The algorithms considered in this paradigm are *RBFNetwork* which implements a normalized Gaussian radial basis function network (Witten & Frank, 2005) and *Multilayer Perceptron* which implements a classifier that uses backpropagation to classify instances (Witten & Frank, 2005).

```
@RELATION moodle

@attribute user-id numeric
@attribute course numeric
@attribute n_assigment numeric
@attribute total_time_assigment numeric
@attribute n_posts numeric
@attribute n_read numeric
@attribute total_time_forum numeric
@attribute n_quiz numeric
@attribute n_quiz_a numeric
@attribute n_quiz_s numeric
@attribute total_time_quiz numeric
@attribute final_mark {fail,pass}

@DATA
1,1,0,0,0,0,0,1,0,1,450,FAIL
2,1,3,8709,5,20,1034,8,5,3,19809,PASS
```

(a) Single Instance Representation

```
@RELATION moodle

@attribute user-id numeric
@attribute course numeric
@attribute bag relational
@attribute type_activity {ASSIGNMENT FORUM_P
FORUM_R QUIZ QUIZ_A QUIZ_R}
@attribute number_activity numeric
@attribute time_activity numeric
@end bag
final_mark {fail,pass}

@DATA
1,1,"QUIZ,1,450\nQUIZ_A,1,450\n",FAIL
2,1,"ASSIGNMENT,3,8709\n FORUM_P,1034\n
   FORUM_R,20,1034\n QUIZ_A,5,19809\n
   QUIZ_R,3,19809",PASS
```

(b) Multiple Instance Representation

Fig. 9. Representation of students' performance problem with WEKA format.

- *Methods based on Support Vector Machines:* support vector machines (SVMs) are one of the most popular methodologies for the design of pattern classification systems with sound theoretical foundations and high generalizing performance. SVMs implement the structural risk minimization principle in order to build large-margin classifiers. The algorithm considered in this paradigm is SMO (Keerthi, Shevade, Bhattacharyya, & Murthy, 2001) which implements John Platt's sequential minimal optimization algorithm to train a support vector classifier.
- *Methods based on Naive Bayes:* Naive Bayes (NB) is a simple probabilistic classifier based on applying Bayes' theorem. NB is a mature and well-known model chosen in many medical data classification tasks. The algorithms considered in this paradigm are NaiveBayesSimple (Duda & Hart, 1973), NaiveBayes (George & Langley, 1995), NaiveBayesMultinomial (Mccallum & Nigam, 1998) and NaiveBayesUpdateable (George & Langley, 1995).

To compare the different proposals we consider three very common measurements used in classification: accuracy, sensitivity and specificity (Bojarczuk, Lopes, & Freitas, 2000; Tan, Tay, Lee, & Heng, 2002). Accuracy measures the proportion of correct predictions out of the total number of cases evaluated, sensitivity measures the proportion of cases correctly identified as meeting a certain condition and specificity is the proportion of cases correctly identified as not meeting a certain condition. Table 7 reports the average results for the measurements considered for all algorithms employed in this study. In traditional supervised learning, a support vector machine algorithm obtains the best accuracy with a 69.76% of students correctly classified. In general, sensitivity values are optimized better at the expense of a decrease in specific values. This demonstrates that the models do not correctly classify negative examples because they identify students as potentially passing the course when actually they finally fail it. In truth, there are cases in this problem that are especially difficult to classify because there are hard-working students that do not pass the course in the end and yet other students that do not complete any activities and are able to pass (although this is not very common cases).

The different paradigms used in the experimentations produce similar results; all the paradigms considered contain some algorithms with results similar to the best proposal, with accuracy values ranging between 64.76% and 69.76%.

5.2. Comparison with multiple instance learning

This study has considered the most representative paradigms used successfully in other applications of multiple instance learning.

Table 7
Experimental results of methods based on traditional supervised learning (using single instance).

Algorithm	Accuracy	Sensitivity	Specificity
Algorithms based on trees			
DecisionStump	0.6690	0.8889	0.3651
RandomForest	0.6667	0.7573	0.5426
RandomTree	0.6476	0.6996	0.5755
J48	0.6857	0.7950	0.5345
Algorithms based on rules			
NNge	0.6952	0.7329	0.6434
Ridor	0.6810	0.8648	0.4310
OneR	0.6476	0.7665	0.4835
ZeroR	0.5810	1.0000	0.0000
Algorithms based on Naive Bayes			
NaiveBayes	0.6857	0.8232	0.4944
NaiveBayesMultinomial	0.6929	0.7662	0.5918
NaiveBayesSimple	0.6810	0.8232	0.4832
NaiveBayesUpdateable	0.6857	0.8232	0.4944
Algorithms based on neural networks			
RBFNetwork	0.6929	0.8227	0.5114
Multilayer Perceptron	0.6881	0.7983	0.5363
Algorithms based on SVMs			
SMO	0.6976	0.8842	0.4374

The main paradigms considered are methods based on diverse density, logistic regression, support vector machines, distance, decision tree, rules and Naive Bayes. They are briefly summarized in continuation. More in-depth information can be consulted on the WEKA workbench (Witten & Frank, 2005) where they were designed.

- *Methods based on Diverse Density:* diverse density (DD), proposed by Maron and Lozano-Pérez (1997), is perhaps the best known framework for MI learning. Given a set of positive and negative bags, the idea behind this approach is to learn a concept that is close to at least one instance in each positive bag, but far from all instances in all the negative bags. Thus, the concept must describe a region of instance space that is dense in instances from positive bags, and is also diverse in that it describes every positive bag. The algorithms considered in this paradigm are: MIDD (Maron & Lozano-Pérez, 1997), MIEMDD (Zhang & Goldman, 2001) and MDD (Witten & Frank, 2005).
- *Methods based on Logistic Regression:* logistic regression is a popular machine learning method in standard single instance learning. The algorithm considered in this paradigm is MILR (Xu,

2003) which adapts standard SI logistic regression to MI learning by assuming a standard single instance logistic model at the instance level and using its class probabilities to compute bag-level class probabilities using the noisy-or model employed by DD.

- *Methods based on Support Vector Machines:* there is an extensive number of ways to adapt this approach to the MIL framework whose results show good performance in different applications. The algorithm considered in this paradigm is the SMO algorithm (Keerthi et al., 2001) for SVM learning in conjunction with an MI kernel (Gartner et al., 2002).
- *Distance-based Approaches:* the k-nearest neighbor (k-NN) in a MIL framework was introduced by Wang and Zucker (2000). The main difference between the different proposals for nearest neighbor algorithms lies in the definition of the distance metrics used to measure the distance between bags. Two schemes extensively employed are the minimal Hausdorff distance and the Kullback–Leibler distance. The algorithms consider in this paradigm are: CitationKNN (Wang & Zucker, 2000) and MIOptimalBall (Auer & Ortner, 2004).
- *Methods based on rules:* these methods employ different classic algorithms of traditional supervised learning adapted to multiple instance learning. There are two possible ways to adapt them: the MIWrapper (Witten & Frank, 2005) method, which assigns the class label of a bag to all its instances and then trains a single instance algorithm on the resulting data, and MISimple (Witten & Frank, 2005) computing summary statistics for a bag to form a single instance from it. The algorithms considered in this paradigm are: PART (MIWrapper), PART (MISimple) and combinations of the different proposals using rule based systems: AdaBoost & PART (MISimple), Bagging & PART (MIWrapper) and AdaBoost & PART (MIWrapper).
- *Methods based on decision trees:* these methods are inspired by AdaBoost which builds a series of weak classifiers using a single instance learner based on appropriately re-weighted versions of the input data, and all instances receive their bags' labels (Xu & Frank, 2004). The algorithms considered in this paradigm are: DecisionStump (Witten & Frank, 2005) and RepTree (Witten & Frank, 2005).
- *Methods based on naive Bayes:* this method is adapted to multiple instance learning using the procedure based on MIWrapper that has been commented on previously. The algorithm considered is Naive Bayes (Witten & Frank, 2005).

The average results of accuracy, sensitivity and specificity are reported in Table 8. On observing the results of the different paradigms, there is a similar situation to that mentioned in the case of the single-instance where the optimization of the sensitivity measurement is at the expense of a decrease in the specificity value. Again, results show an incorrect forecast about which students will pass the subject. This fact could suggest the need for a new method that balances the two measurements. The main problem with classification in this application is the different number and type of activities in each course that makes it much more costly to determine any general relationships among them.

With respect to the different paradigms used, the methods based on rules (PART) or a combination of this method with other proposals obtain the best results for this learning. Concretely, the best algorithm (PART algorithm) reaches a accuracy percentage of 73.57%. Nevertheless, the results obtained by different paradigms are similar and all of them fluctuate between 65.71% and 73.57%.

It can be seen that some paradigms used in single instance and multiple instance representation are similar the two proposals. However, these methods are not directly comparable because they are not exactly the same implementations. Therefore, it is necessary to carry out a statistical test to evaluate the results obtained by each

Table 8
Experimental results of methods based on multiple instance learning.

Algorithm	Accuracy	Sensitivity	Specificity
Algorithms based on rules			
PART (MISimple)	0.7357	0.8387	0.5920
AdaBoostM1 & PART (MISimple)	0.7262	0.8187	0.5992
Bagging & PART (MIWrapper)	0.7167	0.7733	0.6361
AdaBoostM1 & PART (MIWrapper)	0.7071	0.7735	0.6136
PART (MIWrapper)	0.7024	0.7857	0.5842
Algorithms based on support vector machines			
SMO	0.6810	0.8644	0.4270
Algorithms based on Naive Bayes			
NaiveBayes	0.6786	0.8515	0.4371
Algorithms based on decision tree			
DecisionStump	0.6762	0.7820	0.5277
RepTree	0.6595	0.7127	0.5866
Algorithms based on logistic regression			
MILR	0.6952	0.8183	0.5218
Algorithms based on diverse density			
MIDD	0.6976	0.8552	0.4783
MIEMDD	0.6762	0.8549	0.4250
MDD	0.6571	0.7864	0.4757

representation, considering the different algorithms evaluated in order to draw a final conclusion (the next section presents the comparison). At first sight, it is possible to see by means of result showed in previous tables that the algorithms with multiple instance representation generally yield higher accuracy percentages.

5.3. Comparison between single and multiple instance learning

In this section, a statistical study is carried out with the purpose of demonstrating if the representation with multiple instances is more appropriate than single instance representation to solve the problem of predicting students' performance by the work they have done on a virtual learning platform. Both representations have considered fifteen of the most representative algorithms of various paradigms developed to date including systems based on decision tree, logistic regression, neural network, support vector machine and others. Their results concerning accuracy, sensitivity and specificity values have been shown in the previous sections. The idea is to check if, in general, there are significant differences between the accuracy values obtained by different algorithms using multiple instance learning representation as compared to traditional supervised learning representation with single instance. The Wilcoxon rank sum test is used to look for differences between the accuracy values obtained by both representations. This test is a non-parametric one recommended in Demsar's study (Demsar, 2006) which allows us to address the question of whether there are significant differences between the accuracy values obtained by algorithms used in each of the two representations. To do this, the null hypothesis of this test maintains that there are not significant differences between the accuracy values of each representation while the alternative hypothesis assures that there are. This test evaluates the differences in performance of the two representations evaluating the results obtained by the algorithms in each representation. The Table 9 shows mean ranks and the sum of ranks for each representation. The scores are ranked from lowest to highest. Therefore, we can see that algorithms using single instance representation result in a lower mean rank than algorithms using multiple instance representation. This information can be used to ascertain a priori that multiple instance representation has a greater number of algorithms that obtains a higher accuracy value.

Table 9
Sum of ranks and mean rank of the two representations.

	Mean rank	Sum of ranks
Multiple instance representation	18.67	280
Single instance representation	12.33	185

Table 10
Wilcoxon rank-sum test results.

	Wilcoxon W	Z-score	Asymp sig (2-tailed) (*p*-value)
Accuracy for both representations	185	−1.973	0.048

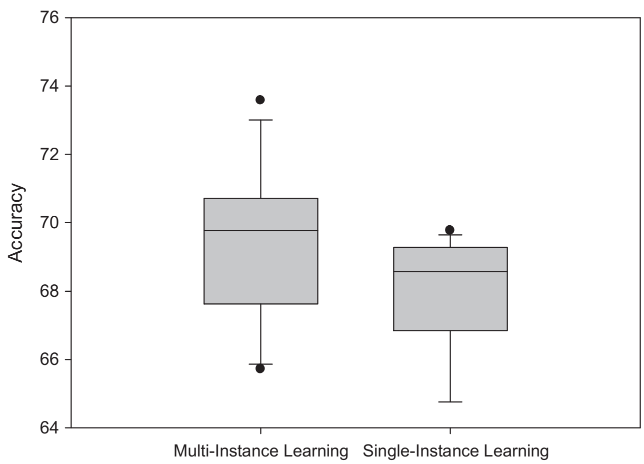


Fig. 10. Box-and-whisker diagram for accuracy measurement.

Table 10 contains the results of Wilcoxon statistical test and its corresponding z-score. Moreover, the significant value of the test is shown which gives a two-tailed probability to determine if we accept or reject the null hypothesis. According to this value, the results are highly significant (p-value < 0.05). Therefore, at a 95% level of confidence we reject the null hypothesis and determine that there are significant differences between the results obtained by these two representations. Consequently, multi-instance representation has significantly higher accuracy values than single-instance representation. This conclusion is reached by noting that for multi-instance representation scores, the mean rank is higher in the algorithms using multi-instance representation (at a value of 18.67) than in single-instance representation (at a value of 12.33).

Fig. 10 shows a box-and-whisker plot with the results on the accuracy obtained by single and multiple instance methods. This graph reveals visually how better results are obtained in multi-instance representation which achieves better values at the extremes as well as in the lower, medium and upper quartiles.

6. Conclusion and future work

This paper describes a first proposal using a MIL representation for the problem of predicting a student's performance based on his/her work in VLE. This problem has been tackled until now using learning based on single instance representation. This representation presents a great scattering of data because there is no homogeneity between the tasks carried out by different students and the activities contained in different courses. It is more than proven that data mining algorithms do not perform appropriately in this scenario (Batista & Monard, 2003; Grzymala-Busse, Hippe, Rzasa, & Vasudevan, 2009; Wang & Wang, 2007; Scheffer, 2002). MIL provides a more flexible representation adapted to the information available at each moment, thus each student is represented by a variable number of instances depending on his/her work. A hard-working student will have a higher number of instances and a lazy-student will have a lower number of instances, eliminating the missing values and overcoming the shortcomings of traditional representation. To check the effectiveness of representation based on multiple instances, the most representative paradigm both of traditional supervised learning and of multiple-instance learning are applied to solve this problem.

Computational experiments show that when the problem is regarded as a traditional supervised learning problem, the results obtained are lower than when the problem is regarded as a multi-instance problem. Concretely, the Wilcoxon rank sum test determines with a confidence of 95% that algorithms using MIL representation achieve significantly better solutions.

Although the results are interesting, there are still quite a few considerations that could surely add even more value to the results obtained. Thus, it would be interesting to design a method that achieves a trade off between the different metrics and allows representative information to be obtained. Another interesting issue consists of expanding the problem to predict students' grades by classifying them in different classes according to the work done on the virtual learning system (and not only considering if a student passes a course or not).

References

Andrews, S., Tsochantaridis, I., & Hofmann, T. (2002). Support vector machines for multiple-instance learning. In *NIPS'02: Proceedings of neural information processing system. Vancouver, Canada* (pp. 561–568).

Anjewierden, A., Kollnöffel, B., & Hulshof, C. (2007). Towards educational data mining: Using data mining methods for automated chat analysis to understand and support inquiry learning processes. In *ADML'07: International workshop on applying data mining in e-learning. Crete, Greece* (pp. 27–36).

Auer, P., & Ortner, R. (2004). A boosting approach to multiple instance learning. In *ECML'04: Proceedings of the fifth European conference on machine learning. Lecture Notes in Computer Science. Vol. 3201. Pisa, Italy* (pp. 63–74).

Batista, G., & Monard, M. (2003). An analysis of four missing data treatment methods for supervised learning. *Applied Artificial Intelligence, 17*, 519–533.

Blockeel, H., Page, D., & Srinivasan, A. (2005). *Multi-instance tree learning. ICML '05: Proceedings of the 22nd international conference on Machine learning*. New York, NY, USA: ACM Press, pp. 57–64.

Bojarczuk, C. C., Lopes, H. S., & Freitas, A. A. (2000). Genetic programming for knowledge discovery in chest-pain diagnosis. *IEEE Engineering in Medicine and Biology Magazine, 19*(4), 38–44.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

Busato, V., Prins, F., Elshout, J., & Hamaker, C. (2000). Intellectual ability, learning style, personality, achievement motivation and academic success of psychology students in higher education. *Personality and Individual Differences, 29*, 1057–1068.

Chai, Y. -M., & Yang, Z. -W. (2007). A multi-instance learning algorithm based on normalized radial basis function network. In *ISSN'07: Proceedings of the fourth international symposium on neural networks. Lecture notes in computer science. Vol. 4491. Nanjing, China* (pp. 1162–1172).

Chanchary, F., Haque, I., & Khalid, S. (2008). Web usage mining to evaluate the transfer of learning in a web-based learning environment. In *WKDD'08: International workshop on knowledge discovery and data mining. Adelaide, Australia* (pp. 249–253).

Chen, C., Hsieh, Y., & Hsu, S. (2007). Mining learner profile utilizing association rule for web-based learning diagnosis. *Expert Systems with Applications, 33*(1). 6-2.

Chen, C.-M., & Chen, M.-C. (2009). Mobile formative assessment tool based on data mining techniques for supporting web-based learning. *Computers and Education, 52*(1), 256–273.

Chen, Y., Bi, J., & Wang, J. (2006). Miles: Multiple-instance learning via embedded instance selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 28*(12), 1931–1947.

Chen, Y., & Wang, J. Z. (2004). Image categorization by learning and reasoning with regions. *Journal of Machine Learning Research, 5*, 913–939.

Chevaleyre, Y. -Z., & Zucker, J.-D. (2001). Solving multiple-instance and multiple-part learning problems with decision trees and decision rules. Application to the mutagenesis problem. In *AI'01: Proceedings of the 14th of the Canadian society for computational studies of intelligence. LNCS Ottawa, Canada* (pp. 204–214).

Chidolue, M. (2001). The relationship between teacher characteristics, learning environment and student achievement and attitude. *Studies in Educational Evaluation, 22*(3), 263–274.

Chiu, C.-H., & Hsiao, H.-F. (2010). Group differences in computer supported collaborative learning: Evidence from patterns of taiwanese students' online communication. *Computers and Education, 54*(2), 427–435.

Delavari, N., Beikzadeh, M., & Shirazi, M. (2004). A new model for using data mining in higher educational system. In *ITHET'04: Proceedings of fifth international conference on information technology based higher education and training. Istanbul, Turkey*.

Demsar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research, 17*, 1–30.

Dietterich, T. G., Lathrop, R. H., & Lozano-Perez, T. (1997). Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence, 89*(1-2), 31–71.

Dringus, L. P., & Ellis, T. (2010). Temporal transitions in participation flow in an asynchronous discussion forum. *Computers and Education, 54*(2), 340–349.

Duda, R., & Hart, P. (1973). *Pattern Classification and Scene Analysis*. New York: Wiley.

Fausett, L., & Elwasif, W. (1994). Predicting performance from test scores using backpropagation and counterpropagation. In *WCCI'94: IEEE world congress on computational intelligence. Washington, USA* (pp. 3398–3402).

Gaines, B., & Compton, P. (1995). Induction of ripple-down rules applied to modeling large databases. *Journal of Intelligence Information System, 5*(3), 211–228.

Gao, S., & Sun, Q. (2008). Exploiting generalized discriminative multiple instance learning for multimedia semantic concept detection. *Pattern Recognition, 41*(10), 3214–3223.

Gartner, T., Flach, P.A., Kowalczyk, A., & Smola, A.J. (2002). Multi-instance kernels. In *ICML'02: Proceedings of the 19th international conference on machine learning. Morgan Kaufmann, Sydney, Australia* (pp. 179–186).

George, H., & Langley, P. (1995). Estimating continuous distributions in bayesian classifiers. In *UAI'95: Proceedings of the 11th conference on uncertainty in artificial intelligence. San Mateo* (pp. 338–345).

Grzymala-Busse, J., Hippe, Z., Rzasa, W., & Vasudevan, S., 2009. A valued tolerance approach to missing attribute values in data mining. In *HSI'09: Proceedings of the second conference on human system interactions. Vol. 3. Catania, Italy* (pp. 220–227).

Gu, Z., Mei, T., Tang, J., Wu, X., & Hua, X. (2008). MILC2: A multi-layer multi-instance learning approach to video concept detection. In *MMM'08: Proceedings of the 14th international conference of multimedia modeling. Kyoto, Japan* (pp. 24–34).

Herman, G., Ye, G., Xu, J., & Zhang, B. (2008). Region-based image categorization with reduced feature set. In *Proceedings of the 10th IEEE workshop on multimedia signal processing. Cairns, Qld* (pp. 586–591).

Holte, R. (1993). Very simple classification rules perform well on most commonly used datasets. *Machine Learning, 11*, 63–91.

Hwang, G.-J., Kuo, F.-R., Yin, P.-Y., & Chuang, K.-H. (2010). A heuristic algorithm for planning personalized learning paths for context-aware ubiquitous learning. *Computers and Education, 54*(2), 404–415.

Keerthi, S., Shevade, S., Bhattacharyya, C., & Murthy, K. (2001). Improvements to platt's smo algorithm for svm classifier design. *Neural Computation, 13*(3), 637–649.

Kotsiantis, S., & Pintelas, P. (2005). Predicting students marks in hellenic open university. In *ICALT'05: The fifth international conference on advanced learning technologies. Kaohsiung, Taiwan* (pp. 664–668).

Lazcorreta, E., Botella, F., & Fernández-Caballero, A. (2008). Towards personalized recommendation by two-step modified apriori data mining algorithm. *Expert Systems with Applications, 35*(3), 1422–1429.

Luan, J., Zhao, C.-M., & Hayek, J. (2004). Exploring a new frontier in higher education research: A case study analysis of using data mining techniques to create NSSE institutional typology. In: California Association for Institutional Research. California, USA, pp. 17–19.

Mangasarian, O. L., & Wild, E. W. (2008). Multiple instance classification via successive linear programming. *Journal of Optimization Theory and Applications, 137*(3), 555–568.

Maron, O., & Lozano-Pérez, T. (1997). A framework for multiple-instance learning. In *NIPS'97: Proceedings of neural information processing system 10. Denver, Colorado, USA* (pp. 570–576).

Martin, B. (1995). *Instance-based learning: Nearest neighbor with generalization*. Ph.D. thesis. Department of Computer Science. University of Waikato, Hamilton, New Zealand.

Martnínez, D. (2001). Predicting student outcomes using discriminant function analysis. In *Annual Meeting of the Research and Planning Group. California, USA* (pp. 163–173).

Mccallum, A., & Nigam, K. (1998). A comparison of event models for naive bayes text classification. In *AAAI'98: Workshop on learning for text categorization. Madison, Wisconsin* (pp. 41–48).

Minaei-Bidgoli, B., & Punch, W. (2003). Using genetic algorithms for data mining optimization in an educational web-based system. *Genetic and Evolutionary Computation, 2*, 2252–2263.

Minnaert, A., & Janssen, P. (1999). The additive effect of regulatory activities on top of intelligence in relation to academic performance in higher education. *Learning and Instruction, 9*, 77–91.

Nagi, K., & Suesawaluk, P. (2008). Research analysis of moodle reports to gauge the level of interactivity in elearning courses at assumption university. In *ICCN'08: Proceedings of the 18th international conference on computer and communication engineering. Kuala Lumpur, Malaysia* (pp. 772–776).

Pang, J., Huang, Q., & Jiang, S. (2008). Multiple instance boost using graph embedding based decision stump for pedestrian detection. In *ECCV'08: Proceedings of the 10th European conference on computer vision. LNCS Vol. 5305* (pp. 541–552). Springer-Verlag Berlin, Heidelberg.

Pao, H. T., Chuang, S. C., Xu, Y. Y., & Fu, H. . (2008). An EM based multiple instance learning method for image classification. *Expert Systems with Applications, 35*(3), 1468–1472.

Qi, X., & Han, Y. (2007). Incorporating multiple SVMs for automatic image annotation. *Pattern Recognition, 40*(2), 728–741.

Quinlan, R. (1993). *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann Publishers.

Rice, W. H. (2006). Moodle e-learning course development. Pack Publishing.

Romero, C., Espejo, P. G., Zafra, A., Romero, J. R., & Ventura, S. (2011). Web usage mining for predicting final marks of moodle students. *Computer Applications in Engineering Education*. doi: 10.1002/cae.20456.

Romero, C., & Ventura, S. (2010). Educational data mining: A review of the state-of-the-art. *IEEE Transaction on Systems, Man and Cybernetics, part C: Applications and Reviews, 40*(6), 610–618.

Ruffo, G. (2000). *Learning single and multiple instance decision tree for computer security applications*. Ph.D. thesis, Department of Computer Science. University of Turin, Torino, Italy.

Scheffer, J. (2002). Dealing with missing data. *Research Letters in the Information and Mathematical Sciences, 3*, 153–160.

Superby, J., Vandamme, J., & Meskens, N. (2006). Determination of factors influencing the achievement of the first-year university students using data mining methods. In *EDM'06: Workshop on Educational Data Mining. Hong Kong, China* (pp. 37–44).

Tan, K.C., Tay, A., Lee, T. H., & Heng, C. M. (2002). Mining multiple comprehensible classification rules using genetic programming. In *CEC'02: Proceedings of the congress on evolutionary computation. Vol. 2. Honolulu, HI, USA* (pp. 1302–1307).

Ventouras, E., Triantis, D., Tsiakas, P., & Stergiopoulos, C. (2010). Comparison of examination methods based on multiple-choice questions and constructed-response questions using personal computers. *Computers and Education, 54*(2), 455–461.

Wang, H., & Wang, S. (2007). A data mining approach to assessing the extent of damage of missing values in survey. *Business Intelligence and Data Mining archive, 2*(3), 249–260.

Wang, J., & Zucker, J. -D. (2000). Solving the multiple-instance problem: A lazy learning approach. In *ICML'00: Proceedings of the 17th international conference on machine learning. Standord, CA, USA* (pp. 1119–1126).

Wiens, T. S., Dale, B. C., Boyce, M. S., & Kershaw, P. G. (2008). Three way k-fold cross-validation of resource selection functions. *Ecological Modelling, 212*(3-4), 244–255.

Witten, I., & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques* (2nd ed.). San Francisco: Morgan Kaufmann.

Xu, X. (2003). *Statistical learning in multiple instance problems*. Ph.D. thesis, Department of Computer Science. University of Waikato, Hamilton, New Zealand.

Xu, X., & Frank, E. (2004). Logistic regression and boosting for labeled bags of instances. In *PAKDD'04: Proceedings of the eigth conference of Pacific-Asia. Vol. 3056 of LNCS. Sydney, Australia* (pp. 272–281).

Yang, C., Dong, M., & Fotouhi, F. (2005). Region based image annotation through multiple-instance learning. In *Multimedia'05: Proceedings of the 13th annual ACM international conference on multimedia. New York, USA* (pp. 435–438).

Zafra, A., Gibaja, E., & Ventura, S. (2011). Multi-instance learning with multi-objective genetic programming for web mining. *Applied Soft Computing, 11*(1), 93–102.

Zafra, A., Ventura, S., Romero, C., & Herrera-Viedma, E. (2009). Multi-instance genetic programming for web index recommendation. *Expert Systems with Applications, 36*, 11470–11479.

Zhang, M.-L., & Zhou, Z.-H. (2004). Improve multi-instance neural networks through feature selection. *Neural Processing Letters, 19*(1), 1–10.

Zhang, M.-L., & Zhou, Z.-H. (2005). Ensembles of multi-instance neural networks. *IIP'04: International conference on intelligent information processing II. International federation for information processing* (Vol. 163). Beijing, China: Springer-Verlag, pp. 471–474.

Zhang, M.-L., & Zhou, Z.-H. (2006). Adapting RBF neural networks to multi-instance learning. *Neural Processing Letters, 23*(1), 1–26.

Zhang, Q., & Goldman, S. (2001). EM-DD: An improved multiple-instance learning technique. In *NIPS'01: Proceedings of neural information processing system 14. Vancouver, Canada* (pp. 1073–1080).

Zhou, Z.-H., & Zhang, M.-L. (2007). Solving multi-instance problems with classifier ensemble based on constructive clustering. *Knowledge and Information Systems, 11*(2), 155–170.