Predicting students' final performance from participation in on-line discussion forums

Cristóbal Romero*, Manuel-Ignacio López, Jose-María Luna, Sebastián Ventura

Department of Computer Sciences and Numerical Analysis, University of Córdoba, 14071 Córdoba, Spain

ABSTRACT

On-line discussion forums constitute communities of people learning from each other, which not only inform the students about their peers' doubts and problems but can also inform instructors about their students' knowledge of the course contents. In fact, nowadays there is increasing interest in the use of discussion forums as an indicator of student performance. In this respect, this paper proposes the use of different data mining approaches for improving prediction of students' final performance starting from participation indicators in both quantitative, qualitative and social network forums. Our objective is to determine how the selection of instances and attributes, the use of different classification algorithms and the date when data is gathered affect the accuracy and comprehensibility of the prediction. A new Moodle's module for gathering forum indicators was developed and different executions were carried out using real data from 114 university students during a first-year course in computer science. A representative set of traditional classification algorithms have been used and compared versus classification via clustering algorithms for predicting whether students will pass or fail the course on the basis of data about their forum usage. The results obtained indicate the suitability of performing both a final prediction at the end of the course and an early prediction before the end of the course; of applying clustering plus class association rules mining instead of traditional classification for obtaining highly interpretable student performance models; and of using a subset of attributes instead of all available attributes, and not all forum messages but only students' messages with content related to the subject of the course for improving classification accuracy.

1. Introduction

Internet forums, web forums, discussion boards, message boards, discussion groups, or bulletin boards are on-line discussion sites where people can hold conversations in the form of posted messages. They are one of the most popular tools for supporting students' communication and collaboration in web-based teaching-learning environments and one of the best ways of sharing ideas, posting problems, commenting on posts by other students, and obtaining feedback (Raghavan, Catherine, Ikbal, Kambhatla, & Majumdar, 2010). An on-line discussion forum, being an asynchronous tool, allows students to maintain discussions related with their learning process at any time and place, to have more time to structure and organize their thoughts, and to communicate simultaneously or even participate in multiple discussions at the same time (Cobo et al., 2011). Forums also enable student interactions and social interactions to occur outside a traditional classroom, and students, especially those of an introvert character, feel less threatened about expressing their views or asking questions (Cheng, Paré, Collimore, & Joordens, 2011). Forums play an important role in students' collaborative learning, in which two main actions are carried by students: writing and reading (Rau, Gao, & Wu, 2008). The students can adopt different attitudes in terms of these two main actions, thus defining different behaviour profiles such as active learners, lurkers, etc. and different type of discussions such as announcement, questioning, clarification, interpretation, conflict and assertion (Rabbany et al., 2011). On the other hand, a synchronous forum allows instructors to analyze student participation in forums, to develop monitoring and even assessment tasks (Lopez, Luna, Romero, & Ventura, 2012). However, the instructor's view of the output of a threaded forum can normally be limited to reviewing a

---

* Corresponding author. Fax: +34 957 218630.
  E-mail address: cromero@uco.es (C. Romero).

transcript or printed version of the written dialogue produced by students. With hundreds of contributions to review in an entire on-line forum, the instructor lacks a comprehensive view of the information embedded in the transcript. In fact, even in on-line courses with a small number of students, there could be thousands of messages generated in a few months using these types of forums. The instructor is faced with the difficulty of interpreting and evaluating the learning and quality of the participation reflected in the students' contributions. Evaluating the participation of students in such a case is very difficult, considering that current e-learning environments do not provide many indicators or information regarding the structure of interactions between students (Dringus & Ellis, 2005). A solution to this problem is the use of quantitative and qualitative participation indicators in forums, such as the number of contributions and the quality of the contributions, together with Data Mining (DM) for discovering and building alternative representation and models for the data underlying discussion forums. DM is a knowledge discovery process to extract information from a dataset and to transform it into an understandable structure for further use (Klosgen & Zytkow, 2002) and is known as EDM (Educational Data Mining) when it uses the unique kinds of data that come from an educational setting (Romero & Ventura, 2013). EDM is unobtrusive or archival research that uses educational data collected in the past to reach a finding and thus the researcher does not interact directly with subjects of the study (Denning, 2008). In this paper, we use an authentic educational setting – a first-year course in Computer Science with 114 university undergraduates – in which, in parallel to the traditional classroom, all students participate in an on-line discussion forum. Therefore, there is no pre-post assessment before-after intervention or treatment, because there is no such intervention or treatment (McMillan & Schumacher, 2006). We propose to use different data mining approaches to predict whether students will pass the course or not, starting from participation in an on-line discussion forum. We have also developed a new Moodle's module for gathering all the forum indicators. The objective of our study is to find a reply to the following research questions:

1. What DM techniques are best for predicting student performance starting from participation in on-line forums? Traditionally, supervised classification algorithms have been used to do this task. However, other unsupervised DM algorithms, such as clustering algorithms, can also be used for the task and provide similar accuracy and comprehensible models.
2. What attributes are the best predictors? Are all attributes or available variables about forum usage data relevant or can similar accuracy be obtained by using just one subgroup of selected attributes?
3. What messages are the best predictors? Are all instances or available messages relevant or perhaps some messages may be irrelevant to the contents of the course and a better prediction can be obtained by using only those messages related to the course.
4. It is possible to make an early prediction? Is it necessary to wait until the end of the course in order to obtain good prediction accuracy or it is possible to do it before the end of the course?

The paper is arranged in the following way. Firstly, the related background is described. Next, the proposed approach is presented together with the material and methods used. Then, the experimental results and discussion are described. Finally, conclusions and future work are dealt with.

2. Background

Predicting students' performance is one of the oldest and most useful applications of EDM and its goal is to estimate the unknown value of students' performance, knowledge, score or mark from other information, aspects or behaviour of those students (Romero & Ventura, 2013). This is a difficult problem to solve due to the large number of factors or characteristics that can bear influence on students' performance, such as demographic, cultural, social, or family factors, socio-economic status, psychological profile, previous schooling, prior academic performance, interactions between student and the faculty, etc. (Araque, Roldan, & Salguero, 2009). Different techniques have been applied for predicting students' performance depending on the variable to be predicted (Hämäläinen & Vinni, 2011): classification (when the predicted variable is a categorical value), regression (when the predicted variable is a continuous value) or density estimation (when the predicted value is a probability density function). It is also important to notice that most of the current research on the application of EDM for predicting student performance has been applied primarily to the specific case of higher education or university students (Kotsiantis, Patriarcheas, & Xenos, 2010) and more specifically to web-based education or e-learning (Romero, Espejo, Romero, &Ventura, 2013). This is mainly due to the fact that the use of Learning Management System (LMS) such as Moodle, Ilias, Claroline, Atutor, Blackboard, WebCT, TopClass, etc. is increasing exponentially due to its ability to create powerful, flexible and engaging on-line courses and experiences. These systems accumulate a vast amount of information in relation to visits and times, resources viewed, assessments, and activities in chat rooms and on-line forums, which is very valuable in analyzing students' behaviour, predicting performance and assisting courseware authors in detecting possible errors, shortcomings and improvements (Romero, Ventura, Espejo, & Hervás, 2008; Romero, Ventura, & García, 2008). One of the most popular LMS activities are on-line discussion forums, because they are a promising strategy for collaboration and higher-order thinking. Some researchers have also detected the predictive relationship between discussion participation and learning (Koschmann, Kelson, Feltovich, & Barrows, 1996). While agreement exists that participation in on-line discussions can enhance student learning, it has also been identified that there is an increasing need to investigate the impact of participation on student course performance (Palmer, Holt, & Bray, 2008). Some examples are: Alstete and Beutell (2004) found that, in on-line courses, the strongest indicator of student performance was the use of discussions forums; Patel and Aghayere (2006) found that forum participation was positively correlated with the student's course grade in two undergraduate civil engineering courses; Shaw (2012) proved that a web-based programming language learning course supported with on-line forums and active student participation increases learning performance as measured by student learning scores; and Cheng et al. (2011) found that students who participated in a forum tended to have better performance in an introductory psychology course. Most of these approaches only use quantitative information such as message frequencies (the number of initiations and replies, the number of messages read, thread lengths, and response time from previous messages) to correlate them with course grades by stepwise multivariate linear regression analysis (Palmer et al., 2008), or only use the number of posts and the number of page views to find relations between forum participation and course performance by regression analysis (Cheng et al., 2011); or use the frequency of access and the duration of sessions to establish several categories of learners by cluster analysis, which depict the differences among the cohort in terms of participation (Khan, Clear, & Sajadi, 2012); or use agglomerative hierarchical clustering for

modelling student activity using the number of messages written and read on the forum by students (Cobo et al., 2011). Some other approaches use qualitative information mainly with a content analysis approach (Peña-Shaff & Nicholis, 2004). Content analysis can specify discussants' intentions by reading their posts and it can reveal latent semantic information in the transcript from the discussion boards for knowledge building or critical thinking (De Wever, Schellens, Valcke, & Van Keer, 2006). Speech acts or question and answer dialogue roles that participants play, together with emotional features covered by LIWC (Linguistic Inquiry and Word Count), have been also used to predict learners' project performance by stepwise regression analysis (Yoo & Kim, 2012). Finally, other approaches use social network information because one of the most important features of Internet forums is their social aspect (Morzy, 2009). Thus, a fuller understanding of students' participation in forums must include a social element to provide a richer explanation of the determinants of student behaviour and performance. These social aspects of a discussion can highlight user interest in a specific topic. Social Network Analysis (SNA) views social relationships in terms of network theory, consisting of nodes (representing individuals or students within the network) and ties (which represent relationships between the students). These networks are often depicted in a social network diagram, where nodes are represented as points and ties are represented as lines (Aggarwal, 2011). SNA examines the structure and composition of ties in a given network and provides insights into its structural characteristics, such as the individuals with the most outgoing connections, the most incoming connections, etc. (Memon, Xu, Hicks, & Chen, 2010). Nowadays, analytic software exists for carrying out social network analysis on the basis of forum usage data. Meerkat-ED (Rabbany, Takaffoli, & Zaïane, 2011) is a tool for analyzing student participation in discussion forums using social network analysis techniques, which automatically discovers relevant network structures and visualizes overall snapshots of interactions between the participants in the discussion forums. SNAPP (Bakharia & Dawson, 2011) is a tool for visualizing the evolution of participant relationships within discussion forums using social network analysis and to assist educators to evaluate student behavioural patterns against learning activity design objectives.

This paper proposes to use together both quantitative, qualitative and social network information about forum usage to predict students' success or failure in a course by applying classification algorithms and classification via clustering algorithms. Although clustering is normally an unsupervised process for grouping similar elements (students in our case) into clusters, classification can be performed via clustering if the class information is used to evaluate the clusters thus obtained. It is important to notice that the number of clusters must be fixed to the same number of class labels in order to obtain a useful model that can be compared with other classification models. This technique of using classification via clustering has been previously applied with success in different domains, for example, to develop an anomaly-based network intrusion detection system (Panda & Patra, 2009), to predict heart disease in medical diagnosis (Soni, Ansari, Sharma, & Soni, 2011), and to develop an effective system for classification of multidimensional data via clustering in image analysis (Krakovsky & Forgac, 2011). However, in educational domain, to our knowledge, only our own previous and initial work (Lopez et al., 2012) has so far used classification via clustering to predict student performance. However, in this paper, we describe the full proposed methodology in detail and a Moodle's module developed to gather indicators about students' interaction with forums. We also use new data mining approaches such as instance and attribute selection together with clustering plus association rules mining. Finally, we perform an early prediction by using the information collected in the middle of the course.

3. Material and methods

This work proposes the use of several data mining approaches (see Fig. 1) to improve prediction of students' final performance (whether they pass or fail the course) starting from student participation in an on-line discussion forum. Firstly, forum interaction data is collected on two different dates during the course, in the middle and at the end of the course. Next, instance and attribute-selection process are applied in order to select just one group of instances/messages and attributes/variables, which are transformed into a proper format to be mined. Then, classification and classification via clustering techniques are applied and compared. Finally, the obtained classification models are described and compared to versus clustering models and additional mining association rules for each cluster.

3.1. Collecting Moodle forum data

We have gathered usage data from on-line discussion forums provided by Moodle (Cole & Foster, 2007) that is one of the most widely-used open source LMS. In fact, we have used three different types of analytic data based on forum activity:

- Quantitative information that uses statistical information such as number of messages read/posted, time spent, etc.
- Qualitative information that uses an evaluation or score of the content of the messages by the teacher.
- Social network information that uses questioning and responding relationships between students.

We have developed a new Moodle's module oriented towards the instructor and specifically geared towards obtaining a summary dataset file that contains all the previous forum usage indicators. This module has two main components or windows. The first component displays and enables all the messages from a selected forum to be assessed (see Fig. 2).
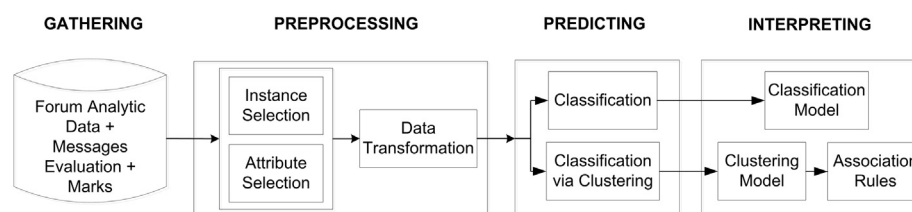

Fig. 1. Data mining approach for improving the prediction of students' final performance from forum data.

Fig. 2. Window for visualizing and evaluating messages.

As we can see in Fig. 2, the instructor can expand or reduce each thread in the forum (if it has several messages) in order to see the text of each message, the name and photo of the student who has written it, the posting date of the message and finally, the instructor can set an evaluation or score for the content of each message. This evaluation is about the relevancy and validity of the contextual meaning of the messages and is done manually by the actual course instructor. In order to facilitate this task, the module highlights in bold type the words most closely related to the subject of the course. The full list of words to highlight in the corresponding forum must be previously provided to the module. Then, the instructor must give each message a value/score between 0 and 3 points, based on grading rubric for on-line discussion participation by Kleinman (2005):

- 0 points to invalid messages that are off-topic or irrelevant to the content of the course.
- 1 point to messages that provide limited or basic information about topics in the course. These messages show that the student has an acceptable level of knowledge of the topic.
- 2 point to messages that provide adequate information about important topics. These messages show that the student has a good level of knowledge of the topic.
- 3 point to messages that provide very complete or precise information about difficult topics. These messages show that the student has an excellent level of knowledge of the topic.

The second component (see Fig. 3) shows a summary or analytic data report of forum participation information and enables the instructor to add the final mark obtained by each student in the course. The instructor can select a single forum or all of them, a specific student or all students, a range of dates, and can use all messages or exclude invalid messages, i.e., messages that scored more than 0 points.

The six qualitative indicators (*Messages*, *Threads*, *Words*, *Sentences*, *Reads*, and *Time*) are well-known and used in forum analysis. A single qualitative indicator (*AvgScoreMsg*) is obtained as the average score of all the messages sent by each student (see Table 1). Finally, the two social network analysis indicators are the degree of centrality (*Centrality*) and the degree of students' prestige (*Prestige*), which come from SNA and are closely related to hyperlink analysis (Memon et al., 2010). Centrality is a value between 0 and 1 that shows the ratio of a student's out-links. In our case, an out-link represents when a student writes a response to another student. A student's centrality i is calculated as the normalized node out-degree of that student:

$$C(i) = \frac{d_o(i)}{n-1}$$

where, $d_o(i)$ is the number of out-links and n is the total number of students.

Fig. 3. Window for viewing forum summary and adding final marks to students.

Table 1
Forum participation indicators used to represent each student.

Indicator	Type	Description
Messages	Quantitative	Number of messages written by the student.
Threads	Quantitative	Number of new threads created by the student.
Words	Quantitative	Number of words written by the student.
Sentences	Quantitative	Number of sentences written by the student.
Reads	Quantitative	Number of messages read on the forum by the student.
Time	Quantitative	Total time, in minutes, spent on forum by the student.
AvgScoreMsg	Qualitative	Average score on the instructor's evaluation of the student's messages.
Centrality	Social	Degree centrality of the student.
Prestige	Social	Degree prestige of the student.

Prestige is a value between 0 and 1 that shows the ratio of a student's in-links. In our case, an in-link represents when a student receives a response from another student. A student's prestige *i* is calculated as the normalized node in-degree of that student:

$$P(i) = \frac{d_I(i)}{n-1}$$

where $d_I(i)$ is the number of in-links and *n* is the number of students.

Both centrality and prestige are measures of the degree of prominence of an actor in a social network (Aggarwal, 2011). On the one hand, central or prominent actors are those that are extensively linked or involved with other actors. A person with extensive communications with many other people in the network is considered more important than a person with relatively few contacts. On the other hand, prestige is a more refined measure of the prominence of an actor than centrality. A prestigious actor is defined as one who is the recipient of extensive ties. Centrality and prestige can be used for detecting hub and authority students respectively (see Fig. 4).

Finally, the last column in Fig. 3 is used to set the final mark obtained by each student in the final exam at the end of the course, i.e. the class or attribute to be predicted in this study (*FinalMarkCourse*). It has two possible values or labels: *PASS* or *FAIL*, which have to be inserted manually by the instructor. We have simplified the performance measure into two binary values, *PASS* and *FAIL*, instead of using the examination scores, because in our case we are not interested in predicting what particular numerical value a student will obtain but rather in predicting whether a student will be successful in the course (pass) or not (fail). With this method, the obtained classification model could be used as an early warning system to alert the instructor about students who are potentially at risk of failing (Marquez-Vera, Cano, Romero, & Ventura, 2013). In this type of system, it is not necessary to predict the exact score obtained by students but only to predict whether the students will pass or fail. Once students were found to be at risk, for example, they could be provided with academic support and guidance to motivate them and attempt to prevent failure in time. For example, when students are identified to be potentially at risk of failing, they should be assigned a personal tutor in order to receive specific academic support, motivation, and guidance to prevent failure. There are also reference texts available which contain academic and behaviour supports for at-risk students (Stormont, Reinke, Herman, & Lembke, 2012) and would complement this particular study, which only aims to predict students' performance.

After inserting all the marks, the instructor can save/download all the information onto a PDF file for reporting purposes or onto an Excel file for data mining purposes.

3.2. Pre-processing data

Data pre-processing is an important step for preparing and filtering data before applying data mining algorithms (Klosgen & Zytkow, 2002). In our case, three main pre-processing tasks have been applied to the previous summarization Excel file: instance selection, attribute selection and data transformation.

- **Instance selection**. This is a data reduction task by choosing just one subset of data instances. In our case, we used two types of message filtering criteria. The first one selects messages posted at different moments/dates during the course, in our case in the middle and at the end of the course. The second one allows us to select all the messages or to exclude the invalid messages, i.e., to select those messages with more than 0 points in the instructor's score.
- **Attribute selection**. This is a feature or variable selection task for reducing the data dimensionality by selecting a subset of relevant attributes. In our case, although we do not have a great number of available attributes/indicators, some of them may be irrelevant for predicting students' performance. We have used a ranking approach to select attributes given that there exists a wide range of attribute-selection algorithms.
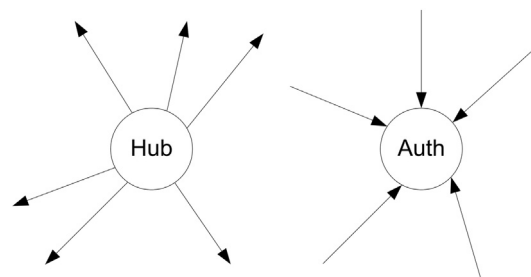

Fig. 4. A hub and an authority node/student.

- **Data transformation**. This task converts a set of data values from the data format of a source data system into the data format of a destination data system. In our case, data is transformed from the Excel format provided by our Moodle's module to the *.ARFF* Weka format (Witten & Frank, 2005), which is the data mining system that we have used for all our executions.

3.3. Predicting final mark

In this study, we are interested in predicting students' success or failure in a course, i.e., if the students will pass or fail the course, and not in predicting the numeric value of their final marks. So, this is a classification problem and not a regression problem. We propose to do it by using different DM methods. On the one hand, we propose to use a traditional classifier given that it is the traditional data mining method used for solving the task. On the other hand, we propose to use classification via clustering as an alternative method.

- **Classification** is a supervised method for identifying to which set of categories (labels) a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known (Duda, Hart, & Stork, 2000). A classifier is a mapping from a (discrete or continuous) feature space X to a discrete set of labels Y of the class to be predicted. In our case, the exploratory variables or feature space are the forum interaction/participation data, and the outcome or class to be predicted is the final grade in the course.
- **Clustering** is a process of grouping objects into groups of similar objects or subsets or clusters (Jain, Murty, & Flynn, 1999). Clustering and classification are both classification methods, although clustering is an unsupervised method and classification is a supervised method. Classification via/based on clustering is an approach for using clustering as a classifier based on the assumption that each cluster corresponds to a class. A clustering algorithm is firstly executed using the training data, after removal of the class attribute, and the mapping between classes and clusters is determined. This mapping is then used to predict class labels for unseen instances in the test data. In other words, the class attribute is not used in clustering but it is used to evaluate the obtained clusters as classifiers. So, it is important to ensure that the number of clusters generated is the same as the number of class labels in the dataset in order to obtain a useful model that relates each cluster with one class. The advantage of using classification via clustering is the ability to obtain a general display of the two groups or clusters which are generated: the students that pass and the students that fail.

3.4. Interpreting the models obtained

The interpretation and employment of the previously obtained classification models can be very useful for instructors to detect the final performance of new students in time and to make decisions about helping students predicted to fail. However, not all the models are equally interpretable. In fact, classification algorithms can be grouped in black and white box models (Klosgen & Zytkow, 2002). Black models such as Bayesian and artificial neural networks normally obtain a high classification accuracy but their explanation of the results is difficult to understand/obscure. On the other hand, white box models such as decision trees and rule-based algorithms are more useful given that they provide a set of IF-THEN classification rules that are one of the most popular forms of depicting knowledge thanks to their simplicity and comprehensibility. These models can be used directly for decision making and provide an explanation for the classification, which can then be reviewed and validated by a human expert.

Clustering algorithms also provide high interpretable models by means of the information of each cluster centroid. Cluster centroids are the mean vectors for each cluster, that is, each attribute/indicators value in the centroid represents the mean value for that attribute in the cluster. However, this centroid-based model is very dissimilar to that obtained by most of the white box classification models that normally provide IF-THEN rules. So, we propose to obtain an additional model for each cluster by means of association rule discovering. We propose to use an association rule mining algorithm on each obtained cluster to discover the most representative IF-THEN rules. Association rule mining is one of the most important and widely-studied data mining methods that aims to extract interesting correlations, associations or casual patterns among sets of items in data repositories (Ceglar & Roddick, 2006).

4. Experimental results and discussion

The data used in this paper was gathered from 114 university students during a first-year course in computer science in 2011–2012. This is an introduction course to computer science from a theoretical and practical point of view that is entitled "Computer Science Fundamentals". The course was enhanced by using a Moodle platform for providing the students with supplementary on-line resources, activities, and a discussion forum. The instructions given to the students were to use the forum for discussing course contents, solving doubts and problems between students, i.e., the instructor does not participate in the discussion. Students could ask questions about theory or exercise problems, replies to previous messages or just to browse through the discussion. The idea was that some students could help other students by using the discussion forum. Although it was not mandatory to join the forum, in order to encourage students to use the forum, the instructor also remarked that the level of participation in the forum would have a positive effect in the case of students that obtain a near-pass mark in the final exam. Finally, at the end of the course, students carried out a final pen and paper exam to evaluate them, and of the 114 students: 68 passed (59.65%) and 46 failed (40.35%) the course.

We carried out different executions to test the proposed approach in order to achieve the next four goals. The first goal was to compare the accuracy and interpretability of classification versus classification via clustering models. The second and third goals were to reduce the number of attributes and instances. And the last goal was to carry out an early prediction. We used the Weka tool (Witten & Frank, 2005), which is open source machine-learning software that provided us with all the data mining algorithms we used in our executions.

4.1. Instance and attribute selection

On the one hand, we carried out an instance selection process by using two different types of message filtering criteria. The first criteria was to collect forum data on different dates during the course. The course started in October 2011 and lasted 60 h of traditional classroom teaching over 4 months. The two specific dates were:

- Middle of the course (30 November 2011), when the summary information of 603 posted messages was included in a data file called Dataset 1.
- End of the course (31 January 2012), when the summary information of 1014 posted messages was included in a data file called Dataset 3.

The second message filtering criterion was to use all the messages or only messages with content truly related to the course. The reason for using this filter is that we noticed that most of the students had also sent messages to the forum not related to the subject of the course, for example, speaking about other courses, meeting other students, arranging to do sports or go to parties, etc. In this way, two new data files were created on the basis of the two previous data files by deleting invalid messages:

- Dataset 2 that contained only 277 messages from the 603 messages in Dataset 1.
- Dataset 4 that contained only 481 messages from the 1014 messages in Dataset 3.

On the other hand, we carried out an attribute-selection process for selecting a subset of relevant attributes from all the available attributes. Before describing the process and in order to reveal some possible correlations or relationships among all the available attributes, we show the matrix correlating (see Table 2) all attributes when using all available data (Dataset 3). This correlation matrix is a symmetrical matrix, on which each entry is the Pearson correlation coefficient or Pearson's r (Rodgers & Nicewander, 1988) that provides the linear dependence between two variables or attributes as a value between 1.0 and -1.0. In our case, all these values have a significance level or p-value < 0.01.

From Table 2, it can be seen that in general, there are a majority of strong positive correlations (r value greater than 0.5) and some medium and small (r value between 0.3 and 0.5, and between 0.1 and 0.3, respectively), positive correlations between the attributes. The most correlated attributes (with highest r values) are Messages&Sentences and Words&Sentences. On the other hand, the least correlated attributes (with lowest r values) are Threads&Time and Time&AvgScoreMsg. However, in order to select the most relevant attributes for predicting student performance, we applied feature-selection algorithms that have been used in similar work (Yoo & Kim, 2012). In our case, we used a ranking procedure of the attributes or forum participation indicators. Firstly, we used several feature-selection algorithms. Then, a voting process in which each algorithm selects or votes a list of attributes. Finally, the attributes that had been selected by more algorithms were selected as the best attributes or attributes that should have a greater effect on students' final performance. Weka provides several feature-selection algorithms (Witten & Frank, 2005) from which we used the ten listed below: CfsSubsetEval (which evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them), ChiSquaredAttributeEval (which evaluates the worth of an attribute by computing the value of the chi-squared statistic with respect to the class), ConsistencySubset-Eval (which evaluates the worth of a subset of attributes by the level of consistency in the class values when the training instances are projected onto the subset of attributes), SignificanceAttributeEval (which evaluates the worth of an attribute by computing the probabilistic significance as a two-way function), SymmetricalUncertAttributeEval (which evaluates the worth of an attribute by measuring the symmetrical uncertainty with respect to the class), GainRatio-AttributeEval (which evaluates the worth of an attribute by measuring the gain ratio with respect to the class), InfoGainAttributeEval (which evaluates the worth of an attribute by measuring the information gain with respect to the class), OneRAttributeEval (which evaluates the worth of an attribute by using the OneR classifier), ReliefFAttributeEval (which evaluates the worth of an attribute by repeatedly sampling an instance and considering the value of the given attribute for the nearest instance of the same and different class), and SVMAttributeEval (which evaluates the worth of an attribute by using an SVM classifier). These algorithms select either one subset of attributes or some of them return a ranked list of all attributes, in which case, we only selected the first 5 attributes. Then, we counted the number of times each attribute was selected by each attribute-selection algorithm. Finally, we selected as the best attributes those attributes selected by 5 or more algorithms, i.e., at least half of the featured selection algorithms (see Table 3).

In Table 3, we can see what the selected attributes were (in bold and between brackets) and the total number of selected attributes (in the last row of each column) for each dataset. Overall, the best or most relevant attributes were the number of messages, the number of words, the average score/evaluation of the messages and the degree of centrality, due to the fact that they had been selected in all datasets by more than 5 algorithms. The least relevant attributes were the number of threads, the number of sentences, the number of reads and the total time in the forum, because they had not been selected by more than 5 algorithms in any dataset. An attribute that behaved differently was the degree of prestige, which was selected only in the two last datasets by more than 5 algorithms, showing that a student's prestige was not a relevant attribute in the first part of the course, but it did become relevant in the second half of the course.

Finally, if we filter each one of the previous four datasets by using all the available attributes (option a) or only the selected best attributes (option b), we obtain the eight datasets (see Table 4) used in the next classification step.

Table 2
Correlation matrix of the selected attributes about activity in forums when using all the available data.

	Messages	Threads	Words	Sentences	Reads	Time	AvgScoreMsg	Centrality	Prestige
Messages	1	0.509	0.657	0,743	0.599	0.434	0.481	0.699	0.604
Threads	0.509	1	0.526	0.461	0.313	0161	0.347	0.553	0.504
Words	0.657	0.526	1	0.773	0.404	0.372	0.474	0.649	0.646
Sentences	0.743	0.461	0.773	1	0.475	0.429	0.497	0.672	0.616
Reads	0.599	0.313	0.404	0.475	1	0.562	0.411	0.606	0.516
Time	0.434	0.161	0.372	0.429	0.562	1	0.255	0.435	0.374
AvgScoreMsg	0.481	0.347	0.474	0.497	0.411	0.255	1	0.544	0.498
Centrality	0.699	0.553	0.649	0.672	0.606	0.435	0.544	1	0.659
Prestige	0.604	0.504	0.646	0.616	0.516	0.374	0.498	0.659	1

Table 3
Frequency of attribute selection by the 10 featured selection algorithms for each dataset.

Attribute	Dataset 1	Dataset 2	Dataset 3	Dataset 4
Messages	**(6)**	**(6)**	**(8)**	**(8)**
Threads	2	4	3	4
Words	**(7)**	**(8)**	**(7)**	**(8)**
Sentences	3	3	3	3
Reads	1	2	3	4
Time	1	2	1	2
AvgScoreMsg	**(5)**	**(7)**	**(5)**	**(7)**
Centrality	**(7)**	**(7)**	**(9)**	**(9)**
Prestige	4	4	**(6)**	**(8)**
Number of Selected Attributes	4	4	5	5

Results in bold and between parenthesis are the selected attributes.

4.2. Classification and classification via clustering

We ran several classification algorithms on the eight previously described datasets (see Table 4). It is important to notice that in all our executions, we always used the entire population (114 students) but with different types of data reduction approaches to predict students' final marks (see Fig. 5).

We used 10 cross fold-validation that means that each dataset is randomly divided into 10 disjointed subsets of equal size in a stratified way (maintaining the original class distribution). Each algorithm is executed 10 times and in each repetition, one of the 10 subsets is used as the test set and the other 9 subsets are combined to form the training set. Finally, the mean evaluation measure of the classification performance is calculated. In our case, we used two well-known measures, namely Accuracy (percentage of correctly classified instances/students) and *F*-measure (harmonic mean of precision and recall). On the other hand, we have executed different types of traditional classification algorithms provided by Weka:

- **Rule-based algorithms** that reveal rules. DTNB that builds a decision table/naive Bayes hybrid classifier (Hall & Frank, 2008). JRip that implements a propositional rule learner as an optimized version of IREP algorithm (Cohen, 1995). Nnge that is a nearest-neighbour-like algorithm using non-nested generalized exemplars which are hyperrectangles that can be viewed as rules (Martin, 1995). And Ridor, which is the implementation of a Ripple-Down Rule learner (Gaines & Compton, 1995).
- **Tree-based algorithms** that reveal a decision tree. ADTree that generates an alternating decision tree based on Freund and Mason algorithm (Freund & Mason, 1999). J48 that is an optimized version of C4.5 decision tree (Quinlan, 1993). LADTree that generates a multi-class alternating decision tree using the LogitBoost strategy (Holmes, Pfahringer, Kirkby, Frank, & Hall, 2002). And RandomForest that constructs random forests based on Leo Breiman algorithm (Breiman, 2001).
- **Function-based algorithms** that reveal a function. Logistics that builds and uses a multinomial logistic regression model with a ridge estimator (Le Cessie & Van Houwelingen, 1992). MultilayerPerceptron that uses a back-propagation network to classify instances (Ruck, Rogers, Kabrisky, Oxley, & Suter, 1990). RBFNetwork that implements a normalized Gaussian radial basis function network (Park & Sandberg, 1991). And SMO, which implements a specific sequential minimal optimization algorithm for training a support vector classifier (Platt, 1999).
- **Bayes-based algorithms** that reveal a probabilistic classifier based on Bayes theorem. BayesNet, which uses a Bayes Network classifier like K2 and B (Bouckaert, 2007). And NaiveBayesSimple, which uses a simple Naive Bayes classifier in which numeric attributes are modelled by a normal distribution (Duda et al., 2000).

Furthermore, we used classification via clustering approach in which we executed clustering algorithms by setting them to generate 2 clusters in order to predict the two classes (Pass/Fail) correctly on the basis of the obtained clusters. We executed the following clustering algorithms provided by Weka:

- **EM** (Expectation Maximization) algorithm (Moon, 1996) that assigns a probability distribution to each instance which indicates the probability of it belonging to each of the clusters.
- **HierarchicalCluster** algorithm (Zhao, Karypis, & Fayyad, 2005) that seeks to build a hierarchy of clusters using an agglomerative or bottom-up approach in which each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.
- **sIB** (sequential Information Bottleneck) algorithm (Slonim, Friedman, & Tishby, 2002) that assigns for each instance the cluster that has the minimum cost/distance to the instance.

Table 4
Dataset description.

| Dataset	Collection date	Instances used	Attributes used	Number of instances	Number of attributes
Dataset 1a	Middle of the course	All	All	603	9
Dataset 1b	Middle of the course	All	Only selected	603	4
Dataset 2a	Middle of the course	Only selected	All	277	9
Dataset 2b	Middle of the course	Only selected	Only selected	277	4
Dataset 3a	End of the course	All	All	1014	9
Dataset 3b	End of the course	All	Only selected	1014	5
Dataset 4a	End of the course	Only selected	All	481	9
Dataset 4b	End of the course	Only selected	Only selected	481	5
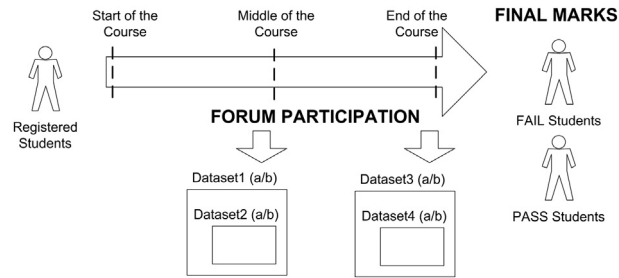
Fig. 5. Data gathering and created datasets.

- **SimpleKMeans** (Kanungo et al., 2000) is the most widely-used, simple and well-known clustering algorithm, which aims to partition n instances into k clusters in which each instance belongs to the cluster with the nearest mean.
- **Xmeans** (Pelleg & Moore, 2000) is an extension of the SimpleKMeans algorithm that efficiently searches the space of cluster locations to optimize the Bayesian Information Criterion (BIC) measure.
- **FarthestFirst** (Hochbaum & Shmoys, 1985) is a variant of SimpleKMeans algorithm that places each cluster centre in turn at the point furthest from the existing cluster centres.

We compared the accuracy and F-measure of traditional classification versus classification via clustering algorithms (see Table 5).

Table 5 shows the accuracy and F-measure values obtained by the classification algorithms (upper rows of the table) and classification via clustering (lower rows of the table) for all datasets. On the basis of the results obtained, we can respond to the four initial research questions posed by this paper:

1. What DM techniques are best for predicting student performance starting from participation in on-line forums? To reply to this first question, we compared the two different approaches to predict student performance. Thus, a set of classification and clustering

Table 5
Accuracy and F-measure of classification algorithms with all the datasets. The best results are provided in bold typeface.

Algorithm	Measure	Dataset 1		Dataset 2		Dataset 3		Dataset 4	
		a	b	a	b	a	b	a	b
DTNB	Accuracy	0.701	0.763	0.798	0.789	0.859	0.833	0.886	0.894
	F-measure	0.698	0.752	0.795	0.786	0.861	0.835	0.889	0.890
JRip	Accuracy	0.684	0.771	0.780	0.763	0.833	0.815	0.833	**0.903**
	F-measure	0.681	0.768	0.777	0.756	0.835	0.817	0.838	**0.895**
NNge	Accuracy	0.710	0.666	0.763	0.807	0.842	0.807	0.869	0.807
	F-measure	0.709	0.667	0.766	0.805	0.841	0.808	0.863	0.807
Ridor	Accuracy	0.719	0.745	0.754	0.780	0.833	0.842	0.886	0.894
	F-measure	0.715	0.742	0.751	0.777	0.835	0.843	0.889	0.890
ADTree	Accuracy	0.684	0.701	0.754	0.807	0.859	0.842	0.804	0.886
	F-measure	0.684	0.698	0.756	0.802	0.860	0.843	0.796	0.886
J48	Accuracy	0.675	0.763	0.771	0.780	0.824	0.807	0.878	0.903
	F-measure	0.663	0.756	0.769	0.776	0.825	0.808	0.881	0.895
LADTree	Accuracy	0.710	0.763	0.771	0.789	0.868	0.850	0.851	0.886
	F-measure	0.707	0.760	0.770	0.784	0.869	0.852	0.853	0.886
RandomForest	Accuracy	0.710	0.719	0.815	0.789	0.850	0.833	0.886	0.894
	F-measure	0.708	0.718	0.816	0.784	0.851	0.834	0.869	0.890
Logistic	Accuracy	0.719	0.745	0.771	0.771	0.859	0.850	0.833	0.877
	F-measure	0.716	0.740	0.769	0.768	0.860	0.852	0.824	0.876
MultilayerPerceptron	Accuracy	0.710	0.736	0.771	0.824	0.842	0.868	0.842	0.886
	F-measure	0.705	0.726	0.768	0.822	0.843	0.869	0.844	0.890
RBFNetwork	Accuracy	0.675	0.728	0.789	0.789	0.868	0.886	0.877	0.886
	F-measure	0.674	0.726	0.784	0.787	0.869	0.887	0.878	0.886
SMO	Accuracy	**0.754**	**0.780**	**0.824**	0.815	0.868	0.886	0.842	0.877
	F-measure	**0.748**	**0.778**	**0.821**	0.805	0.870	0.861	0.824	0.845
BayesNet	Accuracy	0.719	0.719	0.798	0.815	**0.877**	0.842	**0.894**	**0.903**
	F-measure	0.717	0.716	0.796	0.788	**0.878**	0.844	**0.895**	**0.895**
NaiveBayesSimple	Accuracy	0.728	0.736	0.798	**0.824**	0.859	**0.894**	**0.894**	**0.903**
	F-measure	0.720	0.728	0.796	0.822	0.861	**0.896**	**0.895**	**0.895**
EM	Accuracy	**0.736**	**0.771**	**0.780**	**0.807**	0.842	**0.894**	0.886	0.894
	F-measure	**0.715**	**0.719**	**0.738**	**0.755**	0.843	**0.895**	0.883	0.892
HierarchicalClusterer	Accuracy	0.517	0.526	0.535	0.535	0.578	0.570	0.771	0.771
	F-measure	0.442	0.493	0.442	0.442	0.493	0.489	0.772	0.772
sIB	Accuracy	0.614	0.535	0.631	0.526	0.710	0.578	0.842	0.780
	F-measure	0.510	0.442	0.622	0.493	0.689	0.493	0.831	0.769
SimpleKMeans	Accuracy	0.631	0.614	0.622	0.631	0.666	0.640	0.824	0.833
	F-measure	0.622	0.510	0.622	0.622	0.653	0.621	0.822	0.825
Xmeans	Accuracy	0.649	0.631	0.657	0.640	0.666	0.640	0.763	0.807
	F-measure	0.613	0.622	0.627	0.631	0.653	0.621	0.745	0.797
FarthestFirst	Accuracy	0.526	0.535	0.526	0.535	0.526	0.535	0.798	0.798
	F-measure	0.493	0.442	0.493	0.442	0.505	0.472	0.778	0.778

algorithms were used (the algorithms in the upper and lower part of Table 5 respectively), where the accuracy and *F*-measure for different datasets is depicted. The EM algorithm provided the best values both in accuracy and *F*-measure among the clustering algorithms in all datasets, so this algorithm was compared to the best classification algorithms (SMO, BayesNet and NaiveBayesSimple) that obtain the highest accuracy and *F*-measure values in most datasets. Comparing EM and SMO algorithms, SMO obtained a better accuracy and *F*-measure values in 5 out of 8 datasets, whereas EM behaved better in 3 out of 8. The same result was obtained when comparing EM and BayesNets, BayesNet obtained a better accuracy and *F*-measure values in 5 out of 8 datasets, whereas EM behaved better in 3 out of 8. Finally, when comparing EM versus NaiveBayesSimple algorithm, NaiveBayesSimple obtained better accuracy and *F*-measure in 5 of out 8 datasets, EM behaved better in 2 out 8 datasets and both obtained the same accuracy and *F*-measure in one dataset. Thus, although the three classification algorithms were better than the EM clustering algorithm, we could assert that there was not much difference in the way these four algorithms performed. Furthermore, EM provides a white box model that is more interpretable than the models obtained by these three classification algorithms. However, two classification algorithms (J48 and Jrip) obtain the highest accuracy and *F*-measure values (in the last dataset) and also provide a white box model. Furthermore, we calculated the average total error (type I and II) when using all datasets and all algorithms. Type 1 error or False Positive (FP) rate (% of student classified as Fail when they really should be Pass) was 14.529%. Type 2 error or False Negative rate (% of students classified as Pass when really they should be Fail) was 8.519%. The fact that Type 2 error was lower than Type 1 error is a desirable characteristic in a warning system such as ours that is aimed at detecting students at risk of failing. This is due to the fact that, although the warning system may produce some false alarms (type 1 error), it is less likely not to raise an alarm (type 2 error).

2. What attributes are the best predictors? To reply to this second question, the difference between the behaviour of algorithms when using the complete set of attributes or just a subset of attributes was measured (columns a versus columns b in each dataset in Table 5). Thus, we require a statistical test to carry out a pairwise comparison. This kind of statistical testing focuses on comparing the behaviour of two algorithms when they are run on a set of problems. An important test in this sense is the Wilcoxon signed rank test (Blair & Higgins, 1985), which is based on the order of the differences between both algorithms. This non-parametric statistical test calculates the absolute values of differences in the behaviour for each problem, and these differences are ranked in ascending order in order to compare them. Thus, we deem fitting to compare a pair of datasets on a set of algorithms to compare the general behaviour of each dataset. Focussing on the first dataset (Table 5), the Wilcoxon signed rank test calculated a 0.0182 *p*-value for the accuracy measure and a 0.0521 *p*-value for the *F*-measure, so at a significance level of 0.10, i.e., with a probability of 90%, it is possible to assert that the use of a subset of attributes behaved statistically better than the use of the whole dataset. As for the second dataset, the *p*-values obtained were 0.1906 for Accuracy and 0.3556 for the *F*-measure, whereas the third dataset calculated 0.1314 and 0.1088 *p*-values for Accuracy and *F*-measure, respectively. Therefore, it is not possible to assert that the use of a subset of attributes obtained higher accuracy and *F*-measure values for these two datasets. Finally, when considering dataset number four, the Wilcoxon signed rank test obtained a 0.0088 *p*-value for the accuracy measure and a 0.0192 *p*-value for the *F*-measure, so with a probability of 99% and 95%, respectively, it is possible to assert that the use of a subset of attributes behaved statistically better than the use of the whole dataset. So, the statistical analysis shows that only the use of a subset of 4 or 5 attributes (number of messages, number of words, average score messages, degree of centrality and prestige, as we can see in Table 3) saw an increase in Accuracy and *F*-measure values. Although this assertion could not be statistically proven in all cases (for example in datasets number two and three), if we analyze Table 5, we can see that the highest values for these two quality measures were always obtained when the subset of selected attributes was used. We would also like to highlight the fact that using a subset of attributes leads to a more understandable model as the number of attributes to be considered is reduced (Janecek & Gansterer, 2008). This means that the use of the aforementioned attributes is of high interest in the discovery of more understandable and accurate classification models.

3. What messages are the best predictors? In order to reply to this third question about the use of all messages or just messages with a score or messages related to the course subjects, we compared Dataset 1 versus 2, and Dataset 3 versus 4. Using Accuracy and *F*-measure, the Wilcoxon signed rank test revealed that, at a significance level of 0.01 (i.e., with a probability of 99%), it is possible to assert that Dataset 2 behaved statistically better than Dataset 1 for both measures. Thus, it does not matter whether the complete set of attributes is used or not. Similarly, at a significance level of 0.01, it is possible to assert that Dataset 4 behaved statistically better than Dataset 3 in both measures. Thereby, it can be stated that in our case, using only messages related to the subjects of the course obtained better accuracy and *F*-measure performance than when all the available messages were used. The theoretical reasons for using only messages whose content is related to the course is described in more detail in the Discussion section.

4. Is it possible to make an early prediction? In order to reply to this last question, we compared Datasets 1 and 2 versus Datasets 3 and 4. The Wilcoxon signed rank test revealed that, at a significance level of 0.01 (i.e., with a probability of 99%), it was possible to assert that Dataset 3 behaved statistically better than Dataset 1 in accuracy and *F*-measure. In a similar way, at a significance level of 0.01, it was possible to assert that Dataset 4 behaved statistically better than Dataset 2 in both measures. Therefore, although we could assert that the best accuracy and *F*-measure were obtained when predicting at the end of the course, the average values obtained in the middle of the course (Dataset 1 and 2) are between 70 and 80% accurate, which is a good accuracy rate for making an early prediction compared to the final average accuracy values of between 80 and 90% (Dataset 3 and 4). We make this assertion on the basis of the results of previous related works on early student performance prediction systems (Kovacic, 2012; Sabourin, Mott, & Lester, 2012) that obtained lower values for average accuracy than our early approach: between 51 and 59%, between 59 and 60%, respectively. On the other hand, other related works about predicting final student performance at the end of the course (Yoo & Kim, 2012; Zafra, Romero, & Ventura, 2011) obtained lower (between 64 and 73%) and very similar or a little higher (between 86 and 93%) average accuracy values, respectively, than our final approach. So, these results suggest that, in general, our two approaches obtained good accuracy compared to other related researches.

4.3. Interpreting the discovered models

The models discovered by the previous classification algorithms lead us to predict the success or failure of new students, and thus, they are very useful, for example, to identify/detect students that are likely to fail, so that the instructor can offer them appropriate and personalized help in an attempt to avoid failure before the end of the course. Thus, it is very important that the model should be

comprehensible/interpretable for instructors (Romero et al., 2013). However, not all models are equally user-friendly and two different types can be distinguished: black and white models. Black-box models are able to attain very good accuracy rates but their models are more difficult for users to understand. On the contrary, there are also many algorithms such as rule-based and tree-based algorithms whose models are easily understandable by users, named white box models.

Hereunder, we describe and compare the models of two white box classification algorithms (JRip and J48) versus a clustering algorithm (EM) obtained when using the last dataset (Dataset 4b) in which they gave the highest accuracy value. JRip algorithm generates the next set of rules:

```
(Centrality >= 0.018) => FinalMarkCourse = PASS (29.0/2.0)
(Words >= 201) and (Messages > 3) => FinalMarkCourse = PASS (26.0/2.0)
(Messages > 3) and (AvgScoreMsg > 1.2) => FinalMarkCourse = PASS (10.0)
=> FinalMarkCourse = FAIL (49.0/7.0)
```

As we can see, the rule set produced by JRip is an IF-ELSE-THEN structure in which the *THEN* operator is indicated by the symbol "=>", and the numbers between brackets at the end of each rule show the number of instances associated with this rule (coverage) and the number of instances incorrectly classified (error) according to the rule. The first number describes the number of instances/students considered in each rule, whereas the second number shows the number of misclassified instances. The meaning of these specific rules is that if the degree of centrality of a student is higher than 0.018, then the student is predicted to *PASS* the course. In addition, if the number of words and messages written by a student is greater than 201 and 3 respectively, then this student is predicted to *PASS* the course. Furthermore, if the number of messages written is higher than 3 and the average score assigned by the instructor is higher than 1.2, then the student is predicted to *PASS* the course. Finally, if none of the previous conditions are satisfied, then the student is predicted to *FAIL* the course.

J48 algorithm generates the following decision tree:

```
Centrality <= 0.014
| Messages <= 3
| | AvgScoreMsg <= 1.15: FAIL (38.0/5.0)
| | AvgScoreMsg > 1.15: PASS (16.0/3.0)
| Messages > 3
| | Words <= 189: FAIL (9.0/1.0)
| | Words >= 189: PASS (18.0)
Centrality > 0.014: PASS (33.0/2.0)
```

As we can see, a decision tree is a set of rules with the form IF-ELSE-IF, in which the *THEN* operator is indicated by the symbol ":", and again the numbers between brackets mean the coverage/error of this rule or leaf of the tree. As we can see, this model is very similar to the one generated by JRip algorithm. In this specific decision tree, students are divided into two major leafs, students that have a degree of centrality higher and lower than 0.014. Among the students that have a degree of centrality lower than 0.014, the model differentiates between those that sent a number of messages lower than 3 and those whose number of messages is greater than 3. The first subgroup can be divided into those students that obtain an average score greater than 1.15 and are predicted to *PASS* the course, and those that obtain a lower average score and are predicted to *FAIL* the course. Finally, the second subgroup can be divided into those students that write more than 189 words and are predicted to *PASS* the course, and those that write fewer words and are predicted to *FAIL* the course.

Finally, EM algorithm generates two clusters described by their centroids. The centroid represents the most typical case/student or prototype in a cluster, which does not necessarily describe any given case assigned to the cluster. The attribute values for the centroid are mean and standard deviation (see Table 6).

Cluster centroids describe the typical student for each group or cluster (see Table 6). Weka shows that for each centroid, the mean (a measure of central tendency that is obtained as the average of a set of data) and the standard deviation (a measure of the spread of data from the mean, obtained as the square root of the variance) of each attribute. Weka also provides clusters for class mapping because we used classification via clustering, in this case Cluster 0 belongs to the group of *PASS* students and Cluster 1 belongs to the group of *FAIL* students. We can see that the centroids obtained can be very informative from the point of view of classifying *PASS* and *FAIL* students. In fact, students who have shown a great level of participation in the forum (about 5 messages, 300 words) with a good average evaluation (about 2 points) and degree of centrality and prestige (about than 0.04 and 0.02 respectively) are grouped in cluster 1 (student who *PASS* the course). Similarly, students who have shown a very low or level of participation, average evaluation and degree of centrality and prestige are

Table 6
Cluster centroids obtained by EM algorithm.

Attribute	Cluster 0	Cluster 1
Messages	5.122 ± 3.081	0.866 ± 1.075
Words	301.750 ± 202.746	14.752 ± 16.447
AvgScoreMsg	1.908 ± 1.043	0.01 ± 1.191
Centrality	0.043 ± 0.034	0.002 ± 0.011
Prestige	0.027 ± 0.025	0 ± 0.019

classified as *FAIL* students (cluster 0). Finally, with regard to the relationship between the obtained mean and standard deviation values, in cluster 0 (*PASS* students) all of the attributes have a standard deviation less than the mean, i.e. most of the data are centred around the mean, presenting a normal distribution. However, in all attributes in cluster 1 (*FAIL* students) the values of standard deviation are greater than the mean. This is possibly due to different reasons (Freedman, Pisani, & Purves, 2007). On the one hand, almost all of the attributes in cluster 1 have a mean near to 0, and so, any standard deviation will be greater than the mean. On the other hand, this fact reflects the variability and heterogeneity of the data, which demonstrate a great deal of variance (non-normally distributed data) and outliers (observations that are numerically distant from the rest of the data).

However, this centroid-based model is very different to the previous white box classification models which provide rules. So, we plan to discover a set of IF-THEN rules for capturing the main characteristics of the students assigned to each cluster. In order to do so, we propose to use a representative number of association rules discovered for each cluster as an additional model. The problem is that association rule mining algorithms normally discover a huge number of rules. Among the complete set of association rules discovered in a specific problem, it is possible to find a special subset called Class Association Rules. The main feature of this sort of rule is that it includes a consequent limited to a target class label (only one condition), whereas the left-hand side of the rule may contain one or more conditions. This type of rule is represented as IF A THEN C, where A is the antecedent or conditions of the forum participation indicators, and C is the class that in our case is the final mark attribute. This type of rule is more understandable than general association rules, since it only comprises one element in the consequent. In order to obtain a representative set of class association rules that describe the previously obtained clusters, the HotSpot (Agrawal & Choudhary, 2011) algorithm provided by Weka was run on the two cluster data files. However, first we need to create the two data files based on the full dataset which includes each instance along with its assigned cluster. In order to do so, we use the Weka's visualization window to visualize cluster assignment and to save them into new datasets in which each instance now has its assigned cluster as the last attribute value. Next, by simple manipulation of this dataset, we can easily convert it into two separate files – one for each cluster – and then we have to delete the last attribute (the assigned/predicted cluster, which is now the name of the file) so that the last attribute was the current/real *PASS* or *FAIL* class value. Next, HotSpot is executed to discover a set of class association rules that maximize/minimize a target variable/value of interest in each data file or cluster file. This algorithm is very useful for segmentation/profiling because it can find segments where there is a higher or lower than average likelihood of a particular class value occurring. In our case, we want to maximize the *PASS* value in the Cluster 0 file and the *FAIL* value in the Cluster 1 file, which are the main segments of students representing each cluster respectively. In fact, from 63 students grouped in cluster 0: 60 *PASS* and 5 *FAIL* the course, and from 51 students grouped in cluster 1: 43 *FAIL* and 7 *PASS* the course.

The first model induced by the HotSpot algorithm when using the cluster 0 and the value *PASS* is the next:

```
[Messages > 4]: 50 = => [FinalMarkCourse = PASS]: 50 conf: (1)
[Words > 285]: 45 = => [FinalMarkCourse = PASS]: 45 conf: (1)
[Prestige > 0.010]: 43 = => [FinalMarkCourse = =PASS]: 43 conf: (1)
```

As we can see, HotShopt obtains a set of IF-THEN rules in which the *THEN* operator is indicated by the symbol " $= =>$ ", each condition (between square brackets) shows its supports (instances/students in the dataset which contain/match this condition) as a integer number after the symbol ":", and each rule shows its confidence (proportion of instances in the dataset which contain both the antecedent and consequent conditions) at the end of the rule as a real number between brackets ranging between 0 (0%) and 1 (100%). In this first model, HotSpot algorithms discovered 3 rules, all of them having the maximum reliability of 100%. These rules show that from the set of students that *PASS* the course, 50 of them wrote more than 4 messages in the forum, 45 of them wrote more than 285 words and 43 of them obtained a degree of prestige higher than 0.010.

The second model induced by the HotSpot algorithm when using cluster 1 and the FAIL value is as follows:

```
[Words <= 18]: 28 = => [FinalMarkCourse = FAIL]: 28 conf: (1)
[Messages <= 1]: 27 = => [FinalMarkCourse = FAIL]: 27 conf: (1)
[Words <= 18, Centrality <= 0.054]: 24 = => [Mark = FAIL]: 24 conf: (1)
[Words <= 18, Messages <= 1]: 23 = => [Mark = FAIL]: 23 conf: (1)
[Words <= 18, AvgScoreMsg = 0]: 23 = => [FinalMarkCourse = FAIL]: 23 conf: (1)
```

In this case, HotSpot discovered 5 rules, all of them having maximum reliability of 100%. These rules show that from the set of students that *FAIL* the course, 28 of the students wrote less than 18 words, 27 students wrote less than 1 message, 24 students wrote less than 18 words and obtained a centrality degree less than or equal to 0.054, 23 students wrote less than 18 words and less than 1 message, and 23 of them wrote less than 18 words and obtained an average score in messages equal to 0.

4.4. Discussion

Relatively few studies hitherto have focused on the specific and common type of discussion forum used in this work, in which, a forum is implemented as a medium for students to ask questions or discuss anything related to their course (Cheng et al., 2011). Participation in this type of forum is usually voluntary and intrinsic and the instructor only played a passive role. In order to avoid any influence caused by the instructor's subjective guiding methods and ensure objective observations, in this paper, we observed learners' on-line problem-solving

discussions without intervention or guidance from the instructor. Students were told that their activity would be monitored (but not that it would be used as part of a research study) so as not to affect their natural behaviour. However, one debate factor is whether student participation in on-line discussion should be optional or mandatory. Some previous works (Vonderwell, 2003) find that that when students do not see each other, they can avoid answering other students' questions or requests for help. And they may not feel morally obligated or pressured to participate in on-line communication. In order to resolve this problem, the literature suggests that some form of extrinsic motivation is required to ensure a good level of student discussion and participation (Palmer et al., 2008). In this work, the instructor reminded students in class to participate in the on-line discussion forum. And the instructor also explained to students that they had to send at least two new messages/questions and at least two replies to other messages/questions (from other students) in order to help students that obtain a near-pass mark in the final exam.

The data mining models described above (see Section 4.3) show us the discovered relationships between participation in an on-line discussion forum and student's final performance in a computer science course. These models indicate that most of the students who passed the course were the students who participated most actively in the forum, both in quantity (by writing more messages and words) and quality (by obtaining higher values of average score messages, centrality and prestige). Likewise, the opposite was true in that those students who failed in the course were the ones who participated least in the forum. In the literature, there are a great number of related works that have previously studied the causality of students' final performance and participation in on-line discussion forums, in a separate way. On the one hand, there are many factors that may affect final performance and a number of studies have been carried out to identify causal factors of poor academic performance in various institutions worldwide. Most of these studies focus on any intervening elements, such as parents (family causal factors), instructors (academic causal factors), and students (personal causal factors) (Diaz, 2003). The combination of factors influencing academic performance, however, varies from one academic environment to another, from one set of students to the next, and indeed from one cultural setting to another. On the other hand, there are a lot of factors or aspects that may influence the student's participation in on-line discussion forums such as structure of the course, class size, feedback, prior knowledge of computer mediated communication, interface characteristics, content area experience, students roles and instructional tasks, differences in students' demographics and abilities, etc. (Yukselturk, 2010). As a result, the relationship and causality between participation/interaction and learning outcomes is a complex phenomenon and thus an error rate (type I and II errors) can occurs when using only some indicators about participation in on-line discussion forums for predicting student performance. Therefore upgrading its correctness it is a difficult task in that it would also require more human intervention from the instructor. On the one hand, it would be necessary to use a wider variety of students (not only from computer science but from other disciplines where the students are not as familiar with the use of computers): something which we propose as future work in our conclusions. On the other hand, it would also be interesting for instructors to motivate and guide the students in the appropriate use of the on-line forum for other activities on the course. For example, by suggesting that the students discuss or try to resolve possible doubts or difficult topics which the instructor identifies during theoretical or practical classes. In this way, most of the message content in the forum would deal with the most difficult topics and questions on the course, meaning that forum participation indicators would be more closely related with the student's final performance.

Finally, messages or queries posted to a discussion forum often span multiple sentences, are incoherent (in the computational sense), include extra (informal) content and lengthy description, especially in technical discussions (Kim, Shaw, & Ravi, 2010). In fact, some authors (Paredes & Kenneth, 2012) have defined different categories of content richness, in which, a specific type of category is the empty message that contains inexistent content, file exchange without dialogue, greetings messages, etc. Deleting such empty messages that may be irrelevant to discover knowledge from on-line discussions forums has been proposed and used in some earlier related works (Hou, Chang, & Sung, 2008; Li & Wu, 2010). Following the same idea, in this paper we have also proposed to detect and filter empty messages that are about subjects or contents not related to the course that the instructor might consider as spam. Comment spamming (comments which are unsolicited, unrelated, abusive, hateful, commercial advertisements, etc) in on-line discussion forums is a common phenomenon in Web 2.0 applications (Sureka, 2011) and there is an increasing research interest in filtering out spam messages or comment spamming in on-line discussion forums using context-based approaches (Niu, Wang, Chen, Ma, & Hsu, 2007; Sureka, 2011).

5. Conclusions and future work

This paper has investigated how different data mining approaches can be used to improve the prediction of first-year computer science university students' final performance on the basis of their participation in an on-line discussion forum. We have shown how a classification via clustering approach can be used instead of traditional classification algorithms. However, only the EM clustering algorithm obtained a similar accuracy and *F*-measure to the ones obtained with traditional classification algorithms. We have also proposed to discover class association rules on each cluster. An advantage of this kind of representation by class association rules is that the rule-based model obtained is much more specific to each cluster (group of students) than the previous rule-based model obtained with traditional classification algorithms. In this way, more powerful and interpretable models can be obtained using both centroids and class association rules to describe clusters. Our results indicated that two quantitative measures (the number of messages sent and the number of words written), together with the only qualitative measure (the average evaluation obtained in messages) and the two social network measures (the degree of centrality and the degree of centrality and prestige) were the most important attributes for predicting final student performance on the basis of usage data from on-line discussion forums. Using this set of selected features, instead of the whole set of features, obtained the highest accuracy and *F*-measure values. Finally, two message filtering criteria were shown to be very useful in our problem. Firstly, the use of only messages with content related to the course subject matter improved the accuracy of all the algorithms in all cases. Secondly, the use of only messages collected in the middle of the course does not improve the accuracy of any algorithm but the accuracy obtained is sufficient to be used as an early prediction system. In fact, the use of data from the middle of the course allows for early prediction that alerts instructors about those students who are potentially at risk of failing, so that they can provide some type of assistance or help.

Further research should be conducted to address the limitations of this paper. Firstly, in order to generalize on the result obtained, more executions must be backed up by using more data from different forums, from different type of courses, with more qualitative and social network participation indicators gathered on more varied dates during the course. Secondly, it would be very useful to automate the process of evaluating students' messages because evaluating all the messages is a tedious and time-consuming task for instructors. One easy way to

solve this problem would be to use student peer-reviews, i.e., the students themselves evaluate and score the content of the messages written by other students. A more difficult way of solving this problem would be to develop a specific text mining algorithm to attempt to automatically assign a score to a message based on its content.

Acknowledgement

References

Aggarwal, C. C. (2011). *Social network data analytics*. Springer.
Agrawal, A., & Choudhary, A. (2011). Identifying HotSpots in lung cancer data using association rule mining. In *IEEE international conference on data mining workshops* (pp. 995–1002).
Alstete, J. W., & Beutell, N. J. (2004). Performance indicators in online distance learning courses: a case study of management education. *Quality Assurance in Education, 12*(1), 6–14.
Araque, F., Roldan, C., & Salguero, A. (2009). Factors influencing university drop out rates. *Computer & Education, 53*, 563–574.
Bakharia, A., & Dawson, S. (2011). SNAPP: a bird's-eye view of temporal participant interaction. In *Proceedings of the 1st international conference on learning analytics and knowledge* (pp. 168–173).
Blair, R. C., & Higgins, J. J. (1985). *On the relative power of the paired samples t test and Wilcoxon's signed-ranks test*. ERIC Clearinghouse.
Bouckaert, R. R. (2007). *Bayesian network classifiers in Weka for version 3-5-6*. The University of Waikato.
Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.
Ceglar, A., & Roddick, J. F. (2006). Association mining. *ACM Computing Surveys (CSUR), 38*(2), 5.
Cheng, C. K., Paré, D. E., Collimore, L. M., & Joordens, S. (2011). Assessing the effectiveness of a voluntary online discussion forum on improving students' course performance. *Computers & Education, 56*(1), 253–261.
Cobo, G., García, G., Santamaría, E., Morán, J. A., Melenchón, J., & Monzo, C. (2011). Modeling students' activity in online discussion forums: a strategy based on time series and agglomerative hierarchical clustering. In *Proceedings of international conference on educational data mining* (pp. 253–258).
Cohen, W. (1995). Fast effective rule Induction. In *Twelfth international conference on machine learning* (pp. 115–123).
Cole, J., & Foster, H. (2007). *Using Moodle: Teaching with the popular open source course management system*. O'Reilly Media, Inc.
Denning, J. (2008). *The Hechinger Institute guide to education research for journalists*. The Hechinger Institute on Education and the Media, Teacher College, Columbia University.
De Wever, B., Schellens, T., Valcke, M., & Van Keer, H. (2006). Content analysis schemes to analyze transcripts of online asynchronous discussion groups: a review. *Computers & Education, 46*(1), 6–28.
Diaz, A. L. (2003). Personal, family, and academic factors affecting low achievement in secondary schools. *Electronic Journal of Research in Educational Psychology and Psychopedagogy, 1*(1), 43–66.
Dringus, L. P., & Ellis, T. (2005). Using data mining as a strategy for assessing asynchronous discussion forums. *Computers & Education, 45*(1), 141–160.
Duda, R. O., Hart, P. E., & Stork, D. G. (2000). *Pattern classification*. Wiley Interscience.
Freedman, D., Pisani, R., & Purves, R. (2007). *Statistics* (4th ed.). Norton & Company.
Freund, Y., & Mason, L. (1999). The alternating decision tree learning algorithm. In *Machine learning international workshop* (pp. 124–133).
Gaines, B. R., & Compton, P. (1995). Induction of ripple-down rules applied to modeling large databases. *Journal of Intelligent Information Systems, 5*(3), 211–228.
Hall, M., & Frank, E. (2008). Combining Naive Bayes and Decision Tables. In D. L. Wilson, & H. Chad (Eds.), *Proceedings of Twenty-First International Florida Artificial Intelligence Research Society Conference 15–17 May, 2008* (pp. 318–319). Coconut Grove, Florida, USA: AAAI Press.
Hämäläinen, W., & Vinni, M. (2011). *Classifiers for educational data mining*. London: Chapman & Hall/CRC.
Hochbaum, D. S., & Shmoys, D. B. (1985). A best possible heuristic for the k-center problem. *Mathematics of Operations Research, 10*(2), 180–184.
Holmes, G., Pfahringer, B., Kirkby, R., Frank, E., & Hall, M. (2002). Multiclass alternating decision trees. In *Machine learning: ECML 2002* (pp. 105–122).
Hou, H.-T., Chang, K.-E., & Sung, Y.-T. (2008). Analysis of problem-solving-based online asynchronous discussion pattern. *Educational Technology & Society, 11*(1), 17–28.
Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. *ACM Computing Surveys, 31*(3), 264–323.
Janecek, A., & Gansterer, W. (2008). On the relationship between feature selection an classification accuracy. In *JMLR: Workshop and conference proceedings, 4* (pp. 90–105).
Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C., Silverman, R., & Wu, A. Y. (2000, May). The analysis of a simple k-means clustering algorithm. In *Proceedings of the sixteenth annual symposium on computational geometry* (pp. 100–109).
Khan, T. M., Clear, F., & Sajadi, S. S. (2012). The relationship between educational performance and online access routines: analysis of students' access to an online discussion forum. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge* (pp. 226–229). ACM.
Kim, J., Shaw, E., & Ravi, S. (2010). Mining student discussions to profile participation and scaffold learning. In C. Romero, S. Ventura, M. Pchenizky, & R. Baker (Eds.), *Handbook of educational data mining* (pp. 299–310). Routledge.
Kleinman, S. (2005). Strategies for encouraging active learning, interaction, and academic integrity in online courses. *Communication Teacher, 19*(1), 13–18.
Klosgen, W., & Zytkow, J. (2002). *Handbook of data mining and knowledge discovery*. New York: Oxford University Press.
Koschmann, T., Kelson, A. C., Feltovich, P. J., & Barrows, H. S. (1996). Computer-supported problem-based learning: A principled approach to the use of computers in collaborative learning. *CSCL: Theory and Practice of an Emerging Paradigm*, 83–124.
Kotsiantis, S., Patriarcheas, K., & Xenos, M. (2010). A combinational incremental ensemble of classifiers as a technique for predicting students performance in distance education. *Knowledge-Based System, 23*(6), 529–535.
Kovacic, Z. J. (2012). Predicting student success by mining enrolment data. *Research in Higher Education Journal, 15*, 1–20.
Krakovsky, R., & Forgac, R. (2011). Neural network approach to multidimensional data classification via clustering. In *Intelligent systems and informatics (SISY), IEEE 9th international symposium on* (pp. 169–174).
Le Cessie, S., & Van Houwelingen, J. C. (1992). Ridge estimators in logistic regression. *Applied Statistics*, 191–201.
Li, N., & Wu, D. D. (2010). Using text mining and sentiment analysis for online forums hotspot detection and forecast. *Decision Support Systems, 48*, 354–368.
Lopez, M. I., Luna, J. M., Romero, C., & Ventura, S. (2012). Classification via clustering for predicting final marks starting from the student participation in forums (pp. 148–151).
Márquez-Vera, C, Cano, A, Romero, C, & Ventura, S (2013). Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data. *Applied Intelligence, 38*(3), 315–330.
Martin, B. (1995). *Instance-based learning: Nearest neighbour with generalisation* (Doctoral dissertation). University of Waikato.
McMillan, J. H., & Schumacher, S. (2006). *Research in education: A conceptual introduction* (6th ed.). Boston: Pearson Education Inc.
Memon, N., Xu, J. J., Hicks, D. L., & Chen, H. (2010). *Data mining for social network data* (pp. 1–8). Springer.
Moon, T. K. (1996). The expectation-maximization algorithm. *Signal Processing Magazine, IEEE, 13*(6), 47–60.
Morzy, M. (2009). On mining and social role discovery in internet forums. In *Social informatics, 2009. SOCINFO'09. International workshop* (pp. 74–79). IEEE.
Niu, Y., Wang, Y. M., Chen, H., Ma, M., & Hsu, F. A. (2007). Quantitative study of forum spamming using context-based analysis. In *Proceedings of the network and distributed system security (NDSS) symposium*.
Palmer, S., Holt, D., & Bray, S. (2008). Does the discussion help? The impact of a formally assessed online discussion on final student results. *British Journal of Educational Technology, 39*(5), 837–858.
Panda, M., & Patra, M. (2009). A novel classification via clustering method for anomaly based network intrusion detection system. *International Journal of Recent Trends in Engineering, 2*, 1–6.
Paredes, W. C., & Kenneth, K. S. (2012). Modelling learning & performance: a social networks perspective (pp. 34–42).
Park, J., & Sandberg, I. W. (1991). Universal approximation using radial-basis-function networks. *Neural Computation, 3*(2), 246–257.

Patel, J., & Aghayere, A. (2006). Student's perspective on the impact of a web-based discussion forum on student learning.

Pelleg, D., & Moore, A. (2000). X-means: extending k-means with efficient estimation of the number of clusters. In *Proceedings of the seventeenth international conference on machine learning, 1* (pp. 727–734).

Peña-Shaff, J. B., & Nicholls, C. (2004). Analyzing student interactions and meaning construction in computer bulletin board discussions. *Computers & Education, 42*(3), 243–265.

Platt, J. C. (1999). Fast training of support vector machines using sequential minimal optimization. In *Advances in kernels methods: Support vector learning*. MIT Press.

Quinlan, J. R. (1993)*C4. 5: Programs for machine learning*, Vol. 1Morgan Kaufmann.

Rabbany, R., Takaffoli, M., & Zaïane, O. (2011). Analyzing participation of students in online courses using social network analysis techniques. In *Proceedings of educational data mining* (pp. 21–30).

Raghavan, P., Catherine, R., Ikbal, S., Kambhatla, N., & Majumdar, D. (2010). Extracting problem and resolution information from online discussion forums *Management of data* (pp. 77).

Rau, P. L. P., Gao, Q., & Wu, L. M. (2008). Using mobile communication technology in high school education: motivation, pressure, and learning performance. *Computer & Education, 50*(1), 1–22.

Rodgers, J. L., & Nicewander, W. A. (1988). Thirteen ways to look at the correlation coefficient. *The American Statistician, 42*, 59–66.

Romero, C., Espejo, P., Romero, R., & Ventura, S. (2013). Web usage mining for predicting final marks of students that use Moodle courses. *Computer Applications in Engineering Education, 21*(1), 135–146.

Romero, C., & Ventura, S. (2013). Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 3*(1), 12–27.

Romero, C., Ventura, S., Espejo, P., & Hervás, C. (2008). Data mining algorithms to classify students. In *Proceedings of educational data mining* (pp. 20–21).

Romero, C., Ventura, S., & García, E. (2008). Data mining in course management systems: Moodle case study and tutorial. *Computers & Education, 51*(1), 368–384.

Ruck, D. W., Rogers, S. K., Kabrisky, M., Oxley, M. E., & Suter, B. W. (1990). The multilayer perceptron as an approximation to a Bayes optimal discriminant function. *Neural Networks, IEEE Transactions on, 1*(4), 296–298.

Sabourin, J. L., Mott, B. W., & Lester, J. C. (2012). Early prediction of student self-regulation strategies by combining multiple models. In *International conference on educational data mining* (pp. 156–159). Chania, Greece.

Shaw, R. S. (2012). A study of the relationships among learning styles, participation types, and performance in programming language learning supported by online forums. *Computers & Education, 58*(1), 111–120.

Slonim, N., Friedman, N., & Tishby, N. (2002). Unsupervised document classification using sequential information maximization. In *Proceedings of the 25th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 129–136).

Soni, J., Ansari, U., Sharma, D., & Soni, S. (2011). Predictive data mining for medical diagnosis: an overview of heart disease prediction. *International Journal of Computer Applications, 17*(8), 43–48.

Stormont, M., Reinke, W. M., Herman, K. C., & Lembke, E. S. (2012). *Academic and behavior supports for at-risk students: Tier 2 interventions*. The Guilford Press.

Sureka, A. (2011). Mining user comment activity for detecting forum spammers in YouTube (pp. 1–4).

Vonderwell, S. (2003). An examination of asynchronous communication experiences and perspectives of students in an online course: a case study. *The Internet and Higher Education, 6*(1), 77–90.

Witten, I. H., & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Yoo, J., & Kim, J. (2012). Predicting learner's project performance with dialogue features in online Q&A discussions (pp. 570–575).

Yukselturk, E. (2010). An investigation of factors affecting students participation level in an online discussion forum. *The Turkish Online Journal of Educational Technology, 9*(2), 24–32.

Zafra, A., Romero, C., & Ventura, S. (2011). Multiple instance learning for classifying students in learning management systems. *Expert Systems With Applications, 38*, 15020–15031.

Zhao, Y., Karypis, G., & Fayyad, U. (2005). Hierarchical clustering algorithms for document datasets. *Data Mining and Knowledge Discovery, 10*(2), 141–168.