Classification via clustering for predicting final marks based on student participation in forums

M.I. López, J.M Luna, C. Romero, S. Ventura
Department of Computer Science and Numerical Analysis
University of Córdoba
Córdoba, Spain
i32loqum@uco.es, i32luarj@uco.es, cromero@uco.es, sventura@uco.es

ABSTRACT

This paper proposes a classification via clustering approach to predict the final marks in a university course on the basis of forum data. The objective is twofold: to determine if student participation in the course forum can be a good predictor of the final marks for the course and to examine whether the proposed classification via clustering approach can obtain similar accuracy to traditional classification algorithms. Experiments were carried out using real data from first-year university students. Several clustering algorithms using the proposed approach were compared with traditional classification algorithms in predicting whether students pass or fail the course on the basis of their Moodle forum usage data. The results show that the Expectation-Maximisation (EM) clustering algorithm yields results similar to those of the best classification algorithms, especially when using only a group of selected attributes. Finally, the centroids of the EM clusters are described to show the relationship between the two clusters and the two classes of students.

Keywords

classification via clustering, prediction, classification, social networks analysis, forums

1. INTRODUCTION

Forums have recently become one of the leading means of peer communication on the internet. An internet forum is a web application for publishing user-generated content in the form of a discussion. Internet forums are sometimes called web forums, discussion boards, message boards, discussion groups, or bulletin boards [10]. The most important feature of internet forums is their social aspect. Many forums are active for a long period of time and attract a group of dedicated users, who build a tight social community within the forum. These social aspects of a discussion can highlight user interest in a specific topic. Current research activities use data mining to discover this information, especially in educational contexts, where online discussion forums are the best way to share ideas, post problems, comment on posts by other students, and obtain feedback [13]. In fact, mining group activities in a learning context provides quantifiable group profiles, which allow us to (1) evaluate the collaborative activity that the participants carry out, (2) analyse the link structure of the group, (3) compare the collaborative performance of different groups, and (4) predict behaviours and reveal link patterns [6] and collaboration trends. Mining data generated by students communicating using forum-like tools can help reveal aspects of their communication [14]; for example, the more students participate in the forum for a certain course, the more involved they will be in the subject matter of that course. Following this line, in this study we try to test whether or not there is a correlation between the participation of students in Moodle [4] forums and their final course marks. We have developed a new and specific Moodle module in order to obtain directly both statistics and social network information based on student forum usage data. We also propose the use of a classification via clustering approach to predict the final marks on the basis of our forum dataset.

The rest of the paper is organised as follows: a short theoretical background is presented in Section 2, the proposed methodology is outlined in Section 3, Section 4 describes the forum data used, Section 5 presents the experimental results, and conclusions and future research are outlined in Section 6.

2. BACKGROUND

Forums are one of the most commonly used tools in web-based teaching-learning environments because they play an important role in students' collaborative learning [12]. In fact, student activity in discussion threads can be a relevant source of information that facilitates the monitoring of tasks during the course by providing teachers with relevant indicators of student needs and weaknesses [3]. The use of data mining is a potential strategy for discovering and building alternative representations for the data underlying discussion forums [5]. The literature encourages analysis of forum interactions to reveal student characteristics and behaviour [1]; however, there is less published work on the use of data mining to predict student performance based on forum usage data. Classification is one of the oldest and most useful data mining tasks used to predict student outcomes, marks, or scores [15], and some works have used all the tracking data provided by Learning Management Systems (LMSs) in relation to visits and times, resources viewed, assessments, and activities in chat rooms, forums, etc. [2],[16]. However, the use of clustering for classification has not yet been applied in an educational context. Although clustering is normally an unsupervised process for grouping similar elements (students in this case) into clusters, classification can be performed based on clustering if we use the class information to evaluate the obtained clusters. This approach has been used to develop an anomaly-based network intrusion detection system [11], to predict heart disease in medical diagnosis [7], and to develop an effective system for classification of multidimensional data via clustering. [9]. However, we have found no work that uses only forum-usage data to predict final marks or that uses a classification via clustering approach in an educational context.

3. PROPOSED APPROACH

In this work, we propose to use a meta-classifier that uses a cluster for classification approach based on the assumption that each cluster corresponds to a class (see Figure 1). Firstly, the usage and interaction forum data have to be collected and preprocessed. Then, an optional attribute selection process can be applied (B), or not (A), in order to select only a group of attributes/variables or to use all available. Next, a clustering algorithm is 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 test data. In other words, the class attribute is not used in clustering, but it is used to evaluate the obtained clusters as classifiers.

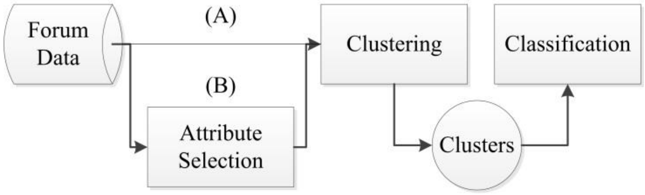


Figure 1: Proposed classification via clustering approach

For all cluster algorithms, 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. We use this approach to test if student participation in forums is related to whether they pass or fail the course.

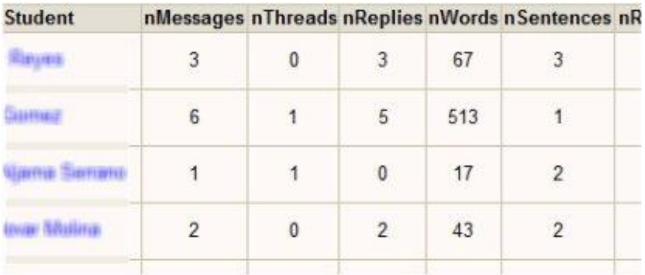
4. DESCRIPTION OF THE DATA USED

The dataset used in this work was gathered from a Moodle forum used by university students during a first-year course in computer engineering in 2011 (see Table 1).

Number of students	**Number of messages**	**Number of threads**	**Number of replies**
114	1014	81	933

Table 1: Some forum statistics

We developed a new module for Moodle specifically to obtain a summary dataset file with basic forum usage statistics (see Figure 2), to perform some analysis of social networks, to facilitate teacher evaluation of the messages, and to add the final marks of the students.



Student	nMessages	nThreads	nReplies	nWords	nSentences	nR
[redacted]	3	0	3	67	3	
[redacted]	6	1	5	513	1	
[redacted]	1	1	0	17	2	
[redacted]	2	0	2	43	2	

Figure 2: Screenshot of Moodle forum module

This tool not only enables us to visualise a list of variables for each student (see Table 2) but also allows us to save this summary information in a PDF file for report purposes or in an Excel file for data mining purposes.

Attribute	**Description**
nMessages	Number of messages sent per student
nThreads	Number of threads created per student
nReplies	Number of replies sent per student
nWords	Number of words written by student
nSentences	Number of sentences written by student
nReads	Number of messages read on the forum
tTime	Total time, in hours, spent on forum
aEvaluation	Average score of the messages
dCentrality	Degree centrality of the student
dPrestige	Degree prestige of the student
fMark	Final mark obtained by the student

Table 2: Variables of a student in a forum

The variables relating to forum usage are nMessages, nThreads, nReplies, nWords, nSentences, nReads, and tTime. The variable aEvaluation is the average score of the messages sent by the student. This evaluation of the contextual meaning of the messages has been done manually by the course teacher, who has read all the messages and assigned a score between 0 (bad) and 3 (very good). The two social network analysis measures are dCentrality and dPrestige, which are closely related to hyperlink analysis [8]. Both centrality and prestige are measures of the degree of prominence of an actor in a social network. Central or prominent actors are those that are extensively linked or involved with other actors. A person with extensive contacts (links) or communications with many other people in the organisation is considered more important than a person with relatively fewer contacts. 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.

Finally, the class or attribute to be predicted in this study is fMark, that is, the final mark obtained in the final exam at the end of the course. It has two possible values or labels: PASS or FAIL.

5. EXPERIMENTAL RESULTS

All our experiments were performed using Weka [17] and the previously described forum dataset. In order to test the accuracy of obtained classification models we used the 10-fold cross-validation method. All classifiers in Weka work in the same way under cross-validation. The model is built using just the instances in the training fold. The classification via clustering approach is based on the "clusters to classes" evaluation routine in the cluster evaluation code, which finds a minimum-error mapping of clusters to classes.

In the first experiment, we executed the following clustering algorithms provided by Weka for classification via clustering using all the available attributes (see Table 2): EM, FarthestFirst, HierarchicalClusterer, sIB, SimpleKMeans, and XMeans.

In the second experiment, we repeated all the previous executions using fewer attributes, based on the assumption that not all the available attributes are discriminative factors in the final marks. A process of feature selection was used to identify which attributes could have the greatest effect on our class (final mark). Weka provides a range of feature-selection algorithms from which we selected ten:CfsSubsetEval, ChiSquaredAttributeEval, ConsistencySubset-Eval, FilteredAttributeEval, FilteredSubsetEval, GainRatio-AttributeEval, InfoGainAttributeEval, OneRAttributeEval, ReliefFAttributeEval, and SVMAttributeEval. To rank the attributes, we counted the number of times each attribute was selected by each attribute-selection algorithm (see Table 3). Finally, we selected as the best attributes the first six attributes in the ranking, because these were selected by at least half (5) of the algorithms.

Attribute	**Frequency**
dCentrality	9
nMessages	8
nReplies, nWords	7
dPrestige	6
aEvaluation	5
nSentences, nReads, nThreads	3
tTime	1

Table 3: Attributes ranked by frequency of appearance

The previous clustering algorithms were then executed for classification via clustering but using only the six selected attributes (see Table 3, above the bold line). Table 4 shows the overall accuracy (rate of correctly classified students) using all the available attributes (A) and using only the six selected attributes (B).

Clustering algorithm	**(A)**	**(B)**
EM	**0.842**	**0.894**
FarthestFirst	0.526	0.535
HierarchicalClusterer	0.578	0.570
sIB	0.710	0.578
SimpleKMeans	0.666	0.640
Xmeans	0.666	0.640

Table 4: Accuracy of classification via clustering approach

An analysis of the results shown in Table 4 reveals that only one algorithm obtained a good level of accuracy. In fact, the EM algorithm obtained the highest accuracy in both cases (A and B) and the best overall accuracy (89.4%) when using only the six selected attributes. All the other clustering algorithms obtained much worse accuracy values (50%–70%) than EM, and, in general, there was no improvement by using only six attributes.

In the third experiment, we compared the accuracy of the previous classification via clustering approach with that of traditional classification algorithms by executing a representative number of classifications of different types:

- Rules-based algorithms: DTNB, JRip, NNge, and Ridor
- Trees-based algorithms: ADTree, J48, LADTree, and RandomForest
- Functions-based algorithms: Logistic, MultilayerPerceptron, RBFNetwork, and SMO
- Bayes-based algorithms: BayesNet and NaiveBayesSimple

Table 5 shows the accuracy obtained by the previous classification algorithms using all the attributes (A) and only the six selected attributes (B).

Algorithms	**(A)**	**(B)**
DTNB	0.859	0.833
JRip	0.833	0.815
NNge	0.842	0.807
Ridor	0.833	0.842
ADTree	0.859	0.842
J48	0.824	0.807
LADTree	0.868	0.850
RandomForest	0.850	0.833
Logistic	0.859	0.850
MultilayerPerceptron	0.842	0.868
RBFNetwork	0.868	0.886
SMO	0.868	0.886
BayesNet	**0.877**	0.842
NaiveBayesSimple	0.859	**0.894**

Table 5: Accuracy of classification algorithms

All the algorithms obtained a good accuracy with more similar values (80%–90%) than those obtained previously by the classification via clustering approach. The results indicate that some algorithms improve when using only six attributes, but others do not. The highest results are obtained by BayesNet when using all the attributes (87.7%) and NaiveBayesSimple when using only six attributes (89.4%), which is the best overall accuracy and is equal to that obtained by the EM algorithm.

Finally, we show the cluster centroids for the EM algorithm when using the six selected attributes that have yielded the best accuracy (see Table 6). The clusters-to-classes mapping done by the EM algorithm is such that cluster 0 is mapped to FAIL class and cluster 1 is mapped to PASS class.

Attributes	**Cluster 0**	**Cluster 1**
nMessages	1.2199	14.8905
nReplies	1.1599	13.6718
nWords	18.4599	668.8039
aEvaluation	0	0.7751
dCentrality	0.0011	0.1565
dPrestige	0	0.1021

Table 6: Cluster centroids obtained by EM algorithm

Cluster centroids describe the typical student for each group or cluster (see Table 6). We can see that the obtained clusters can be very informative from the point of view of classifying good and bad students. In fact, students who show a great level of participation in the forum (cluster 1) are classified as PASS, and students who show a very low level of participation in the forum (cluster 0) are classified as FAIL.

6. CONCLUSIONS

This paper demonstrates the potential of the classification via clustering approach in an educational context, using it to predict students' final marks on the basis of their participation in forums.

Based on the results obtained using several clustering and classification algorithms, we can answer the two initial questions:

a) Yes, student participation in the course forum was a good predictor of the final marks for the course. Another advantage of classification models based on mapping clusters to classes is that they are very simple and interpretable to instructors. In the case presented here, instructors only have to analyse the cluster centroids to know that students active in the forum pass the course and passive students fail.

b) Yes, the proposed classification via clustering approach obtained similar accuracy to traditional classification algorithms using our forum data. However, our proposed approach only had to obtain a good accuracy when using the EM algorithm (compared with traditional classification algorithms). On the other hand, the feature selection process can be useful to in reducing the number of attributes without losing reliability in classification. However, although some algorithms improved their classification performance when using only the selected attributes, the accuracy of other algorithms decreased.

However, in order to generalise the result obtained, the experiments must be repeated using different forum data to test if the same results are obtained or not, that is, if the EM clustering algorithm obtains again a high accuracy comparable with traditional classification algorithms. In the future, we hope to automate the process of evaluating student messages, because evaluating messages manually is a very difficult and time-consuming task for instructors. A data text mining algorithm could be used to automatically detect and classify types of messages and evaluate them. Finally, we are working on improving our Moodle forum module. We hope to develop a network analysis tool to graphically depict the forum interaction (sociograms) and to identify further measures than the two currently used (centrality and prestige) to provide valuable information for predicting students' final marks.

7. ACKNOWLEDGMENTS

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

8. REFERENCES

[1] A. R. Anaya and J. G. Boticario. Content-free collaborative learning modeling using data mining. *User Modeling and User-Adapted Interaction*, pp. 1–36. Springer, 2011.

[2] M. Calvo-Flores, E. Galindo, and M. Jiménez. Predicting students' marks from Moodle logs using neural network models. *Current Developments in Technology-Assisted Education*, 1:586–590, 2006.

[3] G. Cobo, D. García, E. Santamaría, J. A. Morán, J. Melenchón, and C. Monzo. Modeling students' activity in online discussion forums: a strategy based on time series and agglomerative hierarchical clustering. *Proceedings of Educational Data Mining*, 253-258, 2011.

[4] J. Cole and H. Foster. *Using Moodle: Teaching with the popular open source course management system.* O'Reilly Media, Inc., 2007.

[5] L.P. Dringus and T. Ellis. Using data mining as a strategy for assessing asynchronous discussion forums. *Computers & Education*, 45(1):141–160, 2005.

[6] L. Getoor. Link mining: a new data mining challenge. *ACM SIGKDD Explorations Newsletter*, 5(1):84–89, 2003.

[7] S. Jyoti, A. Ujma, S. Dipesh, and S. Sunita. Predictive data mining for medical diagnosis: An overview of heart disease prediction. *International Journal of Computer Applications*, 17(8):43–48, 2011.

[8] N. Memon, J.J. Xu, D.L. Hicks and H. Chen. Data Mining for Social Network Data. 1-8, Springer 2010.

[9] R. Krakovsky and R. Forgac. Neural network approach to multidimensional data classification via clustering. *Intelligent Systems and Informatics (SISY), 2011 IEEE 9th International Symposium on*, 169–174, IEEE2011.

[10] M. Morzy. On mining and social role discovery in internet forums. *Social Informatics, 2009. SOCINFO'09. International Workshop,* 74–79. IEEE, 2009.

[11] M. Panda and M. Patra. A novel classification via clustering method for anomaly based network intrusion detection system. *International Journal of Recent Trends in Engineering*, 2:1–6, 2009.

[12] R. Rabbany, M. Takaffoli and O. Zaïane. Analyzing participation of students in online courses using social network analysis techniques. *Proceedings of Educational Data Mining*, 21-30, 2011.

[13] P. Raghavan, R. Catherine, S. Ikbal, N. Kambhatla, and D. Majumdar. Extracting problem and resolution information from online discussion forums. *Management of Data*, 77, 2010.

[14] P. Reyes and P. Tchounikine, Mining learning groups' activities in forum-type tools. *Proceedings of the 2005 conference on computer support for collaborative learning: Learning 2005: the next 10 years!* 509–513. International Society of the Learning Sciences, 2005.

[15] C. Romero, S. Ventura, P. Espejo, and C. Hervás. Data mining algorithms to classify students. *Proceedings of Educational Data Mining*, 20-21, 2008.

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

[17] I.H. Witten, F. Eibe and M.A. Hall. Data Mining, Practical Machine Learning Tools and Techniques. Third Edition. Morgan Kaufman Publishers, 2001.