DISCOVERING STUDENT INTERACTIONS WITH A COLLABORATIVE LEARNING FORUM THAT PREDICT GROUP COLLABORATION PROBLEMS

Shuangyan Liu, Mike Joy

Department of Computer Science, University of Warwick (United Kingdom)
Shuangyan.Liu@warwick.ac.uk, M.S.Joy@warwick.ac.uk

Abstract

A recent nationwide survey in the UK has revealed that student-induced collaboration problems exist widely in web-based collaborative group work undertaken by undergraduate computing students who are using asynchronous collaboration tools. Assessing these collaboration problems can assist teachers or moderators to understand and evaluate how individual students perform in collaborative groups as well as help students to reflect on their own learning actions.

A number of studies have indicated that quantitative data resulting from student interactions with an asynchronous collaboration tool, such as a forum, can account for the behaviours of individual students and collaborative groups. This poses a question on which aspects of student usage of such a tool predict group collaboration problems.

This paper investigates the roles of various student interactions with a learning forum in order to ascertain the existence of different group collaboration problems. A particular focus of interest has been learning forums, since forums have become broadly adopted tools to support online group collaboration. The types of collaboration problems were drawn from previous research that identified the main student-induced collaboration problems.

A data set was collected for 87 undergraduates who participated in a web-based computer science group project. It consists of two kinds of data. The first comprises student interaction data which were collected from a learning forum system on which the group project was undertaken. The second set of data relates to assessment of group collaboration problems, and was gathered through a questionnaire delivered to the students who participated in the group project.

Multinomial logistic regression analysis has been applied for modelling the relationship between a response variable corresponding to the existence of a group collaboration problem and several predictor variables representing various student interactions with a learning forum.

A set of predictive models were produced by the regression analysis, each representing a statistically significant combination of student interactions that predict the existence of one of the collaboration problems in question. The findings reveal that indicators including the number of posts that were created and replied to by individual students, and the number of times that a student viewed a discussion on a learning forum, contribute significantly in predicting the collaboration problems which were identified. The results also demonstrate how the existence of a problem fluctuates with the alterations in the value of an indicator variable.

The goodness-of-fit of the identified predictive models was measured by the Pearson chi-square test and the results of this test indicate that the models fit the sample data well. The average rate of correct classification by the models was approximately 80%, which means the regression method performed well on the sample data set.

The outcomes of this research can help teachers to assess problems in web-based collaborative group work and also can be used to develop tools for automatically diagnosing group collaboration problems in web-based collaborative learning environments.

Keywords: Group collaboration problems prediction, learning forum, undergraduate group project, multinomial logistical regression, predictive model.

1 INTRODUCTION

A popular tool for supporting online collaborative learning is the discussion forum, which can facilitate the transfer of knowledge and increase the quality of the learning experience [1]. Forums are considered to be powerful as long as students can actively engage with them. A number of research such as [2] [3] [4] has revealed that the majority of students who have experiences of online group work tend to possess the problem of being unwilling to participate actively in online collaboration. This is known as the "lack of individual accountability" problem [2]. The findings from a recent survey in the UK [5] indicate that three sub-categories exist which can reveal the "lack of individual accountability" problem: "not contributing much in online discussions" (denoted as CP-1), "not actively meeting the deadlines" (CP-2), and "not actively completing the assigned work" (CP-3). Several studies in interaction analysis for collaborative activities such as [6] [7] suggest that quantitative data relating to student interactions with a collaborative learning system can reveal the behaviours of individual students. This poses a question that motivates this research, which is not addressed in current literature:

Which aspects of student usage of a collaborative learning forum that can predict the major group collaboration problems relating to students' lack of individual accountability?

This paper aims to investigate the relationship between various student interactions with a collaborative learning forum and the existence of the group collaboration problems in question. A predictive modelling methodology was adopted to draw this relationship.

The remaining sections are structured as below. Section 2 gives a brief explanation of the predictive modelling methodology itself and the related work. Section 3 presents how the predictive modelling process was conducted which includes the identification of the indicators relating to student interactions, the data collection and preparation procedures, which statistical analysis technique was used and the concrete modelling steps. Following that, the results of the predictive modelling are provided in Section 4. Section 5 gives a reflection of the findings obtained from this study. Conclusions are presented in Section 6.

2 PREDICTIVE MODELLING

Predictive modelling [8] [9] offers such a methodology that can produce predictive models which quantitatively define the relationships between the occurrences of an event (i.e. the response or dependent variable) and the factors that can indicate the occurrences of such an event (i.e. the predictors or independent variables). The produced predictive models can then be used to compute values of the response variable for a given set of predictors. The procedure of predictive modelling involves building a data set, which collects empirical data about the response variable and the potential predictors. Then statistical techniques can be applied for estimating and validating the predictive models using the constructed data set.

The methodology of predictive modelling has been widely applied in different fields. In higher education, predictive modelling has been used in a number of areas including but not limited to enrolment management, retention and graduation analysis, and donation prediction [10] [11]. In these areas, the majority of time spent on a modelling project is establishing the dataset to be used. It usually requires at least one year of historical data for building such a dataset.

In the field of online learning, Balaji and Chakrabarti have adopted the methodology to investigate the factors that influence interactions and learning in online discussion forums [12]. The data for this study were collected from two sources. One consisted of the postings relating to the discussions on the content covered in an MBA course. The authors have given no details of what aspects of the postings were examined. The other was a post-course survey that gathered student perceptions of the various factors that affect the effectiveness of the interactions and learning in online discussion forums. Similar data collection methods were adopted for the study presented in this paper. In Liu and Cheng's study regarding the effect of group discussion on web-based collaborative learning [13], predictive modelling was used to investigate the relationship between the discussions categorized as "social talk" and "group-task-related dialogue" and the group learning outcome.

3 METHODOLOGY

3.1 Indicators of Collaboration Problem Existence

In order to discover the types of student interaction data that potentially indicate the existence of the collaboration problems (as mentioned in Section 1: CP-1−CP3), an analysis of the problem scenarios and a further literature review were carried out.

Talavera and Gaudioso [14] suggested that the number of threads started by an individual student can indicate the degree of involvement to produce a contribution and the number of messages that a student replied can form a measure of how they are promoting discussion. In addition to this, Nakahara et al. [7] pointed out another three indicators in their study that can reveal the degree of participation in an online BBS forum: the “number of posts”, the “number of times posts are read” and “ratio of total forum posts created to replies”. In other studies including [15] [6], the number of messages has also been noted as an indication of activity for individual students or groups. Furthermore, Bratitsis and Dimitracopoulou’s study [6] on computer-supported interaction analysis for forums suggested that the proportion of the number of posts made by an individual student to the overall number of posts made by the group that the student belongs to can reveal the contribution status of the student for the group activity and also evince whether the student has actively participated in the group activity or not. Besides, Bratitsis and Dimitracopoulou also noted in [16] that the number of posts made by a student and the number of times that the student read a post during a time period can identify the participation peak for this period.

Apart from the indicators identified from literature, some hypothetical indicators have been proposed to complete the list of indicators. These hypothetical indicators are considered to be related to the existence of the collaboration problems in question. Among these indicators, some are quantitative data related to student interactions with a forum system. One example of a quantitative hypothetical indicator is: the number of times that an individual student logged in to a group forum (noted as “forum_login”). Other hypothetical indicators are qualitative data related to student interactions with a forum system, for example, the pattern of the participation peak over a time period for an individual student (noted as “timeperiod_post_pattern”).

In summary, for each of the three collaboration problems examined, six to seven indicators were identified, and served as the basis for guiding the data collection process.

3.2 Data Collection and Preparation

The data collection procedure aimed to collect two kinds of data: data relating to the indicators and data about the assessment of collaboration problems possessed by students who completed part of a group project. Next, the background of the group project where the data originated from is presented.

The group project was a part of a first year undergraduate module in the authors’ department and lasted for one term. 95 students joined the module and were allocated into 19 groups. The task for each group was to construct a set of questions for other groups to answer and also answer some questions authored by other groups on a collaborative learning forum that was assigned to each group. The questions related to the concepts of the operating system UNIX which were taught in lectures and practised during lab sessions for this module. The private group forum was used for group discussions relating to the group project. A general forum was also set up so that all the groups were able to post their questions and answers. Both the private group forums and the general forum were created and maintained using the Warwick Forums system.

Warwick Forums was capable of capturing data relating to student interactions with the system such as the number of times a user has viewed threads in a forum and the number of times a user has logged in to a forum. This feature enabled collection of the student interaction data required for the predictive modelling.

Apart from the above procedure, a questionnaire was designed for collecting data about the problems that the students experienced in the group project. The questionnaire was targeted at the participating students and completed by them at the end of the term. The ethical consent for the data collection procedure was approved by the author’s department. Data were collected for 87 students who were formed into 18 collaborative groups.

A data set namely *Forum* was constructed based on the collected data. There were three tables defined corresponding to the data prepared for the problems CP-1, CP-2 and CP-3 respectively:

Forum-1, Forum-2 and *Forum-3*. Each of the tables defined values of a response variable (i.e. the categories of problem existence) and values of the predictors (i.e. the indicators) relating to a collaboration problem. These three tables were utilized for the predictive modelling process as presented in Section 3.4.

3.3 Statistical Analysis Technique

Logistic regression has become a standard method of modelling the relationship between a binary or dichotomous response variable and one or more explanatory variables in many fields [17]. Multinomial logistic regression (MLR) is an extension of the logistic regression in the case where the response variable is nominal with more than two levels. In this study, multinomial logistic regression was adopted for building the predictive models for ascertaining the existence of the collaboration problems CP-1, CP-2 and CP-3. This is because the response variables defined have three categories ("yes", "maybe", "no").

3.4 The Modelling Process

Multinomial logistic regression analysis was performed on the three tables *Forum-1, Forum-2* and *Forum-3* in the *Forum* data set using the SPSS statistical software (version 19). The modelling on the table *Forum-1* produced the predictive model I for describing the relationship between the existence of the problem CP-1 and its predictors. Additionally, the modelling on the table *Forum-2* produced the predictive model II for describing the relationship between the existence of the problem CP-2 and its predictors. Last, the modelling on the table *Forum-3* produced the predictive model III for describing the relationship between the existence of the problem CP-3 and its predictors. Next, the results of the modelling process are presented.

4 RESULTS

4.1 Predictive Model I

Table 1 presents the results of the MLR analysis for variables predicting the collaboration problem CP-1. Of the six predictor variables for the problem CP-1, three were able to separate the cases for problem existence: "yes", "maybe", and "no". The three predictors include "post_create" (i.e. the number of posts that were created by a student in the group forum), "post_reply" (i.e. the number of posts that were replied to by a student in the group forum), and "thread_view" (i.e. the number of times that a student viewed the threads in a group forum). This final model was statistically significant [-2 Log likelihood=104.081; $x^2(6)$ =66.895; P=0.000].

Table 1. Summary of multinomial logistic regression analysis for variables predicting the collaboration problem "not contributing much in online discussions" (CP-1) (*N*=87)

Problem CP-1[a]		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp(B)	
								Lower Bound	Upper Bound
Yes	Intercept	1.642	.461	12.665	1	.000			
	thread_view	.118	.049	5.806	1	.016	1.125	1.022	1.239
	post_reply	-.934	.334	7.824	1	.005	.393	.204	.756
	post_create	-4.327	1.348	10.307	1	.001	.013	.001	.185
Maybe	Intercept	.276	.497	.308	1	.579			
	thread_view	.015	.031	.240	1	.625	1.015	.956	1.078
	post_reply	.148	.116	1.644	1	.200	1.160	.925	1.455
	post_create	-1.957	.986	3.940	1	.047	.141	.020	.976

a. The reference category is: No.

The significance of the predictors in the model was measured with the Likelihood ratio tests—"thread_view" [-2 Log likelihood=114.262; $x^2(2)$ =10.182; P=0.006], "post_reply" [-2 Log likelihood=122.930; $x^2(2)$ =18.849; P=0.000], and "post_create" [-2 Log likelihood=135.599; $x^2(2)$

=31.518; P=0.000]. The indicators “forum_view” (i.e. the number of times that a student viewed a group forum), “forum_login” (i.e. the number of times that a student logged in to the group forum) and “ratio_stupost_grpost” (i.e. the ratio of the overall number of posts that a student made to the overall number of posts that a group made) failed to meet the 0.05 significance criterion and were dropped from the final model.

The goodness-of-fit of the model (i.e. how well the model fits a set of observations) was measured by with the Pearson chi-square test. The result of the test was not statistically significant [x^2(136) =139.853, P=0.393], which indicates that the model fits the data well. This is due to the value of P is bigger than 0.05 and therefore the null hypothesis is not rejected. The Pearson chi-square test verifies the null hypothesis that the observed frequency distribution of the outcome categories of the response variable is consistent with a particular theoretical distribution (i.e. the chi-square distribution). Moreover, the overall rate of correct classification for predictive model I on the full dataset (*N*=87) is 74.7%, which is satisfied.

4.2 Predictive Model II

The results of the MLR analysis for variables predicting the collaboration problem CP-2 are shown in Table 2. Of the seven predictor variables for the problem CP-2, two were able to separate the cases for problem existence: “yes”, “maybe”, and “no”. The two predictors are “post_reply” and “post_create”. This final model was statistically significant [-2 Log likelihood=89.591; x^2(4) =71.891; P=0.000].

The significance of the two predictors in the model was “post_reply” [-2 Log likelihood=111.482; x^2(2) =21.891; P=0.000], and “post_create” [-2 Log likelihood=108.269; x^2(2) =18.678; P=0.000]. The indicators “forum_view”, “thread_view”, “forum_login”, “timeperiod_post_pattern” (i.e. the pattern of posting that a student made during a particular time period) and “timeperiod_view_pattern” (i.e. the pattern of viewing that a student had during a particular time period) failed to meet the 0.05 significance criterion and were dropped from the final model.

Table 2. Summary of multinomial logistic regression analysis for variables predicting the collaboration problem “not actively meeting the deadlines” (CP-2) (*N*=87)

Problem CP-2[a]		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp(B)	
								Lower Bound	Upper Bound
Yes	Intercept	3.794	.775	23.939	1	.000			
	post_reply	-.846	.228	13.801	1	.000	.429	.275	.671
	post_create	-1.851	.713	6.740	1	.009	.157	.039	.635
Maybe	Intercept	1.699	.786	4.668	1	.031			
	post_reply	-.307	.172	3.180	1	.075	.736	.525	1.031
	post_create	-.749	.362	4.282	1	.039	.473	.233	.961

a. The reference category is: No.

The goodness-of-fit of the model was not statistically significant [x^2(138) =142.815, P=0.372>0.05], which indicates that the model fits the data well. This is due to the value of P is bigger than 0.05. Therefore, the null hypothesis is not rejected, which states that the observed frequency distribution of the response variable is consistent with the chi-square distribution. The overall rate of correct classification for predictive model II is satisfied (79.3%).

4.3 Predictive Model III

Table 3 presents the results of the MLR analysis for variables predicting the collaboration problem CP-3. Of the six predictor variables for the problem CP-3, two were able to separate the cases for problem existence: “yes”, “maybe”, and “no”. The two predictors include “post_reply” and “post_create”. This final model was statistically significant [-2 Log likelihood=58.203; x^2(4) =107.920; P=0.000].

The two identified predictors were statistically significant: “post_reply” [-2 Log likelihood=95.120; x^2(2) =36.917; P=0.000], and “post_create” [-2 Log likelihood=99.183; x^2(2) =40.980; P=0.000]. The indicators “forum_view”, “thread_view”, “forum_login”, and “ratio_stupost_grpost” failed to meet the 0.05 significance criterion and were dropped from the final model.

Table 3. Summary of multinomial logistic regression analysis for variables predicting the collaboration problem “not actively completing the assigned work” (CP-3) (*N*=87)

Problem CP-3[a]		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp(B)	
								Lower Bound	Upper Bound
Yes	Intercept	7.459	1.921	15.084	1	.000			
	post_reply	-1.637	.426	14.762	1	.000	.195	.084	.449
	post_create	-5.136	1.703	9.098	1	.003	.006	.000	.166
Maybe	Intercept	4.338	1.829	5.628	1	.018			
	post_reply	-.555	.314	3.116	1	.078	.574	.310	1.063
	post_create	-3.137	1.279	6.020	1	.014	.043	.004	.532

a. The reference category is: No.

The goodness-of-fit of the model was not statistically significant [x^2(138) =71.263, P=1.000], which indicates that the model fits the data well. Compared with the other two predictive models, the overall rate of correct classification of predictive model III is the highest (83.9%). This indicates that predictive model III performed well on the full dataset.

5 DISCUSSION

In order to give a reflection on the predictive models, the findings from the study are discussed. It is revealed that each of the established predictive models has identified and prioritized (in terms of relative impact on the final model) the types of student interactions with a collaborative learning forum that contribute to the prediction of the relevant collaboration problem that is examined.

Regarding the predictive model I, the findings reveal that students who have created and replied to more posts in their group forums are less likely to have the problem CP-1 (i.e. “not contributing much in online discussions”). The positive relationship between the number of posts that a student has created or replied to and the level of contribution that the student has made in online discussions is consistent with the results of Talavera and Gaudioso [14] regarding student interactions with forum systems and their contributions to online discussions. Moreover, the finding that students who have viewed the threads in a forum many times were more likely to present the problem CP-1 was unexpected since it was believed that students with much contribution to online discussions should have viewed the threads in their group forums frequently. A possible explanation for this unexpected relationship can be that these students tended to observe the written discourse occurring online between other students but did not actively participate in the group discussion. This type of students is the so-called “witness learner” according to [18].

Concerning the predictive model II, the findings indicate that students who have created and replied to more posts in their group forums are less likely to have the problem CP-2 (i.e. “not actively meeting the deadlines”). These findings agree with Dimitracopoulou [16] with regard to the result that the number of posts made by a student can help identify the participation peak of the student in online discussions. However, the finding that the hypothetical indicator “timeperiod_post_pattern” (i.e. the pattern of posting that a student made during a particular time period) did not significantly affect the prediction of the problem CP-2 and was not included in the final model was somewhat surprising. A further analysis of the data relating to the variable “timeperiod_post_pattern” which were used to generate the predictive model II reveals that the observed data relating to one of the pattern of the variable “timeperiod_post_pattern” (i.e. a student made few posts at the beginning of the group work but created many posts while the deadline was approaching) dispersed evenly in all the categories of the response variable. This indicates that the variable “timeperiod_post_pattern” is not sufficient to classify the existence of the examined collaboration problem.

In terms of the predictive model III, the findings suggest that students who have created and replied to many posts in their group forums are less likely to have the problem CP-3 (i.e. “not actively completing the assigned work”). This is consistent with the results of studies [15] [6] which revealed that the number of messages was an indication of activity for individual students or groups. Moreover, the finding that the hypothetical indicator “ratio_stupost_grpost” did not significantly affect the prediction of the problem CP-3 was unexpected. This is because Bratitsis and Dimitracopoulou’s study [6] on usage interaction analysis in asynchronous discussions suggested that the proportion of the number of posts made by an individual student to the overall number of posts made by the group that the student participated in can reveal the contribution status of the student for the group activity. A further analysis of the development data reveals a special case in the data sample which can lead to this unexpected relationship. There was a student who contributed 8% of the overall group posts but was assessed not having the problem CP-3. However, this case should not be excluded from the data sample, because even though the student made relatively small number of posts (compared with the ones in other groups), he or she contributed the second largest number of posts while the remaining students had no contribution to the overall posts. Thus, the student was believed to be active enough to complete the assigned work in the group.

6 CONCLUSIONS

Although there is much research in identifying the types of student interactions that can indicate student engagement in online group collaboration, no research to our knowledge has quantitatively defined the relationships between the occurrences of student problems that may arise from group collaboration and the types of student interactions that can predict these problems. The present study revealed that a student’s interactions with a learning forum including the number of posts that a student has created or replied to and the number of times that a student has viewed a thread on the forum were all predictive of the major collaboration problems in question. These findings extend the discussions surrounding collaborative process analysis and provide insights about the effect of student interactions with a learning forum on predicting the existence of the collaboration problems examined.

Beyond presenting the initial assessment of the predictive models, we are concluding an evaluation of the constructed predictive models, which comprised several experiments with a test dataset to investigate the reproducibility of the models on independent data which are similar to the data where the models originated from.

Given the fact that current collaborative learning environments provide little or no support for monitoring the collaborative process and thus assessing the collaboration problems encountered by individual students, the established predictive models can assist the development of software mechanism for automatically diagnosing the group collaboration problems in collaborative learning environments.

REFERENCES

[1] H. Kanuka, “An exploration into facilitating higher levels of learning in a text-based internet learning environment using diverse instructional strategies,” Journal of Computer-Mediated Communication, vol.10, no. 3, 2005.

[2] H. An, S. Kim and B. Kim, “Teacher perspectives on online collaborative learning: Factors perceived as facilitating and impeding successful online group work,” Contemporary Issues in Technology and Teacher Education, vol.8, no. 1, pp. 65-83, 2008.

[3] M. Herrick, M.-F. G. Lin and C. Huei-Wen, “Online discussions: The effect of having two deadlines,” in Proceedings of Society for Information Technology & Teacher Education International Conference 2011, 2011, pp. 344-351.

[4] B. Millis. Managing—and motivating!—distance learning group activities. [Online]. Available: http://www.tltgroup.org/gilbert/millis.htm

[5] S. Liu, M. Joy and N. Griffiths, “Students' perceptions of the factors leading to unsuccessful group collaboration,” in Proceedings of the 10th IEEE International Conference on Advanced Learning Technologies (ICALT 2010), 2010, pp. 565-569.

[6] T. Bratitsis and A. Dimitrkopoulou, “Data recording and usage interaction analysis in asynchronous discussions: The D.I.A.S. System,” in Proceedings of the Workshop on Usage Analysis in Learning Systems, 2005.

[7] N. Jun, H. Shinichi, Y. Kazaru and Y. Yuhei, “Itree: Does the mobile phone encourage learners to be more involved in collaborative learning?,” in Proceedings of the 2005 Conference on Computer Supported Collaborative Learning: The next 10 years!, 2005, pp. 470-478.

[8] S. Chatterjee and A. S. Hadi, “Regression analysis by example,” Hoboken, N.J.: Wiley-Interscience, 2006.

[9] R. Nisbet, J. F. Elder and G. Miner, “Handbook of statistical analysis and data mining applications,” Amsterdam; Boston: Academic Press/Elsevier, 2009.

[10] T. Bohannon, “Predictive modelling in higher education,” in SAS Global Forum, 2007.

[11] W. Lan, Z. Wei-Wei and L. Yu-Fen, “An empirical study on the prediction model of postgraduate education in Hebei province,” in Proceedings of the 2010 International Conference on Machine Learning and Cybernetics (ICMLC), pp. 1327-1331.

[12] M. S. Balaji and D. Chakrabarti, “Student interactions in online discussion forum: Empirical research from 'media richness theory' perspective,” Journal of Interactive Online Learning, vol.9, no. 1, pp. 1-22, 2010.

[13] Z. F. Liu and S. S. Cheng, “The student satisfaction and effect of group discussion on networked cooperative learning with the portfolio assessment system,” International Journal of Education and Information Technologies, vol.1, no. 3, pp. 161-166, 2007.

[14] L. Talavera and E. Caudioso, “Mining student data to characterize similar behaviour groups in unstructured collaboration spaces,” in Proceedings of the Workshop on AI in CSCL, 2004, pp. 17-23.

[15] T. Bratitsis and A. Dimitracopoulou, “Indicators for measuring quality in asynchronous discussion forums,” in Proceedings of International Conference on Cognition and Exploratory Learning in Digital Era (CELDA 2006), 2006, pp. 8-10.

[16] A. Dimitracopoulou, “Computer based interaction analysis supporting self-regulation: Achievements and prospects of an emerging research direction,” Technology, Instruction, Cognition and Learning, vol.6, no. 4, pp. 291-314, 2009.

[17] D. W. Hosmer and S. Lemeshow, “Applied logistic regression,” New York: Wiley, 2000.

[18] M. F. Beaudoin, “Learning or lurking? Tracking the 'invisible' online student,” Internet and Higher Education, vol.5, no. 2, pp. 147-55, 2002.