Assessing the Quality of Synchronous Network Learning Activities using Machine Learning Techniques

Georgios Kahrimanis, Eleni Mikroyannidi, Nikolaos Avouris

Human-Computer Interaction Group, E&CE Dept. University of Patras, GR-26500 Rio Patras, Greece
kahrimanis@ece.upatras.gr, elemikrogian@upnet.gr, avouris@upatras.gr

Abstract

During network-mediated synchronous collaborative activities there is need for supporting reflection of the learners involved, by providing them with meaningful feedback on the state of group activity and the quality of their collaborative effort . In order to produce timely feedback to the partners, we need to automate processing activity data and producing meaningful measures of the quality of collaboration to be fed back to the students. This paper presents a study investigating applicability and effectiveness of machine learning techniques in the process. The objective is to use different classification algorithms for assessing quality of collaboration using a set of quantitative indices produced by the collaborative learning environment. Collaboration quality, however, is a term that needs first to be defined using a relevant scheme. After developing a scheme for assessment of collaboration quality in various axes, this study shows encouraging results in the performance of machine learning algorithm prediction scores.

Keywords

Network Supported Collaborative Learning, Machine Learning, Rating Scheme

Introduction

During network-mediated synchronous collaborative activities there is need for supporting reflection of the learners involved, by providing them with meaningful feedback on the state of group activity. Reflection is considered to be crucial for learners in order to improve their practices but providing correct and useful feedback on time, is not a trivial task. In order to produce timely feedback to the partners, we need to automate processing activity data and producing meaningful measures of the quality of collaboration to be fed back to the students or even to supervisors who may make productive use of them. Facilities of network technologies offer new possibilities for automation of interaction analysis. Contemporary collaboration support tools produce and maintain logs of sequential events that can be later manipulated by analysis tools. Then, metrics can be easily calculated that range from simple aggregations of types of events to more sophisticated quantitative indices. However, the rough use of such quantitative measures is not usually enough for making reliable judgements on collaboration quality.

This paper presents a study investigating applicability and effectiveness of machine learning techniques in the process. The objective is to use different classification algorithms for assessing quality of collaboration using a set of quantitative indices produced by the collaboration support environment. Collaboration quality, however, is a term that needs first to be defined using a relevant scheme developed and applied with methodological rigour. Thus, a rating scheme developed for similar purposes (Meier, Spada and Rummel, 2007) was adopted and transformed in order to be useful for the specific study. In the first phases, development of the rating scheme was a trial-and-error process that involved changes of dimensions and scales of rating until researchers in this study reached some methodologically satisfactory results.

After the rating scheme reached its final version for the study, the logfiles of the activity of a number of typical collaborating student dyads were segmented and rated according to it. Then, for each dimension of the rating scheme, a study was conducted in order to find out if the quantitative indices of the activity

could be used for calculating the corresponding dimension accurately enough. Various established classification algorithms were used in order to calculate the value of the dimension. Bayes networks, decision trees, metalearing algorithms, among others, constituted a powerful palette of machine learning tools that produced similar results. The study shows encouraging results in the performance of machine learning algorithm prediction scores for most of the dimensions of collaboration.

In this article, the collaborative activities under investigation and the collaboration tool used are first described, followed by a section on the development of the rating scheme for assessing collaboration quality. After a subsequent section that describes the application of the scheme and argues for the reliability of the method through established statistical measures, the final phase of the application of machine learning algorithm is described. Findings are then illustrated and discussed. The article concludes with some concerns on the applicability of the findings as long as on the shortcomings and further research directions.

Collaborative Activity

The typical collaboration activities studied involved dyads of students following a distance learning computer science course. The dyads were asked to solve an algorithm problem and develop solutions in the form of flowchart diagrams as part of the 'Introduction to Computer Science' module of the Computer Science curriculum of the School of Sciences and Technology, Hellenic Open University. The assignment involved collaborative building of an algorithm of a bus-ticket venting machine with certain constraints (Xenos, Avouris, Komis, Stavrinoudis and Margaritis, 2004). The expected solution is the algorithm of operation of the venting machine in the form o a flowchart diagram.
The dyads studied used the network based collaboration support tool *Synergo* (Avouris, Margaritis & Komis, 2004). Synergo provides a frame of reference, a shared drawing space through which various diagrammatic representations can be built jointly by a group of collaborating partners and a chatting tool for direct text-based communication (fig.1). In addition, Synergo offers a rich set of analysis tools that serve for analysis of collaboration by teachers or researchers. Synergo calculates a number of quantitative metrics of the activity. Statistics of logged actions are automatically provided. The categorization of actions according to a framework such as OCAF (Avouris, Dimitracopoulou, and Komis, 2003) is also helpful for more advanced quantitative results. Messages and different types of actions on the shared workspace can be treated differently and this is important for the investigation of different patterns of activity, e.g. measures like the "symmetry" of collaboration or the "symmetry" of interaction are considered important indicators of balanced contribution of the partners involved.

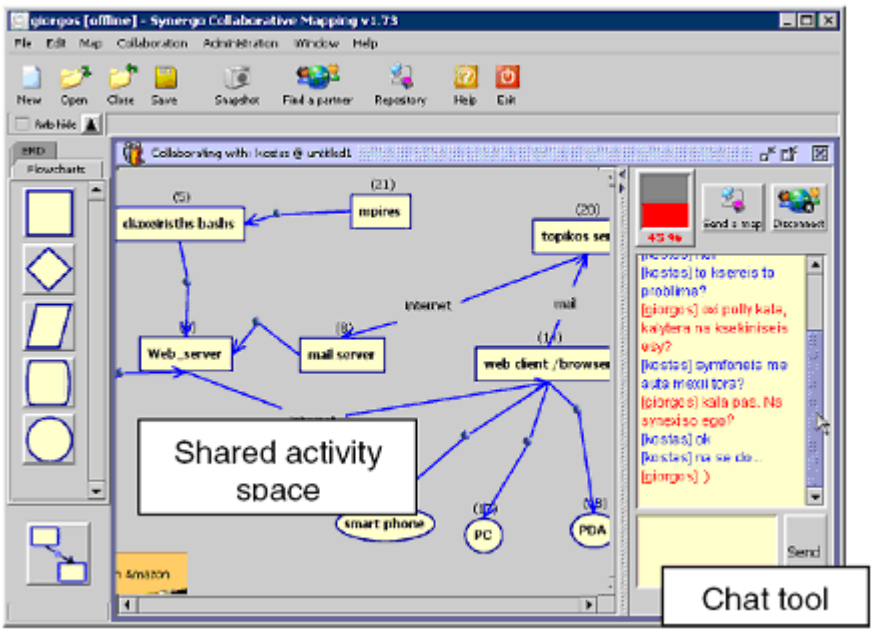

Figure 1: Synergo collaboration environment.

 ISBN No: 978-1-86220-206-1

Moreover, the Synergo analysis tool allows researchers to play back students' actions in a video-like format, reproducing the activity, based on the sequential log files of events that Synergo generates. In that way, the evaluator can review the whole activity, navigate through different episodes that occur and require more detailed inspection. Thus, this facility was very important for the application of the rating scheme as described later in this article.
Synergo gathered a single log file and quantitative measures for each dyad so that, taking into account some data losses due to technical problems, researchers in this study had finaly a consistent dataset of 15 dyads at their disposal.

Quality of Collaboration Rating Scheme

In order to provide a point of reference that is external to machine learning techniques described later in this document, the quality of collaboration is measured using the dimensions of a quality of collaboration rating scheme, developed for manual use by evaluators (Meier, Spada, Rummel, 2007). The scheme includes the following dimensions that were found to represent important aspects of the collaborative activity: Quality of communication has to be ensured in two dimensions. Seamless collaboration flow and the need to sustain mutual understanding are prerequisites for the success of a collaborative activity. Focusing on the task itself, information pooling (relating to the mutual exchange of information between the partners) or tutoring instances are considered important for specific task designs. In-depth negotiation of the solutions and the reach of consensus after critical investigation are also of main significance. In addition, coordination of the whole process is deemed beneficial for the success of the activity. Planning the task and dividing sub-tasks efficiently by handling time constraints are also appreciated. Finally, the scheme gives emphasis on indications that reveal students' motivation towards collaboration and not introvert behaviour and dedicate themselves to the task.

Due to some important differences in the intended use and the collaboration setting studied in this work and the one used for the development of the rating scheme, some modifications were necessary in order to make the scheme applicable and efficient. Apart from changes to the descriptions of dimensions that served for reliable rating in the new setting, two dimensions that were related to cognitive aspects and were sensitive to the specific problem domain were significantly extended. For this reason a well-established schema for the description of learning objectives (Bloom, Englehart, Furst, Hill and Krathwohl, 1956) was combined with the quality of collaboration original rating scheme. The cognitive dimensions were rearranged in four new dimensions that describe desirable cognitive skills at different levels. A further refinement concerned discrimination between indications of cognitive aspects at the level of the individual partner and cognitive activity that involved both partners. For example evaluation of the solution that constitutes one dimension was split into evaluation of solution by one student or joint evaluation of a solution by both partners. As the process went on, it was realised that the dimensions that described cognitive aspects on the level of the individual were not much useful and efficient for the study and only the collaborative versions were finally used. However, the use of both individual and collaborative dimensions when rating the dyads, helped the evaluators to better conceptualise and emphasize on the collaborative aspects of each dimension.

Table 1: The rating scheme used for assessing collaboration quality

Aspect of collaboration	Dimension
Communicational	Sustaining Mutual Understanding
	Coordinating Communication
Cognitive	Knowledge
	Comprehension
	Application
	Evaluation
Technical	Technical coordination
Motivational	Shared task alignment
	Sustaining commitment

ISBN No: 978-1-86220-206-1

Moreover, dimensions related to "Time Management" and "Task Division" that were included in the original scheme were not used because it was decided that they were not so relevant or useful in the specific setting. The participants of the activities did not face any strict time constraints, so the role of proper time management or even intelligent division of the task among partners was not very important in this case. The resulted rating scheme is summarised in Table 1.

Evaluators reached also consensus on the detailed description of each dimension and some empirical anchoring to common examples that helped the assignment of ratings. In this part of the process, a crucial part of the work was to find proper description for dimensions of the original scheme that were adopted but needed to fit in the new circumstances. Especially, dimensions that refer to communicational aspects needed mostly to be redefined in order to serve the new media used.

For example, the core of "Coordinating communication", in the original rating scheme, was the issue of proper turn-taking in dialogue. Transmission delays were common when students communicated over the videoconferencing system, so requiring more explicit turn-taking than in normal face-to-face communication was crucial for the success of the activity (O'Conaill and Whittaker, 1997). However, if chat tool of Synergo is considered as the equivalent, there is a strong difference in the affordances of the chat tool as the communication medium that conducts dialogue and affects students' behaviour and thus performance in collaborative activity. Production costs of utterances in this medium are higher than in spoken communication, while, on the other hand, messages are consistent and reviewable, i.e. they can be inspected and referred back to throughout the collaboration (Clark and Brennan, 1991). However, a problem similar to the turn-taking requirement can occur when producing chat messages "out of sequence" due to missing sequentiality of chat (Clark and Brennan, 1991). It should be expected that some "disorderliness" in the chat messages is acceptable, and should not constitute a negative indication of coordination of communication. On the other hand, some indications included in the original version of the dimension, such as the occurrence of unanswered questions or incoherent replies (which show that coordination of communication is problematic), are also useful in this case. However, in this dataset, coordination of communication is not limited to dialogue through chat. Actions in the shared whiteboard could also be taken as a form of indirect communication, referred as "feedthrough" in Computer Supported Collaborative Work literature (Dix, A., J. Finlay, G. Abowd and R. Beale, 2004). In fact, successful activities are characterised by practices like switching naturally between these two means of communication, for example by referring to actions in the workspace in their chat messages, by carrying out in the workspace suggestions that their partner had made in the chat, or by explaining verbally a piece of the algorithm they had just constructed in the whiteboard.. Having in mind these considerations and empirically inspecting the data set, researchers in this study agreed on a reformed definition of "Coordinating communication" that has broader applicability, than the initial version mainly based on the turn-taking concept. A similar process was followed for the definition of other dimensions in the scheme with the two last dimensions that cover motivational aspects demanding the least extent of modifications.

A scale between 0 and 2 was used in order to characterise collaboration quality in each dimension. This scale was chosen because it was judged as sufficient for the purposes of this study and because a prior pilot use of a scale of 5 levels proved to be too detailed for the evaluators to use for many dimensions.

Application of the rating scheme and reliability

The rating scheme was applied in a dataset containing a number of log files that were initially fragmented according to the duration of the collaboration activity. The unit of analysis was a segment of activity of approximately 20 minutes duration. The last segments of each activity were often longer or shorter in order to conclude the activity without diverging significantly from the 20 minute window. In total, 15 dyads constitute the data set which produced 133 collaboration segments.

In order to ensure methodological reliability of the measures produced when applying the rating scheme, emphasis had to be given on the quality of the rating procedure. A common practice is that the rating is done by multiple evluators and the estimation of agreement for each dimension is a measure of reliability of the process. In this study, a part of the set of collaboration examples was evaluated by two annotators and reliability was measured using two well-established metrics such as Holsti's coefficient (Holsti, 1969)

and Cohen's Kappa (Cohen, 1960). Reliability results per dimension are shown in Figure 2. The general picture was quite satisfactory with the total reliability reaching the value of Cohen's Kappa = 0.92. Poor results were limited just to the dimension of "Knowledge Application" which appears with a low reliability score in Figure 2. This can be attributed to the complicated means of applying knowledge collaboratively on the specific algorithm building task, which makes the effort of rating according to this dimension more difficult. It is also not a trivial task to discriminate between instances of one student applying knowledge individually from instances of applying knowledge collaboratively. Further training of the annotators and refinement of the definition and description of this category should be done in a future study.

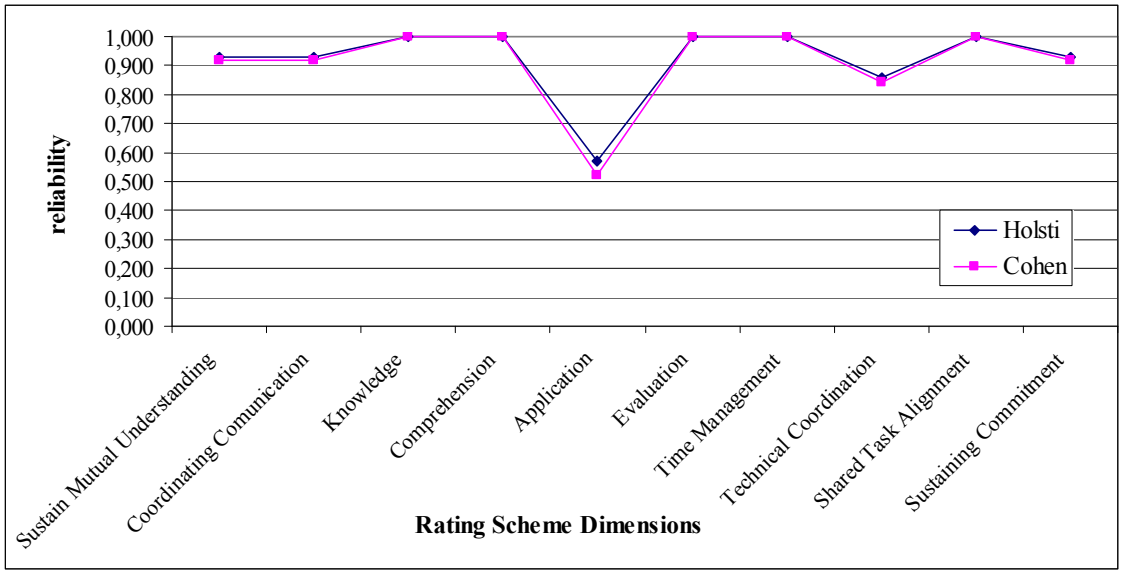

Figure 2: Reliability measures using Holsti's coefficient and Cohen's kappa metrics per dimension of the rating scheme.

Application of Machine Learning Techniques

The data that were reliably annotated as described in previous sections were then used as a point of reference for evaluating the performance of automatic evaluation techniques based on machine learning algorithms. The goal of use of such techniques in this case was to train a model that would be able to classify quite accurately a rich set of examples of episodes of collaborative situations that had been evaluated according to the rating scheme (Dunham, 2003). This model could be then used in the future for classifying other episodes of collaboration that do not need to be rated manually.

Quantitative measures of characteristics of the collaborative activity as they are automatically logged and calculated by the Synergo environment were used as attributes for the machine learning algorithms. These attributes are aggregations and other metrics that are based on sequential event logs produced and maintained by the tool. Many attributes relate to dialogue, such as the total number of exchanged dialogue messages, the average number of words per dialogue message etc., others are related to actions in the shared workspace such as the number of objects inserted, number of alterations of actors in workspace activity etc. Some of the metrics are more sophisticated since they describe the "symmetry" of collaborative activity, i.e. the extent to which the actions logged are equally attributed to each participant. However, in order to make the best use of the techniques used, attributes had to be transformed. In the first phase, the optimal set of attributes was examined with the use of F1 measure (Dunham, 2003). As the number of attributes increases, the performance reaches a stable maximum value and some attributes may be redundant. Further increase in the number of attributes deteriorates performance, which is an indication of a problem well known in machine learning studies as "the curse of dimensionality" (Hand, Mannila and Smyth, 2001). It was estimated that with a selection of 23 attributes the F1 measure reached the maximum value.

Table 2: Classifiers used for training

Category	Classifier
Bayes	*BayesNet*
	NaiveBayesUpdateable
	NaiveBayes
Functions	Logistic
	RBFNetwork
	SimpleLogistic
	SMO
Lazy	Kstar
	LWL
Meta (Metalearning)	*Bagging*
	LogitBoost
Trees	RandomForest
	J48
Rules	*Nnge*
	ZeroR

The final dataset consisted of 133 instances and 23 attributes including those attributes that refer to the dimensions of the rating scheme. Using the open source data environment WEKA (Witten and Frank, 2000) classification algorithms were applied on the dataset in order to predict the quality measures in each dimension in the scale from 0 to 2. In every classification (for each dimension of the scheme), 30% of the dataset was used in order to train the classifier and the rest 70% was used for testing. Table 2 illustrates the classifiers applied, grouped by the category of algorithms that they belong to. The representatives of each "family" of classifiers that gave the best results are shown in italics.
Figure 3 depicts the predictive performance of a set of the five optimal classifiers (Decision tree J48, BayesNet, Simple Logistic Function, Locally Weighted Learning, Nearest Neighbour NNge and a Meta classifier using the Bagging technique).

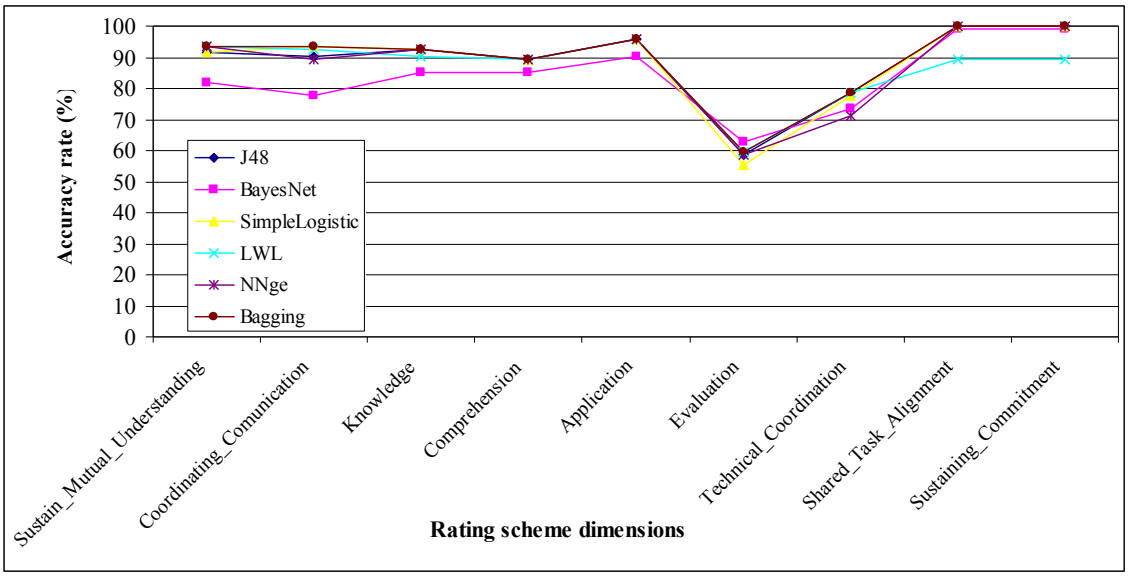

Figure 3: Performance of different classifiers per dimension of the rating scheme.

Results of the study, as shown in Figure 3 are positive for most algorithms. In fact the used classification algorithms performance was comparative in all dimensions, so there is no decisive advantage in using any particular algorithm. In terms of prediction of the dimensions of quality of collaboration, all but two out

of the nine dimensions were predicted with accuracy higher than 80% with the best classifiers scoring better than 90%. Positive results relate to the first two dimensions that refer to communicational aspects of collaboration (Sustaining mutual understanding and Coordinating Communication), to three out of four cognitive task-related dimensions of the scheme (only the Evaluation dimension was predicted with accuracy of around 60%). Finally the last two "motivational" dimensions (Sustaining Commitment and Shared Task Alignment) were predicted with high accuracy of over 90%.

Findings and interpretations

Findings of this study are encouraging although they vary along different dimensions. Prediction of communication dimensions using Machine Learning algorithms was achieved in a satisfactory way. This is expected, since good communication should be correlated with quantitative indices such as the number of messages exchanged or the symmetry of messages between participants.

However, significant diversity was observed in the prediction rates of dimensions of the cognitive domain. Whereas the most elementary cognitive skills, as recognized in the activity, were predicted with satisfactory rates, the accuracy rates of higher order skills were rather poor. The quality of collaboration, as judged according to the *Collaborative evaluation of the solution* dimension could not be predicted reliably. This should be expected since such advanced cognitive skills would not be easily predicted based on quantitative measures such as the total number of messages or the symmetry of shared workspace action. The complex practices of evaluating knowledge collaboratively in such a task are not easily reflected on event counts on which the attributes are based.

What should be also noted concerning the cognitive dimensions is that prediction rates of "Knowledge Application" were high although this dimension was problematic in terms of reliability of manual rating. The reason for that is that the ratings that were used for training the classification algorithms were the ratings of a single evaluator. So, these ratings reflected a consistent conception of the dimension although this conception might be different form the other evaluator. Thus, we are encouraged to believe that quality of this dimension is highly predictable even though further training among evaluators on the application of the rating scheme should be made.

There was also a relative failure in the predictions of the "Technical Coordination" dimension with the best score reaching almost 80%. Difficulties in mastering technical aspects of handling the tools provided are not so easily automatically detectable. However, a shortcoming is that the instances that were rated with a low score in this dimension constituted a very small portion of the whole data set, so not enough data were available for training the classification algorithms .

Finally, dimensions related to orientation of the students toward collaboration and dedication to the task, were predicted with the most success.

Conclusions

This study provided some encouraging results in the attempt to investigate applicability of machine learning techniques for automatic evaluation of collaboration quality. The main finding of the study has been that prediction of many aspects of quality of collaboration seems feasible, based on quantitative measures provided by collaboration tools such as Synergo. Communicational aspects, technical fluency and motivational aspects of collaboration seem to be predictable by techniques like the ones used in our study. On the other hand, when it comes to cognitive task-related aspects of collaboration, the accuracy of prediction deteriorated. Techniques score better with elementary cognitive skills, while advanced cognitive skills, like collaborative knowledge evaluation, seem harder to predict from these data.

In general, further investigation of the method is needed in order to confirm the obtained results. Variation in terms of different kinds of NSCL activities, i.e. involving the use of other synchronous tools or dealing with tasks in other problem domains is necessary if we want to argue for generality of the method. Furthermore, larger scale studies that imply larger training sets and more extended study of the rating scheme, with more established dimensions and systematic training of evaluators would also add

value to the findings reported here. To this direction recent efforts for adapting the original rating scheme that has been used as a starting point in our research, for network learning environments, (e.g. Voyiatzaki et al. 2008) need to be taken in consideration. In this context, dimensions such as "Collaborative Knowledge Evaluation" should be better described and specified in order to explore the possibility of gaining better prediction results, since the assessment of episodes that indicate higher cognitive skills are also very important in NSCL activity evaluation,

Finally, if the method matures enough for different NSCL settings, the next desired step would be to design and develop mechanism for automatic real-time assessment of collaborative episodes and provide estimations of collaboration quality along several dimensions, as feedback to students engaged in collaborative activities or teachers supervising them.

References

Avouris N., Dimitracopoulou, A.,& Komis V. (2003). On analysis of collaborative problem solving: An object-oriented approach, *Computers in Human Behaviour*, 19(2), 147-167.

Avouris, N., Komis, V., Fiotakis, G., Margaritis, M., & Tselios, N. (2003). Tools for Interaction and Collaboration Analysis of learning activities, Proc. CBLIS 2003, Nicosia, Cyprus.

Avouris N., Margaritis M.& Komis V. (2004). Modelling Interaction during small-group synchronous problem solving activities: the Synergo approach 2nd International Workshop on Designing Computational Models of Collaborative Learning Interaction, ITS 2004, Brazil 2004

Bloom, B.S., Englehart, M.D., Furst, E.J., Hill,W.H., & Krathwohl, D.R. (1956). *Taxonomy of Educational Objectives: The Classification of Educational Goals. Handbook 1: Cognitive domain*. New York: McKay.

Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. In L. B. Resnick, J. M. Levine, & S. D. Teasley (Eds.), *Perspectives on socially shared cognition* (pp. 127-148). Washington, DC: American Psychological Association.

Cohen. J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37-46.

Dimitracopoulou, A., & Komis, V. (2005). Design principles for the support of modelling and collaboration in a technology based learning environment, *Int. J. Continuing Engineering Education and Lifelong Learning*, 15 (1/2), 30-55.

Dix, A., Finlay, J., Abowd, G., & Beale, R. (1993). *Human-Computer Interaction*. Prentice Hall.

Dunham, M. (2003), *Data Mining: Introductory And Advanced Topics,* Prentice Hall.

Hand, D., Mannila, H., & Smyth, P. (2001). *Principles of Data Mining.* Cambridge, MA: MIT Press.

Holsti, O. (1979). *Content Analysis for the Social Sciences and Humanities*. Don Mills: Addison-Wesley.

Margaritis, M., Avouris, N., & Kahrimanis, G. (2006). On Supporting Users' Reflection during Small Groups Synchronous Collaboration. In Y. Dimitriadis, I. Zilgurs, & E. Gómez-Sánchez (Eds.), *Lecture Notes in Computer Science* Vol. 4154/2006, 140-154, Springer-Verlag.

Meier, A., Spada, H., & Rummel, N. (2007). A rating scheme for assessing the quality of computer-supported collaboration processes. *International Journal of Computer-Supported Collaborative Learning*, 2 (1), 63-86.

O'Conaill, B., & Whittaker, S. (1997). Characterizing, predicting, and measuring video-mediated communication: A conversational approach. In K. E. Finn, A. J. Sellen, & S. B. Wilbur (Eds.), Video-mediated communication (pp. 107-132). Mahwah, NJ: Lawrence Erlbaum Associates.

Voyiatzaki E., Meier A., Kahrimanis, G.,, Rummel, N., Spada, H., Avouris, N., (2008). Rating the quality of collaboration during networked problem solving activities, Proc. *6th International Conference on Networked Learning,* 5-6 May 2008, Halkidiki, Greece.

Witten, I. H., & Frank, E. (2000). *Data Mining: Practical Machine-Learning Tools*, Academic Press, San Diego, CA.

Xenos, M., Avouris, N., Komis, V., Stavrinoudis D, & Margaritis, M. (2004). Synchronous Collaboration in Distance Education: A Case Study on a CS Course, in Proc. IEEE ICALT 2004, Joensuu, FI.