RESI Revista Eletrônica de Sistemas de Informação, nº1, 2006

SELECTING PEDAGOGICAL PROTOCOLS USING SOM

Fernando Salgueiro, Guido Costa, Fernando Lage, Zulma Cataldi and Ramón García-Martínez
Intelligent Systems Laboratory. School of Engineering. University of Buenos Aires
Educational Informatics Laboratory. School of Engineering. University of Buenos Aires.
Software & Knowledge Engineering Center. Graduate School. Buenos Aires Institute of Technology
liema@fi.uba.ar

Abstract.

During the first semesters of Computer Engineering the amount of human tutors is insufficient: the students/tutors ratio is very high and there is a great difference in the acquired knowledge and backgrounds of the students. The main idea of this paper is to describe a system that could emulate the human tutor and provide to the student with a degree of flexibility for the selection of the most adequate tutorial type. This could be a feasible solution to the stated problem. But a tutorial system should not only emulate the human tutor but besides it should be designed from an epistemological conception of what teaching Basic Programming means specially in an Engineering course due to the profile and identity of the future engineer. The stated solution implement a series of artificial neural networks to determine if there is a relationship between the given initial population of students learning predilections and the different tutoring types. A series of experiences were carried out to validate the current model.
Keywords. Tutor Module, Intelligent Tutoring Systems, SOM.

1. Introduction

Its the main objective of the tutor module of an Intelligent Tutoring System to present the new knowledge to the student in the best way possible. To achieve that our research group [10,1] have designed a series of sub modules and interfaces to avoid the normal overlap in all of the modules of an Intelligent Tutoring System. In the tutor module, the main sub module is the pedagogical protocols, with its two basic components: the profile analyzer and the database of pedagogical protocols available in the system. The system have a database of pedagogical protocols where its use will be subordinated to the availability of the contents in the knowledge module, but the lesson always can be generated for some of the available protocols. In order to collect data about the way in which each student learns, the lists of learning style will be used as the tools for data recollection.

It has been determined the validity and trustworthiness of this instrument through his application by diverse investigators from the date of his creation [2,3] till now. Starting off of the data that provide each student they learning style will be determined and in a second step the learning style will be link to the pedagogical protocol. The Felder list [2] is as well a validated tool, that allows obtaining solid data to give of sustenance to one more an integral methodology to grow from the application of an intelligent tutoring system in a single career to all the university careers. After giving the questionnaire to the students, we will try to get those data records on different sets using the tools that the artificial intelligence provides (AI), such as Neural Networks (NN) in order to obtain the relation of the preferences of the students with the pedagogical protocols. From a statistically significant sample of students of which the lists of complete learning styles had been taken, will try to see if the learning styles can be group according to the education techniques or pedagogical protocols. This will allow correlating the preference of the student with the most suitable pedagogical protocol in the system. As the selection of the pedagogical protocol is one of the elements to determine, is desired to group the students in families with common characteristics.

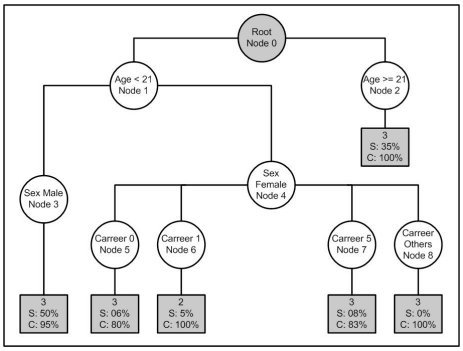

Figure 1. Tree generated by C4.5 algorithm using SOM output as C4.5 Input.

This can be achieved using the Self Organizing Maps (SOM) neural networks (also known with the name Kohonen [7] maps) that make a *"determined clusterization"* or group according to common characteristic of the original set of individuals. Once obtained the resulting groups of SOM network an induction algorithm will be used to find the rules that characterize each one of these groups. In this case the algorithms to be used will belong to the family of Top-Down Induction Trees (TDIT) algorithms. Although several algorithms exist that make these functions, one very complete is Quinlan´s C4.5 [9], an extension of algorithm ID3 (Induction Decision Trees) also proposed by Quinlan [8]. Its objective is to generate a decision tree and the inference rules that characterize this tree. In this particular case, the C4.5 will take as input the data of the students already clusterized by SOM and the output will be the rules describing each cluster.

Once obtained the smaller amount possible of rules by pruning to avoid overfitting, we move to another stage of the analysis in which, by means of an inference process, we found the relation between the SOM clusters and the pedagogical protocols available. In order to carry out the inference additional data of the performance of students in the courses in study with different protocols from education will be used. In Figure 2 the scheme of the solution can be seen: it represents the process of selection in global form, where we start from a student population of which we have they preferences with respect to the learning styles through the lists of Felder, we form groups of students using SOM A table is generated using the previously classified students, using all the attributes that they describe in the and the cluster predicted by SOM Later the C4.5 algorithm is used to generate the rules that best describe each clusters, relating a particular cluster not only with all the his attributes, like in the table of classified students, but with a set of rules.

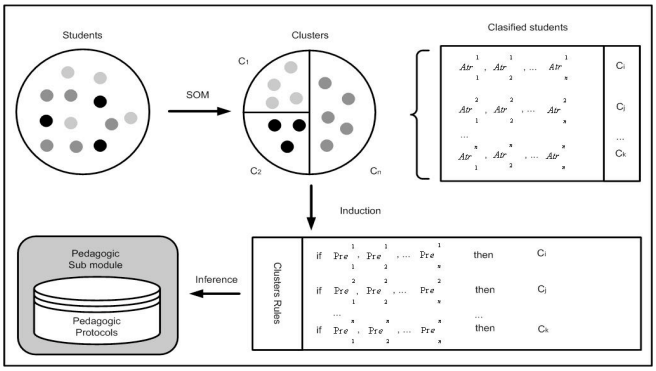

Figure 2. Basic scheme of the solution.

2. Inference of the pedagogic protocol

In this stage we try to relate the groups generated by SOM to the pedagogical protocols by training a Backpropagation type neural network. In order to find the relation between the learning style and the pedagogical protocol that corresponds to each group they took the basic protocols that describes to Perkins [12] in Theory One: [a] The *didactic or skillful instruction*: It satisfies a necessity that arises within the framework from the instruction: the one to expand the repertoire of knowledge of the pupils, [b] *The training*: It satisfies the necessity to assure an effective practice, [c] *Socratic education*: The educational aid to the student to include/understand certain concepts by itself and to give the opportunity him to investigate and to learn how to do it. Therefore the investigation is oriented in the search of the relation between the predilection of the students learning style and the pedagogical protocols used by the human tutors (professors). For it, as orientation parameter the grades of the partial evaluations are used to establish this relation. Two courses (A and B) will be taken pertaining to the Basic area of Programming. The only fundamental change between both was centered in the form of education, that is to say, in the pedagogical protocol used to dictate the classes. From this frame of reference, two courses are evaluated according to the control variables raised by García [5]. The variables raised for the reference courses are the following ones: [a] Similar contents of the courses, [b] Similar schedules, [c] Similar bibliography used for references, [c] Random entrance of the students, without preference defined to some course, [d] Similar previous formation of the assistants and Heads of practical works, [e] Similar didactic tools an [f] Way in the dictation of the class, where each one of the tutors presents the classes based on the pedagogical protocol that turns out more natural to carry out to him between the possible options that they are defined in Theory One and that they are analyzed in this investigation,

independently of the necessities or preferences of the individualized students. From the analysis made by García, it is observed for this study which the only one that it changes is the denominated "way of class dictation", that is to say, the pedagogical protocol for each course. In order to carry out the inference, the following hypothesis will consider: It is possible to relate the learning styles to the pedagogical protocols. This two more particular hypotheses are come off: (1) The composition of styles of learning (necessities and preferences of the students) of each student determines the style of education (or pedagogical protocol) more adapted an (2) Those students in whom the education style does not agree with its preference, present/display difficulties in the approval of the teached subjects. From the second hypothesis it is given off that for the approved students, the majority protocol preferred by the students will have to be the one that agrees with the used one in class by the tutor, whereas for the reprobated ones, the majority protocol must be inverted. In order to validate this affirmation a network of the Backpropagation type trained with the following characteristics: (1) were selected the approved ones of the course with professor who dictates in Socratic style and the most of the reprobated ones of the course with professor who dictates in skillful way and the network is trained considering the output like Socratic protocol. (2) are selected the approved ones of the course with professor who dictates in skillful style plus the reprobated ones of the course with professor who dictates in Socratic way and the network is trained considering the output exit like skillful protocol. In order to suppress the "data noise" the training is make this way due to which they contribute the groups that are outside the analysis (those that approved with any protocol or indifferents and that reprobated by lack of study or other reasons) and hope that the error of the tool is minor than the percentage of elements that are outside the analysis. Therefore, each cluster generated it will be analyzed:

- approved students { Correct protocol ® majority class; Indifferent ® minority class
- reprobated students { crossed protocol ® majority class; Lack of study ® minority class

Now we look to relate the forms of education and the learning styles, being taken as it bases for the analysis the reprobated students. In Figure 3 is observed, taking as bases an example on where single two pedagogical protocols exist and two preferences in the set of students, who the students whose preference agrees with the form or style of education do not have problems to approve. Two subgroups (in red) of students exist whose preference does not agree with the education form and are those that they reprobate, since they are "bad located".

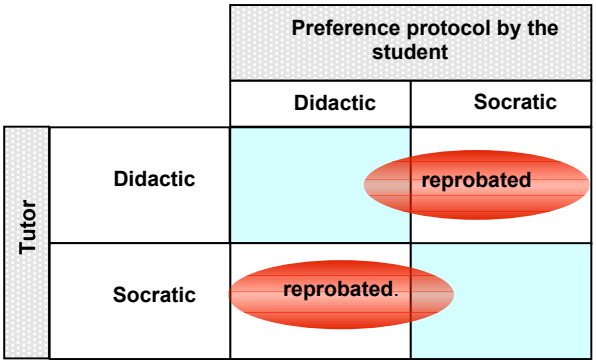

Figure 3. Inference general scheme

Following the hypothesis: The reprobated students who do not belong to the majority cluster predicted by SOM must have a different preference of a pedagogical protocol (inverted in this case) from the one of the professor whereupon they attended the matter. On the other hand, in Figure 3 is the idea of the hypothesis, where the reprobated students, who do not belong when majority cluster, must have a preference of pedagogical protocol different from the received pedagogical protocol in the classes. Therefore, if the data provided by the educational ones is analyzed with both courses with respect to the categorizations made by system SOM, it is possible to be obtained which is the percentage of students who would be badly located in the courses, and that are demonstrated through the results reprobated obtained in the evaluations. So that the obtained result is satisfactory, the Backpropagation neural network must have a classification error smaller than the percentage of elements that were left outside of the analysis, this way this tool will be useful for the classification of the preferences of education (pedagogical protocol) of the students from its styles of learning. This way, this submodule gives a ranking of best suitable pedagogical protocol, in descendent order with respect to its preference for the selected student. Soon, the only thing that is required is to cross all the pedagogical protocols including in the system. The basic scheme of the solution can be seen in Figure 4, where the Backpropagation network provides a ranking of aptitude of the pedagogical protocols available in the system, whereas the general scheme of the original solution, where to trainer a SOM network or use the selecting decision tree to provide only one basic exit of pedagogical protocol

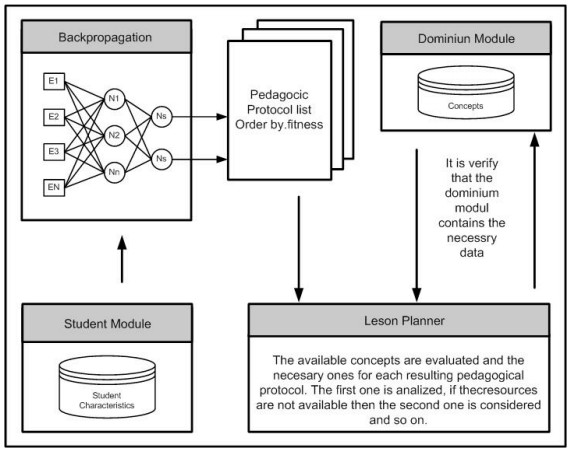

Figure 4. Modified solution structure

3. Experimental results

The experimental results validate the propose solution and will be established the steps to be able to repeat the experiences with another group of students or to apply the method in other different subjects from Basic Programming. The fundamental steps for the experimental design are described in Table 1 where it is begun with the taking of data of the students (to lists of learning styles) and it is used them like entrance for the training of neuronal a network artificial SOM to generate different groups. Soon the rules look for that describes these groups by means of the C4.5 algorithm.

Step	Input	Action	Output
1	Data recollection from students	Use Felder tool on students	Result of the Felder tool.
2	Felder tool result	SOM Training	Students Clusters
3	Cluster + Felder tool results.	Use C4.5 algorithm	Rules describing each generated cluster and the correspondin g decision tree.
4	Academic performance	Academic data grid	Academic grid
5	Result of the Felder tool. + Academic grid + Clusters	Analysis of the cluster and determinatio n of reprobated students.	Reprobated Student List for each cluster.
6	Result of the Felder tool.	Backpropag ation training	Determinatio n of the training error and the data out of analysis. Find the relation between learning style and pedagogic protocol.

Table 1. Steps for the experimental design

4. Validation of the population size

In order to determine the minimum number of elements in the sample we use Hernández Sampieri [6] for the calculation. An initial of 800 student's population has S2 variance of the sample of *n* student that can determine in terms of the probability p where:

$V=0.03$ (5.1)

$V^2= (0.03)^2=0.0009$ (5.2)

The number of samples without any adjustments will be:

$n= (S^2 /\_^2)= 0.09/0.0009= 100$ *students.* (5.3)

Adjusting in order of the real *N* population:

$n = (n´/(1+n´/N))=100/(1+100/800) = 89$ Students. (5.4)

The generalization error is below 3%, with which it is possible to say that the sample size is representative for all the students of the courses. Now we are ready to train the SOM network using the data recollected from the Felder tool. Most of the parameters of network SOM they arise through an iterative process, where the network trains and the results are analyzed. If the results are satisfactory (that is to say, the training error is the sufficiently small), the parameters are modified slightly to try to improve them still more. If the results little satisfactory they are compared with previous set and they are modified in a higher value.

Parameter	Value
Observations	121
Variables	47
Artificial Neurons[1]	10
Cicles	1000
Aleatority	Yes
Learning Parameter	
Initial	0.9
Final	0.1
Decay Function	Exp
Gaussian boundary parameter	
Initial	99,0%
Final	01,0%
Decay Function	Exp

Table 2. Parameters used for SOM with which the data of the students were classified.

It is possible to indicate that for the obtaining of the final values of training of the neuronal network they have been proven more than one hundred combinations, obtaining the best results with the list of parameters that is observed in Table 2.

The amount of clusters**:** If the amount of clusters is very elevated, it can be that it does not exists a correlation between so many pedagogical protocols and clusters, since part of the hypothesis that 3 pedagogical protocols exist (the proposed by Theory One). The number of clusters that looks for to obtain will be annotated between two and three. Summary of the results that the training of the SOM networks give is in Table 3, where the elements of each cluster generated are

totalized and the respective percentage are indicated.

	Cluster 1	Cluster 2
Data with all the attributes	**6** (5.00%)	**114** (95.00%)

Table 3. Summary of resulting elements after applying SOM to the input data.

The result is within the awaited amount of clusters and therefore the experimental data, they agree in the amount of clusters generated. As all the data are categorical, the generated rules will be equalities and it will not be found any range for them (for example: the continuous data). In order to find the attributes with greater gain of information, it is required to use the first N passages of the C4.5 Algorithm. In this case, the first nine were taken and the rules appear in Table 4. Oates [11] among others has analyzed several algorithms of "pruning" to trim the size of the rules generated from a great number of observations. Oates has found that as the size of the initial observations is increased, the size of the rules increases in linear form. This increase in the amount of rules antecedents does not significantly increase the precision in the classification of the rules. Continuing this way we get a result, for this case in individual, as proposed in the works of Quinlan [8,9] and Oates [11] previously mentioned, offered the additional advantage when using the second tree (with less levels and minor amount of nodes), the Intelligent Tutorial System requires minor amount of information to select the pedagogical protocol of the student and with easier access information (it is simpler to know the answers of some key questions in the list that the answers to all the questionnaire).

Rule	Antecedent	Consequent
Rule 1	**If** "Normally they consider me: Extrovert"	**Then** Cluster 2
Rule 2	**If** "Normally they don't consider me Reserved neither Extroverted"	**Then** Cluster 1
Rule 3	**If** "I Remember easily: Something that I have thought much"	**Then** Cluster 2
Rule 4	**If** "I don't remember easily something than I have thought much or something that I did"	**Then** Cluster 1
Rule 5	**If** "I learn: To a normal rate, methodically. If I make an effort , it profit".	**Then** Cluster 2
Rule 6	**If** "I do not learn to a normal rate, not methodically neither disordered"	**Then** Cluster 1
Rule 7	**If** "When I think about which I did yesterday, most of the times I think about: Images"	**Then** Cluster 2
Rule 8	**If** "When I think about which I did yesterday, most of the times I think about: Words"	**Then** Cluster 2
Rule 9	**If** "When I think about which I did yesterday, most of the times I don't think about words neither images"	**Then** Cluster 1

Table 4. Resulting rules to cross the tree generated by the C4.5 Algorithm

Training this way it is managed to suppress the *"noise"* that contributes the groups that are outside the analysis. In Table 5 the results of the students discriminated by courses can be seen, counting total students, students reprobated classified as pertaining to the cluster in opposition to the one of the majority and the percentage that relates the reprobated and approved students that in addition they are bad classified.

Observed Characteristic	Course A	Course B
Total of Students (For this study)	47	53
Students who reprobated the partial evaluation and were in a course with different pedagogical protocol	30	0
Students who approved the partial evaluation were in a course with different pedagogical protocol (inverted)	10	33
Approved students (no mattering about the protocol)	7	20
Reprobated students respect to the approved ones, within the subgroup of badly classified	75%	0%

Table 5. Summary of percentage obtained for the analysis of students, discriminated by courses.

For this experience the network of the Backpropagation type trained and a ranking (scale) of pedagogical protocols and nonsingle the most adapted for a particular situation was obtained, in order to give flexibility to the module that stores the contents.

For the training of the network Backpropagation 67% of the data (qualifications) were used randomly whereas 33% of the remaining data were used to validate the generated model. After more than 100 training of 1000 cycles each one, where it looked for to diminish the error in the resulting network, reached the conclusion that the optimal values for the parameters of the network are those that they are see in Table 6.

Characteristic	Value
% Error (Training group)	3.75%
% Error (Validation group)	2.00%
Network characteristics	
Input neuron	**13**
First hidden layer neurons	**20**
Second hidden layer neurons	**20**
Output neurons	**2**

Table 6. Resultados de los datos de entrenamiento de la red tipo Backpropagation.

This training is valid since the error of the tool (3,75% for the set of training and 2,00 % for the validation set) is minor who the error of the elements that were outside the analysis, that represents the students who did not approve the matter not to study the sufficient thing, although the pedagogical protocol agreed with the preference of the student (who is 25%). Therefore it is possible to be concluded that: [a] course B is related to cluster 1: since the errors induced by elements of cluster 2 within the course a are in a 75% or in other words, the network classifies to 75% of the students reprobated in the course A

like pertaining when cluster 1 and course B is related to cluster 2: since another possible allocation in this case does not exist and in addition the percentage to error of classification and reprobation is of 0%. The obtained results agree with the affirmations of Perkins [12], where the network Backpropagation predicts that most of the reprobated students they must have received classes using another pedagogical protocol. Socratic protocol is related with Cluster 2 and Magistral protocol is related with Cluster 1. This way the same turn out of the inferential step is obtained in order to be able within the framework to incorporate the experimental results to the design of the tutorial module of Intelligent Tutoring System (ITS). One concludes, that controls a module of the tutor able to categorize to the students according to its characteristics, within some of the pedagogical protocols available in the system, for the case in study, controlled data of 2 pedagogical protocols (Magistral and Socratic) and in this case is possible to be categorized automatically to the students within each one of them, according to its preferences to improve the results of a pedagogical session.

5. Conclusions

When validating the model against the real data, as much for the data triangulation as the training of the neural networks that support the model, it was found that the data adapt very satisfactorily to the test conditions, becoming thus, not only a theoretical tool been worth to guide the students in the learning process, but also in an instrument In practice, that allows implantations of an Intelligent Tutorial System able to generate measurable and useful satisfactory results in real environments. It is fulfilled then the primary objective of this work which is to provide an additional tool for the human tutors, who can relegate some of their tasks that, either by lack of time or resources, cannot fulfill in a satisfactory way the student request, whereas it provides a secondary support for the students whom they try to complement their knowledge or to regulate its own rate of learning. Then, it is provided to the field of the Intelligent Tutorial Systems a new tool, to facilitate the selection of the suitable pedagogical protocol, resulting this in a gain, nonsingle for the performance of the STI itself, but in the student, who is the fundamental human component that she makes useful to the system and offers identity to them. Thus it is tried to make a contribution and to improve the academic performance of the different students and therefore its quality from life.

Acknowledgements

This study was supported by LIEMA/LSI-FI-UBA 2002-2005 Educational Informatics Laboratory and Intelligent Systems Laboratory for FI-UBA (Faculty of Engineering, University of Buenos Aires). The authors would thanks to the students who participated in the experience.

References

[1].COSTA, G.; SALGUEIRO, F., CATALDI, Z., GARCÍA-MARTÍNEZ, R. y LAGE, F. 2005. *Intelligent Systems for student modelling.* Proc. GCETE'2005, Global Congress on Engineering and Technology Education CD. march 13-15. Brazil.

[2].FELDER R. & SILVERMAN L. 1988. *Learning Styles and Teaching Styles in Engineering Education*. Engr. Education, Volumen 78, número 7, p. 674-681.

[3].FIGUEROA, N.; LAGE, F.; CATALDI, Z.; DENAZIS, J. 2003. *Evaluation of the experiences for improvement of the process of learning in initial subjet of the computer science careeer.* Proc. Int. Conf. on Eng. and Computer Educ. ICECE 2003, March16-19. San Paulo. Brazil.

[4].FIGUEROA, N.; CATALDI Z.; SALGUEIRO F. A.; COSTA G.; MÉNDEZ P.; LAGE F. J. 2004. *The styles of learning and the university desertion in Computer science Engineering.* CACIC 2004. Arg.

[5].GARCÍA, M. 1995. *El paradigma de Pensamiento del profesor*. Editorial CEAC, Barcelona.

[6].HERNANDEZ SAMPIERI, R. et al.. 2001. *Metodología de la investigación*. Mc Graw Hil. México.

[7].KOHONEN, T. 2001. *Self-Organizing Maps*, third edition. Springer.

[8].QUINLAN, J. 1987. *Simplifying Decision Trees. Simplifying Decision Trees*. International Journal of Man-Machine Studies 27(3): 221-234.

[9].QUINLAN, J. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann.

[10].SALGUEIRO, F., COSTA, G., CATALDI, Z., GARCÍA-MARTINEZ, R. Y LAGE, F. 2005. *Intelligent Systems for tutor modelling.* Proc. Global Congress on Eng. and Tech. Education. march 13-15

[11].OATES, T. 1997. *The Effects of Training Set Size on Decision Tree Complexity*. Proc. 14th International Conference on Machine Learning.

[12].PERKINS, D. 1995. S*mart schools*. The Free Press. A division of Simon & Schuster, Inc.