ESANN'2006 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 26-28 April 2006, d-side publi., ISBN 2-930307-06-4.

Learning what is important: feature selection and rule extraction in a virtual course

T.A. Etchells[1], A. Nebot[2], A. Vellido[2], P.J.G. Lisboa[1], and F. Múgica[3]

[1] Liverpool John Moores University (LJMU). Neural Computation Group.
Byrom St, L3 3AF, Liverpool, U.K
[2] Universitat Politècnica de Catalunya (UPC). Soft Computing Group.
C. Jordi Girona, 1-3, 08034, Barcelona, Spain
[3] Instituto Latinoamericano De la Comunicación Educativa (ILCE)
Calle del Puente 45, México D.F., México

Abstract. Virtual campus environments are becoming a mainstream alternative to traditional distance higher education. The Internet medium they use allows the gathering of information on students' usage behaviour. The knowledge extracted from this information can be fed back to the e-learning environment to ease teachers' workload. In this context, two problems are addressed in the current study: finding which usage features are best at predicting online students' marks, and explaining mark prediction in the form of simple and interpretable rules. To that effect, two methods are used: Fuzzy Inductive Reasoning (FIR) for feature selection and Orthogonal Search-Based Rule Extraction (OSRE). Experiments carried out on the available data indicate that students' marks can be accurately predicted and that a small subset of variables explains the accuracy of such prediction, which can be described through a set of actionable rules.

1 Introduction

The Internet is shaping the next generation of distance education tools. Distance education arises from traditional education to cover the necessities of remote students. Virtual campus environments are fastly becoming a mainstream alternative to traditional distance higher education. The Internet medium they use is a two-way channel: it conveys content, but also allows gathering information on their students' usage behaviour. The knowledge extracted from this information can be fed back to the e-learning environment in order to fit it to the students' needs and requirements, while easing the course advisers' workload. The use of data mining methods in the analysis of e-learning data is still in its infancy, although some strides are being made in this area of research [1]. In general, standards on the application of data mining methods and techniques in this area are yet to be set.

One of the most difficult and time consuming activities for teachers in distance education courses is that of evaluation, as it usually entails using collaborative resources such as e-mail, discussion forums, chats, etc. that may yield high dimensional data sets containing information on students' system usage. It would be helpful to reduce the data dimensionality by identifying and selecting features that are relevant to predict students' performance. This way, teachers could provide feedback to students regarding their learning activities online and in real time. In this paper, we use the FIR methodology [2] for feature selection on a real data set. The interpretability of the mark prediction results would be improved by their description

in terms of simple, actionable rules. This is accomplished in this study through the application of OSRE, a novel overlapping rule extraction method [3].

The rest of the paper is organized as follows: sections 2 and 3 present the FIR and OSRE methodologies, respectively. A description of the data from the analysed e-learning course is provided in section 4. Results from the experiments are presented and briefly discussed in section 5. The paper wraps up with a conclusion section.

2 Fuzzy Inductive Reasoning

The conceptualization of Fuzzy Inductive Reasoning (FIR) arises from General Systems Theory [4]. This modelling and qualitative simulation methodology is based on systems behaviour rather that on structural knowledge. FIR is able to obtain good qualitative relations between the system variables and it is a useful tool for feature selection. There exist many supervised variable selection techniques, but only a few are suitable for arbitrarily non-linear systems. FIR has shown promise in this context [5] and provides a robust alternative to more traditional methods. A study comparing the performance of FIR with other variable selection techniques would be appropriate, but is beyond the scope of this brief paper. FIR consists of four main processes, namely: fuzzification, qualitative model identification, fuzzy forecast and defuzzification. Only the first two of these processes (where feature selection is performed) concern us in this study.

The fuzzification process converts quantitative data stemming from the system into fuzzy data. The qualitative model identification process is responsible for finding relationships between the variables and therefore for obtaining the best model that represents the system. A FIR model is composed of a mask matrix (model structure) and a pattern rule base. Only the mask part of the model is significant for feature selection. The negative elements in this matrix are referred to as *m*-inputs (mask inputs) and denote the input arguments of the qualitative functional relationship. They symbolise causal relationships with the output. The system variable that has one or more *m*-inputs in its matrix column is considered a relevant one, needed for the prediction of system's output. In FIR, a mask candidate matrix is defined as the ensemble of all possible masks from which the best one (maximizing an entropy measure, or quality of the mask) is chosen by either a mechanism of exhaustive search of exponential complexity, or by one of various suboptimal search strategies of polynomial complexity. For further details on FIR, the reader is referred to [2].

3 Orthogonal Search-based Rule Extraction

Orthogonal Search-based Rule Extraction (OSRE: [3]) is an algorithm that efficiently extracts comprehensible rules from smooth models, such as those created by neural networks, that accurately classify data. OSRE is a principled approach and is underpinned by a theoretical framework of continuous valued logic developed in [6]. In essence, the algorithm extracts rules by taking each data item, which the model predicts to be in a particular class, and searching in the direction of each variable to find the limits of the space regions for which the model prediction is in that class (Fig. 1, left). These regions form hyper-boxes that capture in-class data and they are

converted to conjunctive rules in terms of the variables and their values (Fig. 1, right). The obtained set of rules is subjected to a number of refinement steps: removing repetitions; filtering rules of poor specificity and sensitivity; and removing rules that are subsets of other rules [7]. Specificity is defined as one minus the ratio of the number of out-of-class data records that the rule identifies to the total number of out-of-class data. Sensitivity is the ratio of the number of in-class data that the rule identifies to the total number of in-class data. The rules are then ranked in terms of their sensitivity values to form a hierarchy describing the in-class data. Testing against benchmark datasets [3] has showed OSRE to be an accurate and efficient rule extraction algorithm.

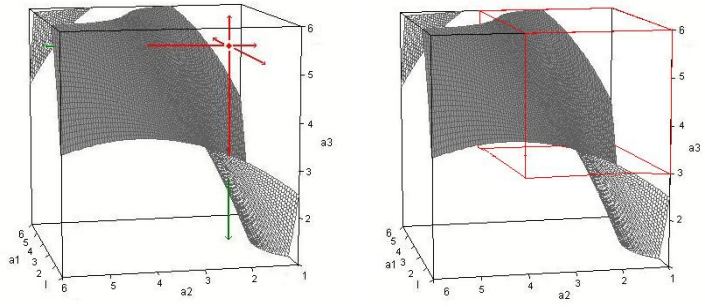

Figure 1: Left: Illustration of orthogonal searching to find decision boundaries; Right: Hyper-boxes constructed from the search results

4 Data from the CECTE virtual campus

The CECTE is a partially virtual campus, part of the international organism known as ILCE (Instituto Latino-Americano de la Comunicación Educativa). The teaching-learning process is semi-presential, as students follow courses online (WCECTE) but also attend weekly TV sessions. Through WCECTE, the students access the course materials and communicate with each other through an e-mail system and a discussion forum. The environment also includes other interactive tools.

For the experiments in this study, a set of 722 students, enrolled in the "Didactic Planning" graduate course, was selected. The course is addressed to high school teachers with the main purpose of learning new teaching methods and strategies. This is the reason why these activities are centred upon the so-called "class plan": a document where a set of strategies are suggested in order to develop a teaching-learning session. The data features available for this study are detailed in Table 1. For these experiments, the final mark (MARK) was categorised, following the advice of the course advisors, as follows: Class 1 (MARK<5: 6.9% of students); Class 2 (5≤MARK<8: 16.6%); Class 3 (8≤MARK<10: 76.5%).

5 Experimental results and discussion

Firstly, we were interested in determining which of the features from Table 1 had the highest relevance in predicting student performance (MARK). For the FIR qualitative

model identification process to take place, it was necessary to provide the mask candidate matrix, which is of depth one (only one row), forbidding the creation of temporal relations. The optimal mask computed by FIR was:

AGE	EXP	G	STD	POS	ACT	ASS	MAIL	COEV	F	FCP	FC	IC	ER	BR
0	0	0	0	0	0	0	0	-1	0	0	0	-2	-3	0

This optimal mask reveals that the average marks of the co-evaluation (COEV), the initial class plan (IC), and the experience report (ER), are the most relevant features to predict the final mark of the course (MARK) for each student. The selection of COEV, a variable scarcely used in e-learning environments, is especially interesting. It reveals that the capability of a student to evaluate the work of others is a good predictor of her/his own final mark. This conclusion was deemed reasonable by the 31 advisors responsible for the course. It is also interesting to note that the variables that describe the personal attributes of the students were not selected by FIR as relevant predictors of the final mark, reflecting that the characteristics of the students' work are far more predictive of the final mark than their personal features.

Feature	**Alias**
Age of the student (minimum of 22, maximum of 67, in this data set)	AGE
Area of expertise.	EXP
Gender.	G
Level of studies.	STD
Position of the student as a teacher.	POS
Percentage of the total of the activities performed by the student (from 0 to 100)	ACT
Percentage of assistance to course sessions (from 0 to 100)	ASS
Average mark of the student in the activities sent by e-mail (from 0 to 10)	MAIL
The average mark of the co-evaluation performed by the student of the class plan of other students (from 0 to 10)	COEV
Average mark of the student's forum participation (from 0 to 10)	F
Average mark of the forum class plan (from 0 to 10)	FCP
Average mark of the final class plan (from 0 to 10)	FC
Average mark of the initial class plan (0 or 10)	IC
Average mark obtained by the student in the experience report (from 0 to 10)	ER
Average mark of the activities performed in the branch (from 0 to 10)	BR
Final mark obtained by the student in the course (from 0 to 10)	MARK

Table 1. Data features collected for the study.

The approach to the OSRE experiments was twofold: Firstly, all data features from Table 1 were used in the classification task; secondly, only the three features selected by FIR were used. Two layered Multi-Layer Perceptrons (MLP) were trained using error back-propagation and weight decay to inhibit overtraining. The data were split into two sets of 361 records, for training and testing the MLPs. In each case, the network parameters were selected by cross-validation and set, for the models using all the variables, to: No. of hidden nodes = 8; learning rate = 0.01; momentum = 0.9; weight decay = 0.01; for the models using the FIR selection of 3 features, all parameters but weight decay = 0.05 were the same. In all cases, the network weights were initialised with random values. Once the number of training epochs that

minimised overtraining was determined, final networks were trained using all 722 data records. OSRE was used to produce a set of rules for each of the classes, shown in Tables 2, 3, and 4. Each rule is a conjunction of the features and their values.

CLASS 1 (all features)		**For this rule only**			**For disjunction of ALL rules up to row *n***		
n	**RULE**	***Spec***	***Sens***	***PPV***	***Spec***	***Sens***	***PPV***
1	(0≤ER≤6) ∧ (0≤COEV≤6)	0.99	0.82	0.85	0.99	0.82	0.85
2	0≤FC≤4	0.99	0.82	0.87	0.98	0.92	0.78
3	(0≤BR≤3) ∧ (0≤ER≤3)	1	0.1	1	0.98	0.96	0.78
CLASS 1 (FIR feat. selection)		**For this rule only**			**For disjunction of ALL rules up to row *n***		
n	**RULE**	***Spec***	***Sens***	***PPV***	***Spec***	***Sens***	***PPV***
1	0≤COEV≤5	0.96	0.94	0.66	0.96	0.94	0.66

Table 2: OSRE rules for Class 1 (MARK < 5). *Spec* stands for Specificity; *Sens* for Sensitivity; *PPV* is the Positive Predictive Value: the ratio of the number of in-class data that the rule predicts to the total number of data the rule predicts. Top table: Results for the full set of features. Bottom table: results for FIR feature selection.

CLASS 2 (all features)		**For this rule only**			**For disjunction of ALL rules up to row *n***		
n	**RULE**	***Spec***	***Sens***	***PPV***	***Spec***	***Sens***	***PPV***
1	(0≤FCP≤7) ∧ (IC=0)	0.96	0.35	0.64	0.96	0.35	0.64
2	(63.4≤ACT≤65.5) ∧ (69.7≤ASS≤100) ∧ (0≤FCP≤8) ∧ ER=0	0.99	0.82	0.87	0.98	0.92	0.78
CLASS 2 (FIR feat. selection)		**For this rule only**			**For disjunction of ALL rules up to row *n***		
n	**RULE**	***Spec***	***Sens***	***PPV***	***Spec***	***Sens***	***PPV***
1	IC=0	0.94	0.43	0.66	0.94	0.43	0.66
2	(4≤COEV≤10) ∧ (0≤ ER ≤2)	0.99	0.33	0.93	0.94	0.66	0.69

Table 3: OSRE rules for Class 2 (5≤MARK<8). *Spec, Sens* and *PPV* as in table 2. Top and bottom tables as in table 2.

Interestingly, in the experiments using all features, those selected as the most relevant by FIR: COEV, IC, and ER figure prominently in the main rules generated by OSRE, especially for classes 1 and 3 (the low and high marks). Therefore, the rule extraction results indirectly validate, at least partially, the FIR selection. Classes 1 and 3 are extremely well captured by their corresponding rules. The students that failed (MARK<5) are defined in very simple terms through low values of ER, COEV, FC and BR. The OSRE results using only the 3 features selected by FIR are quite consistent with those obtained using all features, while providing the most parsimonious rule descriptions of the MARK classes that can be obtained without compromising too much of the classification accuracy.

6 Conclusion

Some inroads have been made into the application of data mining techniques in e-learning environments. In this study, the FIR methodology has been applied to select those features of online usage of a virtual course which best predict students' final

marks. This prediction can be more easily interpreted through rule extraction. The novel OSRE methodology has been applied to obtain simple sets of rules describing the diverse levels of the students' performance. All this newly acquired knowledge can be fed back into the system to ease the workload of the course advisors.

CLASS 3 (all features)		**For this rule only**			**For disjunction of ALL rules up to row *n***		
n	**RULE**	***Spec***	***Sens***	***PPV***	***Spec***	***Sens***	***PPV***
1	(8≤BR≤10) ∧ (3≤F≤10) ∧ (1≤FCP≤10) ∧ IC=10 ∧ (7≤ER≤10) ∧ (8≤COEV≤10)	1	0.86	1	1	0.86	1
2	(8≤BR≤10) ∧ (1≤MAIL≤10) ∧ (9≤ER≤10) ∧ (9≤COEV≤10)	1	0.65	1	1	0.91	1
3	(7≤BR≤10) ∧ (7≤F≤10) ∧ (5≤MAIL≤10) ∧ IC=10 ∧ (7≤FC≤10) ∧ (7≤ER≤10) ∧ (5≤COEV≤10)	1	0.7	1	1	0.94	1
4	(6≤FCP≤10) ∧ (2≤MAIL≤10) ∧ FC=10 ∧ (9≤ER≤10) ∧ (9≤COEV≤10)	1	0.47	1	1	0.95	1
5	BR=10 ∧ IC=10 ∧ FC=10 ∧ (8≤ER≤10) ∧ (6≤COEV≤10)	1	0.49	1	1	0.97	1
CLASS 3 (FIR feat. selection)		**For this rule only**			**For disjunction of ALL rules up to row *n***		
n	**RULE**	***Spec***	***Sens***	***PPV***	***Spec***	***Sens***	***PPV***
1	(9≤ER≤10) ∧ (9≤COEV≤10)	0.91	0.73	0.96	0.91	0.73	0.96
2	IC=10 ∧ (4≤ER≤9) ∧ (7≤COEV≤10)	0.91	0.39	0.94	0.82	0.95	0.95
3	IC=10 ∧ (9≤ER≤10) ∧ (5≤COEV≤9)	0.93	0.35	0.94	0.82	0.99	0.95

Table 4: OSRE rules for Class 3 (8≤MARK<10). *Spec, Sens* and *PPV* as in table 2. Top and bottom tables as in table 2.

Acknowledgements

Alfredo Vellido is a research fellow within the Ramón y Cajal program of the Spanish Ministry of Science and Education.

References

[1] C. Romero and S. Ventura, eds. *Data Mining in E-Learning*. WIT Press, forthcoming, 2006.

[2] A. Nebot, Qualitative Modeling and Simulation of Biomedical Systems Using Fuzzy Inductive Reasoning, Ph.D. thesis, Universitat Politècnica de Catalunya, Barcelona, Spain, 1994.

[3] T.A. Etchells and P.J.G. Lisboa, Orthogonal Search-based Rule Extraction (OSRE) method for trained neural networks: A practical and efficient approach, *IEEE Transactions on Neural Networks*, 17(2), IEEE Neural Networks Society, 2006.

[4] G.J. Klir. *Architecture of Systems Problem Solving*, Springer, New York, 2003 (2nd ed.).

[5] J.M. Mirats i Tur, F.E. Cellier and R.M. Huber, Variable selection procedures and efficient suboptimal mask search algorithms in fuzzy inductive reasoning, *International Journal of General Systems*, 31(5):469-498, Taylor & Francis, 2002.

[6] H. Tsukimoto, Extracting rules from trained neural networks, *IEEE Transactions on Neural Networks*, 11(2):377-389, IEEE Neural Networks Society, 2000.

[7] T.A. Etchells, I.H. Jarman and P.J.G. Lisboa, Empirically derived rules for adjuvant chemotherapy in breast cancer treatment. In proceedings of the *Advances in Medical Signal and Information Processing Int.* Conf. (MEDSIP 2004), 5-8 September, pages 345- 351, Malta, IEE, 2004.