Mining Web-based Educational Systems to Predict Student Learning Achievements

José del Campo-Ávila, Ricardo Conejo, Francisco Triguero, Rafael Morales-Bueno

Universidad de Málaga, Andalucía Tech, Departamento de Lenguajes y Ciencias de la Computación, Campus de Teatinos, Málaga, España

***Abstract* — Educational Data Mining (EDM) is getting great importance as a new interdisciplinary research field related to some other areas. It is directly connected with Web-based Educational Systems (WBES) and Data Mining (DM, a fundamental part of Knowledge Discovery in Databases).**

The former defines the context: WBES store and manage huge amounts of data. Such data are increasingly growing and they contain hidden knowledge that could be very useful to the users (both teachers and students). It is desirable to identify such knowledge in the form of models, patterns or any other representation schema that allows a better exploitation of the system. The latter reveals itself as the tool to achieve such discovering. Data mining must afford very complex and different situations to reach quality solutions. Therefore, data mining is a research field where many advances are being done to accommodate and solve emerging problems. For this purpose, many techniques are usually considered.

In this paper we study how data mining can be used to induce student models from the data acquired by a specific Web-based tool for adaptive testing, called SIETTE. Concretely we have used top down induction decision trees algorithms to extract the patterns because these models, decision trees, are easily understandable. In addition, the conducted validation processes have assured high quality models.

***Keywords* — Data Mining, Decision Trees, Educational technology, Knowledge discovery.**

I. Introduction

SINCE Internet opened a new way to communicate in many different forms, the educational sector adopted such technology and developed the Web-based Educational Systems (WBES). Firstly, they were static systems, mainly dedicated to divulgate contents. But progressively, they extended their capabilities with new characteristics in order to make the systems adaptive and intelligent [1].

At this moment there exist many different systems that combine different elements to achieve some level of intelligence. Therefore, we can find WBES with adaptive techniques [2], some other WBES with intelligent mechanisms [3] and more complex systems that combine both properties (a detailed review of AIWBES was presented by Brusilovsky and Peylo [4]).

What it is evident is the high volume of data that these systems are storing and processing continuously: relations between contents offered to students, interactions with students, number of visits, marks achieved in tests, time used to respond those tests, etc.

Knowledge discovery in databases (KDD) continues extending to almost every field where large amount of data are stored and processed (databases, system logs, activity logs, etc.), so WBES becomes another environment to apply KDD processes.

The data mining techniques are essential for one of the most important points of KDD: they are applied in data analysis phase and machine learning algorithms are used to produce the models that summarize the knowledge discovered [5]. Therefore, it is easy to see that educational tasks can benefit from the knowledge extracted by data mining.

This research field is called Educational Data Mining (EDM) and its main objective is to analyze data stored in WBES in order to resolve educational research issues [6]: validation of the educational system, prediction of students learning achievements, identification of misconceptions [7], assessment and feedback to the authors of courses [8], etc.

In this paper we try to determine that data mining techniques can help to predict students learning achievements, mainly oriented to find relations between continual assessment (or evaluation) and the final grade achieved.

This paper is organized as follow. In Section 2 we describe the materials used and the conducted methodology. Basically, our materials are data collected by SIETTE[8], a Web-based tool for adaptive testing [9] and the framework for data mining called Weka [10]. Then, in Section 3, we present the results and comment the patterns discovered by machine learning algorithms. Finally, in Section 4, we summarize the most relevant conclusions and propose new research lines for futures works.

II. Materials and Methods

Considering the features offered by data mining in order to discover patterns in datasets, in this case extracted from Web-based Educational System, we propose to study the existence of different kinds of relations between the continuous

[8] http://www.siette.org

DOI: 10.9781/ijimai.2015.326

evaluation of students and their final achievements in the subject. For this purpose we work with the following materials and methodologies.

A. Materials

The raw materials of any process of knowledge discovery that uses data mining techniques are data, grouped in subsets called datasets. Every dataset is composed of examples described by attributes and labeled with a class (supervised learning). Values for these attributes can be numerical or nominal.

For this study we have focused in students that took the subject "Principles in Informatics" in two consecutive courses. The skills and competences to be achieved are varied: from basic concepts related with Computer Science (hardware, software, algorithms, etc.) to elementary abilities to develop computer programs using the C programming language.

The evaluation of this subject includes a continuous evaluation during the course (with a weight of 40% in the final grade) that ends with a final evaluation exam (60% weight). The continual assessment (or continuous evaluation) is compound of three tests (20%) and three practical exercises (20%). What we are using in this study are the marks achieved by the students in the tests that have been completed using the SIETTE Web-based Educational System [9]. First test (T1) is used to check how concepts related with Computer Science are assimilated, the second one (T2) focus on initial programming abilities with C (types, expressions, operators and control flow) and the third one (T3) check the knowledge about more advanced concepts in C (functions and structures). The final exam is mostly prepared to check the programming abilities; so basic concepts related with Computer Science are only evaluated with one test (T1).

In Table I we show some statistics related to the real marks achieved in the tests. The maximum value cannot be greater than 100.00, but minimum values can be lower than 0.00 because wrongly answered questions count negatively (if a student answers many questions incorrectly, the mark is lower than 0.00).

Taking this context in consideration, now we can describe the datasets that we have used. In our case the examples summarize the evaluation achieved by the students (116) that took the subject "Principles in Informatics". In a first approach we only consider the marks for every test, but in a second step we added the differences with respect to the average value, in order to establish a relative comparison between the results.

The class attribute is the final grade achieved in the global subject evaluation. We have used the numerical grade, defined in [0,10], and transformed it to the European ECTS grading scale (A for the best grades and F for the worst ones, F corresponds to students that fail) [11].

To carry out the mining process there exist different frameworks that implement multiple machine learning algorithms. We have used Weka [10] because it includes TDIDT (Top Down Induction Decision Trees) algorithms that represent the knowledge extracted in form of decision trees

TABLE I
MARKS ACHIEVED IN TESTS

	T1 (HW, SW, algorithms)	T2 (types, operators, control flow)	T3 (functions, structures)
Minimum	-20.00	-20.00	-18.33
Average	34.18 ± 19.30	35.50 ± 23.42	30.47 ± 27.84
Maximum	76.67	86.67	100.00

Minimum, average and maximum values observed in the tests answered by students. Maximum never can be greater than 100.00, but minimum values can be lower than 0.00 because wrongly answered questions count negatively

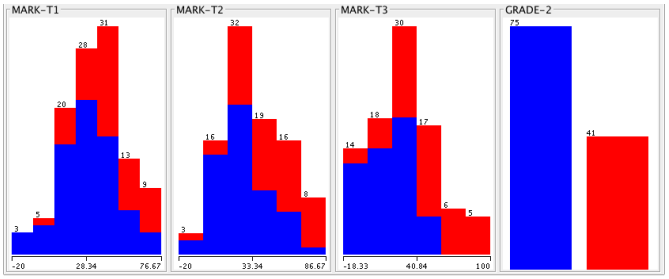

Fig. 1. Class distribution visualizing all marks (for tests T1, T2 and T3) in the first dataset (116 students). Blue color represents students that fail the evaluation (or absent themselves) and red color represents students that pass the evaluation. This chart is plotted by the Weka framework.

[12]: a model easily understandable by humans with some other additional advantages (learning with numerical or nominal data, robustness, verifiable reliability, etc.). Concretely we have selected the J48 algorithm (C4.5 [13] implementation coded in Weka), using it with its default configuration. When plotting the decision trees (Fig. 4, 5, and 6), the numbers present in the nodes (<first> / <second>) represent the number of examples that satisfy the branch (<first>) and the number of examples that, in addition, are incorrectly classified (<second> that it is not present when there is no errors).

B. Methods

Once we have described the datasets and the framework we have used, we can detail which methodology we have followed.

Firstly we have preprocessed the data in order to clean and prepare them. Data extracted from SIETTE are very rich and diverse, but nowadays, they cannot be directly exported to the kind of dataset supported by Weka (ARFF files). Some transformation steps were needed: discretization of numerical grade to ECTS grading scale, calculation of new calculated attributes, identification of missing values, etc.

The datasets used in this study have been progressively transformed to do more detailed mining process. Although the details will be presented in next section, we can advance that we have used 3 datasets derived from the original one.

The first dataset, with 116 examples (students), is described by 3 attributes (marks achieved in every test) and a binary nominal class (passing the subject or failing it – including absent students –). In Fig. 1 it is shown the class distribution for three different marks.

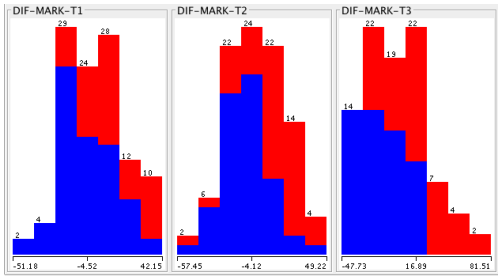

Fig. 2. Class distribution visualizing marks and differences with the average marks (for tests T1, T2 and T3) in the second dataset (116 students). Blue color represents students that fail the evaluation (or absent themselves) and red color represents students that pass the evaluation. This chart is plotted by the Weka framework.

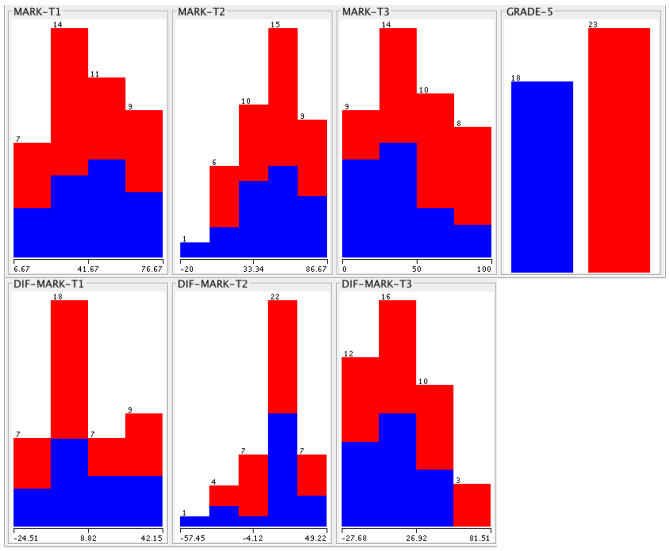

Fig. 3. Class distribution visualizing marks and differences with the average marks (for tests T1, T2 and T3) in the third dataset (41 students). Blue color represents students that have grades D or E and red color represents students that have grades A, B or C. This chart is plotted by the Weka framework.

In the next step we calculated the differences between the mark itself and the average valued achieved in that test by students during their course. Therefore, we incorporated 3 new attributes to the dataset. In Fig. 2 it is shown the class distribution for such new attributes.

Finally, once we have detected patterns to separate students that pass the evaluation and those students that do not pass it, we were interested in inducing some models that could find some pattern to differentiate between best students (with A, B or C grades) and the rest of students that pass the evaluation (D or E grades). In this dataset we only had 41 students so the induction algorithm had some problems with so few examples. To solve it we resample the dataset [14] making it five times bigger (205 examples) and configured J48 to examine a bigger number of examples before expanding (minimum of 20 examples) in order to avoid overfitting and reduce the complexity of the model [12]. In Fig. 3 it is shown the class distribution for this last dataset.

III. RESULTS

In this section we present the results that we have collected

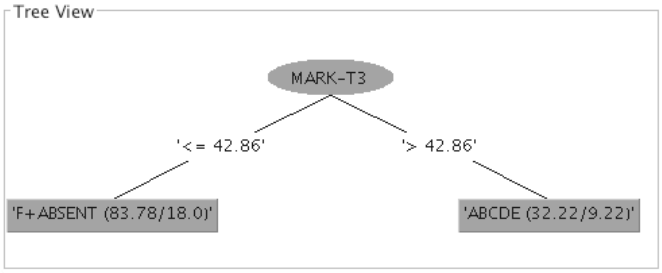

Fig. 4. Decision tree induced by J48 using the first dataset. Attributes are the marks for T1, T2 and T3; the binary class separate between students that pass (ABCDE) or not pass (F+ABSENT) the global evaluation process.

after applying the mining process to the data previously described. As we have explained, we have used a TDIDT algorithm (J48 implementation of C4.5), so the induced models are decision trees, what make possible an easy interpretation of the patterns. In addition, we can rely on the results, because validation processes show high confidence levels. The validation processes we have conducted are 10-fold cross validations.

For the first dataset, that which separates students in a binary class (pass or not pass the evaluation) and only include the marks achieved for every test (T1, T2 and T3), the pattern is easy to understand (even no TDIDT algorithm would be necessary because the class distribution in Fig. 1 shows a similar information). The most important attribute to determine the difference between two student profiles is the mark achieved for the last test (T3), the most close to the final exam. The decision tree, shown in Fig. 4, is not surprising, but reflects the ability of machine learning algorithms to find patterns. Furthermore, the validation shows 80% accuracy, quite reliable considering the number of examples and the class unbalance.

Analyzing the second dataset, extended with new attributes that summarize the differences between the own mark and the average value, some additional knowledge is extracted. Decision tree (Fig. 5) reveals that once we know the mark for T3 (root node), we can detect some other differences. In this case, the new added attributes reveal as important elements to determine the final achievement of the students. Particularly students that are below 42.86 points in T3, need to do best that the average in T1 and T2 to pass. So the requirements are not

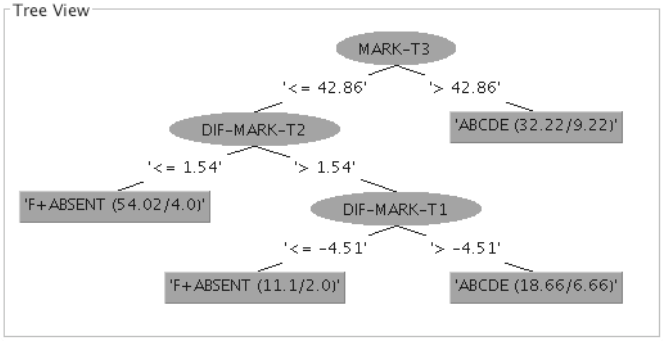

Fig. 5. Decision tree induced by J48 using the second dataset. Attributes are the marks (for T1, T2 and T3) and the difference with the average value; the binary class separate between students that pass (ABCDE) or not pass (F+ABSENT) the global evaluation process.

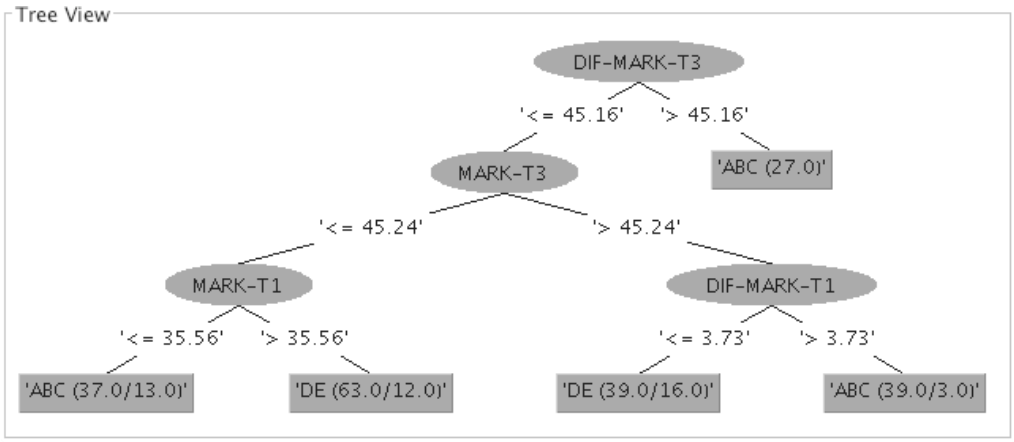

Fig. 6. Decision tree induced by J48 using the third dataset. Attributes are the marks (for T1, T2 and T3) and the difference with the average value; the binary class separate between students that achieve best grades (ABC) and the other ones (DE).

so restrictive for them, note that students do not need to pass the test (50 points out of 100), they only need to do best that the average value (close to 35 out of 100, see Table I)

Once again, the validation presents a quite reliable model (even higher than previous model) because we have 87% accuracy. This makes sense because we have added new attributes that help to better differentiate between student profiles.

Finally, once we have identified some criteria that determine differences between students that pass or not pass the final evaluation, we focus in those that pass the evaluation and how good their results are. Concretely we want to know if there is some element that reveals how they differ. In Fig. 6 we show the decision tree induced by J48 which reliability is relatively high (close to 80% accuracy).

Once again, the last test (T3) seems the most decisive element. It is logical, because this test includes and extends the concepts and abilities needed for the second test (T2). But this time the model differs substantially from previous ones because the actually important attribute is not the mark itself, but the difference with the average value. For every student (there is no exception, see most right-side branch in the decision tree) which T3's mark is beyond the average value in more than 45.16 points (out of 100), the final grade is better than D (A, B or C grade). Note that this difference is even greater than the standard deviation (27.84).

For those students that do not surpass the average value in such quantity, we find both kinds of students. In this case, differences between them are less clear and they could be even misunderstanding at a first moment. As it can be seen in the decision tree, first test information (T1) is selected to expand the tree in the deepest levels. It seems strange that students with lower marks ($\leq$ 35.56) get highest grades in the final evaluation, but we found some explanations that diminish the importance of such strangeness. On one hand, we can see that such asseveration is not so strong, because not all the examples are correctly classified (see <second> number in leaves), so some level of noise is present in that attribute. On the other hand, if we know that first test (T1) is conducted at the beginning of the semester and its relation to final exam is very poor, we can think that dependencies are arguable; even more, we can suppose that good students with a "poor" mark in the first test can detect the necessity of strengthen the efforts because they did an incorrect initial calibration about the difficulty of the subject.

IV. CONCLUSION

In this paper we have studied, by using data mining techniques, the possibility that learning processes in the academic context could incorporate new and relevant knowledge that enables improvements in such processes.

In the conducted analysis we have detected that there are relations between the continual assessment carried out during the semester and the final evaluation. These relations, correctly used, can lead the adaptation of existing strategies or to boost the integration of new methods in subjects for future courses.

To a large extent, such improvements depend on having enough data about the evolution of the evaluations, on analyzing them continuously, on detecting anomalous behaviors; and on developing preventive and corrective actions (new exercises, individual tutorial actions, etc.). At this moment, Web-based Educational Systems offer tools to obtain and process that data, so its usage is highly recommended.

In addition, due to the flexibility of these systems, they can be adapted and extended. New functionalities can be added, and two different developments can be incorporated to progress in the previously mentioned improvements. As a first point, Web-based Educational Systems can collect more data, those that have shown their usefulness for data mining analysis (even calculating new fields). As a second feature, they could incorporate the mining process in the core of the system in order to offer a dual advantage: helping the teacher with the analysis tasks (assessment task) and helping the students by guiding their learning process (adapting task).

This study reveals many future research lines in different dimensions. There exists a wide diversity of techniques in the data mining field, so selecting other paradigms could improve the knowledge acquired (association rules, decision rules,

etc.). If we are not so interested in the understandable knowledge (assessment task) and we prefer to provide the system with better guiding characteristics (adapting task), we have a perspective even broader because we could use many other strategies not so easily human-readable but very accurate (ensembles, neural networks, etc.).

Another promising area is the automatic or semi-automatic tune up of the Web-based Educational Systems. It is interesting to modify the educational system to respond to specific necessities of students [15]. This adaptation could even be implemented in real time, responding during the interaction with the student.

In this sense there are emerging new areas in machine learning and data mining related with data streams [16], very large (even non-ended) datasets that grow increasingly. Its usage fits very well with the dynamic of Web-based Educational Systems that are open constantly and can interact with students (and receive data) at every moment. Therefore, incorporating incremental algorithms [17] that can learn in this context would be positive. Additionally, as the student profile is not static, providing mechanisms to detect concept drift [18, 19] would contribute to create much more adaptable systems.

REFERENCES

[1] P. Brusilovsky, E. Schwarz and G. Weber, "ELM-ART: An intelligent tutoring system on World Wide Web", *Lecture Notes in Computer Science*, vol. 1086, pp. 261-269. 1996.

[2] C. Romero, S. Ventura, A. Zafra and P. de Bra, "Applying Web usage mining for personalizing hyperlinks in Web-based adaptive educational systems", *Computers & Education*, vol. 53, pp. 828–840. 2009.

[3] E. Melis, E. Andrès, J. Büdenbender, A. Frishauf, G. Goguadse, P. Libbrecht, M. Pollet and C. Ullrich, "ActiveMath: A generic and adaptive web-based learning environment", *International Journal of Artificial Intelligence in Education*, vol. 12, pp. 385-407. 2001.

[4] P. Brusilovsky and C. Peylo, "Adaptive and Intelligent Web-based Educational Systems", International Journal of Artificial Intelligence in Education, vol. 13, pp. 156-169. 2003.

[5] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, "From Data Mining to Knowledge Discovery in Databases", AI Magazine, vol. 17, pp. 37-54. 1996.

[6] C. Romero and S. Ventura, "Educational data mining: a review of the state of the art", *IEEE Transactions on Systems, Man, and Cybernetics-Part C*, vol. 40, pp. 601-618. 2010.

[7] E. Guzmán, R. Conejo, and J. Gálvez, "A Data-Driven Technique for Misconception Elicitation", *Lecture Notes in Computer Science*, vol. 6075, pp. 243-254, 2010.

[8] C. Romero, S. Ventura, and P. De Bra, "Knowledge discovery with genetic programming for providing feedback to courseware author", *User Model. User-Adapted Interaction*, vol. 14, pp. 425-464. 2004.

[9] R. Conejo, E. Guzmán, E. Millán, M. Trella, J. L. Pérez-De-La-Cruz and A. Ríos, "SIETTE: A Web-Based Tool for Adaptive Testing", *International Journal of Artificial Intelligence in Education*, vol. 14, pp. 29-61. 2004.

[10] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten, "The WEKA Data Mining Software: An Update", *SIGKDD Explorations*, vol. 11, pp. 10-18. 2009.

[11] European Communities, "*ECTS Users' Guide*", Luxembourg: Office for Official Publications of the European Communities, 2009.

[12] T. M. Mitchell, *Machine Learning*. New York, McGraw-Hill, 1997, ch. 3.

[13] J. R. Quinlan, *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA. 1993.

[14] P. H. Lee, "Resampling Methods Improve the Predictive Power of Modeling in Class-Imbalanced Datasets", *International Journal of Environmental Research and Public Health*, vol. 11, pp. 9776–9789. 2014

[15] A. Kozierkiewicz-Hetmańska and N. Nguyen, "A method for learning scenario determination and modification in intelligent tutoring systems", *International Journal of Applied Mathematics and Computer Science*, vol. 21, pp. 69-82. 2011.

[16] J. Gama, *Knowledge Discovery from Data Streams*. Chapman & Hall/CRC. 2010.

[17] J. del Campo-Ávila, G. Ramos-Jiménez, J. Gama and R. Morales-Bueno, "Improving the performance of an incremental algorithm driven by error margins", *Intelligent Data Analysis*, vol. 12, pp. 305-318. 2008.

[18] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy and A. Bouchachia, "A survey on concept drift adaptation", *ACM Computing Surveys*, vol. 46 (4), Article 44, 37 pages. 2014.

[19] I. Frías-Blanco, J. del Campo-Ávila, G. Ramos-Jiménez, R. Morales-Bueno, A. Ortíz-Díaz and Y. Caballero-Mota, "Online and non-parametric drift detection methods based on Hoeffding's bounds", *IEEE Transactions on Knowledge & Data Engineering*, to be published.

José del Campo-Ávila received the Ph.D. degree in Software Engineering and Artificial Intelligence from the University of Málaga in 2007. He is an Associated Professor in the Languages and Computer Science Department of the University of Málaga, Málaga, Spain, where he has worked since 2003. His research interests include incremental learning, mining data streams for classification, and multiple classifier systems, among others. Dr. del Campo-Ávila is member of the Research and Applications in Artificial Intelligence Group. He is also a regular member of program committees of international conferences, such as Intelligent Data Analysis (IDA) or Data Stream track in ACM Symposium on Applied Computing (DS-SAC).

Ricardo Conejo received the Ph.D. degree in Ingeniero de Caminos, Canales y Puertos (Civil engineer) from the Technical University of Madrid, in 1995. He is a Full Professor in the Languages and Computer Science Department of the University of Málaga, Málaga, Spain, where he has worked since 1986. In addition, he is the Director of the Research and Applications in Artificial Intelligence Group at the same university. He has 25 years of experience in the development of intelligent tutoring systems. He has published extensively on this field. His research interests currently focus on adaptive testing, student knowledge diagnosis and intelligent tutoring systems. He has also worked on fuzzy logic, model-based diagnosis, multiagent systems, and artificial intelligence applied to civil engineering.

Prof. Conejo is a regular member of program committees of international conferences, such as User Modeling Adaptation, and Personalization (UMAP), Intelligent Tutoring Systems (ITS), and Artificial Intelligence in Education (AIED).

Francisco Triguero was born in Málaga, Spain, in 1955. He received the Mathematics Degree, in 1977 from the University of Málaga, Spain, and the Ph.D. degree, in 1982 from the University of Granada, Spain. He works in intelligent tutoring systems, knowledge discovery, and computational learning.

He was a Lecturer in the Mathematics Department of Málaga University (1977–1980), Assistant Professor in the Mathematics Department of Málaga University (1980–1984), Full Professor of Business Mathematics at the University of Seville (1984–1986), and Full Professor of Business Mathematics at the University of Málaga since 1986. He has occupied the position of Director of the Computer Service of Málaga University. Founder and first Director of the Computer Science School and the Telecommunication School, in 1987, first elected Director of the Computer Science School, in 2000, President Chair of the CAEPIA'97 (Conference of AEPIA), and Cochairman of ICALP'2002 and AH'2002.

Dr. Triguero is a member of EATCS (European Association for Theoretical Computer Science) and AEPIA (Spanish Association for Artificial Intelligence), member of the Andalusian Universities Council, and Member of the Advice Council of the Andalusian Institute of Development.

Rafael Morales-Bueno was born in Málaga, Spain, in 1955. He received the Mathematics Degree, in 1978 from the University Complutense of Madrid, Spain, and the Ph.D. degree, in 1991 from the University of Malaga, Spain. He works in data mining, knowledge discovery, computational learning and also in computability and complexity (from classical and fuzzy point of view), among others. He was an Assistant Professor in Computer Science Department of Málaga University (1984–1987), Associate Professor (1987-2009), and Full Professor of Computer Science (2009-). He has occupied the position of Director of the Computer Science School (2000-2004) and Vice-rector of Infrastructures (2004-2012). Chairman of ICALP'2002 and other conferences. Dr. Morales-Bueno is a member of the Research and Applications in Artificial Intelligence Group. He is also a Member of EATCS (European Association for Theoretical Computer Science) and AEPIA (Spanish Association for Artificial Intelligence). Rafael Morales-Bueno is also Vice-president of Momopocket EDE S.L. (Electronic money Entity), a company specialized in mobile payment systems.