Data Mining of Virtual Campus Data

Alfredo Vellido[1], Félix Castro[1, 2], Terence A. Etchells[3], Àngela Nebot[1], and Francisco Mugica[4]

[1] Dept. Llenguatges i Sistemes Informàtics, LSI,
Universitat Politècnica de Catalunya, Campus Nord,
C. Jordi Girona 1-3, Barcelona 08034, Espanya
{avellido, fcastro, angela}@lsi.upc.edu

[2] Centro de Investigación en Tecnologías de Información y Sistemas,
CITIS, Universidad Autónoma del Estado de Hidalgo,
Ciudad Universitaria, Carretera Pachuca-Tulancingo km. 4.5,
Hidalgo, México

[3] School of Computing and Mathematical Sciences,
Liverpool John Moores University,
Byrom St., L3 3AF, Liverpool, UK
t.a.etchells@livjm.ac.uk

[4] Instituto Latinoamericano de la Comunicación Educativa (ILCE),
Calle del Puente 45, México D F 14380, México
fmugica@ilce.edu.mx

Summary. As mentioned elsewhere in this book, e-learning offers "a new context for education where large amounts of information describing the continuum of the teaching–learning interactions are endlessly generated and ubiquitously available". But raw information by itself may be of no help to any of the e-learning actors. The use of Data Mining methods to extract knowledge from this information can, therefore, be an adequate approach to follow in order to use the obtained knowledge to fit the educational proposal to the students' needs and requirements. This chapter provides a case study in which several advanced Data Mining techniques are employed to extract different types of knowledge from virtual campus data concerning students system usage behaviour. The diverse palette of Data Mining problems addressed here include data clustering and visualization, outlier detection, classification, feature selection, and rule extraction. They concern diverse e-learning problems, such as the characterization of atypical students' behaviour and the prediction of students' performance. Different Data Mining techniques from the areas of Statistical and Machine Learning; Fuzzy Logic, and Inductive Reasoning are

A. Vellido et al.: *Data Mining of Virtual Campus Data,* Studies in Computational Intelligence (SCI) **62**, 223–254 (2007)
www.springerlink.com

employed to tackle these problems. Strong emphasis is placed on the interpretability of the results, obtained through rule extraction, so that they can be fed back to the e-learning system in a practical an efficient manner.

1 Introduction

Any e-learning system is, by its own nature, likely to generate large amounts of information describing the continuum of the teaching-learning interactions almost in real time. All this information, gathered from diverse and usually heterogeneous sources, may be of no help by itself to any of the e-learning actors in its raw form. Actually, an excess of such information can become a liability for e-learning tutors and managers unless it is processed according to reasonable goals. Data Mining can provide the adequate tools for such processing; obtaining actionable patterns from large data repositories. The use of Data Mining methods to extract knowledge from the e-learning system available information can, therefore, be an adequate approach to follow in order to use the obtained knowledge to fit the educational proposal to the students' needs and requirements.

Virtual campus environments, such as the one that is the subject of this case study, are fastly becoming a mainstream alternative to traditional distance higher education. The Internet medium they use to convey content, also allows the gathering of information on students' online behaviour. Here, we focus on e-learning systems improvement through the analysis of the data generated by the virtual campus students, aiming to discover their system usage patterns.

The amount of research concerned with the mining of data generated by the usage of e-learning systems is still somehow scarce on the ground (see Castro et al., this book). In this study, we address two main problems concerning virtual campus students' behaviour: the characterization of atypical behaviour and the prediction of students' performance. Several Data Mining problems are concerned, namely: students' data clustering and visualization, behavioural outlier detection, students' classification according to course marks, data feature selection, and rule extraction. The latter becomes of paramount importance as we aim to place a strong emphasis on the interpretability of the obtained results; no matter how accurate these might be: unless they are translated into practical and efficient rules that system managers and tutors can act upon, it will be extremely difficult to feed them back to the e-learning system.

The rest of the chapter is structured as follows: First, in Sect. 2, we provide a detailed description of the data under analysis; they correspond to students of a particular course of the Center of Studies in Communication and Educational Technologies (CECTE, as Spanish acronym) virtual campus. This is followed by successive sections devoted to the different problems at hand. Atypical student behaviour is analysed using the Generative Topographic Mapping (GTM) [1] model for data clustering and visualization. Students' performance prediction and feature selection are dealt through Fuzzy Inductive Reasoning (FIR) [11] and Artificial Neural Networks. The models for rule extraction are Orthogonal Search-based Rule Extraction (OSRE) [4] and FIR. Finally, Sect. 5 concludes the study summarizing its findings.

2 Data from the CECTE Virtual Campus

CECTE is a partially virtual campus, offering postgraduate courses and continuous education (graduate, workshops and specific courses) to Latin-American students. The CECTE is part of the international Latin-American Institute of Educative Communication (ILCE), whose main goal is to offer postgraduate courses.

The ILCE was born in the General Conference of the United Nations Organization for Education Science and Culture, carried out in Uruguay in 1954. Its aim was alleviating the educative needs of the Latin-American region through the use of audiovisual media and technological resources. From inception, the ILCE has tried to apply the most up-to-date tools and technologies (from phonographic records to the Internet) to deliver high-quality distance courses. In February of 1979, ILCE obtained the status of International Organization with juridical personality, its own patrimony and autonomous administration.

Throughout its 50 years of existence, ILCE has undergone structural and technologic changes in order to maintain the high quality of its educative services, taking advantage of the computer, multimedia and telecommunications methodological breakthroughs, and staying at the forefront of the evolution of the distance education field. Moreover, ILCE has signed collaborative agreements with the most prestigious institutions, public and private, offering distance education services around the world, including the British Broadcasting Corporation (BBC), the Open Learning Agency (OLA), the University of British Columbia (UBC), the Public Broadcasting Service (PBS), and the Educational Management Group (EMG), to name just a few.

As previously stated, CECTE is the part of ILCE offering postgraduate courses and continuous education, including master and graduate degrees, workshops and specialized courses. At the moment, up to 2,000 students are registered in any of the offered programs. CECTE has developed 39 specialized courses addressed to educative, public and governmental institutions. The most demanded CECTE courses follow a hybrid, semi-presential model, in which students take courses online through the Web CECTE (WCECTE) but also attend weekly TV sessions through the National System of Educative Television (EDUSAT), which takes advantage of digitalized satellite, optical fibre and telephony technologies. Through WCECTE, students can access the course materials and communicate and interact with each other through an e-mail system and a discussion forum. Moreover, all weekly TV sessions are available to all students through WCECTE. The environment also includes an agenda, a news system, virtual classrooms, a digital library, interactive tutorials, and other related tools.

The educative model of CECTE comprises the interaction and communication between all actors involved in the teaching-learning process: Students, branch coordinators, advisors and tutors, mainly. An important characteristic of the CECTE educative model is the activities calendar, consisting of a list of activities and learning suggestions made by the module advisor.

Advisors are entrusted to deliver a specialized weekly session and interact, communicate, coordinate, and answer possible doubts from course tutors. Branch coordinators are entrusted to evaluate the branch activities and to check if students perform their assigned activities properly. The tutor is a very relevant actor, as he or she interacts directly with students, assigning learning activities, answering doubts, opening topics in discussion forums, evaluating the activities performed by learners, and verifying that the teaching-learning process is adequate, taking advantage of all the tools provided by WCECTE.

The evaluation process is based on a set of activities performed by the students, in both WCECTE and the physical branch. An important evaluation element in the course analysed in this chapter is the students' learning behaviour measured through the analysis of comments and homework posted by them to discussion forums. Two specialized discussion forums were used in the analysed course: general topics about the course and specific class plan topics comments.

Two novel evaluation topics, not often used in e-learning environments, were incorporated in the course: co-evaluation and experience report. In co-evaluation, the advisor grades how well the student evaluates the class plans of his/her course mates. The experience report is a student description of his/her perception of the course. It can be viewed as a self-evaluation of the student's own learning process.

For the experiments in this chapter, a set of 722 students, enrolled in the "Didactic Planning" graduate course, was selected. The course is addressed to second term high school teachers offering specialized subjects, namely Mathematics, Chemistry, Mexican History, Computer Science, English, as well as Reading and Writing, and Ethics and Values workshops. The students are meant to perform a set of activities throughout the course with the main purpose of learning new methods and strategies for planning the classes that they teach. This is the reason why these activities are centred on the so-called "class plan". A class plan is a document where a set of strategies are suggested to develop a teaching–learning session, taking into account different factors that appear in the educational process, such as students' characteristics, teaching style, teachers' experience, etc. The data features available for this study are detailed on Table 1.

3 Characterization of Atypical Virtual Campus Usage

The detection of atypical behaviour in a virtual campus is a research goal on its own. There is much to be learnt from atypical behaviour online, as it can provide clues about what might be failing in the e-learning system, or about virtual campus facilities that might have not been considered.

Here, we approach the discovery of atypical behaviour as a combination of data clustering and outlier detection. Unlike in classification problems (subject of coming sections of this chapter), in data clustering we are not interested in modelling a relation between a set of multivariate data items and a certain set of outcomes for each of them (being this in the form of class membership labels). Instead, we aim to discover and model the groups in which the data items are clustered, according to some item similarity measure.

If a clustering model is used as a Data Mining tool to characterize the groups of online students, we would expect this model to provide a sensible cluster structure despite the presence of data outliers. Outliers are loosely defined here as data items (corresponding to individual students) that are somehow removed from the most populated areas in data space. We assume that outlierness in this context corresponds to atypical student usage behaviours.

Table 1. Data features collected for the experiment

Feature	Alias	Description
Age of the student	AGE	Age of the student
Area of expertise	EXP	Area of expertise of the student (mathematics, chemistry, Mexican history, etc.)
Gender	G	Student's gender
Level of studies	STD	Level of studies (graduate, master, Ph.D., etc.)
Position of the student	POS	Position of the student as a teacher in his/her school
Percentage of the activities performed by the student	ACT	Percentage of the activities performed by the student with respect to the total activities of the course
Percentage of session assistance	ASS	Percentage of student's session assistance with respect to the total number of sessions of the course
Average mark of the e-mail	MAIL	Average mark obtained by the student in the activities sent by e-mail
Average mark of the co-evaluation	COEV	Average mark of the co-evaluation performed by the student of the class plan of other students. The advisor grades how well the student evaluates the class plans of his/her course mates
Average mark of the forum participation	F	Average mark of the student's forum participation (referring to topics related to the course)
Average mark of the forum class plan	FCP	Average mark of the forum class plan (referring only to topics related to the class plan exclusively)
Average mark of the final class plan	FC	Average mark obtained by the student in his/her final class plan
Average mark of the initial class plan	IC	Average mark obtained by the student in his/her initial class plan
Average mark of the experience report	ER	Average mark obtained by the student in the experience report. In this report the student evaluates his/her learning process and describes the main concepts learned
Average mark of the work in the branch	BR	Average mark of the work (activities) performed in the branch
Final mark	MARK	Final mark obtained by the student in the course

The model used in this study for simultaneous data clustering, visualization, and outlier detection, is called GTM [1], which is both a probabilistic alternative to the well-known Self-Organizing Maps (SOM) [8] (also referred to as Kohonen Maps) and a constrained mixture of Gaussians model. Gaussian mixture models are known to lack robustness in the presence of data outliers, and mixtures of multivariate Student *t*-distributions have been suggested as a more robust alternative [14]. In this study, the GTM is applied as a constrained mixture of Student *t*-distributions: the *t*-GTM [17], and used to detect atypical learning behaviour from the actions performed by the CECTE students. This outlier detection is part of a more general clustering process that aims to group students with similar navigational behaviour.

One of the most important phases of a Data Mining process (and one that is usually neglected) is that of data exploration through visualization methods. The interpretability of the clustering results provided by the *t*-GTM, even in terms of this exploratory visualization, can be limited for data sets consisting of too large a number of features. This situation is not uncommon for a wide range of real-world problems concerning clustering methods. Principled methods of feature selection have for long been the preserve of supervised methods. In comparison, feature selection for unsupervised learning has received far less attention. The interpretability of a clustering solution would be greatly improved by its description in terms of a reduced subset of relevant features. Recently, an important advance in feature selection for unsupervised model-based clustering was described in [10] for Gaussian mixture models and extended to GTM and *t*-GTM in [16, 18].

In this section, we first briefly introduce the GTM and its robust *t*-GTM variant. This is followed by the description of an unsupervised method for feature selection associated to it. Outlier detection, clustering, feature selection and visualization results are then, in turn, presented and discussed.

3.1 GTM in a Nutshell

The GTM [1] was originally formulated both as a probabilistic clustering model alternative to the heuristic SOM [8] and as a constrained mixture of Gaussian distributions. It is precisely these constraints what enables it for cluster visualization, overcoming a common limitation of generic mixture models. The GTM can also be seen as a non-linear latent variable model that defines a mapping from a low-dimensional latent (non-observable) space onto the observed data space. The mapping is carried through by basis functions generating a (mixture) density distribution. The functional form of this mapping for each variable *d* can be expressed as

$$y_d(\mathrm{u},\mathrm{W}) = \sum_m^M \phi_m(\mathrm{u}) w_{md}, \tag{1}$$

where $\boldsymbol{\Phi}$ are basis functions $\Phi(\mathrm{u}) = (\phi_1(\mathrm{u}),\ldots,\phi_M(\mathrm{u}))$ that introduce the non-linearity in the mapping; $\mathbf{W}$ is the matrix of adaptive weights w_{md} that defines the mapping; and u is a point in latent space. In order to provide an alternative to the visualization space defined by the characteristic SOM lattice, the latent space of the GTM is discretized as a regular grid of *K* latent points u_k. The mixture density for a data point $\mathbf{x}$, given Gaussian basis functions, can be written as

$$p(\mathrm{x}\,|\,\mathrm{W},\beta) = \frac{1}{K}\sum_{k=1}^{K}\left(\frac{\beta}{2\pi}\right)^{D/2} \exp\left\{-\beta/2\,\|\mathrm{y}_k - \mathrm{x}\|^2\right\}, \tag{2}$$

where the *D* elements of $\mathbf{y}$ are given by (1). This density allows for the definition of a model likelihood, and the well-known Expectation-Maximization (EM: [2]) algorithm can be used to obtain the Maximum Likelihood estimates of the adaptive parameters (W and β) of the model. See [1] for details on these calculations.

3.2 Data Clustering and Visualization Using GTM

Model interpretation usually requires a drastic reduction in the dimensionality of the data. Latent variable models can provide such interpretation through visualization, as they describe the data in intrinsically low-dimensional latent spaces. Each of the latent points u_k in the latent visualization space – which in this study is a 2-dimensional space – is mapped, following (1), as $\mathrm{y}_k = \Phi(\mathrm{u}_k)\mathrm{W}$. The y_k points are usually known as *reference vectors* or *prototypes* and their dimensionality, in the case of this study, is that of the original CECTE virtual campus data, described in Table 1. Each of the reference vector elements corresponds to one of the observed features, and their values over the latent visualization space can be colour-coded to produce *reference maps* that provide information on the behaviour of each variable and its influence on the clustering results.

Each of the latent space points u_k can be considered by itself as a cluster representative (of a cluster containing the subset of course students assigned to it). For simplicity, we use for the GTM a cluster assignment method akin to that of SOM, which is based on a winner-takes-all strategy: Each data observation (student) is assigned to the location in the latent space (cluster) where the mode of the corresponding posterior distribution is highest, i.e. $\mathrm{u}_n^{\mathrm{mode}} = \arg\max_{\mathrm{u}_k} r_{kn}$, where r_{kn} is the probability of student n belonging to cluster k, and it is obtained as part of the EM algorithm.

3.3 Handling Outliers with *t*-GTM

General Gaussian mixture models lack robustness in the presence of data outliers. For the Gaussian GTM, as a constrained mixture model, the presence of outliers is likely to negatively bias the estimation of parameters W and β. To overcome this, several recent studies have suggested the use of multivariate Student *t*-distributions as a robust alternative to Gaussians. The GTM can equally be redefined as a constrained mixture of Student *t*-distributions: namely, the *t*-GTM [17]. Assuming now that the basis functions $\boldsymbol{\Phi}$ are *t*-distributions, the mixture density in (2) can be redefined as

$$p(\mathrm{x}|\mathrm{W},\beta,\nu) = \frac{1}{K}\sum_{k=1}^{K}\frac{\Gamma((\nu/2)+(D/2))\beta^{D/2}}{\Gamma(\nu/2)(\nu\pi)^{D/2}}\left(1+(\beta/\nu)\left\|\mathrm{y}_k-\mathrm{x}\right\|^2\right)^{-(\nu+D)/2}, \quad (3)$$

where $\Gamma(\cdot)$ is the gamma function and the parameter ν can be understood as a tuner that adapts the level of robustness (divergence from normality) for the mixture. Again, the parameters of this model can be estimated using the EM algorithm. Details can be found in [17]. As a result of replacing the Gaussians by *t*-distributions, the impact of outliers on the estimation of the model parameters is effectively minimized.

3.4 Unsupervised Feature Selection with *t*-GTM

The interpretation of the clustering results provided by the *t*-GTM can become extremely difficult – and even impractical – for high-dimensional data sets. Therefore, the development of a method for feature relevance determination (FRD) and a criterion for feature selection (FS) based on it should sensibly increase their interpretability.

Recently, a method for FRD in unsupervised model-based clustering with mixture of Gaussians models was presented in [10] and extended to the *t*-GTM in [18]. This method calculates an unsupervised feature saliency as part of the EM algorithm. Such saliency measures the relevance of a feature on the definition of the cluster structure yielded by the model. It could be argued that the presence of outliers in the data sample is likely to bias the estimation of this saliency: In Gaussian mixtures, too many mixture components tend to fit the atypical data; as a result, those features which define what is atypical in the outliers might be attributed too high a saliency. The use of *t*-GTM should discount the negative effect of data outliers.

Formally, the saliency of feature *d* (its relevance) is defined as ρ_d. A value of $\rho_d = 1$ indicates the full relevance of feature *d*. According to this definition, the mixture density in (3) can be rewritten as

$$p(\mathrm{x} \mid \mathrm{W}, \beta, \nu) = \sum_{k=1}^{K} \frac{1}{K} \prod_{d=1}^{D} \left\{ \rho_d p\left(x_d \middle| \mathrm{u}, \mathrm{W}, \beta, \nu\right) + \left(1 - \rho_d\right) q\left(x_d \middle| \lambda_d\right) \right\}. \quad (4)$$

A feature *d* is irrelevant, with *irrelevance* $(1 - \rho_d)$, if, for all the *t*-GTM mixture components, $p(x_d | \mathrm{u}, \mathrm{W}, \beta, \nu) = q(x_d | \lambda_d)$, where $q(x_d | \lambda_d)$ is a common density followed by feature *d*. Notice that this is tantamount to saying that the distribution for feature *d* does not follow the cluster structure defined by the model.

Once again, we can estimate the model parameters (most importantly, the relevance ρ_d) resorting again to Maximum Likelihood and the EM algorithm. Details can be found in [18].

3.5 Experimental Results and Discussion

The first part of the experiments concerns the identification and characterization of atypical (outlier) students. The *t*-GTM parameters W and w_0 are initialized to fixed values, following a standard procedure [1], that ensures the replicability of the results. The visualization grid of *t*-GTM latent centres is fixed to a square layout of 5×5 nodes (i.e., 25 constrained mixture components). The corresponding grid of basis functions ϕ_m is fixed to a 3×3 layout.

Following Peel and McLachlan [14], a given data observation *n* will be considered to be an outlier if the value of $\Omega_n = \sum_k r_{kn} \beta \left\| \mathrm{y}_k - \mathrm{x}_n \right\|^2$ is sufficiently large. Notice that r_{kn}, as previously mentioned, is the responsibility assumed by a cluster *k* for data observation (student) *n*. A histogram for this statistic is provided in Fig. 1, showing a reasonably well-defined structure that evidences the difference between a majority of students with low values of Ω_n and a minority with large values. The selection of a threshold to differentiate outliers from non-outliers depends on the restrictions of the analysis in a real context. In this study, for illustrative purposes, such threshold has been placed at $\Omega_n = 1.75\times 10^7$, leaving 43 students (just below a 6%) as atypical cases or outliers.

We are interested in characterizing these atypical students. For that, we first visualize how the data have been mapped onto the *t*-GTM visualization space in Fig. 2 (top-left plot). The relative size of each cluster (square) is an indicator of the ratio of data observations assigned to that specific cluster by the model. Such assignment, as described in Sect. 3.2, is based on the modes $\mathrm{u}_n^{\mathrm{mode}} = \arg\max_{\mathrm{u}_k} r_{kn}$, which indicate which cluster bears maximum responsibility for a data point (student). The rather homogeneous cluster spread indicates that outliers do not dominate the mapping, which agrees with what is expected from the *t*-GTM definition. Instead, the mapping based on a Gaussian GTM (on the bottom plot) is clearly dominated by big clusters of students, restricted to a very specific area of the map and surrounded by small clusters corresponding to outliers.

Also in Fig. 2 (top-right plot), we can visualize where the outliers (according to the threshold in Fig. 1) have been mapped onto. With few exceptions, they are mostly on the left-hand side of the map; in other words, they have been located by the *t*-GTM model in a very well-delimited area. This localization simplifies their interpretation in terms of the *t*-GTM reference maps of Fig. 3. Let us focus on the four clusters on the top-left corner of the plot, which have a very strong outlier presence. Their interpretation according to the reference maps is quite straightforward: they are characterized by medium-to-very low values of all features but AGE, which is medium-to-high. This means that the *outlierness* of these particular students has to do with their low involvement with the virtual campus activity, which seems to be characteristic of the most mature students. It is also worth noting that the distribution of the main outliers over the map, as seen on the top-right plot of Fig. 2, resembles quite closely the distribution in Fig. 3 of the low values for the variables MARK, FC, MAIL and of the high values for AGE. These results are coherent with the online teachers' perception that the mature students lack expertise in the use of technology, at least at the beginning of the course.

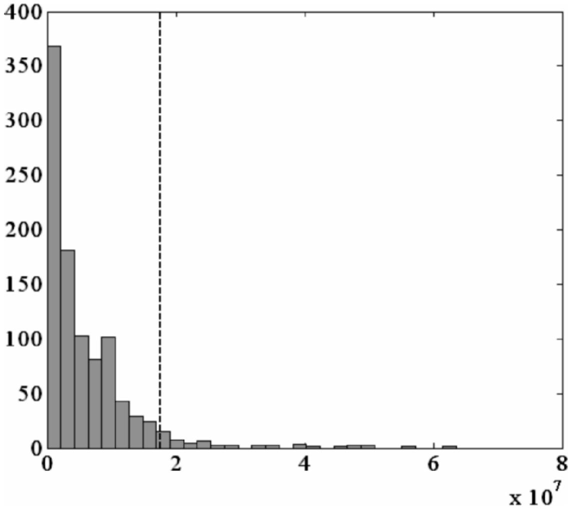


Fig. 1. Histogram of statistic Ω_n for the 722 students. An illustrative threshold is represented as a *dashed line*

The second part of the experiments concerns the estimation of the relative relevance of the data features. In this case, the values of the initial parameters are not fixed.

The feature saliency estimation results for the CECTE data are shown in Fig. 4. They neatly indicate that only one variable has consistently very high relevance: ACT. The average features (BR, F, FC and MAIL) and ASS have high but less consistent relevance. On the contrary, MARKS and, specially, AGE are clearly irrelevant in the definition of cluster structure. According to this, AGE influences cluster structure the least. With this new knowledge, we can now characterize the main outlier clusters described in previous paragraphs in a more parsimonious way: they are mainly defined by low to very low values of the average features, ACT, and ASS.

All these results show that useful knowledge can be extracted from the *t*-GTM combination of outlier detection, FRD and data clustering and visualization [19]. This knowledge could be fed back into the e-learning system in order to provide students with personalized guidance, tailored to their inhomogeneous needs and requirements. As a software tool embedded into the e-learning system, it would also help e-learning tutors to find patterns of student's behaviour.

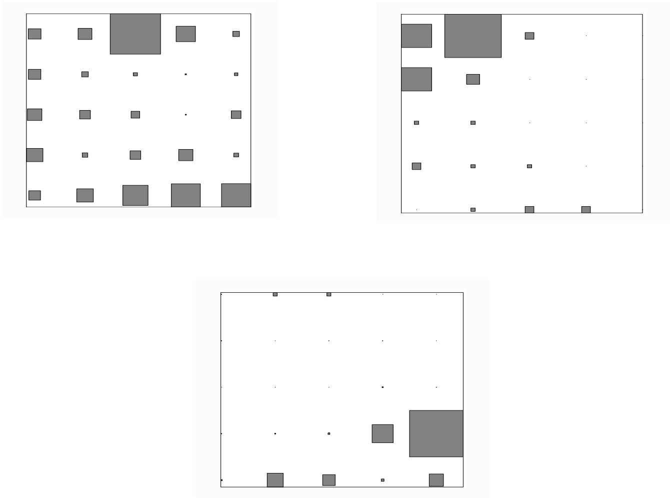


Fig. 2. Clustering of the 722 students on the *t*-GTM (*top-left*) and Gaussian GTM (*bottom*) 5 × 5 visualization spaces. On the *top-right* plot, only the *t*-GTM main outliers (43 students, according to Fig. 1) have been represented. For each plot, the size of the squares is proportional to the ratio of students assigned to the corresponding clusters by the model; therefore, squares of the same size in different plots do not necessarily correspond to the same number of students

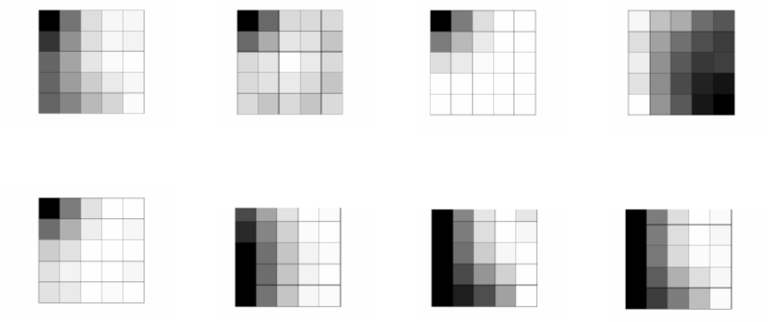


Fig. 3. Each cluster k in the 5 × 5 visualization space is associated to a reference vector $\mathrm{y}_k = \phi(\mathrm{u}_k)\mathrm{W}$. It consists of as many elements as features in Table 1. For brevity, only eight of them are displayed here: From *top* to *bottom* and *left* to *right*, reference maps for: MARK, ACT, ASS, AGE, BR, F, FC and MAIL. They are coded in grey-scale, from black (lowest values) to white (highest values). Given the correspondence between the layout of these maps and that of the clustering results in Fig. 2, the latter can be interpreted according to the former

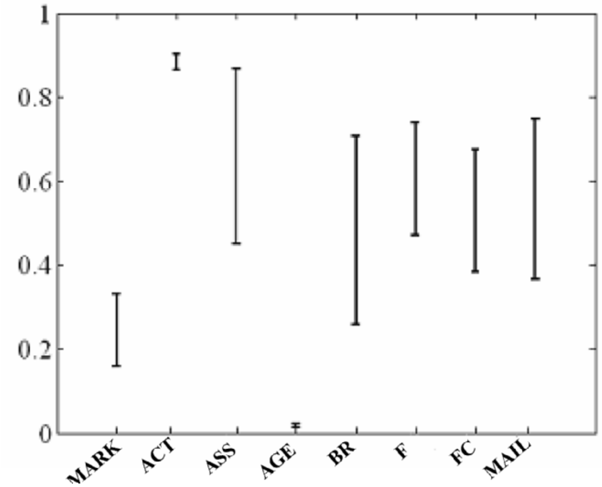


Fig. 4. FRD-GTM estimated values (represented by their means, plus and minus one standard deviation, over 20 runs of the algorithm) of the saliency vector ρ (see Sect. 3.4). Features with uncertain relevance (close to 0.5) show wider variations

4 Prediction of CECTE Students' Performance

In classification problems, we usually aim to model the existing relationships between a set of multivariate data items and a certain set of outcomes for each of them in the form of class membership labels.

One of the most difficult and time consuming activities for teachers in distance education courses is the evaluation process, due to the fact that, in this kind of courses, the review process should be done using collaborative resources such as e-mail, discussion forums, chats, etc. Additional problems are the usually high number of features involved and the complexity to define their influence in the final mark. Therefore, it would be helpful to reduce the dimensionality of the problem by identifying highly relevant features. In this way, it would be possible for teachers to provide feedback to students regarding their learning activities online and in real time. Moreover, the students' performance forecasting would become more interpretable.

In this section, we aim to predict the final mark of the users of the CECTE virtual campus. This prediction is understood as a classification problem, in which marks are grouped by intervals, each identified as a class. As explained in the introduction, we pay preferential attention to the interpretability of the results, trying to translate them into practical and efficient rules that system actors could use. To this end, classification is here intertwined with supervised feature selection and rule extraction.

4.1 Classification and Feature Selection with Fuzzy Inductive Reasoning

The methodological approach used to address the classification task is that of FIR. The conceptualization of FIR arises from the General Systems Theory proposed by Klir [7]. This modelling and qualitative simulation methodology is based on systems behaviour rather that on structural knowledge. It is able to obtain good qualitative relations between the variables that compose the system and to infer the future behaviour of that system. It also has the ability to describe systems that cannot easily be described by classical mathematics (e.g. differential equations), i.e. systems for which the underlying physical laws are not well understood. FIR consists of four main processes, namely: fuzzification, qualitative model identification, fuzzy forecast and defuzzification. Figure 5 describes the structure of the FIR methodology as applied in this study.

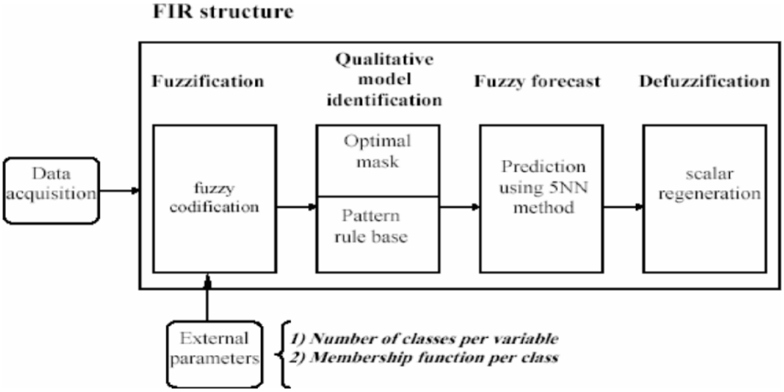


Fig. 5. Fuzzy Inductive Reasoning methodology

The fuzzification process converts quantitative data stemming from the system into fuzzy data. The qualitative model identification process is responsible for finding causal and temporal relations between variables and therefore for obtaining the best model that represents the system. An FIR model is composed of a mask (model structure) and a pattern rule base (behaviour matrix). Once the FIR model is available, the prediction system can take place using the FIR inference engine. This process is called fuzzy forecast. The FIR inference engine is a specialization of the k-nearest neighbour rule, commonly used in the pattern recognition field. Defuzzification is the inverse process of fuzzification. It allows converting the qualitative predicted output into quantitative values that can then be used as inputs to an external quantitative model.

Figure 6 illustrates the process of fuzzification by means of an example. A quantitative value is fuzzified into a qualitative triple, consisting of the class, membership, and side values. The side value, which is specific of the FIR technique and not commonly used in fuzzy logic, is responsible for presserving, in the qualitative triple, the complete knowledge contained in the original quantitative value.

Most fuzzy inference approaches preserve the total knowledge by associating multiple fuzzy rules consisting of tuples of class and membership values with each quantitative data value. In the example of Fig. 6, these rules would represent the temperature of 23° centigrade as being "normal" with likelihood 0.755 and being "warm" with likelihood 0.20. The point where two neighbouring classes match with a membership value of 0.5 is named *landmark*. In the example, the membership function of the class *Normal* is defined by landmarks {13,27}, being this pair the temperature values that specify the limits between the class *Normal* and its adjacent classes, *Fresh* and *Warm*, respectively.

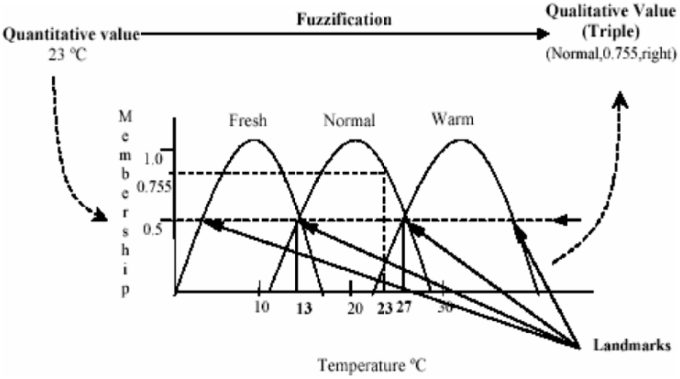


Fig. 6. FIR fuzzification process

The results of the fuzzification process are three matrices of identical size named qualitative data matrices, one containing the class values, the second containing the membership information, and the third containing the side values.

Let us now focus on how feature selection is embedded in the FIR methodology through the concept of masks. In FIR, a mask candidate matrix is the ensemble of all possible masks from which the best is chosen by either a mechanism of exhaustive search of exponential complexity, or by one of various suboptimal search strategies of polynomial complexity, as described in [6]. The mask candidate matrix contains elements of value −1 where the mask has a potential m-input, of value +1 where the mask has its m-output, and of value 0 to denote forbidden connections. A good mask candidate matrix to determine a predictive model for variable y_1 in the example of Fig. 7 might be, for instance:

x t	u_1	u_2	u_3	u_4	y_1	y_2
$t-2\delta t$	−1	−1	−1	−1	−1	−1
$t-\delta t$	−1	−1	−1	−1	−1	−1
t	−1	−1	−1	−1	+1	0

Each of the possible masks is compared to the others with respect to its forecasting power, which is maximal when the associate entropy measure is minimal. The Shannon entropy measure is used to determine the uncertainty associated with forecasting a particular output state given any legal input state. The Shannon entropy relative to one input state is calculated as

$$H_i = \sum_{\forall o} p(o \mid i) \log_2 p(o \mid i), \tag{5}$$

where $p(o|i)$ is the conditional probability of a certain m-output state o to occur, given that the m-input state i has already occurred. It denotes the quotient of the observed frequency of a particular state divided by the highest possible frequency of that state. The overall entropy of the mask is then computed as the weighted sum of the entropy over all input states:

$$H_m = -\sum_{\forall i} p(i) H_i, \tag{6}$$

where $p(i)$ is the probability of that input state to occur. The highest possible entropy $H_{\max}$ is obtained when all probabilities are equal, and zero entropy corresponds to totally deterministic relationships. A normalized overall entropy reduction H_r is defined as

$$H_\mathrm{r} = 1.0 - \left(\frac{H_m}{H_{\max}} \right). \tag{7}$$

H_r is a real-valued number in the range between 0 and 1, where high values indicate an improved forecasting power.

From a statistical point of view, every state should be observed at least five times [9]. Therefore, an observation ratio, O_r, is introduced as an additional contributor to the overall quality measure:

$$O_r = \frac{5n_{5x} + 4n_{4x} + 3n_{3x} + 2n_{2x} + n_{1x}}{5n_{\mathrm{leg}}}, \tag{8}$$

where n_{leg} is the number of legal m-input states, n_{1x} is the number of m-input states observed only once, n_{2x} is the number of m-inputs states observed twice, and so on.

The overall quality of a mask, Q, is then defined as the product of its uncertainty reduction measure, H_r, and its observation ratio, O_r:

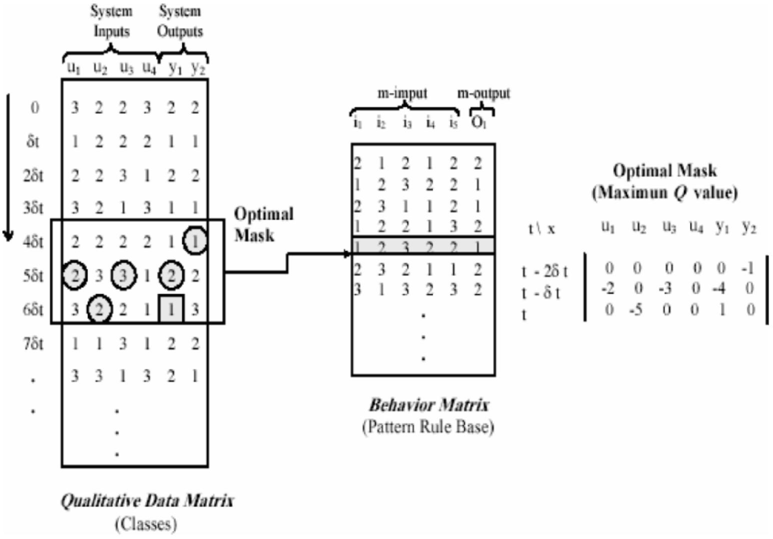
$$Q = H_r O_r. \tag{9}$$

The optimal mask is the mask with the largest Q value. An example of mask corresponding to Fig. 7 is

x t	u_1	u_2	u_3	u_4	y_1	y_2
$t-2\delta t$	0	0	0	0	0	−1
$t-\delta t$	−2	0	−3	0	−4	0
t	0	−5	0	0	+1	0

Each negative element in the mask is called an m-input (mask input). It denotes a causal relation with the output, i.e. it influences the output up to a certain degree. The enumeration of the m-inputs is immaterial and has no relevance. The single positive value denotes the output. The previous example mask contains five m-inputs. In position notation, it can be written as (6,7,9,11,14,17), enumerating the mask cells from top to bottom and from left to right.

Let us now address the second issue. How is the pattern rule base obtained from the mask? This process is illustrated in Fig. 7. The example mask can be used to "flatten" dynamic relationships into pseudo-static relationships. The left-hand side of Fig. 7 shows an excerpt of the quail-tative data matrix that stores the class values. It shows the numerical rather than the symbolic class values. In the example shown in Fig. 7, all the variables were discretized into three classes, with the exception of variable y_1 that was discretized into two classes. The dashed box symbolizes the mask that is shifted downwards along the class value matrix. The round shaded "holes" in the mask denote the positions of the m-inputs, whereas the square shaded "hole" indicates the position of the m-output. The class values are read out from the class value matrix through the "holes" of the mask, and are placed next to each other in the behaviour matrix that is shown on the right-hand side of Fig. 7.



	u_1	u_2	u_3	u_4	y_1	y_2
0	3	2	2	3	2	2
δt	1	2	2	2	1	1
2δt	2	2	3	1	2	2
3δt	3	2	1	3	1	1
4δt	2	2	2	2	1	1
5δt	2	3	3	1	2	2
6δt	3	2	2	1	1	3
7δt	1	1	3	1	2	2
.	3	3	1	3	2	1
.			.			

i_1	i_2	i_3	i_4	i_5	O_1
2	1	2	1	2	2
1	2	3	2	2	1
2	3	1	1	2	1
1	2	2	1	3	2
1	2	3	2	2	1
2	3	2	1	1	2
3	1	3	2	3	2
		.			

t \ x	u_1	u_2	u_3	u_4	y_1	y_2
t - 2δ t	0	0	0	0	0	-1
t - δ t	-2	0	-3	0	-4	0
t	0	-5	0	0	1	0

Fig. 7. FIR qualitative model identification process

Here, each row represents one position of the mask along the class value matrix. It is lined up with the bottom row of the mask. Each row of the behaviour matrix represents one pseudo-static qualitative state or qualitative rule (also called pattern rule). For example, the shaded rule of Fig. 7 can be read as follows: "If the first m-input, i_1, has a value of 1 (corresponding to high), the second m-input, i_2, has a value of 2 (corresponding to medium), etc., then the m-output, O_1, assumes a value of 1 (corresponding to high)".

The FIR inference engine computes a distance measure between the input pattern, for which the output prediction should be obtained, and all patterns stored in the behaviour matrix that match (with regard to the class value) the input pattern. The predicted output is then computed as a weighted mean of the outputs associated with the *k*-nearest neighbours, i.e. those neighbours that exhibit the smallest distance measure in the input space. For a deeper and more detailed insight into the FIR methodology, the reader is referred to [12].

4.1.1 Experimental Results

In accordance with the previous theoretical description, the aim of these experiments was twofold. On the one hand, we aimed to identify FIR models that were capable of predicting students' performance. On the other hand, we were interested in determining which data features had the highest relevance from the students' performance point of view. To this end, a first set of experiments was carried out using sevenfold cross-validation. Each test set was composed of 122 data samples whereas the training sets contained 600 data samples.

Before the model identification process of the FIR methodology can take place, it is necessary to provide the "number of classes" parameter for each system variable (see Fig. 5). In these experiments, due to the reduced size of the data set available, the minimum possible number of classes was chosen to discretize each data feature. Table 2 lists these parameters. The gender (G), area of expertise (EXP) and position (POS) are nominal variables and, therefore, the minimum representation corresponds to the number of values that each one can take. The rest of the variables were discretized into two or three classes. The dependent variable (the one to be predicted), final mark (MARK), was discretized into three classes to allow a better discrimination between bad, regular and good students.

As described in Sect. 4.1, it is also necessary to define the membership function for each class of each system variable. This is accomplished by determining the landmarks associated to each class value. In these experiments the landmarks of the non-nominal variables were given by the experts (the course advisors).

Table 2. "Number of classes parameter for each feature

AGE	EXP	G	STD	POS	ACT	ASS	MAIL	COEV	F	FCP	FC	IC	ER	BR	MARK
2	7	2	2	15	2	2	3	3	3	3	3	2	2	2	3

The next step was the identification of the model. For the qualitative model identification process to take place, it is necessary to provide the mask candidate matrix. In this experiment, all features were considered. The candidate matrix proposed is of depth one (only one row), forbidding the creation of temporal relations between different samples, i.e. students. With the proposed initial mask, the qualitative model identification process computes the optimal and sub-optimal masks. The root mean square (RMS) error between the predicted output, $\hat{y}$, and the observed system output, y, was used to determine the validity of the model. This error is defined as

$$\mathrm{RMS} = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N}}, \tag{10}$$

where N is the total number of data samples. We are now ready to validate the masks by allowing the fuzzy forecasting process of FIR methodology to predict the seven test data sets. The optimal mask encountered by FIR is

t\x	AGE	EXP	G	STD	POS	ACT	ASS	MAIL	COEV	F	FCP	FC	IC	ER	BR	MARK
T	0	0	0	0	0	0	0	0	–1	0	0	0	–2	–3	0	+1

Its associated quality measure has a value of $Q = 0.66$. This mask is then used to perform the prediction through the sevenfold cross-validation. The 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 RMS errors obtained when using this mask to predict the 7 test data sets previously mentioned are shown in Table 3.

Table 3. RMS prediction errors for the 7 test data sets

Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6	Fold 7	Mean
0.6268	0.5239	0.5395	0.7038	0.4413	0.6743	0.6096	0.5884

The mean error value of 0.5884 shows that the optimal mask identified by FIR is able to capture accurately the students' performance using only the COEV, IC and ER variables. Figure 8 shows the best prediction signal obtained by FIR optimal mask (fold #5). Figure 9 shows the worst prediction signal obtained with the same mask (fold #4). From these figures, it is clear that the predicted signals follow very well the real signals, being able to forecast quite accurately low and high marks.

It is important to comment here that the three variables selected by the FIR modelling process (COEV, IC and ER), represent the 50% of the final mark evaluation (the weighted formula used to compute the final mark of the course is: MARK = 0.05*MAIL + 0.20*COEV + 0.05*F + 0.05*FCP + 0.20 *FC + 0.10*IC + 0.20*ER + 0.15*BR). Notice that there are some variables such as the final class plan (FCP) and work in the branch (BR) that, by themselves, constitute 35% of the final mark, yet they have not been included in the optimal mask. This is a relevant and interesting result, as it suggests that the information included in these variables already exists in the selected ones (COEV, IC and ER). Therefore, these variables are somehow redundant from the point of view of the final mark prediction.

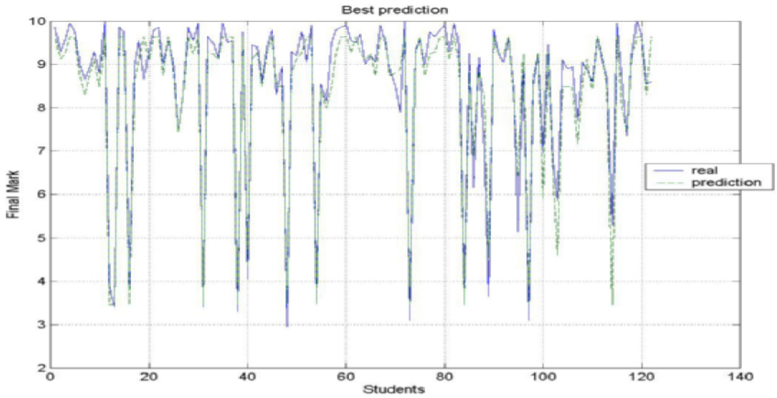


Fig. 8. Real and predicted signals of the MARK variable for fold #5 (RMS = 0.4413)

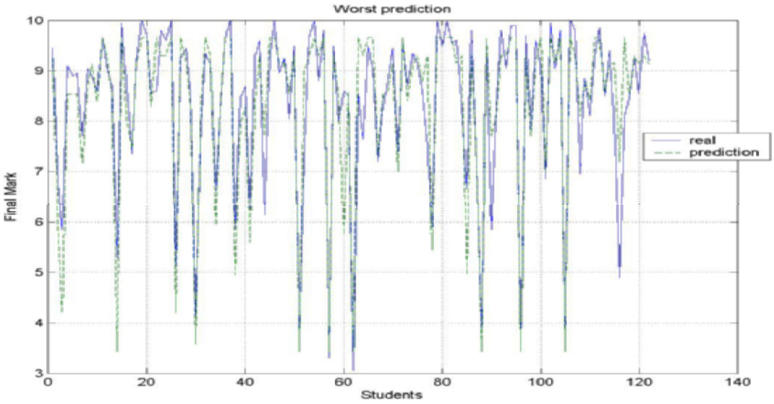


Fig. 9. Real and predicted signals of the MARK variable for fold #4 (RMS = 0.7038)

Another significant result is the selection of the co-evaluation variable as a relevant feature for the prediction of a student final mark. As described before, the advisor grades how well the student evaluates the class plans of his/her course mates. A student that is able to evaluate the work of other people is capable to evaluate correctly his/her own work and, therefore, to execute a good work or a good FCP. Therefore, the information conveyed by this feature is fundamental to predict the final performance of the student in the course. This conclusion has been corroborated by the 31 advisors responsible for the course, and previously published in [13]. It is also worth pointing out that co-evaluation is not a feature commonly used in the e-learning environment. However, its high level of predictive power is manifested in our experiments. Moreover, all the advisors agree that co-evaluation provides valuable leverage to reduce the barriers of distance education.

On the other hand, the experience report (ER) and the initial class plan (IC) also come up as relevant features for the prediction of the students' final mark. The experience report can be viewed as a self-evaluation of his/her own learning process. The average mark of the initial class plan variable is a binary variable that has a value of 0 if the student did not present the initial class plan and a value of 10 if he/she did. Both variables help to predict the peaks of the final mark signal, due to the fact that they contain information not provided by the COEV feature. To some extent, these results can be used by the advisors to adjust the equation that computes the final mark by modifying the weights of COEV and FC variables. At the very least, the data-based feature relevance results would help the course advisors to define a more accurate final mark equation.

It is also interesting to note that the variables that describe the personal attributes of the student, i.e. AGE, EXP, G, STP and POS, are not selected by FIR as relevant to predict the final mark. This makes sense because the work carried out by the students in the course is much more relevant than their personal information. However, we were interested in finding out which of the personal attributes gave us relevant information related to final grades. To this end, a new experiment was carried out in which only the first five features, corresponding to the students' personal attributes, were used in the mask candidate matrix with the main goal of predicting, again, the student final mark. In this case, FIR selected as relevant features the age of the student and his/her area of expertise.

The prediction error (RMS) obtained using this model was 1.893 and the prediction signal was able to follow quite well the minimum and maximum peaks, i.e. the bad and good marks. Although it is not possible to provide a definitive conclusion, it is clear from this experiment that the age of the student, in the first place, and the area of expertise, in the second place, are important personal aspects that influence students' performance. These results agree with some of the conclusions presented in Sect. 3 of this chapter, where GTM, a clustering model, was used to analyse atypical students' behaviour.

4.2 Rule Extraction from Classification Results

The interpretability of the mark prediction results shown in Sect. 4.1.1 would be improved by their description in terms of simple and actionable rules. This is accomplished in this case study through the application of two different methodologies: OSRE, a novel overlapping rule extraction method [4], and, once again, FIR.

4.2.1 OSRE

OSRE: [4] 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 [15]. 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. 10, 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. 10, 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 [3]. 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 [4] has showed OSRE to be an accurate and efficient rule extraction algorithm.

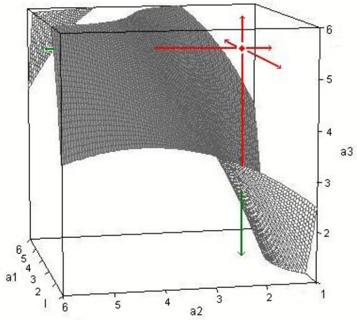


Fig. 10. *Left*: Illustration of orthogonal searching to find decision boundaries; *Right*: Hyper-boxes constructed from the search results

4.2.2 OSRE Results

The approach to the OSRE experiments was twofold: First, all data features from Table 1 were used in the classification task; second, only the three features selected by FIR, as reported in Sect. 4.2, 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: Number of hidden nodes = 8; learning rate = 0.01; momentum = 0.9; weight decay = 0.01; for the models using the FIR selection of three features, all parameters but weight decay = 0.05 were the same. In all cases, the network weights were initialized with random values. Once the number of training epochs that minimized 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 4–6. Each rule is a conjunction of the features and their values.

Interestingly, in the experiments using all features, those selected as the most relevant by FIR, namely COEV, IC, and ER, all 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 three 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. results were validated by educative experts from CECTE [5].

Table 4. 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 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	$(0 \leq \text{ER} \leq 6) \wedge (0 \leq \text{COEV} \leq 6)$	0.99	0.82	0.85	0.99	0.82	0.85
2	$0 \leq \text{FC} \leq 4$	0.99	0.82	0.87	0.98	0.92	0.78
3	$(0 \leq \text{BR} \leq 3) \wedge (0 \leq \text{ER} \leq 3)$	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	$0 \leq \text{COEV} \leq 5$	0.96	0.94	0.66	0.96	0.94	0.66

Table 5. OSRE rules for Class 2 $(5 \le \mathrm{MARK} < 8)$. *Spec*, *Sens* and *PPV* as in table 5. *Top* and *bottom* tables as in table 4

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	$(0 \le \mathrm{FCP} \le 7) \wedge (\mathrm{IC} = 0)$	0.96	0.35	0.64	0.96	0.35	0.64
2	$(63.4 \le \mathrm{ACT} \le 65.5) \wedge (69.7 \le \mathrm{ASS} \le 100) \wedge (0 \le \mathrm{FCP} \le 8) \wedge \mathrm{ER} = 0$	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	$\mathrm{IC} = 0$	0.94	0.43	0.66	0.94	0.43	0.66
2	$(4 \le \mathrm{COEV} \le 10) \wedge (0 \le \mathrm{ER} \le 2)$	0.99	0.33	0.93	0.94	0.66	0.69

4.2.3 Rule Extraction Using FIR

Starting from the description of FIR in Sect. 4.1, we now explain how rule extraction can be implemented as part of this methodology.

Table 6. OSRE rules for Class 3 $(8 \le \mathrm{MARK} < 10)$. *Spec*, *Sens* and *PPV* as in table 3. *Top* and *bottom* tables as in Table 4

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	$(8 \le \mathrm{BR} \le 10) \wedge (3 \le \mathrm{F} \le 10) \wedge (1 \le \mathrm{FCP} \le 10) \wedge \mathrm{IC} = 10 \wedge (7 \le \mathrm{ER} \le 10) \wedge (8 \le \mathrm{COEV} \le 10)$	1	0.86	1	1	0.86	1
2	$(8 \le \mathrm{BR} \le 10) \wedge (1 \le \mathrm{MAIL} \le 10) \wedge (9 \le \mathrm{ER} \le 10) \wedge (9 \le \mathrm{COEV} \le 10)$	1	0.65	1	1	0.91	1
3	$(7 \le \mathrm{BR} \le 10) \wedge (7 \le \mathrm{F} \le 10) \wedge (5 \le \mathrm{MAIL} \le 10) \wedge \mathrm{IC} = 10 \wedge (7 \le \mathrm{FC} \le 10) \wedge (7 \le \mathrm{ER} \le 10) \wedge (5 \le \mathrm{COEV} \le 10)$	1	0.7	1	1	0.94	1
4	$(6 \le \mathrm{FCP} \le 10) \wedge (2 \le \mathrm{MAIL} \le 10) \wedge \mathrm{FC} = 10 \wedge (9 \le \mathrm{ER} \le 10) \wedge (9 \le \mathrm{COEV} \le 10)$	1	0.47	1	1	0.95	1
5	$\mathrm{BR} = 10 \wedge \mathrm{IC} = 10 \wedge \mathrm{FC} = 10 \wedge (8 \le \mathrm{ER} \le 10) \wedge (6 \le \mathrm{COEV} \le 10)$	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	$(9 \le \mathrm{ER} \le 10) \wedge (9 \le \mathrm{COEV} \le 10)$	0.91	0.73	0.96	0.91	0.73	0.96
2	$\mathrm{IC} = 10 \wedge (4 \le \mathrm{ER} \le 9) \wedge (7 \le \mathrm{COEV} \le 10)$	0.91	0.39	0.94	0.82	0.95	0.95
3	$\mathrm{IC} = 10 \wedge (9 \le \mathrm{ER} \le 10) \wedge (5 \le \mathrm{COEV} \le 9)$	0.93	0.35	0.94	0.82	0.99	z

The proposed method is an iterative process that compacts the pattern rule base identified by FIR. On the one hand, we aim to obtain interpretable, realistic and efficient behavioural rules, describing students' learning behaviour. On the other hand, we want to compact the rule base pattern in order to make more parsimonious and, therefore, speed up the prediction process. In order to preserve a model that is congruent with those previously identified by FIR, the proposed algorithm is based on its initial discretization, using only the mask features and the pattern rule base obtained (see Sect. 4.1). Therefore, the rules will maintain the landmarks initially defined.

The model can be summarized as a set of ordered steps:

1. *Basic compactation*. This is an iterative step that evaluates, one at a time, all the rules in a pattern rule base. The pattern rule base, R, is compacted on the basis of the "knowledge" obtained by FIR. A subset of rules R_c can be compacted in the form of a single rule r_c, when all premises P but one (P_a), as well as the consequence C share the same values. Premises, in this context, represent the input features, whereas consequence is the output feature in a rule. If the subset contains all legal values LV_a of P_a, all these rules can be replaced by a single rule, r_c, that has a value of -1 in the premise P_a. When more than one -1 value, P_{ni}, is present in a compacted rule r_c, it is compulsory to evaluate the existence of conflicts by expanding all P_{ni} to all their legal values LV_a, and comparing the resultant rules Xr with the original rules R. If conflicts, Cf, exist, the compacted rule r_c is rejected, and otherwise accepted. In the latter case, the previous subset, R_c is replaced by the compacted one r_c. Conflicts occur when one or more extended rules, Xr have the same values in all its premises, P, but different values in the consequence C.
2. *Improved compactation*. Whereas the previous step only structures the available knowledge and represents it in a more compact form, the improved compactation step extends the knowledge base R to cases that have not been previously used to build the model: R_b. Thus, whereas step 1 leads to a compacted data base that only contains knowledge, the enhanced algorithm contains undisputed knowledge and uncontested belief. Two options are studied: In the first one, using the compacted rule base R' obtained in step 1, all input features P (premises) are visited once more in all the rules r that have non-negative vales (not compactted), and their values are replaced by -1. An expansion to all possible full sets of rules Xr and their comparison with the original rules R are carried out. If no conflicts Cf are found, the compacted rule, r_c , is accepted, and otherwise, rejected. The second option is an extension of the basic compactation, where a consistent and reasonable minimal

ratio, *MR*, of the legal values LV_a should be present in the candidate subset R_c, in order to compact it in the form of a single rule r_c. This latter option seems sensible because, although a reasonable ratio was used to compact R_c in a single rule r_c, the assumed beliefs are minimal and do not compromise the model previously identified by FIR Instead, in option 1, beliefs are assumed to be consistent with the original rules; nevertheless, this could compromise the agreement with model identified, specially when the training data are poor and do not describe well all possible behaviours.

The obtained set of rules is subjected to a number of refinement steps: removal of duplicate rules and conflicting rules; unification of similar rules; evaluation of the obtained rules and removal of rules with low specificity and positive predictive value (PPV: see Sect. 4.2.2).

4.2.4 Experimental Rule Extraction Results Using FIR

The main goal of the rule extraction algorithm described is to endow FIR with a method to describe the analysed system using logical rules that are more comprehensive, readable, and which provide explanations (not only assumptions) that may be validated by domain experts, increasing confidence in the analysis. The experimental results obtained using the rule extraction algorithm described in the previous section are presented in Table 7.

We can see that the specificity and the positive predicted value reaches reasonable values for most rules separately, as well as for the whole set of rules. However, the sensitivity is quite low throughout. Only the sensitivity of rule 7 (highlighted in Table 7) is reasonably high, describing a very common pattern in the analysed data. The best results in this experiment correspond to the second option of the Improved Compactation method.

Although both of the rule extraction methods used in this section resort to the same evaluation metrics, they differ in the way results are presented. Whereas OSRE provides cumulative results for each class, the FIR experimental rule extraction method provides results for all classes together, for each individual rule and for the whole set of rules.

Notwithstanding the inherent difficulty of comparison, the rule extraction results obtained by the experimental FIR extension are, at least globally, not too different from those obtained by OSRE when only the features selected by FIR are used.

FIR obtains a set of seven rules based on its own feature selection. Looking more closely to the rules obtained by FIR and OSRE (with FIR feature selection), it seems that OSRE has a higher compactation capacity, at least for class 1. OSRE models class 1 by a unique rule with a single premise, whereas FIR needs two rules, with two and three premises, respectively. Comparing the set of rules obtained with both methodologies using specificity, sensitivity and positive predictive value metrics, it can be seen that for class 1, FIR rules have a higher specificity and PPV, whereas OSRE rule has a better sensitivity. The three metrics have similar values for both OSRE and FIR rules defining class 2. On the contrary, the OSRE rules obtain higher values in all the metrics for class 3.

The learning behaviour rules obtained by both algorithms were analysed and validated by educative experts of CECTE. They agreed that the obtained results were intuitive, realistic, and mostly consistent with their own perception of the CECTE course students' learning behaviour.

Table 7. Experimental rule extraction results using FIR for both training and test data sets. *Spec* stands for Specificity; *Sens* for Sensitivity; and *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

RULE	Out Class	TRAIN			TEST		
		Spec	Sens	PPV	Spec	Sens	PPV
IF 0<=IC<=5.1 AND 4.9<=COEV<=10 THEN 4.9<=MARK<= 7.9	2	0.96	0.38	0.7	0.95	0.37	0.71
IF 5.1<=IC<=10 AND 0<=ER<= 8.1 THEN 0<=MARK<= 7.9	1–2	0.84	0.51	0.57	0.84	0.43	0.55
IF 0<=COEV<=4.9 AND 8.1<=ER<=10 THEN 0<=MARK<= 4.9	1	1	0.11	0.78	0.97	0.08	0.2
IF 0<=COEV<= 4.9 AND 0<=IC<=5.1 AND 0<=ER<=8.1 THEN 0<=MARK<=4.9	1	1	0.07	1	1	0.08	1
IF 4.9<=COEV<= 7.9 AND 5.1<=IC<=10 THEN 7.9<=MARK<=10	3	0.95	0.03	0.58	0.87	0.04	0.42
IF 7.9<=COEV<=10 AND 8.1<=ER<=10 THEN 7.9<=MARK<=10	**3**	**0.75**	**0.81**	**0.88**	**0.83**	**0.81**	**0.91**
IF 7.9<=COEV<=10 AND 5.1<=IC<=10 AND 0<=ER<=8.1 THEN 7.9<=MARK<=10	**3**	0.78	0.16	0.63	0.8	0.15	0.62
TOTAL RULES		**0.93**	**0.34**	**0.76**	**0.92**	**0.33**	**0.74**

5 Conclusions

The possibility of tracking user behaviour in virtual campus e-learning environments makes possible the mining of the resulting data bases. This opens new possibilities for the pedagogical and instructional designers who create and organize the learning contents.

The presence of outliers in a data set can distort the results obtained from its analysis. Therefore, the data analyst should benefit from models that behave robustly in their presence. One such model, the *t*-GTM, has been introduced. It simultaneously provides robust data clustering and visualization of the results, which become intuitively interpretable. It also neutralizes the negative effects of outliers. Moreover, this model provides a method to assess the unsupervised relative relevance of the data features. Data from the CECTE virtual campus, corresponding to the students of a "Didactic Planning" graduate course, have been analysed using the proposed model. Students with atypical online behaviours (outliers) have been identified and characterized, using the extension of *t*-GTM for FRD. The results have shown that useful knowledge can be extracted from the *t*-GTM combination of outlier detection, FRD and data clustering and visualization. This knowledge could be fed back into the e-learning system in order to provide students with personalized guidance, tailored to their inhomogeneous needs and requirements. As a software tool embedded in the e-learning system, it would also help teachers to find patterns of student's behaviour.

In this case study, we have also addressed the problem of students' marks prediction, using the FIR methodology. The characterization of the students' online behaviour would benefit from a method capable of determining the relevance of the features involved in the analysed data set in terms of this prediction. One such method is also provided by FIR. The experimental results have shown that FIR was able to identify a good model to predict students' final marks and to determine the relevant features involved in the evaluation process. This knowledge could be used for real time student personalization guidance, and to help teachers in finding patterns of student behaviour. For this knowledge to have an intuitive and useful form, results have been described in terms of rules. The novel OSRE methodology and an extension of FIR have been applied here to obtain simple sets of rules describing the diverse levels of the students' performance.

6 Acknowledgments

Alfredo Vellido is a research fellow within the Ramón y Cajal program of the Spanish Ministry of Education and Science. Félix Castro is a research fellow within the PROMEP program of the Mexican Secretary of Public Education.

7 References

1. Bishop, C.M., Svensén, M., Williams, C.K.I.: GTM: The Generative Topographic Mapping. Neural Computation 10(1) (1998) 215–234
2. Dempster, A.P., Laird, M.N., Rubin, D.B.: Maximum Likelihood from Incomplete Data Via the EM Algorithm. Journal of the Royal Statistical Society 39(1) (1977) 1–38
3. Etchells, T.A., Jarman, I.H., Lisboa, P.J.G.: Empirically Derived Rules for Adjuvant Chemotherapy in Breast Cancer Treatment. Proc. of the Advances in Medical Signal and Information Processing International Conference, MEDSIP 2004. 5–8 September, Malta (2004) 345–351
4. Etchells, T.A., Lisboa, P.J.G.: Orthogonal Search-based Rule Extraction (OSRE) Method for Trained Neural Networks: A Practical and Efficient Approach. IEEE Transactions on Neural Networks 17(2) (2006) 374–384
5. Etchells, T.A., Nebot, A., Vellido, A., Lisboa, P.J.G., Mugica, F.: Learning What is Important: Feature Selection and Rule Extraction in a Virtual Course. In: The 14th European Symposium on Artificial Neural Networks, ESANN 2006. Bruges, Belgium (2006) 401–406
6. Jerez, A., Nebot, A.: Genetic Algorithms versus Classical Search Techniques for Identification of Fuzzy Models. Proc. of the 5th European Congress on Intelligent Techniques and Soft Computing, EUFIT'97. Aachen, Germany (1997) 769–773
7. Klir, G.: Architecture of Systems Problem Solving. Plenum Press. New York (1985)
8. Kohonen, T.: Self-Organizing Maps. 3rd edition, Springer, Berlin Heidelberg New York (2000)

9 Law, A., Kelton, D.: Simulation Modeling and Analysis. 2nd edition., McGraw-Hill, New York (1990)

10. Law, M.H.C., Figueiredo, M.A.T., Jain, A.K.: Simultaneous Feature Selection and Clustering Using Mixture Models. IEEE Transactions on Pattern Analysis and Machine Intelligence 26(9) (2004) 1154–1166
11. Nebot, A.: Qualitative Modeling and Simulation of Biomedical Systems Using Fuzzy Inductive Reasoning. Ph.D. thesis, Dept. Llenguatges i Sistemes Informàtics, Universitat Politècnica de Catalunya. Barcelona, Spain (1994)
12. Nebot, A., Cellier, F., Vallerdú, M.: Mixed Quantitative/Qualitative Modeling and Simulation of the Cardiovascular System. Computers Methods and Programs in Biomedicine 55 (1998) 127–155
13. Nebot, A., Castro, F., Vellido, A., Mugica, F.: Identification of Fuzzy Models to Predict Students Performance in an e-Learning Environment. In: Uskov, V. (ed.): The Fifth IASTED International Conference on Web-Based Education, WBE 2006. Puerto Vallarta, Mexico (2006) 74–79
14. Peel, D., McLachlan, G.J.: Robust Mixture Modelling Using the *t*-Distribution. Statistics and Computing 10 (2000) 339–348
15. Tsukimoto, H.: Extracting Rules from Trained Neural Networks. IEEE Transactions on Neural Networks 11(2) (2000) 377–389

16. Vellido, A.: Assessment of an Unsupervised Feature Selection Method for Generative Topographic Mapping. 16th International Conference on Artificial Neural Networks, ICANN 2006. Lecture Notes in Computer Science, Vol. 4132. Springer, Berlin Heidelberg New York (2006) 361–370
17. Vellido, A.: Missing Data Imputation through GTM as a Mixture of *t*-Distributions. Neural Networks, In Press (2006)
18. Vellido, A., Lisboa, P.J.G., Vicente, D.: Robust Analysis of MRS Brain Tumour Data Using *t*-GTM. Neurocomputing 69(7–9) (2006) 754–768
19. Vellido, A., Castro, F., Nebot, A., Mugica, F.: Characterization of Atypical Virtual Campus Usage Behavior Through Robust Generative Relevance Analysis. In: Uskov, V. (ed.): The 5th IASTED International Conference on Web-Based Education, WBE 2006. Puerto Vallarta, Mexico (2006) 183–188