STUDIA UNIV. BABEŞ–BOLYAI, INFORMATICA, Volume **LXIV**, Number 2, 2019
DOI: 10.24193/subbi.2019.2.03

ANALYSING THE ACADEMIC PERFORMANCE OF STUDENTS USING UNSUPERVISED DATA MINING

GEORGE CIUBOTARIU AND LIANA MARIA CRIVEI

Abstract. *Educational Data Mining* is an attractive interdisciplinary domain in which the main goal is to apply data mining techniques in educational environments in order to offer better insights into the educational related tasks. This paper analyses the relevance of two *unsupervised learning* models, *self-organizing maps* and *relational association rule mining* in the context of students' performance prediction. The experimental results obtained by applying the aforementioned unsupervised learning models on a real data set collected from Babeş-Bolyai University emphasize their effectiveness in mining relevant relationships and rules from educational data which may be useful for predicting the academic performance of students.

1. Introduction

Extracting relevant patterns from the educational processes is the main topic in the *Educational data mining* (EDM) field, as it could provide effective methods for understanding students and their learning process and, subsequently, improving the learning outcomes. EDM has lately been under great consideration from the research community since extracting hidden valuable knowledge from educational data is of major interest for the academic institutions and also effective for reviewing and improving their teaching techniques and learning procedures [13].

There is a continuous interest in applying *machine learning* (ML) techniques in the educational field [3]. Within the ML domain, *unsupervised learning* (UL) techniques are broadly applied nowadays in numerous domains including software engineering, medicine, bioinformatics, the financial domain, in order to extract relevant hidden knowledge from the data in the form of rules

Received by the editors: October 28, 2019.

2010 *Mathematics Subject Classification.* 68T05, 68P15.

1998 *CR Categories and Descriptors.* H.2.8[**Database management**]: Database Applications – *Data Mining*; I.2.6[**Computing Methodologies**]: Artificial Intelligence – *Learning*;

Key words and phrases. Educational data mining, Unsupervised learning, Self-organizing map, Relational association rule.

or patterns. Diverse applications using data mining and machine learning algorithms have been implemented, so far, in the EDM domain. Machine learning methods are applied, both from a *supervised* and *unsupervised* perspective, as data mining techniques for developing systems for course planning, predicting the students' progress and academic performance, detecting students' learning type, grouping students in similar classes, supporting instructors in the educational process [8].

The study performed in this paper is aimed to highlight the potential of applying two UL techniques (*self-organizing maps* (SOMs) and *relational association rule mining*) (RARs) in analysing students' academic performance. The main research question we are investigating in this paper is regarding the ability of unsupervised learning models (SOMs and RARs) to detect hidden relationships between the grades received by the students during the semester and their final examination grade category at a certain academic discipline. In addition, we aim to analyse if the unsupervised learning models may reveal some information regarding the quality of the educational processes.

A study on the EDM literature reveals various approaches using unsupervised learning for mining student data in educational environments. Various clustering algorithms, including partitional and hierarchical clustering were applied by Ayers et al. [2] for determining groups of students who have simlar skills. Dutt et al. [6] present a survey on applying *unsupervised learning* techniques for various tasks from the educational setting. *K-means* clustering is applied by Parack et al. [14] for grouping students according to their learning patterns. The identified groups are then used for determining the cognitive styles for each cluster. SOMs were used by Kurdthongmee [11] to group students in clusters according to their academic results. Khadir et al. [10] performed a study based on clustering and SOMs for students' academic performance prediction. Saxena et al. [15] have also applied SOMs for classifying students in categories according to their academic performance. To the best of our knowledge, a study similar to ours has not been performed in the literature.

The rest of the paper is organized as follows. Section 2 presents the *self-organizing maps* and *relational association rule mining* models used in our study. Section 3 introduces our experimental methodology, while Section 4 discusses about the experimental results. The conclusions of our study together with several directions for future research are summarized in Section 5.

2. Unsupervised machine learning models used

Unsupervised learning models are known in the ML literature as *descriptive* models, due to their ability to detect how data are organized. Unsupervised learning algorithms receive only unlabeled examples and learn to detect hidden patterns from the input data based on their features. UL methods are useful for discovering the underlying structure of the data.

2.1. Relational association rule mining.

Relational association rules (RARs) [4, 16] were introduced in the data mining literature as an extension of the classical *association rules* with the goal of uncovering various types of relationships, both ordinal and non-ordinal, between data attributes.

The concept of *RARs* will be further presented. We consider $\mathcal{O} = \{o_1, o_2, \ldots, o_n\}$ a set of *instances* (objects) and $\mathcal{A} = (at_1, \ldots, at_m)$ a sequence of relevant attributes characterizing the instances from $\mathcal{O}$. Each attribute at_i takes values from a non-empty domain D_i. The value of attribute at_i for instance o_j is denoted by $\eta(o_j, at_i)$ and by $\mathcal{T}$ is denoted the set of all possible relations which can be defined between the domains D_i and D_j.

Definition 1. *A* relational association rule [16] *is an expression* $(at_{i_1}, at_{i_2}, at_{i_3}, \ldots, at_{i_h}) \Rightarrow (at_{i_1} \tau_1\ at_{i_2} \tau_2\ at_{i_3} \ldots \tau_{h-1}\ at_{i_h})$, *where* $\{at_{i_1}, at_{i_2}, at_{i_3}, \ldots, at_{i_h}\} \subseteq \mathcal{A}$, $at_{i_k} \neq at_{i_p}$, $k, p = 1, \ldots, h$, $k \neq p$ *and* $\tau_k \in \mathcal{T}$ *is a relation over* $D_{i_k} \times D_{i_{k+1}}$, D_{i_k} *representing the domain of the attribute* at_{i_k}.

a) *If* $at_{i_1}, at_{i_2}, \ldots, at_{i_h}$ *are non-missing in* l *instances from* $\mathcal{O}$ *then we call* $s = \frac{l}{n}$ *the* support *of the rule.*

b) *If we denote by* $\mathcal{O}' \subseteq \mathcal{O}$ *the set of instances where* $at_{i_1}, at_{i_2}, at_{i_3}, \ldots, at_{i_h}$ *are non-missing and all the relations* $\eta(o_j, at_{i_1})\ \tau_1\ \eta(o_j, at_{i_2})$, $\eta(o_j, at_{i_2})\ \tau_2\ \eta(o_j, at_{i_3})$, $\ldots$, $\eta(o_j, at_{i_{h-1}})\ \tau_{h-1}\ \eta(o_j, at_{i_h})$ *hold for each instance* $o_j \in \mathcal{O}'$ *then we call* $c = \frac{|\mathcal{O}'|}{n}$ *the* confidence *of the rule.*

The notion of *interestingness* was introduced in [16] as the property of RARs to have their *support* and *confidence* greater than or equal to certain thresholds. The algorithm *DRAR* (*Discovery of Relational Association Rules*) for uncovering interesting RARs was introduced in [5]. *DRAR* is an Apriori-like algorithm consisting of a RAR generation process that starts from the 2-length rules which are filtered such that to preserve only the interesting rules. The iterative process continues with 3-length rules, then with 4-length rules and so on. At a certain iteration, the RARs of length n are generated by joining [16] interesting RARs of length n-1. The obtained set is then filtered for preserving only the interesting n-length rules. When no new interesting RARs are identified at a certain iteration, the generation process stops.

2.2. **Self-organizing maps.** *Self-organizing maps* (SOMs), also known in the literature as *Kohonen maps*, were introduced by Teuvo Kohonen and are *unsupervised learning* models widely used as tools for visualizing high-dimensional data. SOMs are connected to the *artificial neural networks* (ANNs) in literature and to *competitive learning*. In *competitive learning*, the output neurons compete themselves to be activated. A *self-organizing map* [17] is trained using an unsupervised learning algorithm (Kohonen's algorithm) to map, using a non-linear mapping, the continuous input space of high-dimensional instances into a discrete (usually two-dimensional) output space called a *map* [7]. The *topology preserving mapping* is the main characteristic of the mapping provided by a SOM. This means that the input samples which are close to each other in the input space will be also close to each other on the map (output space). Thus, a SOM is able to provide clusters of similar data instances [12].

3. Methodology

As previously stated, our study aims at investigating the relevance of unsupervised SOM and RAR models in analysing the academic performance of students.

We are introducing the following theoretical model. We denote by $\mathcal{S}tud = \{stud_1, stud_2, \ldots, stud_n\}$ a data set in which the instances $stud_i$ describe the performance of students during an academic semester, at a given academic discipline $\mathcal{D}$. Each instance $stud_i$ is characterized by a set of *grades* received during the semester $\mathcal{G} = \{g_1, g_2, \ldots, g_m\}$ representing attributes for measuring the performance of the student for the given discipline. Thus, each $stud_i$ is represented as an m-dimensional vector $stud_i = (stud_{i1}, stud_{i2}, \ldots, stud_{im})$, $stud_{ij}$ representing the value of attribute g_j for student $stud_i$.

The goal of the current study is to investigate if two unsupervised data mining models, *self-organizing maps* and *relational association rule mining*, are able to discover some rules and relationships which would be useful for predicting the *final performance* for the students, based on their grades obtained during the academic semester. Since predicting the exact final examination grade for a student is a difficult task, considering the uncertainty in the learning and evaluation processes, we are considering in this paper four categories of final grades: (1) *excellent* (denoted by E and representing the final grades 9 and 10); (2) *good* (denoted by G and representing the final grades 7 and 8); (3) *satisfactory* (denoted by S and representing the final grades 5 and 6); and (4) *fail* (denoted by F and representing the final grade 4). Let us denote by $\mathcal{C} = \{E, G, S, F\}$.

We note that our unsupervised analysis does not include the grades of the students' at the written exam (obtained in the examination session), which

are also part of their final examination grade. Thus, we aim to analyse if only the grades received by the students during the semester are enough to discriminate their written exam grade and, accordingly, the students' final examination grade category.

The problem investigated in this paper, from an *unsupervised learning* perspective, is that of assigning to each student (characterized by its grades received during an academic semester) the category corresponding to its final grade. This assignment may be formalized by a mapping $f : \mathbb{R}^m \to \mathcal{C}$.

3.1. **Data set.** In our experiment we are considering a real data set [1], denoted by D, containing the grades obtained by students at a Computer Science undergraduate course offered at Babeş-Bolyai University collected during six academic years (2011-2017) at the regular and retake examination sessions. The data set consists of 905 students characterized by 4 attributes, denoted by a_1, a_2, a_3, a_4. Attributes a_1, a_2, a_3, a_4 represent scores obtained by the students at several evaluations during the academic semester: seminar score (a_1), project score (a_2), project status score (a_3) and written test score (a_4). For each student $s \in D$, its *final examination grade* ($f(s)$), received at the end of the academic semester after the final examination is known. In our experiments, the final examination grade of a student is transformed into a category (E, G, S, F), as previously shown. However, in our unsupervised learning based experiments, the students' final grade will be used only for evaluating the learning performance, without using it for building the SOM and RAR models. Figure 1 depicts a histogram of grades (4-10) built on the data set D.

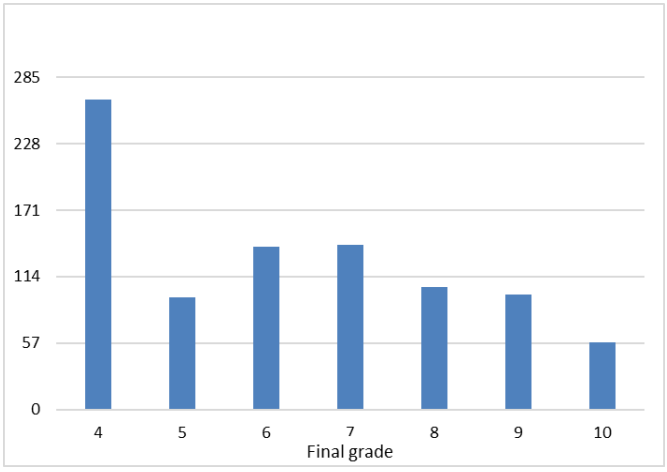


FIGURE 1. Histogram of grades from the data set D.

The histogram from Figure 1 reveals a distribution of passing grades (grades without 4) close to the normal one. For analysing how correlated are the attributes a_1, a_2, a_3, a_4 with the target output (category corresponding to the final examination grade), the Pearson correlation coefficients are computed. The correlation values are shown in Table 1.

a_1	a_2	a_3	a_4
0.631	0.676	0.461	0.607

TABLE 1. The Pearson correlation coefficient between attributes a_1, a_2, a_3, a_4 and the target output.

From Table 1 we observe that there is a good enough correlation between the attributes a_1, a_2, a_4 and the category corresponding to the final examination grade. The *project score* (attribute a_2) shows the maximum correlation with the final category. The smallest correlation is observed for attribute a_3.

3.2. **Experiments.** The experiments described in this section are aimed to test the ability of SOMs and RARs, as unsupervised learning models, to detect relevant relationships in the students' grades (received during the semester) which are well correlated with their final grade category. For a certain grade category (class) $c \in \{E, G, S, F\}$ we denote by $D_c \subset D$ the subset of students from D whose final grade category is c. We note that $\bigcup_{c \in \mathcal{C}} D_c = D$.

The *first experiment* is conducted for obtaining, using a SOM, a two dimensional representation of the data set. Two SOM visualizations will be provided: one for the entire data set (characterized by all attributes a_1, a_2, a_3, a_4) and the second for the data set without attribute a_3 (i.e the data set characterized only by the attributes a_1, a_2, a_4). After the SOM was unsupervisedly built, the U-Matrix method [9] will be used for visualization. For the SOM, a torus topology is used, with the following parameters: 200000 training epochs and a learning rate of 0.1. The *Euclidian distance* is used as a distance metric between the input instances.

The goal of our *second experiment* is to uncover in each subset D_c, using the $DRAR$ algorithm, a set RAR_c of interesting RARs. We aim to verify the hypothesis that the sets RAR_c are able to discriminate between the classes of students having different final grades.

For the RAR mining experiment, five additional attributes were added to the data sets D_c ($a_i = i, \forall i, 5 \leq i \leq 9$) and we used the following parameters for the mining process: 1 for the minimum support threshold, 0.6 for the minimum confidence threshold and two possible binary relations between the

attributes ($<$ and $\geq$). Attributes a_5, a_6, a_7, a_8, a_9 represent some thresholds for the grades (i.e. $5, 6, 7, 8, 9$) and were added with the goal of enlarging the set of uncovered RARs, allowing the discovery of binary RARs such as $a_i \leq 5$.

For evaluating how well the set RAR_c characterizes the set of students from the set D_c we use the average confidence of all the subrules from the rules from RAR_c, denoted by $Prec_c = \frac{\sum_{r \in RAR_c} \sum_{sr \in \mathcal{S}_r} conf(sr)}{|RAR_c|}$. By $conf(r)$ we denote the confidence of the rule r in the data set D_c and $\mathcal{S}_r$ represents the set of subrules of r (including itself). We note that $Prec_c$ ranges from 0 to 1, higher values for $Prec_c$ indicating that the set RAR_c better characterizes the data set D_c.

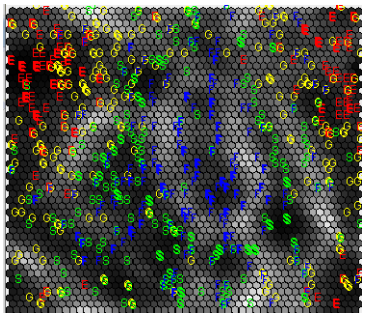
4. Results and discussion

This section presents the experimental results obtained following the experimental methodology introduced in Section 3.2 and discusses about the patterns unsupervisedly discovered using the SOM and RAR models.

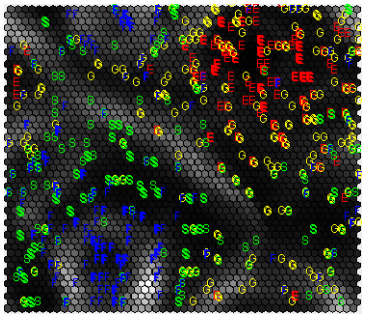
4.1. **Experiments using SOMs.** The left hand side image from Figure 2 illustrates the SOM obtained on the data set from Section 3.1 using attributes a_1, a_2, a_3, a_4, while the right hand side image from Figure 2 depicts the SOM trained on the instances characterized only by attributes a_1, a_2, a_4. On both images, the students with the same final class (final grade category which is depicted on the map) are marked with the same colour: red for the E labels, yellow for G, green for S and blue for F. As expected, Figure 2a depicts a good enough mapping, but still there is no clear separation between the grades. It seems that a slightly better mapping and separation between the grades is provided by the map from Figure 2b when attribute a_3 (the project status score) has not been considered.

The SOMs from Figure 2 reveal the difficulty of the task for predicting the final examination grade category for the students, based on the grades they received during the semester. However, the unsupervisedly built SOMs are able to uncover some patterns regarding the students' final grade category. We observe two main areas on both maps, a cluster of students with the final categories F and S, which is well enough delimited and one containing the categories G and E. Inside the first cluster, we observe a well distinguishable subclass containing students with the final category F.

For evaluating the quality of the SOMs built, the *average quantization error* (AQE) is computed during the unsupervised training process. The AQE [18] is computed by averaging the mean Euclidean distance between the input samples and their best-matching units. Figure 3 comparatively illustrates how AQE decreases during the training of the maps built for the entire data set



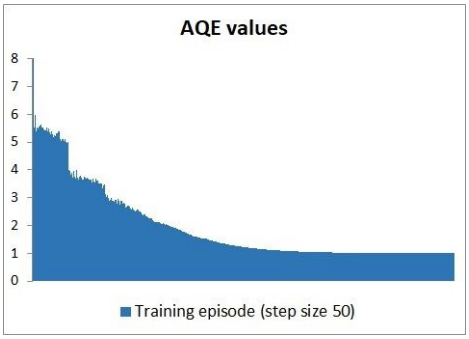
(A) SOM visualization considering attributes a_1, a_2, a_3, a_4.



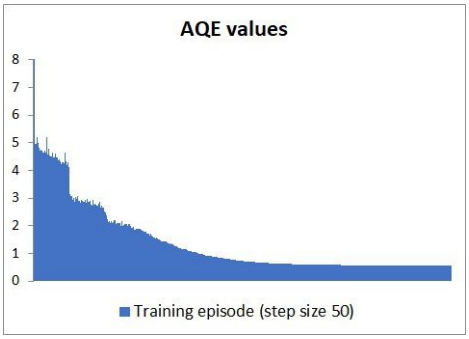
(B) SOM visualization considering attributes a_1, a_2, a_4.

FIGURE 2. U-Matrix visualization of the SOM built on the data set D using attributes a_1, a_2, a_3, a_4 (left) and a_1, a_2, a_4 (right).

(left side image) and for the data set without attribute a_3 (right side image). We note that the AQE reached after the training was completed is 0.997 for the SOM from Figure 3a and 0.559 for the SOM from Figure 3b. The final AQEs which are close to 0 confirm the accuracy of the trained SOMs. In addition, the SOM built on the data set without attribute a_3 has the smallest AQE, indicating a better mapping.



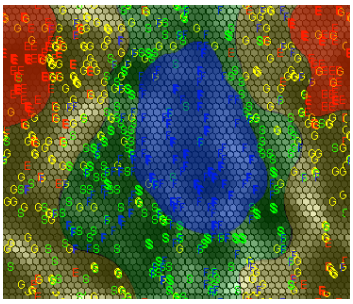
(A) AQE visualization for the SOM built using attributes a_1, a_2, a_3, a_4.



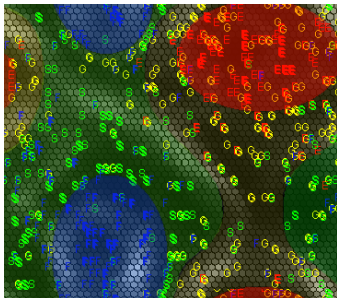
(B) AQE visualization for the SOM built using attributes a_1, a_2, a_4.

FIGURE 3. Evolution of AQE values during training the SOMs built on the data set D using attributes a_1, a_2, a_3, a_4 (left) and a_1, a_2, a_4 (right).

Figure 4 illustrates a detailed visualisation of the SOMs from Figure 2, considering the torus topology used for building the SOMs. The left side image from Figure 4 corresponds to the SOM from Figure 2a, while Figure 4a corresponds to the visualization from Figure 2b. On both SOMs, the distinguishable classes of students are highlighted and coloured according to their class label (E - red, G - yellow, S - green, F - blue).



(A) Detailed visualisation of the SOM from Figure 2a.



(B) Detailed visualisation of the SOM from Figure 2b.

FIGURE 4. Detailed visualisation of the SOMs from Figure 2.

A comparative analysis of the two images from Figure 4 and the highlighted areas reveal the following. In Figure 4a we observe that the F labeled cluster is well disgtinguishable. However, there are a few outliers that go beyond the cluster's border, entering in the S class zone. Moreover, the E labeled cluster also interferes with the outer regions and creates noisy zones on the map. The flaw with the image from Figure 4a is that, even though the centres of the E and F labeled groups are compact and solid, the margins of each one tend to be more fuzzy, exchanging different grades with neighbouring regions. As we can see, there are overlapping grades, especially regarding the F labels, that incline towards a defiance of the boundaries, which suggests that the unsupervised classification model is not capable of clearly discriminating all the students and, consequently, misinterprets some of the patterns of their performance. Regardless these misclassifications that may be due to the small number of attributes characterizing the instances and the presence of outliers, the SOM model is confident enough to make correct prediction most of the time. In Figure 4b there are two contrasting, well separated, areas of high/low grades with a sharp gap between them. The regions corresponding to the average grades surround tightly these two clusters. Few exceptions still occur when separating the grades.

On the other hand, the SOM from Figure 4b seems to provide a better classification of the data. The two opposed areas (of students with high/low grades) now are more compact and clearly separated. Their margins tend to be smoother, especially for those labeled F that are now more compact than in Figure 4a, with less perturbations from the other grades, as a virtual median barrier keeps them apart. However, the class of average grades is still difficult to be separated, and, there has been a trade-off between size and accuracy, since the higher grades are now more compact, but less separated from each other.Nonetheless, if we would combine the two previous interpretations, we may analyse and classify the data better by using both of their strengths, as in each model the data is more or less scattered across the map, which would be of use in cases when we desire a greater confidence on a particular class of grades. While the model from Figure 4a may offer us a better understanding of the students with passing grades (as they belong to a rather compact group), the SOM from Figure 4b may show us an antithetical approach of the highest grades and the lowest ones. Moreover, the SOM model built without using attribute a_3 is particularly good at classifying lower grades with greater accuracy, and, even if there still is noise in the data categorised as E or G, the model can confidently predict the performance of a good student. What makes it so difficult to classify all the students is the fact that there is a discrepancy mainly among the F labeled class, as there is an inconsistent progress for each one, that would result in a more scattered pattern that interferes with better classified data.

Analysing both maps from Figure 2 we also note that most of the students belonging to category F (i.e. having the final grade 4) are, on both maps, well enough delimited from the students from other categories (S, G, E). Overall, we observe as a main pattern that neighboring students belong to near categories (F-S, G-E). But several outliers may be observed on the map: neighboring students having very different categories (e.g. E and F). A possible explanation for such incorrect mappings may be that the data set includes the examination results not only for the normal session, but also for the retake session. Thus, it is very likely to have the same instance but with different final labels (i.e. the categories from the normal and retake session) which may be very different (e.g. F and S). Besides the previously mentioned cause for outliers, another possible one is given by the intrinsic uncertainty of the educational processes. The data set includes instances for which there is a visible uncorrelation between the grades received during the semester and the final examination grade. Such discordances may appear due to a bias evaluation or some unexpected events in the students learning process. To avoid introducing noise in the data set which will affect the performance of

the learning, we will further investigate preprocessing techniques for detecting such outliers.

4.2. **Experiments using RARs.** The results of the RAR mining experiment are further presented, with the aim of highlighting the relevance of the relational association rules mining process in distinguishing between classes of students having different final grade category. For each category $c \in \mathcal{C}$ we present in Figure 5 the set RAR_c of maximal RARs mined from D_c using the experimental methodology from Section 3.2. In addition, we indicate the value of $Prec_c$ which evaluates the quality of the set RAR_c.

c	Length	Rule	Confidence	Prec
E	2	$a_1 > 8$	0.898	0.716
	2	$a_2 > 9$	0.713	
	2	$a_4 > 8$	0.771	
	3	$a_1 \leq a_2 > 8$	0.707	
	3	$a_1 \leq a_3 > 8$	0.650	
	3	$a_1 \leq a_4 > 7$	0.643	
	3	$a_2 \leq a_3 > 9$	0.618	
G	2	$a_1 > 5$	0.87	0.689
	2	$a_1 \leq 8$	0.675	
	2	$a_2 > 7$	0.728	
	2	$a_2 \leq 9$	0.679	
	2	$a_3 > 7$	0.683	
	2	$a_3 \leq 9$	0.630	
	2	$a_4 > 6$	0.630	
	2	$a_4 \leq 8$	0.626	
	3	$a_1 \leq a_2 > 6$	0.614	
	3	$a_1 \leq a_3 > 6$	0.606	

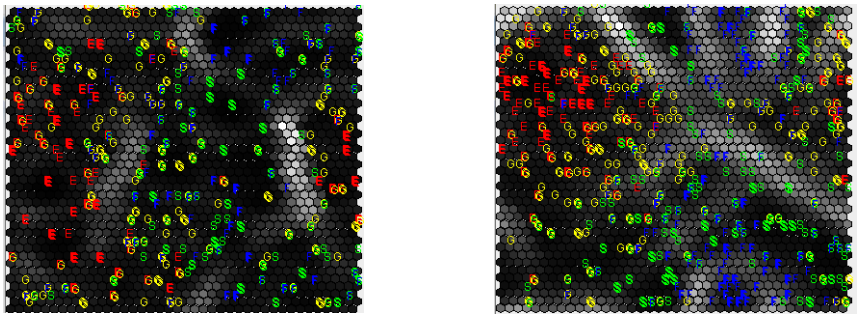
c	Length	Rule	Confidence	Prec
S	2	$a_1 \leq a_3$	0.686	0.69
	2	$a_1 \leq 6$	0.737	
	2	$a_2 \leq 6$	0.661	
	2	$a_3 > 5$	0.623	
	2	$a_3 \leq 7$	0.619	
	3	$a_1 \leq a_2 \leq 9$	0.640	
	3	$a_1 \leq a_3 > a_4$	0.636	
	3	$a_1 > a_4 \leq 5$	0.653	
	3	$a_2 > a_4 \leq 5$	0.640	
	3	$a_3 > a_4 \leq 7$	0.631	
	4	$a_1 \leq a_2 > a_4 \leq 7$	0.619	
F	2	$a_1 \leq a_3$	0.643	0.695
	2	$a_1 \leq 5$	0.624	
	2	$a_2 \leq a_3$	0.654	
	2	$a_2 \leq 5$	0.680	
	2	$a_3 \leq 6$	0.661	
	3	$a_1 > a_4 \leq 5$	0.628	
	3	$a_3 > a_4 \leq 8$	0.617	

FIGURE 5. The sets of maximal interesting RARs mined for each category of grades: E, G, S, F.

From Figure 5 we observe that the sets of RARs characterizing the classes of students with different final grade category are disjoint, in general. For example, the fourth RAR of length three from the left side table indicate that for 61.8% of the students who have received a final examination grade of 9 or 10 (category E), the project score is less than or equal to project status score, which is greater than 9. We note that this RAR does not characterize the other categories of students, thus it is very likely to be useful for discriminating students according to their final category.

For facilitating the interpretation of the RARs, we decided to build a SOM for having a visual representation of the rules and highlighting how well they characterize the classes of students. Let us denote by *Rules* the sequence of all distinct mined RARs given in Figure 5, including all their subrules. If a RAR

of a certain length appears in more than one category, it will appear in *Rules* only once (e.g. the RAR $a_1 \leq a_3$ of length 2 appear as an interesting rule for both categories S and F). Thus, an additional data set D_{RAR} is created by characterizing each student $stud_i$ from the original data set D by a 48-length binary vector $V_i = (v_1^i, v_2^i, \ldots, v_{35}^i)$, where 48 is the size of the sequence *Rules*, i.e the number of distinct RARs mined. An element v_j^i from V_i is set to 1 if the j-th RAR from *Rules* is verified in $stud_i$ and 0 otherwise. Figure 6 depicts the SOMs built on the data set D_{RAR} using all attributes a_1, a_2, a_3, a_4 (left) and using only attributes a_1, a_2, a_4 (right). On each SOMs, the instances are labeled with their final grade category (E, G, S, F).



(A) SOM visualization considering attributes a_1, a_2, a_3, a_4.

(B) SOM visualization considering attributes a_1, a_2, a_4.

FIGURE 6. U-Matrix visualization of the SOM built on the data set D_{RAR} using attributes a_1, a_2, a_3, a_4 (left) and a_1, a_2, a_4 (right).

From the interesting RARs depicted in Figure 5 and visualized in Figure 6a we also observe that there is an overlapping, in general, between the set of RARs characterizing near categories (F/S, G/E). This is expectable, as previously shown in Section 4.1 where the SOM mapping highlighted that there are instances, mainly from near categories, that are hard to discriminate. For instance, the rule $a_1 \leq a_3$ appears for both S and F categories with highly similar confidences (0.686 for S and 0.643 for F). Another example is the rule $a_2 > 7$ from G and $a_2 > 9$ from E which is also explainable due to some instances that are on the border between the two categories. Certainly, such overlapping rules are not useful for discriminating between classes. A post-processing step would be useful for detecting and removing such rules from the mining process and will be further investigated. However, we observe interesting RARs, such as the 4-length rule from the S category, which characterizes only this category of students. On the other hand, the

RARs expressing the E category have the higher precision (0.716) and this is also observable on the SOM, as this category is easily distinguishable from other classes, sustaining the conclusions from Section 4.1.

Regarding the usefulness of attribute a_3 in mining relevant RARs, the following were observed by analysing the RARs depicted in Figure 5 and visualized in Figure 6b. On the one hand, a_3 creates some misleading relationships between the features, such as the rule $a_3 > a_4$, creating overlapping values with other grades (i.e. it is interesting for both F and S categories). That would cause some misclassifying when interpreting border cases, students that are in between two classes. When removing the a_3 feature, the RAR model is improved, as there are less overlapping areas, even though there are not as many relationships between the features (but the number of RARs may be increased by reducing the minimum confidence threshold). This can also be seen in the experiment from Section 4.1, when comparing the two SOMs (built with and without attribute a_3). On the SOM from Figure 6b built without considering a_3 feature, there is a tendency of the higher and lower grades to create distinct clusters that have little or no noise. The class of F labeled instances is more clearly distinguishable in Figure 6b than in Figure 6a.

The results previously presented highlight the potential of the sets RAR_c to differentiate the students according to their final examination grade category, based on their grades received during the academic semester. As previously shown, RARs are able to express interesting patterns in academic data sets and are useful for providing a better insight into the problem of students' academic performance prediction. However, we have a small number of attributes in our case study. By increasing the number of relevant attributes, it is very likely that more informative and meaningful RARs would be mined.

It is worth mentioning that the results obtained using both SOM and RAR models conducted to similar conclusions, which were detailed in Section 4.1 and 4.2. For obtaining a more accurate representation of the input instances (students) using both unsupervised learning models investigated in this paper (SOMs and RARs), the attribute set characterizing the students must be enlarged with other relevant characteristics. It would be useful to have multiple attributes in the mining process and to extend the set of relations used in the mining process in order to obtain much more informative and relevant RARs as well as a better separation using the SOM model.

A more in depth analysis of the outlier instances provided by the SOM and RAR models may provide valuable information regarding the improvement of the educational processes. For instance, the results of the unsupervised learning processes may reveal the following: (1) the examination grades for some of the evaluations received during the academic semester may be incorrect due to

the variations within the instructors evaluation criteria or standards, as well as possible cheating methods used by a few students; (2) some of the partial examinations may be redundant; (3) a change of the computation method for the partial grades may be required; (4) it could be necessary to increase the number of the examinations performed during the academic semester.

5. Conclusions and future work

This paper examined two unsupervised learning models, *self-organizing maps* and *relational association rule mining*, in the context of analysing data sets related to students' academic performance. Experiments performed on a real data set collected from Babeş-Bolyai University, Romania highlighted the potential of unsupervised learning based data mining tools to detect meaningful patterns regarding the academic performance of students.

We may conclude that the grades received by the students during the semester may be relevant in predicting their final performance. However, several outliers were observed in the data set. Such anomalous instances may be due to: (1) a small number of students' evaluations during the semester (attributes); (2) the students' learning process which is not continuous during the academic semester; (3) the difference between the evaluation standards of the instructors from the laboratory and seminar activities. As a consequence, an increased number of evaluations during the academic semester would be useful, for stimulating students to study during the semester and not only for the final examination.

Future work will be performed in order to extend the experiments and the analysis of the obtained results. For increasing the performance of the unsupervised learning process, methods for detecting anomalies and outliers in data will be further investigated. In addition, a post-processing phase for filtering the set of mined RARs will be analysed for removing rules which overlap with multiple classes.

References

[1] Academic data set, 2018. http://www.cs.ubbcluj.ro/∼liana.crivei/AcademicDataSets/ThirdDataSet.txt.

[2] Elizabeth Ayers, Rebecca Nugent, and Nema Dean. A comparison of student skill knowledge estimates. In *Educational Data Mining - EDM 2009, Cordoba, Spain, July 1-3, 2009. Proceedings of the 2nd International Conference on Educational Data Mining.*, pages 1–10, 2009.

[3] Alejandro Bogarín, Rebeca Cerezo, and Cristóbal Romero. A survey on educational process mining. *Wiley Interdisc. Rew.: Data Mining and Knowledge Discovery*, 8(1), 2018.

[4] Alina Câmpan, Gabriela Şerban, and Andrian Marcus. Relational association rules and error detection. *Studia Universitatis Babes-Bolyai Informatica*, LI(1):31–36, 2006.

[5] Gabriela Czibula, Maria-Iuliana Bocicor, and Istvan Gergely Czibula. Promoter sequences prediction using relational association rule mining. *Evolutionary Bioinformatics*, 8:181–196, 04 2012.
[6] Ashish Dutt, Saeed Aghabozrgi, Maizatul Akmal Binti Ismail, and Hamidreza Mahroeian. Clustering algorithms applied in educational data mining. *Intern. J. of Information and Electronics Engineering*, 5(2):112–116, May 2015.
[7] N. Elfelly, J.-Y. Dieulot, and P. Borne. A neural approach of multimodel representation of complex processes. *International Journal of Computers, Communications & Control*, III(2):149–160, 2008.
[8] Syed Tanveer Jishan, Raisul Islam Rashu, Naheena Haque, and Rashedur M. Rahman. Improving accuracy of students' final grade prediction model using optimal equal width binning and synthetic minority over-sampling technique. *Decision Analytics*, 2(1):1, Mar 2015.
[9] S. Kaski, and T. Kohonen. Exploratory data analysis by the self-organizing map: Structures of welfare and poverty in the world. *Neural Networks in Financial Engineering. Proceedings of the Third International Conference on Neural Networks in the Capital Markets*, pages 498–507, World Scientific, 1996.
[10] S.A. Khadir, K.M. Amanullah, and P.G. Shankar. Student's academic performance analysis using SOM. *International Journal for Scientific Research and Development*, 3(02):1037–1039, 2015.
[11] Wattanapong Kurdthongmee. Utilization of a self organizing map as a tool to study and predict the success of engineering students at Walailak University. *Walailak Journal of Science and Technology*, 5(1):111–123, 2008.
[12] J. Lampinen and E. Oja. Clustering properties of hierarchical self-organizing maps. *Journal of Mathematical Imaging and Vision*, 2(3):261–272, 1992.
[13] Siti Khadijah Mohamad and Zaidatun Tasir. Educational data mining: A review. *Procedia - Social and Behavioral Sciences*, 97:320–324, 2013.
[14] Suhem Parack, Zain Zahid, and Fatima Merchant. Application of data mining in educational databases for predicting academic trends and patterns. *22012 IEEE International Conference on Technology Enhanced Education (ICTEE)*, pages: 1–4, 2012.
[15] K. Saxena, S. Jaloree, R.S. Thakur, and S. Kamley. Self organizing map (SOM) based modelling technique for student academic performance prediction. *Intern. Journal on Future Revolution in Computer Science and Communication Engineering*, 3(9): 115–120, 2017.
[16] Gabriela Şerban, Alina Câmpan, and Istvan Gergely Czibula. A programming interface for finding relational association rules. *International Journal of Computers, Communications & Control*, I(S.):439–444, June 2006.
[17] Panu Somervuo and Teuvo Kohonen. Self-organizing maps and learning vector quantization for feature sequences. *Neural Processing Letters*, 10: 151–159, 1999.
[18] Yi Sun. On quantization error of self-organizing map network. *Neurocomputing*, 34(1-4): 169–193, 2000.

Babeş-Bolyai University, Department of Computer Science, 1 M. Kogălniceanu Street, 400084 Cluj-Napoca, Romania

Email address: cgir2476@scs.ubbcluj.ro, liana.crivei@cs.ubbcluj.ro