Comput Syst Sci & Eng (2019) 3: 131–144
© 2019 CRL Publishing Ltd

International Journal of
Computer Systems
Science & Engineering

# Non-deterministic outlier detection method based on the variable precision rough set model

**Alberto Fernández Oliva**[1], **Francisco Maciá Pérez**[2], **José Vicente Berná-Martinez**[2]*, **Miguel Abreu Ortega**[3]

[1] *Department of Computer Science, Faculty of Mathematics and Computer Science, University of Havana, Cuba*
[2] *Department of Computer Technology, University of Alicante, Spain.*
[3] *Department of E-Commerce, Carnival Cruise Line, Florida, United States*

This study presents a method for the detection of outliers based on the Variable Precision Rough Set Model (VPRSM). The basis of this model is the generalisation of the standard concept of a set inclusion relation on which the Rough Set Basic Model (RSBM) is based. The primary contribution of this study is the improvement in detection quality, which is achieved due to the generalisation allowed by the classification system that allows a certain degree of uncertainty. From this method, a computationally efficient algorithm is proposed. The experiments performed with a real scenario and a comparison of the results with the RSBM-based method demonstrate the effectiveness of the method as well as the algorithm's efficiency in diverse contexts, which also involve large amounts of data.

Keywords: Outliers, Rough Sets (RS), RS Basic Model (RSBM), Variable Precision Rough Set Model (VPRSM), data set, Data Mining.

## 1. INTRODUCTION

The detection of exceptional cases (i.e., outlier detection) is a field of growing relevance within Data Mining. If the detection goal is to extract the most probable patterns of knowledge from large volumes of data (i.e., trend expressions that ignore the marginality or the exception), the opposite view is used in outlier detection. This process could report knowledge findings of strategic importance in a wide range of applications: fraud detection, the detection of illegal access to corporate networks, the detection of errors in input data, etc. This makes us consider the following aspect: although the term outlier could lead us to assume that it always implies a negative interpretation of the phenomenon due to the exceptionality implied by its conception, this is not always the case in practice. In diverse occasions, the detection of outliers, such as cases distinguished by their exceptionality, can be the fundamental objective of certain processes, analyses or studies. Therefore, the term "outlier" should not always be connected with a negative connotation.

Currently, researchers create models, algorithms and functions that are defined in increasingly abstract cases; the development of investigations is thus more conditioned by the nature of the data they investigate. As a result, it is necessary to conceive increasingly novel and efficient data analysis techniques. Data Mining has been established as a subfield of artificial intelligence that can provide techniques, theories and tools that efficiently allow the analysis of the complex datasets of today's world [1]. Essential aspects that justify the transcendence of outlier detection in the context of Knowledge Discovery on Data-Data Mining (KDD-DM) are important.

From the KDD-DM perspective, outlier detection is generally approached from two different points of view: outliers as undesirable objects that must be addressed or removed in the data preparation phase because their presence in a dataset can

*Corresponding Author. E-mail: jvberna@ua.es, Telephone +34 96 590 34 00 – ext. 2114, Fax: +34 96 590 9643

significantly hinder the detection of reliable patterns or outliers as objects that can be identified from the implicit interest that they have for the process itself. In the latter case, they should not be removed from a dataset. For certain applications, these objects are more representative and interesting than the most common events from the perspective of information discovery. Examples of applications in this sense can include those related to fraud detection in the use of credit cards, where the detection of outliers could provide information to typify misconduct patterns, or in the electronic business field, where the detection of outliers could provide useful information for Customer Relationship Management (CRM).

These concepts highlight that KDD-DM processes require increasingly efficient methods for the detection of outliers. In today's datasets, increasingly sophisticated data representation structures and forms of storage tend to appear. Therefore, work must be performed based on obtaining effective detection models based on the challenges imposed by such particularities and on the use of new technologies in general.

The investigation of state-of-the-art techniques in this study has allowed us to identify the extent of the outlier detection problem based on its application in multiple contexts. Our conclusion is that its scope of application is wide and diverse. This diversity of application fields, in which the nature of the data and the contexts in which they are defined acquire different particularities, is perhaps one of the reasons that explain the wide variety of existing detection methods. Each method adjusts to the data and the contexts in which they will be applied; thus, it is challenging to conceive increasingly flexible detection methods that can be applied in different contexts.

With the goal of making outlier detection more efficient, researchers tend to apply new techniques. The Rough Set Basic Model (RSBM) proposed by Professor Z. Pawlak [2] in 1982 is based on a simple and solid mathematical basis: the equivalence relation theory, which describes partitions constituted by indiscernible types of objects. In recent years, this model has been successfully applied in diverse contexts. In [3], we proposed a method based on the RSBM that demonstrated the validity and potential of this method for the detection of outliers. However, it could also be confirmed that the RSBM only allows accurate classifications, and many problems generally require uncertainty to be admitted into a given classification along with having the capacity to generalise the conclusion obtained from more reduced datasets.

In this study, the initial hypothesis is that the Variable Precision Rough Set Model (VPRSM) [4] can provide a solution to the abovementioned problem. Relying on the non-deterministic character provided by the VPRSM and by the relaxation of the set inclusion concept that allows the management of certain thresholds set by the user, we propose a new model in this study based on the VPRSM and create a new algorithm based on the algorithm presented in [3], which shows significant improvements in its generalisation and detection capacity while maintaining the spatial and temporal complexity levels that make it viable in practice.

The remainder of this article is structured as follows. Section 2 presents the most significant aspects obtained as a result of the state-of-the-art study performed with regard to outlier detection and the previous studies that constitute the background of this proposal. Section 3 presents the outlier detection proposal from a model based on the VPRSM properties and an algorithm based on this method; a detection example is also used to illustrate the execution of the proposed algorithm. In Section 4, the validation of the results is performed using multiple experiments that show improvements in the detection quality and in the computational feasibility of the algorithm; these experiments also allow for comparisons to be made with other methods. Finally, Section 5 presents the primary conclusions of this study along with the suggestions for future work.

## 2. BACKGROUND

In recent decades, the outlier detection problem has acquired special relevance in multiple and diverse contexts [5], among which the following can be highlighted as examples: fraud detection in the use of credit cards or in cellular telephony; the identification of conflicting users in the processing of bank loan applications; the detection of intruders in computer networks; the monitoring of traffic in computer networks; the diagnosis of faults or flaws in the operation of engines, generators, pipelines and measuring instruments; the detection of structural defects; the automated control of production lines to detect faulty productions; the automated monitoring of medical parameters; and the identification of new molecular structures [6, 7, 8, 9, 10, 11, 12, 13, 14, 15]. All these challenges highlight the interest of the scientific community in solving the outlier detection problem with efficient techniques and methods.

In general, outlier detection methods are based on two fundamental fields of mathematics and computer science: statistics and artificial intelligence (AI). To establish an adequate outlier detection process, it is necessary to select an algorithm that precisely models the analysed data, that accurately highlights the objects' exceptional nature based on certain specific technique, and that is computationally efficient and scalable for the datasets on which it will be applied within a context or vicinity of interest [16].

Statistical models were the first to be used to address the outlier detection problem. A large subset of outlier detection methods, even the more recent methods, incorporate certain statistical techniques. In general, these techniques are closely linked to the KDD-Data Mining processes [17] [18]. From the statistical perspective, outliers can be indicative of the characteristics of a population segment that would be discovered in the normal course of the analysis or can be interpreted as objects that are not representative of the population and oppose the objectives of the analysis, allowing them to seriously distort the results of statistical tests. Their detection presupposes an analysis of the data to be able to determine the type of influence they exert on the data. Current AI techniques make certain traditional statistical detection methods obsolete [19]. In general, statistical methods are appropriate for processing datasets with real continuous values that are composed of quantitative data or at least qualitative data with ordinal values. However, it is currently necessary to process increasing amounts of categorical data. Parametric methods assume that the data must follow a parametric distribution, and in such a case, the abovementioned methods do not work correctly in multivariate contexts. As a

solution to this problem, the methods used in this study are non-parametric methods that consider distance, clustering or density. However, the temporal complexity of most of the distance-based methods is quadratic, which is important if working with large or dynamic sets of data [20]. Another problem to consider in the use of statistical methods is the dimensionality increment of the dataset [21]; this can also cause an increase in the processing time and a distortion of its distribution. When the dimension of the dataset increases, the effectiveness of certain algorithms can be severely compromised [22]. For example, the concept of distance in a space of different dimensions tends to vary. High dimensionality can also affect the effectiveness of the methods based on proximity criteria between the data. In general, in high-dimensionality contexts, data are scattered; some authors [23, 24] have worked to develop techniques and approaches that can solve this problem, while others focus on selecting the most outstanding attributes to reduce a dataset's dimensionality [25]. Other techniques project the data onto a space with smaller dimensions. In the data mining context, the distribution of the attributes' values is almost always unknown. When the dataset dimension increases significantly, it is difficult to estimate multi-dimensional distributions from it [26]. In such cases, estimation methods can have a much higher margin of error, which limits the application of distribution-based detection methods. With regard to AI-based methods, the fundamental techniques on which they are based are those related to Machine Learning [27], such as Decision Trees and Neural Networks.

A wide set of outlier detection methods implemented from algorithms of diverse types are described in [26, 28]. In general, the algorithms or outlier detection methods are classified based on the technique on which they are based [29]. Among the most noteworthy are the following: distribution-based techniques [30, 31, 32, 33, 34]; techniques based on different depth criteria [35, 26]; distance-based techniques [37] [38, 39, 40, 41, 42, 43]; density-based techniques [44, 45, 46, 47]; methods based on clustering techniques [48, 48, 59, 51, 52, 43]; and techniques based on the use of neural networks [54, 55, 56, 57]. Performing a more in-depth review, we have been able to confirm that there are many more techniques and that there is a tendency for them to proliferate based on the accelerated and constant development of new information technologies. Each technique highlights the new aspects that distinguish them from the others [58].

Researchers tend to apply new techniques to the outlier detection problem to improve the efficiency of the detection process; these new techniques include the Rough Set Basic Model (RSBM) [2]. The RSBM consists of a model with a simple and solid mathematical basis: the equivalence relation theory. These relations describe partitions constituted by indiscernible types of objects. This model has been successfully applied in diverse contexts, including Knowledge Discovery from Data, Data Mining, Machine Learning, Expert Systems, and Decision Support Systems. The RSBM has gained the attention of academics and researchers at the international level. However, as a basis for outlier characterization and detection,**Rough Sets** consider a new perspective and have great potential with regard to theory and practical applicability.

In recent years, proposals have appeared from which efficient algorithms can be built for the detection of outliers based on Rough Set (RS) theory [59, 60, 61]. Jiang et al. [62] proposed a new approach to the outlier detection problem based on the RSBM in which outliers are defined as elements of **non-redundant exceptional sets** that have a**degree of marginalisation** greater than an established threshold. Although the underlying idea is intuitive, it leads to an intractable problem due to it being of exponential order. In a previous work [3], we reached a solution to this computational intractability problem via an extension of the theoretical framework proposed, from which an outlier detection method was established with a simple and rigorous formal theoretical approach based on the existing definition of outliers. The method is computationally feasible for large datasets, and to demonstrate this, we proposed an RSBM-based outlier detection algorithm with a non-exponential temporal and spatial complexity order. The proposed algorithm is linear with respect to the cardinality of the data universe on which it is applied and quadratic with respect to the number of equivalence relations used for describing said universe; however, such a number actually represents a constant because its value is typically significantly much lower than the cardinality of the analysed universe. The method is based on an original and novel approach of RS theory, which has not been previously used in any of the classification categories for the outlier detection methods [63, 64]. The method is applicable to data expressed in tabular form (i.e., the data structure of the Relational Model). The table must be at least in the $1^{st}$ normal form to guarantee that there are no redundancies, and its attributes must be single-valued; otherwise, they will be in contradiction with the essence of the method because the possibility of establishing equivalence relations from them would not exist. The above-mentioned explanation directs the application field towards outlier mining in large datasets. This method is applicable to both continuous and discrete data, and the fact that the datasets can contain a mixture of attribute types (e.g., continuous and categorical attributes mixed) is not a limitation for its application.

The fundamental contribution of the RS theory is to facilitate classification analysis. The approximation, both upper and lower, becomes necessary because of the inability to establish complete classifications of objects that belong to a certain category with the knowledge available [65].

With a certain frequency, the information available only allows partial classifications to be made, and RS theory can be efficiently used to model this type of classification. However, from this theory, such a classification must be true [66], limiting the possibility of conceiving a classification with a controlled degree of uncertainty (i.e., the possibility that there is a certain error in the classification). This is not possible with the RSBM. Paradoxically, in practice and in many cases, it is convenient to admit a certain degree of uncertainty in the classification process, which can allow for better comprehension and use of the properties of the data being analysed [80].

Another limitation of the RSBM is that it assumes that the universe U of objects or data considered is known and that all the conclusions derived from the application of such a model are only applicable to that set of objects. However, in practice, generalising the conclusions obtained from a small set of objects (U) to a larger universe (e.g., the real world) is typically required . The RSBM allows hypotheses to be obtained that are only based on error-free classification rules, which are expressed in the lower approximation, $\underline{X}$, obtained from the analysis of the

data involved (U); thus, the RSBM is a deterministic model. However, in reality, there are multiple situations that require the need for considering incorrect partial classifications. An incorrect partial classification rule also provides useful information and can establish the tendency of values if most of the available data to which the rule is applied can be correctly classified.

A generalisation of the RSBM was proposed by W. Ziarko and is called the Variable Precision Rough Set Model (VPRSM) [4]. The VPRSM model rectifies the deterministic character with regard to classification presented by the basic RS model by starting from a simple idea: the relaxation of the set inclusion concept. This concept effectively manages certain thresholds established by the user. Thus, the VPRSM provides the possibility of detecting or establishing this information trend and to perform analyses on a universe of objects or data. Thus, the VPRSM is a statistical model [67]. The most relevant aspects of the VPRSM are that it is fundamentally aimed towards solving the limitations of the RSBM. Its basis lies in a new conception or generalisation of the standard concept of set inclusion relaxation.

The primary objective of this study is to improve the method based on RSBM [3] with the creation of a non-deterministic outlier detection method based on the VPRSM. This new method must remain computationally feasible for what we conceive an algorithm that allows us to validate it. The starting hypothesis is that the VPRSM model broadens the application of the original method, which is based on the RSBM, to contexts in which a classification with a certain degree of uncertainty is required.

## 3. PROPOSAL FOR VPRSM-BASED OUTLIER DETECTION

### 3.1 Detection method based on the VPRSM properties

The VPRSM is a generalisation of the RSBM [68] [69] and is derived from the RSBM without assuming anything additional. From this generalisation, the management of information with a certain degree of uncertainty is allowed [70]. Numerous investigations related to its application have been produced since the emergence of this model [71, 72, 73, 74, 75, 76], which demonstrates its usefulness and viability.

The essence of the VPRSM model is given by the generalisation of the standard concept of set inclusion relaxation [77]. This concept is too rigorous for representing a nearly complete set inclusion. Based on an extended concept for this relation defined in the VPRSM model [78], a certain degree of error is allowed to be established or foreseen.

In this section, we construct the proposed outlier detection method as we present and analyse the mathematical tools provided by the VPRSM model [4].

It becomes evident from the definition of the standard inclusion relation (see **Definition 1**) that there is no possibility of contemplating any type of declassification.

**Definition 1 —Standard inclusion relation**: Let U be a finite universe of objects and X, Y $\subset$ U; $X \neq \phi$; and $Y \neq \phi$. Then, X is included in Y, or X$\subseteq$Y. If $\forall$x$\in$X, then x$\in$Y. Fig. 1 graphically illustrates this definition.


**Figure 1** Standard inclusion relation.

The first step to overcome the limitations imposed by the RSBM consists of breaking free of the need of explicitly defining the universal quantifier. The "measure of the degree of declassification" (see **Definition 2**) proposed in the VPRSM makes this possible.

**Definition 2 —Measure of the degree of declassification:** The measure of the degree of declassification relative to the set X with respect to set Y, c(X, Y), is the existing relative error when classifying a set of objects and is defined as:

$$c(X,Y) = \begin{cases} 1 - |X \cap Y|/|X| & \text{if } |X| \neq 0 \\ 0 & \text{if } |X| = 0 \end{cases}$$

This definition is evident because it can be observed that:

- if X$\subseteq$Y $\Rightarrow$ |X$\cap$Y|=|X|, then c(X, Y) = 1 – 1/1 = 0 $\Rightarrow$ there is no error in the classification.
- if c(X, Y) $\approx$ 1 $\Rightarrow$ X, Y are nearly disjointed.
- if c(X, Y) = 1 $\Rightarrow$ |X$\cap$Y|=0 $\Rightarrow$ X, Y are disjointed.

The numerical expression c(X, Y) is indicative of the relative classification error. The product c(X, Y)*|X| will indicate the absolute classification error (i.e., the number of misclassified objects).

If the measure of relative declassification is used as a reference, the inclusion relation can be defined to obviate the need to explicitly set the general quantifier as follows: X$\subseteq$Y $\Leftrightarrow$ c(X, Y) = 0. Based on this definition, c(X, Y) can have values greater than 0 without being too high when the relation represents a majority. Thus, a majority of the objects of X must be classified in Y. The concept of the majority imposes the setting of a threshold, and in such a case, it is assumed that the majority implies that more than 50% of the elements of X should be common with Y. Thus, the specification of an admissible threshold of error in the classification is added to the definition of the inclusion relation [18].

**Definition 3 —Majority inclusion relation**: Let U be a finite universe of objects; $0 \leq \beta < 0.5$, where $\beta$ is the admissible declassification error; and X, Y $\subset$ U, X$\neq \phi$, Y$\neq \phi$. Then, X is said to be primarily included in Y, or X is included in Y with a $\beta$-error, X$\subseteq^{\beta}$Y, if and only if c(X, Y)$\leq \beta$. From the same definition, it can be shown that $\beta$=0 expresses a standard inclusion relation, which is called the total inclusion in this model.

In the example shown in Fig. 2, it is assumed that the following sets are present: $X_1$= {$x_1$, $x_2$, $x_3$, $x_4$}, $X_2$={$x_1$, $x_2$, $x_5$}, $X_3$= {$x_1$, $x_6$, $x_7$} and Y = {$x_1$, $x_2$, $x_3$, $x_8$}. The majority inclusion relation is illustrated between $X_1$, $X_2$, $X_3$ and Y. Note the degree of declassification existing between those sets and set Y. Additionally, note that from the given definition of majority inclusion, $X_3 \subseteq^{\beta}$Y is not fulfilled because the declassification error $\beta > 0.5$ between these two sets.

From the new definition of the inclusion relation, the most representative concepts of the RSBM can be redefined as follows.

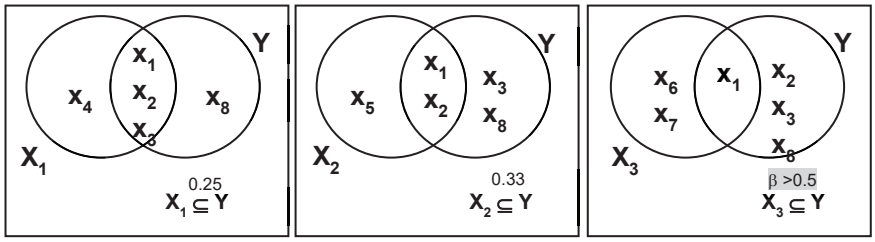



**Figure 2** Example of majority inclusion.

**Definition 4**: Let X be an arbitrary subset of the universe U, and $\Phi \subseteq U \times U$ be an equivalence relation that divides U into a finite set of equivalence classes $<x>_\Phi$, defining:

a) $\underline{X}_\beta = \cup\{\langle x\rangle_\Phi : \langle x\rangle_\Phi \subseteq^\beta X\}$ It is known that $\langle x\rangle_\Phi \subseteq^\beta X \Leftrightarrow c(\langle x\rangle_\Phi, X) \leq \beta$.

b) $\overline{X_\beta} = \cup\{\langle x\rangle_\Phi : \langle x\rangle_\Phi \not\subset^\beta X^c\}$ It is demonstrated that $\langle x\rangle_\Phi \not\subset^\beta X^c \Leftrightarrow c(\langle x\rangle_\Phi, X) < 1-\beta$.

c) $BN_\beta$ ($\beta$-boundary region) = $\overline{X_\beta} - \underline{X}_\beta$

d) $B^\beta$ ($\beta$-inner boundary region) = $X \cap BN_\beta$

e) $NEG_\beta$ ($\beta$-negative region) = U- $\overline{X_\beta}$

In Fig. 3, it can be observed that the RSBM is a particular case of the variable precision model. The figure shows the representative regions of the basic model with a classification error $\beta$=0. In such a situation, the VPRSM corresponds to the RSBM.

In Fig. 4, significant regions are shown to vary if a certain classification error is allowed. In this case, $\beta$=0.1 is assumed. Additionally, note that the $\beta$-negative region of X is the union of all the equivalence classes that can be classified within $X^c$ with a classification error not higher than $\beta$.

Considering that when $\beta$=0, the standard RS model is a particular case of the VPRSM, the following proposition can be established, where other relations that are also fulfilled are expressed.

**Proposition 5:**

a) $\underline{X} \subseteq \underline{X}_\beta$: the lower approximation is a subset of the $\beta$-lower approximation

b) $\overline{X}_\beta \subseteq \overline{X}$: the $\beta$-upper approximation is a subset of the upper approximation.

c) $BN_\beta \subseteq BN$: the $\beta$-boundary region is a subset of the boundary region.

d) $NEG \subseteq NEG_\beta$: the negative region is a subset of the $\beta$-negative region.

When the classification error $\beta$ increases, the sizes of the positive and negative regions of X increases, while that of the boundary region decreases. Fig. 5 shows the variation of the approximated regions based on the variation of the $\beta$-error and illustrates and summarizes many of the properties that have been mentioned.

Based on the concept of the majority inclusion relation defined in the VPRSM, we have developed a new outlier detection method that allows for classification with a certain degree of error when calculating significant regions.

## 3.2 Outlier detection algorithm

From the method proposed in the previous section, an algorithm must be built that can improve the detection quality and provide a wider range of applications while maintaining the spatial and temporal complexity levels obtained to date. Such a method would ensure its own viability in real environments where large amounts of data must be considered.

For the design of this new algorithm, we have started from the RSBM algorithm, which has already been tested and validated [3]. Using the theoretical framework provided by the VPRSM to implement the proposed method, we have modified the calculation of significant regions of the original algorithm, particularly with regard to the determination of the $\beta$-inner boundaries ($B^\beta{}_i$, $1 \leq i \leq m$). As already noted, in such a model, a certain $\beta$-error is allowed in the classification, which objectively translates into relaxing the inclusion relations when establishing the significant regions of the model in the analysis framework. Thus, the possibility of a nearly complete classification is given by relaxing its deterministic character based on the RSBM conception.

The $\beta$-error is added at the inputs of the algorithm implemented for the RSBM; therefore, the inputs for the VPRSM-based algorithm include the following parameters: the universe $U$, the concept $C$ (represented by variable X in the algorithm), the criteria that distinguish the equivalence relations considered in the analysis ($r_i$, $1 \leq i \leq m$), the established detection threshold value $\mu$, and the $\beta$-error. The same data structures described for the RSBM-based algorithm [3] are maintained. The fundamental data structure used in the algorithm is the dictionary, which contains a set of pairs (i.e., *keys* and *values*), where the key is an arbitrary object to which one and only one object of the value-type object is associated. In the algorithm, *keys* are described by the results of applying a classifier to an arbitrary element of the universe. Such a classifier is associated with a particular equivalence relation $r_i$, where $1 \leq i \leq m$, and it allows classification of the members of the equivalence classes defined by said relation. The *values* associated with the *keys* are lists of elements that belong to the equivalence class identified by the *key* associated with said value. For each equivalence relation, a dictionary is built, and from all of these dictionaries, a list of dimension $m$ is built, where $m$ is the number of equivalence relations considered. Based on the data structures used in this study, the spatial complexity of the algorithm is $O(n*m)$ because each dictionary can contain a maximum of all the elements ($n$) of the universe.

Following the strategy of the original algorithm, the new algorithm is composed of two stages: the formation of the $\beta$-inner boundaries and an outlier detection process. In the following, each of these stages is shown and analysed using its pseudocode.

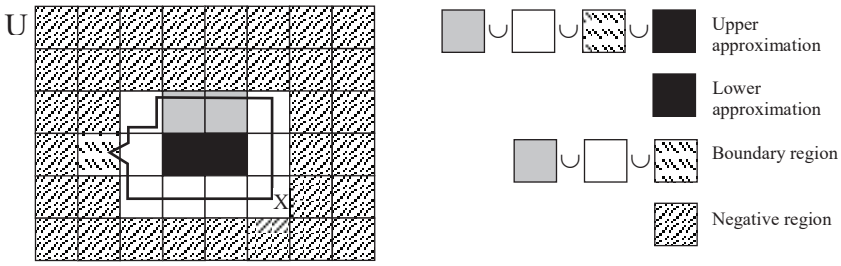



**Figure 3** Representative regions for $\beta$=0. Corresponding to the RSBM

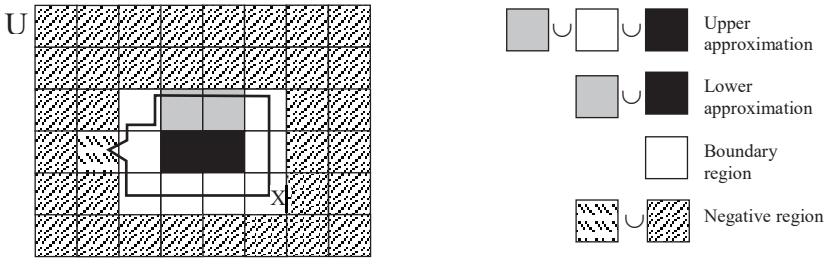



**Figure 4** Variation of the significant regions allowing a declassification error of $\beta$=0.1.

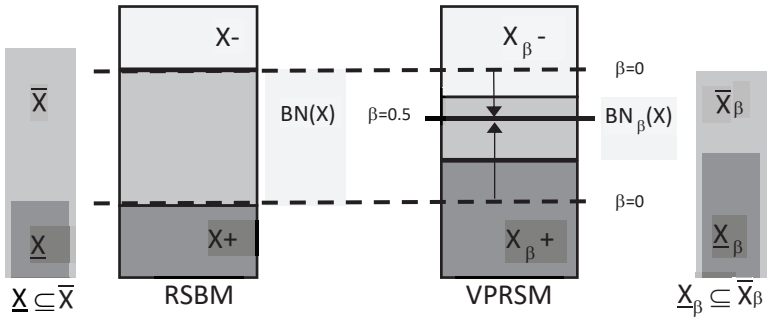



**Figure 5** Variation of the significant regions based on the $\beta$-error variation.

**Stage 1 — Formation of $\beta$-inner boundaries:** the classifiers are applied (one per equivalence relation considered in the analysis) to the elements of U to form the $\beta$-inner boundaries.

**BUILD-REGIONS (U, X, R, $\beta$)**

```
for each r ∈ R
P_r = CLASSIFY-ELEMENTS (U,          // P_r is the partition induced by the
r)                                   equivalence relation <r>
for each class ∈ P_r
if c(class, X)≤ β
X_β = X_β ∪ class                    // By definition 4a: class ⊆^β X ⇒
                                     class ∈ X_β
else if c(class, X) ≥ 1- β
NEG_β = NEG_β ∪ class                // By definition 4b: class ⊄^β X ⇒
                                     class ∈ NEG_β
else
∪ (class ∩ X)                        // By definition 4c: (class ∩ X) ⊆
                                     add the elements of <class>
                                     // that meet the <concept> to the
                                     inner boundary relative to <r>
```

**Alg.1**. Stage 1: Formation of $\beta$-*inner boundaries*

The temporal complexity of this stage is *O(n*m*c),* where $c$ is the cost of classifying each element, $n$ is the cardinality of the universe, and $m$ is the number of equivalence relations considered in the analysis.

**Stage 2 — Outlier detection process:** The set that contains all the elements that meet the concept and can be outlier candidates is made up. From this set, all elements with a *degree of exceptionality* greater than the established detection threshold $\mu$ are classified as such.

**VPRS-OUTLIER-DETECTION (U, X, R, $\beta$, $\mu$)**

The temporal complexity of this stage is $O(n * m^2)$.

Considering stages 1 and 2, the execution cost for the entire algorithm *is O(max(O(stage 1), O(stage 2)))= O(stage 2)=O(n*m$^2$)*.

In general, the number of equivalence relations involved in the analysis in the vast majority of cases is not large compared to the number of elements in the dataset. For this reason, the quadratic dependence of the execution time with respect to the amount of equivalence relations does not markedly affect the algorithm execution time. As shown in the results below, this quadratic dependence is nearly linear for small values (m≤ 20).

With regard to the spatial complexity, the same order is also

```
BUILD-REGIONS (U, X, R, β)
for each r ∈ R                          // For each equivalence relation <r>
containsAnother = FALSE                 // There is no inner boundary that is a subset of
for each q ∈ R – {r}                    // For each equivalence relation <q> different from <r>
if ⊂                                    // If the inner boundary of <q> is a subset of the inner boundary of <r>,
                                        // then its elements are ruled out as members of the Set of possible
                                        // OUTLIERS: E
containsAnother = TRUE
break                                   // It is not necessary to continue
if not containsAnother                  // If no inner boundary is a subset of the one being analysed, then all the
                                        // elements of the inner boundary of < r> are members of the Set of possible
E = E ∪                                 // OUTLIERS: E
for each e ∈ E
if EX-DEGREE(e) ≥ μ                     // The elements of E that exceed a certain degree of exceptionality belong // to the set of outliers
OUTLIERS = OUTLIERS ∪ {e}
```

**Alg.2**. Stage 2: Outlier detection process

maintained, $O(n * m)$, because the data structures described for the version of the algorithm based on the RSBM are maintained.

### 3.3 Example of outlier detection in a data set by the VPRSM algorithm

The operation of the proposed algorithm is shown using an example that highlights the way in which the significant regions vary when a certain $\beta$-error is allowed; this example also describes how the classification is relaxed. In Section 4, the test and validation of the proposal will be addressed with a real dataset.

A universe U that represents 25 patients is considered (Table 1). In this table, a diagnostic is established for whether each patient suffers from flu or not as a function of the patient's temperature and from the presence of a headache or not.

Two criteria are defined, where each divides U into a determined number of equivalence classes:

$$r_1 = \left\{ x \in U \left\{ \begin{array}{l} 1\_if\_headache(x) \\ 0\_otherwise \end{array} \right\} \right\}$$

$$r_2 = \left\{ x \in U \left\{ \begin{array}{l} 0\_if\_Normal\_temperature(x) \\ 1\_if\_High\_temperature(x) \\ 2\_otherwise \end{array} \right\} \right\}$$

A concept is defined as those patients who suffer from the flu:

**CONCEPT** C = {x∈U ^ flu(x)}

Fig. 6-a shows the equivalence classes that are in the partition of U that is created from $r_1$. In both classes, there are elements that fulfil $C$ (i.e., patients with flu) and elements that do not fulfil $C$; therefore, both classes are within the boundary of C with respect to $r_1$. The elements of both classes that fulfil $C$ are those that make up the inner boundary. Fig. 6-b shows how the classification is made when $\beta$=*0*, which is equivalent to the RSBM, and when allowing a declassification error of $\beta$=*0.25* (i.e., the VPRSM). Note that for $r_1$, none of the boundaries change even if the value of $\beta$ varies.

However, when analysing what occurs with regard to $r_2$, it is observed that the introduction of a classification error can vary the boundary elements. Thus, the relation $r_2$ produces 3 equivalence classes for the universe U (Fig. 7-a). In equivalence class 2, 80% of the elements belong to the concept $C$. When the boundaries are built with $\beta$=*0,* equivalence class 2 is within the boundaries between the elements that belong to the concept and those that do not. This occurs because there are many elements that are in equivalence class 2 that do not belong to the concept because they are not patients with flu. However, when a classification error is introduced (i.e., $\beta$=*0.25*), equivalence class 2 enters the positive region because 80% of its elements belong to the concept (Fig. 7-b). This fact makes sense because equivalence class 2 can be considered to be positive with a degree of error of $\beta$=*0.25* if many elements of the class meet the concept $C$.

As shown in the example, the introduction of an error in the classification of the elements that are or are not part of the concept can relax the relation definition and classify the elements with a certain margin of error.

## 4. VALIDATION OF RESULTS

The fundamental objective of the experiments in this study is to validate the proposed hypothesis that the incorporation of the precision variable to the proposed outlier detection algorithm improves the results. However, given the large amounts of data with which work is typically performed for this type of problem, another of the objectives of the tests is aimed at verifying that the temporal complexity of the algorithm remains linear in practice.

We will still incorporate an additional object into the proposed test, where the obtained results can be contrasted and compared to those of other methods, algorithms and strategies. To accomplish this goal, a dataset provided by the UCI Machine Learning Repository of the Center for Machine Learning and Intelligent Systems of the University of California, Irvine [79] was chosen. This dataset contains data from the Census Bureau Database of the United States, has already been used in more than 50 diverse scientific articles, and is therefore considered to be a good reference dataset. In [79], the most outstanding characteristics of this set and a detailed explanation of its attributes can be obtained.

### 4.1 Experiments to determine detection quality

To demonstrate that the proposed method is valid with regard to the detection capacity in real datasets, we have designed certain tests in which we define a concept and a series of equivalence relations and intentionally introduce a set of outliers to the dataset. Then, we use the proposed method for the detection of outliers and analyse the results. The elements defined include the following:

**Table 1** Example data that represent the U universe.

| ID | Headache | Temperature | Diagnostic | ID | Headache | Temperature | Diagnostic |
|---|---|---|---|---|---|---|---|
| 1 | YES | NORMAL | UNKNOWN | 14 | YES | NORMAL | HEADACHE |
| 2 | NO | VERY HIGH | FLU | 15 | NO | VERY HIGH | FLU |
| 3 | YES | HIGH | FLU | 16 | NO | VERY HIGH | FLU |
| 4 | NO | NORMAL | UNKNOWN | 17 | NO | NORMAL | - |
| 5 | YES | VERY HIGH | FLU | 18 | NO | VERY HIGH | FLU |
| 6 | NO | HIGH | UNKNOWN | 19 | YES | HIGH | FLU |
| 7 | NO | HIGH | INSOLATION | 20 | YES | HIGH | FLU |
| 8 | NO | VERY HIGH | FLU | 21 | YES | HIGH | FLU |
| 9 | YES | NORMAL | - | 22 | YES | HIGH | FLU |
| 10 | YES | NORMAL | INSOLATION | 23 | YES | HIGH | FLU |
| 11 | YES | VERY HIGH | FLU | 24 | YES | HIGH | FLU |
| 12 | NO | NORMAL | - | 25 | YES | HIGH | FLU |
| 13 | YES | NORMAL | HEADACHE | | | | |

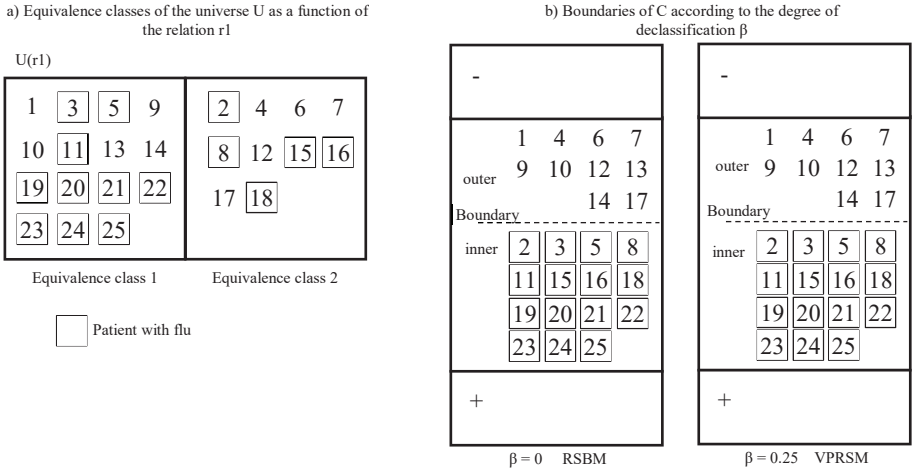



**Figure 6** a) Partition established by $r_1$ on $U$. b) Boundary of C with respect to $r_1$. $\beta$=0; $\beta$=0.25.

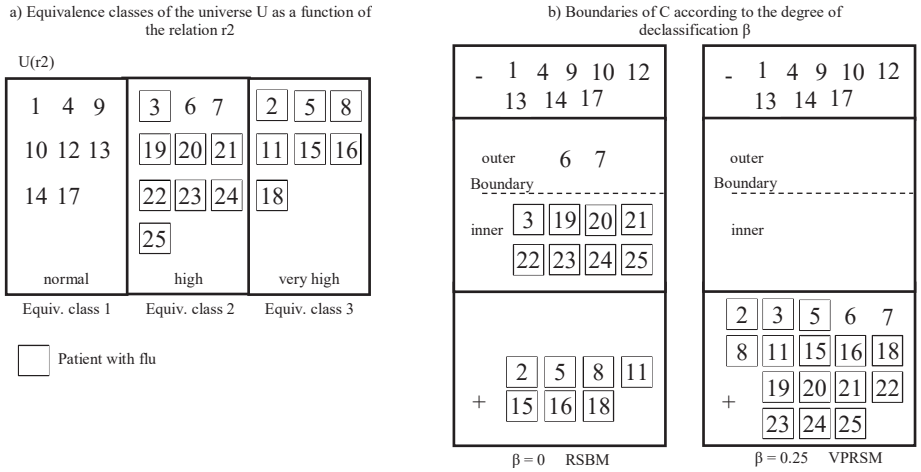



**Figure 7** a) Partition established by $r_2$ on $U$. b) Boundary of C with respect to $r_2$. $\beta$=0; $\beta$=0.25.

- The individuals of the dataset that were the subject of study are those that meet the following CONCEPT: $1 \leq$ people_with_age $\leq 10$.
- The criteria for performing the analysis were established using the following equivalence relations:

```
  r1:  defined from the categorical attribute
"workclass"
  -c1.1:  workclass =['private' OR
'self-emp-not-inc' OR 'self-emp-inc' OR 'federal-gov
```

```
local-gov' OR 'state-gov without-pay']
  -c1.2:  workclass = ['never-worked']
  r2:  defined from the categorical attribute
"education"
  -c2.1:  education = ['bachelors' OR 'some-college'
OR '11th' OR '9th' OR '7th-8th' OR '12th' OR '10th'
OR 'HS-grad' OR 'prof-school' OR 'assoc-acdm' OR
'assoc-voc' OR 'masters' OR 'doctorate']
  -c2.2:  education = ['preschool' OR '1st-4th' OR
'5th-6th']
  r3:  defined from the categorical attribute
"marital-status"
  -c3.1:  marital-status = ['married-civ-spouse'
OR 'divorced' OR 'separated' OR 'widowed' OR
'married-spouse-absent' OR 'married-AF-spouse']
  -c3.2:  marital-status = ['never-married']
  r4:  defined from the categorical attribute
"occupation"
  -c4.1:  occupation = ['tech-support' OR
'craft-repair' OR 'other-service' OR 'sales'
OR 'exec-managerial' OR 'prof-specialty' OR
'handlers-cleaners' OR 'machine-op-inspct'
OR 'adm-clerical' OR 'farming-fishing' OR
'transport-moving' OR 'priv-house-serv' OR
'protective-serv' OR 'armed-Forces']
  -c4.2:  occupation = ['student']
```

Therefore, any element that satisfies the concept and belongs to the class *cx*.1 (x = 1, 2, 3, 4) is contradictory with the relation *rx* because the individuals subjected to the analysis are children between 1 and 10 years of age.

Table 2 shows the set of outliers that were intentionally introduced into the dataset, showing only the attributes that are relevant for the analysis. Values that contradict the concept have been introduced.

In this table, the values marked with an asterisk (*) are contradictory for children between 1 and 10 years old. In this set of outliers, the contradiction levels of the individuals vary. In certain cases, they are contradictory with one or two attributes; in other cases, they are contradictory with three or four and thus represent the most contradictory elements.

Fig. 11 shows the amount of outliers detected for different values of the thresholds $\beta$ (i.e., the declassification error) and $\mu$ (i.e., the degree of exceptionality). The results that correspond to the RSBM are those found when $\beta$*=0*. The values $\beta$*=0.10, 0.20, 0.30, 0.40,* and 0.50 establish errors that are admitted in the classification and therefore correspond to the VPRSM.

The goal of this test is to describe the variation in the amount of outliers detected when the thresholds $\beta$ and $\mu$ are varied. This test also allows the comparison of results that are produced when working with the RSBM ($\beta$*=0*) and the VPRSM($\beta \neq 0$).

When interpreting the results, it has to be noted that in all cases, within the set of outliers detected, there were always some outliers that had been intentionally introduced into the data set. When the amount of outliers detected was higher than the amount of outliers introduced, then all the introduced outliers were within the detected set. When the number of outliers detected was lower than the amount introduced, then those that were detected were always the most contradictory outliers. For example, when $\mu$*=0.02* and $\beta$*=0.0*, 24 outliers were detected, among which 13 had been introduced. Additionally, when $\mu$*=0.6* and $\beta$*=0.2*, only 4 outliers were detected, of which 2 belong to the set of 13 introduced and thus represent the two most contradictory outliers; four attributes were contradictory in those cases. The interpretation of the tests performed also allows us to draw the following conclusions:

- An adequate choice of equivalence relations or classification criteria ensures good detection efficiency.

- For small values of $\mu$ and $\beta$, the number of detected outliers can be high, and elements that are not actually outliers can detected as such. For example, when $\mu$ = 0.2 and $\beta$ = 0.0, 24 outliers were detected, which reaffirms an important aspect of the statistic view of the outlier detection problem for the final designation of a case as exceptional. When the considered candidate observations have been identified by a given detection method, then the investigator must perform an analysis of these results and select those observations that demonstrate real contradictions with respect to the studied sample.

- When gradually increasing the value of the detection threshold ($\mu$), a refinement in the detection is achieved. In general, when the value of this parameter increases, the number of outliers detected decreases. Given this decrease, it can be observed that those that remain in each case are those that are contradictory with a higher number of attributes. However, in certain cases and for certain variations in $\mu$, such refinement is not achieved. For example, 24 outliers are detected when $\mu$ is varied from 0.2 to 0.4 and $\beta$ *= 0.0*. The same results are found when $\mu$ is varied from 0.8 to 1.0 and $\beta$ *= 0.0*. Additionally, in both cases, the number of outliers detected was 9. Note that in the two examples, the value of $\beta$ *= 0.0*, which implies that no degree of declassification has been allowed; therefore, these results are indicative of the RSBM. Additionally, note that when a certain degree of declassification (i.e., $\beta \neq 0.0$) is allowed for the same variations in $\mu$ as in the previous example, the amount of detected outliers is different.

- After $\mu$ reaches its highest possible value (i.e., $\mu$*=1.0*), the number of detected outliers is 9; however, a higher detection refinement can be achieved if $\beta$ is varied until the most contradictory outliers are identified. Thus, detection quality can be improved if a controlled degree of declassification ($\beta$) is allowed and increased gradually. However, we must be cautious with the variation of $\beta$ because allowing a high degree of declassification can result in all elements that are near boundaries going into the positive or negative region, leaving the inner boundaries with no elements because all of them are removed. In the tests performed, for example, it is evident that this phenomenon occurs above $\beta$ *= 0.3* because no outliers are detected above this value.

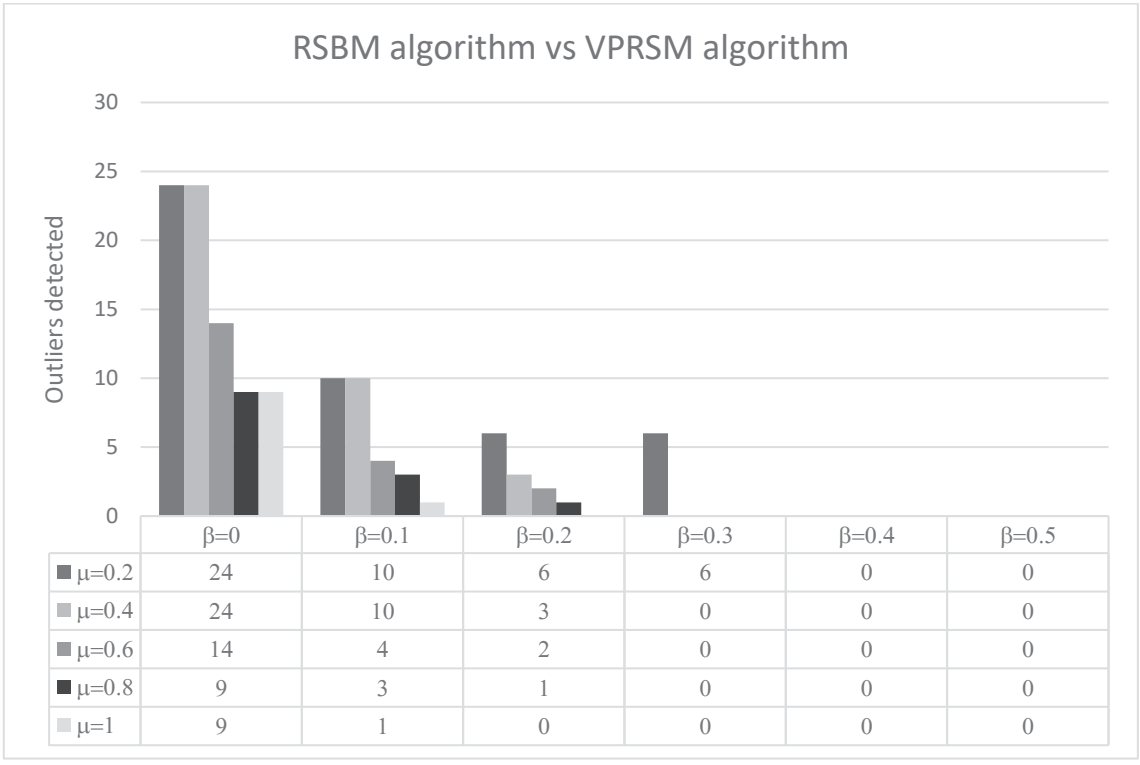
## 4.2 Experiments to determine the algorithm's feasibility

To describe the behaviour of the proposed algorithm, we will analyse its behaviour when considering the variation of all the parameters that define the size of the algorithm input, which include the number of rows and columns of the dataset and the number of equivalence relations considered in the analysis. Additionally, the behaviour of the VPRSM algorithm was compared with that of the original RSBM model.

In Fig. 8, the algorithm execution can be observed while maintaining a constant number of equivalence relations (i.e., 5 relations) and a constant number of rows in the dataset (i.e., 30,000 rows) and varying the number of columns of the dataset from 5 to 14 columns.

**Table 2** Outliers introduced into the dataset. The values marked with * are contradictory with the concept.

| Age | WorkClass | Education | Marital-Status | Occupation |
|---|---|---|---|---|
| 7 | **self-emp-inc*** | 1st-4th | never-married | student |
| 6 | never-worked | **masters*** | never-married | student |
| 9 | never-worked | **doctorate*** | never-married | student |
| 9 | never-worked | 5th-6th | never-married | **Armed-Forces*** |
| 7 | never-worked | 1st-4th | never-married | **Adm-clerical*** |
| 8 | **self-emp-inc*** | **masters*** | never-married | Student |
| 8 | never-worked | **doctorate*** | **married-civ-spouse*** | Student |
| 6 | never-worked | 1st-4th | **divorced*** | **Armed-Forces*** |
| 9 | **federal-gov*** | 5th-6th | never-married | **Adm-clerical*** |
| 3 | self-emp-inc | **masters*** | **married-civ-spouse*** | Student |
| 7 | never-worked | **doctorate*** | **divorced*** | **Adm-clerical*** |
| 2 | **federal-gov*** | **masters*** | **divorced*** | **Armed-Forces*** |
| 8 | **self-emp-inc*** | **doctorate*** | **married-civ-spouse*** | **Armed-Forces*** |




| | β=0 | β=0.1 | β=0.2 | β=0.3 | β=0.4 | β=0.5 |
|---|---|---|---|---|---|---|
| μ=0.2 | 24 | 10 | 6 | 6 | 0 | 0 |
| μ=0.4 | 24 | 10 | 3 | 0 | 0 | 0 |
| μ=0.6 | 14 | 4 | 2 | 0 | 0 | 0 |
| μ=0.8 | 9 | 3 | 1 | 0 | 0 | 0 |
| μ=1 | 9 | 1 | 0 | 0 | 0 | 0 |

**Figure 8** Basic RS algorithm (RSBM) vs. VPRSM with regard to outlier detection.

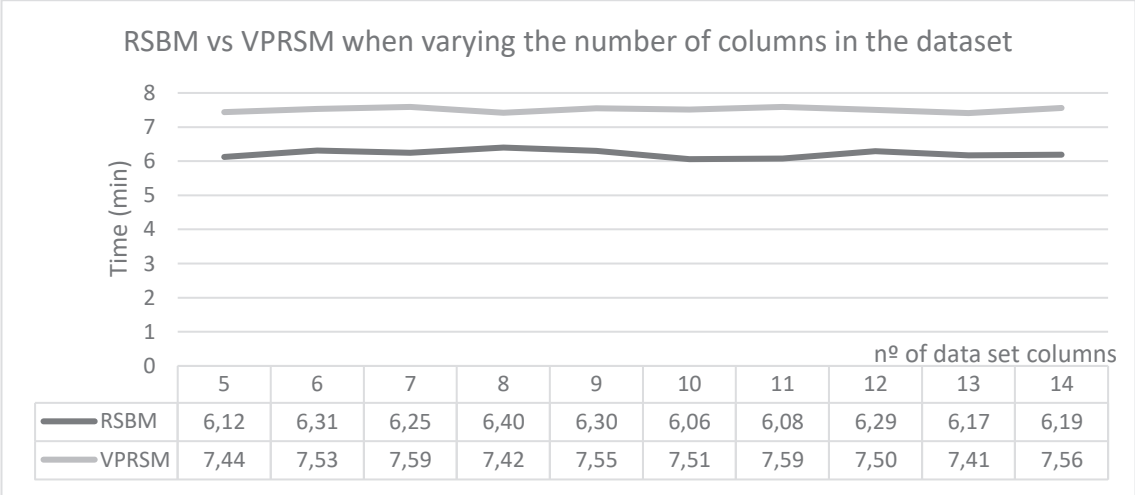
The same experiment was performed, results shown in Fig. 9, while maintaining a fixed number of rows (i.e., 30,000 rows) and columns (i.e., 14 columns) in the dataset and modifying the number of equivalence relations from 2 to 14 relations. The cost is shown to be nearly linear.

As shown in Fig. 10, an experiment was performed that maintained a fixed number of columns in the dataset (i.e., 14 columns) and a fixed number of equivalence relations (i.e., 5 equivalence relations) while the cardinality of the dataset was varied from 5,000 to 30,000 rows.

These results confirm that the temporal complexity orders under execution correspond to those of the algorithm that were justified from the theoretical perspective. The results also demonstrate that the constants of the orders are reasonable and allow such algorithms to be used for realistic datasets. Another important aspect to highlight is that the execution times for both versions of the algorithm (i.e., the RSBM and the VPRSM) do not differ significantly when considering systems that are classified as basic.

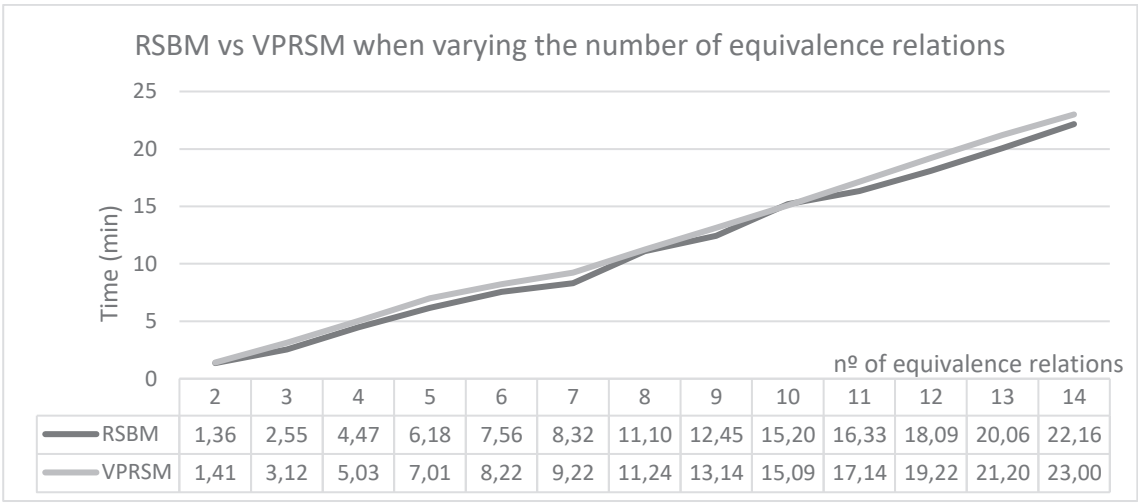
## 5. CONCLUSIONS

The results obtained from the tests performed in this study demonstrate that the proposed VPRSM-based algorithm can eliminate the deterministic character with regard to the classi-



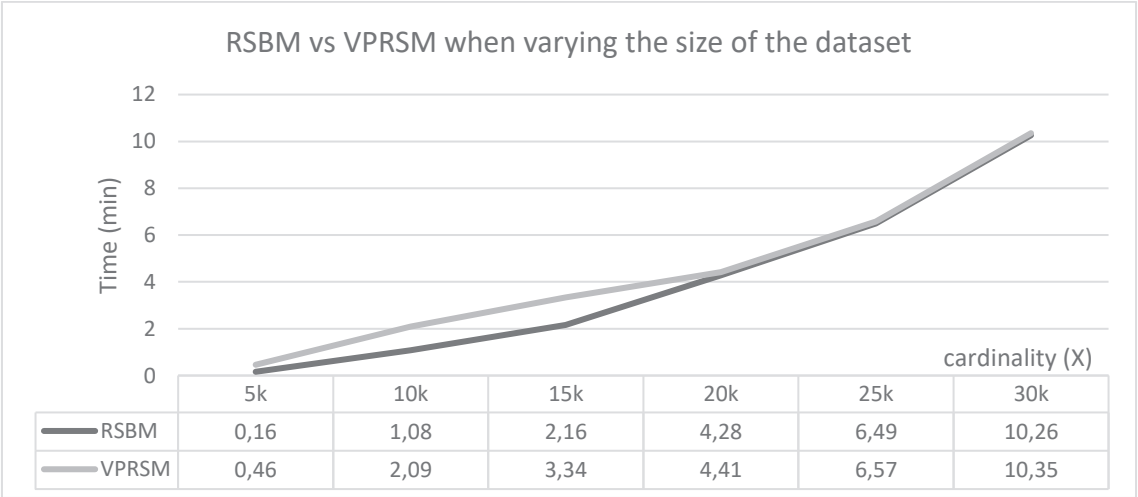

| | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
|---|---|---|---|---|---|---|---|---|---|---|
| RSBM | 6,12 | 6,31 | 6,25 | 6,40 | 6,30 | 6,06 | 6,08 | 6,29 | 6,17 | 6,19 |
| VPRSM | 7,44 | 7,53 | 7,59 | 7,42 | 7,55 | 7,51 | 7,59 | 7,50 | 7,41 | 7,56 |

**Figure 9** Execution time of the RSBM vs. VPRSM when modifying the number of columns in the dataset while maintaining a constant number of rows and equivalence relations.




| | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| RSBM | 1,36 | 2,55 | 4,47 | 6,18 | 7,56 | 8,32 | 11,10 | 12,45 | 15,20 | 16,33 | 18,09 | 20,06 | 22,16 |
| VPRSM | 1,41 | 3,12 | 5,03 | 7,01 | 8,22 | 9,22 | 11,24 | 13,14 | 15,09 | 17,14 | 19,22 | 21,20 | 23,00 |

**Figure 10** Execution time of the RSBM vs. VPRSM while modifying the number of equivalence relations and maintaining a constant number of rows and columns in the dataset.




| | 5k | 10k | 15k | 20k | 25k | 30k |
|---|---|---|---|---|---|---|
| RSBM | 0,16 | 1,08 | 2,16 | 4,28 | 6,49 | 10,26 |
| VPRSM | 0,46 | 2,09 | 3,34 | 4,41 | 6,57 | 10,35 |

**Figure 11** Execution time of the RSBM vs. VPRSM while varying the number of rows in the dataset and maintaining a constant number of columns and a constant number of equivalence relations.

fication that limits the algorithm based on the basic RS model. Thus, a higher precision is achieved in the detection while maintaining the most contradictory elements as outliers.

Additionally, the version of the VPRSM-based algorithm achieves better results in the detection of outliers by refining the candidates and focusing on detecting those outliers that are more contradictory. The proposed method achieves this result while maintaining the same temporal and spatial complexities as the RSBM-based algorithm. The proposed method is shown to provide a computationally efficient solution, offering the possibility of using quasi-linear algorithms, which is an advantage that any data analyst or engineer will value, given the typically elevated complexity of the procedures in the KDD-DM field and the typically large size of datasets.

In the long term, our investigation seeks a much more ambitious objective: to provide a tool that allows the probabilistic prediction of an outlier condition for all elements of a given dataset in a computationally feasible manner. To achieve this goal, the next step in this field of research should consist of creating an algorithm that can automatically calculating the $\mu$ and $\beta$ thresholds involved in the proposed method that must be defined by the user. Based on this algorithm, our investigation will be focused on the creation of a new method that allows the set of such thresholds under which a certain element of a dataset would be an outlier to be determined.

## Compliance with Ethical Standards

Funding: This study was funded by grant TIN2016-78103-C2-2-R and University of Alicante GRE14-02.

The authors declare that they have no conflict of interest.

## REFERENCES

1. J. Han, M. Kamber, J. Pei, Data Mining: Concepts and Techniques, 3rd.ed, ISBN 978-0-12-381479-1, Morgan Kaufmann Publisher (imprint of ELSEVIER), 2012
2. Z. Pawlak, Rough Sets. International Journal of Computer and Information Sciencies, 11(5):341-356, 1982
3. F. Maciá, J.V. Berna, A. Fernández, M.A. Abreu; Algorithm for the detection of outliers based on the theory of rough sets; Decision Support Systems 75, ELSEVIER, pp 63-75, 2015
4. W. Ziarko, Variable Precision Rough Set Model, Journal of Computer and System Sciences, 46(1), pp 39–59, 1993
5. I. Ben-Gal, Outlier detection, Data mining and knowledge discovery handbook, pp 117-130, Springer, 2010
6. J. W. Branch, Ch. Giannella, B. Szymanski, R. Wolff, H. Kargupta, In-network outlier detection in wireless sensor networks, Knowledge and Information Systems, January 2013, Volume 34, Issue 1, pp 23-54, First online: 18 January 2012
7. E.W.T. Ngai, Y. Hu, Y.H. Wong, Y. Chen, X. Sun, The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature, Decision Support Systems, Elsevier, Volume 50, Issue 3, pp 559–569, 2011
8. C.C. Aggarwal, Y. Zhao, P.S. Yu, Outlier detection in graph streams, Data Engineering (ICDE), IEEE 27th International Conference on, pp 399 – 409, ISSN: 1063-6382, Publisher: IEEE, 2011
9. P. Gogoi, D.K. Bhattacharyya, B. Borah and J.K. Kalita, A Survey of Outlier Detection Methods in Network Anomaly Identification, The Computer Journal, Published by Oxford University Press, 2011
10. O.P. Popoola, K. Wang, Video-Based Abnormal Human Behavior Recognition—A Review, IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, Volume: 42, pp 865 – 878, ISSN: 1094-6977, 2012
11. B. Du, L. Zhang, Random-Selection-Based Anomaly Detector for Hyperspectral Imagery, IEEE Transactions on Geoscience and Remote Sensing, Volume: 49, Issue: 5, pp 1578 – 1589, ISSN: 0196-2892, 2011
12. M. Gupta, J. Gao, Y. Sun, J. Han, Integrating community matching and outlier detection for mining evolutionary community outliers, Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '12), pp 859-867, ISBN: 978-1-4503-1462-6, 2012
13. A. Sharma, P. K. Panigrahi, A review of financial accounting fraud detection based on data mining techniques, International Journal of Computer Applications, 2012
14. M.H. Bhuyan, D.K. Bhattacharyya, J.K. Kalita, AOCD: An Adaptive Outlier Based Coordinated Scan Detection Approach, International Journal of Network Security, Vol.14, No.6, pp 339-351, 2012
15. A. Freitas, T. Silva-Costa, F. Lopes, Factors influencing hospital high length of stay outliers, BMC Health Services Research, DOI: 10.1186/1472-6963-12-265; 2012
16. A. Zimek, R.J.G.B. Campello, J. Sander, Ensembles for unsupervised outlier detection: challenges and research questions a position paper, ACM SIGKDD Explorations, Volume 15, Issue 1, Pages 11-22, 2014
17. Y. Zhang, N.A.S. Hamm, N. Meratnia, A. Stein, Statistics-based outlier detection for wireless sensor networks, International Journal of Geographical Information Science, Volume 26, Issue 8, pp 1373-1392, 2012
18. S. Hido, Y. Tsuboi, H. Kashima, M. Sugiyama, T. Kanamori, Statistical outlier detection using direct density ratio estimation, Knowledge and Information Systems, Volume 26, Issue 2, pp 309-336, 2011
19. G. Buzzi-Ferraris, F. Manenti, Outlier detection in large data sets, Computers & Chemical Engineering, Volume 35, Issue 2, pp 388–390, ELSEVIER, 2011
20. V.J. Hodge and J. Austin. A survey of outlier detection methodologies. Artificial Intelligence Review, 22 (2). pp. 85-126. 2004
21. N. Pham, R. Pagh, A near-linear time approximation algorithm for angle-based outlier detection in high-dimensional data, Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD'12), pp 877-885, ISBN: 978-1-4503-1462-6, 2012
22. A. Zimek, E. Schubert, H.P. Kriegel, A survey on unsupervised outlier detection in high-dimensional numerical data, Statistical Analysis and Data Mining, Volume 5, Issue 5, pp 363–387, 2012
23. C.C. Aggarwal, Towards Exploratory Test Instance Centered Diagnosis in High Dimensional Classification, IEEE Transactions on Knowledge and Data Engineering, 19(8), pp 1001-1015, 2015
24. C. C. Aggarwal, P.S. Yu, Outlier Detection with Uncertain Data, Proceedings of the SIAM Conference on Data Mining, pp 483-493, 2008
25. S.K.S. Fan, H.K. Huang, Y.J. Chang, Robust Multivariate Control Chart for Outlier Detection Using Hierarchical Cluster Tree in SW2, Quality and Reliability Engineering International, Volume 29, Issue 7, pp 971–985; Wiley Online Library, 2013
26. V.J. Hodge, J. Austin, Survey of Outlier Detection Methodologies. Artificial Intelligence Review, 22, pp 85-126, 2004
27. M. Last, A. Kandel, Automated Detection of Outliers in Real-

World Data, Proceedings of the Second International Conference on Intelligent Technologies, Bangkok, Thailand, pp 292-301, 2001
28. J. Tang, Z. Chen, A. Fu, D. Cheung, A Robust Outlier Detection Scheme in Large Data Sets, Proceedings of the 6th Pacific-Asia Conference on Knowledge Discovery and Data Mining, Taipei, Taiwan, 2002
29. HP. Kriegel, P. Kröger, A. Zimek, Outlier Detection Techniques, The 2010 SIAM International Conference on Data Mining, Tutorial Notes, Columbus, Ohio, 2010
30. D. Hawkins, Identification of outliers, Chapman and Hall, Reading, 1980
31. P. Rousseeuw, A. Leroy, Robust Regression and Outlier Detection, John Wiley & Sons, Inc. New York, USA, 329 pp, 1987
32. V. Barnett, T. Lewis, Outliers in Statistical Data, 3rd. ed., John Wiley & Sons. Chichester, ISBN 0-471-93094-6, 1994
33. K. Yamanishi, J. Takeuchi, G. Williams, On-line Unsupervised Outlier detection using finite mixtures with discounting Learning Algorithms, Proceedings of the sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD'00, pp 320-325, 2000
34. K. Yamanishi, J. Takeuchi, Discovering outlier filtering rules from unlabeled data: combining a supervised learner with an unsupervised learner, Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD'01, pp 389-394, 2001
35. I. Ruts, P. Rousseeuw, Computing depth contours of bivariate point clouds, Computational Statistics and Data Analysis, 23, No. 1/ 1996, ISSN 0167-9473, pp 153-168, 1996
36. T. Johnson, I. Kwok, R.T. Ng, Fast Computation of 2d depth contours, Proceedings of the ACM SIG KDD'98, pp 224-228, 1998
37. E. Knorr, R. Ng, Algorithms for mining distance-based outliers in large datasets, Proceedings of the 24th Int. Conf. on Very Large Database VLDB'98, New York, pp 392-403, 1998
38. E. Knorr, R. Ng, Finding intentional knowledge of distance-based outliers, Proceedings of the VLDB'99, pp 211-222, 1999
39. F. Angiulli, C. Pizzuti, Fast outlier detection in high dimensional spaces, Proceedings of the 6th European Conference on Principles of Data Mining and Knowledge Discovery PKDD'02, pp 15-26, 2002
40. S.D. Bay, M. Schwabacher, Mining distance-based outliers in near lineal time with randomization and a simple pruning rule, Proceedings of the 9th annual ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2003
41. F. Angiulli, C. Pizzuti, Outlier Mining in Large High-Dimensional Data Sets, IEEE Transactions on Knowledge and Data Engineering, Vol. 17, Art. No. 2, pp 203-215, 2005
42. G.H. Orair, C.H.C. Teixeira, W. Meira Jr, Y. Wang, Distance-based outlier detection: consolidation and renewed bearing, Proceedings of the VLDB Endowment, Volume 3 Issue 1-2, pp 1469-1480, 2010
43. K. Bhaduri, B.L. Matthews, C.R. Giannella, Algorithms for speeding up distance-based outlier detection; Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '11), pp 859-867, ISBN: 978-1-4503-0813-7, 2011
44. W. Jin, A.K. Tung, J. Han, Mining top-n local outliers in large databases, Proceedings of the KDD'01, pp 293-298, 2001
45. A. L. Chiu, A. W. Fu, Enhancements on local outlier detection. Proceedings of the IDEAS'03, 2003
46. T. Hu, S.Y. Sung, Detecting pattern-based outliers, Pattern Recognition Letters, 24 (16), pp 3059-3068, 2003
47. D. Ren, B. Wang, W. Perrizo, RDF: A density-based Outlier Detection Method using Vertical Data Representation, Proceedings of the Fourth IEEE International Conference on Data Mining - ICDM'04, Brighton, UK, 2004
48. L. Kaufman, P. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis, New York, Wiley, 1990
49. G. Sudipto, R. Rajeev, S. Kyuscok, ROCK: A robust clustering algorithm for categorical attributes, Information Systems, 25(5), pp 345-366, 2002
50. F. Jiang, S.S. Tseng, C.M. Su, Two-phase Clustering Process for Outliers Detection, Pattern Recognition Letters, pp 691-700, 2001
51. V. Hautamäki, S. Cherednichenko, I. Kärkkäinen, T. Kinnunen, P. Fränti, Improving k-means by outlier removal, Proceedings of the 14th Scandinavian Conference on Image Analysis (SCIA'05), Joensuu, Finland, pp 978–987, 2005
52. H.M. Koupaie, S. Ibrahim, Outlier detection in stream data by clustering method, International Journal of Advanced Computer Science and Information Technology (IJACSIT), Vol. 2, No. 3, pp 25-34, ISSN: 2296 – 1739, 2013
53. S. Vijayarani, P. Jothi, An Efficient Clustering Algorithm for Outlier Detection in Data Streams, International Journal of Advanced Research in Computer and Communication Engineering, Vol. 2, Issue 9, ISSN (Print): 2319 – 5940, ISSN (Online): 2278 – 1021, 2013
54. H. Simon, H. Hongxing, B. Rohan, W. Graham, Outlier Detection Using Replicator Neural Networks, Proceedings of the 4th International Conference on Data Warehousing and Knowledge Discovery, pp 170-180, 2002
55. Z. He, S. Deng, X. Xu, Outlier detection integrating semantic knowledge, Proceedings of the International Conference for Web Information Age WAIM'02, Lecture Notes in Computer Science, Springer, 2002
56. Z. He, X. Xu, J. Z. Huang, S. Deng, Mining class outlier: concepts, algorithms and applications in CRM. Expert System with Applications, pp. 681–697, 2004
57. L. Toth, G. Gosztolya, Replicator Neural Networks for Outlier Modeling in Segmental Speech Recognition, LNCS, Springer Berlin/ Heidelberg, ISSN 0302-9743, Vol. 3173/2004, pp 996-1001, 2004
58. K. Singh, S. Upadhyaya, Outlier Detection: Applications and Techniques, IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No 3, ISSN (Online): 1694-0814, January 2012
59. F. Jiang, Y. Sui, C. Cao, Some issues about outlier detection in rough set theory, Expert Systems with Applications, Elsevier, Volume 36, Issue 3, Part 1, pp 4680–4687, 2009
60. F. Shaari, A.A. Bakar, A.R. Hamdan, Outlier detection based on rough sets theory, Journal of Intelligent Data Analysis, Volume 13, Issue 2, pp 191-206, 2009
61. Z. Xue, Y. Shang, A. Feng, Semi-supervised outlier detection based on fuzzy rough C-means clustering, Mathematics and Computers in Simulation, Volume 80, Issue 9, pp 1911–1921, ELSEVIER, 2010
62. F. Jiang, Y. Sui, C. Cao, Outlier detection using rough sets theory, Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing (RSFDGrC 2005). Springer, 2005
63. Z. Pawlak, J. Grzymala-Busse, R. Slowinski, W. Ziarko, Rough sets, Communications of the ACM, Volume 38 Issue 11, pp 88-95, ACM New York, NY, USA, 1995
64. G. Wang, Y. Yao, H. Yu, A Survey on Rough Set Theory and Applications, Chinese Journal of Computers, issue No.7, pp 1229—1246, 2009
65. Z. Pawlak, Rough Sets: Theoretical Aspects of Reasoning About Data, s.l.: Spinger, 1991
66. Z. Pawlak, A. Skowron, Rudiments of rough sets, Information Sciences, ELSEVIER, Volume 177, Issue 1, pp 3–27, 2007
67. W. Ziarko, Probabilistic Decision tables in the Variable Precision Rough Set Model, Computer Science Department, University of Regina, Regina, Saskatchewan, S4S 0A2, Canada, 2001
68. J.D. Katzberg, W. Ziarko, Variable precision extension of rough

sets; Fundamenta Informaticae, vol. 27, no. 2-3, pages 155-168, 1996
69. Y. Shen, F. Wang, Variable precision rough set model over two universes and its properties, Soft Computing, Volume 15, Issue 3, pp 557-567, 2011
70. X. Zhang, Z. Mo, F. Xiong, W. Cheng, Comparative study of variable precision rough set model and graded rough set model, International Journal of Approximate Reasoning, Volume 53, Issue 1, pp 104–116, 2012
71. I.T.R. Yanto, P. Vitasari, T. Herawan, M.M. Deris, Applying variable precision rough set model for clustering student suffering study's anxiety, Expert Systems with Applications, Volume 39, Issue 1, Pages 452–459, 2012
72. M. Ningler, G. Stockmanns, G. Schneider; H.D. Kochs, E. Kochsa, Adapted variable precision rough set approach for EEG analysis, Artificial Intelligence in Medicine, ELSEVIER, Volume 47, Issue 3, Pages 239–261, 2009
73. M. J. Beynon, N. Driffield, An illustration of variable precision rough sets model: an analysis of the findings of the UK Monopolies and Mergers Commission, Computers & Operations Research, ELSEVIER, Volume 32, Issue 7, pp 1739–1759, 2005
74. X. Pan, S. Zhang, H. Zhang, X. Na, X. Li, A variable precision rough set approach to the remote sensing land use/cover classification, Computers & Geosciences, ELSEVIER, Volume 36, Issue 12, pp 1466–1473, 2010
75. H.Y. Zhang, Y. Leung, L. Zhou, Variable-precision-dominance-based rough set approach to interval-valued information systems, Information Sciences, ELSEVIER, Volume 244, pp 75–91, 2013
76. B.Q. Hu; Generalized interval-valued fuzzy variable precision rough sets determined by fuzzy logical operators, International Journal of General Systems, Volume 44, Issue 7-8, 2015
77. Z. Pawlak, A. Skowron, Rough sets: Some extensions, Information Sciences, ELSEVIER, Volume 177, Issue 1, pp 28–40, 2007
78. W. Ziarko, Rough Sets, Fuzzy Sets and Knowledge Discovery, Springer Verlag, pp 326-334, 1994
79. UCI Machine Learning Repository, http://cml.ics.uci.edu, last accessed: 10/01/2017
80. J. Liu, H. Deng. Outlier detection on uncertain data based on local information. Knowledge-Based Systems. Vol 51, 2013.