JOURNAL OF INFORMATION SCIENCE AND ENGINEERING 36, 671-685 (2020)
DOI: 10.6688/JISE.202005_36(3).0012

A Method Non-Deterministic and Computationally Viable for Detecting Outliers in Large Datasets

ALBERTO FERNÁNDEZ OLIVA[1], FRANCISCO MACIÁ PÉREZ[2],
JOSÉ VICENTE BERNÁ MARTINEZ[2] AND MIGUEL ALFONSO ABREU ORTEGA[3]
[1]*Department of Computer Science*
University of Havana
Havana, 10100 Cuba
E-mail: afdez@matcom.uh.cu
[2]*Department of Information Technology and Computer Science*
University of Alicante
Alicante, 03690 Spain
E-mail: {pmacia; jvberna}@dtic.ua.es
[3]*Georgia Institute of Technology*
Atlanta, GA 30332 USA
E-mail: mabreu@gatech.edu

This paper presents an outlier detection method that is based on a Variable Precision Rough Set Model (VPRSM). This method generalizes the standard set inclusion relation, which is the foundation of the Rough Sets Basic Model (RSBM). The main contribution of this research is an improvement in the quality of detection because this generalization allows us to classify when there is some degree of uncertainty. From the proposed method, a computationally viable algorithm for large volumes of data is also introduced. The experiments performed in a real scenario and a comparison of the results with the RSBM-based method demonstrate the efficiency of both the method and the algorithm in diverse contexts that involve large volumes of data.

Keywords: outliers, rough sets (RS), RS basic model (RSBM), variable precision rough set model (VPRSM), data set, data mining.

1. INTRODUCTION

From the perspective of Knowledge Data Discovery and Data-Data Mining (KDD-DM), *outliers* usually represent undesirable objects that must be addressed or eliminated in the data preparation phase, to not hinder the detection of reliable patterns. However, for some applications, these objects are even more representative and interesting than the most common events. Some of these applications would be, for example, credit card fraud detection, where outlier detection could provide information for examining patterns of misconduct; or electronic business data analysis, where outlier detection could be useful for Customer Relationship Management.

KDD-DM processes require increasingly effective methods for outlier detection. The current data sets include ever more bulky and sophisticated data, representation structures, and data storage means.

After examining the current state of the art [1], it is concluded that the scope of application is wide and diverse, where the nature of the data and the spaces in which they

Received April 30, 2018; revised July 22 & August 21, 2018; accepted September 6, 2018.
Communicated by Tzung-Pei Hong.

are defined acquire very different characteristics. This circumstance has led to the development of a variety of detection methods that are appropriate to each problem. The challenge is to devise increasingly flexible detection methods that can be valid in different environments.

In this paper, we rely on the non-deterministic nature of *VPRSM* [2], based on the fuzzification of the set inclusion concept, which allows managing certain thresholds that are established by the user. From this idea, we propose an extension of the theoretical *VPRSM* to create a new algorithm for outlier detection. This algorithm shows remarkable improvements in its capacity for generalization and detection but maintains the levels of spatial and temporal complexity that make it viable in practice.

The remainder of the article is structured as follows: In Section 2, the most significant aspects of the state of the art and the background of this proposal are discussed. In Section 3, a *VPRSM*-based theoretical framework is constructed, and an algorithm is proposed for outlier detection based on this computationally viable method, which admits some degree of uncertainty or misclassification. In Section 4, the results are validated by various experiments that show an improvement in both the quality of detection and the computational viability of the algorithm, and in addition, Section 4 presents a comparison with other methods. Finally, in Section 5, the main conclusions of the study are presented together with future lines of research.

2. STATE OF THE ART AND BACKGROUND

The *outlier detection* problem is currently gaining more and more importance in multiple and diverse contexts [3], where some noteworthy examples are credit card or cellular phone fraud detection; identification of conflicting users in the assessment of bank loan applications; computer network intruder detection; computer network traffic monitoring; diagnosis of faults or malfunctions in the operation of motors, generators, pipes, and measuring instruments; detection of structural defects; automated control of manufacturing lines to detect faulty batches; automated monitoring of medical parameters and the identification of new molecular structures.

Statistical models were the first to address the outlier detection problem. However, current artificial intelligence (AI) techniques are giving better results [4] in environments where increasingly large volumes of categorical data must be processed. In [5, 6], and [7], an extensive set of outlier detection methods based on different algorithms is described. In general, outlier detection algorithms or methods are classified according to the technique on which they are based. The most important techniques are based on distributions, different depth criteria, distances, densities, *clustering*, neural networks, or machine learning.

By increasing the size of the data set, the effectiveness of certain algorithms can be seriously compromised. For example, the concept of distance in a space of different dimensions varies. The temporal complexity of most distance-based methods is quadratic, which causes an increase in the processing time and a distortion of the processing time distribution. In the context of data mining, the distribution of attribute values is almost always unknown, which affects those methods that are based on data proximity. By greatly increasing the size of the data set, it is very difficult to estimate its multidimen-

sional distributions [5].

In recent years, proposals have emerged from which efficient algorithms can be constructed for outlier detection based on *Rough Set* Theory [8-10]. The use of Rough Sets produces new problems such as computational viability, which is why many works tackle the question of generating more efficient algorithms [11] exploiting the mathematical qualities of the Rough Sets theory [12, 13]. Some works delve into the structuring of the data and pre-detection processes to optimize the algorithms [14], in the use of the knowledge that we have about data to combine rough sets with other methods [15, 16] or in the characteristics of the data set to be analyzed [17] and thus make the detection more efficient. In other works they talk about achieve an approach to fuzzy logic in terms of classification [18] which opens other lines of work to achieve efficient and effective classification algorithms [19].

In [20], an alternative approach to *Rough Set* Theory is proposed, which constitutes a novel approach to the outlier detection problem compared to that based on *RSBM*. Under this approach, the outliers are defined as elements of **non-redundant exceptional sets** that have a greater **degree of marginality** than an established threshold. The fundamental contribution of *RS* theory is to facilitate the analysis of the classification. The approximation (upper and lower) becomes necessary because of the inability to establish, with the available knowledge, a complete classification of objects that belong to a certain category [21].

With some frequency, the available information allows only a partial classification, and *RS* theory can be used effectively to model this type of classification, but according to this theory, this classification must be completely correct or certain [21]. This arrangement limits the possibility of producing a classification result that has a controlled degree of uncertainty, in other words, the possibility that there is a certain specific amount of error in the classification. This arrangement is not what occurs with *RSBM*. Paradoxically, in practice, and in many cases, it is convenient to admit some degree of uncertainty in the classification process, which can allow better understanding and use of the properties of the data that is being analyzed.

Another limitation ascribed to *RSBM* is that it assumes that the universe U of objects or data under consideration is known and that all conclusions derived from the application of the above-mentioned model are applicable only to that set of objects. However, in practice, there is a need to generalize the conclusions that are obtained from a small set of objects (U) to a larger universe, for example, the real world. *RSBM* allows us to obtain hypotheses based only on error-free classification rules (which are expressed in the lower approximation, $\underline{X}$) obtained from the analysis of the data involved (U); in other words, this model is deterministic. However, there are multiple situations in the real world that require accounting for a partially incorrect classification. A partially incorrect classification rule also provides useful information. It can establish the trends in the values if most of the available data to which the rule applies can be classified correctly.

A generalization of *RSBM* is *VPRSM* [2]. This model overcomes the deterministic nature (in terms of classification) based on a very simple idea that allows us to manage certain thresholds that are established by the user: the fuzzification of set inclusion. *VPRSM* offers the possibility of detecting or establishing this information trend and, based on it, performing some analyses on a particular universe of objects or data; in other words, it is a statistical model.

The main objective of this research is to create a *VPRSM*-based non-deterministic outlier detection method that is computationally viable and that is based on the hypothesis that *VPRSM* allows us to extend the application of the original *RSBM*-based method [22] to contexts in which there is a need for classification with some degree of uncertainty.

3. *VPRSM*-BASED OUTLIER DETECTION PROPOSAL

3.1 Detection Method based on the Properties of *VPRSM*

In this section, we will construct the new *outlier* detection method as we analyze and propose the mathematical tools that are needed and that arise from *VPRSM*.

It is evident that no classification can be performed based on the definition of a standard inclusion relation. The first step to overcome the limitations imposed by *RSBM* is to free ourselves from the need to explicitly define the universal quantifier. The *measure of the degree of misclassification* proposed in *VPRSM* will allow us to do so: the measure of the relative degree of misclassification of the set X with respect to set Y, $c(X, Y)$. In other words, the relative error present when classifying a set of objects is defined as

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

This definition is evident because we can observe 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 $c(X, Y) \approx 1 \Rightarrow X, Y$ come close to being disjoint sets; and
– if $c(X, Y) = 1 \Rightarrow |X\cap Y| = 0 \Rightarrow X, Y$ are disjoint.

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

If the relative degree of misclassification is taken as the basis measure, then the inclusion relation can be redefined, which avoids having an explicit definition of the universal quantifier, such as: $X\subseteq Y \Leftrightarrow c(X, Y) = 0$. According to this definition, $c(X, Y)$ can take values that are greater than 0 (without being *too large*) in a case in which the relationship would represent a *majority*. In other words, it is necessary that a *majority* of objects in X be classified in Y. It is obvious that the concept of *majority* requires establishing a threshold. 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. The definition of the inclusion relation is redefined by adding the specification of an admissible error limit in the classification.

Definition 1: *majority inclusion relation*: Let U be a finite universe of objects. Let β, $0 \leq \beta < 0.5$, the admissible misclassification error. Let $X, Y\subset U, X\neq\phi, Y\neq\phi$. We say that X is

for the most part included in Y, or that X is included in Y with a β-*error*, $X \subseteq^{\beta} Y$, if and only if $c(X, Y) \leq \beta$. From the same definition, it can be seen that $\beta = 0$ expresses a *standard inclusion relation*, which is called, in this model, total inclusion.

Based on this new definition of an inclusion relation, the most representative concepts of *RSBM* are redefined:

Definition 2: Let X be an arbitrary subset of the universe U. Let $\Phi \subseteq U \times U$ be an equivalence relation that divides U into a finite set of equivalence classes $\langle x \rangle_{\Phi}$. Let us define the following:

a) $\underline{X}_{\beta} = \cup \{\langle x \rangle_{\Phi} : \langle x \rangle_{\Phi} \subseteq^{\beta} X\}$, and it is known that $\langle x \rangle_{\Phi} \subseteq^{\beta} X \Leftrightarrow c(\langle x \rangle_{\Phi}, X) \leq \beta$;
b) $\overline{X}_{\beta} = \cup \{\langle x \rangle_{\Phi} : \langle x \rangle_{\Phi} \not\subset^{\beta} X^{c}\}$. Therefore, $\langle x \rangle_{\Phi} \not\subset^{\beta} X^{c} \Leftrightarrow c(\langle x \rangle_{\Phi}, X) < 1 - \beta$;
c) BN_{β}(β-boundary region) = $\overline{X}_{\beta} - \underline{X}_{\beta}$;
d) B^{β}(β-internal boundary region) = $X \cap BN_{\beta}$; and
e) NEG_{β}(β-negative region) = $U - \overline{X}_{\beta}$.

RSBM is a specific case of *VPRSM*: the representative regions of both models correspond to a classification error $\beta = 0$. However, these regions vary if a certain classification error is allowed. Note also that the β-*negative* region of X is the union of all equivalence classes that can be classified within X^{c}, with an error in the classification no greater than β.

Bearing in mind that when $\beta = 0$ the standard *RS* model [23] is a specific case of *VPRSM*, the following proposition is established, whereby other relationships that are also fulfilled are expressed.

Proposition:
a) $\underline{X} \subseteq \underline{X}_{\beta}$: lower approximation is a proper subset of the β-lower approximation.
b) $X_{\beta} \subseteq \overline{X}$: β-upper approximation is a proper subset of the upper approximation.
c) $BN_{\beta} \subseteq BN$: the β-boundary region is a proper subset of the boundary region.
d) $NEG \subseteq NEG_{\beta}$: the negative region is a proper subset of the β-negative region.

Intuitively, it can be seen how, by decreasing the classification error β, the size of the positive and negative regions of X decreases while that of the boundary region increases.

In conclusion, based on the concept of majority inclusion defined in *VPRSM*, we have devised a new outlier detection method that allows performing classifications with some degree of error when calculating the significant regions; in other words, it allows an *almost* complete classification.

3.2 Outlier Detection Algorithm

From the method proposed in the previous section, in this section, a new algorithm is conceived that improves the quality of detection and provides a greater range of application. Likewise, this algorithm maintains the levels of spatial and temporal complexity that ensure its viability in real environments in which we must work with large volumes of data.

To design the new algorithm, we started from the *RSBM* algorithm, which was already tested and validated in [13], but we added the admissible *β-error* to the input. Therefore, the inputs are the following: the universe U, the dataset X (sub-set of objects of the Universe that constitute the study set or interest), the criteria that distinguish the equivalence relations R:(r_i, $1 \leq i \leq m$) to be accounted for in the analysis, the value of the established detection threshold μ, and the admissible *β-error* in the classification. The same data structures as those described for the original algorithm [22], based on *RSBM*, are maintained.

The fundamental data structure used in the algorithm is that of a *dictionary*, where a dictionary is a set of pairs (*key*, *value*), where *key* is any object to which only one *value* object is associated. In the algorithm, the key*s* are obtained by applying a classifier to any element of the universe. This classifier is associated with a particular equivalence relation r_i, with $1 \leq i \leq m$, and allows us to classify the members of the equivalence classes that are defined by the relation. The *values* that are associated with the *keys* are lists of elements that belong to the equivalence class identified by the *key* associated with said *value*. A dictionary is built for each equivalence relation, and from all of them, a list of dimension m is formed, where m is the number of equivalence relations under consideration. According to the data structures used, it can be said that the spatial complexity of the algorithm is $O(n \times m)$ because each dictionary, at most, can contain all of the elements (n) of the universe.

The algorithm consists of two steps: (1) the formation of *β-internal boundaries* and (2) the outlier detection process itself. Next, each of these steps is shown and analyzed from its pseudo-code.

Step 1: Formation of *β-internal boundaries*: the classifiers are applied, one for each equivalence relation that is accounted for in the analysis, to the elements of U to form the *β-internal boundaries*.

Algorithm 1: Pseudo-code of the BUILD_REGIONS algorithm, representing step 1 for the formation of the *β-internal boundaries.*

```
BUILD_REGIONS (U, X, R, β): B
1  for each r ∈R
2    P_r = CLASSIFY-ELEMENTS (U, r) // P_r is the partition induced by the equivalence relation <r>
3    for each class ∈ P_r
4        if c(class, X) ≤ β
5            X_β = X_β ∪ class              // By proposition 4a: class ⊆^β X ⇒ class ∈ X_β
6        else if c(class, X) ≥ 1− β
7            NEG_β = NEG_β ∪ class          // By proposition 4b: class ⊄^β X ⇒ class ∈ NEG_β
8        else
9            B_r^β = B_r^β ∪ (class ∩ X)    // By proposition 4c: (class ∩ X) ⊆ B_r^β add the elements of <class>
                                            // that meet the <dataset X> to the internal boundary relative to <r>
10 return B
```

The temporal complexity of this step is $O(n \times m \times c)$, where c is the cost of classifying each element, n is the cardinality of the universe, and m is the number of equivalence relations that are accounted for in the analysis. The cost of classifying each element "c", is constant, since the cost of classifying an element from an equivalence relation is constant.

Step 2: *Outlier* detection process: the set of OUTLIERS that contain all of the elements that meet the dataset X and could be outlier candidates is calculated. Of these, all of those whose *degree of exceptionality* is greater than the established *detection threshold* μ are classified as such.

Algorithm 2: Pseudo-code of the VPRS_OUTLIER_DETECTION algorithm that implements step 2: outlier detection procedure.

```
VPRS_OUTLIER_DETECTION (U, X, R, β, μ): OUTLIERS
1   B = BUILD_REGIONS (U, X, R, β)                  // Step 1
2   for each r ∈ R                                  // For each equivalence relation <r>
3       contains Another = FALSE                    // There is no internal boundary subset of B_r^β
4       for each q ∈ R – {r}                        // For each equivalence relation<q>different from <r>
5           if B_r^β ⊂ B_r^β                        // If the internal boundary of <q> is a subset of the
                                                    // internal boundary of <r>, then its elements are discarded
                                                    // as members of the set of possible OUTLIERS: E
6               contains Another = TRUE
7               break                               // It is no longer necessary to continue
8       if not contains Another                     // If no internal boundary is a subset of the one analyzed,
                                                    // all of the elements of the internal boundary of <r>
9           E = E ∪ B_r^β                           // members of the set of possible OUTLIERS: E
10  for each e ∈ E
11      if EX-DEGREE(e) ≥ μ                         // The elements of E above the exceptionality threshold
12          OUTLIERS = OUTLIERS ∪ {e}               // belong to the set of outliers
13  return OUTLIERS
```

The temporal complexity of step 2 is $O(n \times m^2)$. When accounting for steps 1 and 2, the computational cost of the whole algorithm is $O(max(O(step\ 1), O(step\ 2))) = O(step\ 2) = \boldsymbol{O(n \times m^2)}$.

In general, the number of equivalence relations that occur in the analysis, in the vast majority of cases, is not very large in relation to the number of *data set* elements. For this reason, the quadratic dependence of the execution time with respect to the number of equivalence relations does not affect, to a large extent, the execution time of the algorithm. As will be seen later in the results obtained, this quadratic dependence is almost linear for small values ($m \leq 20$).

Because the data structures described for the *RSBM*-based algorithm are maintained, the same order $O(n*m)$ is also maintained for the spatial complexity. The method is applicable to data in the form of a table, with non-redundant data and mono-valued attributes, and data can be both continuous and categorical.

4. VALIDATION OF THE RESULTS

The fundamental objectives of the experiments are (1) to validate the incorporation of variable precision into the outlier detection algorithm to improve the results obtained; (2) given the large volumes of data used in this type of problem, to verify that the temporal complexity of the algorithm remains linear in practice; and (3) to compare the results obtained with other methods, algorithms, and strategies.

To achieve this last objective (3), experiments (1) and (2) were performed using the data set provided by the *UCI Machine Learning Repository* of the *Center for Machine*

Learning and Intelligent Systems of the University of California, Irvine [24], which contains the data extracted from the *US Census Bureau Database*. This data set has already been used in more than 50 different scientific articles and is therefore endorsed as a reference data set.

4.1 Experiments to Determine the Quality of Detection

The purpose of this test is to demonstrate that the method is valid, to show the variation in the quantity of *detected outliers* as the value of the threshold is varied (β and μ), and to compare the *RSBM* ($\beta = 0$) and *VPRSM* ($\beta \neq 0$) algorithms. To demonstrate that the method is valid in terms of its detection capacity on real data sets, a set of tests were designed, while defining a dataset X for the data set under study and a series of equivalence relations and intentionally introducing a set of outliers. Afterward, we use the proposed method to detect outliers and analyze the results. The elements defined are the following:

– Individuals in the data set that were studied are those that met the following criterion: *DATASET X*: 1 ≤ subjects_aged ≤ 10.
– The criteria for conducting the analysis were established by the following equivalence relationships

```
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 'han-
         dlers-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']
```

Code 1. Equivalence relationship for analysis of data set.

Therefore, any element that meets the dataset *X* and belongs to class $c_x.1$ (x = 1, 2, 3, 4) is contradictory for the relation r_x because the individuals under analysis are children between 1 and 10 years old.

A set of 13 outliers with contradictory values in different fields of each record is intentionally introduced in the data set, which were always for children aged 1-10 years. In the set of introduced outliers, the level of contradiction of individuals varies. In some cases, they are contradictory for one or two attributes, while in others they are contradictory for three or four, and these are precisely the most contradictory elements.

Table 1 shows the number of detected outliers for different thresholds β (misclassification error) and μ (degree of outlier). The results that correspond to *RSBM* are exactly

those achieved for $\beta = 0$. The values set were $\beta = 0.10$; 0.20; 0.30; 0.40; 0.50; this approach establishes admissible error values in the classification and therefore corresponds to *VPRSM*.

Table 1. Basic *RS* algorithm (*RSBM*) vs. *VPRSM* for detecting outliers.

	RSBM	*VPRSM*				
	$\beta = 0$	$\beta = 0.1$	$\beta = 0.2$	$\beta = 0.3$	$\beta = 0.4$	$\beta = 0.5$
$\mu = 0.2$	24	10	6	6	0	0
$\mu = 0.4$	24	10	3	0	0	0
$\mu = 0.6$	14	4	2	0	0	0
$\mu = 0.8$	9	3	1	0	0	0
$\mu = 1$	9	1	0	0	0	0

When interpreting the results, it should be noted first that in all cases, within the set of detected outliers, some of those intentionally introduced into the data set were always found. When the number of detected outliers was greater than the number of introduced *outliers*, those detected included all introduced. When the number of detected outliers was lower than the number of introduced outliers, then of them, those detected were always the most contradictory. For example, with $\mu = 0.02$ and $\beta = 0.0$, 24 outliers were detected, and among them were the 13 that were introduced, and with $\mu = 0.6$ and $\beta = 0.2$, although only four outliers were detected, two of them belong to the set of the 13 introduced, especially the two most contradictory because they were so for the four attributes under consideration.

The interpretation of the tests performed also allows us to draw the following conclusions:

An adequate choice of equivalence relations or classification criteria ensures effectiveness in the detection.

For small values of the parameters μ and β, the number of detected outliers is usually high, and elements that really are not outliers are identified as such. For example, for $\mu = 0.2$ and $\beta = 0.0$, 24 outliers were detected. This finding reaffirms an important aspect of the statistical view of the outlier detection problem for identifying a case as exceptional: when the candidate observations to be considered as such have been identified by some method, the researcher must make an analysis of those results to select those observations that show a real contradiction with respect to the sample studied.

By successively increasing the detection threshold (μ), a refinement of the result is achieved. In general, the higher the threshold is, the smaller the number of detected outliers. In addition, those that remain are contradictory for a greater number of attributes. However, in some cases and for certain variations of the value of μ, such a refinement is not achieved. For example, by varying the value of μ from 0.2 to 0.4, with $\beta = 0.0$, the number of detected outliers is in both cases 24. The same circumstance occurs when μ varies from 0.8 to 1.0, with $\beta = 0.0$. In both cases, again, the number of detected outliers was nine. Note the value of $\beta = 0.0$ in both examples, which implies that no degree of misclassification has been allowed. Therefore, the results are referred to as *RSBM.* Note also that by allowing a certain degree of misclassification (values of $\beta \neq 0.0$) for the same variations of the values of μ referred to in the previous example, the quantity of detected results is different.

Another element to note is that once μ reaches the highest possible value ($\mu = 1.0$), the number of detected outliers is nine. However, again, by making variations in the value of β, a greater detection refinement is achieved, detecting only the most contradictory elements. This finding shows that controlled and progressive misclassification (β) improves the quality of detection. Nevertheless, one should be cautious when varying β because a high degree of misclassification can imply that all of the elements in the boundary pass to the positive or negative regions, leaving the internal boundaries without elements. In the tests performed, for example, this arrangement is evident from $\beta = 0.3$.

4.2 Experiments to Determine the Viability of the Algorithm

To observe the performance of the algorithm, its behavior is analyzed by varying all of the parameters that define the size of the input, that is, the number of rows and columns of the data set and the number of equivalence relationships that are accounted for in the analysis. To contextualize the results, the results are compared with the *RSBM* algorithm.

Fig. 1 (a) shows the execution of the algorithm when the number of equivalence relations (5 relationships) and the number of rows of the data set (30,000 rows) are constant and the number of columns in the data set to process is varied (from 5 to 14 columns). Fig. 1 (b) shows the same experiment but keeping the number of rows (30,000

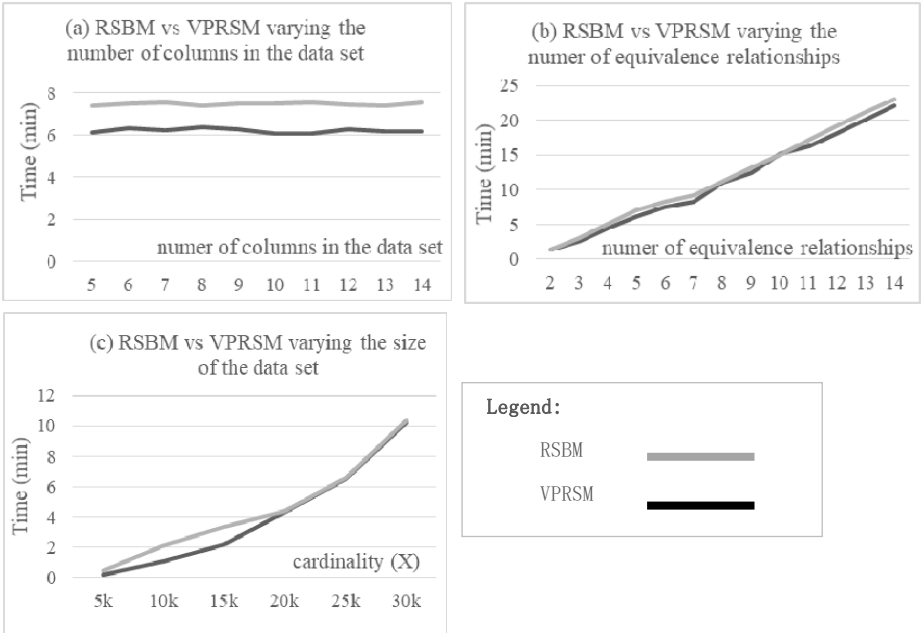


Fig. 1. (a) Execution time of *RSBM* vs. *VPRSM* by varying the number of columns of the data set and keeping the number of rows and equivalence relationships constant; (b) Execution time of *RSBM* vs. *VPRSM* by varying the number of equivalence relations and keeping the number of rows and columns of the data set constant; (c) Execution time of *RSBM* vs. *VPRSM* by varying the number of rows of the data set and keeping the number of columns and equivalence relations constant.

records) and columns (14 columns) of the data set constant and varying the number of equivalence relations (from 2 to 14 relationships). It can be observed that the cost is almost linear. Finally, in Fig. 1 (c), the experiment was performed while keeping the number of columns of the data set (14 columns) and the number of equivalence relations (5 equivalence relations) constant and varying the cardinality of the data set from 5,000 to 30,000 rows.

The results confirm that the levels of temporal complexity in the execution correspond to those of the algorithm that were justified from a theoretical point of view. The results also show that level constants are reasonable, which allows the algorithm to be used with realistic data sets. Finally, the execution times for both versions of the algorithm, *RSBM* and *VPRSM*, do not differ significantly, and thus, the new capacity of variable precision can be incorporated into the basic classification system.

4.3 Comparison with Other Outlier Detection Algorithms

Most state-of-the-art techniques and algorithms for detecting outliers are conceived, to a greater or lesser extent, to solve a certain type of problem, even for a specific case. It is difficult to make valid comparisons between these algorithms because they will depend strongly on what the user is looking for. However, it would be interesting to conduct a comparative study of the different existing methods involving the advantages presented by the current proposal in the area in which it is useful – to provide, in an unsupervised manner, general results with respect to all the elements of the *universe* of the data, by determining some initial conditions: *dataset X* and *equivalence relations*. When accounting for these concerns, Table 2 details how the created algorithm can overcome the limitations of the methods studied when generalization is required.

Table 2. Characteristics of the proposed method against the limitations of conventional methods.

Against STATISTICAL METHODS and DISTANCE-BASED METHODS
• Applicability to data sets with a mix of continuous and discrete attributes. In some cases equivalence relations are useful to discretize continuous data [25]. • It is not necessary to know the data distribution nor establish a distance criterion on them. • The problem of temporal complexity, quadratic order, which most distance-based methods present, is solved. • The dimensionality and size of the data set is not a limitation for the algorithm's execution.
Against DENSITY– and DEPTH-BASED METHODS
• It is not necessary to establish criteria on the *density* of the data in the data set. • The dimensionality of the data set is not a limitation for the algorithm's execution. • It is not necessary to perform previous calculations that consume a large amount of time, such as the calculation of the *convex envelope*, which is necessary in most *depth*-based methods. • RSBM allow obtaining the results in an unsupervised way, without the need for the user to establish, as a step prior to their execution, the value of certain parameters that are involved in the analysis, which is necessary in *density*-based methods such as

DBSCAN.
- *RSBM* and *VPRSM* represent improvements in terms of the temporal complexity compared to *depth*-based methods.

Against NEURAL NETWORK BASED METHODS

- It is not necessary to establish previous processes that consume a large amount of time, such as training the network, which is necessary in some neural network models to ensure their learning.
- The dimensionality of the data set is not a limitation for the algorithm's execution.
- Algorithm functionality does not depend on data *density* criteria, as is required in some supervised models.
- It is not necessary to model the data *distribution*, as required in some supervised models.
- Some approaches based on supervised networks establish the use of thresholds for various purposes in the outlier detection process. This concern is resolved from the conception of the *RSBM* algorithm.

The great advantage of the proposed VPRS_OUTLIER_DETECTION algorithm is its generalist nature. Of course, an algorithm that is specifically designed to detect a specific type of outlier is usually better, both in the quality of detection and in the spatial and temporal complexity. However, having a generic algorithm that is capable of addressing different types of problems and different types of data and capable of behaving reasonably with large volumes of data is a very interesting option that avoids having to design different algorithms each time that we face new problems or when the conditions of the problems that are already solved change.

5. CONCLUSIONS

The results obtained from the tests performed show that the proposed *VPRSM*-based algorithm eliminates the deterministic character, in terms of classification, which limits the *RSBM*-based algorithm. This improvement results in greater detection accuracy, identifying only the most contradictory elements as outliers.

Additionally, the *VPRSM*-based algorithm achieves this improvement while maintaining the same level of temporal and spatial complexity of the *RSBM*-based algorithm. As has been demonstrated, our method provides a computationally efficient solution, offering the possibility of using quasi-linear algorithms, which is an advantage over the usually high procedural complexity in the field of *KDD-DM* and the enormous volume of the data sets.

In the medium term, our research is aimed at the creation of a tool that allows us to predict probabilistically the condition of the outliers for all of the elements of a given data set, while this prediction is computationally viable. To achieve this goal, we are working on the creation of an algorithm that is capable of automatically calculating the thresholds μ and β, which are involved in the proposed method (currently, they must be defined by the user). From this algorithm, our research will focus on the creation of a new method to determine the set of threshold values under which a certain element of a data set would be an outlier.

ACKNOWLEDGEMENTS

This work has been supported by grant TIN2016-78103-C2-2-R, and University of Alicante projects GRE14-02 and Smart University.

REFERENCES

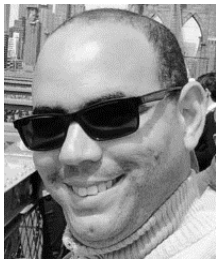
1. A. Fernández, "Estimación probabilística del grado de excepcionalidad de un elemento arbitrario en un conjunto finito de datos. Aplicación de la Teoría de Conjuntos Aproximados de Precisión Variable (VPRSM)," Ph.D. Dissertation, Department of Information Technology and Computing, University of Alicante, 2010.
2. W. Ziarko, "Probabilistic decision tables in the variable precision rough set model," *Computational Intelligence*, Vol. 17, 2001, pp. 593-603.
3. I. Ben-Gal, "Outlier detection," *Data Mining and Knowledge Discovery Handbook*, Springer, Berlin, 2010, pp. 117-130.
4. G. Buzzi-Ferraris and F. Manenti, "Outlier detection in large data sets," *Computers & Chemical Engineering*, Vol. 35, 2011, pp. 388-390.
5. V. J. Hodge and J. Austin, "Survey of outlier detection methodologies," *Artificial Intelligence Review*, Vol. 22, 2004, pp. 85-126.
6. M. Last and A. Kandel, "Automated detection of outliers in real-world data," in *Proceedings of the 2nd International Conference on Intelligent Technologies*, 2001, pp. 292-301.
7. J. Tang, Z. Chen, A. Fu, and D. Cheung, "A robust outlier detection scheme in large data sets," in *Proceedings of the 6th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2002, pp. 6-8.
8. F. Jiang, Y. Sui, and C. Cao, "Some issues about outlier detection in rough set theory," *Expert Systems with Applications*, Vol. 36, Part 1, 2009, pp. 4680-4687.
9. F. Shaari, A. A. Bakar, and A. R. Hamdan, "Outlier detection based on rough sets theory," *Journal of Intelligent Data Analysis*, Vol. 13, 2009, pp. 191-206.
10. Z. Xue, Y. Shang, and A. Feng, "Semi-supervised outlier detection based on fuzzy rough C-means clustering," *Mathematics and Computers in Simulation*, Vol. 80, 2010, pp. 1911-1921.
11. F. Jiang and Y.-M. Cheng, "Outlier detection based on granular computing and rough set theory," *Applied Intelligence*, Vol. 42, 2015, pp. 303-322.
12. S. Liu, J. Wang, and G. Xing, "The review of outlier mining based on granular computing," in *Proceedings of IEEE International Conference on Granular Computing*, 2008, pp. 462-465.
13. F. Jiang, G. Liu, J. Du, and Y. Sui, "Initialization of K-modes clustering using outlier detection techniques," *Information Sciences*, Vol. 332, 2016, pp. 167-183.
14. X. Ahang, J. Dai, and Y. Yu, "On the union and intersection operations of rough sets based on various approximation spaces," *Information Sciences*, Vol. 292, 2015, pp. 214-229.
15. H. Hannah, A. Taher, and G. Jothi, "Supervised hybrid feature selection based on PSO and rough sets for medical diagnosis," *Computer Methods and Programs in Biomedicine*, Vol. 113, 2014, pp. 175-185.

16. S. Maldonado, G. Peters, and R. Weber, "Credit scoring using three-way decisions with probabilistic rough sets," *Information Sciences*, Vol. 507, 2018, pp. 700-714.
17. J. Zhang, T. Li, and H. Chen, "Composite rough sets for dynamic data mining," *Information Science*, Vol. 257, 2014, pp. 81-100.
18. D. Liang and D. Liu, "A novel risk decision making based on decision-theoretic rough sets under hesitant fuzzy information," *IEEE Transactions on Fuzzy Systems*, Vol. 23, 2015, pp. 237-247.
19. X. Ma, Q. Liu, and J. Zhan, "A survey of decision making methods based on certain hybrid soft set models," *Artificial Intelligence Review*, Vol. 47, 2017, pp. 507-530.
20. F. Jiang, Y. Sui, and C. Cao, "Outlier detection using rough sets theory," in *Proceedings of the 10th International Conference on Rough Sets*, *Fuzzy Sets*, *Data Mining*, *and Granular Computing*, Vol. Part II, 2005, pp. 79-87.
21. Z. Pawlak and A. Skowron, "Rudiments of rough sets," *Information Sciences*, Vol. 177, 2007, pp. 3-27.
22. F. Maciá, J. V. Berna, A. Fernández, and M. A. Abreu, "Algorithm for the detection of outliers based on the theory of rough sets," *Decision Support Systems*, Vol. 75, 2015, pp. 63-75.
23. Z. Pawlak, "Rough sets," *International Journal of Computer and Information Sciencies*, Vol. 11, 1982, pp. 341-356.
24. C. Meek, B. Thiesson, and D. Heckerman, "US census data," *UCI Machine Learning Repository*, 1990, http://cml.ics.uci.edu.
25. F. Jiang and Y. Sui, "A novel approach for discretization of continuous attributes in rough set theory," *Knowledge-Based Systems*, Vol. 73, 2015, pp. 324-334.

Alberto Fernández Oliva was born in Cuba in 1955. He received his Bachelor and Master degree in in Computer Science from the Havana University in 1979 and 1997 respectively. He received his Ph.D. degree in Computer Science from the University of Alicante in 2010. Full Professor of Computer Science Department at Havana University. His research interests since 2007 are in the areas of data mining and knowledge discovery on data (outlier detection methods).

Francisco Maciá Pérez was born in Spain in 1968. He received his engineering degree and the Ph.D. degree in Computer Science from the University of Alicante in 1994 and 2001 respectively. He worked as System's Administrator at the University of Alicante from 1996 to 2001. He was an Associate Professor from 1997 to 2001. Since 2001, he is a Professor and currently he is the Vice President for Information Technologies at the University of Alicante. His research interests are in the areas of network management, computer networks, smart sensor networks and distributed systems, which are applied to industrial problems.

José Vicente Berná Martinez was born in Spain in 1978. He received his engineering degree and the Ph.D. degree in Computer Science from the University of Alicante in 2004 and 2011 respectively. From 2006 to 2013, he was an Assistant Professor at the University of Alicante, currently he is an Associate Professor. His research interests are in the areas of computer networks, distributed systems, bio-inspired systems and robotics which are applied to industrial problems.



Miguel Alfonso Abreu Ortega was born in Cuba in 1987. He is graduated with honors in Computer Science at Havana University. He has been working in subjects relative to data mining and knowledge data discovery since 2007. He was a training professor at Havana University.