IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 1, 2020, accepted October 18, 2020, date of publication October 30, 2020, date of current version November 16, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3034885

Exceptional in so Many Ways–Discovering Descriptors That Display Exceptional Behavior on Contrasting Scenarios

JOSÉ MARÍA LUNA[1], MYKOLA PECHENIZKIY[2], WOUTER DUIVESTEIJN[2], AND SEBASTIÁN VENTURA[1], (Senior Member, IEEE)

[1]Department of Computer Science and Numerical Analysis, University of Cordoba, 14071 Córdoba, Spain
[2]Department of Computer Science, Eindhoven University of Technology, 5600 Eindhoven, The Netherlands

Corresponding author: Sebastián Ventura (sventura@uco.es)

This work was supported in part by the Spanish Ministry of Economy and Competitiveness, under Project TIN2017-83445-P, in part by the FEDER funds, and in part by the UCO-FEDER under Project 18 REF.1263116 MOD.A.

ABSTRACT The current state of the art in supervised descriptive pattern mining is very good in automatically finding subsets of the dataset at hand that are exceptional in some sense. The most common form, subgroup discovery, generally finds subgroups where a single target variable has an unusual distribution. Exceptional model mining (EMM) typically finds subgroups where a pair of target variables display an unusual interaction. What these methods have in common is that one specific exceptionality is enough to flag up a subgroup as exceptional. This, however, naturally leads to the question: can we also find multiple instances of exceptional behaviour simultaneously in the same subgroup? This paper provides a first, affirmative answer to that question in the form of the SPEC (Subsets of Pairwise Exceptional Correlations) model class for EMM. Given a set of predefined numeric target variables, SPEC will flag up subgroups as interesting if multiple target pairs display an unusual rank correlation. This is a fundamental extension of the EMM toolbox, which comes with additional algorithmic challenges. To address these challenges, we provide a series of algorithmic solutions whose strengths/flaws are empirically analysed.

INDEX TERMS Exceptional model mining, exceptional patterns, supervised descriptive pattern mining, rank correlation.

I. INTRODUCTION

We are living in a Golden Age of data science, where data mining techniques designed to discover valuable insights from a collection of records [1] are employed to transform tons of facts into useful information in fields as diverse as education [2], health care [3], and Internet of Things [4]. Nowadays, the quantity of data gathered on different domains is so high that it is a common practice not only to provide specific algorithms for such huge quantity of data [5], but also to reduce such enormous amount of data in order to be able to process it. In this regard, identifying subsets of a dataset that are somehow of great interest to researchers in a specific field is a key point for different discovering and filtering tasks [6]. Traditional pattern mining methods discover coherent nuggets of information that somehow deviate from the norm, i.e., where something interesting is going on. This deviation is quantified according to different measures: in terms of a relatively high/low occurrence, which is known as frequent/infrequent itemset mining [7], or an unusual distribution for a specific target variable, known as Subgroup Discovery (SD) [8], or even considering patterns of high utility for a specific aim [9].

Exceptional model mining (EMM) [10], [11] was proposed as a supervised descriptive pattern mining framework [3] to encompass different forms of interesting behaviour on a pair of target attributes. Unlike SD [12], which typically seeks unusual target *distributions*, EMM typically looks for unusual pairwise *interactions* between two targets (reported as a model class). This framework allows users to define the model class they are interested in, and to search for interesting data subsets according to such model, pointing out reasons to understand why a specific subset causes such unusual interaction. Taking the example widely used in EMM

The associate editor coordinating the review of this manuscript and approving it for publication was Feng Xia.

about the analysis of the housing price per square meter [11], the general know-how is that a larger size of the lot coincides with a higher sales price. At this point, an investor might wonder whether it is possible to find specific data subsets where the price of an additional square meter is significantly less than the norm, or even zero. Finding out such subsets may ease the speculation and increase the benefits thanks to the knowledge provided by EMM, e.g., the price of a house in the higher segments of the market is mostly determined by its location and facilities. The desirable location may provide a natural limit on the lot size, such that this is not a factor in the pricing.

While EMM is quite successful in automatically determining which subgroups of the dataset at hand feature unusual interactions between the predefined pair of target variables, one does need to preset these target variables before running the algorithm. A typical EMM instance does not allow for the algorithm to determine itself which target variables are the relevant ones. Perhaps there may be multiple unusual interactions at play in a larger set of target variables simultaneously. As an example, let us go back to the analysis of the housing price where the aim now is to discover subsets that present an unusual interaction on multiple models at the same time, e.g. housing price and any other target variable. As we will describe in the experimental section, houses with a recreational room cause exceptional pairwise interactions between the sales price and multiple other targets: (1) supplementary square meters are not related to a rise in sales price now; (2) a higher number of full bathrooms does not cause an increment in the sales price; (3) additional stories excluding basement does not imply less affordable houses. The finding of such a subset (houses with a recreational room) determines various exceptional interactions with regard to the housing price that should be known by an investor to apply corrective actions that increase the financial benefits, e.g., adding a recreational room is a luxury that has a huge impact in the sales price in such a way that neither additional square meters, full bathrooms or stories excluding basement will increase the sales price. All in all, a typical EMM instance requires a pair of fixed target variables to look for unusual interactions but, in many application fields, the target variables are not predefined since there is no previous knowledge about the problem at hand, requiring to look for several unusual pairwise interactions between sets of targets.

Any exceptional interaction between any combination of variables is a challenging research agenda; this paper sets a specific step on that path. A new model class for EMM is proposed, assessing the rank correlation between each pair of numeric targets from a set. In other words, extending the EMM model class [13] to highlight subgroups where several pairwise correlations are exceptional. The concept of pairwise exceptional pattern is therefore introduced as an extension of the already known concept of exceptional pattern. Hence, SPEC fundamentally extends the typical EMM toolbox, and as such, requires fundamental algorithmic contributions as well, which this paper will provide. In this regard, the contribution of this research work can be summarized as follows:

1) SPEC describes reasons to understand the cause of unusual interaction among multiple targets in data. It looks for good descriptors extracting interesting subsets of data on contrasting scenarios.
2) A typical EMM instance does not allow for the algorithm to determine itself which target variables are the relevant ones; conversely, SPEC can determine exactly which targets are relevant for the subgroup at hand, and subgroups may be interesting only if they can display exceptionality in several such pairwise settings simultaneously.
3) SPEC is moved towards unsupervised learning tasks where no knowledge about data is required in terms of targets (data subsets to be analysed). Instead, users search for useful information among a wide set of attributes at the same time.

The final aim of this paper is to set the bases for further research studies on the ideas here presented, so the presented algorithmic solutions are just adaptations of well-known algorithms demonstrating that all the proposed ideas are feasible to be carried out. Some data are analysed to demonstrate the usefulness of the proposal and how interesting is the discovery of exceptional data subsets on many different pairs of targets. The rest of the paper is organized as follows. Section II provides some key concepts and descriptions to understand the EMM problem. Section III describes the SPEC model class as well as some algorithmic solutions. Section IV includes some experimental analysis. Finally, a lesson learned is described in Section V and some concluding remarks are outlined in Section VI.

II. PRELIMINARIES

In general terms, a pattern (itemset) is the key element in any process of eliciting useful knowledge [14] since it defines subsequences or substructures representing any type of homogeneity and regularity in data [7]. Formally, given a set of items $I = \{i_1, i_2, \ldots, i_l\}$ in a database Ω, a pattern P is defined as a subset of I, i.e., $P = \{i_j, \ldots, i_k\} \subseteq I \in \Omega \ : \ 1 \leq j \wedge k \leq l$. Pattern mining [15] is a broad subtask of data mining [1] that aims at describing intrinsic and important properties of data by finding novel, significant, unexpected, nontrivial and actionable elements hidden in data. This task identifies and describes chunks of data [11] that are of great interest to researchers in a specific field.

Nowadays, trending supervised local pattern mining [3] frameworks such as exceptional model mining (EMM) [10], which looks for unusual interactions on a pair of targets and describes reasons to understand the cause of such unusual interaction, are being considered to capture different forms of interesting behaviour. EMM [10] is defined as a multi-target generalization of subgroup discovery [8]. Rather than denoting the unusual distribution of a single target variable t as subgroup discovery does, EMM considers the unexpected interaction between a pair of target variables t_x, t_y [11].

EMM is highly related to the discovery of more actionable insights by finding coherent subsets that behave differently when they are compared to either the whole dataset (the interest is focused on deviations from a possibly inhomogeneous norm) or the complement (the attention is paid to dichotomies) of such subsets. In a formal way, let us assume a dataset Ω consisting of a bag of records $r \in \Omega$ in the form $r = (a_1, \ldots, a_k, t_1, t_2) : k \in \mathbb{N}^+$. Here, $A = \{a_1, \ldots, a_k\}$ denotes the *descriptive attributes* or *descriptors*, and $T = \{t_1, t_2\}$ denotes the *target variables* or *targets*. EMM aims at discovering a subset of data $G_D \subset \Omega$ corresponding to a description given by a set of descriptors $D(a_1, \ldots, a_k)$, satisfying that $G_D = \{\forall r^i \in \Omega \ : \ D(a_1^i, \ldots, a_k^i) = 1\}$, and G_D showing an unusual interaction on two specific target variables t_x and t_y.

In one of the three original EMM model classes, the concept of interest was based on the *Pearson's* standard correlation coefficient ρ between t_x and t_y for both the subset ρ^{G_D} and the whole dataset ρ^{Ω} (or the complement $G_D^C \equiv \Omega \setminus G_D$, so the correlation is denoted as $\rho^{G_D^C}$). Thus, a quality measure for this model class was determined by $\varphi(G_D) = |\rho^{G_D} - \rho^{\Omega}|$ or $\varphi(G_D) = |\rho^{G_D} - \rho^{G_D^C}|$ so the higher the value $\varphi(G_D)$ the more exceptional G_D was [10]. However, $G_D \subseteq \Omega$ so it is not possible to statistically compare the exceptionality [13] in terms of $\varphi(G_D) = |\rho^{G_D} - \rho^{\Omega}|$. Besides, *Pearson's* standard correlation coefficient ρ includes some drawbacks that should be considered in the EMM problem [13]: (1) it is sensitive to non-normality and without the normality assumption (it usually happen in many real-life examples), many statistical tests on ρ become meaningless or at least hard to interpret; (2) it is easily affected by outliers; and (3) it assumes there is always a linear relationship between targets. In order to overcome the aforementioned drawbacks, two different correlation coefficients based on ranking were already considered. One of them is the *Spearman's* rank correlation coefficient [16] which is a nonparametric measure of rank correlation (statistical dependence between the ranking of two variables t_x and t_y). It assesses how well the relationship between the two variables (t_x and t_y) can be described using a monotonic function. The other one is *Kendall's* τ coefficient [17] which is used to measure the statistical dependence by determining the similarity of the orderings of the data when ranked by each of the quantities. In general terms, both coefficients usually produce almost the same solutions and, therefore, they can be used interchangeably in this problem as it was already pointed out [13].

Recently, to allow subgroups to be exceptional in subsets of a predefined set of target columns at hand, two approaches to EMM were introduced. As such, these works are the closest related to the current paper. Reference [18] introduced Exceptional Preferences Mining, where a subgroup is interesting if it features unusual preference relations between a predefined set of labels. These relations can be gauged over the whole set, but also in a labelwise or even pairwise manner: in the last case a subgroup is deemed interesting if a single pair of labels is ranked in an unusual way, compared to the overall dataset. Reference [19] defined a local pattern mining task on an olfactory dataset. Since it is generally not well understood exactly which properties influence our olfactory perceptions, the method allows for the discovery of rules where any subset of the predefined label set is relevant. Two facts set these two related works apart from the current paper, that is, neither of the related papers allows for numeric targets, and neither of the related papers explicitly rewards multiple simultaneous exceptional pairwise interactions.

III. SUBSETS OF PAIR-WISE EXCEPTIONAL CORRELATIONS

The SPEC (Subsets of Pairwise Exceptional Correlations) model class fundamentally extends the exceptional model mining (EMM) toolbox. Unusual interactions between target variables are not measured on a single model (pairs of targets) but multiple models (several pairs of targets). Hence, SPEC solves a much broader task that looks for good descriptors extracting interesting subsets of data on contrasting scenarios: a pattern or itemset (set of descriptive attributes) might be deemed to be exceptional if it describes an exceptional behaviour on disparate models in the same dataset Ω.

Definition 1 (descriptors and targets): Suppose a dataset Ω comprising a bag of N records $r \in \Omega$ of the form $r = (a_1, \ldots, a_k, t_1, t_2)$, where k is a positive integer, i.e., $k \in \mathbb{N}^+$. We call the first k attributes of the dataset the *descriptors*, whose set we denote by A and whose collective domain (which is a Cartesian product of k single-attribute domains, each of which can be binary, categorical, or numerical) we denote by $\mathcal{A}$. The last two attributes of the dataset, conversely, are denoted as a set by T, and its collective domain in this paper must be $\mathbb{R}^2$: all targets are numeric. If necessary, we will refer to the i^{th} record of the dataset and its components by superscript i.

Definition 2 (descriptions and subgroups): A *description* D is a function $D : \mathcal{A} \rightarrow \{0, 1\}$. For every record in Ω, a description considers its values for the descriptors, and makes the binary decision whether the description (e.g.: $a_1 \leq 3 \wedge a_2 = \text{red}$) *covers* the record or not. A description D induces a *subgroup* G_D: the set of transactions $G_D \subset \Omega$ that D covers, i.e., $G_D = \{r^i \in \Omega \mid D(a_1^i, \ldots, a_k^i) = 1\}$. We denote by n the number of transactions in a subgroup, i.e., $n = |G_D|$; the capital N is used to define the number of transactions in the whole dataset, i.e., $N = |\Omega|$.

Definition 3 (complement): The complement of a subgroup corresponding to a description D is the set of transactions $G_D^C \subset \Omega$ not covered by D, i.e., $G_D^C = \{r^i \in \Omega \mid D(a_1^i, \ldots, a_k^i) = 0\}$. We also denote by n^C the number of transactions in that set: $n^C = |G_D^C| = N - n$ where $n = |G_D|$ and $N = |\Omega|$.

Definition 4 (exceptional pattern): A description D and its associated subgroup G_D are deemed as exceptional in a dataset Ω if the subset $G_D \subset \Omega$ obtained by D describes an unusual interaction between a pair of targets. This exceptionality is governed by a *quality measure* φ. Even though technically a quality measure has access to the whole description,

the spirit of EMM suggests that it assesses the exceptionality of the interaction between the targets in the induced subgroup. Hence, a description is *defined* in terms of the descriptors, it *selects* a subset of the dataset known as a subgroup, and the quality measure *evaluates* the subgroup by contrasting the interaction behaviour of the selected two targets with target interaction behaviour outside of the subgroup (complement).

Definition 5 (pairwise exceptional pattern): A description D and its associated subgroup G_D are considered as pairwise exceptional in a dataset Ω if the subset $G_D \subset \Omega$ obtained by D describes an unusual interaction between multiple pairs of targets. Similarly to exceptional patterns, the exceptionality of a pair is governed by a *quality measure* φ. In general terms, a pairwise exceptional pattern P_1 is better than another pairwise exceptional pattern P_2 if the number of target pairs in which P_1 is exceptional is much higher than P_2.

A. TASK COMPLEXITY

The traditional exceptional model mining task has a nontrivial computational cost [11] and it is even higher for the proposed SPEC model class. Let us, for the moment, fix two specific target variables t_x and t_y as exceptional model mining does. At least $2^k - 1$ different descriptions exist for this model (assuming that all descriptors are binary; if descriptors are not binary, the number increases so this is the best-case scenario. It is already exponential). Each of these subsets (and their complements) should quantify the rank correlation coefficient for the targets t_x and t_y so a brute force approach requires a total of $2 \times (2^k - 1) = 2^{k+1} - 2$ different evaluations. Let us now consider the proposed SPEC model class and several targets m that is larger than two. Here, the combinations of target pairs for which this procedure needs to be applied is $C_{m,2} = \binom{m}{2} = \frac{m!}{2!(m-2)!} = \frac{1}{2}m(m-1)$ pairs of target variables in Ω. Thus, when dealing with the problem of mining exceptional patterns, a total of $\frac{1}{2}m(m-1)(2^{k+1}-2)$ different evaluations are required; the complexity is exponential in the number of descriptors and quadratic in the number of targets. Additionally, Kendall's τ rank correlation requires a total of $(n \times (n-1))/2$ evaluations in a subgroup of n records, so the τ rank correlation for each pair of targets will require $(|G_D| \times (|G_D| - 1))/2$ operations for the subgroup G_D and $(|G_D^C| \times (|G_D^C| - 1))/2$ for its complement. Hence, the final computational complexity is $\frac{1}{2}m(m-1) \times (2^{k+1} - 2) \times ((|G_D| \times (|G_D| - 1))/2 + (|G_D^C| \times (|G_D^C| - 1))/2)$. In such cases where $m = 2$, the complexity collapses to that of a traditional exceptional model mining with the rank correlation model class.

B. PROPOSED APPROACHES

To demonstrate the usefulness of the proposed ideas and how they can be accomplished, disparate methodologies including exhaustive search, random search and evolutionary approaches are proposed. All these approaches are just adaptations of well-known and widely recognised techniques with the aim of serving as a demonstration of the usefulness of the provided foundations for future research works as well as more efficient and specifically designed algorithms. Adapted algorithms belong to two different methodologies (exhaustive search and heuristic-based approaches). Due to some space restrictions, a more detailed description of such algorithms is accordingly available at http://www.uco.es/kdis/spec together with the source code (they were implemented in Python).

Algorithm 1 Exhaustive Search Approach

Require: Ω, α, β ▷ Dataset Ω, minimum threshold values α and β
Ensure: $\mathcal{P}$
1: $\mathcal{P} \leftarrow \emptyset$
2: $\mathcal{F} \leftarrow$ *Apply a FIM Algorithm*(Ω) ▷ Generate all the frequent itemsets
3: $\mathcal{P} \leftarrow$ *GetExceptinalPatterns*$(\mathcal{F}, \Omega, \alpha, \beta)$ ▷ Calculate exceptionality of $\forall D \in \mathcal{F}$
4: **return** $\mathcal{P}$
5: **procedure** Get exceptional patterns$(\mathcal{F}, \Omega, \alpha, \beta)$
6: $\mathcal{P} \leftarrow \emptyset$
7: **for all** descriptors $D \in \mathcal{F}$ **do** ▷ Analise all descriptors D (patterns) in $\mathcal{F}$
8: $\mathcal{S}_{G_D} \leftarrow \emptyset$
9: $G_D \leftarrow \emptyset$
10: $G_D^C \leftarrow \emptyset$
11: **for all** record $r \in \Omega$ **do** ▷ Save records in G_D or its complement G_D^C
12: **if** $D \subseteq r$ **then**
13: $G_D \leftarrow G_D \cup r$
14: **else**
15: $G_D^C \leftarrow G_D^C \cup r$
16: **end if**
17: **end for**
18: **for all** $t_x, t_y \in \Omega$ **do** ▷ Obtain $z_\tau^{t_x t_y}(G_D)$ for all the pairs of targets
19: calculate $\varphi_\tau^{t_x t_y}(G_D) = 1$-$p$-value of $z_\tau^{t_x t_y}(G_D)$
20: **if** $\varphi_\tau^{t_x t_y}(G_D) \geq \alpha$ **then** ▷ An unusual interaction is discovered
21: $\mathcal{S}_{G_D} \leftarrow \mathcal{S}_{G_D} \cup p$-value
22: **end if**
23: **end for**
24: **if** quality$(\mathcal{S}_{G_D}) \geq \beta$ **then** ▷ More than β unusual interactions
25: $\mathcal{P} \leftarrow \mathcal{P} \cup \mathcal{S}_{G_D}$
26: **end if**
27: **end for**
28: **return** $\mathcal{P}$
29: **end procedure**

1) EXHAUSTIVE SEARCH

Algorithm 1 illustrates the pseudo-code of a simple exhaustive search approach. The proposal works in two different phases. In the first one (see Line 2, Algorithm 1), a frequent

itemset mining (FIM) algorithm is applied to obtain all the frequent solutions (only considering descriptive attributes) in Ω. Here, any of the most widely used FIM algorithms [20] are provided to be used: Apriori, FP-Growth, Eclat, LCM, Sam, Relim. In the second phase (see Line 3, Algorithm 1), the exceptionality of each of the previous solutions is computed (see Lines 5 to 29, Algorithm 1). Here, each descriptor D is analyzed to obtain the subset of data $G_D \subset \Omega$ and its complement G_D^C (see Lines 11 to 17, Algorithm 1). Once G_D and G_D^C are obtained, the procedure calculates the unusual interaction for each pair of targets $t_x, t_y \in T$ measured by one minus the p-value of $z_\tau^{t_x t_y}(G_D)$. This value denotes the probability that there exist statistical differences in the *Kendall's* rank correlation by taking the null-hypothesis $H_0 : \tau_{t_x t_y}^{G_D} = \tau_{t_x t_y}^{G_D^C}$ against $H_1 : \tau_{t_x t_y}^{G_D} \neq \tau_{t_x t_y}^{G_D^C}$. Hence, according a minimum threshold value α, D denotes an unusual interaction on $t_x, t_y \in T$ if and only if $\varphi_\tau^{t_x t_y}(G_D)$, which is calculated as one minus the p-value of $z_\tau^{t_x t_y}(G_D)$, is greater than a minimum α value (see Line 20, Algorithm 1). Finally, a descriptor D (pattern or itemset) will be considered as interesting if it denotes an unusual interaction in more than β pairs of targets $t_x, t_y \in T$ (see Line 24, Algorithm 1). The algorithm ends by returning the set P of exceptional patterns from the whole set $\mathcal{F}$ of feasible patterns or descriptors. Considering the Apriori algorithm as baseline, it is widely studied [21] that its order of the time complexity is exponential, that is, $O(2^{d+1})$ and it runs slowly with regard to the number of attributes d.

2) RANDOM SEARCH

The proposed algorithm (see Algorithm 2) works on *ite* iterations and each iteration is responsible for generating M random solutions (descriptions). To generate a random description (see Lines 4 to 14, Algorithm 2), the algorithm first produces a random number l between 1 and k (number of descriptors in data) to determine the length of the solution (number of items in the description). For such a number, random descriptors are taken to form a chromosome of lenght l. Each gene within the chromosome represents a chosen descriptor. Once the description D is randomly formed (the chromosome), a procedure to calculate whether the descriptor is exceptional or not (see Lines 15 to 27, Algorithm 2) is performed. This procedure is exactly the same as the one shown in Algorithm 1, i.e., it obtains the subset of data $G_D \subset \Omega$ (and its complement G_D^C) given the description D and it calculates the unusual interaction for each pair of targets $t_x, t_y \in T$ measured based on the p-value of $z_\tau^{t_x t_y}(G_D)$. Finally, since this algorithm can work with continuous attributes, those descriptors that include the same descriptors but different range of values should be considered as equal. Hence, a procedure to avoid repeated subsets is included (see Line 29, Algorithm 2), which checks whether the records covered are the same even when their range of values is different. Finally, we also propose a variant of random search approach in which attributes that are not selected yet would be more probable to be selected than those already chosen, that is, future samples are dependent of the previous samples. In other words, line 7 in Algorithm 2 is the only modification of this version, and the probability of choosing an attribute $a_k \in \Omega$ depends on the number of times it was already selected. The order of the time complexity of a

Algorithm 2 Random Search Proposal

Require: $n, ite, M, \Omega, \alpha, \varphi$ ▷ number n of best solutions, number iterations *ite* to be perform and solutions M per iteration, dataset Ω, minimum threshold values α and φ
Ensure: $\mathcal{P}$
1: $\mathcal{P} \leftarrow \emptyset$
2: **for** i **from** 1 **to** *it* **do** ▷ Iterate *ite* times seeking solutions
3: **for** j **from** 1 **to** M **do** ▷ Generate M descriptors in each iteration
4: Random number $l \in [1, k]$ ▷ k is the number of attributes in Ω
5: $D \leftarrow \emptyset$
6: **for** j **from** 1 **to** l **do** ▷ Iterate l times to generate a description D
7: Select a random attribute $a_k \in \Omega$
8: **if** a_k is continuous **then**
9: $d_j \leftarrow$ Two random (uniform) values within $[min(a_k), max(a_k)]$
10: **else**
11: $d_j \leftarrow$ A random (uniform) discrete value for a_k
12: **end if**
13: $D \leftarrow D \cup d_j$
14: **end for**
15: **for all** record $r \in \Omega$ **do** ▷ Save records in G_D or its complement G_D^C
16: **if** $D \subseteq r$ **then**
17: $G_D \leftarrow G_D \cup r$
18: **else**
19: $G_D^C \leftarrow G_D^C \cup r$
20: **end if**
21: **end for**
22: **for all** $t_x, t_y \in \Omega$ **do** ▷ Obtain $z_\tau^{t_x t_y}(G_D)$ for all the pairs of targets
23: calculate $\varphi_\tau^{t_x t_y}(G_D) = 1$-$p$-value of $z_\tau^{t_x t_y}(G_D)$
24: **if** $\varphi_\tau^{t_x t_y}(G_D) \geq \alpha$ **then** ▷ An unusual interaction is discovered
25: $\mathcal{S}_{G_D} \leftarrow \mathcal{S}_{G_D} \cup p$-value
26: **end if**
27: **end for**
28: **if** $\varphi_{\text{SPEC}i}(\mathcal{S}_{G_D}) \geq \varphi$ **then** ▷ $i \in$ {one, some, all}
29: Check and update range of values of D
30: Update set $\mathcal{P}$ with D considering the maximum number n of solutions
31: **end if**
32: **end for**
33: **end for**
34: **return** $\mathcal{P}$

random search algorithm depends on the number of solutions m to be found, that is, $O(m)$.

3) EVOLUTIONARY COMPUTATION

The proposal follows a well-known generational schema where, in each generation of the evolutionary process, solutions are crossed and mutated, and new offspring are obtained. The algorithm (see Algorithm 3) starts by encoding patterns (descriptions) through a similar process of the already described random search approach (see Lines 4 to 31, Algorithm 2). Finally, the generational schema is performed over G generations (see Lines 6 to 30, Algorithm 3), returning the set $\mathcal{E}$ comprising those best solutions found so far. A fitness function is proposed to define how promising a solution s_1 is, in such a way that it is defined as the average of one minus the p-value of $z_\tau^{t_x t_y}(s_1)$ for any pair of targets t_x and t_y.

In order to obtain new solutions along the evolutionary process, two genetic operators have been proposed and applied (see Line 8, Algorithm 3). The crossover genetic operator is based on the assertion that extremely high or low values of $|G_D|$ tend to produce a high exceptionality since $|G_D|$ and $|G_D^C|$ are dissimilar [13]. Thus, having a solution with a $|G_D|$ (frequency of the pattern) value close to 0.5 (in per unit basis) tend to obtain a new solution whose $|G_D|$ value is far from 0.5. Here, given two patterns (solutions or set of descriptors) p_1 and p_2 and each pattern including a set of items (variables of the dataset), the item having the lowest frequency from the pattern having the highest frequency is swapped by the item with the highest frequency from the pattern with the lowest frequency. In other words, being the frequency of p_1 0.7 and the frequency of p_2 0.3, the aim is to increase the frequency of p_1 and reduce the frequency of p_2 so both solutions tend to be far from 0.5 (the value of $|G_D|$ to be avoided). As for the mutator genetic operator, it follows the same idea as the crossover operator but it now replaces the worst item within the pattern by a random one. It is widely studied [21] that the order of the time complexity of evolutionary algorithms follows a quadratic distribution, being equal to $O(N \times d)$ for d attributes and N instances or transactions. Last but not least, this proposal cannot be compared to existing high-performance evolutionary-based solutions for mining frequent patterns [22], [23] since the goal is completely different. In proposal for mining exceptional patterns the aim is not to find frequent data subsets so neither the fitness function nor the genetic operators are similar to those provided by authors in [22], [23]. Hence, the proposals cannot be fairly compared.

Algorithm 3 Evolutionary Algorithm

Require: $G, M, \Omega, \alpha, \varphi$ ▷ Number of generations G, population size M, dataset Ω, minimum threshold values α and φ

Ensure: $\mathcal{E}$

1: $\mathcal{P} \leftarrow \emptyset$ ▷ General population
2: $\mathcal{E} \leftarrow \emptyset$ ▷ Elite population with the best solutions found so far
3: **for** j **from** 1 **to** M **do** ▷ Generate M solutions (descriptors) to form the general population
4: Follow the same procedure as the one shown in Lines 4 to 31, Algorithm 2
5: **end for**
6: **for** g **from** 1 **to** G **do** ▷ Iterate G times seeking solutions
7: *parents* $\leftarrow$ apply parent selector on $\mathcal{P}$ ▷ Typical tournament selector (the size of tournament is 3)
8: *offspring* $\leftarrow$ apply crossover and mutation on *parents*
9: **for all** *ind* $\in$ *offspring* **do** ▷ Evaluate all the new individuals
10: **for all** record $r \in \Omega$ **do** ▷ Save records in G_D or its complement G_D^C
11: **if** $D \in ind \subseteq r$ **then**
12: $G_D \leftarrow G_D \cup r$
13: **else**
14: $G_D^C \leftarrow G_D^C \cup r$
15: **end if**
16: **end for**
17: **for all** $t_x, t_y \in \Omega$ **do** ▷ Obtain $z_\tau^{t_x t_y}(G_D)$ for all the pairs of targets
18: calculate $\varphi_\tau^{t_x t_y}(G_D) = 1$-$p$-value of $z_\tau^{t_x t_y}(G_D)$
19: **if** $\varphi_\tau^{t_x t_y}(G_D) \geq \alpha$ **then** ▷ An unusual interaction is discovered
20: $\mathcal{S}_{G_D} \leftarrow \mathcal{S}_{G_D} \cup p$-value
21: **end if**
22: **end for**
23: **if** $\varphi_{\text{SPEC}i}(\mathcal{S}_{G_D}) \geq \varphi$ **then** ▷ $i \in$ {one, some, all}
24: Check and update range of values of $D \in ind$
25: Update set $\mathcal{P}$ with *ind* considering the population size
26: **end if**
27: **end for**
28: $\mathcal{P} \leftarrow$ update the general population considering the set *offspring*
29: $\mathcal{E} \leftarrow$ update the elite population considering the sets $\mathcal{P}$ and *offspring*
30: **end for**
31: **return** $\mathcal{E}$

IV. EXPERIMENTAL ANALYSIS

The aim of this experimental analysis is threefold: demonstrating the usefulness of the genetic operators for this problem when evolutionary algorithms are considered; describing the performance of the proposed algorithms when data dimensionality varies; demonstrating the importance of using SPEC by analysing different solutions. The reader should consider that insights discovered by SPEC on real-word data should be analysed in collaboration with experts in the specific application field. Three different methodologies, implemented in Python programming language, are given and all their variants are freely available to be downloaded at http://www.uco.es/kdis/spec. It should be noted that all the experiments presented in this section were

TABLE 1. Datasets (ordered according to the number of targets) used in the experimental study. Descriptors label shows the number of discrete and continuous descriptors (discrete, continuous).

Dataset	$\#Transactions$	#Targets	λ	$\#Descriptors$
Iris	150	4	6	1 (1, 0)
Housing	546	5	10	7 (7, 0)
AirQuality	9,358	6	15	6 (0, 6)
StockRange	951	7	21	9 (6, 3)
Pollution	61	12	66	4 (0, 4)
WaterQuality	1,060	13	78	10 (3, 7)
Emotions	593	65	2,080	5 (0, 5)
Yeast	2,417	99	4,851	18 (14, 4)

run on an Intel(R) Core(TM) i7 CPU at 2.67GHz with 12GB main memory and running CentOS 5.4. To carry out this experimental analysis a collection of datasets are considered and described, considering a varied number of transactions, targets and descriptors.

As any evolutionary approach, the proposed evolutionary model should be configured with a set of adjustable parameters. All these parameters require previous study to determine those considered optimal, that is, those that allow us to obtain the best global results. It is worth mentioning that no single combination of parameter values performs better for any data sets, and sometimes, it depends on the problem under study. In this regard, the best results for the evolutionary approach are described in the following subsection.

A. DATASETS AND EXPERIMENTAL SET-UP

The experimental analysis has been carried out by considering a varied set of datasets (see Table 1) which is publicly available at http://www.uco.es/kdis/spec. These datasets were selected to be as varied as possible, comprising either continuous and discrete attributes, including a varied number of target attributes (λ combinations of pairs of targets), and containing a diverse number of transactions. As for the target variables, the λ value is also provided (feasible pairs of targets to be considered). Finally, since exhaustive search algorithms cannot handle continuous descriptors, any descriptor variable that is defined in a continuous domain will be discretized in 3 and 5 bins of equal width and equal frequency. Datasets with no descriptor defined in a continuous domain (*Iris* and *Housing*) will not be discretized.

Exhaustive search approaches have been run by considering different support thresholds (0.05, 0.10 and 0.15) so any pattern (G_D) that overcomes these thresholds is analysed in terms of exceptionality ($\varphi_\tau^{t_i t_j}(G_D)$ value). Even when three manners of gauging exceptionality were described in this work (φ_{SPECone}, $\varphi_{\text{SPECsome}}$, and φ_{SPECall}), the experimental analysis aims to compare the runtime and performance of the proposed algorithms and, therefore, the behaviour is measured regardless the three metrics or manners of gauging exceptionality. Hence, the average $\varphi_\tau^{t_i t_j}(G_D)\ \forall t_i t_j \in T$ is considered to quantify the solutions. Random search approaches consider 2,000 iterations in which the best 20 solutions are returned. Finally, as for the evolutionary computation approach, it considers a population size of 100 individuals and the algorithm is running till there are 150 generations without any improvement in the average results (considering the best 20 solutions found so far). As for the genetic operators' probabilities, the algorithm self-adapts these values, that is, the values increase or decrease according to the average value of the 20 best solutions (elite population). In the beginning, both probabilities (mutation and crossover) are the same (a 0.5 value is considered). The average value of the elite population is analysed every 5 iterations and if there is an improvement, then the crossover probability is increased in 0.05 and the mutation probability is decreased in 0.05. On the contrary, if no improvement is achieved after 5 iterations, then the crossover probability is decreased in 0.05 and the mutation probability is increased in 0.05.

TABLE 2. Average number of solutions (and percentage of improvement) that are better than their predecessors for each dataset.

	Crossover			Mutation		
Dataset	Random	Proposed	%	Random	Proposed	%
Housing	780.9	870.9	11.5	708.1	755.2	6.6
StockRange	629.7	760.0	20.7	604.0	778.3	28.9
Pollution	557.1	676.2	21.4	601.5	724.8	20.5
Emotions	710.0	770.6	8.5	592.5	650.0	9.7
Yeast	733.8	916.7	24.9	622.2	765.3	23.0

B. ANALYSIS OF THE GENETIC OPERATORS

This section aims to demonstrate that the proposed genetic operators are well-suited for this problem, so a comparison with random genetic operators is performed. By random genetic operator, we mean that items within a solution are randomly selected to be swapped by those of other solution (crossover) or to be replaced by new ones (mutation). In this analysis, the *Iris* dataset has been discarded since it only includes a single descriptor with three different values and, therefore, only three solutions are available (one for each of the three values). Five different datasets (considering a different number of descriptors) have been considered in this analysis, comprising few descriptors (*Pollution* and *Emotions* datasets) as well as a high number of descriptors (*Yeast* dataset).

First, each genetic operator is analysed in isolation, studying the number of new solutions obtained that are better than their predecessors. In other words, how many times each genetic operator can produce new solutions whose $\varphi_\tau^{t_i t_j}(G_D)$ values are higher than their predecessors. The average results obtained after running 30 times each genetic operator and dataset, and considering the same number of individuals to be obtained, are shown in Table 2. Second, we combine both versions of crossover and mutation genetic operators (random and proposed) and run the whole evolutionary proposal 30 times per dataset. Results are shown in Table 3, considering the 20 best solutions found so far. The aim is to demonstrate which combination is better in terms of average $\varphi_\tau^{t_i t_j}(G_D)$ for each solution provided by the evolutionary proposal.

As a result, the proposed crossover/mutation genetic operators are really suitable for this problem. Thus, given a

TABLE 3. Average $\varphi_\tau^{t_i t_j}(G_D)$ (for the resulting set of solutions) obtained for each dataset.

Dataset	(1)	(2)	(3)	(4)
Housing	0.7673	0.7782	0.7724	0.7815
StockRange	0.5402	0.5446	0.5381	0.5477
Pollution	0.5075	0.5155	0.5154	0.5188
Emotions	0.5213	0.5238	0.5241	0.5251
Yeast	0.5082	0.5093	0.5086	0.5090

(1) Random crossover and mutation
(2) Random crossover and proposed mutation
(3) Proposed crossover and random mutation
(4) Proposed crossover and mutation

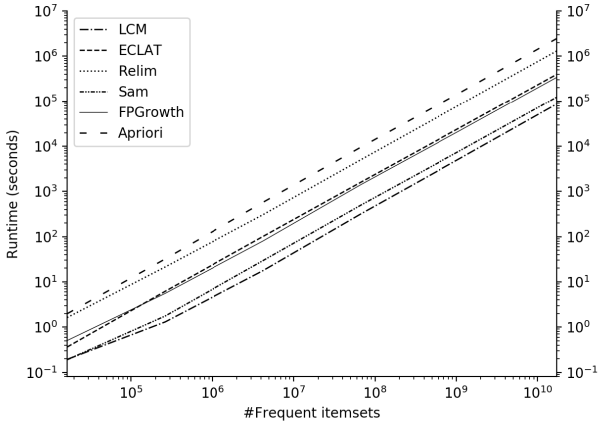


FIGURE 1. Performance of the exhaustive search approaches when the search space increases.

description D comprising a set of descriptors $a_i, \ldots, a_j$, it is much more interesting to modify those descriptors that produce a frequency of G_D close to 0.5 than to randomly choose a descriptor within D. This issue was already formally described in the previous section according to some preliminary studies described in [13].

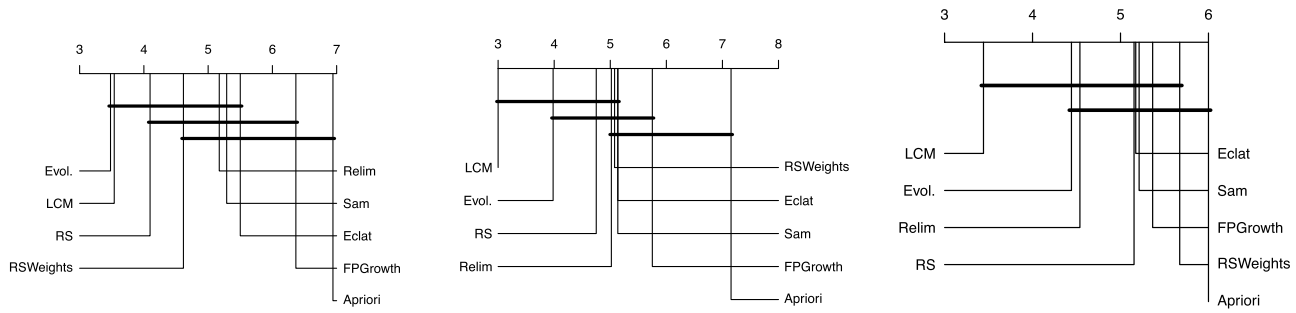
C. ANALYSIS OF THE PERFORMANCE

The goal of this analysis is twofold. First, it describes the performance in terms of runtime. Second, it studies the results in terms of average $\varphi_\tau^{t_i t_j}(G_D)$ for the set of solutions returned by each algorithm. In order to carry out a fair comparison, this second analysis is performed by taking the best 20 solutions provided by the exhaustive search algorithms and comparing them to the set of 20 solutions provided by heuristic-based approaches. It is important to remark that any other comparison is unfair since exhaustive search approaches return the whole set of solutions and, therefore, the average is biased.

Focusing first on the analysis of the runtime (see Figure 1), let us focus on how the exhaustive search approaches perform when data dimensionality increases (in terms of attributes/descriptors), that is, from 10^4 to 10^{10} items. These exhaustive search algorithms are based on well-known FIM algorithms (LCM, ECLAT, Relim, Sam, FP-Growth and Apriori) [24] and they all return exactly the same set of solutions. In general terms and similarly to some studies already carried out by the FIM community [20], the LCM algorithm is the one that best performs, but Sam also obtains really good results in terms of runtime. Apriori, on the contrary, is the algorithm that worst performs and this behaviour was expected since it was the first algorithm proposed to obtain frequent itemsets.

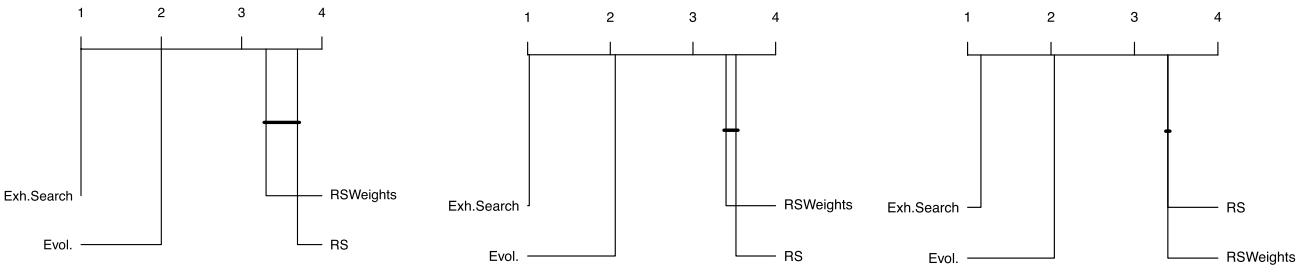
Considering now the whole set of datasets provided in the experimental set-up, the runtime is analysed by considering all the exhaustive search approaches (LCM, ECLAT, Relim, Sam, FP-Growth and Apriori), two versions of a Random Search (RS as a traditional random search and RSWeights as the version in which the probability to be chosen is inversely proportional to the number of times the attribute was chosen), and an evolutionary computation approach (Evol.). Three different support threshold values were also considered. Additionally, due to exhaustive search algorithms (LCM, ECLAT, Relim, Sam, FP-Growth and Apriori) cannot deal with continuous descriptors/attributes, different discretization methods (equal-width EW, and equal-frequency EF) have been applied to those datasets having continuous features. Here, 3 and 5 bins have been considered for both EW and EF methods. In total, 26 different scenarios (8 datasets with 4 different discretization methods, except for Iris and Housing that did not require any discretization) were considered. To analyse and validate the results of a series of nonparametric statistical tests were considered, applying the Friedman's test [25] to evaluate whether there are significant differences in the results of the algorithms. If Friedman's test indicated that the results were significantly different, the Rom post-hoc test [26] was used to perform multiple comparisons among all methods. This test is a modification to Hochberg's procedure [27] to increase its power and it was well-studied and highly recommended [28] to perform multiple comparisons in experimental studies. As a result, Friedman's test rejected the null hypothesis in all cases analysed, considering a significance level $\alpha = 0.05$. A Rom post-hoc test [26] is therefore performed for all pairwise comparisons and results are illustrated in Figure 2. LCM and the evolutionary computation approach appeared as the best algorithms in runtime, the former being better when the search space decreases (a higher support threshold value). Apriori, on the contrary, appeared as the worst algorithm in terms of runtime. In general, there is no statistical difference among LCM, Evol., RS, RSWeights, Relim, Sam and ECLAT.

Focusing on the analysis of the results in terms of average $\varphi_\tau^{t_i t_j}(G_D)$ for the set of solutions returned by each algorithm, the best 20 solutions provided by each algorithm were analysed. It is important to remark again that any other comparison is unfair since exhaustive search approaches return the whole set of solutions and, therefore, the average is biased. Similarly to the previous analysis, three different studies based on different support threshold values were performed. In order to analyse and validate the results, a series of nonparametric statistical tests were also considered, applying the Friedman's test [25] to evaluate whether there are significant differences in the results of the algorithms. Since all the exhaustive search approaches return exactly the same set of



(a) Support threshold value of 5% (b) Support threshold value of 10% (c) Support threshold value of 15%

FIGURE 2. Runtime analysis. Results for the Rom post-hoc test on three different support threshold values (5%, 10% and 15%).



(a) Support threshold value of 5% (b) Support threshold value of 10% (c) Support threshold value of 15%

FIGURE 3. Average $\varphi_\tau^{t_i t_j}(G_D)$ analysis. Results for the Rom post-hoc test on three different support threshold values (5%, 10% and 15%).

results, we have gathered all the algorithms under the label *Exh.Search*. Additionally, two versions of a Random Search (RS as a traditional random search and RSWeights as the version in which the probability to be chosen is inversely proportional to the number of times the attribute was chosen), and an evolutionary computation approach (Evol.) were considered. The Friedman's test rejected the null hypothesis in all cases analysed, considering a significance level of $\alpha = 0.05$. A Rom post-hoc test [26] is therefore performed for all pairwise comparisons and results are illustrated in Figure 3. As it was expected, exhaustive search approaches obtain the best results since they can mine the whole search space and, therefore, the 20 best solutions analysed are always those with the maximum $\varphi_\tau^{t_i t_j}(G_D)$ value. The evolutionary computation approach appears as the second-best approach, being better than RS and RSWeights. There is no statistical difference between RSWeights and RS, but the former performs statistically better. It is important to remark that the statistical differences between Exh.Search and Evol. are reduced with the reduction of the search space (higher support threshold value). Finally, shall us remark that all these analyses were performed on 26 different scenarios (8 datasets with 4 different discretization methods, except for Iris and Housing that did not require any discretization).

D. ANALYSIS OF THE SOLUTIONS

This subsection aims to illustrate and analyse some insights obtained in the experimental analysis, demonstrating the utility of SPEC on real scenarios. The two datasets with the lowest number of targets have been considered here as a matter of shortening the study. It is important to remark that this analysis should be done in collaboration with experts to fully understand the extracted insights.

1) IRIS DATASET

In the first real-world experiment, we analyse the *Iris* dataset, which is perhaps the best-known dataset to be found in the data mining literature. *Iris* consists of 50 samples from each of three species of Iris (Iris setosa, Iris virginica and Iris versicolor). Four features were measured from each sample, including the length and the width of the sepals and petals, in centimetres. The information for each sample includes four attributes of interest that are taken as target variables (sepal-length, petal-length, sepal-width and petal-width), and an additional attribute to define candidate subgroups, i.e. the species of Iris. In this analysis, the total number of pairs of targets is 6, i.e. $\lambda = C_{4,2} = 6$. Since only one attribute is considered to generate candidate subgroups, the search space is equal to the number of values for this single attribute, i.e. three subgroups, one per species of *Iris*. The one that provided a higher $\Delta\varphi_\tau^{t_i t_j}(G_D)$ is Iris-setosa (value 0.8108), which is much higher than Iris-versicolor (value 0.4238) and Iris-virginica (value 0.3946). Considering, therefore, Iris-setosa as the baseline, Figure 4 shows the $\varphi_\tau^{t_i t_j}(G_D)$ for each of the 6 pairs of targets, obtaining that the general trend between most of the pairs of targets is completely different

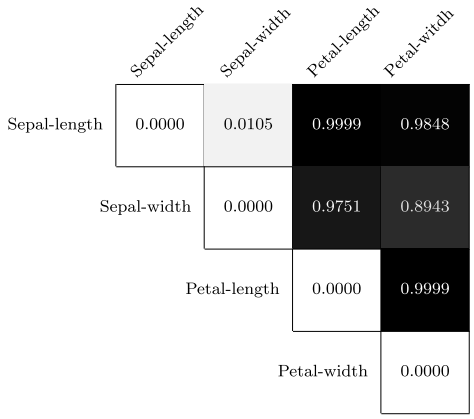


FIGURE 4. Heatmaps of the $\varphi_\tau^{t_i t_j}(G_D)$ values for each pair of targets and considering the descriptor D={*class = Iris-setosa*} found on the *Iris* dataset.

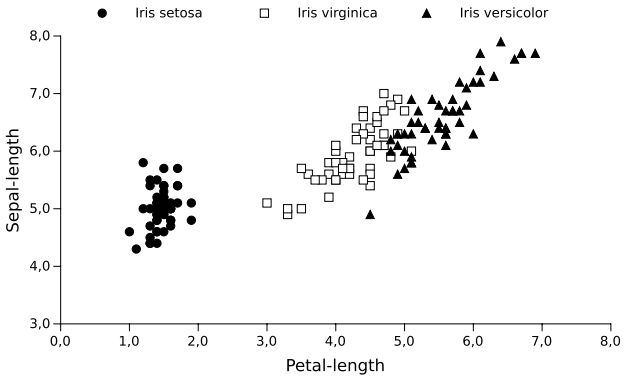


FIGURE 5. Iris species by petal and sepal length.

when setosa is considered. It is possible to statistically assert with a probability $\alpha = 0.95$ that the trend between petal-length and any other feature is affected by the setosa class. In other words, the length of the petal and sepal denotes an exceptional correlation with a probability of 0.9999; the length of the petal and the width of the sepal exceptionally correlates with a probability of 0.9751; and the length and width of the petal also denote an exceptional correlation with a probability of 0.9999. Hence, D={*class = Iris-setosa*} is an exceptional pattern that denotes exceptional correlations between different pairs of measures. For a matter of clarification, let us focus on the length of the petal and sepal, which denoted an exceptional correlation with a probability of 0.9999. This behaviour is clearly illustrated in Figure 5, where the length of both the sepal and the petal seems to be positively correlated (an increment of the petal length implies an increment of the sepal length). This behaviour, however, is different when only the setosa class is considered since an increment of the petal length does not imply an increment in the sepal length.

The strength of using SPEC is the ability to obtain subsets that denote an exceptional behaviour on more than a single pairs of targets. The example provided in this section illustrates the capability of SPEC to provide the user with a more general knowledge, much more rewarding than the one obtained by EMM. For example, considering any algorithm

FIGURE 6. Heatmaps of the $\varphi_\tau^{t_i t_j}(G_D)$ values for each pair of targets and considering the descriptor D={*recroom = no, fullbase = yes, gashw = no, prefarea = no*} found on the *Windsor Housing* dataset.

for mining exceptional models [11], it is possible to obtain that the subset given by the descriptor D={*class = Iris-virginica*} is really interesting for the targets sepal-length and petal-width, since with a probability $p = 0.9999$ it is possible to statistically assert that its correlation is exceptional. Even when this assertion is fine, this knowledge is partial and does not illustrate all the information that can be obtained, i.e. descriptor D={*class = Iris-setosa*} is more general and provides additional unusual interactions among target variables.

2) WINDSOR HOUSING DATASET

In this second analysis on a real-world dataset we analyse the *Windsor Housing* dataset. It contains information on 546 houses that were sold in Windsor, Canada, in the summer of 1987. The information for each house includes 5 attributes of interest that are taken as target variables (price of a house, the lot size of a property in square feet, number of bedrooms, number of full bathrooms, and the number of stories excluding basement), and 7 additional attributes to define candidate subgroups: *Driveway* — does the house include a driveway?; *Recroom* — does the house have a recreational room?; *Fullbase* — does the house have a full finished basement?; *Gashw* — does the house use gas for hot water heating?; *Airco* — does the house have central air conditioning?; *Garagepl* — number of garage places, which can be 0, 1, 2, or 3; and, finally, *Prefarea* — is the house located in a preferred neighbourhood of the city?. It is important to remark that, due to some space limitations, the whole set of obtained results are available at http://www.uco.es/kdis/spec. The most important results are described below.

Considering the solution with the highest $\Delta\varphi_\tau^{t_i t_j}(G_D)$ value, it refers to a descriptor D={*recroom = no, fullbase = yes, gashw = no, prefarea = no*} where the subset given by D includes 77 records, and the complement 469. Analysing each pair of targets (see Figure 6), it is obtained that the sales price denotes a behaviour that is statistically different on any of the following variables: lot size, number of bedrooms, number of full bathrooms, and the number of stories

excluding basement. This issue is quite interesting since it statistically shows that the correlation between price and the aforementioned variables is modified when D comes to play. In other words, it means that when a house includes a full finished basement, even when it is not located in a preferred neighbourhood and it lacks of some extras (recreational room and gas for hot water heating), then the price is not related to the lot size since extra square meters are meaningless for the purchaser. Here, the fact of having a full finished basement is a huge motivation to invest in houses. The same behaviour is described for additional bedrooms, bathrooms or stories. According to the general trend, all of them are positively correlated to the sales price. However, this correlation disappears with D and it may be caused by the full finished basement, which is considered as a luxury in some neighbourhoods. This knowledge is essential for any real estate agency, so corrective actions can be taken to avoid that unusual correlation with regard to the sales price.

Since the *Windsor Housing* dataset has been previously used in EMM, some of the provided descriptors [11] are analysed again by considering now the proposed SPEC model class. In this sense, the variables *driveway* and *recroom* are analysed, which were previously obtained by EMM [11] as good descriptors on the basis of targets *Price* and *Lot size*. The aim now is to provide a more general knowledge by studying whether these descriptors (*driveway* and *recroom*) provide exceptional behaviour on the whole set of target variables. Analysing the results obtained by SPEC in this sense (see Figure 7), it is obtained that the price is only affected by the descriptor $D = \{driveway = yes,\ recroom = yes\}$ on a single target variable, i.e. *Lot size*. It means that the price is not positively correlated (as expected according to the general trend) to the lot size for those houses having a driveway and a recreational room. This exceptional behaviour is perhaps caused by the fact that these features are considered luxury extras for a house so the fact of adding extra square meters does not heavily affect to the sales price. Additionally, it should be noted that the price increases (as expected by the general trend of data) with the increasing number of bedrooms, bathrooms and stories excluding basement even when the house includes a driveway and a recreational room.

Finally, we aim to analyse each of the attributes within the aforementioned descriptor, i.e. $D = \{driveway = yes,\ recroom = yes\}$. Figure 8(a) shows the p-values for each pair of targets when considering the subset given by the feature $driveway = yes$. As shown, this subset does not affect the general data trend and it is difficult to find an exceptional behaviour on any pair of targets. On the contrary, considering the feature $recroom = yes$, it is more probable to find unusual interactions between different pairs of targets (see Figure 8(b)). For example, it is illustrated that it is highly probable that those houses that include a recreational room do not increase their price with additional square meters (considering a probability of 97.67%) neither with additional bathrooms (considering a probability of 95.96%). It means that the fact of having a recreational room is enough for many purchasers so they do not consider the lot size or the number of bathrooms as extra features that someone should pay for them.

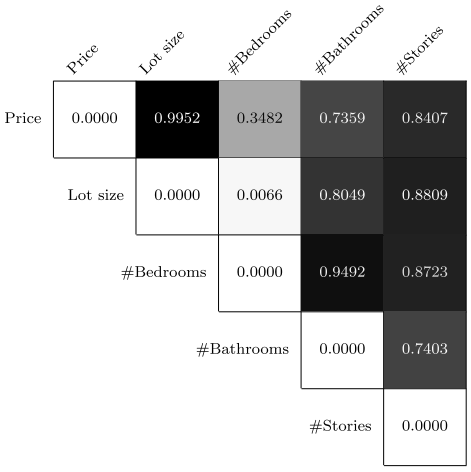
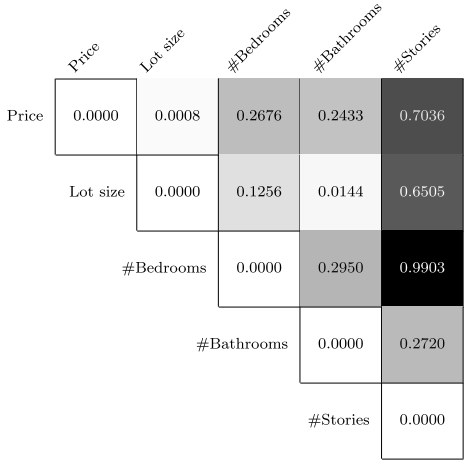


FIGURE 7. Heatmaps of the $\varphi_{\tau}^{t_i t_j}(G_D)$ values for each pair of targets and considering the descriptor D={*driveway* = *yes*, *recroom* = *yes*} found on the *Windsor Housing* dataset.



(a) D={*driveway* = *yes*}

(b) D={*recroom* = *yes*}

FIGURE 8. Heatmaps of the probability values (1-p-values) for two descriptors D, found on the *Windsor Housing* dataset.

V. LESSONS LEARNT

Unlike traditional EMM, which aims at discovering interesting data subsets that denote some unusual interaction between a fixed pair of target variables, SPEC mines data subsets on any combination of target variables. Hence, the task of the SPEC model class for EMM is computationally harder than traditional EMM. Additionally, since SPEC extracts data subsets with an unusual interaction between any pair of variables, the knowledge provided is potentially more powerful, looking beyond where EMM looks.

The experimental analysis has demonstrated the good performance of the proposed algorithms for different data dimensionalities, also obtaining a diverse set of solutions. Results obtained on various real-world datasets have demonstrated the usefulness of this new model class. On the dataset including the sales price of houses, it is discovered that houses with a recreational room are quite interesting since they denote exceptional behaviour on multiple scenarios. For example, the price of a house usually increases when the lot size also increases, or when the number of bathrooms increases (it is a luxury to have some extra bathrooms so people pay for that). However, when the house includes a recreational room, it is discovered that with a high probability (more than 95%), the price is not affected by the lot size or the number of bathrooms. A recreational room is a luxury by itself so some extra square feet or some additional bathrooms are meaningless for the price — it usually happens that houses with a recreational room are located in the best districts of the city so the price is not affected so much by the aforementioned variables.

As demonstrated, the aforementioned knowledge cannot be provided by traditional EMM tasks since they are only focused on a pair of targets. Thus, any comparison with regard to EMM algorithms is unfair. It is also important to highlight that a major drawback of SPEC is its computational time, which is higher than the one of traditional EMM due to the huge number of combinations of targets that it requires to be analysed. In this sense, some proposals based on heuristic search methodologies were provided to reduce the computational time. These algorithms include some additional features such as the ability to extract subsets on continuous domains (exhaustive search approaches require a preprocessing step to transform continuous variables into discrete ones). This feature implies, just in numerical domains, that the whole set of features can be used as tentative target variables. It is therefore not required to predefine a set of descriptive attributes and a set of target variables. A major drawback of the heuristic approaches, however, is the lack of guarantee that all the feasible solutions are analysed so better solutions can be still hidden for the user.

To sum up, the proposed approaches have some advantages and flaws. If the runtime is meaningless and the user requires to analyse the whole search space to take any existing solution, then the exhaustive search approaches are much more appropriate and, within them, LCM seems to be the faster one. On the contrary, if the runtime is paramount, then heuristic-based solutions should be used, especially the evolutionary computation approach. Nevertheless, the user should be aware that the obtained solutions might not be the best ones.

VI. CONCLUSION

This paper presents the Subsets of Pairwise Exceptional Correlations (SPEC) model class for EMM. The proposed model class shares the basic concepts of exceptional model mining, but it also includes some additional features that are required to improve the knowledge on the user's side. In any case, any comparison of SPEC with regard to EMM algorithms is unfair since the aim of both are completely different (a pair of targets vs a wide set of targets). Additionally, it is important to remark that SPEC is a supervised local pattern mining task that describes reasons to understand the cause of unusual interaction among any combination of target variables in data. When the set of target variables is wide enough, SPEC moves towards an unsupervised learning task. Finally, since SPEC extracts data subsets with an unusual interaction between any pair of variables, the knowledge provided is more useful since it looks not only for exceptional subsets on a particular case (a predefined set of targets as exceptional model mining does) but on the general dataset, denoting an exceptional behaviour on the whole dataset.

The formal definitions and major features of this novel framework are described, and multiple algorithmic solutions are presented (different exhaustive search and heuristic-based approaches). All these approaches are just adaptations of well-known and widely recognised techniques with the aim of serving as a demonstration of the usefulness of the provided foundations for future research works as well as more efficient and specifically designed algorithms.

REFERENCES

[1] P. N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*, 1st ed. Reading, MA, USA: Addison-Wesley, 2005.

[2] A. Y. Noaman, J. M. Luna, A. H. M. Ragab, and S. Ventura, "Recommending degree studies according to students' attitudes in high school by means of subgroup discovery," *Int. J. Comput. Intell. Syst.*, vol. 9, no. 6, pp. 1101–1117, Nov. 2016, doi: 10.1080/18756891.2016.1256573.

[3] S. Ventura and J. M. Luna, *Supervised Descriptive Pattern Mining*, 1st ed. Cham, Switzerland: Springer, 2018.

[4] Y. Luo and Y. Xiang, "Application of data mining methods in Internet of Things technology for the translation systems in traditional ethnic books," *IEEE Access*, vol. 8, pp. 93398–93407, 2020, doi: 10.1109/ACCESS.2020.2994551.

[5] J. M. Luna, F. Padillo, M. Pechenizkiy, and S. Ventura, "Apriori versions based on MapReduce for mining frequent patterns on big data," *IEEE Trans. Cybern.*, vol. 48, no. 10, pp. 2851–2865, Oct. 2018.

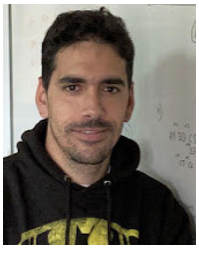
[6] J. M. Luna, M. Pechenizkiy, M. J. del Jesus, and S. Ventura, "Mining context-aware association rules using grammar-based genetic programming," *IEEE Trans. Cybern.*, vol. 48, no. 11, pp. 3030–3044, Nov. 2018.

[7] S. Ventura and J. M. Luna, *Pattern mining With Evolution Algorithms*, 1st ed. Cham, Switzerland: Springer, 2016.

[8] F. Herrera, C. J. Carmona, P. González, and M. J. del Jesus, "An overview on subgroup discovery: Foundations and applications," *Knowl. Inf. Syst.*, vol. 29, no. 3, pp. 495–525, Dec. 2011.

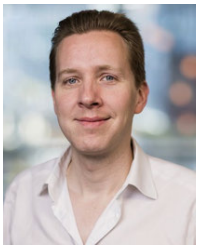
[9] J. C. Lin, Y. Li, P. Fournier-Viger, Y. Djenouri, and J. Zhang, "Efficient chain structure for high-utility sequential pattern mining," *IEEE Access*, vol. 8, pp. 40714–40722, 2020, doi: 10.1109/ACCESS.2020.2976662.

[10] D. Leman, A. Feelders, and A. J. Knobbe, "Exceptional model mining," in *Proc. Eur. Conf. Mach. Learn. Knowl. Discovery Databases*, vol. 5212. Antwerp, Belgium: Springer, 2008, pp. 1–16.
[11] W. Duivesteijn, A. J. Feelders, and A. Knobbe, "Exceptional model mining: Supervised descriptive local pattern mining with complex target concepts," *Data Mining Knowl. Discovery*, vol. 30, no. 1, pp. 47–98, Jan. 2016.
[12] W. Klösgen, "Explora: A multipattern and multistrategy discovery assistant," in *Advances in Knowledge Discovery and Data Mining*, U. M. Fayyad, P. Smyth, G. Piatetsky-Shapiro, and R. Uthurusamy, Eds. Palo Alto, CA, USA: American Association for Artificial Intelligence, 1996, pp. 249–271.
[13] L. Downar and W. Duivesteijn, "Exceptionally monotone models—The rank correlation model class for exceptional model mining," *Knowl. Inf. Syst.*, vol. 51, no. 2, pp. 369–394, 2017.
[14] J. M. Luna, P. Fournier-Viger, and S. Ventura, "Extracting user-centric knowledge on two different spaces: Concepts and records," *IEEE Access*, vol. 8, pp. 134782–134799, 2020, doi: 10.1109/ACCESS.2020.3010852.
[15] C. C. Aggarwal and J. Han, *Frequent Pattern Mining*, 1st ed. Cham, Switzerland: Springer, 2014.
[16] C. Spearman, "The proof and measurement of association between two things," *Amer. J. Psychol.*, vol. 15, no. 1, pp. 72–101, 1904, doi: 10.2307/1412159.
[17] M. G. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, nos. 1–2, pp. 81–93, Jun. 1938.
[18] C. R. de Sá, W. Duivesteijn, C. Soares, and A. J. Knobbe, "Exceptional preferences mining," in *Discovery Science* (Lecture Notes in Artificial Intelligence). Bari, Italy: Springer, 2016, pp. 3–18.
[19] G. Bosc, J. Golebiowski, M. Bensafi, C. Robardet, M. Plantevit, J. Boulicaut, and M. Kaytoue, "Local subgroup discovery for eliciting and understanding new structure-odor relationships," in *Discovery Science* (Lecture Notes in Artificial Intelligence). Bari, Italy: Springer, 2016, pp. 19–34.
[20] P. Fournier-Viger, J. C.-W. Lin, B. Vo, T. T. Chi, J. Zhang, and H. B. Le, "A survey of itemset mining," *Wiley Interdiscipl. Rev., Data Mining Knowl. Discovery*, vol. 7, no. 4, p. e1207, Jul. 2017.
[21] I. Tahyudin, H. Haviluddin, and H. Nanbo, "Time complexity of *a priori* and evolutionary algorithm for numerical association rule mining optimization," *Int. J. Sci. Technol. Res.*, vol. 8, no. 11, pp. 483–485, 2019.
[22] X. Yan, C. Zhang, and S. Zhang, "Genetic algorithm-based strategy for identifying association rules without specifying actual minimum support," *Expert Syst. Appl.*, vol. 36, no. 2, pp. 3066–3076, Mar. 2009.
[23] J. M. Luna, J. R. Romero, and S. Ventura, "Design and behavior study of a grammar-guided genetic programming algorithm for mining association rules," *Knowl. Inf. Syst.*, vol. 32, no. 1, pp. 53–76, Jul. 2012, doi: 10.1007/s10115-011-0419-z.
[24] J. M. Luna, P. Fournier-Viger, and S. Ventura, "Frequent itemset mining: A 25 years review," *Wiley Interdiscipl. Rev., Data Mining Knowl. Discovery*, vol. 9, no. 6, Nov. 2019.
[25] M. Friedman, "A comparison of alternative tests of significance for the problem of *m* rankings," *Ann. Math. Statist.*, vol. 11, no. 1, pp. 86–92, Mar. 1940.
[26] D. M. Rom, "A sequentially rejective test procedure based on a modified Bonferroni inequality," *Biometrika*, vol. 77, no. 3, pp. 663–665, 1990.
[27] Y. Hochberg, "A sharper Bonferroni procedure for multiple tests of significance," *Biometrika*, no. 75, pp. 800–803, 1988.
[28] S. García, A. Fernández, J. Luengo, and F. Herrera, "Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power," *Inf. Sci.*, vol. 180, no. 10, pp. 2044–2064, May 2010.



JOSÉ MARÍA LUNA received the Ph.D. degree in computer science from the University of Granada, Spain, in 2014. He is currently an Assistant Professor with the Department of Computer Science and Numerical Analysis, University of Cordoba, Spain. He is author of the two books related to pattern mining, published by Springer. He has published more than 30 articles in top ranked journals and international scientific conferences. He is author of two book chapters. He has also been involved in four national and regional research projects. He has contributed to three international projects. His research is focused on evolutionary computation and pattern mining

MYKOLA PECHENIZKIY is currently a Full Professor and the Chair of the Data Mining Research Group that is part of the Data and Artificial Intelligence cluster with the Department of Mathematics and Computer Science, Eindhoven University of Technology, The Netherlands. His core expertise and research interests include predictive analytics and knowledge discovery from evolving data, and in their application to real-world problems in industry, medicine and education. He has been a Principal Investigator of several nationally funded and industry funded projects that being inspired by challenges of the real-world applications aim at developing foundations for next generation of informed and responsible predictive analytics. Over the past decade, he has coauthored more than 100 peer-reviewed publications. He serves on several program committees and editorial boards of leading data mining and AI conferences (AAAI, ECMLPKDD, and IJCAI) and journals (*Data Mining and Knowledge Discovery* and *Machine Learning*).



WOUTER DUIVESTEIJN received the B.Sc. degrees in mathematics and computer science, and the M.Sc. degrees in fundamental mathematics and applied computing science from Universiteit Utrecht, The Netherlands, in 2005, 2007, and 2008, respectively, and the Ph.D. degree from Universiteit Leiden, The Netherlands, in 2013, with a thesis titled Exceptional Model Mining. He spent the subsequent three years as a Postdoctoral Researcher at Technische Universität Dortmund, Germany, the University of Bristol, U.K., and Universiteit Gent, Belgium, before moving to this current position in 2016. His research revolves around all aspects of Exceptional Model Mining: finding subgroups in datasets that are interpretable and display some kind of unusual behavior. Lately, he has started working on some fundamental problems in clustering. He is currently an Assistant Professor Data Mining with Technische Universiteit Eindhoven, The Netherlands. He has contributed to the organization of seven conferences and workshops, four of which as the General (co-)Chair, and he has been providing reviews as a member of more than 30 program committees and for ten journals, including participating in the DAMI Guest Editorial Board for the ECMLPKDD journal track.

SEBASTIÁN VENTURA (Senior Member, IEEE) received the B.Sc. and Ph.D. degrees in sciences from the University of Córdoba, Spain, in 1989 and 1996, respectively. He is currently a Full Professor with the Department of Computer Science and Numerical Analysis, University of Córdoba, where he also heads the Knowledge Discovery and Intelligent Systems Research Laboratory. He has published three books and about 300 articles in journals and scientific conferences, and he has edited three books and several special issues in international journals. He has also been involved in 15 research projects (being the coordinator of seven of them) supported by the Spanish and Andalusian governments and the European Union. His main research interests include data science, computational intelligence, and their applications. He is a Senior Member of the IEEE Computer, the IEEE Computational Intelligence, and the IEEE Systems, Man, and Cybernetics Societies, as well as the Association of Computing Machinery (ACM).