Noname manuscript No.
(will be inserted by the editor)

Outlying Property Detection with Numerical Attributes

Fabrizio Angiulli · Fabio Fassetti · Giuseppe Manco · Luigi Palopoli

the date of receipt and acceptance should be inserted later

Abstract The *outlying property detection problem* (OPDP) is the problem of discovering the properties distinguishing a given object, known in advance to be an outlier in a database, from the other database objects. This problem has been recently analyzed focusing on categorical attributes only. However, numerical attributes are very relevant and widely used in databases. Therefore, in this paper, we analyze the OPDP within a context where also numerical attributes are taken into account, which represents a relevant case left open in the literature. As major contributions, we present an efficient parameter-free algorithm to compute the measure of object exceptionality we introduce, and propose a unified framework for mining exceptional properties in the presence of both categorical and numerical attributes.

1 Introduction

Anomaly and outlier detection is a prominent research topic in data mining that focuses on approaches to discover unexpected elements in data populations. Historically, this research topic has been extensively investigated and several methods have been proposed which find outliers based on either statistical modeling or spatial proximity.

Despite the wide attention that anomaly detection has received in the literature, the related problem of anomaly *justification* happened to be largely underestimated. Typically, the result of an outlier detection algorithm over a population

This research has been partially supported by the PRIN project 20122F87B2 "Compositional Approaches for the Characterization and Mining of Omics Data" co-financed by the Italian Ministry of Education, University and Research.

F. Angiulli, F. Fassetti and L. Palopoli
DIMES Department, University of Calabria, 87036 Rende - Italy.
E-mail: {angiulli,fassetti,palopoli}@dimes.unical.it

G. Manco
Institute of High Performance Computing and Networks (ICAR-CNR), 87036 Rende - Italy.
E-mail: manco@icar.cnr.it.

of objects is a score associated with each object. The score, either binary or numerical, quantifies whether the related object significantly deviates from the rest of the population. Scoring the objects enables comparison and ranking, and ultimately the *detection* of the outlier objects; however, a score is a mere quantitative information, which provides little or no insight about the structural reasons why a given object is deemed as an outlier.

In order to cope with this problem, a possible approach would be to reformulate outlier detection algorithms in ways to allow them to provide, besides the outlier detection, also an interpretation of discovered outlierness in terms of discriminative features [13, 14]. A main disadvantage in this approach is the lack of generality, as it would require to reconsider the several anomaly detection algorithms proposed in the literature and to suitably reformulate them so that their output is not only the outlier objects and related scores, but also their justification in terms of discrimintative features.

Alternatively, one can formalize the problem as a more general, supervised learning task: given an object already deemed as an outlier, the objective is that of discovering the properties distinguishing such an outlier from the other database objects. We call this the *outlying property detection problem* (OPDP) [5, 16, 32, 39]. Notice that, under this perspective, OPDP is different from the outlier detection problem, as it allows also to focus on objects which, in principle, might not be outliers at all, and we would simply like to single out those features distinguishing the object under observation from the rest of the population. For example, [16] describes the case of candidates who apply for a position, for whom we would like to highlight weaknesses and strengths.

In the paper [5], the OPDP was studied and instantiated as follows: given a dataset characterized by certain attributes and a single input object known in advance to be anomalous in that dataset, the goal is to find a set of attributes explaining why this object is actually anomalous or, in other terms, detect the unexpected properties (if any) this anomalous object exhibits. The cited paper only considers the case where attributes whose values justify the given object anomaly are categorical. In several scenarios, though, the input dataset has numerical attributes which may well account for the anomaly of a given input anomalous object. The appropriate handling of such non-categorical attributes is a non-trivial problem left open in [5] and it is precisely the problem we face in this paper.

To see why this problem is relevant, consider the case of patient data, characterized by health parameters including several numerical features such as body temperature, blood pressure measurements, or cholesterol level. If a history of patients is available, then it is relevant to single out that subset of those parameters that mostly differentiate a sick patient from the healthy population. It is important to highlight here that the abnormal individual, whose peculiar characteristics we want to detect, is provided as an input to the problem, that is, this individual has been recognized as anomalous in advance by the virtue of some external information, mean or procedure.

This paper generalizes the approach proposed in [5], by extending it to the case of numerical attributes. Similar to the mentioned paper, the basic idea is to focus on a *property* featured by a given input anomalous object, where this property characterizes the outlierness of the object if there is a high *imbalance* between the density of the value exhibited by the object under consideration and the densities of the rest of the database values.

To elucidate, given a dataset DB and a query object q deemed to be abnormal (on the basis of available external knowledge), we claim that a property, that is a set of attributes, witnesses the abnormality of the object q if the combination of values q exhibits on these attributes is *anomalously rare* according to the joint distribution of the same attributes in the whole data set.

This rarity, or unbalance, can be unveiled by analyzing the curve of the cumulative distribution function (*cdf*) associated with the occurrence probability of the domain values. As explained in [5], relying on the *cdf* allows to correctly recognize exceptional properties independently of the form of the underlying probability density function (*pdf*): the former compares the occurrence probabilities of the domain values rather than directly comparing the domain values themselves.

When dealing with numerical attributes, a key aspect is being able to efficiently estimate both the *cdf* and the related *pdf*, as well as to exploit them to measure the associated imbalance. This is in fact the main contribution of the paper, which can be hence summarized as follows.

- We refine the *outlierness* measure proposed in [5], which is able to quantify the exceptionality of a property featured by the query object as a function of the underlying *cdf*. We analyze the main characteristics of the proposed measure, as well as its relationships and differences with related measures from the literature.
- We then present a parameter-free algorithm for computing both *pdf* and *cdf* for numerical attributes in time $O(n \log n)$, and show how the latter can be employed in the detection of outlier explanations.
- This result, combined with the results of [5], enables a general methodology for uniformly mining exceptional properties in the presence of both categorical and numerical attributes. This way, a fully automated support is provided to decode those properties determining the abnormality of the given object within the reference data context.

The rest of the paper is organized as follows. Section 2 introduces our mining task and discusses the relationships and differences with the outlier detection mining task. Section 3 introduces the outlierness measure and the concept of explanation. Section 4 describes the method for computing outlierness and determining associated explanations. Section 5 discusses experimental results, including a real-life case study. Finally, Section 6 presents conclusions.

2 Background and Related Work

To begin with, we next introduce some preliminary definitions and fix the notation. An *attribute* a is an identifier with an associated domain, also denoted $\mathbb{D}(a)$. Let $\mathbf{A} = a_1, \ldots, a_m$ be a set of m attributes[1]. Then, an *object* o on $\mathbf{A}$ is a tuple $o = \langle v_1, \ldots, v_m \rangle$ of m values, such that each v_i is a value in the domain of a_i. The *value* v_i associated with the attribute a_i in o will be denoted by $o[a_i]$. A *database* DB on a set of attributes $\mathbf{A}$ is a multi-set (that is, duplicate elements are allowed) of objects on $\mathbf{A}$.

[1] For the sake of simplicity and without loss of generality, we are assuming that an arbitrary ordering of the attributes in $\mathbf{A}$ has been fixed.

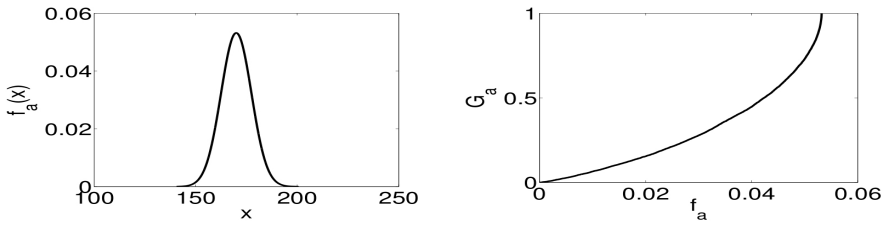

Fig. 1 Example of function $G_a(\cdot)$.

2.1 Outlier detection

Given a database DB over an attribute schema $\mathbf{A}$, an *outlier* is an object $o \in DB$ that is "exceptional", as it significantly differs from the rest of the data in DB. The notion of outlierness has been extensively studied in recent literature and, in this context, approaches to outlier detection can be classified as supervised, semi-supervised, and unsupervised.

Supervised methods exploit the availability of a labeled data set, containing observations already labeled as normal and abnormal, in order to build a model of the normal class [11]. Since usually normal observations are the great majority, these data sets are unbalanced and specific classification techniques must be designed to deal with the presence of rare classes.

Semi-supervised methods typically assume that only normal examples are given. The goal is to find a description of the data, that is a rule partitioning the object space into an accepting region, containing the normal objects, and a rejecting region, containing all the other objects [37]. These methods are also called one-class classifiers or domain description techniques, and they are related to novelty detection since the domain description is used to identify objects significantly deviating from the training examples.

Unsupervised methods search for outliers in an unlabelled data set by assigning to each object a score which reflects its degree of abnormality. Scores are usually computed by comparing each object with objects belonging to its neighborhood. Following [1], we can classify the main unsupervised approaches to outlier detection as probabilistic and statistical models, (e.g. [6,7,17,28,42]), where the outliers are modeled as data points which poorly fit the underlying data distribution; linear models, (e.g., [10,35,41]), where data points are embedded into a lower dimensional subspace in terms of linear relationships and outliers are modeled as data points exhibiting large residuals; proximity-based models (e.g. [4,9,23,25,27,34]), which model outliers as data points isolated from the remaining data; models for high-dimensional data (e.g. [2,20,22,31,33,40]) where it is assumed that outliers are characterized by unusual local behavior in lower dimensional subspaces.

2.2 Outlying property detection

All of the above mentioned methods focus on outlier identification and they do not provide *explanations* of why an identified outlier is exceptional. Indeed, in a sense,

the problem addressed here is to be considered orthogonal to the unsupervised outlier detection task, as we are interested in unveiling the specific properties that make an object $o \in DB$ special w.r.t. a population in DB. To this purpose, we assume that a set $o_1, \ldots, o_k$ of outliers is already given as input, and we are interested in characterizing each o_i. This can be accomplished by:

1. Detecting the subset $S_i \subseteq DB$ that represents a population, and such that $o_i \in S$. Intuitively, S represents a set of objects that share similar features.
2. Identifying a set $\{a_{i_1}, \ldots, a_{i_{n_i}}\} \in \mathbf{A}$ (with $n_i \leq m$) where $o_i[a_{i_1}, \ldots, a_{i_{n_i}}]$ substantially differentiates o_i from the other objects in S_i.

Subspace outlier mining techniques [2, 31] could in principle be used to extract information about outlier properties. However, the originary task considered thereof is different from the task investigated here, since subspaces in those approaches highlight the outlierness, whereas in our approach they represent a homogenous subpopulation upon which to compare a given property. The approaches [13, 14] consider the problem of detecting and interpreting local outliers, i.e., objects which are outliers *relative* to a subpopulation of neighbors, rather than the entire dataset. The outlierness is measured in a low-dimensional subspace capable of preserving the locality around the neighbors while at the same time maximizing the distance from the outlier candidate. Incidentally, the low-dimensional transformation also provides the insights for the relevant features which contribute most to the outlierness. Again, the problem tackled in these papers is different, since our aim is to characterize the outlierness of an input object, rather than to discover outliers in a population.

In [26], the authors focus on the identification of the intensional knowledge associated with distance-based outliers. First, they detect the distance-based outliers in the full attribute space and then, for each outlier, they search for the subspaces that better explain why it is exceptional. The exceptional object is not provided in input, but it belongs to the set of distance-based outliers of the dataset in the full attribute space. Furthermore, this setting models outliers which are exceptional with respect to the whole population, but it does not capture objects which are exceptional only with respect to homogeneous subpopulations.

Since the problem specification requires outliers to be given in the first place, it is in principle possible to divide the data into outliers and normal objects. Based on this partitioning, it would be possible to use contrast set mining techniques (like, e.g., [8]), in order to explain those outliers. However, this would result in high class imbalance, where the few instances in the rare class would trigger poor quality contrasts. Solutions to the rare case problem have been proposed, based on the enrichment of the originary dataset. In particular, [32] proposes an approach which follows this strategy. The authors assume that outliers are given as input, and their objective is to find an *explanatory subspace*, that is a subspace of the original numerical attribute space where the outlier shows the greatest deviation from the other points. The basic idea of the algorithm is to encode the notion of outlierness as separability: given an object o deemed as an outlier, one can devise an artificial set of points $\mathbf{x}$ oversampled from a gaussian distribution centered in o. Then, an outlierness of o can be measured in terms of the accuracy in separating the artificial points $\mathbf{x}$ from the other points in DB. Having encoded the outlierness as a classification problem, the explanatory subspace can hence be reduced to feature selection relative to such a classification problem.

In [16], the authors propose a method based on ranking and searching. In short, given a subspace, the authors rank the query object within the subspace according to a density-based outlierness measure. Then explanations are provided as those minimal subspaces for which the rank in minimum. Since the number of possible subspaces is exponential in the dimensionality of the data, the authors propose a heuristic to reduce the complexity of the search.

The techniques [32] and [16] can be considered paradigmatic of two different categories of approaches: those based on feature selection and those based on score and search. The authors in [39] discuss the connection between these two approaches and propose a hybrid solution. As for the feature selection phase, they aim at determinining the subspaces where a kernel density estimate of the data at the query point is minimized, by formulating a quadratic integer programming problem. Since solving the associated objective function is NP-hard, they relaxed it to a problem in the real domain that provides a ranking of the features. Having obtained the feature ranking they perform a score-and-search on the top-ranked features.

Besides other technical details, there are some substantial differences between the approaches [16, 32, 39] and the approach devised in this paper. First of all, it is assumed that outlierness is relative to the whole population. Their method returns individual subspaces where the query object is mostly outlying comparing to the other subspaces. By contrast, we are interested in modeling the scenario where outlierness can be expressed relative to a homogeneous subpopulation. That, is, we are interested in finding contextual rule-based explanations relative to a subpopulation of homogeneous objects. In this respect, the meaning of the two types of explanations is fundamentally different. Formally, with reference to the problem statement above, these methods assume in condition 1 that $S_i = DB$ and they only focus on condition 2.

To see why this can be problematic, onsider the dataset in Figure 2a, showing a skill-age relationship. According to the data, skill is directly proportional to age, with the highest level of skill reached on the mean at about 35 years of age. The red point on the left upper corner is a clear outlier, since it represents a young individual (eighteen years old) exhibiting a high skill score. In this scenario, the only way to characterize the outlierness of this individual is to look at the full feature space, since neither the age nor the skill subspace alone are able to explain abnormality. Hence, a subspace selection method will return the whole set of attributes, a not compeletely satisfactory explanation, since a deeper analysis should disclose the fact that, within the subpopulation of high skilled people, there's an individual which is characterized by a young age.

Furthermore, relying on separability can be misleading, as the accuracy of the method can be reduced in those subspaces where several points exhibit low density (and hence they do not properly characterize the exceptionality of the outlier). Consider the situation depicted in 2b. In this situation, the individual denoted by the red point can be clearly separated by all other points (the separation is the black line). Yet, the subspace Height/Skill does not properly characterize its outlierness, since the data is sparse in this subspace and all points exhibit a similar (low) density. This is further exhacerbated in situations where separability can be expressed in a non-linear fashion, as shown in 2c. Here, we can clearly see how the contour of the red point would allow to produce artificial point that enhance the separation within the Height/Skill subspace. Still, the density of the red point

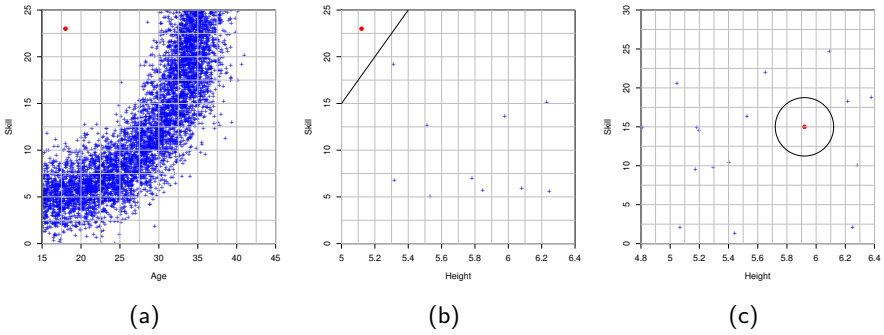

Fig. 2 Outlier explanation in problematic situations. (a) Although the whole *Age*/*Skill* feature space is suitable for the outlier (located left upper corner), the explanation should characterize it as the only individual exhibiting high skill within the subpopulation of individuals with young age. (b-c) The candidate outlier is clearly separable from all other individuals (by means of a line or a circle). However, the feature subspace does not properly characterize the outlier, as all individuals exhibit low density within the given subspace.

is not substantially different than those of the other points, and hence expressing exceptionality through the subspace is clearly inappropriate.

We claim that a more robust solution can be devised, in the style of [5]. In that paper, each subset of attributes is intended to represent a *property* of individuals. A property witnesses the abnormality of an object if the combination of values the object assumes on these attributes is very infrequent with respect to the overall distribution of the attribute values in the dataset, and this is measured my means of the so called *outlierness* score. This latter is based on measuring how much the frequency of the combination of values assumed by that object on those attributes is rare as compared to the frequencies associated with the combinations of values assumed on the same attributes by the other objects in the population.

A major problem with the outlierness score presented in [5] is that it was specifically designed and shown effective for categorical attributes. Hence the question arises on how to adapt that idea to a more general setting with both categorical and numerical attributes. Discretizing numerical attributes and applying the above technique to the discretized attributes makes more difficult to discover meaningful knowledge than working directly on the numerical data, for several reasons. First of all, the result of the analysis will strongly depend on the kind of discretization. This drawback is further exacerbated by the peculiarities of the outlierness measure, which assigns higher scores to very unbalanced distributions, and by contrast provides low scores to uniform frequency distributions. In this sense, the discretization process should be supervised by the outlierness score, in order to detect in the first place the bins capable of magnifying the score itself.

The notion of outlierness introduced here shares a common rationale with that already proposed in [5], but aims at overcoming the aforementioned drawbacks in presence of numerical data, as accounted for in the following sections.

Before concluding the section, we point out that there is a major difference between our measure and all of those employed in the techniques above discussed. Indeed, methods based on traditional outlier scores or on density estimation take

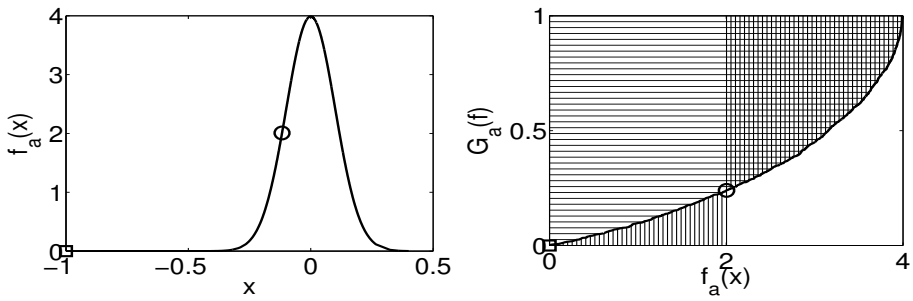

Fig. 3 Example of outlierness measure.

into account only the distribution of the data in the neighborhood of the outlier, while methods based on the concept of separability are not able to discriminate the amount of the deviation of the outlier from the rest of the population. Conversely, our measure is able to consider the data distribution in its entirety and, as such, it is specifically tailored for detecting outlying properties. We will substantiate the above claim in the following section.

3 Outliers and Explanations

In the following, we shall characterize populations in a "rule-based" fashion, by denoting the subset of DB that embodies them.

Formally, a *condition* on $\mathbf{A}$ is an expression of the form $a \in [l, u]$, where (i) $a \in \mathbf{A}$, (ii) $l, u \in \mathbb{D}(a)$, and (iii) $l \leq u$, if a is numeric, and $l = u$, if a is categorical. If $l = u$, the interval $I = [l, u]$ is sometimes abbreviated as u and the condition as $a \in I$ or $a = I$.

Let c be a condition $a \in [l, u]$ on $\mathbf{A}$. An object o of DB satisfies the condition c, if and only if $o[a]$ equals l, if a is categorical, or $l \leq o[a] \leq u$, if a is numerical. Moreover, o satisfies a set of conditions C if and only if o satisfies each condition $c \in C$. Given a set C of conditions on $\mathbf{A}$. The *selection* DB_C of the database DB w.r.t. C is the database consisting of the objects $o \in DB$ satisfying C.

Next, the definition of outlierness (Section 3) and of explanation (Section 3.3) are introduced.

3.1 Outlierness

The *outlierness* is a measure used to quantify the exceptionality of a property. The intuition underlying this measure is that an attribute makes an object exceptional if the relative likelihood of the value assumed by that object on the attribute is rare if compared to the relative likelihood associated with the other values assumed on the same attribute by the other objects of the database.

Let a be an attribute of A. We assume that a random variable X_a is associated with the attribute a, which models the domain of a. Then, with $f_a(x)$ we denote the pdf associated with X_a. The pdf provides a first indication on the outlierness

degree of a given value x, as usually we would expect low pdf values associated to outliers. However, the sole pdf value is not enough. A given pdf value represents a hypothetical "frequency" for that value in the sample under consideration. How typical is that "frequency" provides a better insight on the outlierness degree: a low pdf value in a population exhibiting low values only is not an indicator of an outlier, whereas an anomalous low pdf value in a population of significantly higher values denotes that the value under observation represents an outlier. Thus, analyzing how the values distribute on a pdf is the key for measuring the degree of outlierness.

Let X_a^f denote the random variable whose pdf represents the relative likelihood for the pdf f_a to assume a certain value. The cdf G_a of X_a^f is:

$$G_a(\varphi) = Pr(X_a^f \leq \varphi) = \int_0^{\varphi} Pr(X_a^f = v)\ \mathrm{d}v. \tag{1}$$

Example 1 Assume that the height of the individuals of a population is normally distributed with mean $\mu = 170cm$ and standard deviation $\sigma = 7.5cm$. Then, let a be the attribute representing the height, X_a is a random variable following the same distribution of the domain and $f_a(x)$ is the associated pdf, reported in the first graph of fig. 1. The pdf $f_a(x)$ assumes values in the domain $[0, f_a(\mu) = 0.0532] \subset \mathbb{R}$. Consider, now, the random variable X_a^f. The cdf $G_a(v)$ associated with X_a^f denotes the probability for f_a to assume value less than or equal to v. Then, $G_a(v) = 0$ for each $v \leq 0$ and $G_a(v) = 1$ for each $v \geq 0.0532$. To compute the value of $G_a(v)$ for a generic v, the integral reported in Equation (1) has to be evaluated. The resulting function is reported in the second graph of fig. 1.

The *Outlying Property Factor* $\mathrm{OPF}_a(o, DB)$ (or, simply, $\mathrm{OPF}_a(o)$) of the attribute a in o w.r.t. DB is defined as follows:

$$\mathrm{OPF}_a(o) = \Omega\left(\int_{f_a(o[a])}^{\sup(f_a)} (1 - G_a(f))\ \mathrm{d}f - \int_0^{f_a(o[a])} G_a(f)\ \mathrm{d}f\right). \tag{2}$$

Here, Ω denotes a function from $\mathbb{R}$ to $[0, 1]$ such that (i) $\Omega(x) = 0$ for $x < 0$, and (ii) $\Omega(x) = \Omega^+(x)$ for $x \geq 0$, where Ω^+ is any monotone increasing function mapping $\mathbb{R}_0^+$ to $[0, 1]$. In the following we employ the mapping:

$$\Omega^+(x) = \frac{1 - \exp(-x)}{1 + \exp(-x)}.$$

The first integral measures the *area above* the cdf $G_a(f)$ for $f > f_a(o[a])$, while the second integral measures the *area below* the cdf G_a for $f \leq f_a(o[a])$. Intuitively, the larger the first term, the larger the degree of unbalanceness between the occurrence probability of $o[a]$ and that of the values that are more probable than $o[a]$. As for the second term, the smaller it is, the more likely the value $o[a]$ to be rare. Thus, the outlierness value ranges within $[0, 1]$ and, in particular, it is close to zero for usual properties. By contrast, values closer to one denote exceptional properties.

Example 2 Consider fig. 3, reporting on the left a Gaussian distribution $f_a(x)$ (with mean $\mu = 0$ and standard deviation $\sigma = 0.1$). Consider the values $v_1 = -1$ and $v_2 = -0.12$, for which $f_a(v_1) \approx 0$ and $f_a(v_2) \approx 2$ hold. Assume that an outlier object o exhibits value v_1 on a. The associated outlierness $\mathrm{OPF}_a(o)$ corresponds to

the whole area (filled with horizontal lines) above the cdf curve, that is $\Omega(3.06) = 0.91$. For an object o' exhibiting value v_2 on a, instead, the associated outlierness corresponds to the difference between two areas (filled with vertical lines) detected at frequency 2, that is $\Omega(1.17 - 0.10) = 0.49$.

For the sake of clarity, in the above example we considered a pdf having a simple form. However, we wish to point out that our measure is able to correctly recognize exceptional properties irrespectively of the form of the underlying pdf, since it compares the occurrence probabilities of the domain values rather than directly comparing the original domain values.

3.2 Properties and comparison with outlier detection scores

We notice that if the argument of the function Ω approached to $+\infty$ it would be mapped to one. However, for any bounded probability density function f next we show that the argument of Ω is finite.

Theorem 1 *The argument of Ω in Equation* (2) *is upper (lower, resp.) bounded by* $\sup(f_a)$ *(*$-\sup(f_a)$*, resp.).*

Proof The argument of Ω in Equation (2) can be rewritten as:

$$\int_{f_a(o[a])}^{\sup(f_a)} (1 - G_a(f))\ \mathrm{d}f - \int_0^{f_a(o[a])} G_a(f)\ \mathrm{d}f =$$
$$= \int_{f_a(o[a])}^{\sup(f_a)} \mathrm{d}f - \int_0^{\sup(f_a)} G_a(f)\ \mathrm{d}f =$$
$$= \sup(f_a) - f_a(o[a]) - \int_0^{\sup(f_a)} G_a(f)\ \mathrm{d}f.$$

Since, $G_a(f) \in [0, 1]$ and $f_a(o[a]) \in [0, \sup(f_a)]$, the upper bound can be obtained by considering $f_a(o[a]) = 0$ and $G_a(f) = 0$:

$$\sup(f_a) - f_a(o[a]) - \int_0^{\sup(f_a)} G_a(f)\ \mathrm{d}f \leq \sup(f_a),$$

while the lower bound can be obtained by considering $f_a(o[a]) = \sup(f_a)$ and $G_a(f) = 1$:

$$\sup(f_a) - f_a(o[a]) - \int_0^{\sup(f_a)} G_a(f)\ \mathrm{d}f \geq -\sup(f_a).$$

We point out that the OPF measure is carefully tailored to the task at hand, since it is able to compare the specificity of the value assumed by the outlier on the property under analysis with the specificity of all the other values.

This is due to the fact that OPF considers the data distribution in its entirety as opposed to the majority of the outlier scores designed within the data mining literature that approach the problem by taking into account only the distribution of the data in the neighborhood of the object. Indeed, traditional outlier detection

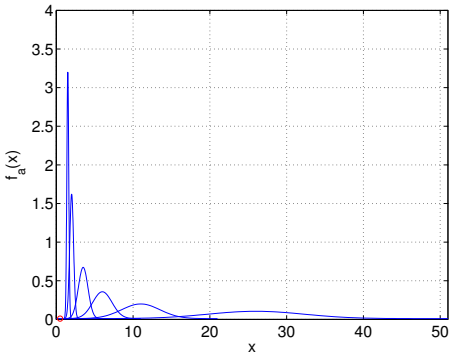

Fig. 4 Example density plots: each curve is associated with a different truncated Gaussian distribution with its own μ and σ parameter. The red circle (located in 1) is an outlier point sampled from a uniform distribution separated from the truncated Gaussians.

measures are designed to rank objects according to their exceptionality with respect to a given set of properties, while our measure is designed to rank properties according to their exceptionality with respect to a given object.

Notice that the above characteristics is shared both by the so called "global" methods (such as the distance-based ones) and by those known as "local" ones (such as the density-based LOF, MDEF, and others). Indeed, despite their names, the difference between those techniques relies in the fact that to construct the outlier score the former consider only the neighborhood of the object, while the latter take also advantage of the neighborhood of its neighbors. Thus, in order to declare points either as inliers or outliers both rely uniquely on the distribution of the objects within their neighborhood.

One can argue that the ranking of the outlier scores accomplished by the above methods takes anyway into account the whole distribution of the data, but we notice this is not really the case since traditional outlier detection techniques usually assign the same score to outliers immersed in very different data distributions. Consequently, they are not suitable to rank properties with respect to their exceptionality To illustrate this behavior, consider the distributions reported in Figure 4. We assume that each curve is associated with a different property and is built as follows: 1% of the data (call it $\mathcal{N}_1$) is uniformly distributed in the range $[0, 2]$, while the remainder 99% of the data (call it $\mathcal{N}_2$) is distributed according to a truncated normal distribution with mean μ and standard deviation σ having support $[\mu - 4\sigma, \mu + 4\sigma]$. The standard deviation σ is set to $(\mu - 2)/4$, so that $\mathcal{N}_2$ ranges from 2 to $2+8\sigma$. As for μ, we consider six different values: $1.5, 2, 3.5, 6, 11$ and 25.5.

For each data distribution, we generated a dataset of 100,000 objects and considered as outlier *out* an object whose value approaches 1, namely an object laying in the middle of $\mathcal{N}_1$. Then, according to the data distribution, the neighborhood of the oulier is not affected by $\mathcal{N}_2$.

The following table reports, for the object *out*, its outlier scores according to the proposed measure and two known ones a global (KNN [4]) and a local (LOF [9]) score.

μ	OPF	KNN	LOF
1	0.798	0.350	1.007
2	0.493	0.350	1.007
5	0.203	0.350	1.007
10	0.094	0.350	1.007
20	0.038	0.350	1.007
50	0.009	0.350	1.007

As reported in table, the values of OPF are strongly affected by the distribution of the whole population and, in particular, the larger is the variance of $\mathcal{N}_2$, the more the values of $\mathcal{N}_1$ and $\mathcal{N}_2$ tend to be equipossible and, then, the smaller become the outlierness of *out*. Conversely, both KNN and LOF consider solely the neighborhood of *out*. The resulting outlier score associated with *out* does not depend on changing $\mathcal{N}_2$. As a consequence, using such methods to rank properties appears inappropriate.

3.3 Explanations

Explanations are used in our framework to provide a justification of the anomalous value characterizing an outlier. Intuitively, an attribute $a \in \mathbf{A}$ of o that behaves normally with respect to the database as a whole, may be unexpected when the attention is restricted to a portion of the database. We shall again call this anomalous attribute a *property* of o. Relevant subsets of the database upon which to investigate outlierness can be hence obtained by selecting the database objects satisfying a condition, and such that a property is exceptional for o w.r.t. that data subset.

A condition c (set of conditions C, resp.) is, intuitively, an *explanation* of the property a if $o \in DB_c$ ($o \in DB_C$, resp.) and a is exceptional for o w.r.t. DB_c (DB_C, resp.) (i.e., the value $\mathrm{OPF}_a(o, DB_C)$ is close to 1). Finally, the *outlierness* of the set property a in o w.r.t. DB with *explanation* C is defined as $OPF_a^C(o, DB) = \mathrm{OPF}_a(o, DB_C)$.

It is worth noticing that, according to the relative size of DB_C, not all the explanations should be considered equally relevant. In the following, we concentrate on σ-explanations, i.e., conditions C such that $\frac{|DB_C|}{DB} \geq \sigma$, where $\sigma \in [0, 1]$ is a user-defined parameter.

Thus, given an object o of a database DB on a set of attributes $\mathbf{A}$ and parameter $\sigma_\theta \in [0, 1]$ and $k_n > 0$, the problem of interest here is: *Find the k_n pairs (E, p), such that $E \subseteq \mathbf{A}$, $p \in \mathbf{A} \setminus E$, and E is a σ_θ-explanation, scoring the highest values of $OPF_p^E(o, DB)$*. Such an attribute p is also called an *outlying property* (with explanation E).

4 Detecting Outlying Properties

In order to detect outlying properties and their explanations, we need to solve two basic problems: (1) computing the outlierness of a certain multiset of values and (2) determining the conditions to be employed to form explanations. The strategies we have designed to solve these two problems exploit a common framework, which is based on Kernel Density Estimation (KDE). Specifically, given a numerical

Function *ComputePDF*$(\mathbf{x}, h, \mathbf{w})$

Input: $\mathbf{x} = x_1, \ldots, x_n$: a set of values
h : a bandwidth
$\mathbf{w} = w_1, \ldots, w_n$: a set of weights
Output: $\hat{\mathbf{f}} = \hat{f}_1, \ldots, \hat{f}_n$: the density estimate at points $\mathbf{x}$

Sort the sequence $L = x_1^l, \ldots, x_n^l$, according to the values $\{x_i - \frac{w_i h}{2} : 1 \leq i \leq n\}$, and record the associated indexes $l_1, \ldots, l_n$;
Sort the sequence $U = x_1^u, \ldots, x_n^u$, according to the values $\{x_i + \frac{w_i h}{2} : 1 \leq i \leq n\}$, and record the associated indexes $u_1, \ldots, u_n$;
for $i = 1$ **to** n **do**
  Find the last element $x_{l^*}^l$ of L not greater than x_i;
  Find the first element $x_{u^*}^u$ of U not smaller than x_i;
  Set J to $\{l_1, l_2, \ldots, l^*\} \cap \{u^*, \ldots, u_{n-1}, u_n\}$;
  Set $\hat{f}_i$ to $\frac{1}{nh} \sum_{j \in J} \frac{1}{w_j}$;
return $(\hat{f}_1, \ldots, \hat{f}_n)$;

Function *EstimatePDF*$(\mathbf{x})$

Input: $\mathbf{x} = x_1, \ldots, x_n$
Output: $\hat{\mathbf{f}} = \hat{f}_1, \ldots, \hat{f}_n$

Set h to $1.06 \cdot \mathrm{std}(\mathbf{x}) \cdot n^{-1/5}$ `// Rule of thumb`
Set $\boldsymbol{\beta}$ to $(1, \ldots, 1)$;
for $t = 1$ **to** 5 **do**
  $\hat{\mathbf{f}} = \mathit{ComputePDF}(\mathbf{x}, h, \mathbf{w})$;
  $f_m = (\prod_{i=1}^n \hat{f}_i)^{1/n}$;
  for $i = 1$ **to** n **do**
    Set β_i to $(f_m / \hat{f}_i)^{1/2}$;
return $(\hat{f}_1, \ldots, \hat{f}_n)$;

attribute a, in order to estimate the pdf f_a we exploit *generalized kernel density estimation* [24], according to which the estimated density at point $x \in \mathbb{D}(a)$ is

$$\hat{f}_{\mathbf{m},\mathbf{w},\mathbf{b}}(x) = \left(\sum_{i=1}^{k} w_i \right)^{-1} \sum_{i=1}^{k} \frac{w_i}{b_i} K \left(\frac{x - m_i}{b_i} \right), \tag{3}$$

Here, K is a kernel function, and $\mathbf{m} = (m_1, \ldots, m_k)$, $\mathbf{w} = (w_1, \ldots, w_k)$ and $\mathbf{b} = (b_1, \ldots, b_k)$ are k-dimensional vectors denoting the *kernel location*, *weight*, and *bandwidth*, respectively. The above mentioned strategies are detailed next, together with the method for mining outlying properties.

4.1 Outlierness computation.

In order to compute the outlierness, we specialize formula in Equation (3) by setting $\mathbf{m} = (x_1, \ldots, x_n)$ and $\mathbf{w} = \mathbf{1}$, thus obtaining

$$\hat{f}_a(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{b_i} K \left(\frac{x - m_i}{b_i} \right), \tag{4}$$

Function *ComputeOutlierness*(o, a, DB)

Input: o : an outlier object
a : a dataset attribute
DB : a dataset
Output: *out* : the outlierness of the attribute a in o w.r.t. DB

1 Set $\mathbf{x}$ to $DB[a]$;
2 Set $\hat{\mathbf{f}}$ to *EstimatePDF*$(\mathbf{x})$;
3 Determine the sequence $\tilde{f}_1, \ldots, \tilde{f}_n$, by sorting the elements of the set $\{\hat{f}_i : 1 \leq i \leq n\}$;
4 **for** $i = 1$ **to** n **do**
5 Set G_i to $|\{f_j \leq \tilde{f}_i : 1 \leq j \leq n\}|/n = i/n$;
6 Let i^* be such that $\tilde{f}_{i^*}$ is the value in $\hat{\mathbf{f}}$ associated with $o[a]$;
7 Set *out* to 0;
8 **for** $i = i^* + 1$ **to** n **do**
9 Set $out = out + (\tilde{f}_i - \tilde{f}_{i-1})(2 - G_i - G_{i-1})/2$;
10 **for** $i = 2$ **to** i^* **do**
11 Set $out = out - (\tilde{f}_i - \tilde{f}_{i-1})(G_i + G_{i-1})/2$;
12 **return** $\Omega(out)$;

where $x_1, \ldots, x_n$ are the values in $\{y[a] : y \in DB\}$, each term b_i is equal to $h\beta_i$, with h a global bandwidth and $\prod_{i=1}^{n} \beta_i = 1$. The rationale underlying this choice is that we want that each value at hand ($\mathbf{m} = \mathbf{x}$) contributes in equal manner ($\mathbf{w} = \mathbf{1}$) to the estimation of the underlying pdf. Moreover, we employ the *Parzen window* kernel function, that is $K(x) = 1$, for $|x| \leq 1/2$, and $K(x) = 0$ otherwise, since this kernel represents a good trade-off between simplicity of computation and accuracy. Indeed, we are able to provide a parameter-free function that computes an accurate estimate $\hat{f}_a$ of the pdf f_a in time $O(n \log n)$. We also notice that, since the outlierness depends on the cdf of the pdf values, this greatly mitigates the impact of the non-smoothness of the estimate of the pdf through Parzen windows, other than making the measure robust w.r.t. deviations of the estimate from the real distribution.

Let $\mathbf{x}$ denote the vector $(x_1, \ldots, x_n)$, and $\boldsymbol{\beta}$ denote the vector $(\beta_1, \ldots, \beta_n)$. The function *ComputePDF* computes the vector $\hat{\mathbf{f}}$, whose generic element $\hat{f}_i$ represents the value of density $\hat{f}_a(x_i)$ at point x_i, as computed by exploiting Equation (4). In particular, when the Parzen window is employed, the computation of $\hat{f}_a(x)$ reduces to determine the value $\frac{1}{nh} \sum_{j \in J} \frac{1}{\beta_j}$, where J is the set containing the indexes j of the elements x_j of $\mathbf{x}$ such that $\left|\frac{x - x_j}{\beta_j h}\right| \leq \frac{1}{2}$ or, in other words, such that $x_j - \frac{\beta_j h}{2} \leq x$ and $x \leq x_j + \frac{\beta_j h}{2}$. The set J associated with a specific value x, can be determined by performing two binary searches and one intersection, as shown in the pseudo-code. Since this computation is executed n times, this leads to an overall cost $O(n \log n)$.

The function *EstimatePDF* is in charge of computing the right values for the parameters h and $\boldsymbol{\beta}$. It exploits the algorithm for calculating a variable bandwidth KDE [36]. The method starts with a density estimate by using a fixed-bandwidth kernel, with h determined by means of a rule of thumb [38] (see Function *EstimatePDF*, line 1) and $\boldsymbol{\beta} = \mathbf{1}$. Then, the bandwidths β_i are updated to a value which is inversely related to the density estimate. It was observed [19] that itera-

tions produce little changes: hence, we execute it a fixed number of times in order to have a computational cost of $O(n \log n)$.

The function *ComputeOutlierness* exploits *EstimatePDF* to compute the numerical estimate $\hat{\mathbf{f}}$ of the pdf f_a. Then, it computes the distribution function G_a (see Equation (1)) by setting G_i to $|\{f_j \leq \tilde{f}_i : 1 \leq j \leq n\}|/n$, that is i/n, and, finally, the outlierness value *out* (see Equation (2)), by performing a numerical integration, which costs $O(n)$. Thus, the dominating operations of *ComputeOutlierness* are the call to the function *EstimatePDF* and the sorting of the elements of $\hat{\mathbf{f}}$, with a resulting overall cost $O(n \log n)$.

4.2 Condition building

Proper conditions are the basic building blocks for the explanations. To single them out, our strategy consists in finding, for each attribute a, the "natural" interval I_a including $o[a]$, namely, an interval of homogeneous values on a. The rationale underlying this choice is to avoid the risk of overfitting: a guided search for a proper condition can easily yield an ad-hoc fragment of the data where the outlierness measure is "artificially" maximized. On the other side, proper conditions which encode the genuine intervals for each attribute domain can have a relevant impact on the detection of significant outlier explanations.

The search for feasible intervals still relies on adopting the kernel density family introduced above. In practice, for each attribute a, we estimate f_a by means of $\hat{f}_{\mathbf{m},\mathbf{w},\mathbf{b}}$. This latter function can be interpreted as a mixture density over the parameter sets $\mathbf{m}, \mathbf{w}, \mathbf{b}$. Also, the adoption of a Gaussian kernel

$$K(x) = \phi(x) = (2\pi)^{-1/2} \exp(x^2/2)$$

allows to estimate the parameter set via a standard EM-based maximum likelihood approach. In particular, the approach will partition the values $\{y[a] : y \in DB\}$ into a set of j^* disjoint intervals $I_a^1, \ldots, I_a^{j^*}$ and, then, the interval $I_a^{j_o}$ which $o[a]$ belongs to will be selected as the proper condition I_a for the attribute a.

The resulting iterative scheme draws from [24], and updates locations and bandwidths according to the following equations:

$$m_j = \frac{1}{\sum_i \gamma_{ij}} \sum_{i=1}^{n} x_i \gamma_{ij}, \tag{5}$$

$$b_j^2 = \frac{1}{\sum_i \gamma_{ij}} \sum_{i=1}^{n} \gamma_{ij} (x_i - m_j)^2. \tag{6}$$

Here, γ_{ij} represents the mixed probability that value i is associated with the j-th interval and, in its turn, is computed at each iteration as:

$$\gamma_{ij} = \frac{w_j \phi_{b_j}(x_i - m_j)}{\hat{f}_{\mathbf{m},\mathbf{w},\mathbf{b}}(x_i)} \tag{7}$$

We also adapt the annihilation procedure proposed in [18], which allows for an automatic estimation of the optimal number j^* of intervals, as well as to ignore

Function *ComputeInterval*(o, a, DB)

Input: o : an outlier object
a : a dataset attribute
DB : a dataset
Output: *out* : the natural interval of the attribute a in o w.r.t. DB

```
set x to DB[a];
set j* to √n;
initialize γ_ij randomly, ∀i ∈ [1..n] and ∀j ∈ [1..j*];
repeat
    for j = 1 to j* do
        update w_j // Equation (8)
        if w_j > 0 then
            update m_j, b_j // Equation (6)
            update γ_ij, ∀i ∈ [1..n] // Equation (7)
        else
            eliminate the jth component;
            set j* to j* − 1;
until increase in likelihood is negligible;
assign x_i to the interval I_a^{j_i} s.t. j_i = arg max_j γ_ij;
let I_a^{j_o} be the interval which o[a] belongs to;
set l_a to min_i{x_i | x_i ∈ I_a^{j_o}};
set u_a to max_i{x_i | x_i ∈ I_a^{j_o}};
return [l_a, u_a]
```

initialization issues. The estimation of the parameters is accomplished iteratively for each interval I_a^j, where each weight is computed as

$$w_j = \frac{\max\{0, \sum_{i=1}^{n} \gamma_{ij} - \frac{n}{2}\}}{\sum_{j=1}^{j^*} \max\{0, \sum_{i=1}^{n} \gamma_{ij} - \frac{n}{2}\}} \tag{8}$$

Whenever a weight equals to 0, the contribution of its component annihilates in the density estimation. As a consequence, the iterative procedure can start with a high initial value j^*, and the initialization of each mixed probability can be done randomly without compromising the final result. Function *ComputeInterval* reports the overall scheme. We also call the interval reported by this function the *natural interval* of a in o w.r.t. DB.

4.3 The mining method

Given a dataset DB on the set of attributes $\mathbf{A} = \{a_1, \ldots, a_m\}$, an outlier object o, parameters $\sigma_\theta \in [0, 1]$, a positive integer k_n (denoting the desired number of outlying pairs), and positive integer $k_\theta \leq m$ (representing an upper bound to

Algorithm 1: *OutlyingPropertyDetector*(o, DB)

Input: o : an outlier object
DB : a dataset
Output: $\mathcal{OP}$: the set of minimal explanation-property pairs of o in DB

```
1  set Ω_θ to 0 /* outlierness threshold */
   // First phase
2  foreach attribute a_i ∈ A do
3      Compute interval I_{a_i};
   // Second phase
4  foreach attribute p ∈ A do
5      set L_1 to {c_i ≡ a_i ∈ I_{a_i} s.t. |DB_{c_i}|/|DB| ≥ σ_θ};
6      set j to 2;
7      while j ≤ k_θ and L_{j-1} ≠ ∅ do
8          set E_j to {C ∪ {c} s.t. C ∈ L_{j-1} and c ∈ ⋃L_{j-1} and c ∉ C};
9          foreach C ∈ E_j do
10             if |DB_C|/|DB| ≥ σ_θ then
11                 if out_p^C(o, DB) > Ω_θ then
12                     if |OP| = k_n then
13                         let (C_min, p_min) be minpair(OP);
14                         set OP to OP \ {(C_min, p_min)};
15                     set OP to OP ∪ {(C, p)};
16                     if |OP| = k_n then
17                         let (C_min, p_min) be minpair(OP);
18                         set Ω_θ to OPF_{p_min}^{C_min}(o, DB);
19                 else
20                     set L_j to L_j ∪ {C};
21         set j to j + 1;
22 return OP
```

the size of an acceptable explanation [2]), the algorithm *Outlying-Property Detector* computes the top k_n pairs (E, p), with $|E| \leq k_\theta$ and $p \in \mathbf{A} \setminus E$, such that:

1. E is a σ_θ-explanation, and
2. (E, p) is *minimal*, that is, there is no pair (E', p) with $E' \subset E$ for which both points 1 and 2 hold.

The algorithm consists of two main phases. During the first phase, for each attribute $a_i \in \mathbf{A}$, the interval I_{a_i} and, hence, the associated condition $a_i \in I_{a_i}$, is determined by means of the procedure described in Section 4.2. Given the set of conditions $S = \{a_1 \in I_{a_1}, \ldots, a_m \in I_{a_m}\}$ on the m attributes in $\mathbf{A}$, the second phase exploits an apriori-like strategy [3] in order to search for the pairs (E, p) with $E \subseteq S$ satisfying the above mentioned conditions. The computed pairs are accumulated in the set $\mathcal{OP}$, which represents the output of the algorithm. In particular, we should store in $\mathcal{OP}$ the top k_n pairs. So the algorithm starts by setting Ω_θ to 0 and, then, all the pairs scoring an outlierness greater than 0 are stored in $\mathcal{OP}$ as long as the size of $\mathcal{OP}$ reaches k_n. When k_n pairs are in $\mathcal{OP}$, the algorithm exploits

[2] We point out that k_θ has a twofold function: it allows the analyst to control the complexity of the mined patterns and it speeds up the algorithm execution. However, by setting k_θ to m the algorithm is able to detect explanations of any length, while pruning the search space and avoiding overfitting by means of the threshold support.

the function $minpair(\mathcal{OP})$ in order to find the pair belonging to $\mathcal{OP}$ scoring the lowest value of outlierness, sets Ω_θ to the outlierness of such a pair and inserts a new pair in $\mathcal{OP}$ if and only if its outlierness is greater than Ω_θ. In such a case the pair belonging to $\mathcal{OP}$ scoring the lowest value of outlierness is removed from $\mathcal{OP}$.

As for the cost of the above procedure, the first step basically depends on the rate of convergence of the EM algorithm. By assuming that the number k of kernel locations is initially set to $\sqrt{n}$, the basic iteration is $O(n^{3/2})$. Notice, however, that interval components annihilate early in the first iterations, so that we can assume that the number of intervals k^* is bounded to a constant value. Thus, the overall complexity of the first step is linear in the size of the data and the number of iterations. Clearly, the rate of convergence of the algorithm is also of practical interest, and it is usually faster than the quadratic convergence typically available with Newton-type methods. [15] shows that the rate of convergence of the EM algorithm is linear.

As far as the second step is concerned, computing the outlierness costs $O(n \log n)$. Since these two sub-steps are executed at most $O(m^{k_\theta})$ times, the overall cost of step 2 is $O(m^{k_\theta} n \log n)$. However, notice that the apriori-like strategy greatly reduces the size of portion of the search space to be explored, so that the total number of conditions explored in practice is much smaller.

5 Experimental results

In this section, experimental results conducted by employing the proposed technique are described.

Experiments are organized as follows. In Section 5.1 we evaluate our approach on some datasets from the UCI Machine Learning repository with the aim of showing its effectiveness and of studying its scalability. In Section 5.2 we study trustability of the technique by performing a sensitivity analysis of the algorithm to the parameter σ_θ with reference to good and bad outliers. In Section 5.3 we perform a comparison of our approach with that proposed in [5] designed for categorical attributes. Finally, in Section 5.4 we describe some specific real-life case study where the technique was profitably exploited.

In these experiments, whenever a ground truth is not readily available, it is represented by outlier tuples detected by resorting to the feature bagging algorithm described in [29]. Briefly, the technique detects outliers by iteratively running a base outlier detection algorithm on a subset of the available attributes. Outlier detected in the various runs are then scored by adopting a *combine* function which assigns a score to each outlier.

The bagging technique was instantiated by exploiting the base OD method described in [4], where the parameters are set to produce just a single outlier. Further, the adopted *combine* technique simply scores outliers on the basis of the positive responses they get within the iterations.[3]

Notice that the feature bagging technique boosts the robustness of base outlier detection techniques. At the same time, it makes quite difficult to manually infer (e.g., by means of visualization techniques) the reasons why a specific tuple was

[3] In practice, if a tuple is detected as an outlier in a given iteration, it gets a positive score. Scores are then summarized in the combine function, and tuples are sorted according to the scores.

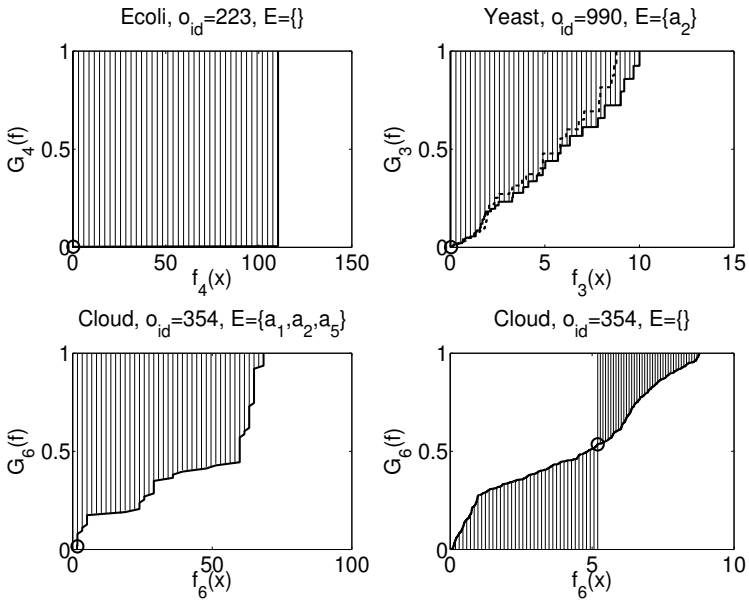

Fig. 5 Experimental results on the *Ecoli*, *Yeast*, and *Cloud* datasets.

detected to be an outlier. In fact, a tuple can be singled out as an outlier for a combination of factors which, in turn, depend on different attribute subsets. As a consequence, the analysis of the outliers produced with such a technique provides a significant benchmark of the effectiveness of our outlier explanation technique.

5.1 Evaluation and execution time

We employed three real datasets from the UCI Machine Learning repository [30]. The first two datasets, namely *Ecoli* (with 336 instances and 7 attributes) and *Yeast* (with 1,484 instances and 8 attributes), contain information about protein localization sites. The third database, called *Cloud*, contains information about cloud cover and includes 1,024 instances with 10 attributes.

The support threshold σ_θ was set to 0.2, the maximum number k_θ of conditions in the explanation was set to 3, and the number k_n of top explanation-property pairs to 5. The following table reports the explanation-property pairs scoring the maximum outlierness value.

DB	**o**	$\mathbf{out_p^E(o)}$	**p**	**E**
Ecoli	223	1.000	a_4	$\emptyset$
Yeast	990	0.997	a_3	$\{\ a_2 \in [0.13, 0.38]\ \}$
Cloud	354	1.000	a_6	$\{\ a_1 \in [1.0, 6.7],$ $a_2 \in [134.9, 255.0],$ $a_5 \in [2{,}450.5, 3{,}211.5]\ \}$

In the third column, we report the outlierness value, in the fourth column the attribute associated with the property, and in the fifth column the explanation.

Figure 5 reports the functions $G_a(f)$ associated with the objects considered in the experiments.

Figure 5, at its top left, reports the area associated with the property a_4 and empty explanation for the object 223 in the *Ecoli* database. The property a_4 is the attribute *Presence of charge on N-terminus of predicted lipoproteins*. The object 223 is the only object assuming value 0.5 on this attribute, while all the other objects assume value 1.0. As a consequence, this attribute is a clear outlying property with respect to the whole database and, in fact, the associated explanation is empty.

Figure 5, at its top right, reports the area associated with the property a_3 for the object 990 in the *Yeast* database. The attribute a_3 is *Score of the ALOM membrane spanning region prediction program*. The solid line represents the curve $G_{a_3}(f)$ obtained when the explanation $\{a_2\}$ is taken into account, while the dashed line represents the curve $G_{a_3}(f)$ obtained for the empty explanation. We note that, by taking the explanation into account, an improvement of the outlierness value is achieved, even if the property a_3 is quite interesting also with respect to the whole database.

Finally, Figure 5, at its bottom left, reports the area associated with the property a_6 and the explanation $\{a_1, a_2, a_5\}$ for the object 354 in the *Cloud* database. The attribute a_6 is the *Visible entropy*, while the explanation attributes are *Visible mean*, *Visible max* and *Contrast*. Figure 5 on the bottom right reports the area associated with the same property, but for the empty explanation. In this case, it is worth noting that the property a_6 turns out not to be exceptional with respect to the whole database (the outlierness value is approximatively 0.3) but it becomes rather exceptional with respect to the subpopulation selected by the explanation.

The following table reports the execution times associated with the experiments.[4]

DB	**Condition Building**	**Top Pairs Computation**
Ecoli	3.19 *sec*	0.19 *sec*
Yeast	23.19 *sec*	0.83 *sec*
Cloud	167.53 *sec*	1.20 *sec*

We also study the effects of the k_θ parameter on the performance. Greater values of k_θ trigger larger explanations. Since the support for large explanations decreases as well, the performance tends to become stable for large values of k_θ. Figure 6 reports this phenomenon: the curves tend to flatten for increasing values, on all datasets.

5.2 Trustability and sensitivity

A nice feature of our algorithm is the robustness of the outlying property factor to the support σ_θ. If an explanation that characterizes the outlier exists then a property will eventually be associated to such an explanation, with an high *OPF* for all values of σ_θ less or equal than the actual support of explanation. That is, if we start with the highest value of σ_θ and progressively decrease this value, the algorithm will eventually detect all relevant pairs and it will associate an high

[4] Experiments have been performed on an Intel Core i7 2.3GHz based computer by using the Java programming language.

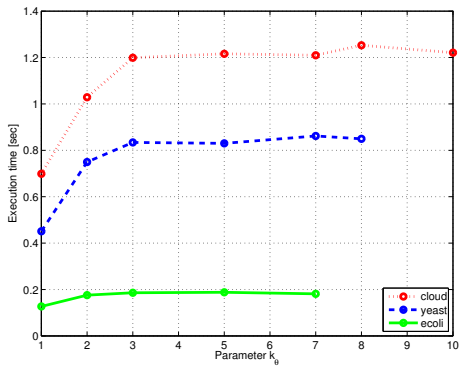

Fig. 6 Total Outlier Computation Time for *Ecoli*, *Yeast*, and *Cloud* datasets.

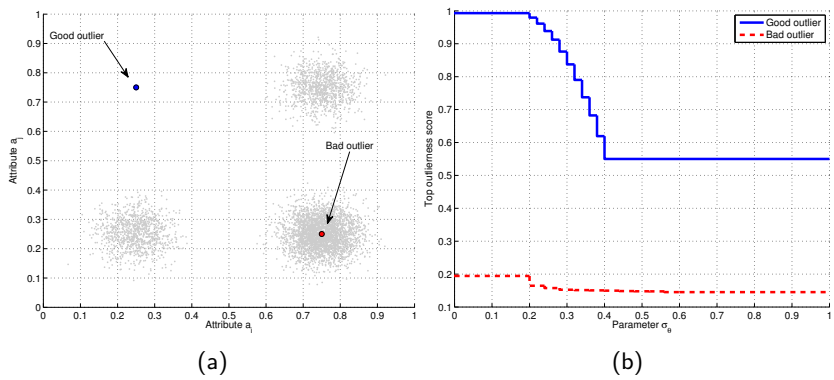

Fig. 7 Sensitivity to σ_θ. (a) Structure of the datasets exploited for the experiments. The red point (located in $(0.75, 0.25)$) and blue point (located in $(0.25, 0.75)$) represent respectively a bad and good outlier. (b) Outlierness degree of the good and bad outlier, for increasing values of σ_θ.

OPF to them. By contrast if an object does not exhibit any relevant explanation the *OPF* value of each property will be low regardless of σ_θ.

To verify this capability of the algorithm to detect faithful explanations we devise the following experiment. We considered a family of synthetically generated datasets, both consisting of 5,000 points and 10 attributes. The values in 8 attributes are randomly chosen from a uniform distribution with range $[0, 1]$. For the remaining 2 attributes, we group values in three gaussian clusters, cluster $C1$ having mean (μ, μ), cluster $C2$ having mean $(1-\mu, 1-\mu)$, and cluster $C3$ having mean $(1-\mu, \mu)$, where $\mu = 0.25$. All of the three clusters have a standard deviation 0.05.

On each dataset, we consider two artificial points. One point represents a "good outlier" o_g with an associated explanation-property pair (E_{o_g}, p_{o_g}). The "good outlier" o_g is located in $(\mu, 1-\mu)$. Thus, the explanation property pair exists in the selected subspace (a_i, a_j) and it has the form $(E_{o_g} = \{a_i \in [0, 0.5]\}, p_{o_g} = \{a_j\})$, meaning that among the objects having a low value for a_i, the object o_g exhibits

an exceptional value for a_j. The exceptionality is relative to the size of $C1$ and $C2$, which characterize the explanation, and no exceptionalily is likely to exhist within the remaining dimensions.

We also take into account a "bad outlier" by adding an object o_b located close to the center of cluster $C3$. In principle, the exceptionality of this object can only exhibited in the same subspace (a_i, a_j), and it is inversely proportional to the size of $C3$.

For each dataset we varied the size of the three clusters. Specifically, we varied a parameter $\varrho \in [0.2, 0.4]$ and enforced each of the clusters $C1$ and $C2$ to contain the ϱ fraction of the dataset points, whereas $C3$ contains the remaining $1 - 2\varrho$ fraction. Figure 7a reports, for such an example dataset, the projection on the two selected dimensions. All datasets exhibit a similar structure, but the cluster have a different population. The "good outlier" o_g is colored in blue, , while the "bad outlier" o_b is colored red.

Figure 7b reports the top outlierness score associated with the objects o_g (solid red line) and o_b (dashed blue line). Values are obtaining by averaging the outcomes of 100 runs on randomly generated datasets of the above family. The figure shows that the outlierness is increasing for decreasing values of σ_θ and, in particular, each outlier admits a threshold under which the outlierness degree is maximal. Also, the minimal outlierness value is given by a threshold σ_θ for which no explanations can be found. With the given cluster structure and the varying size of the clusters, no interval can be detected for the good outlier when $\sigma_\theta \geq 0.4$ In such a case the only viable explanations are $(\emptyset, a_i)$ and $(\emptyset, a_j)$. In both cases the outlierness of the good outlier is determined by the difference in density of the two clusters projected on the dimension under focus. Within this dimension in fact, the good outlier is still located in the region with the lowest density. this explains a relatively high outlierness value even in the case $\sigma_\theta \geq 0.4$.

A similar situation happens for the bad outlier. However, the outlierness exhibited by this value is extremely low, even for low values of σ_θ. The peak in the region $\sigma_\theta \leq 0.2$ is due to statistical fluctuations: the only explanation is characterized by $(\{a_i \in [0.5, 1], a_j \in [0, 0.5]\}, a_k)$ and a_k is one of the dimension where the values are uniformly distributed. Besides that, it is clear that the outlierness degree for the "good outlier" and the "bad outlier" differs significantly irrespective of the σ_θ parameter. This shows that the technique is robust and, in particular, we can expect high values of outlierness for good outliers and low values for bad outliers.

To conclude, the experiments witness the stability of the algorithm and in particular it suggests a procedure for finding the optimal σ_θ by progressively decreasing its value.

5.3 Comparison to other approaches

We already discussed the main differences with regard other approaches found in the literature in Section 2. As a matter of fact, the proposed approach is a direct extention of the approach proposed in [5]. The technical development however is completely different. Specifically, as for the measure here described, its computation requires to perform density estimation of the data, which is not needed in the case of [5]. As for the explanations, the part concerning the computation of

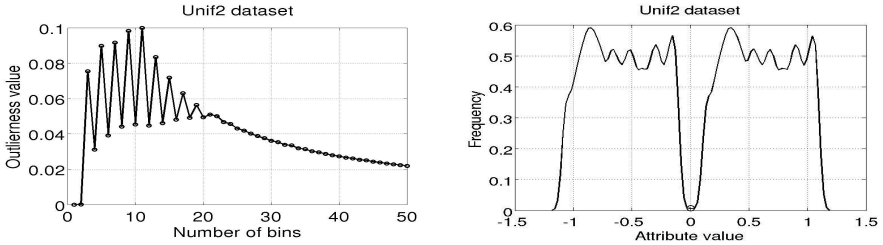

Fig. 8 *Unif2* dataset: outlierness of A_{U2} in o_{U2} computed using the method in [5] (on the left), and density estimate of the same attribute carried out by our method (on the right).

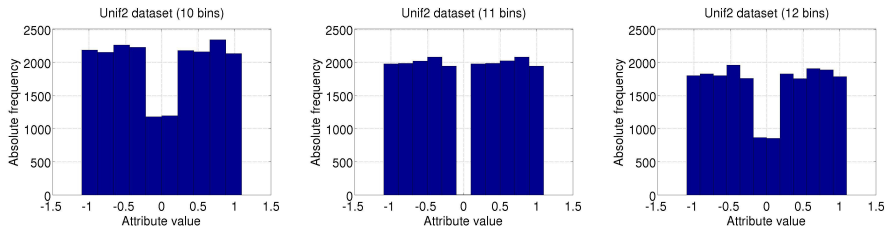

Fig. 9 Different equi-width histograms associated with the attribute A_{U2} of the *Unif2* data set.

intervals is not required in [5], where attribute-values are directly used instead. Moreover, pruning rules already applied in [5] cannot be applied here, due the different nature of the two approaches.

That said, it is worth considering whether a simple adaptation of the approach [5] can be compared to the one we illustrated above in terms of quality. And, in fact, we show next that the strategy based on first discretizing numerical data and then applying on the transformed dataset an algorithm specifically designed for mining outlying properties in categorical data is unfaithful.

To this aim, we consider a synthetic data set (*Unif2* in the following) consisting of 20,000 objects. This dataset contains an outlier o_{U2} which is distinguished from the rest of the population from the value it assumes on a particular attribute A_{U2}. Specifically, almost all values of this attribute belong to two equally-sized uniformly distributed clusters, the first one in the range $[-1.1, -0.1]$ and the second one in the range $[0.1, 1.1]$. The only exception is represented by the object o_{U2}, for which $o_{U2}[A_{U2}] = 0$ holds.

In the following, we focus on the analysis of the behavior of the two methods on the attribute A_{U2}, in order to demonstrate that, while A_{U2} is naturally perceived as an outlying property by the technique presented here, it is very unlikely this is discovered if the technique developed in [5] is employed. In order to apply the method [5] to the *Unif2* dataset, its attributes have been discretized by grouping attribute values in equi-width bins (we note that, in our scenario, using equi-sized bins is meaningless, since such a kind of discretization corresponds to make the attribute values uniformly distributed).

Figure 8 reports, on the left, the value of the outlierness (as defined in [5]) on the value $o_{U2}[A_{U2}]$ of o_{U2} according to different bins sizes employed in discretizing the

data. Specifically, the number of bins has been varied from 2 to 50. The experiment highlights that when the method in [5] is applied, the outcome of the analysis strongly depends on the kind of adopted discretization. In particular when the number of bins is in the range $[4, 20]$ the outlierness measure is quite unstable, in that it basically fluctuates between its maximum value and its middle value. This means that by increasing or, alternatively, decreasing the number of bins of as low as one unit, the result of the analysis may dramatically change. This is a very undesirable property, since determining the right number of bins for the analysis at hand might be quite a challenging task.

Figure 9, showing different frequency histograms associated with the attribute A_{U2}, should further clarify the matter. The histogram associated with the best outlierness value, namely outlierness 0.1, is the one using 11 bins (at the center of the figure). In this case, the central bin (centered in zero) scores a low value of absolute frequency. On the other hand, for both 10 bins (reported on the left in the same figure) and 12 bins (reported on the right), the fact that the outlierness of A_{U2} in o_{U2} is sensibly smaller can be explained by looking at the displayed histograms. In both cases, the value of o_{U2} is grouped with some more frequent values and, hence, the corresponding outlierness value gets sensibly smaller.

Providing a large number of bins does not solve the problem: indeed, as already pointed out in Section 2, the scoring function assigns a score close to 1 to very unbalanced distributions, while its value rapidly decreases when frequencies spread. And, indeed, with a large bin size the number of different categorical values (each associated with a different bin) becomes large, and these values score about the same absolute frequency. The consequence is that the outlierness values get small as well.

It can be concluded that, in order to enable the method [5] to discover meaningful knowledge, the bins that maximize the score should be detected in the first place. However, the interaction with explanations (which select subsets of the overall population) makes it rather difficult to provide optimal a-priori intervals, since the distribution of the property attribute are likely to change when switching from one explanation to another.

This is clearly not the case with the technique proposed in this paper. Since the outlierness measure defined here directly exploits the density estimate of the object value, it is completely adaptive with respect to numerical data and does not suffer of the aforementioned drawbacks. The outlierness of the object o_{U2} on the attribute A_{U2} computed by our method is 0.775. Figure 8 on the right shows the density estimate of attribute A_{U2}, together with the value associated to o_{U2} (notice the circle on the curve), which is exploited in order to compute the outlierness associated with o_{U2}.

5.4 Case Studies

5.4.1 Interplanetary magnetic field data

This experiment has been conducted in order to validate our technique on a well-established ground truth. In this experiment, researchers of the Physics Department of University of Calabria provided us with a dataset of interplanetary mag-

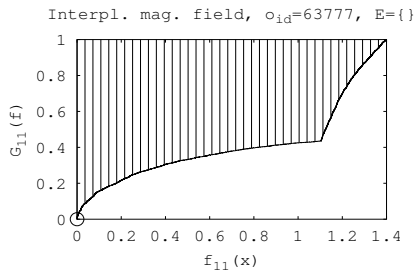

Fig. 10 Outlierness of observation 63,777 of Interplanetary magnetic field dataset.

netic field measurements [21]. The measurements concern MESSENGER magnetic field data in the solar wind at a heliocentric distance of about 0.3 AU.

The dataset consists of 236,183 observations each composed of 11 attributes. The attributes are: *time*, the instant of time of the observation; *bx*, *by*, and *bz*, the three components of the magnetic field; *bmag*, the magnetic field amplitude; *angle2* and *angle4*, the magnetic field angle variation after 2 and 4 seconds respectively; *pvi2* and *pvi4*, the partial variance of increments of the magnetic field after 2 and 4 seconds respectively; *lim2* and *lim4*, the local intermittency measure after 2 and 4 seconds respectively.

They also pointed out 2,153 observations (about the 0.91% of the data) that they manually marked as anomalous on the basis of the domain knowledge. Specifically, these observations correspond to regions of high shear stress and high magnetic field compressibility. Moreover, the analysis they conducted on these anomalous observations has revealed that they correspond to bursts in the *PVI* and *LIM* series.

The OPD technique has been applied to each of these observations in order to determine the exceptional properties they possess. Interestingly, almost all of these observations exhibit the same exceptional properties, while no explanation is able to improve their outlierness, thus confirming that these properties are exceptional in a global sense. Specifically, they differ from the rest of the population on the basis of the values assumed on the four last attributes, thus confirming that our technique is able to detect meaningful outlying properties. Figure 10 reports the outlierness computation for the observation 63,777, which evaluates 0.49.

5.4.2 Medical prescriptions

We further tested the technique on a real-life dataset about doctors and their associated medical prescriptions. In the scenario under consideration, each doctor is associated with a group of patients, and can prescribe drugs to people belonging to that group. There are several situations in which the detection of anomalous prescriptions can be of interest in this scenario: from fraud detection (doctors prescribing more than expected, e.g., with regards to a specific pharmaceutical company) to the diagnosis of unknown epidemiological issues. The specific goal is to find doctors whose behaviour is different than expected. Outlier explanation plays a crucial role here, since we are interested in knowing both the reference

population of doctors with similar prescribing behavior, and the reason why a doctor behaviour is considered anomalous in that population. For example, a doctor behavior can be considered anomalous because the number of prescriptions for a given drug is significantly higher than average or he/she is prone to prescribe drugs produced by a particular company.

The data we analyze contains information about three different entities:

- *doctors*: biographic information, along with information about their patients;
- *drugs*: the active element and the pharmaceutical company that produces the drug;
- *prescriptions*: this is the facts table containing information about prescriptions made by doctors to their patients.

The resulting table contains 2020 tuples, where each tuple represents the number of prescriptions that a specific doctor made on 106 drugs. To better model patients' influence on prescriptions, prescriptions were weighted according to their age and sex. In practice, tuples are normalized in order to make fair comparisons among doctors associated with different classes of patients.

By analyzing the data with feature bagging algorithm [29] we found 5 top outliers exhibiting a significant outlierness score. Two of these outliers are particularly interesting to analyze with the explanation techniques, namely tuples 35 and 651.

In particular, the outlierness of tuple 35 is characterized by attributes a_{26}, a_{102} and a_{103}. Within the population detected by the intervals for the attributes a_{102} and a_{103}, the tuple however exhibits a significantly low value for a_{26}. This is clearly shown in Figure 11.

A different behavior is instead exhibited by tuple 651, characterized by attributes a_1, a_2, a_5, a_6. By selecting the population by means of attributes a_1, a_2, a_5 and studying the distribution for attribute a_6 in this population, we can notice that the value exhibited by tuple 651 is at the upper extreme. Again, this represents a deviation from the normal behavior in that population, as shown in the leftmost graph of Figure 11.

6 Conclusions and Future Work

In this paper we devised a technique by which the outlying properties detection problem can be solved in the presence of numerical attributes, which represents a step forward with respect to available techniques. The core of our approach was the definition of a sensible outlierness measure, representing a refined generalization of that proposed in [5], which is able to quantify the exceptionality of a given property featured by the given input anomalous object with respect to a reference data population. Also, we developed algorithms to detect properties characterizing the anomalous object provided in input. The experiments we conducted confirm that the presented approach is indeed rather effective.

There are several application scenarios where the proposed technique can be profitably applied. In all of these scenarios, data basically express measurements on empirical situations, and the underlying data is made of several numerical attributes describing such measurements. In the *doctors* scenario, for example, it can be used to find explanations for anomalous or frauding behavior. Further scenarios include rank learning problems like in [12]: there, we investigate the problem of

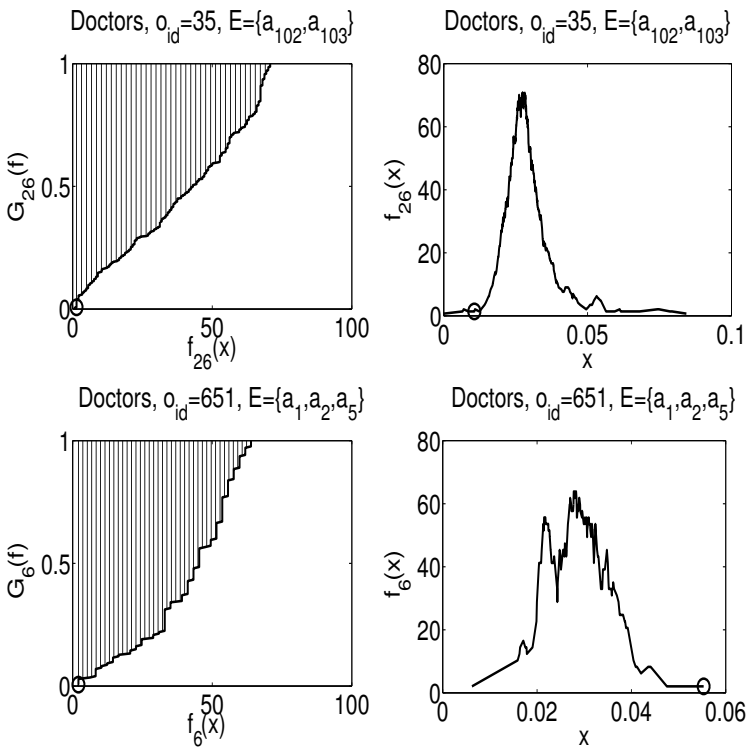

Fig. 11 Outlier explanations in the `Doctors` Dataset, for tuples a_{34} and a_{651}.

detecting rules for characterizing individuals who are selected as exceptional according to a specific scoring function (like, e.g., the amount of fraud they commit in a fraud detection scenario). It is clear that if exceptional objects are outliers, then the outlier explanation technique described in this paper is a basic building block for rule learning in that domain.

With reference to the problem statement discussed in the introductory section, we remark that here we only focus on a single property attribute. We leave to a future work the investigation of how to extend the technique to cope with multi-attribute properties.

References

1. C. Aggarwal. *Outlier Analysis*. Springer, New York, 2013.
2. C. C. Aggarwal and P.S. Yu. Outlier detection for high dimensional data. In *Proceeding of the ACM SIGMOD Conference on Managment of Data (SIGMOD'01)*, pages 37–46, 2001.
3. Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules in large databases. In *IProceedings of the nternational Conference on Very Large Data Bases (VLDB'94)*, pages 487–499, 1994.
4. F. Angiulli and F. Fassetti. Dolphin: an efficient algorithm for mining distance-based outliers in very large datasets. *ACM Transactions on Knowledge Discovery from Data*, 3(1):Article 4, 2009.

5. F. Angiulli, F. Fassetti, and L. Palopoli. Detecting outlying properties of exceptional objects. *ACM Transactions on Database Systems*, 34(1):1–62, 2009.
6. A. Arning, C. Aggarwal, and P. Raghavan. A linear method for deviation detection in large databases. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'96)*, pages 164–169, 1996.
7. V. Barnett and T. Lewis. *Outliers in Statistical Data*. John Wiley & Sons, 1994.
8. S. Bay and M. Pazzani. Detecting change in categorical data: Mining constrast sets. In *Proceedings of the ACM Conf. On Knowledge Discovery in Data (KDD'99)*, pages 302–306, 1999.
9. M. M. Breunig, H. Kriegel, R.T. Ng, and J. Sander. Lof: Identifying density-based local outliers. In *Proceedings of the ACM SIGMOD Conference on Managment of Data (SIGMOD'00)*, pages 93–104, 2000.
10. C. Caroni. Outlier detection by robust principal component analysis. *Communications in Statistics Simulation and Computation*, 29:129–151, 2000.
11. N. V. Chawla, N. Japkowicz, and A. Kotcz. Editorial: special issue on learning from imbalanced data sets. *SIGKDD Explorations*, 6(1):1–6, 2004.
12. Gianni Costa, Fabio Fassetti, Massimo Guarascio, Giuseppe Manco, and Riccardo Ortale. Mining models of exceptional objects through rule learning. In *Proceedings of the ACM Symposium on Applied Computing (SAC'10)*, pages 1078–1082, 2010.
13. X. H. Dang, I. Assent, R.T. Ng, A. Zimek, and E. Schubert. Discriminative features for identifying and interpreting outliers. In *Proceedings of the IEEE International Conference on Data Engineering, (ICDE'14)*, pages 88–99, 2014.
14. X. H. Dang, B. Micenkov, I. Assent, and R. T. Ng. Local outlier detection with interpretation. In *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML-PKDD'13)*, volume 8190 of *Lecture Notes in Computer Science*, pages 304–320, 2013.
15. A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, B*, 39, 1977.
16. L Duan, G. Tang, J. Pei, J.Bailey, A. Campbell, and C. Tang. Mining outlying aspects on numeric data. *Data Mining and Knowledge Discovery*, 29(5):1116–1151, 2015.
17. E. Eskin. Anomaly detection over noisy data using learned probability distributions. In *Proceedings of the International Conference on Machine Learning (ICML'00)*, pages 255–262, 2000.
18. M. A. T. Figueiredo and A. K. Jain. Unsupervised learning of finite mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24:381–396, 2002.
19. J. Fox. Describing univariate distributions. In J. Fox and J. S. Long, editors, *Modern Methods of Data Analysis*, pages 58–125. CA: Sage Publications, 1990.
20. A. Ghoting, S. Parthasarathy, and M.E. Otey. Fast mining of distance-based outliers in high-dimensional datasets. volume 16, pages 349–364, 2015.
21. A. Greco and S. Perri. Identification of high shears and compressive discontinuities in the inner heliosphere. *The Astrophysical Journal*, 784(2):163, 2014.
22. H.P.Kriegel, P.Kroger, E.Schubert, and A.Zimek. Outlier detection in axis-parallel subspaces of high dimensional data. In *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'09)*, pages 831–838, 2009.
23. W. Jin, A. K. H. Tung, and J. Han. Mining top-n local outliers in large databases. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'01)*, pages 293–298, 2001.
24. M. C. Jones and D. A. Henderson. Maximum likelihood kernel density estimation: On the potential of convolution sieves. *Computational Statistics & Data Analysis*, 53:3726–3733, 2009.
25. E. Knorr and R. Ng. Algorithms for mining distance-based outliers in large datasets. In *Proceedings of the International Conference on Very Large Databases (VLDB'98)*, pages 392–403, 1998.
26. E. Knorr and R. Ng. Finding intensional knowledge of distance-based outliers. In *Proceedings of the International Conference on Very Large Databases (VLDB'99)*, pages 211–222, 1999.
27. H.-P. Kriegel, P. Kröger, E. Schubert, and A. Zimek. Loop: local outlier probabilities. In *Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM'09)*, pages 1649–1652, 2009.
28. H.-P. Kriegel, M. Schubert, and A. Zimek. Angle-based outlier detection in high-dimensional data. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'08)*, pages 444–452, 2008.

29. Aleksandar Lazarevic and Vipin Kumar. Feature bagging for outlier detection. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery in Data (KDD'05)*, pages 157–166, 2005.
30. M. Lichman. UCI machine learning repository [`http://archive.ics.uci.edu/ml`]. Irvine, CA: University of California, School of Information and Computer Science, 2013.
31. F.T. Liu, K.M. Ting, and Z.-H. Zhou. Isolation forest. In *Proceedings of the IEEE International Conference on Data Mining (ICDM'08)*, pages 413–422, 2008.
32. B. Micenková, R. T. Ng, X. H. Dang, and I Assent. Explaining outliers by subspace separability. In *Proceedings of the IEEE International Conference on Data Mining (ICDM'13)*, pages 518–527, 2013.
33. H. Nguyen, V. Gopalkrishnan, and I. Assent. An unbiased distance-based outlier detection approach for high dimensional data. In *Proceedings of the International Conference on Database Systems for Advanced Applications (DASFAA)*, pages 138–152, 2011.
34. S. Papadimitriou, H. Kitagawa, P.B. Gibbons, and C. Faloutsos. Loci: Fast outlier detection using the local correlation integral. In *Proceedings of the IEEE International Conference on Data Enginnering (ICDE'03)*, pages 315–326, 2003.
35. P. Rousseeuw and A. Leroy. *Robust Regression and Outlier Detection*. Wiley, 2003.
36. I. H. Salgado-Ugarte and M. A. Pérez-Hernández. Exploring the use of variable bandwidth kernel density estimators. *Stata Journal*, 3(2):133–147, 2003.
37. B. Schölkopf, C. Burges, and V. Vapnik. Extracting support data for a given task. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 252–257, 1995.
38. B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, 1986.
39. N. X. Vinh, J. Chan, J. Bailey, C. Leckie, K. Ramamohanarao, and J. Pei. Scalable outlying-inlying aspects discovery via feature ranking. In *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data (PAKDD'15)*, pages 422–434, 2015.
40. T. De Vries, S. Chawla, and M. Houle. Finding local anomalies in very high dimensional space. In *Proceedings of the IEEE International Confence on Data Mining (ICDM'10)*, pages 128–137, 2010.
41. L. Xiong, X. Chen, and J. Schneider. Direct robust matrix factorization for anomaly detection. In *Proceedings of the IEEE International Confence on Data Mining (ICDM'11)*, pages 844 – 853, 2011.
42. X. Yang, L. J. Latecki, and D. Pokrajac. Outlier detection with globally optimal exemplar-based GMM. In *Proceedings of the SIAM Conference on Data Mining (SDM'09)*, pages 145–154, 2009.