Applied Intelligence (2022) 52:17534–17556
https://doi.org/10.1007/s10489-022-03271-3

A density estimation approach for detecting and explaining exceptional values in categorical data

Fabrizio Angiulli[1] · Fabio Fassetti[1] · Luigi Palopoli[1] · Cristina Serrao[1]

Accepted: 18 January 2022 / Published online: 2 April 2022

Abstract

In this work we deal with the problem of detecting and explaining anomalous values in categorical datasets. We take the perspective of perceiving an attribute value as anomalous if its frequency is exceptional within the overall distribution of frequencies. As a first main contribution, we provide the notion of *frequency occurrence*. This measure can be thought of as a form of Kernel Density Estimation applied to the domain of frequency values. As a second contribution, we define an *outlierness* measure for categorical values that leverages the cumulated frequency distribution of the frequency occurrence distribution. This measure is able to identify two kinds of anomalies, called *lower outliers* and *upper outliers*, corresponding to exceptionally low or high frequent values. Moreover, we provide interpretable *explanations* for anomalous data values. We point out that providing interpretable explanations for the knowledge mined is a desirable feature of any knowledge discovery technique, though most of the traditional outlier detection methods do not provide explanations. Considering that when dealing with explanations the user could be overwhelmed by a huge amount of redundant information, as a third main contribution, we define a mechanism that allows us to single out *outstanding explanations*. The proposed technique is *knowledge-centric*, since we focus on explanation-property pairs and anomalous objects are a by-product of the mined knowledge. This clearly differentiates the proposed approach from traditional outlier detection approaches which instead are *object-centric*. The experiments highlight that the method is scalable and also able to identify anomalies of a different nature from those detected by traditional techniques.

Keywords Outlier Detection; Outlier Explanation; Categorical Data

1 Introduction

An outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which it occurs. Their detection can identify system faults and frauds before they escalate with potentially catastrophic consequences; it turns out that, in some applications, the rare events can be more interesting than the more regularly occurring ones [2, 26].

As outliers are interesting because they are suspected of not being generated by the same mechanisms as the rest of the data, it is important to justify why detected outliers are generated by some other mechanisms [9, 22, 23]. However, the border between data normality and abnormality is often not clear cut; consequently, while some outlier detection methods assign to each object in the input data set a label as either "normal" or "outlier", in this paper we describe a method able to single out anomalous data bunches by working on attributes and associated values.

We deal with categorical data where it is generally more difficult to devise criteria able to discriminate normal and abnormal data [46, 48]. Moreover, the notion of outlierness in the field of numerical data has long been analysed [20] and many approaches have already been designed to define the exceptional nature of a property, but most of them cannot be easily generalized to deal with categorical or mixed categorical/numerical dataset.

✉ Cristina Serrao
c.serrao@dimes.unical.it

Fabrizio Angiulli
f.angiulli@dimes.unical.it

Fabio Fassetti
f.fassetti@dimes.unical.it

Luigi Palopoli
l.palopoli@dimes.unical.it

[1] DIMES, University of Calabria, 87036 Rende (CS), Italy

Some algorithms exist that build an anomaly detection model specially devised for categorical variables and transform any numerical variable into a categorical space through a previous discretization phase [32, 49]; however, the main drawback of such a strategy is that the result of the analysis strongly depends on the results of the discretization process.

As an alternative, some mixed criteria techniques have been developed which manage numerical and categorical data separately and then merge the two by providing a method which encompasses the analysis of an element in both spaces [28, 30, 39].

In this paper, we specifically focus on problems which are typical of categorical data. We do that by taking the perspective of perceiving an attribute value as anomalous if its frequency occurrence is exceptionally typical or untypical within the distribution of frequencies occurrences of any other attribute value.

However, within the categorical scenario the process of comparing frequencies poses several challenges. Indeed, if we take the point of view that the data at hand is the result of a sampling procedure in which data values are associated with some pre-defined occurrence probabilities, then the fact that a certain categorical value is observed exactly f times is a matter of chance rather than being a hard property of that value.

This has led us to the definition of the concept of *soft frequency occurrence* which, intuitively, consists in the *estimate of the density associated with frequency occurrences*. We obtain this measure by specializing the classical Kernel Density Estimation technique to the domain of frequency values.

As a second contribution, we leverage the cumulated frequency distribution of the above density estimate to decide if the frequency of a certain value is rare when compared to the frequencies associated with the other values. In particular, we are able to identify two kinds of anomalies, namely *lower outliers* and *upper outliers*. A *lower outlier* is a value whose frequency is small while, typically, the dataset objects assume a few similar values, namely the frequencies of the other values are large. An *upper outlier* is a value whose frequency is large while, typically, the dataset objects assume almost distinct values, namely the frequencies of the other values are small.

Note that both these definitions look for unexpected behaviors by establishing a comparison between the frequency of the outlier value and that of the other values in the attribute domain. While the notion of lower outlier shares similarities with the classical concept of anomaly, the notion of upper outlier is conceptually different and should not be confused with the concept of mode. Indeed, for a value, a high frequency is not enough to be an upper outlier, as the rest of the values must appear with low frequencies.

Consider the example reported in Table 1: as for the attribute $A1$, value c is a lower outlier, since it occurs only once and the other values occur many times; as for the attribute $A2$, value e is an upper outliers, since it occurs many times and the other values occur once or at most twice; as for the attribute $A3$, although the value e has the same frequency both in $A2$ and $A3$ and it is the most frequent in $A3$, it is not an (upper) outlier since the other values in $A3$ have comparable frequencies.

Thus, we are able to single out, by one unified outlierness measure, both exceptionally infrequent and exceptionally frequent values; this peculiarity clearly differentiates our proposal from almost all the existing measures of outlierness.

Although values can show exceptional behavior with respect to the whole population, it must be pointed out that very often a value emerges as exceptional only when we restrict our attention to a subset of the whole population [7]. In particular, our method, differently from many others, is capable of returning to the user not only data bunches which are anomalous with respect to the entire data population, but also those records which are normal in general but anomalous only when contrasted to a sub-population. Therefore, our technique has been designed to output the so-called *explanation-property pairs* (E, p), where E, called *explanation*, denotes a condition used to determine the target subpopulation and p, called *property*, represents an attribute p_a and a value p_v such that p_v is exceptionally frequent or infrequent within the subpopulation selected by the explanation E.

The output of the algorithm corresponds to the so-called *explanation-property pairs*, which allows us to provide an interpretable explanation for the abnormal values discovered.

However, it must be noticed that, when dealing with explanations, there exists a risk that the user is overwhelmed

Table 1 An example highlighting the notion of lower (A1) and upper (A2) outliers

	A1	A2	A3
$\mathbf{obj}_1$	a	**e**	**e**
$\mathbf{obj}_2$	a	**e**	**e**
$\mathbf{obj}_3$	a	**e**	**e**
$\mathbf{obj}_4$	b	**e**	**e**
$\mathbf{obj}_5$	b	**e**	**e**
$\mathbf{obj}_6$	b	g	f
$\mathbf{obj}_7$	**c**	h	f
$\mathbf{obj}_8$	d	h	f
$\mathbf{obj}_9$	d	i	g
$\mathbf{obj}_{10}$	d	j	g
$\mathbf{obj}_{11}$	d	k	g

by a huge amount of redundant explaining patterns. Thus, as a further contribution, we define a subtle mechanism that allows us to single out the explanations encoding the outstanding exceptionalities in the data, carrying out no redundant information. Loosely speaking, for a condition to encode a significant explanation we require the frequency distribution of the associated sub-population to be unexpected given the knowledge of the frequency distribution of any other of its super-population, where unexpectedness is measured by means of the chi-squared goodness-of-fit test. Moreover, for an attribute and a value to encode a significant property, we require that the outlierness measured within the sub-population associated with a given significant explanation improves the one measured within any of its super-populations associated with significant explanations. Outlierness improvement must be greater than a factor which is inversely related to the unexpectedness of the sub-population frequency distribution. Maximal significant explanation-property pairs are said to be *outstanding*. The output of the algorithm precisely corresponds of the so-called *outstanding explanation-property pairs*.

Finally, our technique is *knowledge-centric* as the search space we visit is formed by explanation-property pairs and the outliers we provide can be seen as a product of the knowledge mined. This is clearly different from traditional outlier detection approaches which are object-centric.

The rest of the work is organised as follows. Section 2 discusses work related with the present one. Section 3 introduces the frequency occurrence function. Section 4 describes the outlierness function for ranking categorical values. Section 5 introduces the concept of outstanding explanation-property pair and describes the goal of our mining method. Section 6 describes experimental results.

2 Related work

Categorical data has received relatively little attention as compared to quantitative data because detecting anomalies in categorical data is a challenging problem [48]. Generally, traditional approaches do not handle categorical data in a satisfactory manner, due to the fact that, in most cases, there is no concept of sorting for the set of values a categorical variable can assume; so the development of specific techniques is needed. We start by noting that there is little literature about detecting anomalous properties, and/or related outlier objects, equipped with explanations which face the task of the identification of both *features* and *subpopulations* which characterize anomalies. Moreover, to the best of our knowledge, no technique is able to natively detect *upper* outliers.

There exist several approaches to detect outliers in the certain setting, namely statistical - based [14, 25], distance - based [4, 12, 13, 29, 34], density - based [18, 45], isolation - based [38], subspace - based [1, 6, 7], knowledge - based [5], neural network - based [31, 42], and many others [3, 19].

Some anomaly detection techniques depend on the identification of a representative pattern suggested by the majority observations so that objects that result to be far from it, according to a suitable distance measure, can be perceived as anomalies. However, designing such a measure in presence of categorical data is challenging [17].

Different strategies have been proposed to face with the above problem. In [21] some methods are presented to map categorical data on numerical data together with a framework for their analysis. However, the effectiveness of these techniques is strongly related to the choice of the mapping function. A different prospective is that of exploiting for categorical data some traditional approaches designed for the quantitative domain by choosing an appropriate distance measure. This is done in [8, 13], where anomalies are defined as the N observations whose average distance to the k nearest neighbors are the greatest; instead, [35] considers as anomalous those observations with fewer than p observations within a certain distance d.

Many other methods exploit the Hamming distance to identify anomalies among categorical data [15, 16, 37] together with a pruning strategy to cope with the quadratic complexity of evaluating distances.

Recently a new idea of distance suitable for the categorical domain has been introduced to detect and characterize outliers in a semi-supervised fashion [33]. The key intuition is that the distance between two values of a categorical attribute can be determined by the way in which they co-occur with the values of other attributes in the data set: if two values are similarly distributed with respect to a certain set of attributes, their distance must be low. A model defining the distances between categorical values is defined on the training set and is used to evaluate the outlier score associated with each test instance t as the sum the distances between t and a subset of objects known to be *normal*.

The family of density-based aqpproaches includes those methods that identify observations having outlying behavior in local areas, thus result to be inconsistent within their neighborhood and not necessary with the pattern suggested by the majority of all other observations.

Local anomaly detection methods for categorical data include the k-Local Anomalies Factor k-LOF [51] and, ROAD [47] and WATCH [36] methods.

The k-LOF is a local anomaly detection method for both categorical and quantitative data [51]. It extends Local Anomalies Factor (LOF) [18] to categorical domain. The k-LOF identifies an observation as a local outlier if its relationships with its neighbors are weaker than the relationships between its neighbors and its neighbors' neighbors. It does that by building a similarity graph and by using the

concept of k-walk, i.e the paths of length k on the similarity graph joining two observations, to provide an outlierness score.

The ROAD algorithm [47] exploits both distances and densities. The Hamming distance is used to group objects into clusters and highlight observations located in sparse regions; then a density measure is calculated for each object on the basis of the frequencies of its values, in order to identify objects whose values are almost infrequent in the dataset.

Both measures present some limitations: as for Hamming distance, outliers with few exceptional attributes are not captured, as for density, they are not compared with expected values and attributes associated with distinct values, as primary keys, can affect results.

The WATCH method [36] has been recently designed to find out outliers in high dimensional categorical datasets using feature grouping. First, it groups correlated features, then it looks for outliers in each feature group by calculating a weighting factor for each categorical variable that takes into account the correlation between this variable and the others in the same group.

A completely different prospective is taken by methods that exploit information - theoretic measures. The idea behind these approaches relies on the direct relationship between the existence of anomalies and the amount of noise in the dataset. This led some authors [32] to formulate the outlier detection task in terms of an optimization problem, i. e. finding a subset of k objects such that the expected entropy of the resultant dataset after the removal of this subset is minimized. This strategy has to be intended from a global point of view as the outlierness measure involves simultaneously all the attributes and is neither able to detect outliers in sub-populations nor to identify outliers characterised by one (or few) outlying attributes.

To overcome with the last issue, in [24] an outlier factor is designed on the basis of the ratio between the probability of co-occurrence of two sets of attributes and the product between the probabilities of occurrence of the two sets taken separately. Here, the authors are interested in properties consisting in at least two attributes and do not address sub-populations.

The importance of learning value interactions has shown to be effective when handling categorical data and some recent contributions are found in the literature that exploit such a strategy to detect anomalies.

CBRW [40] estimates the outlierness of each feature value which can either detect outliers directly or determine feature selection for subsequent outlier detection. The value is computed by *comparing the frequency of each value with the most frequent value* (*the mode*). However, this is just a measure of deviation, explanations are not provided and, by definition, upper outliers cannot be detected.

As a further drawback, noisy values may significantly influence the performance of CBRW, thus the same authors propose HOUR [41], a new outlier detection framework for data with noisy features. A noise-resilient outlier scoring function is defined to rank objects based on their outlierness in a given feature subset and an outlier ranking evaluation function is proposed to evaluate the quality of the ranking w.r.t the feature subset. Feature selection and ranking evaluation are iteratively performed until the best feature subset is obtained.

Nevertheless, this approach shows to be not particularly suitable for high dimensional data, thus two further main improvement have been proposed, namely POP [44] and OUVAS [50].

The idea of investigating feature subsets has been taken into account also in [43] but in a completely different way. Here, the main intuition is that, provided with a random subsample of the main dataset, those instances with rare combinations of values on any attribute subset have also a higher probability of having zero appearances in subsamples of any size.

Subspaces managment seems to meet our concept of explanation, but the semantic is completely different. The problem of outlier explanation [10] we deal with in this paper consists in finding features that can justify the outlierness of an object, in a sense that its anomalous state emerges only as a consequence of the selection we have made.

Some previous works [9, 10] follow this path and propose a technique for categorical and numerical domains respectively that, *given in input one single object known to be outlier*, provides features justifying its anomaly and subpopulations where its exceptionality is evident. A generalization is proposed in [11] where a set, required to be small, of outliers is provided in input.

3 Frequency occurrence

In this section we give same preliminary definitions and introduce the notation employed throughout the paper.

A *dataset* $\mathcal{D}$ on a set of *categorical* attributes $\mathcal{A}$ is a set of objects o assuming values on the attributes in $\mathcal{A}$. By $o[a]$ we denote the value of o on the attribute $a \in \mathcal{A}$. $\mathcal{D}[a]$ denotes the multiset $\{o[a] \mid o \in \mathcal{D}\}$.

A *condition* $\mathcal{C}$ is a set of pairs (a, v) where each a is an attribute and each $v \in \mathcal{D}[a]$. A singleton condition is said to be *atomic*. By $\mathcal{D}_\mathcal{C}$ we denote the new dataset $\{o \in \mathcal{D} \mid o[a] = v, \forall (a, v) \in \mathcal{C})\}$.

A *condition* $\mathcal{C}$ is said to be *valid* if $\mathcal{D}_\mathcal{C} \neq \emptyset$. It follows from this definition that in a valid condition $\mathcal{C}$, for all $(a, v), (a, u) \in \mathcal{C}$, it holds that $u = v$. Thus, $\mathcal{C}$ is valid, the number $|\mathcal{C}|$ of atomic conditions in $\mathcal{C}$ is equal to the number

of attributes involved in $\mathcal{C}$. In the following, if not otherwise stated, we will take into account only valid conditions.

Definition 1 (Frequency distribution) A *frequency distribution* $\mathcal{H}$ is a multiset of the form $\mathcal{H} = \{f_1^{(1)}, \ldots, f_1^{(w_1)}, \ldots, f_n^{(1)}, \ldots, f_n^{(w_n)}\}$ where each $f_i^{(j)} \in \mathbb{N}$ is a distinct frequency, $f_i^{(j)} = f_i^{(k)} = f_i$ for each $1 \leq j, k \leq w_i$, and w_i denotes the number of occurrences of the frequency f_i. By $N(\mathcal{H})$ (or simply N whenever $\mathcal{H}$ is clear from the context) we denote $w_1 \cdot f_1 + \ldots + w_n \cdot f_n$.

For the sake of simplicity, we will refer to a frequency distribution as a set $\mathcal{H} = \{f_1, f_2, \ldots, f_n\}$ and to the number of occurrences w_i of f_i as $w(f_i)$. To ease the writing of expressions, we also assume that the dummy frequency $f_0 = 0$ with $w_0 = 0$ is always implicitly part of any frequency distribution.

Given a multiset V, the frequency f_v^V of the value $v \in V$ is the number of occurrences of v in V.

The frequency distribution of the dataset $\mathcal{D}$ on the attribute a is the multiset $\mathcal{H}_a^{\mathcal{D}} = \{f_v^{\mathcal{D}[a]} \mid v \in \mathcal{D}[a]\}$. Note that $N(\mathcal{H}_a^{\mathcal{D}}) = |\mathcal{D}|$.

Theorem 1 *Let $\mathcal{H} = \{f_1, \ldots, f_n\}$ be a frequency distribution. Then, $n \leq \sqrt{N(\mathcal{H})}$.*

Proof Since $N(\mathcal{H}) = w_1 \cdot f_1 + w_2 \cdot f_2 + \cdots + w_n \cdot f_n$, n is maximized when (i) $f_1 = 1$, (ii) $\forall i$, $w_i = 1$, and (iii) $\forall i > 1$, $f_{i+1} = f_i + 1$. Thus, the maximum n is such that $1+2+\cdots+n = N(\mathcal{H})$ and, since $1+2+\cdots+n = \frac{n(n+1)}{2}$, it follows that $n \cdot (n+1) = 2 \cdot N(\mathcal{H})$ and, then, that $n = O\left(\sqrt{N(\mathcal{H})}\right)$. □

From the above theorem, it immediately follows that the number of distinct frequencies in $\mathcal{H}^D$ is at most $\sqrt{|\mathcal{D}|}$.

Now we define the notion of *frequency occurrence* as a tool for quantifying how frequent is a certain frequency.

Definition 2 (**Hard frequency occurrence**) Given a frequency distribution $\mathcal{H}$, the *frequency occurrence* $\mathcal{F}_{\mathcal{H}}(f_i)$ of f_i, also denoted by $\mathcal{F}(f_i)$ whenever $\mathcal{H}$ is clear from the context, is the product $w_i \cdot f_i$.

The above definition allows us to associate with each distinct value in $\mathcal{D}[a]$ a score that is related not only to its frequency in the dataset but also to how many other values have its same frequency.

A major drawback of the previous definition is that close frequency values do not interact with each other and, as a consequence, small variations of the frequency distribution may cause sensible variations in the *frequency occurrence* values. E.g., consider the case in which the frequencies $f_i = 49$, $w_i = 1$ and $f_{i+1} = 51$, $w_{i+1} = 1$ are replaced with $f_i' = 50$, $w_i' = 2$. While in the former case $\mathcal{F}(f_i) = 49$ and $\mathcal{F}(f_{i+1}) = 51$, in the latter case we have that $\mathcal{F}(f_i') = 100$ that is about twice the frequency occurrence associated with f_i and f_{i+1}. Intuitively, we do not desire a similar small variation in the frequency distribution to impact so largely on the outcome of the measure. Indeed, if we take the point of view that the data at hand is the result of a sampling procedure in which data values are associated with some pre-defined occurrence probabilities, then the fact that a certain categorical value is observed exactly f times is a matter of chance, rather than being an hard property of that value.

Thus, we refine the previous definition of frequency occurrence in order to cope with the scenario depicted above. Specifically, to overcome the mentioned drawback, we need to force close frequency values to influence each other in order to jointly contribute to the frequency occurrence value. With this aim, we inspired to Kernel Density Estimation (KDE) methods to design an ad-hoc density estimation procedure.

First of all, we point out that we are working in a discrete domain composed of frequency values, a peculiarity that differentiates it from the standard framework of KDE. We start by illustrating the proposed density estimation procedure.

A (*discrete*) *kernel function K_{f_i} with parameter f_i* is a probability mass function having the property that $\sup_{f \geq 0} K_{f_i}(f) = K_{f_i}(f_i)$.

Given an interval $I = [f_l, f_u]$ of frequencies, a frequency f_i, and a kernel function K, the *volume* of K_{f_i} in I, denoted as $V_I(K_{f_i})$, is given by $\sum_{f=f_l}^{f_u} K_{f_i}(f)$. The following expression

$$\mathcal{F}(f) = \sum_{\varphi \in I(f)} \left\{ \sum_{i=1}^{n} w_i \cdot f_i \cdot K_{f_i}(\varphi) \right\}.$$

where $I(f)$ represents an interval of frequencies centred in f, provides the density estimate of the *frequency occurrence* of the frequency f.

Since $K_{f_i}(\cdot)$ is a probability mass function, the frequency f_i provides a contribution to the *frequency occurrence* of f corresponding to the portion of the volume of K_{f_i} which is contained in $I(f)$, that is $V_{I(f)}(K_{f_i})$. Hence, if the interval $I(f)$ contains the entire domain of K_{f_i} then f_i provides its maximal contribution $w_i \cdot f_i$. Frequencies f_i whose domain do not intersect $I(f)$ do not contribute to the *frequency occurrence* of f at all.

The above definition needs to properly calibrate the width $I(f)$ of the interval to be centred in f. To eliminate the dependence of the formulation from an arbitrary interval, we resort to the following alternative formulation in which frequencies φ are not constrained to belong to the interval $I(f)$. However, since the generic kernel $K_{f_i}(\cdot)$

could be arbitrarily far from the frequency of interest f, now its contribution has to be properly weighted

$$\mathcal{F}(f) = \sum_{\varphi \geq 0} \left\{ \sum_{i=1}^{n} \left[w_i \cdot f_i \cdot K_{f_i}(\varphi) \cdot \frac{Pr[X_{f_i} = f]}{Pr[X_{f_i} = f_i]} \right] \right\}.$$

Let X_{f_i} denote the random variable distributed according to K_{f_i} and, hence, having f_i as the value that is most likely to be observed. The ratio $\frac{Pr[X_{f_i}=f]}{Pr[X_{f_i}=f_i]} \leq 1$ represents a weight factor for the kernel $K_{f_i}(\cdot)$ which is maximum, in that evaluates to 1, for $f = f_i$. Hence, the closer the kernel $K_{f_i}(\cdot)$ to the frequency of interest f, the larger its contribution to the *frequency occurrence* of f. Since the above probabilities can be directly obtained from the associated kernel, it can be rewritten as follows

$$\mathcal{F}(f) = \sum_{\varphi \geq 0} \left\{ \sum_{i=1}^{n} \left[w_i \cdot f_i \cdot K_{f_i}(\varphi) \cdot \frac{K_{f_i}(f)}{K_{f_i}(f_i)} \right] \right\}. \tag{1}$$

Equation (1) can be rewritten as

$$\mathcal{F}(f) = \sum_{i=1}^{n} \left[w_i \cdot f_i \cdot \frac{K_{f_i}(f)}{K_{f_i}(f_i)} \cdot \sum_{\varphi \geq 0} K_{f_i}(\varphi) \right].$$

Since $K_{f_i}(\cdot)$ is a probability mass function, the summation over the domain of all its values is equal to 1 and, thus, the above expression can be finally simplified in

$$\mathcal{F}(f) = \sum_{i=1}^{n} \left[w_i \cdot f_i \cdot \frac{K_{f_i}(f)}{K_{f_i}(f_i)} \right]. \tag{2}$$

Since $\mathcal{F}$ represents a notion of density function associated with frequency occurrences, it is preferable that its volume evaluated in the frequencies $\mathcal{H} = \{f_1, \ldots, f_n\}$ evaluates to $N(\mathcal{H})$. This leads to the following final form of the *frequency occurrence* function.

Definition 3 (Soft occurrence function) Given a frequency distribution $\mathcal{H}$, the *frequency occurrence* $\mathcal{F}_{\mathcal{H}}(f_i)$ of f_i, also denoted by $\mathcal{F}(f_i)$ whenever $\mathcal{H}$ is clear from the context, is given by the following expression

$$\mathcal{F}(f) = \frac{N(\mathcal{H})}{N_{\mathcal{F}}(\mathcal{H})} \cdot \sum_{i=1}^{n} \left[w_i \cdot f_i \cdot \widehat{K}_{f_i}(f) \right], \tag{3}$$

where

$$\widehat{K}_{f_i}(f) = \frac{K_{f_i}(f)}{K_{f_i}(f_i)} \quad \text{and} \quad N_{\mathcal{F}}(\mathcal{H}) = \sum_{j=1}^{n} \left\{ \sum_{i=1}^{n} \left[w_i \cdot f_i \cdot \widehat{K}_{f_i}(f_j) \right] \right\}.$$

Figure 1 reports the frequency occurrence values according to the hard and soft definition. Note that frequencies 48 and 50 are closer to each other and the soft frequency occurrence definition for them lead to a value that is more like the one we would get if we observed a value between 48 and 50 twice in the distribution. The values of the soft occurrences are not normalized to improve intelligibility of the example.

As for the kernel selection, interestingly we can take advantage of the peculiarity of the frequency domain to base our estimation on a very natural kernel definition. Indeed, as kernel $K_{f_i}(\cdot)$ we will exploit the binomial distribution $binopdf(f; n, p)$ with parameter n, denoting the number of independent trials, equal to $N(\mathcal{H})$, and parameter p, denoting the success probability, equal to $p = f_i/N(\mathcal{H})$. We argue that this kind of kernel is particularly natural for our setting and, moreover, note that its use relieve us from the problem of selecting a suitable kernel bandwidth, a problem that affects almost all the kernel density estimation procedures. Indeed, the fact that we observe a certain number f_i of occurrences for a given value v can be assimilated to the outcome of a sequence of Bernoulli trials each having success probability p_v, which is modeled by a binomial random variable. Note that, for large sample sizes $N(\mathcal{H})$, the frequency f_i tends to the expected value $N(\mathcal{H}) \cdot p_v$ of the above random variable. Hence, $f_i/N(\mathcal{H})$ closely approximates p_v and the binomial function with parameters $n = N(\mathcal{H})$ and $p = f_i/N(\mathcal{H})$ represents the distribution of that variable. This distribution clearly provides the probability to observe any other frequency $f_i' \neq f_i$ for the same value v.

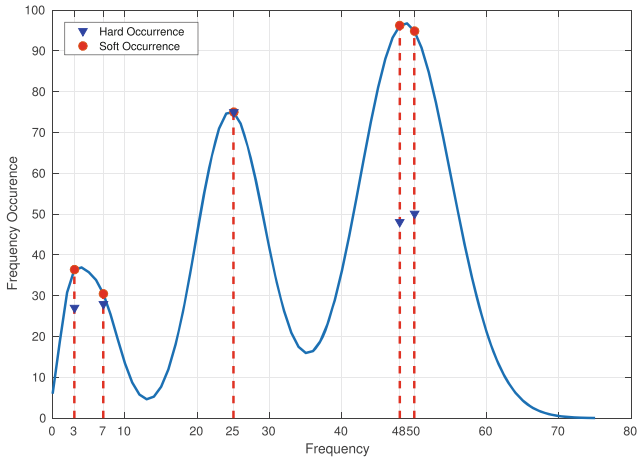
3.1 Computational cost

Computational complexity of the proposed measure can be estimated by taking into account the cost of computing the set of frequency occurrences for each attribute in $\mathcal{A}$ and the cost of evaluating $\widehat{K}_{f_i}(\cdot)$ for each frequency in $\mathcal{H}^D$. Theorems discussed in the following provide some results about time complexity.

Theorem 2 *Let $\mathcal{D}$ be a dataset and let $\mathcal{H}^D = \{f_1, \ldots, f_n\}$. Then, the cost of computing the set of frequency occurrences $\{\mathcal{F}_{\mathcal{H}^D}(f_1), \ldots, \mathcal{F}_{\mathcal{H}^D}(f_n)\}$ is $O(|\mathcal{D}| \cdot C_K)$, where C_K represents the cost of evaluating $\widehat{K}_{f_i}(\cdot)$.*

Proof Consider (3). The cost of evaluating this Equation for a given f involves the computation of a summation of n terms, with n equals to the number of different frequencies. Due Theorem 1, $n = O\left(\sqrt{|\mathcal{D}|}\right)$. Since we have to evaluate (3) for any distinct f in the dataset, the Equation has to be computed n times. Then, since each evaluation costs C_k, the overall cost is $O\left(\sqrt{|\mathcal{D}|} \cdot \sqrt{|\mathcal{D}|} \cdot C_k\right) = O(|\mathcal{D}| \cdot C_k)$. □

Now we take into account the cost of computing the term $\widehat{K}_{f_i}(f)$ when $K_{f_i}(\cdot)$ is the binomial kernel.



f	Hard $\mathcal{F}$	Soft $\mathcal{F}$
3	27	36.37
7	28	30.54
25	75	75.02
48	48	96.19
50	50	94.89

Fig. 1 Comparison between hard and soft occurrence

Theorem 3 *Given a dataset $\mathcal{D}$ and two frequencies f_i and f_j in $\mathcal{H}^{\mathcal{D}}$, the cost of computing $\widehat{K}_{f_i}(f_j)$ is $O(1)$ with a pre-processing $O(|\mathcal{D}|)$.*

Proof First of all, we prove that $\widehat{K}_{f_i}(f_j)$ can be obtained by evaluating the following expression:

$$\exp\left[\left(\sum_{k=1}^{f_i}\ln k\right)+\left(\sum_{k=1}^{N-f_i}\ln k\right)-\left(\sum_{k=1}^{f_j}\ln k\right)-\left(\sum_{k=1}^{N-f}\ln k\right)+(f_i-f)\ln\left(\frac{N}{f_i}-1\right)\right].$$

To get this equality, consider the logarithm of $\widehat{K}_{f_i}(f_j)$. Since $K_{f_i}(f_j)$ is a binomial probability function, by exploiting the properties of logarithms we obtain:

$$\begin{aligned}
&\ln\widehat{K}_{f_i}(f_j)=\ln\frac{K_{f_i}(f_j)}{K_{f_i}(f_i)}=\\
&=\ln\frac{\binom{N}{f_j}\left(\frac{f_i}{N}\right)^{f_j}\left(1-\frac{f_i}{N}\right)^{N-f_j}}{\binom{N}{f_i}\left(\frac{f_i}{N}\right)^{f_i}\left(1-\frac{f_i}{N}\right)^{N-f_i}}=\ln\frac{\binom{N}{f_j}}{\binom{N}{f_i}}+(f_j-f_i)\ln\left(\frac{f_i}{N}\right)\\
&\quad+(f_i-f_j)\ln\left(1-\frac{f_i}{N}\right)=\\
&=\ln\frac{\binom{N}{f_j}}{\binom{N}{f_i}}+(f_i-f_j)\ln\left(\frac{1-f_i/N}{f_i/N}\right)\\
&=\ln\frac{\binom{N}{f_j}}{\binom{N}{f_i}}+(f_i-f_j)\ln\left(\frac{N}{f_i}-1\right).
\end{aligned}$$

Since

$$\begin{aligned}
\ln\frac{\binom{N}{f_j}}{\binom{N}{f_i}}&=\ln\left(\frac{N!}{f_j!(N-f_j)!}\cdot\frac{f_i!(N-f_i)!}{N!}\right)\\
&=\sum_{k=1}^{f_i}\ln k+\sum_{k=1}^{N-f_i}\ln k-\sum_{k=1}^{f_j}\ln k-\sum_{k=1}^{N-f_j}\ln k,
\end{aligned}$$

the statement follows. We note that all the above terms can be pre-computed. Indeed, during the pre-processing step we can build an array of N elements such that the generic i^{th} entry stores $\sum_{k=1}^{i}\ln k$. □

4 Categorical outlierness

In this section we introduce the concept of *outlierness* and discuss about the measure we have designed to discover outlier properties in categorical datasets.

Definition 4 (Cumulated frequency distribution) Given a frequency distribution $\mathcal{H}=\{f_1,\ldots,f_n\}$, the associated *cumulated frequency distribution* H is

$$H(f)=\sum_{f_j\le f}\mathcal{F}_{\mathcal{H}}(f_j).$$

In the following, we refer to the value $H(f_i)$ also as to H_i.

The idea behind the measure we will discuss in the following is that an object in a categorical dataset can be considered an outlier with respect to an attribute if the frequency of the value assumed by this object on such an attribute is rare if compared to the frequencies associated with the other values assumed on the same attribute by the other objects of the dataset.

We are interested in two relevant kinds of anomalies referring to two different scenarios.

Lower Outlier. An object o is anomalous since for a given attribute a the value that o assumes in a is rare (its frequency is low) while, typically, the dataset objects assume a few similar values, namely the frequencies of the other values are large.

Upper Outlier. An object o is anomalous since for a given attribute a the value that o assumes in a is usual (its frequency is high) while, typically, the dataset objects

assume almost distinct values, namely the frequencies of the other values are small.

In order to discover outliers, we exploit the cumulated frequency distribution associated with the dataset. With this aim, we use the area above and below the curve of the cumulated frequency distribution to quantify the degree of anomaly associated with a certain frequency.

Intuitively, the larger the area above the portion of the curve included from a certain frequency f_i to the maximum frequency $f_{\max}$, and the more f_i differs from frequencies that are greater than f_i. At the same time, the larger the area below the portion of the curve included from the minimum frequency $f_{\min}$ and a certain frequency f_i, and the more f_i differs from frequencies that are smaller than f_i.

You can evaluate the contribution given by the area *above* the *cumulated frequency distribution curve* to the outlierness of a certain frequency f_i, using the following expression

$$A^{\uparrow}(f_i) = \sum_{j>i} (f_j - f_{j-1}) \cdot \left(H_n - H_{j-1}\right). \tag{4}$$

The *lower outlier score* $out^{\downarrow}(f_i)$ is given by the the normalised area

$$out^{\downarrow}(f_i) = A^{\uparrow}(f_i)/A^{\uparrow}_{\max}(f_i), \tag{5}$$

obtained by dividing the area $A^{\uparrow}(f_i)$ by

$$A^{\uparrow}_{\max}(f_i) = (A^{\uparrow}(f_0) - A^{\uparrow}(f_i)) + (f_n - f_i) \cdot (H_n - H_{i-1}) \tag{6}$$

corresponding to the area above the cumulated frequency histogram up to the frequency f_i, represented by the term $(A^{\uparrow}(f_0) - A^{\uparrow}(f_i))$, plus an upper bound to the area above the cumulated frequency histogram starting from f_i, represented by the term $(f_n - f_i) \cdot (H_n - H_{i-1})$. Notice that the former term is minimised for $f_i \to 1$, while the latter term tends to $A^{\uparrow}(f_i)$ for $f_n \to \infty$ and, hence, in this case $out^{\downarrow}(f_i)$ tends to its maximum value 1.

The second scenario we are interested in aims to highlight the *upper outliers*, namely those objects that, for a given attribute, assume a value whose frequency is high, while typically, the dataset objects assume distinct values, that is the frequencies of the other values are low.

In order to discover such a kind of anomaly we take into account the area *below* the cumulated frequency distribution, starting from the lowest frequency up to the target frequency f_i. The bigger this area, the more this frequency can be highlighted as anomalous. The contribution of the frequency f_i is computed as

$$A^{\downarrow}(f_i) = \sum_{j \le i} (f_j - f_{j-1}) \cdot H_{j-1}. \tag{7}$$

The *upper outlier score* $out^{\uparrow}(f_i)$ is given by the the normalised area

$$out^{\uparrow}(f_i) = A^{\downarrow}(f_i)/A^{\downarrow}_{\max}(f_i), \tag{8}$$

obtained by dividing the area $A^{\downarrow}(f_i)$ by the term

$$A^{\downarrow}_{\max}(f_i) = (f_i - 1) \cdot H_i \tag{9}$$

representing an upper bound to the area below the cumulated frequency histogram up to the frequency f_i. Notice that $A^{\downarrow}(f_i)$ tends to $A^{\downarrow}_{\max}(f_i)$ for $f_{i-1} \to 1$ and $H_i \to H_{i-1}$, or equivalently $\mathcal{F}(f_i) \ll \mathcal{F}(f_{i-1})$, so in this case $out^{\uparrow}(f_i)$ tends to its maximum value 1.

The outlierness, or abnormality score, associated with the frequency f_i is a combined measure of the above two normalised areas:

$$out(f_i) = \frac{W_i^{\uparrow} \cdot out^{\uparrow}(f_i) + W_i^{\downarrow} \cdot out^{\downarrow}(f_i)}{W_i^{\uparrow} \cdot \Delta(out^{\uparrow}(f_i)) + W_i^{\downarrow} \cdot \Delta(out^{\downarrow}(f_i))} \tag{10}$$

Specifically, the (global) outlierness score of f_i is the weighted mean of the upper and lower outliernesses associated with f_i, with weights $W_i^{\uparrow} = H_i$ and $W_i^{\downarrow} = (H_n - H_{i-1})$, respectively. Note that H_i represents the fraction of the frequencies having value less or equal than f_i, while $(H_n - H_{i-1})$ represents the fraction of the frequencies having value greater or equal than f_i. Thus, when both the contributions $out^{\uparrow}(f_i)$ and $out^{\downarrow}(f_i)$ are greater than 0, the two weights provide their relative importance in terms of the fraction of the data population used to compute each of them. As for the function $\Delta(x)$, it evaluates to 0 if $x = 0$, and to 1 otherwise. Thus, it serves the purpose of ignoring the weight associated with the lower or upper outlierness if it evaluates to 0 and, otherwise, of taking it into account in its entirety.

In order to clarify areas employed for outlierness computation, let us refer to the following example. Consider a single attribute dataset whose associated set of distinct frequencies is $\{f_1 = 1, f_2 = 2, f_3 = 3, f_4 = 4, f_5 = 5, f_6 = 6\}$ and the set of weights is $\{w_1 = 3, w_2 = 2, w_3 = 1, w_4 = 2, w_5 = 1, , w_6 = 2\}$. Assume that we want to compute the outlierness associated with the frequency $f_3 = 3$. Figure 2a and b represent the areas exploited to compute such outlierness.

On the left the area $A_i^{\downarrow}$ together with the area used for normalisation, $A_{\downarrow}^{\max}$, is reported, while, on the right, the area $A_i^{\uparrow}$ together with the area used for normalisation, $A_{\uparrow}^{\max}$, is reported.

If $W_i^{\downarrow} > W_i^{\uparrow}$ we say that the global score is an *upper score*. Conversely if $W_i^{\uparrow} \le W_i^{\downarrow}$ we say that the global score is a *lower score*.

We will use the notation $out_{\mathcal{H}}(\cdot)$ whenever it is needed to highlight the frequency distribution $\mathcal{H}$ used to compute the outlierness. The outlierness $out_a(v, D)$ of the value $v \in$

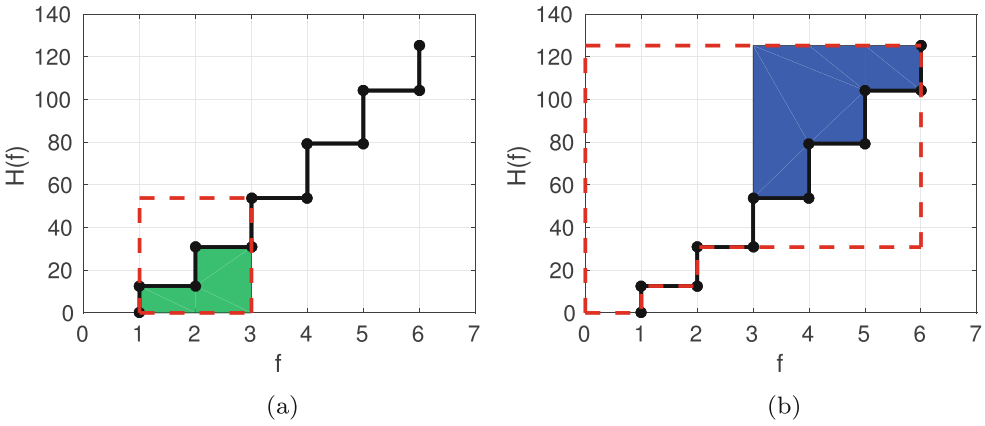


Fig. 2 Outlierness computation example

$D[a]$ with respect to the attribute a in the dataset $\mathcal{D}$, is given by $out_{\mathcal{H}^{\mathcal{D}[a]}}(f_v^{\mathcal{H}^{\mathcal{D}[a]}})$.

5 Outstanding explanation-property pairs

Exceptional values v for an attribute a, are those associated with large values of outlierness $out_a(v, D)$. Thus, we are interested in detecting such exceptional values. However, it must be pointed out that very often a value emerges as exceptional for a certain attribute only when we restrict our attention to a subset of the whole population.

This intuition leads to the definition of the notion of explanation-property pair.

Definition 5 An explanation-property pair (E, p), or simply *pair* for the sake of conciseness, consists of condition E, also called *explanation*, and of an atomic condition $p = (p_a, p_v)$, also called *property*. By p_a (p_v, resp.) we denote the attribute (value, resp.) involved in the atomic condition p.

Given a pair $\pi = (E, p)$, $\mathcal{D}_\pi$ denotes the set of objects $\mathcal{D}_{E\cup\{p\}}$. The *outlierness* $out(\pi)$ of an explanation-property pair $\pi = (E, p)$ is the outlierness $out_{p_a}(p_v, D_E)$ of the value p_v with respect the attribute p_a in the dataset D_E.

To illustrate definitions, we exploit a running example derived from the *Breast Cancer Wisconsin* dataset [32]. The example is based on the *Clump Thickness* attribute, referred to as CT. Specifically, Table 2 refers to the frequency distribution of values in the domain of CT in the full dataset before any explanation is taken into account. Conversely, Table 2b shows the frequency distribution of the same attribute when we focus on the subset of benign tumors, together with the scores gained by each value (fourth column). The row there highlighted concerns the pair (E_1, p), with explanation $E_1 = \{(Type, 2)\}$ and property $p = (CT, 7)$.

We note that the number of possible explanation-property pairs is usually very large. So we need a mechanism that allows us to single out the subset of these pairs encoding the outstanding exceptionalities in the data and carrying no redundant information; we do that through the notions of unexpected, significant, and outstanding pairs defined in the following.

Given pairs (E, p) and (E', p), we say that (E, p) is *more specific than* (E', p), or equivalently that (E', p) is *more general than* (E, p), if $E' \subset E$. In this case, we also say that the two pairs are *related*.

Given two related pairs (E, p) and (E', p), with (E, p) more specific than (E', p), we wonder if the frequency distribution $\mathcal{H}_{p_a}^{\mathcal{D}_E}$ (also called observed distribution) is statistically different from the frequency distribution $\mathcal{H}_{p_a}^{\mathcal{D}_{E'}}$ (also called reference distribution). We can ask the question by leveraging a goodness-of-fit test, which establishes if the observed frequency distribution is unexpected given the reference distribution.

A suitable test for categorical values is the chi-square test. The chi-square test relies on the chi-square statistic X^2, which is the sum of the squared difference between the observed frequencies and the reference frequencies, normalized on the value of the expected frequencies:

$$X^2(E', E, p) = \sum_{i=1}^{r} \left[\frac{\left(h_i - h'_i \left(\frac{m}{n}\right)\right)^2}{h'_i \left(\frac{m}{n}\right)} \right] \tag{11}$$

where: (i) h'_i (h_i, resp.) is the number of occurrences in $\mathcal{D}_{E'}[p_a]$ ($\mathcal{D}_E[p_a]$, resp.) associated with the i-th distinct categorical value of $\mathcal{D}_{E'}[p_a]$, (ii) r is the number of these distinct categorical values, and (iii) n and m represent the number of objects in $\mathcal{D}_{E'}$ and $\mathcal{D}_E$, respectively.

It is known that the X^2 statistics asymptotically approaches the χ^2 distribution with $r-1$ degrees of freedom, hence, the X^2 value can be used to calculate a p-value by comparing its value to the proper chi-squared distribution. The p-value is the probability of obtaining a test statistic at least as extreme

Table 2 Behaviour of *Clump Thickness* attribute in the dataset portions selected by different set of conditions (CT = *Clump Thickness*, UCS = *Uniformity of Cell Shape*, SECS = *Single Epithelial Cell Size*, T = *Type*)

$E_0 = \emptyset$

$Value$	f	$\mathcal{F}$	$out_{CT}(\cdot, \mathcal{D}_{E_0})$
1	139	140.12	0.3323
2	50	54.32	0.3416
3	104	75.68	0.2601
4	79	74.69	0.2672
5	128	143.55	0.2496
6	33	31.87	0.4335
7	23	20.53	0.4969
8	44	56.70	0.3666
9	14	11.76	0.5693
10	69	73.78	0.2885

(a)

$E_1 = \{(T, 2)\}$

$Value$	f	$\mathcal{F}$	$out_{CT}(\cdot, \mathcal{D}_{E_1})$
1	136	106.56	0.3421
2	46	36.90	0.3325
3	92	107.26	0.1536
4	67	63.13	0.2172
5	83	113.23	0.1553
6	15	11.75	0.5466
7	1	1.94	0.6530
8	4	3.23	0.6308

(b)

$E_2 = \{(T, 2), (UCS, 2)\}$

$Value$	f	$\mathcal{F}$	$out_{CT}(\cdot, \mathcal{D}_{E_2})$
1	7	7.61	0.3468
2	3	2.81	0.5629
3	15	15.38	0.1923
4	9	8.79	0.2789
5	16	14.89	0.2676
7	1	1.51	0.6866

(c)

$E_3 = \{(T, 2), (UCS, 2), (SECS, 2)\}$

$Value$	f	$\mathcal{F}$	$out_{CT}(\cdot, \mathcal{D}_{E_3})$
1	7	8.33	0.2157
2	2	1.35	0.6032
3	11	10.68	0.1782
4	8	9.81	0.1687
5	12	9.60	0.2627
7	1	1.12	0.6759

(d)

as the one that was actually observed, assuming that the null hypothesis, that is that the observed distribution complies with the reference one, holds. Thus the value $F_{\chi^2_{r-1}}(X^2)$, where $F_{\chi^2_{r-1}}$ denotes the cumulative distribution function of the χ^2 distribution with $r-1$ degrees of freedom, provides the desired p-value.

Definition 6 (**Unexpected pair**) Given related pairs $\pi = (E, p)$ and $\pi' = (E', p)$, with (E, p) more specific than (E', p), we say that (E, p) *is unexpected given* (E', p), if $out(\pi) \geq (1 + \alpha) \cdot out(\pi')$, where $\alpha = 1 - F_{\chi^2_{r-1}}\big(X^2(E', E, p)\big)$.

Note that, the smaller the p-value provided by the cdf value $F_{\chi^2_{r-1}}\big(X^2(E', E, p)\big)$ and the less expected the distribution of categorical values in $\mathcal{D}_E[p_a]$ given the knowledge of the distribution of the categorical values in $\mathcal{D}_{E'}[p_a]$ and, consequently, $\alpha \in [0, 1]$ is inversely related to the unexpectedness of the former distribution. Hence, in order for the pair (E, p) to represent non-redundant information, we require that the outlierness of (E', p) must be improved of a quantity which is inversely related to unexpectedness of the distribution of categorical values in $\mathcal{D}_E[p_a]$.

Therefore, the concept of unexpected pair serves the purpose of avoiding that statistical fluctuations of the frequencies may slightly favor the un-typicality of the property value. Indeed, while outlierness improvements might be observed by augmenting the current explanation with some attribute-value pairs, this circumstance does not necessarily imply that the augmented explanation-property pair is more relevant than the original one, since the improvement could be so slight to be associated with statistical fluctuations of the property distribution within the two explanations. Hence, when an explanation is augmented with some attribute-value pairs, the more is preserved the property distribution with respect to the original explanation, the larger the outlierness improvement required to declare the augmented explanation-property pair as unexpected.

Consider the running example and suppose you want to extend the explanation $E_1 = \{(Type, 2)\}$ with a further condition to get the *more specific* $E_2 = \{(Type, 2), (Uniformity\ of\ Cell\ Shape, 2)\}$.

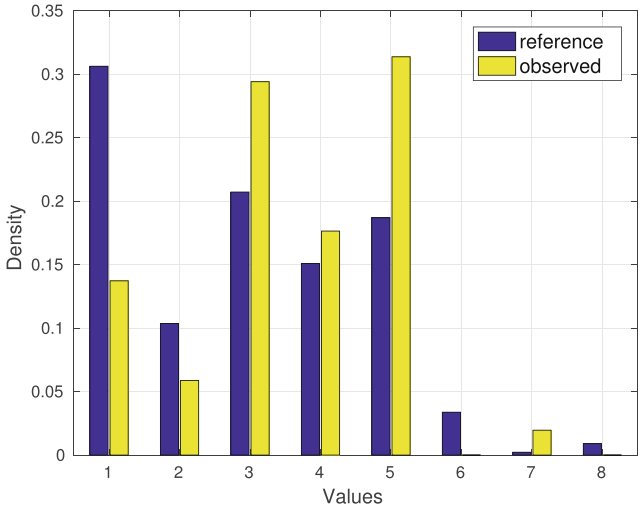
To verify if the pair (E_2, p) is unexpected given the pair (E_1, p), the frequency distributions $\mathcal{H}_{CT}^{\mathcal{D}_{E_2}}$ and $\mathcal{H}_{CT}^{\mathcal{D}_{E_1}}$ have to be compared. The plots reported Fig. 3 show the above distributions and the associated α value demonstrates that a very weak correlation exists. Thus, for the pair $(E2, p)$ to be considered unexpected given $(E1, p)$ just a slight increase in the outlierness score is necessary to satisfy Definition 6:

$$\begin{aligned} out_{CT}(7, \mathcal{D}_{E_2}) &= 0.6866 > (1 + 0.0065) \cdot out_{CT}(7, \mathcal{D}_{E_1}) \\ &= 1.0065 \cdot 0.6530. \end{aligned}$$

Consider now the property $p' = (CT, 2)$. Also the pair (E_2, p') is unexpected given (E_1, p'), indeed

$$\begin{aligned} out_{CT}(2, \mathcal{D}_{E_2}) &= 0.5629 > (1 + 0.0065) \cdot out_{CT}(2, \mathcal{D}_{E_1}) \\ &= 1.0065 \cdot 0.3325. \end{aligned}$$

However, if we consider the explanation $E_3 = E_2 \cup \{(Single\ Epithelial\ Cell\ Size, 2)\}$ although the score of the pair (E_3, p') is greater than that of the pair (E_2, p'), the



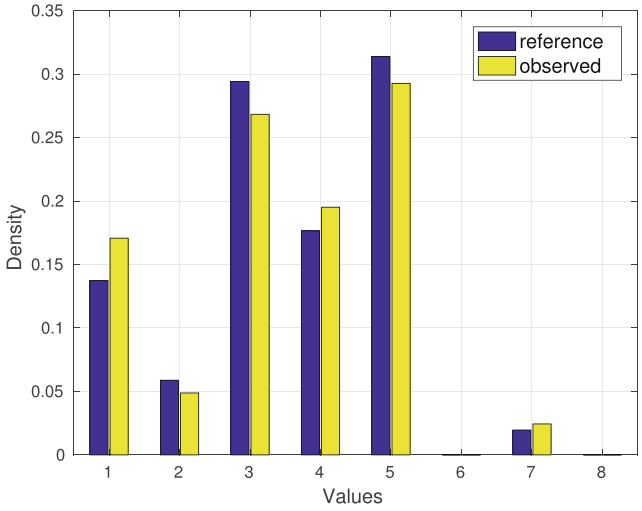
X^2	α
19.5973	0.0065

Fig. 3 $\mathcal{H}_{CT}^{\mathcal{D}_{E_2}}$ (yellow) vs $\mathcal{H}_{CT}^{\mathcal{D}_{E_1}}$ (blue)

former pair is not unexpected given the latter since (Fig. 4):

$$out_{CT}(2, \mathcal{D}_{E_3}) = 0.6031 < (1 + 0.9871) \cdot out_{CT}(2, \mathcal{D}_{E_2}) = 1.9871 \cdot 0.5629.$$

It is important to note that, given any pair (E', p), often it suffices to augment E' with a random attribute r to observe a slight outlierness improvement, that is to have $out_{p_a}(p_v, \mathcal{D}_{E' \cup \{r\}}) \geq out_{p_a}(p_v, \mathcal{D}_{E'})$. To understand why this is not unusual, assume that r is uncorrelated with the attributes in E', then the distribution $\mathcal{D}_{E'}[p_a]$ will be almost preserved in $\mathcal{D}_{E' \cup \{r\}}[p_a]$. Thus, we compute α since it precisely provides a measure of the unexpectedness of the sub-population distribution: being the value p_v fixed, larger explanations that preserve the original distribution represent redundant information. Moreover, since statistical fluctuations of the frequencies may slightly favor the untypicality of the property value, the more preserved the above distribution, the larger the outlierness improvement required. In the scenario depicted above, this almost certainly filters out the expended pair $(E \cup \{r\}, p)$, even if the associated outlierness gets larger.



X^2	α
0.6200	0.9871

Fig. 4 $\mathcal{H}_{CT}^{\mathcal{D}_{E_3}}$ (yellow) vs $\mathcal{H}_{CT}^{\mathcal{D}_{E_2}}$ (blue)

Definition 7 (**Significant pair**) A pair $(\emptyset, p)$ is *significant* by definition. Moreover, a pair (E, p) with $E \neq \emptyset$ is said to be *significant* if there exists a more general significant pair (E', p) such that (E, p) is unexpected given (E', p).

Thus, a pair is significant if it is unexpected given at least one other more general significant pair. The notion of significant pair is implemented by requiring unexpectedness of pairs with respect to a subset of the given explanation that resulted to be unexpected itself. To better understand this notion, consider the explanations E'', $E' = E'' \cup \{(a', v')\}$, and $E = E' \cup \{(a, v)\}$ and the three related pairs (E'', p), (E', p), and (E, p), with (E'', p) a significant pair. Often neither (E, p) is unexpected given (E', p) nor (E', p) is unexpected given (E'', p). This can be understood since atomic conditions may not in general carry sufficient additional correlations.

However, a more complex condition, like $E \setminus E''$, could sensibly alter the initial distribution of values. Thus, if (E, p) is unexpected given the pair (E'', p), and (E'', p) is itself significant, we say that also (E, p) is significant, no matter of the expectedness of (E, p) given (E', p) which is more specific than (E'', p), for otherwise chains as the one described above will prevent almost all pairs to be identified as significant.

Definition 8 (**Strongly significant pair**) A pair (E, p) is said to be *strongly significant* if, for any more general significant pair (E', p), the pair (E, p) is unexpected given (E', p).

In the running example, the pair (E_3, p') is significant since it is unexpected given the pair $(\emptyset, p')$, which is significant by definition, indeed:

$$out_{CT}(2, \mathcal{D}_{E_3}) = 0.6032 > (1 + 0.0288) \cdot out_{CT}(2, \mathcal{D}) = 1.0288 \cdot 0.3416.$$

However, it is not strongly significant, since (E_3, p') is not unexpected given the more general significant pair (E_2, p'). Note that the pair (E_2, p') is also significant (since it is

unexpected given $(\emptyset, p')$) and, hence, it will be strongly significant provided that any other more specific significant pair is not unexpected given (E_2, p').

Strongly significant pairs are those unexpected given any other more general significant pair. While the definition of significant pair serves the purpose of identifying those pairs that represent interesting information, the definition of strongly significant pair serves the purpose of identifying those pairs that carry no redundant information. Indeed, if a significant pair (E, p) is not unexpected given a more specific significant pair (E', p) then, from what has been stated above, it follows that the same information can be obtained from the more general pair (E', p).

Definition 9 (**Outstanding**) Maximal strongly significant pairs (E, p) are said to be *outstanding* pairs. Given threshold θ, a pair is said θ-outstanding if it is outstanding and has outlierness not smaller than θ.

Outstanding pairs can be thought of as the maximal interesting non-redundant explanation-property pairs in the dataset at hand. Outstanding pairs form the output of the technique. Specifically, we are interested in outstanding pairs associated with the highest outlierness values. Thus, given parameter N and threshold θ, the technique outputs the top–N θ-outstanding explanation-property pairs $\pi = (E, p)$ together with the associated sets of objects $\mathcal{D}_\pi$.

This strategy greatly reduces sensitivity of our outlierness measure to variations of the explanation. Moreover, by means of the strongly significant and outstanding definitions, we present to the analyst only the explanation-property pairs that are unexpected with respect to any other more general pair, thus avoiding she/he to be overwhelmed by useless information.

5.1 The *FDEOut* algorithm

In this section we describe the *FDEOut* algorithm. In order to detect outstanding pairs, the algorithm performs the exploration of the set enumeration tree associated with the explanations according to a depth-first strategy.

At a given iteration, the algorithm analyses pairs composed by an explanation E and, simultaneously, all the possible properties p concerning attributes not involved in E. The depth-first strategy is adopted for reducing the cost of the search.

In order to evaluate the score of a given pair (E, p), the dataset objects have to be grouped according to E and this can be directly exploited also to group objects according to $E \cup \{e\}$ due to the depth-first visit.

It follows from the adopted strategy that the algorithm works with partial information, since when the algorithm analyzes the pair $\delta = (E, p)$ it has not yet explored all the pairs (E', p) with $E' \subset E$.

Algorithm 1: *FDEOut*.

Input: Dataset: $\mathcal{D}$, Thresholds: ϕ, θ
Output: θ-outstanding pairs (E, p) composed of at most ϕ atomic conditions, *upper* and *lower* outliers
let $\mathcal{A}$ be the set of attributes of $\mathcal{D}$;
set $\mathcal{E}$ to $\emptyset$;
foreach $e \in \mathcal{A}$ **do**
  insert e in $\mathcal{E}$;
  sort objects in $\mathcal{D}$ according to values in e;
  `// Start exploring possible explanations, calling the recursive function checkExplanation which updates the set of outstanding pairs` $\mathcal{OP}$
  call checkExplanation($\mathcal{D}, \mathcal{E}, \mathcal{OP}$);
  remove e from E;
end
remove from $\mathcal{OP}$ all non-maximal pairs;
`// Find upper outliers`
let $\Delta_U \subseteq \mathcal{OP}$ be the of pairs associated with upper scores;
set U to $\emptyset$;
foreach $\delta = (E, p) \in \Delta_U$ **do**
  add to U the objects satisfying each condition in E and the condition p
end
`// Find lower outliers`
let $\Delta_L \subseteq \mathcal{OP}$ be the of pairs associated with lower scores;
set L to $\emptyset$;
foreach $\delta = (E, p) \in \Delta_L$ **do**
  add to L the objects satisfying each condition in E and the condition p
end
return *pairs in* $\mathcal{OP}$*, U and L*

The algorithm builds the result set composed by outstanding pairs, denoted as $\mathcal{OP}$, as follows. When $\delta = (E, p)$ is evaluated, consider the set $\Delta^{\subset}$ of pairs (E', p) with $E' \subset E$ currently in $\mathcal{OP}$ and the set $\Delta^{\supset}$ of pairs (E', p) with $E' \supset E$ currently in $\mathcal{OP}$. Note that each pair $\Delta^{\subset}$ is more general than δ and that each pair $\Delta^{\supset}$ is more specific than δ, thus, $\delta' \in \Delta^{\subset}$ can be exploited to evaluate the significance of δ and δ can be exploited to evaluate the significance of $\delta' \in \Delta^{\supset}$. In more details:

- (i) if the score of δ is lower than the threshold θ or lower than at least one pair in $\Delta^{\subset}$ then δ is dropped since it is not significant;
- (ii) if δ is significant with respect to at least one pair in $\Delta^{\subset}$ then δ is candidate to be strongly significant and is inserted in $\mathcal{OP}$ as marked;
- (iii) if δ is significant with respect to no pairs in $\Delta^{\subset}$, then δ is not strongly significant and is inserted in $\mathcal{OP}$

as not marked; note that it has to be inserted since, being significant, it can be relevant to disprove the significance of an other pair;

(iv) if the score of δ is larger than that of a $\delta' \in \Delta^{\supset}$ then δ' is dropped since it is no more significant;

(v) if $\delta' \in \Delta^{\supset}$ is marked and is not significant with respect to δ then the mark of δ' is removed since δ disprove the strongly significance of δ'.

Once the explanation set enumeration tree is explored, only the maximal marked pairs are kept in $\mathcal{OP}$.

Function checkExplanation(E).

Input: $\mathcal{D}$: the data set, $\mathcal{E}$: the set of attributes involved in explanation conditions, $\mathcal{OP}$: the set of pairs
let $\mathcal{A}$ be the set of attributes of $\mathcal{D}$;
// compute the oulierness of pairs associated with explanation involving attributes in $\mathcal{E}$ by calling checkProperty which updates the set of outstanding pairs $\mathcal{OP}$
foreach $a \in \mathcal{A} \setminus \mathcal{E}$ **do**
  checkProperty($\mathcal{D}, a, \mathcal{E}, \mathcal{OP}$)
end
// recursively check other pairs
foreach $e \in \mathcal{A} \setminus \mathcal{E}$ **do**
  insert e in $\mathcal{E}$;
  sort object in $\mathcal{D}$ according to values in $\mathcal{E}$;
  checkExplanation($\mathcal{D}, \mathcal{E}, \mathcal{OP}$);
  remove e from E;
end

6 Experimental results

In this section, we describe experimental results obtained by using the *FDEOut* algorithm.

First of all, to study the applicability of our method to real datasets, we have tested its scalability by varying the number of objects, the number of attributes, and the depth of the analysis.

Then, to clarify the different nature of the anomalies we detect w.r.t those returned by classical outlier detection methods, we have performed two families of experiments to test whether classical techniques are able to detect anomalies pointed out by our approach and to compare the detection ability of our method with related ones on known outliers.

Specifically, in Section 6.2 we employ as target the top-10 lower and upper anomalies detected by our approach for different datasets and compute their outlier scores according to the distance-based and density-based detection approaches we choose as competitors.

In Section 6.3 we inject outliers into the datasets by selecting objects within the majority class and replacing, for each attribute, the value each selected object assumes with a different value randomly picked from the attribute domain. Then, we evaluate the change in the outlierness score of the alterated objects to highlight the sensitivity of each techniques in identifying such anomalies.

Function checkProperty($\mathcal{D}$, p, E).

Input: $\mathcal{D}$: the data set, a: the property attribute, $\mathcal{E}$: the set of attributes involved in explanation conditions, $\mathcal{OP}$: the set of pairs
Output:
// Since objects of $\mathcal{D}$ are sorted according to $\mathcal{E}$, they are automatically partiotitioned in groups, where each group contains objects sharing the same values for the attributes of $\mathcal{E}$
foreach *group G of objects* **do**
  let V be the set of values objects of G assume on a;
  let E_G be the set of conditions concerning attributes in $\mathcal{E}$ defined by group G;
  // evaluate possible conditions associated with attribute a
  foreach *value v in V* **do**
    let p be the condition (a, v);
    compute the outlierness out_δ of $\delta = (E_G, p)$;
    // check significance and strongly significance for (E_G, p)
    let $\Delta^{\subset}$ be the subset of pairs $\langle E', p\rangle$ with $E' \subset E$ currently stored in Δ;
    let $\Delta^{\supset}$ be the set of pairs $\langle E', p\rangle$ with $E' \supset E$ currently in Δ;
    let $out_{\min}$ be the minimum score of pairs in $\Delta^{\subset}$;
    if $out_\delta \geq \theta$ *and* $out_\delta \geq out_{\min}$ **then**
      foreach $\delta^{\subset} \in \Delta^{\subset}$ **do**
        if δ *is significant with respect to* $\delta^{\subset}$ *according to Definition 7* **then**
          insert δ in $\mathcal{OP}$ and mark it;
        else
          insert δ in $\mathcal{OP}$ as not marked;
        end
      end
    end
    foreach $\delta^{\supset} \in \Delta^{\supset}$ **do**
      let $out^{\supset}$ be the score of $\delta^{\supset}$;
      if $out_\delta > out^{\supset}$ **then**
        remove $\delta^{\supset}$ from $\mathcal{OP}$;
      end
      if $\delta^{\supset}$ *is not significant with respect to* δ *according to Definition 7* **then**
        unmark $\delta^{\supset}$
      end
    end
  end
end

Finally, in Section 6.4 we discuss knowledge mined by means of our approach.

6.1 Scalability

Figure 5 shows the scalability analysis of our method on the *Mushrooms* dataset from *UCI Machine learning repository*

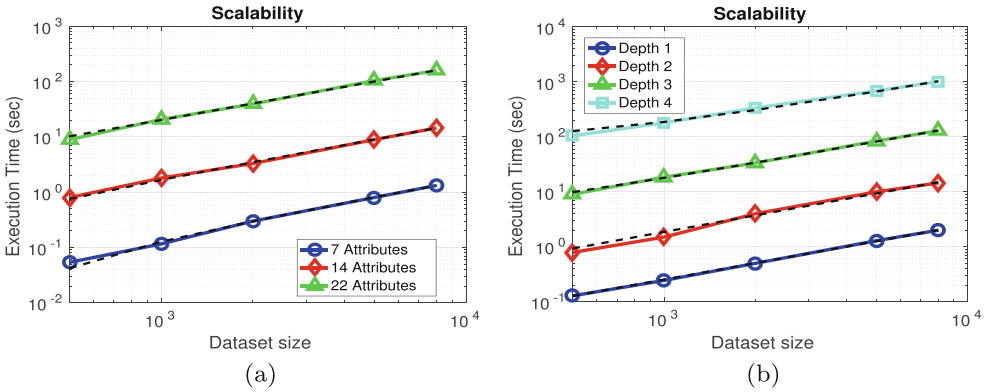


Fig. 5 Scalability analysis

[27]. In the experiment reported in Fig. 5a, we varied the number of objects n in {500, 1000, 2000, 5000, 8000} and the number of attributes m in {7, 14, 22} , while the depth parameter has been held fixed to $\delta = 3$. The dashed lines represent the trend of the linear growth estimated exploiting regression. This estimation confirms that the algorithm scales linearly with respect to the dataset size. As for the number of attributes, as expected for a given number of objects, the execution time increases due to the growth of the associated search space. On the full dataset the execution time is rather limited, as it amounts to about 2 minutes. In the experiment reported in Fig. 5b, we varied both the number of objects n and the depth parameter δ in {1, 2, 3, 4}, while considering the full feature space. Also in this case, the linear growth is represented by the dashed lines, so similar considerations can be drawn.

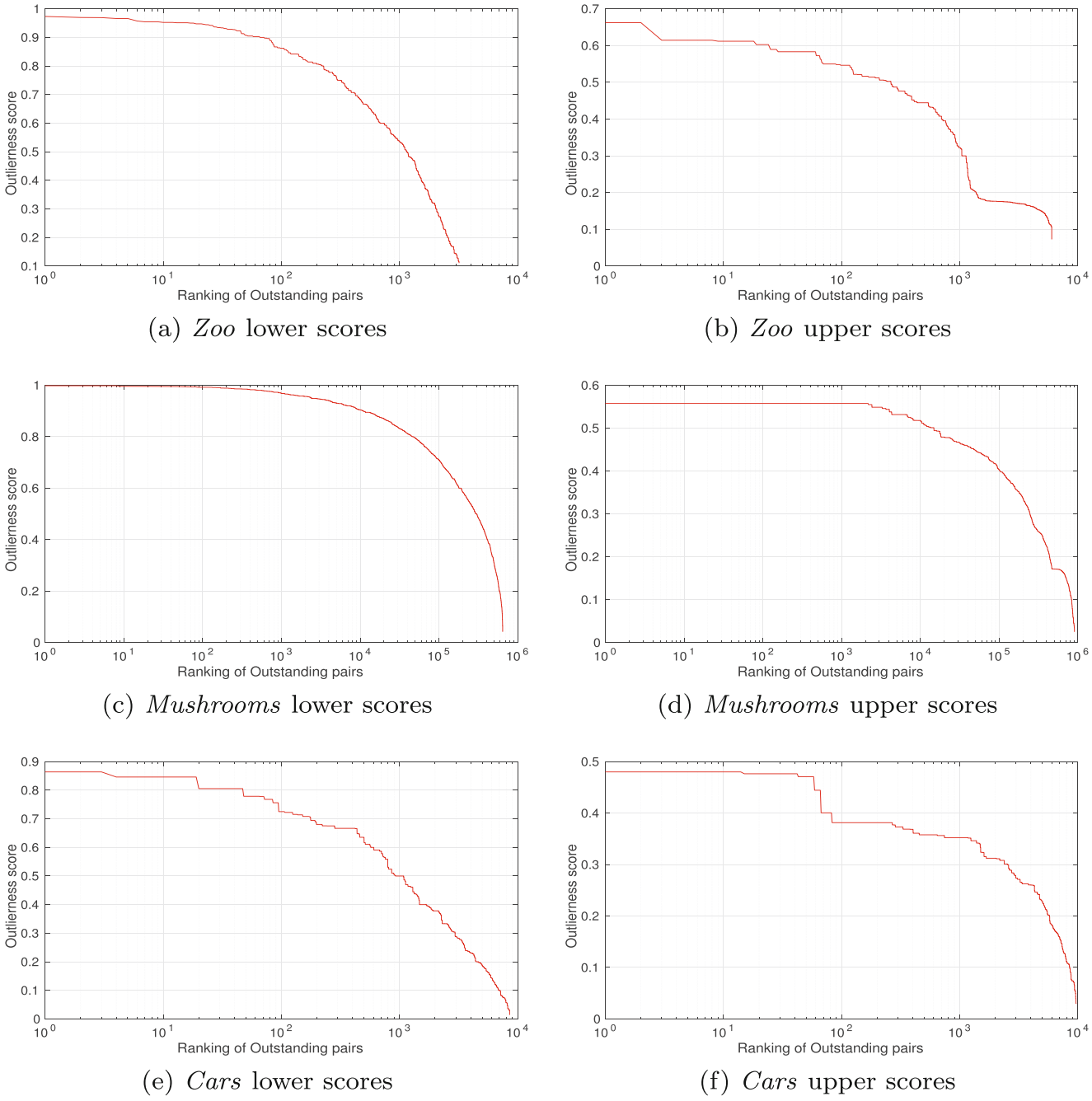
Table 3 reports the number of outstanding explanation-property pairs returned by the algorithm on the following dataset: *Zoo* ($n = 101$ objects and $m = 18$ attributes), *Mushrooms* ($n = 8,124$ objects and $m = 22$ attributes), *Cars* ($n = 1,728$ objects and $m = 8$ attributes).

We consider increasing values of the depth parameter $\delta \in \{1, 2, 3, 4, 5\}$. The column $\#Pairs$ reports the total number of pairs forming the search space up to the depth level δ. The column $\#Outstanding$ reports the number of outstanding pairs and the percentage of these pairs on the total number of pairs (within brackets). The latter column reports the number of outstanding pairs of size δ, that are the novel pairs introduced by exploring the last level of the current search space, together with their percentage on the total number of pairs (within brackets). From the table, it can be seen that the fraction of novel patterns rapidly decreases with the depth, thus suggesting that meaningful analyses do not require large values for the parameter δ. Indeed, the number of outstanding pairs settles for depth $\delta = 3$ on all the datasets. Moreover, it can be seen that the notion of outstanding pair is able to greatly reduce the number of potential explaining patterns to be presented to the user, since the percentage of these patterns on the whole search space rapidly decreases with the depth. E.g., for *Mushrooms* the outstanding pairs represent the 0.96% of the pairs of size up to $\delta = 5$.

We further note that outstanding pairs are not necessarily associated with large values of outlierness. Thus, it is sensible to determine how many of these outstanding pairs are indeed θ-outstanding, for suitable values of the threshold parameter θ. Figure 6 reports the distribution of the outliernesses associated with the outstanding pairs obtained for $\delta = 3$. Plots on the left concern the lower scores (that we recall are outlierness scores having weights $W_i^{\uparrow} \geq W_i^{\downarrow}$ in (10)), while plot on the right concern the upper scores (having weights $W_i^{\downarrow} > W_i^{\uparrow}$ in (10)), We can notice that the top lower scores are always likely to

Table 3 Outstanding pairs vs depth analysis δ

δ	#Pairs	#Outstanding	#Outstanding (size=δ)
		Zoo	
1	7,138	5,887 (82.47%)	5,320 (74.53%)
2	67,318	8,686 (12.90%)	3,958 (8.02%)
3	237,169	9,199 (3.88%)	899 (0.38%)
4	603,998	9,284 (1.54%)	181 (0.03%)
5	1,209,569	9,296 (0.77%)	43 (0.00%)
		Cars	
1	24,373	12,643 (51.87%)	12,277 (50.37%)
2	71,081	18,307 (25.76%)	14,462 (20.35%)
3	141,830	18,231 (12.85%)	4,719 (3.33%)
4	192,973	18,142 (9.4%)	1,280 (0.66%)
5	207,829	18,233 (8.77%)	534 (0.26%)
		Mushrooms	
1	1,292,236	1,119,911 (86.66%)	1,113,508 (86.17%)
2	7,475,244	1,457,732 (19.5%)	1,004,863 (13.44%)
3	27,672,066	1,551,032 (5.61%)	454,084 (1.64%)
4	75,453,759	1,566,458 (2.08%)	158,423 (0.21%)
5	163,210,877	1,567,825 (0.96%)	39,730 (0.02%)



(a) *Zoo* lower scores (b) *Zoo* upper scores

(c) *Mushrooms* lower scores (d) *Mushrooms* upper scores

(e) *Cars* lower scores (f) *Cars* upper scores

Fig. 6 Outstanding pairs outlierness distribution

reach high values of outlierness, close to 1. As for the upper scores, they are in general less pronounced, due to the different nature of these two kinds of anomalies. In any case, it is interesting to notice that only a little fraction of the outstanding pairs is associated with the largest score values and, hence, only a reduced fraction of the outstanding pairs are indeed θ-outstanding pairs. Usually, the 0.1% or 1% of the outstanding pairs are associated with outlierness values comparable to the maximum outlierness scores by any pair. It is easy to check that the number of these pairs amounts to some hundreds in the general, being about 100 for *Zoo* and *Cars*, and about 1,000 for *Mushrooms*.

6.2 Comparison with classic distance and density based outlier detection methods

We compare our method with two of the main categories of outliers: (*i*) *distance-based* approaches, that are used to discover *global* outliers, i.e. objects showing abnormal

behaviour when compared with the whole dataset population; (*ii*) *density-based* approaches, which are able to single out *local* outliers, i.e. objects showing abnormal behaviour when compared with a certain subset of the data with their neighbourhood.

As distance-based definition, we use the average KNN score, representing the average distance from the k-nearest neighbours of the object [13]. As density-based, we use Local Outlier Factor or LOF [18]. Both methods employ the Hamming distance. Moreover, we compare our method with the ROAD algorithm [47] that exploits both densities and distances, namely it establishes two independent rankings: (i) each data object is assigned to a frequency score and objects with low scores are considered outliers (Type-1 outliers); (ii) the k-mode clustering is performed in order to isolate those objects that are far from *big clusters* (i.e. clusters containing at least $\alpha\%$ of the whole dataset) according to Hamming distance (Type-2 outliers). The goal of these experiments is to highlight that we are able to detected anomalies of different nature and to provide evidence that our method is knowledge-centric, since it concentrates on anomalous values, as opposed to classical methods which are instead object-centric.

To compare the approaches, we ranked the dataset objects o by assigning to each of the them the largest outlierness of a pair π such that $o \in \mathcal{D}_\pi$. We determined our top–10 outliers by selecting the objects associated with the largest outliernesses. Then we selected these objects, containing values deemed to be exceptional by our method, with the purpose of verifying how they are ranked by popular object-centric techniques. Hence, we computed their outlier scores according to the KNN, LOF and ROAD definitions.

All the chosen competitors require an input parameter k, representing the number of k nearest-neighbors or the number of clusters to be taken into account. Since selecting the right value of k is a challenging task, we computed the KNN, LOF and ROAD outlier scores for all the possible values of k and determined the ranking positions associated with our top–10 outliers. Particularly, all the integers from 1 to the number of objects n have been considered for KNN while 30 log-space values between 1 and n have been considered for LOF due to its higher temporal cost. For ROAD algorithm, we stopped at the value of k such that at least one *big cluster* is obtained and use the frequency score to rank those objects having the same distance from their nearest big cluster.

Figures 7 and 8, report the box-plots for k varying in $[1, n]$ of the KNN, LOF and ROAD Type-2 outliers rankings associated with our top–10 outliers. Plots on the top concern lower outliers, while plots on the bottom concern upper outliers. From these plots it can be seen that the median ranking associated with our outliers can be far away from the top and also that, within the whole ranking distribution, the same outlier can be ranked in very different positions. In general, it seems that lower outliers are likely to be ranked better than upper outliers by our competitors, and this witnesses for the peculiar nature of upper outliers. On the *Zoo* dataset there is no apparent correlation between our outliers and KNN, LOF and ROAD outliers. On the *Mushrooms* dataset some of our lower outliers are, on the average, ranked very high also by the other algorithms. Some of them are almost always top outliers for all methods (see the top 1st, 2nd, 5th, and 7th outliers) thus witnessing that these outliers have both global and local nature. However, most of our outliers are not detected by these techniques.

Before concluding this comparison, it must be pointed out that the best rankings associated with the selected objects are obtained for very different values of the parameter k. Since, the output of the KNN, LOF and ROAD methods are determined for a selected value of k, it is very unlike that, even in presence of some agreement between our top outliers and local and global outliers, they are simultaneously ranked in high positions for the same provided value of k.

6.3 Comparison with other techniques

In this section, the proposed technique is compared with other methods.

To this aim, for a given dataset, we selected the objects within its majority class and then generated a family of altered datasets as follows. For each of the above objects and for each attribute, we generated a novel dataset by altering the value the object assumes on that attribute with a different value randomly picked from the attribute domain. Thus, the total number of datasets of the family is given by $n \cdot d$, where n is the number of objects within the majority class and d is the number of dataset attributes.

Due to the heavy computations required by this kind of experiment, we selected the following datasets from the *UCI Machine learning repository* [27]: *Zoo* ($n = 101$ objects and $m = 18$ attributes), *Breast cancer* ($n = 286$ objects and $m = 9$ attributes), *House votes* ($n = 232$ objects and $m = 16$ attributes).

On each family, we then ran the *FDEOut* algorithm and KNN [13], LOF [18], ROAD [47], CBRW [40], WATCH [36], and KLOF [51].

To make scores comparable, we determined the standardized outlier score $z = (sc - \mu_{sc})/\sigma_{sc}$ of the altered object both before (say z_0 this value) and after (say z_1 this value) the alteration, where μ_{sc} and σ_{sc} are, respectively, the mean and standard deviation of the outlier score distribution before the alteration. As for *FDEOut* we used as outlier score the maximum outlierness associated with an outstanding pair involving the object. Moreover, since no

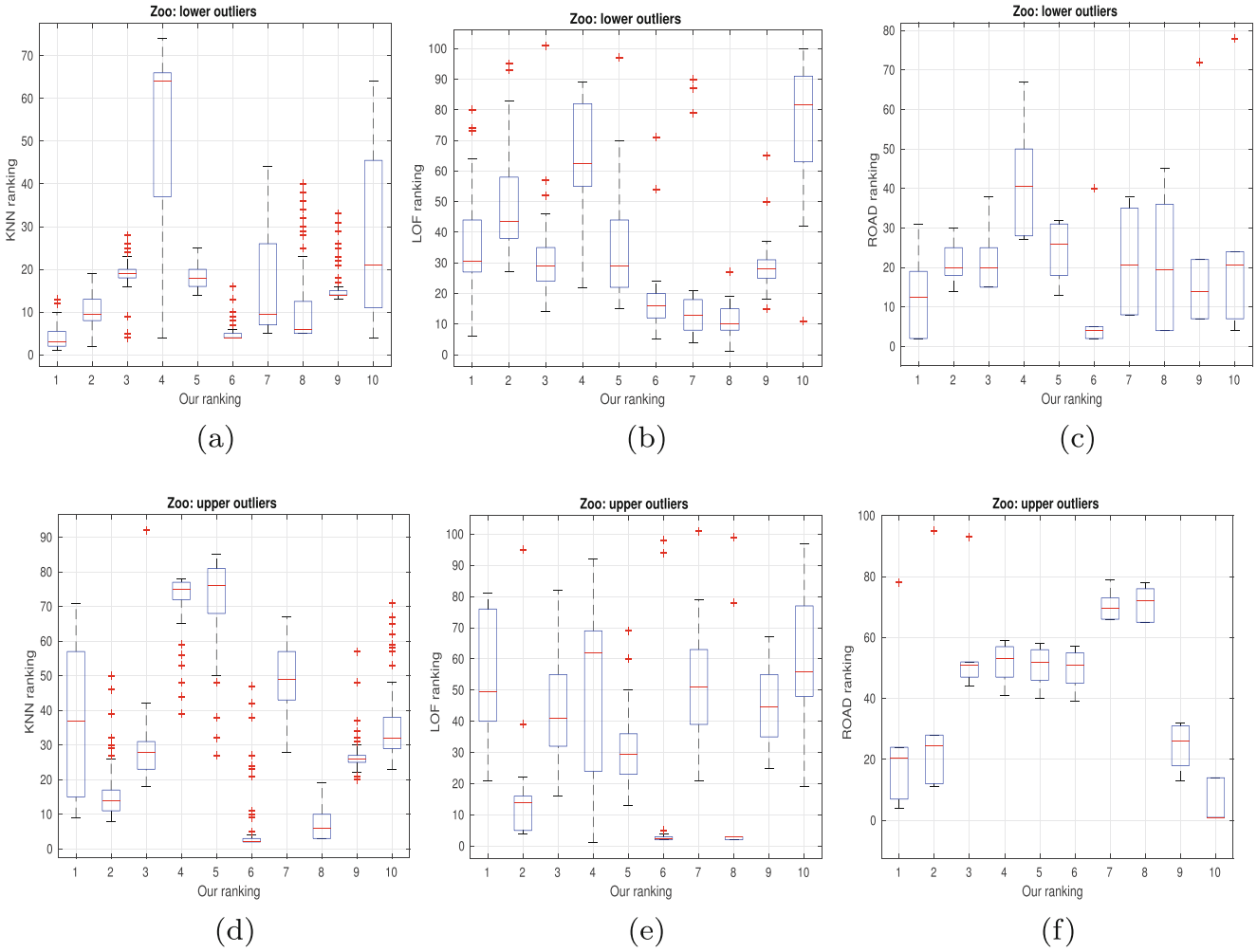


Fig. 7 Comparison with KNN, LOF and ROAD on *Zoo*

other method is designed to detect upper outliers, in the comparison we considered only our lower outliers.

In order to evaluate the ability of the method to detect the contaminated data, we measured the increase $\Delta z = z_1 - z_0$ of standardized outlier score associated with altered objects and compared the Δz of each method with that of *FDEOut*. We note that objects with large standardized outlier score z_0 in almost one of the two compared methods are less interesting for this analysis, since they already show an exceptional value of outlierness and their Δz is unlike to achieve large values. Clearly this will unfairly favor one of the two methods and, hence, objects whose standardized outlier score exceeds the mean by more than one standard deviation, i.e. such that $z_0 > 1$, are not considered in the comparison. We further note that this corresponds to focus on the normal objects that become anomalous due to the performed alteration.

Figures 9, 10 and 11 show the comparison between the distribution of the Δzs associated with *FDEOut* and the same distribution associated with each other method. On the ordinate there are the Δz values sorted in decreasing order, while each value on the abscissa corresponds to a dataset of the family.

The plots highlight that *FDEOut* is more sensitive to perturbations of the data, since in all cases the most pronounced variations of standardized score associated with *FDEOut* amount to about 3 standard deviations, while rarely the other methods exceed 1.5. As for the other methods, their quality vary with the data characteristics and, hence, they show a comparable performance on altered data. This seems to suggest that our method is able to detect also subtle anomalies.

6.4 Knowledge mined

In this section we present some knowledge mined by our method. For ease of interpretation, we report the outstanding pairs mined on the *Zoo* and *Breast cancer*. It is worth pointing out that our measure is defined at the value level, thus it does not allow us to state that an object is anomalous

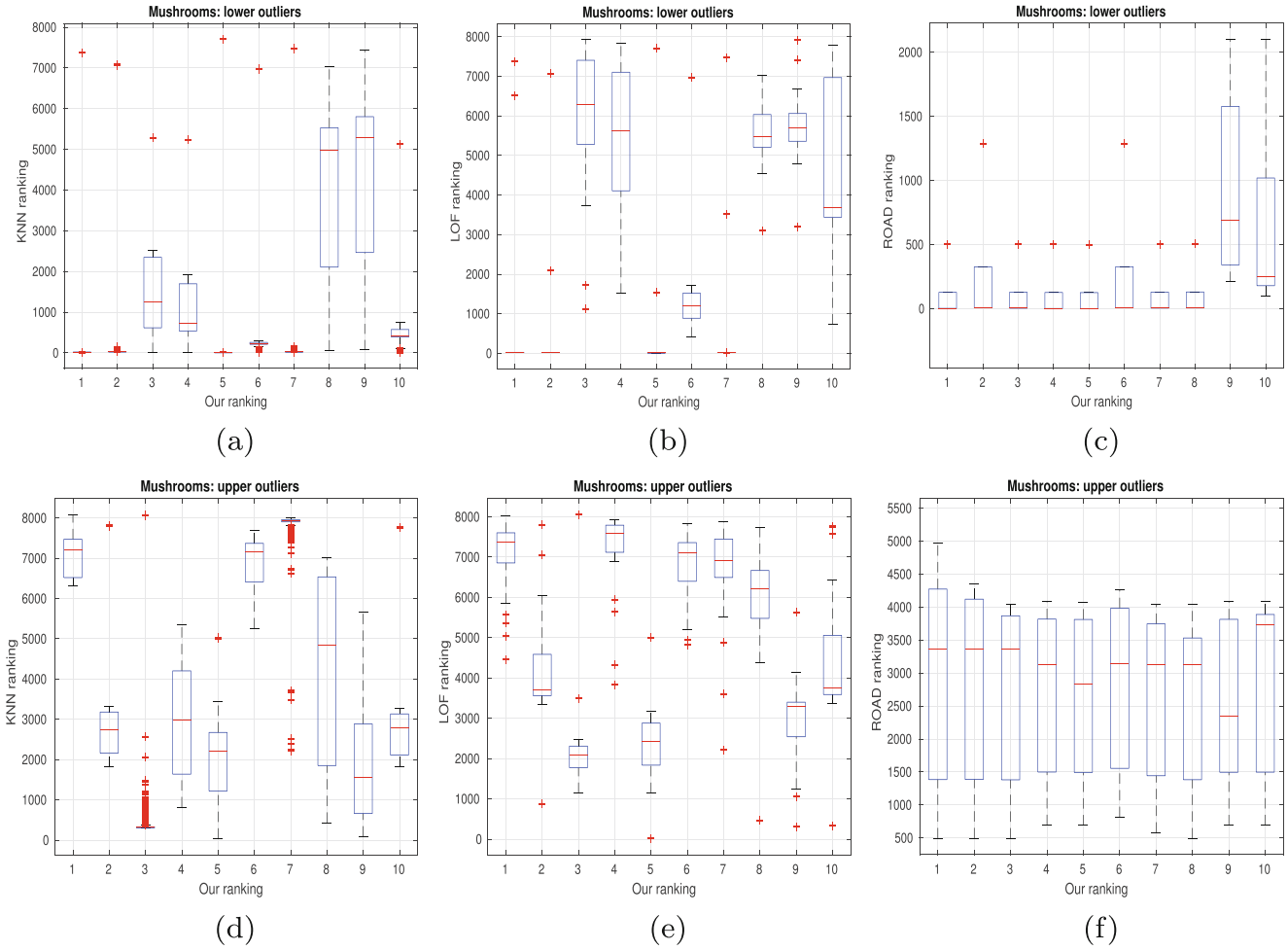


Fig. 8 Comparison with KNN, LOF and ROAD on *Mushrooms*

in an absolute sense. As our score refers to ⟨*Explanation, Property*⟩ pairs, we label as anomalous all objects, in the dataset portion isolated by the explanation, whose value p_v for attribute p_a receives a reasonable high score in such a dataset projection. Note that these objects result to be anomalous w.r.t that specific pair and not in a general sense.

Information provided by the top lower pairs we get from the *Zoo* dataset is summarized below:

- The *scorpion* is the only invertebrate with a tail.
- Among vertebrates without fins, the *seasnake* is the only one that does not breathe.
- Among non-acquatic animals, the *clam* is the only one that breathes.
- The *platypus* lays eggs although it provides milk.
- Among predators without feathers, the *ladybird* is the only airborne.
- Among catsized animals, the *octopus* is the only invertebrate.
- The *stingray* is a catsize animal, but it is venomous.
- Among animals which don't lay eggs, the *seasnake* and the *scorpion* are the only ones that do not breastfeed offspring.
- The *crab* is the only invertebrate having four legs.
- Among vertebrate breathing animals, the *pitviper* and the *frog* are the only venomous ones in the dataset.

Table 4 clarifies how such knowledge is mined. We report the outstanding pairs from which we have deduced the information above, together with the objects selected by each pair.

It is interesting to note that the same object can be considered anomalous with respect to different pairs, as in the case of the *scorpion* and the *seasnake*. Furthermore, a pair can isolate multiple objects as for the 8^{th} and 10^{th} pair.

As for the upper outliers, we find out that the dataset contains the frog twice, the former is venomous and the latter is non-venomous. However, the animal names are like primary keys for the dataset, so having the same name twice can be pointed out as anomalous. Our technique is able to highlight such a situation. Other curiosities about the

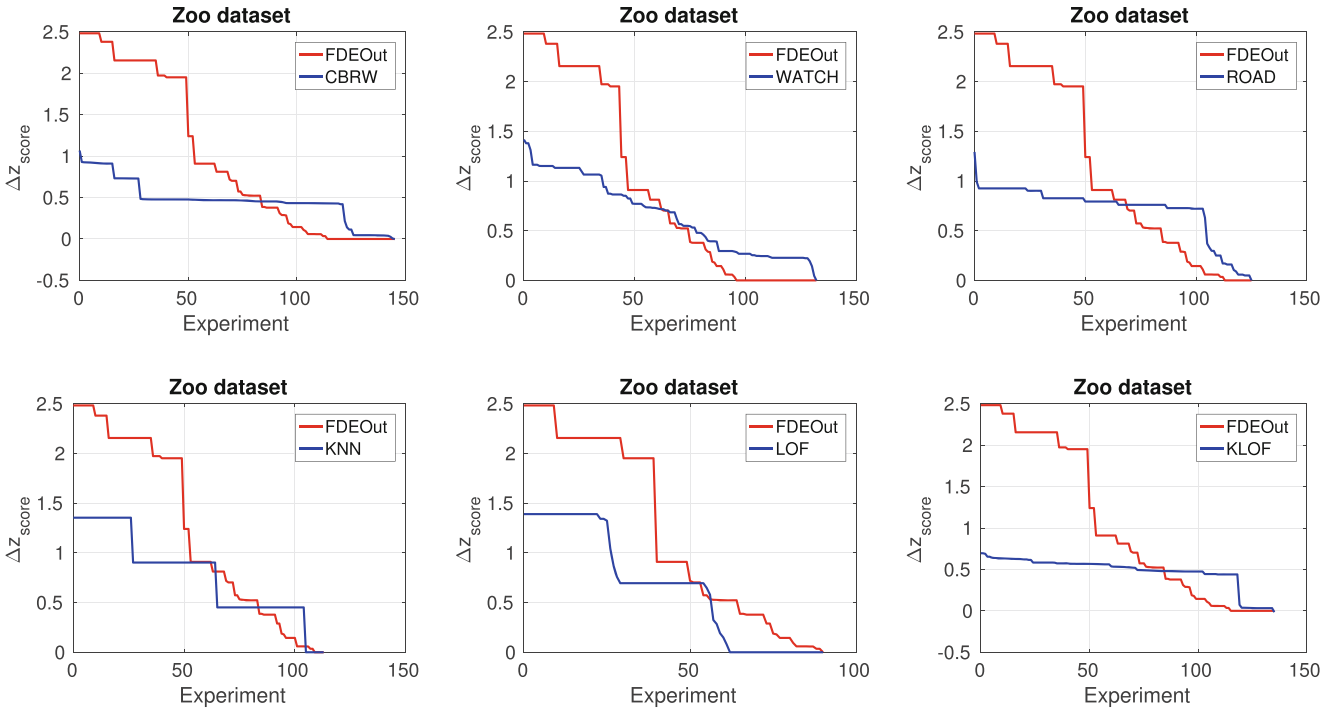


Fig. 9 Comparison on the *Zoo* dataset

animal world are spotted by our upper outliers, including the following:

- Among breathing not catsized predators, the most frequent are *non-flying birds*.
- Most no-feathers no-toothed animals have *six legs*.
- Among no-flying breathing catsized animals, the most frequent are *mammals*.
- Most *gastropods* have no legs.
- Most *no-toothed* have two legs.

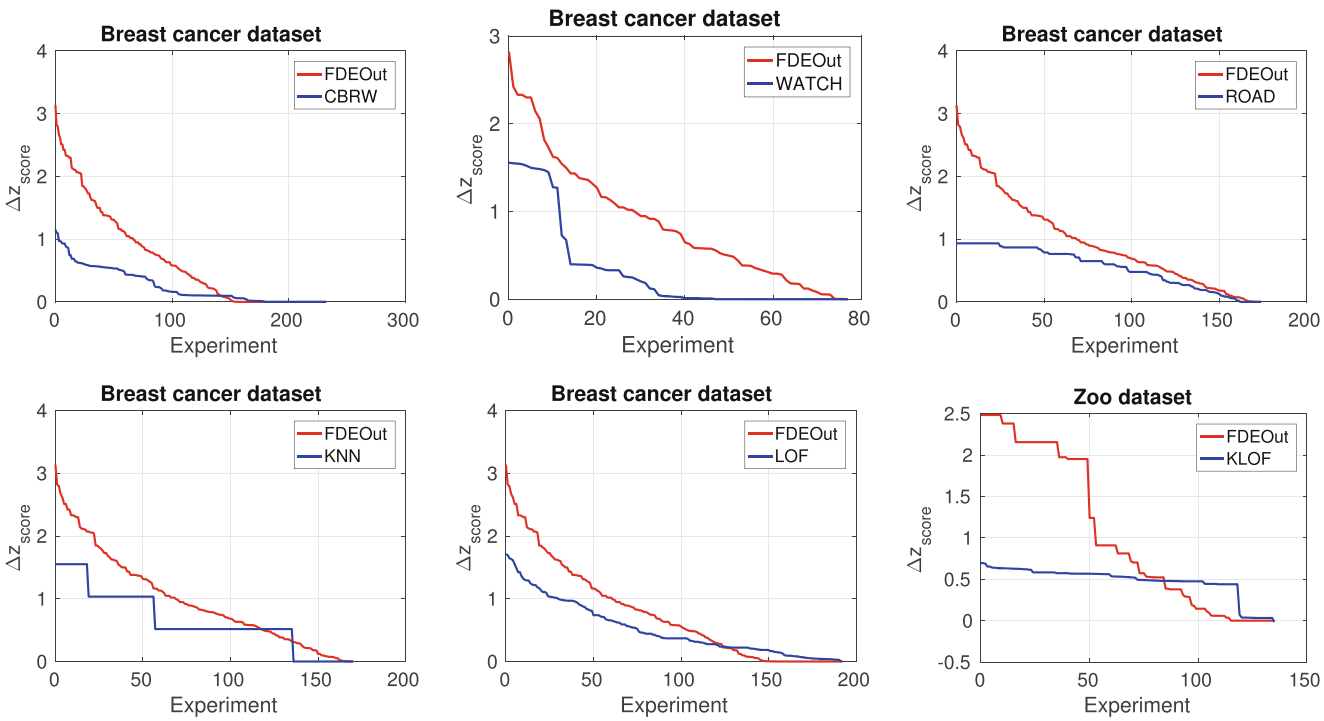


Fig. 10 Comparison on the *Breast cancer* dataset

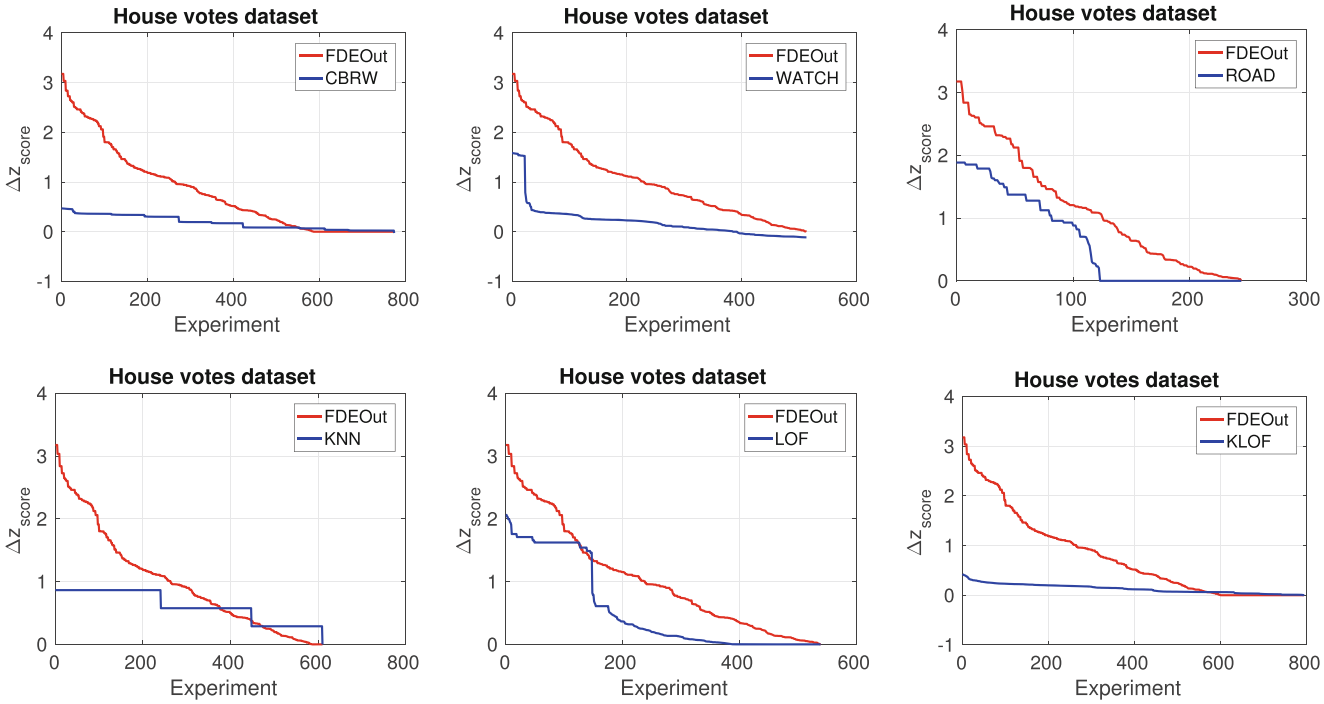


Fig. 11 Comparison on the *House votes* dataset

Table 4 Lower Outlier pairs detected no the Zoo dataset

Property	Explanation	Animals
Backbone=NO	Tail=YES	Scorpion
Breathes=NO	Backbone=YES, Fins=NO	Seasnake
Breathes=NO	Aquatic=NO	Clam
Milk=YES	Eggs=YES	Platypus
Airborne=YES	Feathers=NO, Predator=YES	Ladybird
Backbone=NO	Catsize=YES	Octopus
Venomous=YES	Catsize=YES	Stingray
Eggs=NO	Milk=NO	Seasnake, Scorpion
Backbone=NO	Legs=4	Crab
Venomous=YES	Breathes=YES, Breathes=YES	Pitviper, Frog

Table 5 Upper Outlier pairs detected no the Zoo dataset

Property	Explanation
Name=Frog	Ø
Type=No-flying birds	Predator=YES, Breathes=YES, Catsize=NO
Legs=6	Feathers=NO, Toothed=NO
Type=Mammals	Airborne=NO, Breathes=YES, Catsize=NO
Legs=0	Type=Gastropods
Legs=2	Toothed=YES

Table 6 Lower Outlier pairs detected on the Breast Cancer Wisconsin dataset (UCZ = Uniformity Cell Size; UCS = Uniformity Cell Shape; BN = Bare Nuclei; BC = Bland Chromatin; MA = Marginal Adhesion; NN = Normal Nucleoli)

Property	Explanation	Number of Objects
Mitoses=5/7/8	UCZ=1	1 object each
Type=malign	UCS=1	2 objects
Type=malign	BN=1, BC=2	1 objects
Type=malign	BN=1, MA=1	3 objects
Type=malign	BN=1, NN=1	3 objects

The table reports the number of objects sharing the anomalous property value

Table 7 Upper Outlier pairs detected no the Breast Cancer Wisconsin dataset (UCZ = Uniformity Cell Size; UCS = Uniformity Cell Shape; BN = Bare Nuclei; BC = Bland Chromatin; MA = Marginal Adhesion; NN = Normal Nucleoli, CT=Clump Thickness)

Property	Explanation
ID=1276091	Ø
ID=1182404	Ø
NN=10	UCS=10
BN=10	MA = 1, Type=malign
NN=10	UCZ=10
BN=10	CT=8
MA=10	UCS = 10

To infer this type of knowledge we consider the best ranked outstanding pairs identifying upper outliers. Table 5 outlines the ⟨*Explanation, Property*⟩ pairs we take into account. Note that in this case more objects have the value reported as anomalous, but what makes them *special* is the fact that the frequency of such values is atypical within the distribution.

As for the *Breast cancer Wisconsin*, information provided by the top lower outlier pairs is summarized below:

- When uniformity of cell size is 1, the value of attribute mitoses is almost equal to 1, except for three samples having mitoses equal to 5, 7, and 8.
- When uniformity of cell shape is 1, the tumor is always benign except for two samples.
- Among samples having bare nuclei equal to 1 and Bland Chromatin equal to 2, only one is malign.
- Among samples having both bare nuclei and marginal adhesion equal to 1, only three are malign.
- Among samples having bare nuclei and normal nucleoli equal to 1, only three are malign.

Mining upper outliers, the technique identified that there are two duplicated identifiers. Other upper outliers are discussed below:

- Among samples having uniformity of cell shape equal to 10, most have normal nucleoli equal to 10.
- Among malignant tumor having marginal adhesion equal to 1, most have bare nuclei equal to 10.
- Among samples with uniformity of cell size equal to 10, most have normal nucleoli equal to 10.
- Among samples with clump thickness equal to 8, most have bare nucleoli equal to 10.
- Among samples with uniformity cell size equal to 10, most have marginal adhesion equal to 10.

The explanation-property pairs we take into account to discuss the knowledge are reported in Table 6 for lower outliers and in Table 7 for the upper ones.

7 Conclusions

In this work we have provided a contribution to single out and explain anomalous values in categorical domains. We perceive frequencies of attribute values as samples of a distribution whose density has to be estimated. This leads to the notion of frequency occurrence we exploit to build our definition of outlier: an attribute value is suspected to be an outlier if its frequency occurrence is exceptionally typical or un-typical within the distribution of frequencies occurrences of any other attribute value. As a second contribution, our technique is able to provide interpretable explanations for the abnormal values discovered. Thus, the outliers we provide can be seen as a product of the knowledge mined, making the approach knowledge-centric rather than object centric.

The performances have been evaluated on some popular benchmark categorical datasets and a comparative view is proposed.

We notice that our method could be possibly combined with techniques that deal with outlier detection in numerical domains, in the spirit of what was done in [28, 30, 39], and we leave this as a subject of future research.

Funding Open access funding provided by Università della Calabria within the CRUI-CARE Agreement.

References

1. Aggarwal CC, Yu P (2001) Outlier detection for high dimensional data. In: SIGMOD
2. Aggarwal CC (2017) An Introduction to Outlier Analysis, pp 1–34 Springer
3. Aggarwal CC (2017) Outlier Detection in Categorical, Text, and Mixed Attribute Data, pp 249–272. Springer International Publishing, Cham
4. Angiulli F, Fassetti F (2009) Dolphin: an efficient algorithm for mining distance-based outliers in very large datasets. ACM Trans Knowl Disc Data 3(1 Article):4
5. Angiulli F, Fassetti F (2014) Exploiting domain knowledge to detect outliers. Data Min Knowl Discov 28(2):519–568
6. Angiulli F, Fassetti F, Palopoli L (2009) Detecting outlying properties of exceptional objects. ACM Trans. Database Syst 34(1)
7. Angiulli F, Fassetti F, Palopoli L (2013) Discovering characterizations of the behavior of anomalous subpopulations. IEEE Trans. Knowl. Data Eng. 25(6):1280–1292
8. Angiulli F, Basta S, Pizzuti C (2006) Distance-based detection and prediction of outliers. IEEE Trans Knowl Data Eng 18(2):145–160
9. Angiulli F, Fassetti F, Manco G, Palopoli L (2017) Outlying property detection with numerical attributes. Data Min Knowl Discov 31(1):134–163
10. Angiulli F, Fassetti F, Palopoli L (2009) Detecting outlying properties of exceptional objects. ACM Trans Database Syst (TODS) 34(1):7
11. Angiulli F, Fassetti F, Palopoli L (2013) Discovering characterizations of the behavior of anomalous subpopulations. IEEE TKDE 25(6):1280–1292
12. Angiulli F, Pizzuti C (2002) Fast outlier detection in high dimensional spaces. In: Principles of data mining and knowledge discovery, 6th european conference, PKDD 2002, helsinki, finland, august 19-23, 2002, proceedings. pp 15–26

13. Angiulli F, Pizzuti C (2005) Outlier mining in large high-dimensional data sets. IEEE Trans Knowl Data Eng 17(2):203–215
14. Barnett V, Lewis T (1994) Outliers in statistical data. Wiley, NJ
15. Bay SD, Schwabacher M (2003) Mining distance-based outliers in near linear time with randomization and a simple pruning rule. In: Proceedings of the 9th ACM SIGKDD international conference on Knowledge discovery and data mining. pp 29–38. ACM
16. Bhaduri K, Matthews BL, Giannella CR (2011) Algorithms for speeding up distance-based outlier detection. In: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. pp 859–867. ACM
17. Boriah S, Chandola V, Kumar V (2008) Similarity measures for categorical data: a comparative evaluation. In: Proceedings of the 2008 SIAM international conference on data mining. pp 243–254. SIAM
18. Breunig MM, Kriegel H, Ng R, Sander J (2000) Lof: Identifying density-based local outliers. In: SIGMOD, pp 93–104
19. Chandola V, Banerjee A, Kumar V (2009) Anomaly detection: A survey. ACM Comput. Surv 41(3)
20. Chandola V, Banerjee A, Kumar V (2009) Anomaly detection: a survey. ACM computing surveys (CSUR) 41(3):15
21. Chandola V, Boriah S, Kumar V (2009) A framework for exploring categorical data. In: SIAM Int. Conf. on data mining (SDM). pp 187–198
22. Dang XH, Assent I, Ng RT, Zimek A, Schubert E (2014) Discriminative features for identifying and interpreting outliers. In: 2014 IEEE 30Th international conference on data engineering. pp 88–99. IEEE
23. Dang XH, Micenková B., Assent I, Ng RT (2013) Local outlier detection with interpretation. In: Joint european conference on machine learning and knowledge discovery in databases. pp 304–320. Springer
24. Das K, Schneider J (2007) Detecting anomalous records in categorical datasets. In: ACM Int. Conf. on knowl. Discovery and data mining (KDD). pp 220–229
25. Dave D, Varma THR, Méan AM (2014) A review of various statestical methods for outlier detection
26. Domingues R, Filippone M, Michiardi P, Zouaoui J (2018) A comparative evaluation of outlier detection algorithms: Experiments and analyses. Pattern Recogn 74:406–421
27. Dua D, Graff C (2017) UCI machine learning repository. http://archive.ics.uci.edu/ml
28. Eiras-Franco C, Martinez-Rego D, Guijarro-Berdinas B, Alonso-Betanzos A, Bahamonde A (2019) Large scale anomaly detection in mixed numerical and categorical input spaces. Inf Sci 487:115–127
29. Ghoting A, Parthasarathy S, Otey M (2006) Fast mining of distance-based outliers in high-dimensional datasets. In: SDM. Bethesda, MD, USA
30. Ghoting A, Otey ME, Parthasarathy S (2004) Loaded: Link-based outlier and anomaly detection in evolving data sets. In: Fourth IEEE international conference on data mining (ICDM'04). pp 387–390. IEEE
31. Hancock JT, Khoshgoftaar TM (2020) Survey on categorical data for neural networks. Journal of Big Data 7(1):1–41
32. He Z, Deng S, Xu X (2005) An optimization model for outlier detection in categorical data. In: International conference on intelligent computing. pp 400–409. Springer
33. Ienco D, Pensa RG, Meo R (2016) A semisupervised approach to the detection and characterization of outliers in categorical data. IEEE Trans Neural Netw Learn Syst 28(5):1017–1029
34. Knorr EM, Ng RT (1999) Finding intensional knowledge of distance-based outliers. In: Int. Conf. on very large data bases. pp 211–222. VLDB
35. Knorr EM, Ng RT, Tucakov V (2000) Distance-based outliers: Algorithms and applications. The VLDB Journal 8(3-4):309–338
36. Li J, Zhang J, Pang N, Qin X (2018) Weighted outlier detection of high-dimensional categorical data using feature grouping. IEEE Transactions on Systems, Man and cybernetics: Systems
37. Li S, Lee R, Lang SD (2007) Mining distance-based outliers from categorical data. In: Seventh IEEE int. Conf. on data mining workshops (ICDMW 2007). pp. 225–230. IEEE
38. Liu F, Ting K, Zhou ZH (2012) Isolation-based anomaly detection. ACM Trans on Knowledge Discovery from Data (TKDD) 6(1)
39. Otey ME, Ghoting A, Parthasarathy S (2006) Fast distributed outlier detection in mixed-attribute data sets. Data Min Knowl Discov 12(2):203–228
40. Pang G, Cao L, Chen L (2016) Outlier detection in complex categorical data by modelling the feature value couplings. In: IJCAI. pp 1902–1908
41. Pang G, Cao L, Chen L, Liu H (2017) Learning homophily couplings from non-iid data for joint feature selection and noise-resilient outlier detection. In: IJCAI. pp 2585–2591
42. Pang G, Shen C, Cao L, Hengel AVD (2021) Deep learning for anomaly detection: A review. ACM Comput. Surv 54(2) (mar)
43. Pang G, Ting KM, Albrecht D, Jin H (2016) Zero++: Harnessing the power of zero appearances to detect anomalies in large-scale data sets. J Artif Intell Res 57:593–620
44. Pang G, Xu H, Cao L, Zhao W (2017) Selective value coupling learning for detecting outliers in high-dimensional categorical data. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. pp 807–816
45. Papadimitriou S, Kitagawa H, Gibbons P, Faloutsos C (2003) Loci: Fast outlier detection using the local correlation integral. In: ICDE. pp 315–326
46. Ranga Suri NNR, Murty M, Athithan N (2019) Outlier Detection in Categorical Data, pp 69–93. Springer International Publishing, Cham
47. Suri NR, Murty MN, Athithan G (2012) An algorithm for mining outliers in categorical data through ranking. In: IEEE Int. Conf. on hybrid intelligent systems (HIS). pp 247–252
48. Taha A, Hadi AS (2019) Anomaly detection methods for categorical data: a review. ACM Computing Surveys (CSUR) 52(2):38
49. Wei L, Qian W, Zhou A, Jin W, Jeffrey XY (2003) Hot: Hypergraph-based outlier test for categorical data. In: Pacific-asia conf. on knowledge discovery and data mining. pp 399–410. Springer
50. Xu H, Wang Y, Cheng L, Wang Y, Ma X (2018) Exploring a high-quality outlying feature value set for noise-resilient outlier detection in categorical data. In: Proceedings of the 27th ACM International Conference on Information and Knowledge Management. pp 17–26
51. Yu JX, Qian W, Lu H, Zhou A (2006) Finding centric local outliers in categorical/numerical spaces. Knowl Inf Syst 9(3):309–338

Fabrizio Angiulli is currently a professor of computer science at DIMES, University of Calabria, Rende, Italy. His research interests include data mining, machine learning, and artificial intelligence, with a focus on the design of anomaly detection approaches for various scenarios, efficient and effective large and high-dimensional data analysis, and explainable learning. He has authored more than one hundred papers appearing in premier journals and conference proceedings. He regularly serves on the program committee of several conferences and, as an associate editor, on the editorial board of AI Communications.

Fabio Fassetti received the Laurea degree in computer engineering in 2004 and the PhD degree in system engineering and computer science in 2008, both from the University of Calabria, Cosenza, Italy. He has been an assistant professor of computer engineering at DIMES Dept., University of Calabria, Italy, since 2012. His research interests include bioinformatics, machine learning, data mining, artificial intelligence, knowledge representation and reasoning.

Luigi Palopoli is full professor of Computer Engineering since the year 2000. He is affiliated with DIMES, Università della Calabria. His research interests are in the areas of artificial intelligence, databases, bioinformatics. In his career he published over 150 papers that appeared in prestigious international scientific journals and conference proceedings. Palopoli is an associate editor of the journals *AI Communications* (IOS press) and *Network Modeling Analysis in Health Informatics and Bioinformatics* (Springer).

Cristina Serrao received the Master Degree in Computer Engineering at University of Calabria, Rende, Italy in 2018 and she is currently a PhD Student in Information and Communication Technology at the same University. She has been a research fellow since 2015 and the main topics she is working on are about Anomaly Detection and Bioinformatics.