Detecting and Explaining Exceptional Values in Categorical Data
DISCUSSION PAPER

Fabrizio Angiulli, Fabio Fassetti, Luigi Palopoli, and Cristina Serrao

DIMES, University of Calabria, 87036 Rende (CS), Italy
{f.angiulli,f.fassetti,palopoli,c.serrao}@dimes.unical.it

Abstract. In this work we deal with the problem of detecting and explaining exceptional behaving values in categorical datasets by perceiving an attribute value as anomalous if its frequency occurrence is exceptionally typical or un-typical within the distribution of frequencies occurrences of any other attribute value. The notion of *frequency occurrence* is provided by specialising the Kernel Density Estimation method to the domain of frequency values and an outlierness measure is defined by leveraging the cdf of such a density. This measure is able to simultaneously identify two kinds of anomalies called *lower outliers* and *upper outliers*, namely exceptionally low or high frequent values.
Moreover, data values labeled as outliers come with an interpretable *explanations* for their abnormality, which is a desirable feature of any knowledge discovery technique.

CEUR-WS.org/Vol-2646/29-paper.pdf

1 Introduction

An outlying observation is one that appears to deviate markedly from other members of the sample in which it occurs. Such rare events can be even more interesting than the more regularly occurring ones as they are suspected of not being generated by the same mechanisms as the rest of the data.

We deal with categorical data and, specifically, we perceive an attribute value as anomalous if its frequency occurrence is exceptionally typical or un-typical within the distribution of frequencies occurrences of any other attribute value. To quantify such a *frequency occurrence* we specialize the classical Kernel Density Estimation technique to the domain of frequency values and get to the concept of *soft frequency occurrence*. The cumulated frequency distribution of the above density estimate is used to decide if the frequency of a certain value is anomalous when compared to the other values' frequencies. In particular, we are able to identify two kind of anomalies, namely *lower outliers* and *upper outliers*. A *lower outlier* is a value whose frequency is low while, typically, the dataset objects

Copyright © 2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0). This volume is published and copyrighted by its editors. SEBD 2020, June 21-24, 2020, Villasimius, Italy.

assume a few similar values, namely the frequencies of the other values are high. An *upper outlier* is a value whose frequency is high while, typically, the dataset objects assume almost distinct values, namely the frequencies of the other values are low. Note that the measure detects both scenarios and to automatically characterizes the target value as a lower or upper outlier.

A value can show exceptional behaviour only when we restrict our attention to a subset of the whole population. Thus, we design our technique to output the so-called *explanation-property pairs* (E, p), where E denotes a condition used to determine the target subpopulation and p represents an attribute p_a and a value p_v such that the p_v is exceptionally frequent or infrequent within the subpopulation selected by the explanation E.

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 describes experimental results.

2 Related works

Categorical data has received relatively little attention as compared to quantitative data because detecting anomalies in categorical domain is a challenging problem [11].

We start by noting that there is little literature about outlier explanation [3], i.e the problem of detecting anomalous properties and/or related outlier objects equipped with features justifying their outlierness. Moreover, to the best of our knowledge, no technique is able to natively detect *upper* outliers.

Among traditional outlier detection methods explored in the context of numerical data[8] you can consider two main clusters: distance-based[6] and density-based[7]. These ideas have been properly adapted to the categorical domain.

An example of distance-based method is discussed in [6, 1]. Outliers are defined as the N observations whose average distance to the k nearest neighbors are the greatest; in order to do that an appropriate distance has to be chosen.

Detecting local anomalies, i.e. observations having outlying behavior in local areas, is another interesting discovery problem. Local anomaly detection methods for categorical data include the k-LOF [12] and the WATCH method [9].

The first one is an extension of Local Anomalies Factor (LOF) method [7] to the categorical domain while the WATCH method [9] has been designed to find out outliers in high dimensional categorical datasets using feature grouping.

Both distance and density are taken into account by the ROAD algorithm [10]. It considers the Hamming distance to compute cluster-based outliers and the density, evaluated as the mean frequency of the values, is used to identify frequency-based outliers.

With regard to the outlier explanation, [2],[3] 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 [4]

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 $\mathbf{A}$ is a set of objects o assuming values on the attributes in $\mathbf{A}$. By $o[a]$ we denote the value of o on the attribute $a \in \mathbf{A}$. $\mathcal{D}[a]$ denotes the multiset $\{o[a] \mid o \in \mathcal{D}\}$. Given a multiset V, the frequency f_v^V of the value $v \in V$ is the number of occurrences of v in V. A *condition* $\mathcal{C}$ is a set of pairs (a, v) where each a is an attribute and each $v \in \mathcal{D}[a]$. By $\mathcal{D}_\mathcal{C}$ we denote the new dataset $\{o \in \mathcal{D} \mid o[a] = v, \forall(a, v) \in \mathcal{C})\}$.

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 $w(f_i)$.*

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. However, close frequency values do not interact with each other, e.g. having $f_i = 49, w_i = 1$ and $f_{i+1} = 51, w_{i+1} = 1$ is completely different from $f_i' = 50, w_i' = 2$, as in the former case $\mathcal{F}(f_i) = 49$ and $\mathcal{F}(f_{i+1}) = 51$ while in the latter we have $\mathcal{F}(f_i') = 100$ that is about twice the previous case. Intuitively, we do not desire a similar small variation in the frequency distribution to impact so largely on the *frequency occurrence* values. To force close frequency values to influence each other in order and jointly contribute to the frequency occurrence we design an ad-hoc density estimation method inspired to Kernel Density Estimation (KDE).

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

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

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 $V_{I(f)}(K_{f_i})$.

It is possible to eliminate the dependence from $I(f)$ by properly weighting the contribution of $K_{f_i}(\cdot)$ based on its distance from the target frequency f. Such a weight can be directly obtained from the associated kernel as the ratio between the probability of observing frequency f and the probability of observing f_i, when such frequencies are realization of a random variable distributed according to $K_{f_i}(\cdot)$. This allow us to rewrite equation 1 as follows:

$$\mathcal{F}(f) = \sum_{\varphi \geq 0} \left\{ \sum_{i=1}^{n} \left[w_i \cdot f_i \cdot K_{f_i}(\varphi) \cdot \frac{K_{f_i}(f)}{K_{f_i}(f_i)} \right] \right\}. \tag{2}$$

Note that the summation over the domain of all $K_{f_i}(\cdot)$ values is equal to 1 since it is a probability mass function. Moreover, as $\mathcal{F}$ 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

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

As for the kernel selection we 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 $p = f_i / N(\mathcal{H})$.

4 Categorical Outlierness

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 high.

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 low.

In order to discover outliers, we exploit the cumulated frequency distribution associated with the estimated density.

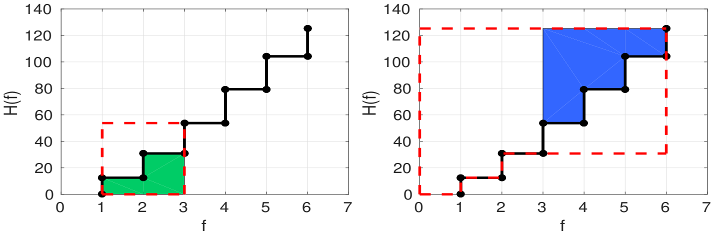
Definition 4 (Cumulated frequency distribution). *Given a frequency distribution $\mathcal{H} = \{f_1, \ldots, f_n\}$, the associated cumulated frequency distribution H is*

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

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

To quantify the degree of anomaly associated with a certain frequency, we use the area above and below the curve of the cumulated frequency distribution. Intuitively, the larger the area $A^{\uparrow}(f_i)$ 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. Thus, this area is exploited to associate *lower outlier score* $out^{\downarrow}(f_i)$ to the target frequecy f_i. At the same time, the larger the area $A^{\downarrow}(f_i)$ 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. This area can be used to associate an *uppper outlier score* $out^{\uparrow}(f_i)$ to f_i.

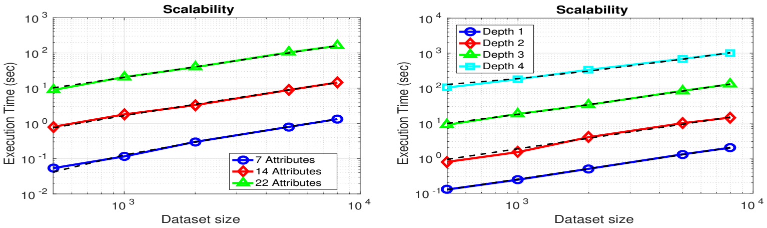
The outlierness associated with the frequency f_i is a combined measure of the above two normalised areas and exceptional values for an attribute a, are those associated with large values of outlierness. More details are available in[5].



f	w
1	3
2	2
3	1
4	2
5	1
6	2

Fig. 1: Outlierness computation example. Areas involved in the computation of the score are highlighted. They are normalized properly to get the upper and lower outlierness score for the target frequency

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.



(a) (b)

Fig. 2: Scalability analysis.

Definition 5. *An explanation-property pair* (E, p)*, 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*.*

The *outlierness* of an explanation-property pair (E, p) is the outlierness score associated with of the value p_v with respect the attribute p_a in the dataset D_E.

We implemented an algorithm that receives in input a dataset $\mathcal{D}$ and a depth parameter $\delta \geq 1$, and returns all the pairs (E, p) among those composed of at most δ atomic conditions. The algorithm analyzes explanations of length less or equal than δ according to a depth-first strategy that allows an efficient selection of sub-populations exploiting an approach similar to the one described in [3].

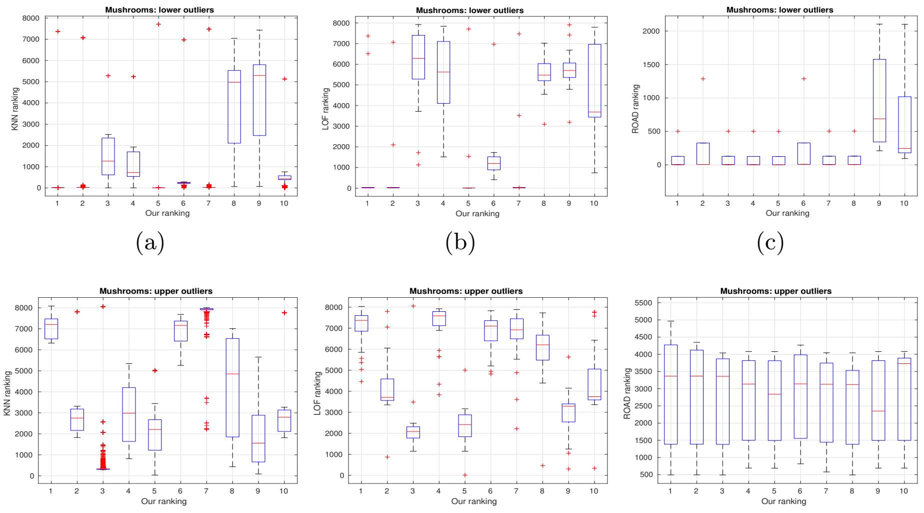
5 Experimental results

First of all, to study the applicability of our method to real datasets, we have tested its scalability. Then, to clarify the different nature of our anomalies with those returned by other outlier detection methods, we compared our method with traditional distance-based and density-based outlier detection approaches and with a method tailored for categorical data. Here, the results obtained on the *Mushrooms* dataset (n = 8,124 objects and m = 22 attributes) from UCI ML Repository are reported. More experiments are discussed in [5].

Scalability. In the experiment reported in Figure 2a, 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 very contained, as it amounts to about 2 minutes. In the experiment reported in Figure 2b, 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, and similar considerations can be drawn.

Comparison with 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; (ii) *density-based* approaches, which are able to single out *local* outliers. As distance-based definition, we use the average KNN score, representing the average distance from the k-nearest neighbours of the object. As density-based, we use Local Outlier Factor or LOF [7]. Both methods employ the Hamming distance. Moreover, we compare our method with the ROAD algorithm [10] that exploits both densities and distances.

We ranked the dataset objects o by assigning to each of the them the largest outlierness of a pair π such that $o \in \mathcal{D}_\pi$ and we determined our top–10 outliers as the objects associated with the largest outliernesses; then we computed their outlier scores according to the KNN, LOF and ROAD definitions. Note that all the chosen competitors require an input parameter k whose selection may be challanging, so we have computeted their scores for all the possible values of k and report the ranking positions associated with our top–10 outliers as a box-plots generated when k changes.



(a) (b) (c)

(d) (e) (f)

Fig. 3: Comparison with KNN, LOF and ROAD on *Mushrooms*.

The plots in fig. 3 highlight 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 *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.

Note 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 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 lead to the notion of frequency occurrence we exploit to build our definition of outlier. 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.

References

1. Angiulli, F., Basta, S., Pizzuti, C.: Distance-based detection and prediction of outliers. IEEE transactions on knowledge and data engineering **18**(2) (2006)
2. Angiulli, F., Fassetti, F., Manco, G., Palopoli, L.: Outlying property detection with numerical attributes. Data Min. Knowl. Discov. **31**(1) (2017)
3. Angiulli, F., Fassetti, F., Palopoli, L.: Detecting outlying properties of exceptional objects. ACM Transactions on Database Systems (TODS) **34**(1) (2009)
4. Angiulli, F., Fassetti, F., Palopoli, L.: Discovering characterizations of the behavior of anomalous subpopulations. IEEE TKDE **25**(6) (2013)
5. Angiulli, F., Fassetti, F., Palopoli, L., Serrao, C.: A density estimation approach for detecting and explaining exceptional values in categorical data. In: Discovery Science - 22nd International Conference, Proceedings. Springer (2019)
6. Angiulli, F., Pizzuti, C.: Outlier mining in large high-dimensional data sets. IEEE transactions on Knowledge and Data engineering **17**(2) (2005)
7. Breunig, M.M., Kriegel, H.P., Ng, R.T., Sander, J.: Lof: identifying density-based local outliers. In: ACM sigmod record. vol. 29. ACM (2000)
8. Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection: A survey. ACM computing surveys (CSUR) **41**(3) (2009)
9. Li, J., Zhang, J., Pang, N., Qin, X.: Weighted outlier detection of high-dimensional categorical data using feature grouping. IEEE SMC (2018)
10. Suri, N.R., Murty, M.N., Athithan, G.: An algorithm for mining outliers in categorical data through ranking. In: IEEE HIS (2012)
11. Taha, A., Hadi, A.S.: Anomaly detection methods for categorical data: A review. ACM Computing Surveys (CSUR) **52**(2) (2019)
12. Yu, J.X., Qian, W., Lu, H., Zhou, A.: Finding centric local outliers in categorical/numerical spaces. Knowledge and Information Systems **9**(3), 309–338 (2006)