DOI: `10.1007/978-3-319-03524-6_36`.

Outlier Detection with Arbitrary Probability Functions

Fabrizio Angiulli and Fabio Fassetti

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

Abstract. We consider the problem of unsupervised outlier detection in large collections of data objects when objects are modeled by means of arbitrary multidimensional probability density functions. Specifically, we present a novel definition of outlier in the context of uncertain data under the attribute level uncertainty model, according to which an uncertain object is an object that always exists but its actual value is modeled by a multivariate pdf. The notion of outlier provided is distance-based, in that an uncertain object is declared to be an outlier on the basis of the expected number of its neighbors in the data set. To the best of our knowledge this is the first work that considers the unsupervised outlier detection problem on the full feature space on data objects modeled by means of arbitrarily shaped multidimensional distribution functions. Properties that allow to reduce the number of probability distance computations are presented, together with an efficient algorithm for determining the outliers in an input uncertain data set.

1 Introduction

Traditional knowledge discovery techniques deal with feature vectors having *deterministic* values. Thus, data *uncertainty* is usually ignored in the analysis problem formulation.

However, it must be noted that *uncertainty* arises in real data in many ways, since the data may contain errors or may be only partially complete [1]. The uncertainty may result from the limitations of the equipment, indeed physical devices are often imprecise due to *measurement errors*. Another source of uncertainty are *repeated measurements*, e.g. sea surface temperature could be recorded multiple times during a day. Also, in some applications data values are *continuously changing*, as positions of mobile devices or observations associated with natural phenomena, and these quantities can be approximated by using an uncertain model.

Simply disregarding uncertainty may lead to less accurate conclusions or even inexact ones. This has raised the need for uncertain data management techniques [2], that are techniques managing data records typically represented by probability distributions [3–8]. In this work it is assumed that an *uncertain object* is an object that always exists but its actual value is uncertain and modeled by a multivariate probability density function [9]. This notion of uncertain object has been extensively adopted in the literature and corresponds to the *attribute level uncertainty model* viewpoint [9].

In particular, we deal with the problem of *detecting outliers in uncertain data*. An *outlier* is an observation that differs so much from other observations as to arouse suspicion that it was generated by a different mechanism [10]. As a major contribution, we introduce a novel definition of uncertain outlier representing the generalization of the classic distance-based outlier definition [11–13] to the management of uncertain data modeled as arbitrary pdfs.

There exists several approaches to detect outliers in the certain setting [14], namely, statistical-based [15], distance-based [16], density-based [17, 18], MDEF-based [19], and others [14]. However, as far as the uncertain setting is concerned, the investigation of the problem of detecting outliers is still in its infancy. Indeed, only recently some approaches to outlier detection in uncertain data have been proposed [8, 20, 21].

The method described in [8] is a density based approach designed for uncertain objects which aims at selecting outliers in subspaces. The underlying idea of the method is to approximate the density of the data set by means of kernel density estimation and then to declare an uncertain object as an outlier if there exists a subspace such that the probability that the object lies in a sufficiently dense region of the data set is negligible. We note that, differently from our approach, in [8] the density estimate does not take directly into account the form of the pdfs associated with uncertain objects, since it is performed by using equi-bandwidth Gaussian kernels centered in the means of the object distributions. Pdfs are then taken into account to determine the objects lying in regions of low density, where the density is computed as before mentioned.

In [20] authors present a distance-based approach to detect outliers which adopts a completely different model of uncertainty than ours, that is the existential uncertainty model, according to which an uncertain object x assumes a specific value v_x with a fixed probability p_x and does not exist with probability $1 - p_x$. According to this approach, uncertain objects are not modeled by means of distribution functions, but are deterministic values that may either occur or not occur in an outcome of the dataset. Hence, although [20] deals with distance-based outliers, the scenario there considered is completely different from that considered here, and the two methods are not comparable at all.

In [21] authors assume that the space of attributes is partitioned in a space of conditioning attributes and a space of dependent attributes. An uncertain object consists of a pair (l, r), where l is a tuple on a set of conditioning attributes and r is a set of tuples on a dependent attributes, also called instances. To each instance $r_j \in r$ a measure of normality is assigned, consisting in the probability of observing r_j given that both r and l have been observed. The normality of an object is then obtained as the geometric mean of the normality of all its instances. We notice that the approach presented in [21] essentially aims at detecting the abnormal instances, that, loosely speaking, are the abnormal outcomes of the uncertain objects. Thus, the task on interest in [21] is not comparable to that considered here. Moreover, uncertain objects are modeled in a way which is completely different from that considered here.

The contributions of this work are summarized next:

- To the best of our knowledge, this is the first work that considers unsupervised outlier detection on the full feature space on data objects modeled by means of arbitrarily shaped multidimensional distribution functions;
- We introduce a novel definition of uncertain outlier representing the generalization of the classic distance-based outlier definition [11–13] to the management of uncertain data modeled as pdfs;
- Specifically, our approach consists in declaring an object as an outlier if the probability that it has at least k neighbors sufficiently close is low. Hence, it corresponds to perform a nearest neighbor density estimate on all the possible outcomes of the dataset. As such, its semantics is completely different from previously introduced unsupervised approaches for outlier detection on uncertain data;
- We provide an efficient algorithm for the computation of the uncertain distance-based outliers, which works on any domain and with any distance function.

The rest of the paper is organized as follows. Section 2 introduces the definition of uncertain outlier and some other preliminary definitions and properties. Section 3 details how to compute the outlier probability. Section 4 presents the outlier detection method. Section 5 illustrates experimental results. Finally, Section 6 concludes the work.

2 Preliminaries

2.1 Uncertain Objects

Let $(\mathbb{D}, \mathrm{d})$ denote a metric space, where $\mathbb{D}$ is a set, also called *domain*, and d is a *metric distance* on $\mathbb{D}$. (e.g., $\mathbb{D}$ is the d-dimensional real space $\mathbb{R}^d$ equipped with the Euclidean distance d).

A *certain object* v is an element of $\mathbb{D}$. An *uncertain object* x is a random variable having domain $\mathbb{D}$ with associated probability density function f^x, where $f^x(v)$ denotes the density of x in v.

We note that a certain object v can be regarded as an uncertain one whose associated pdf f^v is $\delta_v(t)$, where $\delta_v(t) = \delta(0)$, for $t = v$, and $\delta_v(t) = 0$, otherwise, with $\delta(t)$ denoting the Dirac delta function.

Given a set $S = \{x_1, \ldots, x_N\}$ of uncertain objects, an *outcome* I_S of S is a set $\{v_1, \ldots, v_N\}$ of certain objects such that $f^{x_i}(v_i) > 0$ $(1 \leq i \leq N)$. The pdf f^S associated with S is

$$f^S(v_1, \ldots, v_N) = \prod_{i=1}^{N} f^{x_i}(v_i).$$

Given two uncertain objects x and y, $\mathrm{d}(x, y)$ denotes the continuous random variable representing the *distance* between x and y.

2.2 Uncertain Outliers

Given an uncertain data set $\mathbf{DS}$, $D_k(x, \mathbf{DS} \setminus \{x\})$ (or $D_k(x)$, for short) denotes the continuous random variable representing the distance between x and its k-th nearest neighbor in $\mathbf{DS}$.

We are now in the position of providing the definition of uncertain distance-based outlier.

Definition 1. Given an uncertain data set $\mathbf{DS}$, an *uncertain distance-based outlier* in $\mathbf{DS}$ according to parameters k, R and $\delta \in (0, 1)$, is an uncertain object x of $\mathbf{DS}$ such that the following relationship holds:

$$Pr(D_k(x, \mathbf{DS} \setminus \{x\}) \leq R) \leq 1 - \delta.$$

That is to say, an uncertain distance-based outlier is a data set object for which the probability of having k data set objects besides itself within distance R is smaller than $1 - \delta$.

Let N be the number of objects in $\mathbf{DS}$. In order to determine the probability $D_k(x)$, the following multi-dimensional integral has to be computed, where $\mathbf{DS}'$ denotes the data set $\mathbf{DS} \setminus \{x\}$:

$$\int_{\mathbb{D}^N} f^x(v) \cdot f^{\mathbf{DS}'}(I_{\mathbf{DS}'}) \cdot \mathbf{I}[D_k(v, I_{\mathbf{DS}'}) \leq R] \; \mathrm{d}I_{\mathbf{DS}'} \; \mathrm{d}v,$$

where the function $\mathbf{I}(\cdot)$ outputs 1 if the probability of its argument is 1, and 0 otherwise.

It is clear that deciding if an object is an uncertain distance-based outlier is a difficult task, since it requires to compute an integral involving all the outcomes of the data set. However, in the following sections we will show that this challenging task can be efficiently addressed.

2.3 Further Definitions and Properties

W.l.o.g. it is assumed that each uncertain object x is associated with a finite region $\mathrm{SUP}(x)$, containing the support of x, namely the region such that $Pr(x \notin \mathrm{SUP}(x)) = 0$ holds. For example, SUP could be defined as an hyper-ball or an hyper-rectangle (e.g. the minimum bounding rectangle or MBR).

If the support of x is infinite, then $\mathrm{SUP}(x)$ is such that $Pr(x \notin \mathrm{SUP}(x)) \leq \pi$, for a fixed small value π, and the probability for x to exist outside $\mathrm{SUP}(x)$ is considered negligible. In this case the error involved in the calculation of the probability $Pr(\mathrm{d}(x, y) \leq R)$ is the square of π.

For example, assume that the data set objects x are normally distributed with mean μ_x and standard deviation σ_x. If the region $\mathrm{SUP}(x)$ is defined as $[\mu_x - 4\sigma_x, \mu_x + 4\sigma_x]$ then the probability $\pi = Pr(x \notin \mathrm{SUP}(x))$ is $\pi = 2 \cdot \Phi(-4) \approx 0.00006$ and the maximum error is $\pi^2 \approx 4 \cdot 10^{-9}$.

The *minimum distance* $mindist(x, y)$ between x and y is defined as $\min\{\mathrm{d}(u, v) : u \in \mathrm{SUP}(x) \wedge v \in \mathrm{SUP}(y)\}$, while the *maximum distance* $maxdist(x, y)$ between x and y is defined as $\max\{\mathrm{d}(u, v) : u \in \mathrm{SUP}(x) \wedge v \in \mathrm{SUP}(y)\}$.

Consider the two following definitions.

Definition 2. Let $D_k^m(x)$ denote the smallest distance for which there exists exactly k objects y of $\mathbf{DS}$ such that $maxdist(x,y) \leq D_k^m(x)$.

Definition 3. Let $d_k^m(x)$ denote the smallest distance for which there exists exactly k objects y of $\mathbf{DS}$ such that $mindist(x,y) \leq d_k^m(x)$.

The following two properties hold.

Property 1. Let x be an uncertain object for which $d_k^M(x)$ is less or equal than R. Then x is not an outlier.

As a matter of fact, if the condition of the statement of Property 1 is verified, then each outcome of x has certainly k neighbors within radius R in every outcome of the dataset.

Property 2. Let x be an uncertain object for which $d_k^m(x)$ is greater than R. Then x is an outlier.

3 Outlier probability

In this section we show how the value of $Pr(D_k(x) \leq R)$ can be computed, for x a generic uncertain object of $\mathbf{DS}$.

Given a certain object v and an uncertain object y, let $p_v^y(R) = Pr(\mathrm{d}(v,y) \leq R)$ denote the cumulative density function representing the relative likelihood for the distance between objects v and y to assume value less or equal than R, that is

$$p_v^y(R) = Pr(\mathrm{d}(v,y) \leq R) = \int_{\mathcal{B}_R(v)} f^y(u) \ \mathrm{d}u, \tag{1}$$

where $\mathcal{B}_R(v)$ denotes the hyper-ball having radius R and centered in v.

Let v be an outcome of the uncertain object x. For $k \geq 1$, it holds

$$\begin{aligned} &Pr(D_k(v, \mathbf{DS} \setminus \{x\}) \leq R) = \\ &\qquad = 1 - \left(\sum_{S \subseteq \mathbf{DS} : |S| < k} \left(\prod_{z \in S} p_v^z(R) \cdot \prod_{z \in \mathbf{DS} \setminus S} (1 - p_v^z(R)) \right) \right), \end{aligned} \tag{2}$$

that is one minus the probability that less than k data set objects lie within distance R from v.

Thus, the probability $Pr(D_k(x) \leq R)$ can be eventually obtained as follows:

$$Pr(D_k(x) \leq R) = \int_{\mathbb{D}} f^x(v) \cdot Pr(D_k(v, \mathbf{DS} \setminus \{x\}) \leq R) \ \mathrm{d}v. \tag{3}$$

Probability values $p_v^y(R)$ depend on the objects v and y, and on the real value R, and involve the computation of one integral with domain of integration $\mathbb{D}$ (more precisely, the hyper-ball in $\mathbb{D}$ of center v and radius R).

Algorithm 1: *UncertainDBOutlierDetector*

```
   // Candidate selection phase
1: Determine the set $OutCands$ of candidate outliers by detecting objects $x$ such
   that $d_k^M(x) > R$
   // Candidate filtering phase
2: Set $Outliers$ to the empty set
3: foreach $x$ in $OutCands$ do
4:     if $d_k^m(x) > R$ then
5:         Insert $x$ into $Outliers$;
6:     else
7:         if $Pr(D_k(x) \leq R) \leq 1 - \delta$ then
8:             Insert $x$ into $Outliers$;
9: return the set $Outliers$
```

It is known [22] that given a function g, if m points w_1, w_2, ..., w_m are randomly selected according to a given pdf f, then the following approximation holds:

$$\int g(u)\,\mathrm{d}u \approx \frac{1}{m}\sum_{i=1}^{m}\frac{g(w_i)}{f(w_i)}. \tag{4}$$

Thus, in order to compute the value $p_v^y(R)$ reported in Equation (1), the function $g_v^y(u)$ such that $g_v^y(u) = f^y(u)$ if $\mathrm{d}(v,u) \leq R$, and $g_v^y(u) = 0$ otherwise, can be integrated by evaluating the formula in Equation (4) with the points w_i randomly selected according to the pdf f^y. This procedure reduces to compute the relative number of sample points w_i lying at distance not greater than R from v, that is

$$p_v^y(R) = \frac{|\{w_i : \mathrm{d}(v,w_i) \leq R\}|}{m}. \tag{5}$$

Let $\mathbf{DS}_v$ be the subset of $\mathbf{DS} \setminus \{x\}$ such that $\mathbf{DS}_v = \{y \in \mathbf{DS} \setminus \{x\} : \mathit{mindist}(v,y) \leq R\}$. Probability $Pr(D_k(v, \mathbf{DS} \setminus \{x\}) \leq R)$ depends on probabilities $p_v^y(R)$ for the objects y belonging to the set $\mathbf{DS}_v$.

Equation (2) can be computed on the objects in the set $\mathbf{DS}_v$ by means of a dynamic programming procedure, as that reported in [23], in time $O(k \cdot |\mathbf{DS}_v|)$, that is linear both in k and in the size $|\mathbf{DS}_v|$ of $\mathbf{DS}_v$.

4 Algorithm

In this section we describe the algorithm *UncertainDBOutlierDetector* that mines the distance-based outliers in an uncertain data set **DS** of N objects.

The algorithm is reported in figure. It basically consists of two phases: the candidate selection phase, and the candidate filtering phase.

The *candidate selection phase* (see step 1 in the figure) is described next. As stated in Section 2 the uncertain objects x of **DS** satisfying $d_k^M(x) > R$ are a superset of the outliers in $\mathbf{DS}$, a suitable set $OutCands$ of uncertain *candidate outliers* can be obtained by considering $\mathbf{DS}$ as a set of certain objects equipped with the certain distance $maxdist$. Indeed, in this case the certain outliers in $\mathbf{DS}$ are precisely the objects x of $\mathbf{DS}$ such that $d_k^M(x) > R$. Thus, an efficient certain outlier detection algorithm can be exploited to select the candidate outliers. In particular, in step 2 the state of the art certain distance-based outlier detection algorithm DOLPHIN is employed [24].

After having determined the set of candidate outliers $OutCands$, the *candidate filtering phase* (see steps 3-9 in the figure) determines the set $Outliers$ of uncertain outliers in $\mathbf{DS}$. The objects x of $OutCands$ such that $d_k^m(x) > R$ can be safely inserted into $Outliers$ since, as stated in Section 2, they are outliers for sure. We call these objects *ready outliers*. As for the non-ready outliers x, it has to be decided whether $Pr(D_k(x) \leq R) \leq 1 - \delta$ or not. We note that the sets $\mathbf{DS}_x$ associated with the object x in $OutCands$, which are needed in order to compute the probability $Pr(D_k(x) \leq R)$, are computed during the candidate selection phase.

Next we analyze the temporal cost of the algorithm. Let d denote the cost of computing the distance d between two certain objects of $\mathbb{D}$ and also the distances $maxdist$ and $mindist$ between two uncertain objects of $\mathbb{D}$. Let N_c denote the number of outlier candidates, let m denote the number of samples employed to evaluate integrals by means of the formula in Equation (5), let $m_c \leq m$ denote the mean number of samples needed to decide if $Pr(D_k(x) \leq R) \leq 1-\delta$, and let N_n denote the mean size of the sets $\mathbf{DS}_x$, for x a candidate outlier. The worst case cost of the candidate selection phase (step 1) corresponds to $O(\frac{k}{p}Nd)$, where $p \in (0, 1]$ is an intrinsic parameter of the data set at hand (for details, we refer the reader to [24], where it is shown that the method has linear time performance with respect to the data set size). As for step 7, for each outcome v of x, computing the probability $Pr(\mathrm{d}(v, y) \leq R)$, with $y \in \mathbf{DS}_v$, costs $O(md)$, while computing the probability $Pr(D_k(v, \mathbf{DS} \setminus \{x\}) \leq R)$ costs kN_n. Thus, deciding for $Pr(D_k(x) \leq R) \leq 1 - \delta$ costs $m_cN_n(md + k)$. As a whole, the candidate filtering phase costs $O(N_cm_cN_n(md+k))$. Thus, the last phase of the algorithm is the potentially heaviest one, since it involves integral calculations needed to compute the outlier probability. In order to be practical, the algorithm must be able to select a number of outlier candidates N_c close to the value αN of expected outliers ($\alpha \in [0, 1]$) and to keep as lower as possible the number $m_c \ll m$ of integral computations associated with the terms $Pr(D_k(v, \mathbf{DS} \setminus \{x\}) \leq R)$.

5 Experimental results

In this section, we describe experimental results carried out by using the *UncertainDBOutlierDetector* algorithm. In all the experiments, we employ the following parameters: the number of neighbors was set to $k = \lfloor \varrho N \rfloor$, with $\varrho = 0.001$, the probability threshold δ is set to 0.9, m_0 to 30 and m to 100. The experiments are conducted on a Intel Xeon 2.33 GHz based machine under the Linux operating system.

5.1 Data sets employed

In order to evaluate the performance of the introduced method, we performed two sets of experiments.

Firstly, we considered a family of synthetic data in order to show the scalability of the approach when the number of objects and the number of dimensions of the data set increases. Secondly, we considered two families of real data sets in order to study how parameters influence the number of candidate outliers and of ready outliers.

Each data set is characterized by a parameter γ (also called *spread*) used to set the degree of uncertainty associated with data set objects.

As for the *Synthetic* data sets, a family differing for the number N of uncertain objects and the number D of attributes is generated according to the following strategy. The uncertain objects in each data set form two normally distributed separated clusters with mean $(-10, 0, \ldots, 0)$ and $(10, 0, \ldots, 0)$, respectively. Moreover, the 3‰ of the data set objects are uniformly distributed in a region lying on the hyper-plane $x = 0$ (that is to say, their first coordinate is always zero). Uncertain objects are randomly generated and may use a normal, an exponential or a uniform distribution whose spread is related to the standard deviation of the overall data by means of the parameter γ.

As far as the real data sets are concerned, we employed two 2-dimensional data sets from the R-Tree Portal[1], that are *Cities*, containing 5,922 city and village locations in Greece, and *US Point*, containing 15,206 points of populated places in USA. For both data sets, a family of uncertain data sets has been obtained as follows. An uncertain object x_i has been associated with each certain object v_i in the original data set, whose pdf $f^{x_i}(u)$ is a two-dimensional normal, uniform or exponential randomly selected distribution centered in x_i and whose spread is, again, related to the standard deviation of the overall data by means of the parameter γ.

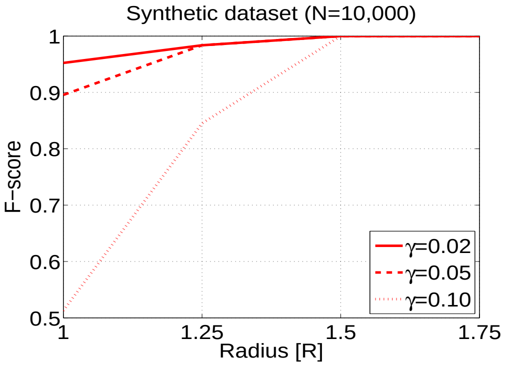
5.2 Scalability analysis

These experiments are intended to study the scalability of the method when the size of the dataset increases both in terms of the number of objects and in terms of the number of dimensions.

All the experiments are conducted with three different values of the parameter γ, namely 0.02, 0.05 and 0.1, in order to show how the method behaves for different levels of uncertainty.

Figure 2 on the left shows the scalability of the method with respect to the number N of data set objects. In this experiment, N has been varied between 10,000 and 1,000,000, while the number of dimensions D has been held fixed to 3. These curves show that the method has very good performances for different values of spread. In particular, the execution time is below 1,000 seconds even for one million of objects, confirming that the method is able to manage large data sets.

[1] See `http://www.rtreeportal.org`.



(a) Accuracy of the *Synthetic* data set family.

γ\R	1.0	1.25	1.5	1.75
0.02	3.3‰	3.1‰	3.0‰	3.0‰
0.05	3.7‰	3.1‰	3.0‰	3.0‰
0.1	8.7‰	4.1‰	3.0‰	3.0‰

(b) Outlier detected.

Fig. 1: *Synthetic* data set family.

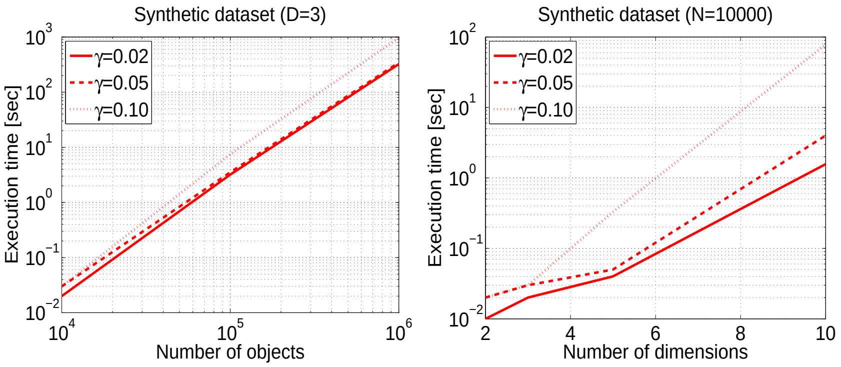


Fig. 2: Scalability with respect to the data set size and the number of dimensions for the *Synthetic* data set family.

Conversely, Figure 2 on the right shows the scalability of the method with respect to the number of dimensions D of the data set. In this case the number of objects has been held fixed to 10,000. Also in this case, time performances are good. The execution time clearly increases with the dimensionality, due to the increasing cost of evaluating outcomes of the distributions, but in this experiments remained below 100 seconds even for 10-dimensional data sets, confirming that the method can be profitably employed to analyze multidimensional data.

We studied also the accuracy of the method. Figure 1a reports the F-score as a function of the radius R. The F-score is a well known measure used to evaluate the accuracy of a method. Specifically, the F-score is a combined measure of precision and recall, where the former is the ratio between the number of outliers returned by the method and the total number of objects returned by the method, while the latter is the ratio between the number of outliers returned by the method and the total number of outliers in the dataset. In order to compute

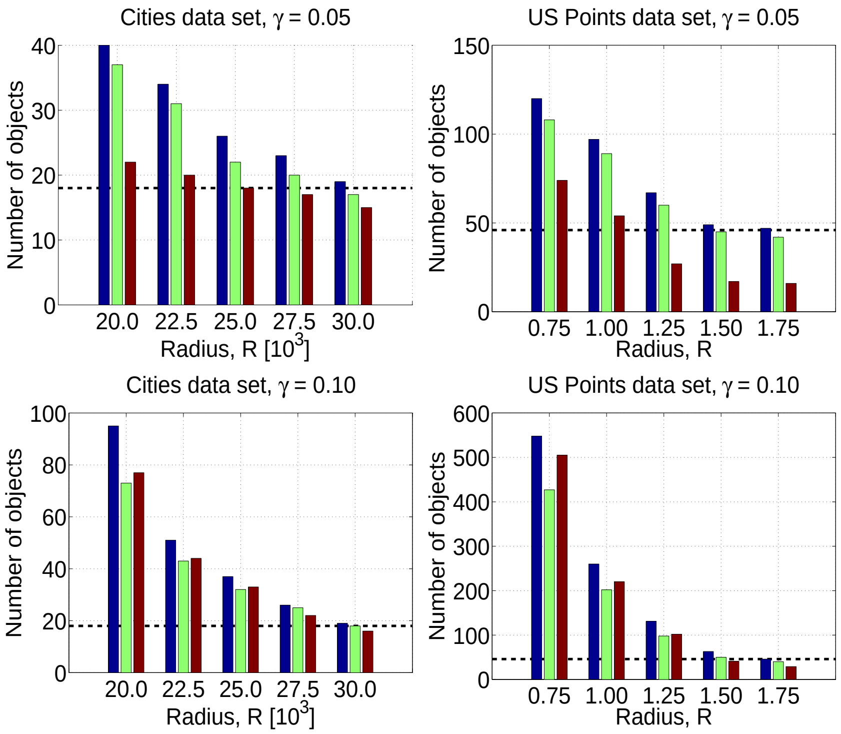


Fig. 3: Experimental results.

such measures, it is assumed that the actual outliers are all and only the objects lying in the hyper-plane $x = 0$.

In the figure three curves are reported, each referred to a different value of spread. Such curves highlight the efficiency of the approach. Indeed, for values of radius above 1.5 the F-score is equal to 1 for every considered spread, and for spread equal to 0.02 and 0.05 the F-score is almost always above 0.9 for every radii considered. For the highest spread and the lowest radius considered, the F-score lowers. This situation can be understood by considering table in Figure 1b which reports the number of outliers returned by the method.

It can be seen that for spread equal to 0.1 and radius set to 1, the number of outliers returned by the method is notably larger than the actual number of outliers. However, all the clear outliers (those lying in the hyper-plane) are correctly retrieved by the method but the method start to consider as outliers the objects lying in the tails of the distributions associated with the clusters (that we have not considered as outliers).

5.3 Sensitivity analysis

In this section, we study the behavior of the algorithm on the data sets *Cities* and *US Points* in order to study how parameters influence the number of candidate outliers and of ready outliers. Different values for the radius R and for the spread γ ($\gamma \in \{0.05, 0.1\}$) have been considered.

Figure 3 shows the result of these experiments. The first column shows results on the *Cities* data set, while the second column shows results on the *US Points* data set. The first row concerns experiments for $\gamma = 0.05$, while the second one for $\gamma = 0.10$. The diagrams report the number of candidate outliers detected at the end of the candidate selection phase (bar on the left), the actual number of outliers detected (middle bar), and the number of non-ready outliers (bar on the right). From the figure it is clear the effect of the radius on the efficiency of the method and on the number of actual outliers. The dashed line represents the number αN, with $\alpha = 0.003$. It is clear that a proper value for the radius has to be selected in order to control the actual number of outliers and, moreover, that if the radius is properly determined, the computational effort of the method results negligible (see the first two rows of the following table).

The following table shows the execution times (in seconds) of the candidate filtering phase of the algorithm.

R	*Cities*		R	*US Points*	
	$s = 0.05$	$s = 0.10$		$s = 0.05$	$s = 0.10$
30,000	0.01	0.05	1.75	0.07	0.29
27,500	0.02	0.09	1.50	0.08	0.97
25,000	0.03	0.17	1.25	0.14	9.17
22,500	0.03	0.47	1.00	0.30	18.77
20,000	0.04	2.04	0.75	0.41	64.07

Since the time sensibly increases only when the number of candidate outliers is very different from the desired one, the table confirms that by properly tuning the value of the radius the *UncertainDBOutlierDetector* algorithm is able to solve very efficiently the computationally heavy uncertain distance-based outlier detection problem.

6 Conclusions

In this work, a novel definition of uncertain outlier has been introduced to deal with multidimensional arbitrary shaped probability density functions and representing the generalization of the classic distance-based outlier definition.

Specifically, our approach corresponds to perform a nearest neighbor density estimate on all the possible outcomes of the dataset and, to the best of our knowledge, has no counterpart in the literature.

Moreover, it has been presented a method to efficiently compute the uncertain outliers, thus overcoming the difficulties raised by the introduced definition. Experiments have confirmed the effectiveness and the efficiency of the approach.

References

1. Lindley, D.: Understanding Uncertainty. Wiley-Interscience (2006)
2. Aggarwal, C., Yu, P.: A survey of uncertain data algorithms and applications. IEEE Trans. Knowl. Data Eng. **21**(5) (2009) 609–623
3. Mohri, M.: Learning from uncertain data. In: Proc. Conf. on Learning Theory (COLT). (2003) 656–670
4. Ngai, W., Kao, B., Chui, C., Cheng, R., Chau, M., Yip, K.: Efficient clustering of uncertain data. In: Proc. Int. Conf. on Data Mining (ICDM). (2006) 436–445
5. Kriegel, H.P., Pfeifle, M.: Density-based clustering of uncertain data. In: Proc. Int. Conf. on Knowledge Discovery and Data Mining (KDD). (2005) 672–677
6. Ren, J., Lee, S., Chen, X., Kao, B., Cheng, R., Cheung, D.: Naive bayes classification of uncertain data. Proc. Int. Conf. on Data Mining (ICDM) (2009) 944–949
7. Bi, J., Zhang, T.: Support vector classification with input data uncertainty. In: Proc. Conf. on Neural Information Processing Systems (NIPS). (2004) 161–168
8. Aggarwal, C., Yu, P.: Outlier detection with uncertain data. In: Proc. Int. Conf. on Data Mining (SDM). (2008) 483–493
9. Green, T., Tannen, V.: Models for incomplete and probabilistic information. IEEE Data Eng. Bull. **29**(1) (2006) 17–24
10. Hawkins, D.: Identification of Outliers. Monographs on Applied Probability and Statistics. Chapman & Hall (May 1980)
11. Knorr, E., Ng, R., Tucakov, V.: Distance-based outlier: algorithms and applications. VLDB Journal **8**(3-4) (2000) 237–253
12. Ramaswamy, S., Rastogi, R., Shim, K.: Efficient algorithms for mining outliers from large data sets. In: Proc. Int. Conf. on Management of Data (SIGMOD). (2000) 427–438
13. Angiulli, F., Pizzuti, C.: Outlier mining in large high-dimensional data sets. IEEE Trans. Knowl. Data Eng. **2**(17) (February 2005) 203–215
14. Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection: A survey. ACM Comput. Surv. **41**(3) (2009)
15. Barnett, V., Lewis, T.: Outliers in Statistical Data. John Wiley & Sons (1994)
16. Knorr, E., Ng, R.: Algorithms for mining distance-based outliers in large datasets. In: Proc. Int. Conf. on Very Large Databases (VLDB98). (1998) 392–403
17. Breunig, M.M., Kriegel, H., Ng, R., Sander, J.: Lof: Identifying density-based local outliers. In: Proc. Int. Conf. on Managment of Data (SIGMOD). (2000)
18. Jin, W., Tung, A., Han, J.: Mining top-n local outliers in large databases. In: Proc. ACM Int. Conf. on Knowledge Discovery and Data Mining (KDD). (2001)
19. Papadimitriou, S., Kitagawa, H., Gibbons, P., Faloutsos, C.: Loci: Fast outlier detection using the local correlation integral. In: Proc. Int. Conf. on Data Enginnering (ICDE). (2003) 315–326
20. Wang, B., Xiao, G., Yu, H., Yang, X.: Distance-based outlier detection on uncertain data. In: Proc. Computer and Information Technology (CIT). (2009) 293–298
21. Jiang, B., Pei, J.: Outlier detection on uncertain data: Objects, instances, and inference. In: Proc. Int. Conf. on Data Engineering (ICDE). (2011)
22. Lepage, G.: A new algorithm for adaptive multidimensional integration. Journal of Computational Physics **27** (1978)
23. A.M. Rushdi, A.A.Q.: Efficient computation of the p.m.f. and the c.d.f. of the generalized binomial distribution. Microeletron. Reliab. **34**(9) (1994) 1489–1499
24. Angiulli, F., Fassetti, F.: Dolphin: an efficient algorithm for mining distance-based outliers in very large datasets. ACM Trans. Knowl. Disc. Data **3(1)** (2009) Art. 4