Cross-Outlier Detection

Spiros Papadimitriou and Christos Faloutsos*

Computer Science Department,
Carnegie Mellon University,
5000 Forbes Ave, Pittsburgh, PA, USA
{spapadim,christos}@cs.cmu.edu

Abstract. The problem of outlier detection has been studied in the context of several domains and has received attention from the database research community. To the best of our knowledge, work up to date focuses exclusively on the problem as follows [1]: "given a *single* set of observations in some space, find those that deviate so as to arouse suspicion that they were generated by a different mechanism."
However, in several domains, we have more than one set of observations (or, equivalently, as single set with class labels assigned to each observation). For example, in astronomical data, labels may involve types of galaxies (e.g., spiral galaxies with abnormal concentration of elliptical galaxies in their neighborhood; in biodiversity data, labels may involve different population types, e.g., patches of different species populations, food types, diseases, etc). A single observation may look normal both within its own class, as well as within the entire set of observations. However, when examined with respect to other classes, it may still arouse suspicions.
In this paper we consider the problem "given a set of observations with class labels, find those that arouse suspicions, taking into account the class labels." This variant has significant practical importance. Many of the existing outlier detection approaches cannot be extended to this case. We present one practical approach for dealing with this problem and demonstrate its performance on real and synthetic datasets.

1 Introduction

In several problem domains (e.g., surveillance and auditing, stock market analysis, health monitoring systems, to mention a few), the problem of detecting rare

* This material is based upon work supported by the National Science Foundation under Grants No. IIS-9817496, IIS-9988876, IIS-0083148, IIS-0113089, IIS-0209107 IIS-0205224 by the Pennsylvania Infrastructure Technology Alliance (PITA) Grant No. 22-901-0001, and by the Defense Advanced Research Projects Agency under Contract No. N66001-00-1-8936. Additional funding was provided by donations from Intel. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation, DARPA, or other funding parties.

events, deviant objects, and exceptions is very important. Methods for finding such outliers in large data sets are drawing increasing attention [2,3,4,5,6,7,8,9,10,11].

As noted in [1], "the intuitive definition of an outlier would be 'an observation that deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism'." The traditional and—to the best of our knowledge—exclusive focus has been on the problem of detecting deviants in a single set of observations, i.e.,

Problem 1 (Outlier detection—single set). Given a set of objects, find these that deviate significantly from the rest.

However, there are several important practical situations where we have two collections of points. Consider the following illustrative example: Assume we have the locations of two types of objects, say vegetable patches and rabbit populations. If we consider, say, rabbit populations in isolation, these may be evenly distributed. The same may be true for food locations alone as well as for the union of the two sets.

Even though everything may look "normal" when we ignore object types, there is still the possibility of "suspicious" objects when we consider them in relation to objects of the other type. For example, a group of patches with far fewer rabbits present in the vicinity may indicate a measurement error. A population away from marked food locations may hint toward the presence of external, unaccounted-for factors.

The above may be considered a "toy" example that only serves illustrative purposes. Nonetheless, in several real-world situations, the spatial relationship among objects of two different types is of interest. A few examples:

- Situations similar to the one above actually do arise in biological/medical domains.
- In geographical/geopolitical applications, we may have points that represent populations, land and water features, regional boundaries, retail locations, police stations, crime incidence and so on. It is not difficult to think of situations where the correlations between such different objects are important.
- In astrophysics, it is well known that the distributions of different celestial objects follow certain laws (for example, elliptical and exponential galaxies form small clusters of one type and these clusters "repel" each other). There are *vast* collections of astrophysical measurements and even single deviant observations would potentially be of great interest.

In brief, we argue that the following outlier detection problem is of practical importance:

Problem 2 (Cross-outlier detection). Given two sets (or classes) of objects, find those which deviate with respect to the other set.

In this case we have a primary set $\mathbb{P}$ (e.g., elliptical galaxies) in which we want to discover *cross-outliers* with respect to a reference set $\mathbb{R}$ (e.g., spiral galaxies). Note that the single set case is always a special case, where $\mathbb{R} = \mathbb{P}$.

However, the converse is *not* true. That is, approaches for the single-set problem are not immediately extensible to cross-outlier detection. First off, several outlier definitions themselves cannot be extended (see also Section 5.1), let alone the corresponding methods to apply the definitions and compute the outliers. In summary, the contributions of this paper are two-fold:

- We identify the problem of cross-outlier detection. To the best of our knowledge, this has not been explicitly studied in the past, even though it is of significant practical interest. In general, an arbitrary method for the single-set problem cannot be easily extended to cross-outlier detection (but the opposite is true).
- We present a practical method that solves the problem. The main features of our method are:
  - It provides a meaningful answer to the question stated above, using a statistically intuitive criterion for outlier flagging (the local neighborhood size differs more than three standard deviations from the local averages), with no magic cut-offs.
  - Our definitions lend themselves to fast, single-pass estimation using box-counting. The running time of these methods is typically linear with respect to both dataset size and dimensionality.
  - It is an important first step (see also Section 5.3) toward the even more general problem of multiple-class cross-outliers (where the reference set $\mathbb{R}$ may be the union of more than one other class of objects).

The rest of the paper is organized as follows: Section 2 briefly discusses related work for the single class case, as well as more remotely related work on multiple dataset correlations and clustering. Section 3 presents our definition of a cross-outlier and briefly discusses its advantages. Section 4 demonstrates our approach on both synthetic and real datasets. Section 5 discusses some important issues and possible future directions. Finally, Section 6 gives the conclusions.

2 Background and related work

In this section we present prior work on the problem of single class outlier detection. To the best of our knowledge, the multiple class problem has not been explicitly considered.

2.1 Single dataset outlier detection

Previous methods for single dataset outlier detection broadly fall into the following categories.

Distribution based Methods in this category are typically found in statistics textbooks. They deploy some standard distribution model (e.g., normal) and flag as outliers those points which deviate from the model [4,1,12].

For arbitrary data sets without any prior knowledge of the distribution of points, we have to perform expensive tests to determine which model fits the data best, if any.

Clustering Many clustering algorithms detect outliers as by-products [13]. However, since the main objective is clustering, they are not optimized for outlier detection. Furthermore, the outlier-ness criteria are often implicit and cannot easily be inferred from the clustering procedures.

An intriguing clustering algorithm using the fractal dimension has been suggested by [14]; however it has not been demonstrated on real datasets.

Depth based This is based on computational geometry and finds different layers of k-d convex hulls [7]. Points in the outer layer are potentially flagged as outliers. However, these algorithms suffer from the dimensionality curse.

Distance based This was originally proposed by E.M. Knorr and R.T. Ng [8,9,10,11]. A point in a data set $\mathbb{P}$ is a *distance-based outlier* if at least a fraction β of the points in $\mathbb{P}$ are further than r from it.

This outlier definition is based on a single, global criterion determined by the parameters r and β and cannot cope with local density variations.

Density based This was proposed by M. Breunig, et al. [5]. It relies on the *local outlier factor* (LOF) of each point, which depends on the local density of its neighborhood. The neighborhood is defined by the distance to the *MinPts*-th nearest neighbor. In typical use, points with a high LOF are flagged as outliers.

This approach was proposed primarily to deal with the local density problems of the distance based method. However, selecting *MinPts* is non-trivial; in order to detect outlying clusters, *MinPts* has to be as large as the size of these clusters.

2.2 Multiple class outlier detection

To the best of our knowledge, this problem has not received explicit consideration to this date. Some single class approaches may be modified to deal with multiple classes, but the task is non-trivial. The general problem is open and provides promising future research directions. In this section we discuss more remotely related work.

Multi-dimensional correlations The problem of discovering general correlations between two datasets has been studied to some extent, both in the context of data mining, as well as for the purposes of selectivity estimation of spatial queries. However, none of these approaches deal with single points and identification of outlying observations.

[15] deals with the problem the general relationship of one multi-dimensional dataset with respect to another. This might be a good first step when exploring correlations between datasets. However, even when two datasets have been found to be correlated as a whole and to some extent co-located in space, this method cannot identify single outlying points.

Prior to that, [16] considers the problem of selectivity estimation of spatial joins across two point sets. Also, [17,18] consider the selectivity and performance of nearest neighbor queries within a single dataset.

Non-spatial clustering Scalable algorithms for extracting clusters from large collections of spatial data are presented in [19] and [20]. The authors also combine this with the extraction of characteristics based on non-spatial attributes by using both spatial dominant and non-spatial dominant approaches (depending on whether cluster discovery is performed first or on subsets derived using non-spatial attributes). It is not clear if these results can be extended to deal with the multiple class outlier detection problem. In the single class case, clusters of one or very few points can be immediately considered as outliers. However, this is not necessarily the case when dealing with multiple classes.

3 Proposed method

In this section we introduce our definition of an outlier and discuss its main properties. Our approach is based on the distribution of distances between points of the primary set and a reference set with respect to which we want to discover outliers. We use an intuitive, probabilistic criterion for automatic flagging of outliers.

3.1 Definitions

We consider the problem of detecting outlying observations from a primary set of points $\mathbb{P}$, with respect to a reference set of points $\mathbb{R}$. We want to discover points $p \in \mathbb{P}$ that "arouse suspicions" with respect to points $r \in \mathbb{R}$. Note that single-set outliers are a special case, where $\mathbb{R} = \mathbb{P}$.

Table 1 describes all symbols and basic definitions. To be more precise, for a point $p \in \mathbb{P}$ let $\hat{n}_{\mathbb{P},\mathbb{R}}(p, r, \alpha)$ be the average, over all points $q \in \mathbb{P}$ in the r-neighborhood of p, of $n_{\mathbb{R}}(q, \alpha r)$. The use of two radii serves to decouple the neighbor size radius αr from the radius r over which we are averaging.

We eventually need to *estimate* these quantities (see also Figure 1). We introduce the following two terms:

Definition 1 (Counting and sampling neighborhood). *The* counting neighborhood *(or αr-neighborhood) is the neighborhood of radius αr, over which each $n_{\mathbb{R}}(q, \alpha r)$ is estimated. The* sampling neighborhood *(or r-neighborhood) is the neighborhood of radius r, over which we collect samples of $n_{\mathbb{R}}(q, \alpha r)$ in order to estimate $\hat{n}_{\mathbb{P},\mathbb{R}}(p, r, \alpha)$. The* locality parameter *is α.*

The locality parameter α determines the relationship between the size of the sampling neighborhood and the counting neighborhood. We typically set this value to $\alpha = 1/2$ (see also Section 5.1).

Our outlier detection scheme relies on the standard deviation of the αr-neighbor count of points in the reference set $\mathbb{R}$. Therefore, we also define the quantity $\hat{\sigma}_{\mathbb{P},\mathbb{R}}(p, r, \alpha)$ to be precisely that, for each point $p \in \mathbb{P}$ and each sampling radius r.

Table 1. Symbols and definitions.

Symbol	Definition
$\mathbb{P}$ p_i	Primary set of points $\mathbb{P} = \{p_1, \ldots, p_i, \ldots, p_N\}$.
$\mathbb{R}$ r_i	Reference set of points $\mathbb{R} = \{r_1, \ldots, r_i, \ldots, r_M\}$.
N, M	Point set sizes.
k	Dimension of the data sets.
$d(p, q)$	Distance between points p and q.
$R_{\mathbb{P}}$, $R_{\mathbb{R}}$	Range (diameter) of each point set—e.g., $R_{\mathbb{P}} := \max_{p,q \in \mathbb{P}} d(p, q)$.
$\mathcal{N}_P(p, r)$	The set of r-neighbors of p from the point set P, i.e., $$\mathcal{N}(p, r) := \{q \in P \mid d(p, q) \leq r\}$$ Note that p does not necessarily belong to P.
$n_P(p, r)$	The number of r-neighbors of p_i from the set P, i.e., $n_P(p, r) := \lvert\mathcal{N}_P(p, r)\rvert$. Note that if $p \in P$, then $n_P(p, r)$ cannot be zero.
α	Locality parameter.
$\hat{n}_{\mathbb{P},\mathbb{R}}(p, r, \alpha)$	Average of $n_{\mathbb{R}}(p, \alpha r)$ over the set of r-neighbors of $p \in \mathbb{P}$, i.e., $$\hat{n}_{\mathbb{P},\mathbb{R}}(p, r, \alpha) := \frac{\sum_{q \in \mathcal{N}_{\mathbb{P}}(p,r)} n_{\mathbb{R}}(q, \alpha r)}{n_{\mathbb{P}}(p, r)}$$ For brevity, we often use $\hat{n}$ instead of $\hat{n}_{\mathbb{P},\mathbb{R}}$.
$\hat{\sigma}_{\mathbb{P},\mathbb{R}}(p, r, \alpha)$	Standard deviation of $n_{\mathbb{R}}(p, \alpha r)$ over the set of r-neighbors of $p \in \mathbb{P}$, i.e., $$\hat{\sigma}_{\mathbb{P},\mathbb{R}}(p, r, \alpha) := \sqrt{\frac{\sum_{q \in \mathcal{N}_{\mathbb{P}}(p,r)} \left(n_{\mathbb{R}}(q, \alpha r) - \hat{n}_{\mathbb{P},\mathbb{R}}(p, r, \alpha)\right)^2}{n_{\mathbb{P}}(p, r)}}$$ where $p \in \mathbb{P}$. For brevity we often use $\hat{\sigma}$ instead of $\hat{\sigma}_{\mathbb{P},\mathbb{R}}$.
k_σ	Determines what is *significant* deviation, i.e., a point $p \in \mathbb{P}$ is flagged as an outlier with respect to the set $\mathbb{R}$ iff $$\lvert\hat{n}_{\mathbb{P},\mathbb{R}}(p, r, \alpha) - n_{\mathbb{R}}(p, \alpha r)\rvert > k_\sigma \hat{\sigma}_{\mathbb{P},\mathbb{R}}(p, r, \alpha)$$ Typically, $k_\sigma = 3$.

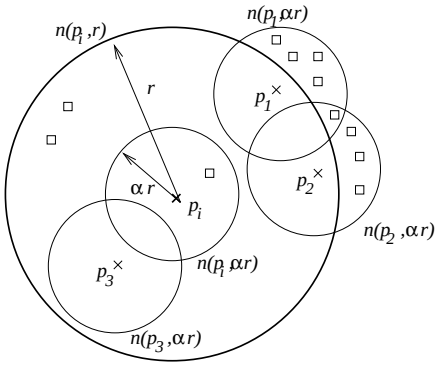

Fig. 1. Definitions for n and $\hat{n}$. Points in the primary set $\mathbb{P}$ are shown with "$\times$" and points in the reference set $\mathbb{R}$ with "$\square$". For instance, $n_{\mathbb{P}}(p_i, r) = 4$ (including itself), $n_{\mathbb{R}}(p_i, \alpha r) = 1$, $n_{\mathbb{R}}(p_1, \alpha r) = 6$ and $\hat{n}_{\mathbb{P},\mathbb{R}}(p_i, r, \alpha) = (1{+}5{+}4{+}0)/4 = 3.25$.

Definition 2 (Cross-outlier criterion). *A point $p \in \mathbb{P}$ is a* cross-outlier at scale (or radius) r *with respect to the reference set $\mathbb{R}$ if*

$$|\hat{n}_{\mathbb{P},\mathbb{R}}(p, r, \alpha) - n_{\mathbb{R}}(p, \alpha r)| > k_\sigma \hat{\sigma}_{\mathbb{P},\mathbb{R}}(p, r, \alpha)$$

Finally, the average and standard deviation with respect to radius r can provide very useful information about the vicinity of a point.

Definition 3 (Distribution plot). *For any point $p \in \mathbb{P}$, the plot of $n_{\mathbb{R}}(p, \alpha r)$ and $\hat{n}_{\mathbb{P},\mathbb{R}}(p, r, \alpha)$ with $\hat{n}_{\mathbb{P},\mathbb{R}}(p, r, \alpha) \pm 3\hat{\sigma}_{\mathbb{P},\mathbb{R}}(p, r, \alpha)$, versus r (for a range of radii of interest), is called its* (local) distribution plot.

3.2 Advantages of our definitions

Among several alternatives for an outlier score (such as $\max(\hat{n}/n, n/\hat{n})$, to give one example), our choice allows us to use probabilistic arguments for flagging outliers.

The above definitions and concepts make minimal assumptions. The only general requirement is that a distance is defined. Arbitrary distance functions are allowed, which may incorporate domain-specific, expert knowledge, if desired.

A final but very important point is that distance distributions can be quickly estimated in time that is linear with respect both to dataset sizes and dimensionality. Therefore, the above definitions lend themselves to fast, single-pass estimation algorithms, based on box-counting [21]. The only further constraint imposed in this case is that all points must belong to a k-dimensional vector space (either inherently, or after employing some embedding technique).

Table 2. Box-counting symbols and definitions.

Symbol	Definition
$\mathcal{C}(p, r, \alpha)$	Set of cells in some grid, with cell side $2\alpha r$, each fully contained within $\mathcal{L}^\infty$-distance r from point p.
C_i	Cell in some grid.
$c_{P,i}$	The count of points *from set* P within the corresponding cell C_i.
$S_P^q(p, r, \alpha)$	Sum of box counts (from set P) to the q-th power, i.e., $$S_P^q(p, r, \alpha) := \sum_{C_i \in \mathcal{C}(p,r,\alpha)} c_{P,i}^q$$
$P_{P,R}^q(p, r, \alpha)$	Sum of box count products (from sets P and R); in particular, $$P_{P,R}^q(p, r, \alpha) := \sum_{C_i \in \mathcal{C}(p,r,\alpha)} c_{P,i} c_{R,i}^q$$ Note that, $S_P^q = P_{P,P}^{q-1}$.

The main idea is to approximate the r-neighbor counts for each point p with pre-computed counts of points within a cell[1] of side r which contains p.

In a little more detail, in order to quickly estimate $\hat{n}(p, r, \alpha)$ for a point $p_i \in \mathbb{P}$ (from now on, we assume $\mathcal{L}^\infty$ distances), we can use the following approach. Consider a grid of cells with side $2\alpha r$ over both sets $\mathbb{P}$ and $\mathbb{R}$. Within each cell, we store separate counts of points it contains from $\mathbb{P}$ and $\mathbb{R}$. Perform a *box count* on the grid: For each cell C_j in the grid, find the counts, $c_{\mathbb{R},j}$ and $c_{\mathbb{P},j}$, of the number of points from $\mathbb{R}$ and $\mathbb{P}$, respectively, in the cell. There is a total number of $c_{\mathbb{P},j}$ points $p \in \mathbb{P} \cap C_j$ (counting p itself), each of which has $c_{\mathbb{P},j}$ neighbors from $\mathbb{R}$. So, the total number of $\mathbb{R}$ neighbors over all points from $\mathbb{P}$ in C_j is $c_{\mathbb{P},j} c_{\mathbb{R},j}$. Denote by $\mathcal{C}(p, r, \alpha)$ the set of all cells in the grid such that the entire cell is within distance r of p_i. We use $\mathcal{C}(p, r, \alpha)$ as an approximation for the r-neighborhood of p_i. Summing over all these cells, we get a total number of $\mathbb{P}$-$\mathbb{R}$ pairs of $P_{\mathbb{P},\mathbb{R}}(p, r, \alpha) := \sum_{C_j \in \mathcal{C}(p,r,\alpha)} c_{\mathbb{P},j} c_{\mathbb{R},j}$. The total number of objects is simply the sum of all box counts for points in $\mathbb{P}$, i.e., $S_{\mathbb{P}}^1(p, r, \alpha)$

$$\hat{n}_{\mathbb{P},\mathbb{R}}(p, r, \alpha) = \frac{P_{\mathbb{P},\mathbb{R}}^1(p, r, \alpha)}{S_{\mathbb{P}}^1(p, r, \alpha)}$$

[1] In practice, we have to use multiple cells in a number randomly shifted grids and use some selection or voting scheme to get a good approximation; see [21] for more details.

A similar calculation can be done to estimate

$$\hat{\sigma}_{\mathbb{P},\mathbb{R}}(p,r,\alpha) = \sqrt{\frac{P^2_{\mathbb{P},\mathbb{R}}(p,r,\alpha)}{S^1_{\mathbb{P}}(p,r,\alpha)} - \left(\frac{P^1_{\mathbb{P},\mathbb{R}}(p,r,\alpha)}{S^1_{\mathbb{P}}(p,r,\alpha)}\right)^2}$$

4 Experimental results

In this section we give examples of our method and discuss some important observations related to our approach, as well as the problem in general.

Gap In this case (see Figure 2, top row) the primary set consists of 340 points with a uniformly random distribution within a square region. In single-set outlier detection ($\mathbb{R} = \mathbb{P}$) some fringe points are flagged with a positive deviation (i.e., at some scale, their neighbor count is below the local average). Also, a few interior points in locally dense regions are flagged with a negative deviation.

In cross-outlier detection, we use a reference set $\mathbb{R}$ of 1400 points, again uniformly random in a slightly larger square region, but with a central square gap. As expected, the points of $\mathbb{P}$ that fall within well within the gap of $\mathbb{R}$ are detected as cross-outliers with a positive deviation. Also, very few[2] other points are flagged.

Core In this case (see Figure 2, middle row), the primary set again consists of 300 points with a uniformly random distribution within a square region. The single-set outliers are similar to the previous case.

In cross-outlier detection, we use a reference set $\mathbb{R}$ of 250 points uniformly random within a central square "core." As expected again, the points of $\mathbb{P}$ that fall within the reference "core" are all detected as outliers. Also, some fringe points are still detected as outliers (see Section 4.1).

Lines The primary set $\mathbb{P}$ consists of 100 points regularly spaced along a line (Figure 2, bottom row). The single-set outliers ($\mathbb{P} = \mathbb{R}$) consist of eight points, four at each end of the line. Indeed, these points are "special," since their distribution of neighbors clearly differs from that of points in the middle of the line.

In cross outlier detection, the reference set $\mathbb{R}$ consists of two lines of 100 points each, both parallel to $\mathbb{P}$ and slightly shifted downward along their common direction. As expected, the points at the bottom-left end of $\mathbb{P}$ are no longer outliers, with respect to $\mathbb{P}$. Note that the *same* four points along the top-right end are flagged (see discussion in Section 4.1).

[2] Since $\mathbb{R}$ is significantly denser than $\mathbb{P}$, this is expected.

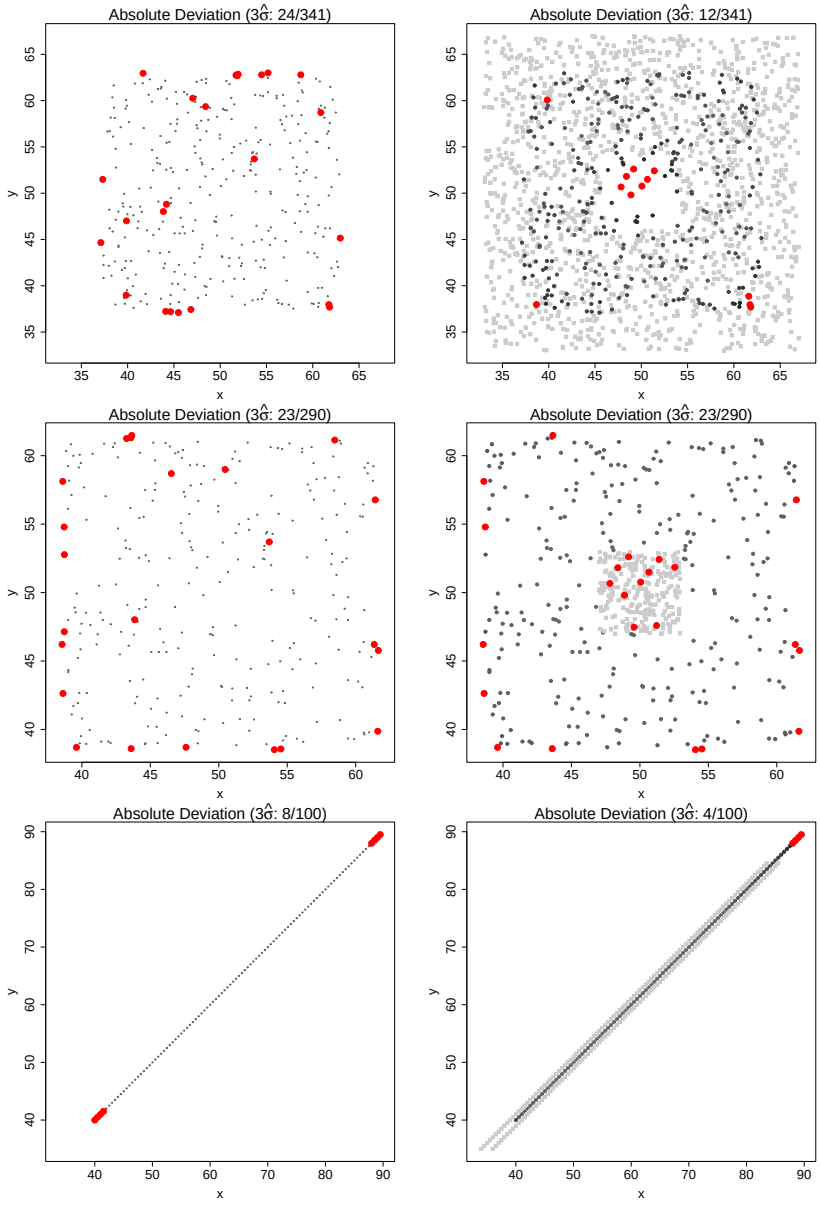

Fig. 2. "Plain" outliers (left, $\mathbb{R} = \mathbb{P}$) and cross-outliers (right). The reference set is shown with square, gray points in the right column. Outliers are marked with larger, red points in each case. In all cases, $\alpha = 1/4$.

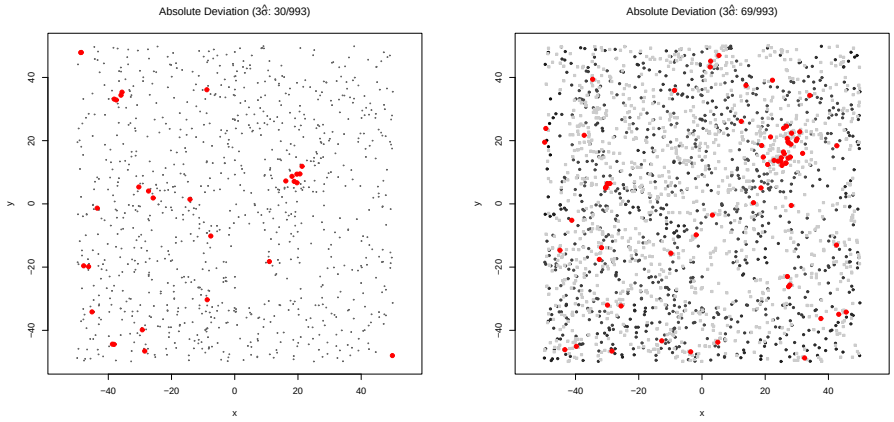

Fig. 3. "Plain" outliers (left, $\mathbb{R} = \mathbb{P}$) and cross-outliers (right) for the galaxy datasets. In all cases, $\alpha = 1/4$.

Galaxy The primary set consists of a section with 993 spiral galaxies and the reference set of a section with 1218 elliptical galaxies, both from the Sloan Digital Sky Survey (Figure 3). Although not shown in the figure, all cross-outliers are flagged with a *negative* deviation (except two at the very edge of the dataset). Also (see Figure 4 and Section 4.1) all are flagged by a narrow margin. This is indeed expected: elliptical galaxies form clusters, intertwined with clusters of spiral galaxies. The distribution is overall even (as evidenced by the consistently wide standard deviation band); however, a few of the elliptical galaxies are within unusually dense clusters of spiral galaxies.

4.1 Observations

Fringe points The points located along the fringes of a data set are clearly different from the rest of the points.

One could argue that outlier definitions such as the one of the depth-based approach [7] rely *primarily* on this observation in order to detect outliers. Our method goes beyond that and can also capture isolated central points (as can be seen, for example, from the `Gap` example), but can still distinguish fringe points.

With respect to pairwise distances upon which our approach is based, the first observation is that fringe points have fewer neighbors than interior points. More than that, however, all neighbors of fringe points lie on the *same* half-plane. It is a consequence of this *second* fact that the standard deviation of neighbor counts is (comparatively) smaller at certain scales for fringe points.

This explains why in the `Core` example more fringe points are detected as cross-outliers than in `Gap`. The reference set in `Gap` is chosen to cover a slightly larger region than the primary set in order to illustrate this point. The fringe points of $\mathbb{P}$ in `Gap` are not fringe points *with respect to* $\mathbb{R}$: they have $\mathbb{R}$-neighbors

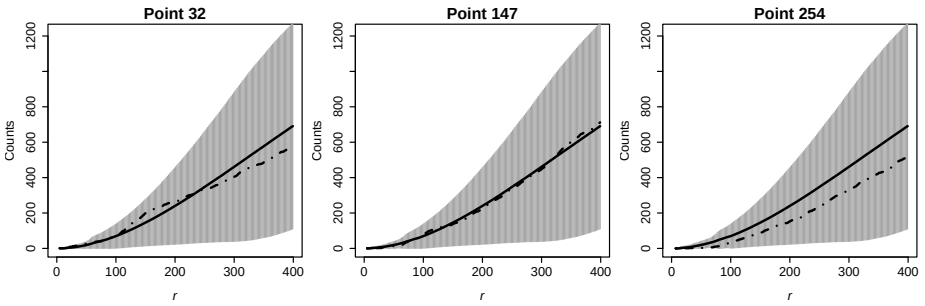

Fig. 4. Distribution plot for cross-outliers in `Galaxy`. The horizontal axis is scale (or, sampling radius r). The solid line is $\hat{n}_{\mathbb{P},\mathbb{R}}(p, r, \alpha)$ and the dashed line is $n_{\mathbb{R}}(p, \alpha)$. The gray bands span $\pm 3\hat{\sigma}_{\mathbb{P},\mathbb{R}}(p, r, \alpha)$ around the average. The galaxy on the right is flagged with positive deviation, the other two with negative. All are flagged at small scales by a narrow margin.

on all sides of the plane. However, the fringe points of $\mathbb{P}$ in `Core` have $\mathbb{R}$-neighbors only on one half-plane. Thus, the fringe points of $\mathbb{P}$ in `Core` are indeed different than the interior points (always *with respect to* $\mathbb{R}$).

Role of each distribution In this paragraph we further discuss the sampling and counting neighborhoods. In particular, the former contains points of the primary set $\mathbb{P}$, while the latter of the reference set $\mathbb{R}$. Thus, the distribution of points in *both* sets plays an important role in cross-outlier detection (but see also Section 5.1).

This explains the fact that in `Lines` the same four endpoints are flagged as cross-outliers. We argue that this is a desirable feature. First, the points near the top-right end that are closer to $\mathbb{R}$ are indeed less "distinct" than their neighbors at the very end. This fact depends on the distribution of $\mathbb{P}$, not $\mathbb{R}$! Furthermore, consider extending $\mathbb{P}$ toward the top-right: then, neither of the endpoints are suspicious (whether surrounded or not by points of $\mathbb{R}$). This, again, depends on the distribution of $\mathbb{P}$! Indeed, in the latter case, our method does not detect any outliers.

Digging deeper As hinted in the discussion of the results, the sign of the deviation can give us important information. However, we can go even further and examine the *distribution plots*, which we discuss very briefly here. Figure 5 is included as an example. We can clearly see that a point within the gap belongs to a sparse region (with respect to $\mathbb{R}$). Moreover, we can clearly see that the point within the gap is flagged by a much wider margin and at a wider range of scales, whereas a fringe point is marginally flagged. Thus, the distribution plots provide important information about *why* each point is an outlier, as well as its vicinity.

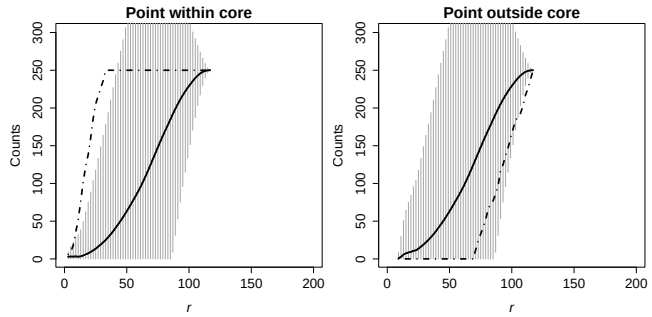

Fig. 5. Distribution plot for cross-outliers in `Core`. Again, the horizontal axis is scale (or, sampling radius r). The solid line is $\hat{n}_{\mathbb{P},\mathbb{R}}(p, r, \alpha)$ and the dashed line is $n_{\mathbb{R}}(p, \alpha)$. The gray bands span $\pm 3\hat{\sigma}_{\mathbb{P},\mathbb{R}}(p, r, \alpha)$ around the average.

5 Discussion

In this section we first discuss why the problem of cross-outlier detection is different from the single-set case, even though the two may, at first, seem almost identical. We also discuss some directions for future research. These relate to the fast, single-pass estimation algorithms that our definitions admit.

5.1 Differences to single class outlier detection

The intuitive definition of [1] implies two important parts in any definition of an outlier: *what* is considered a deviation (i.e., where or how we look for them) and *how* do we determine *significant* deviations. Therefore, all outlier definitions employ a model for the data and a measure of correlation, either explicitly or implicitly.

The first difference in the case of cross-outliers follows directly from the problem definition. What we essentially estimate is not a single probability distribution or correlation, but either some (conditional) probability with respect to the reference set or the covariance among sets. However, several of the existing definitions do not make their model assumptions clear or employ a model that cannot be easily extended as described above. These outlier detection approaches are hard to modify.

It should be noted that our definition employs a very general and intuitive model which is based on pairwise distanced and makes minimal assumptions.

The second major difference again follows from the fact that we are dealing with two *separate* sets. Simply put, in the "classical" case ($\mathbb{R} = \mathbb{P}$), we can obviously assume that a point set is co-located in space with respect to itself. However, this need not be the case when $\mathbb{R} \neq \mathbb{P}$. This assumption is sometimes implicitly employed in outlier definitions.

Tools such as that of [15] are useful here as a first step to determine the *overall* spatial relationship between the two sets. It must further be noted that, in our approach, the locality parameter α is tunable and typically two values should be sufficient: $\alpha \approx 1/2$ (or 1) and any $\alpha \leq r_{min} / \max\{R_{\mathbb{P}}, R_{\mathbb{R}}\}$ where r_{min} is the smallest distance between any two points (irrespective of type)[3].

5.2 Efficiency considerations

Our definitions are based on pairwise distance distributions. As demonstrated in [15,21], these can be estimated very quickly with a single pass over the data, in time that is practically linear with respect to both data set size and dimensionality. The only minor restriction imposed by these algorithms is that $\alpha = 1/2^k$ for some integer k.

Furthermore, if we have more than two classes of points, the pre-processing step for box counting can be modified to keep separate counts for each class. This does not increase computational cost (only space in proportion to the number of classes) and allows fast outlier detection where the reference set $\mathbb{R}$ is the *union* of points from several classes (rather than a single class).

5.3 Generalizations

The observation in the last paragraph of the previous section naturally leads to the problem of multi-class outlier detection. As pointed out, the fast algorithms can easily detect outliers when the reference set $\mathbb{R}$ is any *given* combination of classes, without incurring any extra computational cost.

An interesting future research direction is to extend these algorithms with heuristic pruning approaches (e.g., similar to those in association rule[4] algorithms [22]; in our case, items correspond to point classes) to efficiently search the entire space of all class combinations (i.e., pointset unions) in the place of $\mathbb{R}$.

6 Conclusions

In this paper we present the problem of *cross-outlier* detection. This is the first contribution; we argue that this is a non-trivial problem of practical interest and certainly more than an immediate generalization. We discuss several aspects of the problem that make it different from "classical" outlier detection. The former is a special case of cross-outliers (with $\mathbb{R} = \mathbb{P}$) but the converse is not true.

Beyond introducing the problem, we present a method that can provide an answer. Furthermore, our definitions use a statistically intuitive flagging criterion and lend themselves to fast, single-pass estimation. We demonstrate our approach using both synthetic and real datasets.

[3] The second choice for α formally implies that, at every scale, the sampling neighborhood completely covers both datasets.

[4] This is one potential approach; regions with *no* co-located classes can probably be ignored. Of course, this far from exhausts all possible pruning techniques.

References

1. Hawkins, D.: Identification of Outliers. Chapman and Hall (1980)
2. Aggarwal, C., Yu, P.: Outlier detection for high dimensional data. In: Proc. SIGMOD. (2001)
3. Arning, A., Agrawal, R., Raghavan, P.: A linear method for deviation detection in large database. In: Proc. KDD. (1996) 164–169
4. Barnett, V., Lewis, T.: Outliers in Statistical Data. John Wiley (1994)
5. Breunig, M., Kriegel, H., Ng, R., Sander, J.: Lof: Identifying density-based local outliers. In: Proc. SIGMOD Conf. (2000) 93–104
6. Jagadish, H., Koudas, N., Muthukrishnan, S.: Mining deviants in a time series database. In: Proc. VLDB. (1999) 102–113
7. Johnson, T., Kwok, I., Ng, R.: Fast computation of 2-dimensional depth contours. In: Proc. KDD. (1998) 224–228
8. Knorr, E., Ng, R.: A unified notion of outliers: Properties and computation. In: Proc. KDD. (1997) 219–222
9. Knorr, E.M., Ng, R.: Algorithms for mining distance-based outliers in large datasets. In: Proc. VLDB 1998. (1998) 392–403
10. Knorr, E., Ng, R.: Finding intentional knowledge of distance-based outliers. In: Proc. VLDB. (1999) 211–222
11. Knorr, E., Ng, R., Tucakov, V.: Distance-based outliers: Algorithms and applications. VLDB Journal **8** (2000) 237–253
12. Rousseeuw, P., Leroy, A.: Robust Regression and Outlier Detection. John Wiley and Sons (1987)
13. Jain, A., Murty, M., Flynn, P.: Data clustering: A review. ACM Comp. Surveys **31** (1999) 264–323
14. Barbará, D., Chen, P.: Using the fractal dimension to cluster datasets. In: Proc. KDD. (2000) 260–264
15. Traina, A., Traina, C., Papadimitriou, S., Faloutsos, C.: Tri-Plots: Scalable tools for multidimensional data mining. In: Proc. KDD. (2001) 184–193
16. Faloutsos, C., Seeger, B., Jr., C.T., Traina, A.: Spatial join selectivity using power laws. In: Proc. SIGMOD. (2000) 177–188
17. Belussi, A., Faloutsos, C.: Estimating the selectivity of spatial queries using the 'correlation' fractal dimension. In: Proc. VLDB. (1995) 299–310
18. Berchtold, S., Böhm, C., Keim, D., Kriegel, H.P.: A cost model for nearest neighbor search in high-dimensional data space. In: Proc. PODS. (1997) 78–86
19. Ng, R., Han, J.: Efficient and effective clustering methods for spatial data mining. In: Proc. VLDB. (1994) 144–155
20. Knorr, E., Ng, R.: Finding aggregate proximity relationships and commonalities in spatial data mining. IEEE TKDE **8** (1996) 884–897
21. Papadimitriou, S., Kitagawa, H., Gibbons, P., Faloutsos, C.: LOCI: Fast outlier detection using the local correlation integral. In: Proc. ICDE. (2003)
22. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules in large databases. In: Proc. VLDB. (1994) 487–499