A Unified Notion of Outliers: Properties and Computation

Edwin M. Knorr and **Raymond T. Ng**
Department of Computer Science
University of British Columbia
Vancouver, B.C., V6T 1Z4, Canada
rng@cs.ubc.ca

Abstract

As said in signal processing, "One person's noise is another person's signal." For many applications, such as the exploration of satellite or medical images, and the monitoring of criminal activities in electronic commerce, identifying exceptions can often lead to the discovery of truly unexpected knowledge. In this paper, we study an intuitive notion of outliers. A key contribution of this paper is to show how the proposed notion of outliers unifies or generalizes many existing notions of outliers provided by discordancy tests for standard statistical distributions. Thus, a unified outlier detection system can replace a whole spectrum of statistical discordancy tests with a single module detecting only the kinds of outliers proposed. A second contribution of this paper is the development of an approach to find all outliers in a dataset. The structure underlying this approach resembles a data cube, which has the advantage of facilitating integration with the many OLAP and data mining systems using data cubes.

Introduction[1]

It has been widely recognized that knowledge discovery tasks can be classified into 4 main categories: (a) dependency detection, (b) class identification, (c) class description, and (d) exception/outlier detection. The first 3 categories of tasks correspond to patterns that apply to many objects, or to a large percentage of objects, in the dataset. In contrast, the 4th category focuses on a very small minority of data objects, so small that these objects are often discarded as noise. For many applications, identifying exceptions can often lead to the discovery of truly unexpected knowledge. Some existing algorithms in machine learning and data mining have considered outliers, but only to the extent of tolerating outliers in whatever the algorithms are supposed to do (Ester *et al.* 1996; Ng & Han 1994; Zhang, Ramakrishnan, & Livny 1996).

Almost all studies that consider outlier identification as their primary objective are in statistics. Barnett and Lewis provide a comprehensive treatment, listing about 100 discordancy tests for normal, exponential, Poisson, and binomial distributions (Barnett & Lewis 1994). The choices of appropriate discordancy tests depend on: (a) the distribution, (b) whether or not the distribution parameters (e.g., mean and variance) are known, (c) the number of expected outliers, and (d) even the types of expected outliers (e.g., upper or lower outliers, in an ordered sample). For example, for a normal distribution with known mean and known variance, there are separate discordancy tests for: single upper outliers, upper outlier pairs, k upper outliers, single lower outliers, lower outlier pairs, k lower outliers, upper and lower outlier pairs, etc. There are other tests if the mean is not known, or if the variance is not known. Furthermore, there are separate tests for other distributions. Yet, despite all of these options and decisions, there is no guarantee of finding outliers, either because there may not be any test developed for a specific combination, or because no standard distribution can adequately model the observed distribution. In a data mining context, the distributions of the values of the attributes are almost always unknown. To fit the observed distributions into standard distributions, and to choose suitable tests, requires non-trivial computational effort for large datasets. This motivates our study of a unified notion of outliers, defined as follows:

An object O in a dataset T is a $UO(p, D)$-outlier if at least fraction p of the objects in T are $\geq$ distance D from O.

This notion is intuitive because it captures the general spirit of an outlier. Hawkins eloquently describes an outlier as "an observation that deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism" (Hawkins

1980). As such, this notion is a natural candidate for situations where the observed distribution does not fit any standard distribution, or where no discordancy test has been developed. In this paper, we show that for many discordancy tests, if an object O is an outlier according to a specific discordancy test, then O is also a $UO(p, D)$-outlier for some suitably defined p and D.

Data mining, by definition, implies large quantities of data; therefore, efficiency or scalability is an important goal. A key contribution of this paper is the development of a partitioning-based approach for detecting $UO(p, D)$-outliers. Its efficiency, particularly for datasets with low dimensionality, will be demonstrated by the performance results presented at the end of this paper. Another advantage of our approach is that its underlying data structure resembles a data cube (Gray *et al.* 1995). This makes our method easily integratable with the many OLAP systems and data mining systems using data cubes.

Properties of $UO(p, D)$-Outliers: Relationships with Existing Notions

In this section, we show how our notion of $UO(p, D)$-outliers relates to existing notions of outliers, most of which turn out to be specific instances of ours.

Definition 1 We say that $UO(p, D)$ *unifies* or *generalizes* another definition *Def* for outliers, if there exist specific values p_0, D_0 such that object O is an outlier according to *Def* iff O is a $UO(p_0, D_0)$-outlier. □

Let N be the number of objects in the input dataset T. Each object is identified with the same k attributes. k is called the dimensionality of the dataset. Suppose there is an underlying distance metric function F that gives the distance between any pair of objects in T. For an object O, the D-neighbourhood of O contains the set of objects $Q \in T$ that are within distance D from O, i.e., $\{Q \in T \mid F(O, Q) \leq D\}$. The fraction p is the minimum fraction of objects in T that must be *outside* the D-neighbourhood of an outlier.

Outliers in Normal Distributions

For a normal distribution, outliers can be considered to be points that lie 3 or more standard deviations (i.e., $\geq 3\sigma$) from the mean μ (Freedman, Pisani, & Purves 1978).

Definition 2 Let T be a set of values that is truly normally distributed with mean μ and standard deviation σ. Define Def_{Normal} as follows: $t \in T$ is an outlier iff $\frac{t-\mu}{\sigma} \geq 3$ or $\frac{t-\mu}{\sigma} \leq -3$. □

Lemma 1 $UO(p, D)$ unifies Def_{Normal} with $p_0 = 0.9988, D_0 = 0.13\sigma$, i.e., t is an outlier according to Def_{Normal} iff t is a $UO(0.9988, 0.13\sigma)$-outlier.

Proof Outline: We use probabilities to reflect the number of points lying in a D-neighbourhood. Specifically, the probability is $1-p$ that the distance between 2 randomly selected points, O and Q, is less than or equal to D. Let T_1 and T_2 be random variables that are normally distributed with parameters μ and σ^2 (i.e., $T_1, T_2 \sim N(\mu, \sigma^2)$). Define $Z_1 = \frac{T_1-\mu}{\sigma}$ and $Z_2 = \frac{T_2-\mu}{\sigma}$ as standard normal variables (i.e., $Z_1, Z_2 \sim N(0, 1)$). Using a table that lists areas for ranges $-\infty \leq Z_i \leq z$ under the standard normal curve to 4 decimal places, and given $p_0 = 0.9988$ and $D_0 = 0.13\sigma$, we get:

$$
\begin{array}{lcl}
 & Pr(|T_1 - T_2| \geq 0.13\sigma) & \geq 0.9988 \\
\iff & Pr(|T_1 - T_2| \leq 0.13\sigma) & \leq 0.0012 \\
\iff & Pr(T_1 - 0.13\sigma \leq T_2 \leq T_1 + 0.13\sigma) & \leq 0.0012 \\
\iff & Pr(Z_1 - 0.13 \leq Z_2 \leq Z_1 + 0.13) & \leq 0.0012 \\
\iff & Z_1 \leq -3.0000 \quad \text{or} \quad Z_1 \geq 3.0000 & \\
\iff & T_1 \leq \mu - 3\sigma \quad \text{or} \quad T_1 \geq \mu + 3\sigma & \square
\end{array}
$$

Note that if the value 3σ in Def_{Normal} is changed to some other value, such as 4σ, the above proof procedure can easily be modified with the corresponding p_0 and D_0 to show that $UO(p, D)$ still unifies the new definition of Def_{Normal}.

Barnett and Lewis give many specialized tests for identifying outliers in normal distributions with known or unknown means and/or standard deviations. Instead of using the standard normal curve, some of those tests use a t-distribution curve, which has a fatter tail than the normal curve. Nevertheless, the same proof procedure shown above still works with appropriate modifications. As a concrete example, Barnett and Lewis give Def_{N6} for testing for an upper and lower outlier pair in a normal distribution with unknown mean and variance. In a more detailed report (Knorr & Ng 1997), we give a proof showing that $UO(p, D)$ unifies Def_{N6} with $p_0 = 0.999$ and $D_0 = 0.2$. Due to very limited space, we omit those details. Details and proofs are also available for the following lemmas.

Outliers in Other Distributions

Consider the following test for finding a single upper outlier in an exponential sample with unknown parameter (Barnett & Lewis 1994). Let the sampled values be ordered as $x_{(1)}, \ldots, x_{(n)}$, and let the test statistic T_1 be defined as $\frac{x_{(n)}}{(\sum_{i=1}^{n} x_{(i)})}$. Let τ be the observed value for T_1, and $SP(\tau)$ be the significance probability that T_1 takes values more discordant than τ.

Barnett and Lewis provide an example using 131 excess cycle times in steel manufacture, whose distribution reasonably approximates an exponential distribution with parameter 0.14. Let us consider this example.

Definition 3 Define Def_{Exp} as follows. Observation t is an outlier (according to test statistic T_1 defined above) iff $SP(\tau) \leq 0.01$. For example, for the 131 excess cycle times, $x_{(131)} = 92$ is an outlier because $SP(\tau) \leq 0.00017$. □

Lemma 2 For an exponential distribution with parameter 0.14, $UO(p, D)$ unifies Def_{Exp} with $p_0 = 0.999$ and $D_0 = 0.0001$. □

For a binomial distribution, t can be defined as an outlier iff τ is greater than or equal to a specific critical value (listed in tables of such values (Barnett & Lewis 1994)). Again, a proof is available showing that $UO(p, D)$ unifies this definition of outliers in a binomial distribution.

The following test can be used to find outliers in a Poisson distribution with parameter $\mu = 3.0$.

Definition 4 Define $Def_{Poisson}$ as follows: t is an outlier iff $t \geq 8$. □

Lemma 3 $UO(p, D)$ unifies $Def_{Poisson}$ with $p_0 = 0.9962$ and $D_0 = 1$. □

Comparison with Sequential Exceptions and with Data Clustering Algorithms

All of the outlier identification studies that we have come across are in statistics expect for the "sequential exceptions" approach (Arning, Agrawal, & Raghavan 1996), whereby a dataset is searched for implicit redundancies. Specifically, data items called *sequential exceptions* are extracted, which maximize the reduction in Kolmogorov complexity. Unlike $UO(p, D)$-outlier detection (and all statistical discordancy tests, and most distance-based data mining works), the approach of Arning, et al. does not require a metric distance function. Thus, because of model differences, their approach and ours will not necessarily identify the same outliers, even for a dataset that is applicable to both approaches.

Data clustering algorithms assign similar objects in a dataset to the same classes; however, they provide little support for identifying outliers (Kaufman & Rousseeuw 1990; Fisher 1987). CLARANS, BIRCH, and DBSCAN are clustering algorithms designed specifically for data mining applications. In CLARANS (Ng & Han 1994), an object is (removed as) "noise" if its removal raises the silhouette coefficient of the clusters. In BIRCH (Zhang, Ramakrishnan, & Livny 1996), an object is (removed as) noise if it is "too far from its closest seed," where a seed is some representative object such as the centroid of a cluster. DBSCAN classifies objects as core, border, or outlying, based on the reachability and connectivity of the object being clustered (Ester *et al.* 1996). A key here is that DBSCAN, as a clustering algorithm, aims to produce maximal-size clusters and is reluctant to label objects as outliers. Neither DBSCAN nor CLARANS nor BIRCH is designed to unify the kinds of distribution-dependent discordancy tests described earlier.

An Approach for Finding All $UO(p, D)$-Outliers

Although limited space does not permit the inclusion and complexity analysis of our algorithm for finding all $UO(p, D)$-outliers, we present the underlying properties and the general approach for computing those outliers, followed by preliminary experimental results.

Underlying Cell Structure and Properties

A naive algorithm for detecting all $UO(p, D)$-outliers for given values of p and D would be to count, for each object Q, the number of objects in the D-neighbourhood of Q. A more optimized algorithm is to build and search a spatial indexing structure, such as a k-d tree (Bentley 1975). Instead, our approach for finding all $UO(p, D)$-outliers relies on an underlying cell structure. As will become obvious later, the idea is to reduce object-by-object processing to cell-by-cell processing, thereby gaining efficiency. For ease of presentation, we show the cell structure and its properties for the 2-dimensional case, i.e., $k = 2$. Later, we describe what changes are required to generalize to higher dimensions.

The 2-dimensional space spanned by the data objects is partitioned into cells or squares of length $l = \frac{D}{2\sqrt{2}}$. Let $C_{x,y}$ denote the cell that is at the intersection of row x and column y. The Layer-1 (L_1) neighbours of $C_{x,y}$ are all the immediate neighbouring cells of $C_{x,y}$ as defined in the usual sense, i.e.,

$$L_1(C_{x,y}) = \{C_{u,v} \mid u = x \pm 1,\ v = y \pm 1,\ C_{u,v} \neq C_{x,y}\}. \quad (1)$$

In the 2-dimensional case, a typical cell (except for cells that are on the boundary of the cell structure) has 8 L_1 neighbours.

Property 1 Any pair of objects within the same cell is at most distance $\frac{D}{2}$ apart. □

Property 2 If $C_{u,v}$ is a L_1 neighbour of $C_{x,y}$, then any object $P \in C_{u,v}$ and any object $Q \in C_{x,y}$ are at most distance D apart. □

Property 1 is valid because the length of a diagonal of a cell is $\sqrt{2}l = \sqrt{2}\frac{D}{2\sqrt{2}} = \frac{D}{2}$. Similarly, Property 2 is true because the distance between any pair of objects in the two cells cannot exceed twice the length of a diagonal of a cell. As will become obvious later, these two properties are useful in ruling out many objects as outlier candidates. The Layer-2 (L_2) neighbours of $C_{x,y}$ are all the cells within 3 cells of $C_{x,y}$, i.e.,

$$L_2(C_{x,y}) = \{C_{u,v} \mid u = x \pm 3,\ v = y \pm 3,\\ C_{u,v} \notin L_1(C_{x,y}),\ C_{u,v} \neq C_{x,y}\}. \quad (2)$$

While Layer 1 is 1 cell thick, Layer 2 is 2 cells thick. This is significant because of the following property.

Property 3 If $C_{u,v} \neq C_{x,y}$ is neither an L_1 nor an L_2 neighbour of $C_{x,y}$, then any object $P \in C_{u,v}$ and any object $Q \in C_{x,y}$ must be at least distance D apart. □

Because the combined thickness of L_1 plus L_2 is 3 cells, the distance between P and Q must exceed $3l = \frac{3D}{2\sqrt{2}} > D$. In the 2-dimensional case, a typical cell, except for boundary cells, has $7^2 - 3^2 = 40$ L_2 cells.

The key idea of our approach is summarized in the following property. For convenience, let M denote the maximum number of objects that can be *inside* the D-neighbourhood of an outlier, i.e., $M = N(1-p)$, and let m be the total number of cells.

Property 4 **(a)** If there are $> M$ objects in $C_{x,y}$, *none* of the objects in $C_{x,y}$ is an outlier. **(b)** If there are $> M$ objects in $C_{x,y} \cup L_1(C_{x,y})$, *none* of the objects in $C_{x,y}$ is an outlier. **(c)** If there are $\leq M$ objects in $C_{x,y} \cup L_1(C_{x,y}) \cup L_2(C_{x,y})$, *every* object in $C_{x,y}$ is an outlier. □

Properties 4(a) and 4(b) are direct consequences of Properties 1 and 2, and 4(c) is due to Property 3. Note that these properties help to identify outliers or non-outliers in a cell-by-cell manner, rather than on an object-by-object basis. This kind of "bulk processing" reduces execution time significantly.

Generalization to Higher Dimensions

To generalize from the 2-dimensional case to $k > 2$, we use the same algorithm; however, to maintain Properties 1-4, (i) the length of a cell changes from $l = \frac{D}{2\sqrt{2}}$ to $l = \frac{D}{2\sqrt{k}}$, and (ii) Layer 2 is no longer 2 cells thick, but rather $\lceil 2\sqrt{k} \rceil$ cells thick. We defer proofs and complexity analysis to a more detailed report (Knorr & Ng 1997). In ongoing work, we are studying how to optimize $UO(p, D)$-outlier detection for high dimensional cases, e.g., $k > 10$.

Preliminary Experimental Results

We ran our algorithm on a 16-attribute, 856-record dataset consisting of 1995-96 performance statistics for players in the National Hockey League. For a 3-dimensional cell structure containing 1000 cells, and using parameters $0.99 \leq p \leq 0.979$ and $70 \leq D \leq 125$, search times ranged from 0.11 to 0.17 seconds—approximately 50% of k-d tree search times. When we ran our algorithm against much larger, synthetic datasets, we obtained results even more dramatic. For 100000 tuples and $0.99 \leq p \leq 0.99999$ (approximately), our algorithm yielded search times ranging from 0.52 to 95.09 seconds. In comparison, a somewhat optimized brute force approach required 4.54 to 447.68 seconds. Searches using various kinds of k-d trees exceeded the CPU timer limit of 2147 seconds.

Acknowledgments

This research has been partially sponsored by NSERC Grant OGP0138055 and IRIS-2 Grants HMI-5 & IC-5.

References

Arning, A.; Agrawal, R.; and Raghavan, P. 1996. A linear method for deviation detection in large databases. In *Proc. KDD*, 164–169.

Barnett, V., and Lewis, T. 1994. *Outliers in Statistical Data.* John Wiley & Sons.

Bentley, J. L. 1975. Multidimensional binary search trees used for associative searching. *CACM* 18(9):509–517.

Ester, M.; Kriegel, H.-P.; Sander, J.; and Xu, X. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proc. KDD*, 226–231.

Fisher, D. 1987. Knowledge acquisition via incremental conceptual clustering. *Machine Learning* 2(2):139–172.

Freedman, D.; Pisani, R.; and Purves, R. 1978. *Statistics.* New York: W.W. Norton.

Gray, J.; Bosworth, A.; Layman, A.; and Pirahesh, H. 1995. Data cube: A relational aggregation operator generalizing group-by, cross-tab and sub-totals. Technical Report MSR-TR-95-22, Microsoft Research.

Hawkins, D. 1980. *Identification of Outliers.* London: Chapman and Hall.

Kaufman, L., and Rousseeuw, P. 1990. *Finding Groups in Data.* John Wiley & Sons.

Knorr, E. M., and Ng, R. T. 1997. A unified notion of outliers. Unpublished Manuscript, Dept. of Computer Science, University of British Columbia.

Ng, R., and Han, J. 1994. Efficient and effective clustering methods for spatial data mining. In *Proc. 20th VLDB*, 144–155.

Zhang, T.; Ramakrishnan, R.; and Livny, M. 1996. Birch: An efficient data clustering method for very large databases. In *Proc. ACM SIGMOD*, 103–114.