Principles of Knowledge Discovery in Data

Winter 2007

Chapter 7: Outlier Detection

Dr. Osmar R. Zaïane

University of Alberta

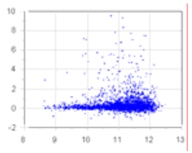
Course Content

- Introduction to Data Mining
- Association Analysis
- Sequential Pattern Analysis
- Classification and prediction
- Contrast Sets
- Data Clustering
- Outlier Detection
- Web Mining
- Other topics if time permits (spatial data, biomedical data, etc.)

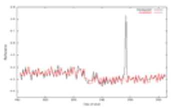
Chapter 7 Objectives

Learn basic techniques for data clustering.

Understand the issues and the major challenges in clustering large data sets in multi-dimensional spaces



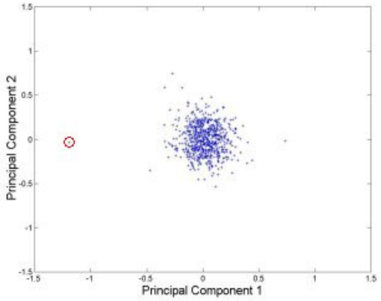
What is an Outlier?



- An observation (or measurement) that is unusually different (large or small) relative to the other values in a data set.
- Outliers typically are attributable to one of the following causes:
  - **Error**: the measurement or event is observed, recorded, or entered into the computer incorrectly.
  - **Contamination**: the measurement or event comes from a different population.
  - **Inherent variability**: the measurement or event is correct, but represents a rare event.

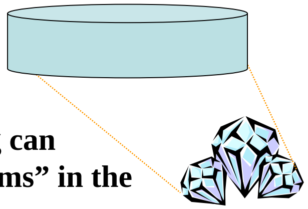
Many Names for Outlier Detection

- Outlier detection
- Outlier analysis
- Anomaly detection
- Intrusion detection
- Misuse detection
- Surprise discovery
- Rarity detection
- Detection of unusual events



Finding Gems

- If Data Mining is about finding gems in a database, from all the data mining tasks: characterization, classification, clustering, association analysis, contrasting…, outlier detection is the closest to this metaphor.



Data Mining can discover “gems” in the data

Lecture Outline

Part I: What is Outlier Detection *(30 minutes)*

- **Introduction to outlier analysis**
  - **Definitions and Relative Notions**
  - **Motivating Examples for outlier detection**
  - **Taxonomy of Major Outlier Detection Algorithms**

Part II: Statistics Approaches

- **Distribution-Based (Univariate and multivariate)**
- **Depth-Based**
- **Graphical Aids**

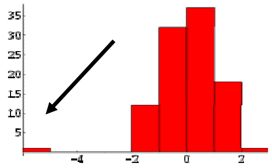
Part III: Data Mining Approaches

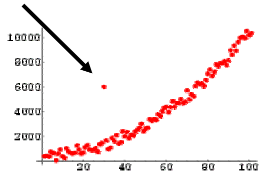
- **Clustering-Based**
- **Distance-Based**
- **Density-Based**
- **Resolution-Based**

Global versus Local Outliers

- Global outliers

Vis-à-vis the whole dataset

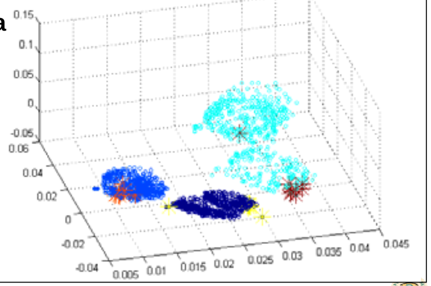




- Local outliers

Vis-à-vis a subset of the data

- Is there an anomaly more outlier than other outliers?
- Could we rank outliers?



Different Definitions

- An observation that deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism. **[Hawkins, 1980]**

- An outlier is an observation (or subset of observations which appear to be inconsistent with the remainder of that dataset **[Barnet & Lewis,1994]**

- An outlier is an observation that lies outside the overall pattern of a distribution **[Moore & McCabe, 1999]**

- Outliers are those data records that do not follow any patter in an application. **[Chen, Tan & Fu, 2003]**

More Definitions

- An object *O* in a dataset is a DB(*p*,*D*)-outlier if at least a fraction *p* of the other objects in the dataset lies greater than distance *D* from *O*. **[Knorr & Ng, 1997]**

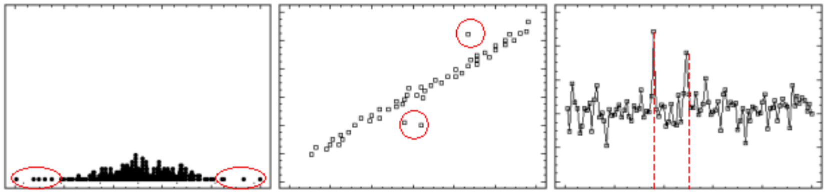
- An outlier in a set of data is an observation or a point that is considerably dissimilar or inconsistent with the remainder of the data **[Ramaswany, Rastogi & Shim, 2000]**

- Given an input data set with N points, parameters n and k, a point *p* is a D^k_N outlier if there are no more than n-1 other points *p'* such that $D^k(d') < D^k(p)$ where $D^k(p)$ denotes the distance of point *p* from its kth nearest neighbor. **[Ramaswany, Rastogi & Shim, 2000]**

- Given a set of observations X, an outlier is an observation that is an element of this set X but which is inconsistent with the majority of the data or inconsistent with a sub-group of X to which the element is meant to be similar. **[Fan, Zaïane, Foss & Wu, 2006]**

Relativity of an Outlier

- The notion of outlier is subjective and highly application-domain-dependant.



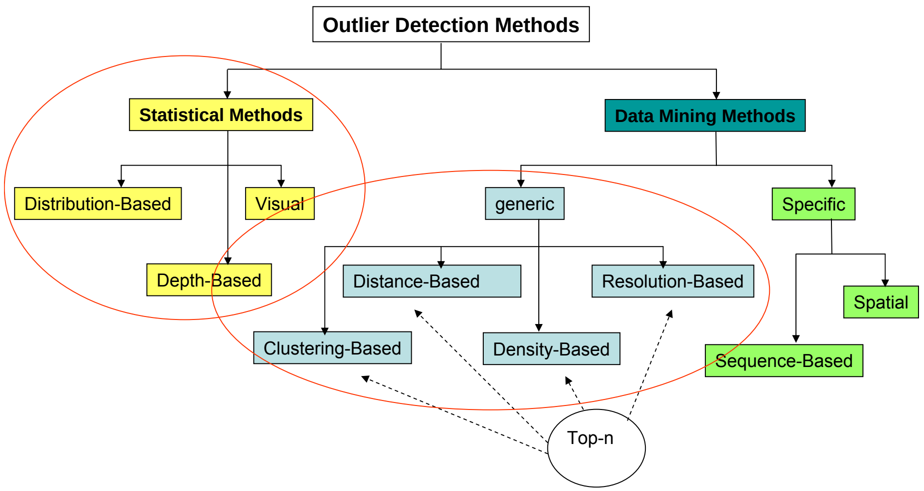
(a) Outliers w.r.t. a distribution (b) Outliers w.r.t. a pattern (c) time series outliers

There is an ambiguity in defining an outlier

Application of Anomaly Detection

- Data Cleaning - Elimination of Noise (abnormal data)
  - Noise can significantly affect data modeling (Data Quality)
- Network Intrusion (Hackers, DoS, etc.)
- Fraud detection (Credit cards, stocks, financial transactions, communications, voting irregularities, etc.)
- Surveillance
- Performance Analysis (for scouting athletes, etc.)
- Weather Prediction (Environmental protection, disaster prevention, etc.)
- Real-time anomaly detection in various monitoring systems, such as structural health, transportation;

Topology for Outlier Detection



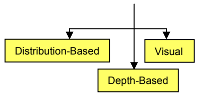
Lecture Outline

Part I: What is Outlier Detection *(30 minutes)*

- **Introduction to outlier analysis**
  - **Definitions and Relative Notions**
  - **Motivating Examples for outlier detection**
  - **Taxonomy of Major Outlier Detection Algorithms**

Part II: Statistics Approaches

- **Distribution-Based (Univariate and multivariate)**
- **Depth-Based**
- **Graphical Aids**



Part III: Data Mining Approaches

- **Clustering-Based**
- **Distance-Based**
- **Density-Based**
- **Resolution-Based**

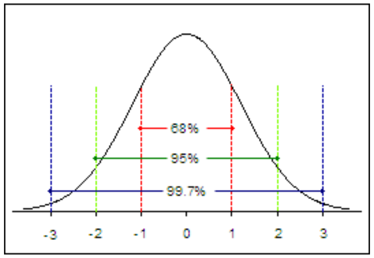
Outliers and Statistics

- Currently, in most applications outlier detection still depends to a large extent on traditional statistical methods.
- In Statistics, prior to the building of a multivariate (or any) statistical representation from the process data, a pre-screening/pre-treatment procedure is essential to remove noise that can affect models and seriously bias and influence statistic estimates.
- Assume statistical distribution and find records which deviate significantly from the assumed model.

Chebyshev Theorem

- Univariate

The definition is based on a standard probability model (Normal, Poison, Binomial) Assumes or fits a distribution to the data.



- The Russian mathematician P. L. Chebyshev (1821- 1894) discovered that the fraction of observations falling between two distinct values, whose differences from the mean have the same absolute value, is related to the variance of the population. Chebyshev's Theorem gives a conservative estimate to the above percentage.

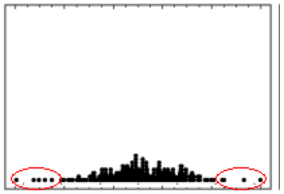
Theorem: The fraction of any data set lying within k standard deviations of the mean is at least $1 – 1/k^2$

- For any population or sample, at least $(1 - (1 / k^2)$ of the observations in the data set fall within k standard deviations of the mean, where $k >= 1$.
- For $k = 1$ It is not helpful to say that 0 is the proportion of all observations between + or – a standard deviation. However, for above 1, the theorem provides a lower bound for the proportion in question given k.

Distribution-Based Outlier Detection

- Univariate

According to Chebyshev’s theorem almost all the observations in a data set will have a z-score less than 3 in absolute value. – i.e. all data fall into interval [μ-3 σ, μ+3σ]
μ is the mean and σ is the standard deviation.



Z-score z=(x- μ)/σ

The z-score for each data point is computed and the observations with z-score greater than 3 are declared outliers.

- **Any problem with this?**

μ and σ are themselves very sensitive to outliers. Extreme values skew the mean. Consider the mean of {1,2,3,4,5} is 3 while the mean of {1, 2, 3, 4, 1000} is 202.

Covariance Matrix and Mahalanobis Distance

- The shape and size of multivariate data are quantified by the variance-covariance matrix.
- Given a dataset with *n* rows and *d* columns the variance-covariance matrix is a *d*x*d* matrix calculated as follows:
  - Center the data by subtracting the mean vector from each row
  - Calculate the dot product between columns
  - Multiply the matrix by the constant 1/(*n*-1)
- A well-known distance measure which takes into account the covariance matrix is the Mahalanobis distance.
- For a *d*-dimensional multivariate sample x_i (*i* = 1; …; *n*) the Mahalanobis distance is defined as

$$MD_i = \sqrt{(x_i - t)^T C^{-1} (x_i - t)}$$

for *i* = 1; …; *n*

where *t* is the multivariate arithmetic mean, and *C* is the sample covariance matrix.

Distribution-Based Outlier Detection

- **Multivariate**

For multivariate normally distributed data, the values are approximately chi-square distributed with p degrees of freedom (χ_p^2)

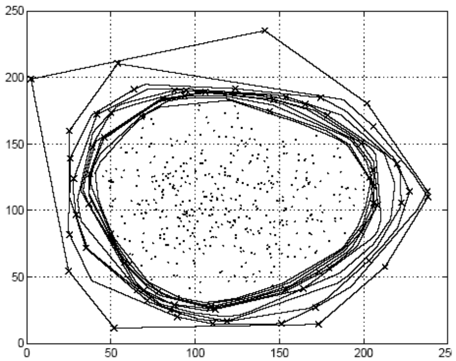
Multivariate outliers can now simply be defined as observations having a large (squared) Mahalanobis distance.

However, Mahalanobis distance needs robustification due to sensitivity of mean and variance to outliers

Use MCD – Minimum Covariance Determinant (a subset of points which minimizes the determinant of variance-covariance matrix. Compute mean and C based on this subset.)

Depth-Based Outlier Detection

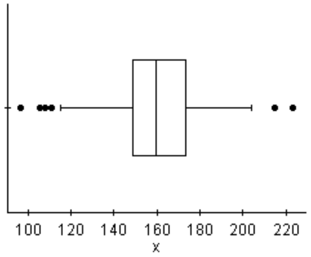
- Minimum Volume Ellipsoid Estimation: An ellipsoid is fitted around the dense area of the data. Outside ellipsoid ➔ outlier.
- Convex Peeling: based on computational geometry. Assigns a depth to each point. Points on the convex hull are labeled outliers.
- No assumption of probability distribution. No distance function required.
- Convex hull expensive.



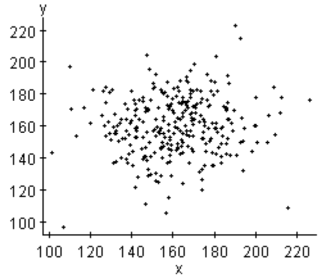
Graphical Aids

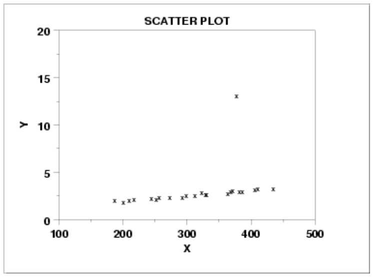
- Box-plot [Tukey 1977]

One bar at the median; box edges at lower quartile (25%) and upper quartile (75%); two whiskers at +- 1.5* interquartile range. Beyond whiskers are outliers.



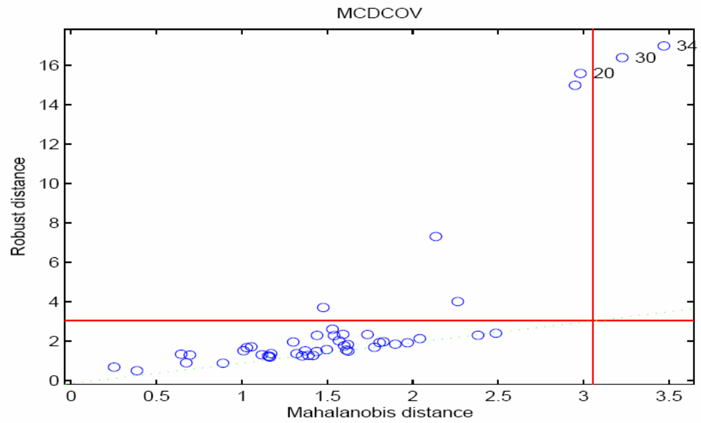
- Scatter-plot





Graphical Aids -2

- dd-plot



Lecture Outline

Part I: What is Outlier Detection *(30 minutes)*

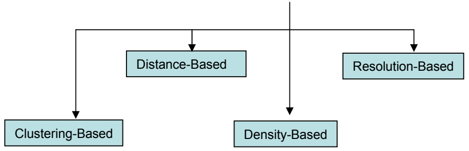
- **Introduction to outlier analysis**
  - **Definitions and Relative Notions**
  - **Motivating Examples for outlier detection**
  - **Taxonomy of Major Outlier Detection Algorithms**

Part II: Statistics Approaches

- **Distribution-Based (Univariate and multivariate)**
- **Depth-Based**
- **Graphical Aids**

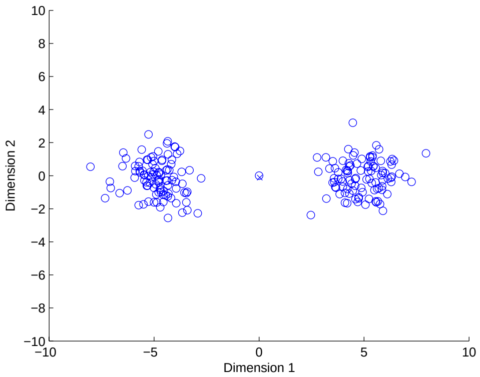
Part III: Data Mining Approaches

- **Clustering-Based**
- **Distance-Based**
- **Density-Based**
- **Resolution-Based**



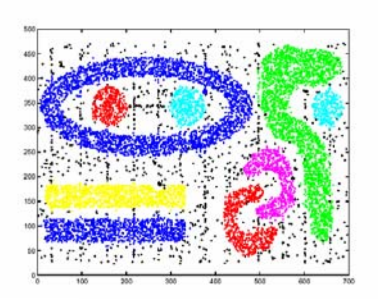
Problems with Statistical Solutions

- Consider the following case where the mean is itself an outlier.

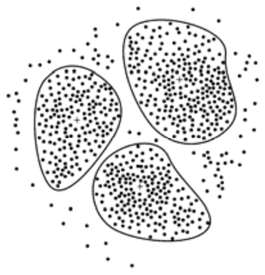


Clustering-Based Outlier Mining

- Some clustering techniques distinguish between isolated points and clustered points – non sensitive to noise. Example DBSCAN and TURN*.
- Identified small clusters and singletons are labeled outliers.

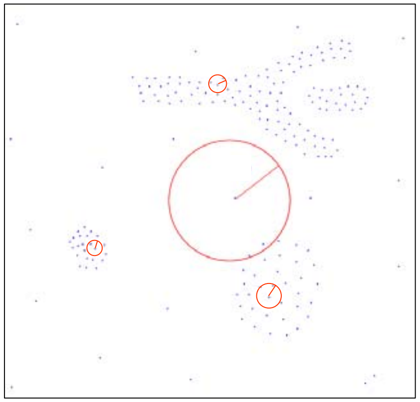


*TURN**'s clustering result on t7.10k.dat



k-Nearest Neighbor Approach

- Given k, for each point calculate the average distance to its k nearest neighbours. The larger the average distance the higher the likelihood the point is an outlier.
- Could sort in descending order the points based on their average distance to their k-NN.



K=3

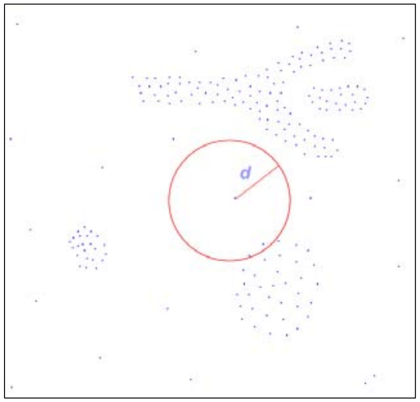
Distance-Based Outlier Mining

- A typical distance-based outlier notion--DB(*p,d*) outlier, was first introduced by Knorr and Ng.
- Definition of DB(*p*,*d*) outlier:

> an object *o* is an outlier if at least a fraction *p* of the objects in *S* lies at a distance greater than *d* from *o*

— [Knorr and Ng CASCON1997]

- Can effectively identify outliers which deviate from the majority.



Distance-Based Approach

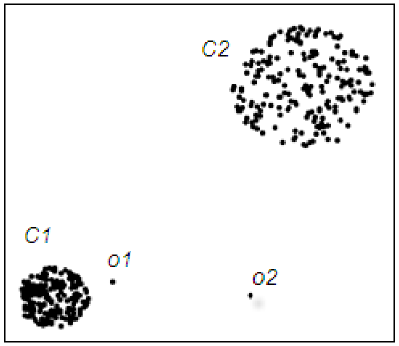
- DB(*p*,*d*) outliers tend to be points that lie in the sparse regions of the feature space and they are identified on the basis of the nearest neighbour density estimation. The range of neighborhood is set using parameters *p* (density) and *d* (radius).
- If neighbours lie relatively far, then the point is declared exceptional and labeled outlier.
- Distance between points is calculated iteratively in a Nested-loop (NL Algorithm). Improved upon later.

```
Simple Nested Loop Algorithm (O(N2))
For each object $x \in D$, compute distance to each $q \neq x \in D$
until $p + 1$ neighbors are found with distance $\leq d$.
If |Neighbors(o)|  $\leq p$, Report o as DB($p$,$d$) outlier.
```

- Possible use of index.

Distance-Based Issues

- Tend to find outliers global to the whole dataset.
- Not good for dataset consisting of clusters of diverse density.
- In the example, C_1 is significantly denser than C_2 and o_1 is not found as outlier since C_1 is too dense relative to C_2.



Density-Based Outlier Mining

- M. Berunig et al. [SIGMOD2000] proposed a Local Outlier Factor (LOF) to measure the degree of “outlying” of an object with regard to its surrounding neighborhood.
- LOF basically scores outliers on the basis of the density of their neighbourhood.
- LOF-outlier mining algorithm can effectively identify local outliers which deviate from their “belong-to” clusters (it is relative to their local neighborhood).

LOF Approach

- LOF is based on a number of new concepts. The formula below seems daunting but when broken down it makes sense.

$$LOF_k(p) = \frac{\sum_{o \in N_k(p)} \frac{lrd_k(o)}{lrd_k(p)}}{|N_k(p)|}$$

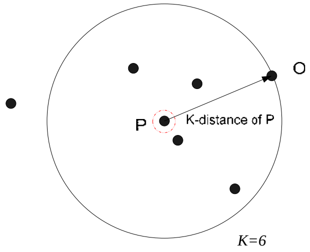
LOF(p) = Average of the ratio of the **local reachability density** of p and local reachability density of points in p **k-distance neighborhood**.

- Let’s see what these concepts mean.

k-distance Neigbourhood

k-distance of p: is the furthest distance among the k-nearest neighbours of a data point p.
k-distance(p) is defined as the distance d(p,o) between p & o such that:
-for at least k objects q ∈ D\{p} it holds that d(p,q) ⩽ d(p,o)
-for at most k-1 objects q ∈ D\{p} it holds that d(p,q) < d(p,o)

k is similar to MinPt in DBSCAN except that there is no radius ε and the number of points is always k. k-distance represents a variable ε.

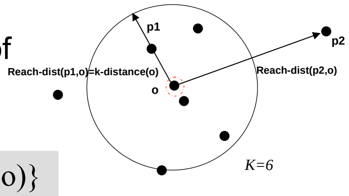


k-distance neighborhood of p: is the set of k-nearest neighbours i.e. the data Points closer to p than k-distance(p).

$$N_k(p) = \{q \in D \setminus \{p\} \mid d(p,q) \le \text{k-distance}(p)\}$$

Local Reachability Density

reachability distance of p w.r.t. o: is either the radius of the neighborhood of o if p in the neighborhood of o or the real distance from p to o.



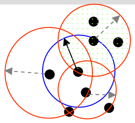
$$\text{reach-dist}_k(p, o) = \max\{\text{k-distance}(o), d(p,o)\}$$

Local reachability density of p: Inverse of the average reachability distance from the k-nearest-neighbors of p

$$lrd_k(p) = \frac{1}{\left[\frac{\sum_{o \in N_k(p)} reach-dist_k(p,o)}{|N_k(p)|}\right]}$$

$$LOF_k(p) = \frac{\sum_{o \in N_k(p)} \frac{lrd_k(o)}{lrd_k(p)}}{|N_k(p)|}$$

Average of the ratio of the local reachability density of p and those of p's k-NN.



LOF Issues

- For a data point p deep in a cluster, LOF(p)=1. The original paper gives an upper and lower bound for LOF(p) but it is simply a very large number >1 for outliers.
- Complexity is in the order O(N^2).
- Selecting k is not obvious. LOF does not change monotonically with k and the relationship between k and LOF is inconsistent from dataset to dataset and even from cluster to cluster within a dataset.

We define an Outlier as:

Given a set of observations X, an outlier is an observation that is an element of this set but which is inconsistent with the majority of the data or inconsistent with a sub-group of X to which the element is meant to be similar.

The above definition has two implications:
<u>outlier vis-à-vis the majority</u>; and
<u>outlier vis-à-vis a group of neighbours</u>.

There is global view and there is a local view

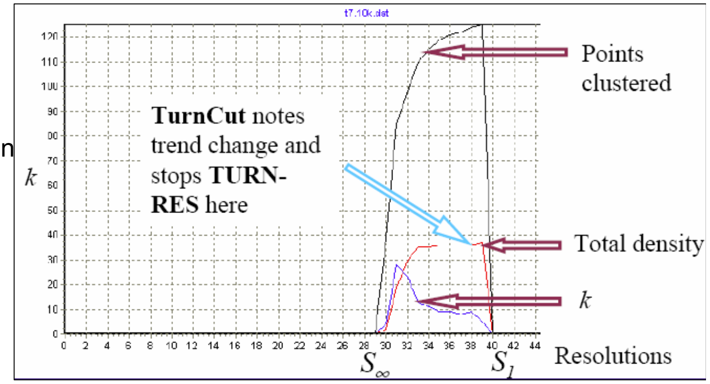
Resolution-Based Outlier

- Lemma: Given a set of data objects, the underlying "clusters" and "outliers" change when increasing or decreasing the resolution of data objects.
- This makes it possible to identify "clusters" or "outliers" by consecutively changing the resolution of a set of data objects and collect pre-defined statistical properties.

Clustering Using Resolution Change

- TURN*, A non-parametric clustering algorithm based on this principle is introduced by Foss and Zaïane [ICDM 2002]
- Two sub-algorithms
  - TURN-RES Clusters at a given resolution
  - TURN-CUT Selects the resolution

ROF uses the same principle



Neighbourhood and Resolution Change

- When the resolution changes for a dataset and the clustering is performed again, different outliers show different behavior in the re-distribution of clusters (i.e. vis-à-vis their neighbourhood)
- Definition of neighborhood
  - If an Object O has a nearest neighboring points P along each dimension in k-dimensional dataset D and the distance between P and O is less or equal to d, then P is defined as the close neighbor of O, all the close neighbors of P are also classified as the close neighbors of O, and so on. All these connected objects are classified as the same neighborhood.

Note: d can be any pre-defined value such as 1, it has no influence on the results, because the pair-wise distances between points are relative measurements during resolution change.

Resolution-Based Outlier Factor

- Definition of Resolution-based Outlier Factor (ROF)
  - If the resolution of a dataset changes consecutively between **maximum resolution** *where all the points are <u>non-neighbours</u>*, and **minimum resolution** *where all the points are <u>neighbours</u>*, the resolution-based outlier factor of an object is defined as the accumulated ratios of sizes of clusters containing this object in two consecutive resolutions.

Resolution-Based Outlier

- Definition of Resolution-based Outlier Factor (ROF)

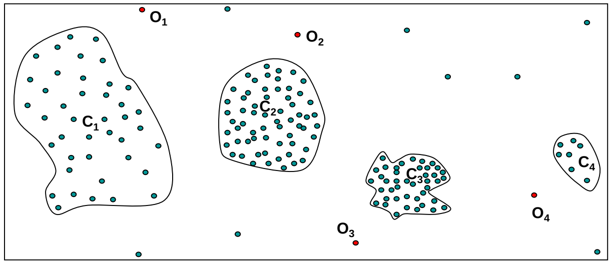
$$ROF(p) = \sum_{i=1}^{n} \frac{\text{ClusterSiz e}_{i-1}(p) - 1}{\text{ClusterSiz e}_{i}(p)}$$

Where

- r_1, r_2 ... r_i...r_n. — Resolution at each step
- n —Total number of resolution change steps from S_{max} to S_{min}
- CluserSize $_{i-1}$ — Number of objects in the cluster containing object p at the previous resolution
- ClusterSize $_i$ — Number of objects in the cluster containing object p at the current resolution

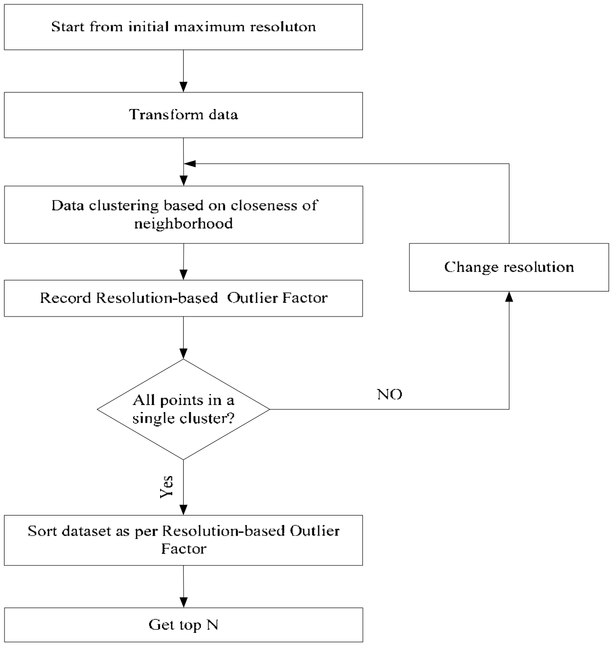
Synthetic 2D Example

|C1|=41
|C2|=61



- The most isolated objects get merged later than cluster points. They tend to get smaller ROF values. The last merged has the lowest ROF.
- The objects with enough neighbours as well as those affiliated with large size clusters (C2) increase their ROF values (approximately equal to 1) faster than smaller, isolated clusters (C4). The size of a cluster is its cardinality. Objects in C2 have higher ROF than those in C1 (61 vs 41)
- The definition can measure the degree of outlying for an object against its genuine cluster. This can be explained by comparing the outlying of O2 against its genuine cluster C2 versus O1 against its genuine cluster C1

Flowchart for resolution-based outlier mining algorithm


Start from initial maximum resoluton
Transform data
Data clustering based on closeness of neighborhood
Record Resolution-based Outlier Factor
All points in a single cluster?
NO
Change resolution
Yes
Sort dataset as per Resolution-based Outlier Factor
Get top N

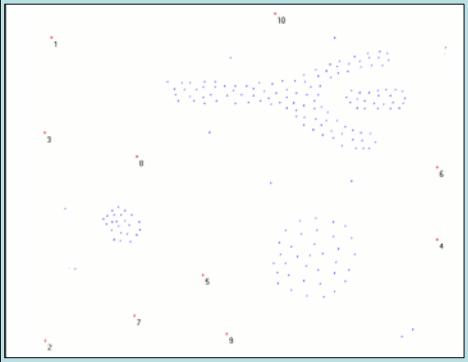
Comparison with DB-outlier and LOF-outlier

	DB-outlier	LOF-outlier	RB-outlier
Outlier Notion	Evaluates the degree of outlying of an object by looking a specified number of nearest objects	Measures how an object is deviated from its "best-guess" cluster	Measure how an object is deviated from its neighborhood with consideration to the surrounding community (reachable neighborhoods)
Outlier Mining Algorithm	Search the specified number of nearest objects to each object	Search the nearest objects and calculate the "local reachability density" of its neighborhood and LOF for each object	Change resolution of the dataset and collect properties of each object with respect to its clustering behavior at each changed resolution.

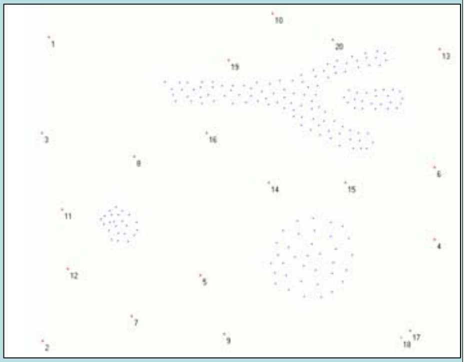
Comparison with DB-outlier and LOF-outlier

	DB-outlier	LOF-outlier	RB-outlier
Implementation and Application	Easy to implement, hard to use	Fair to implement, fair to use.	Easy to implement, easy to use
Outlier Mining Results	Best suited for datasets with a single cluster. Some local outliers are missed in case of multiple clusters	Good for datasets with multiple clusters with different densities and shapes. Good identification of local outliers.	Good for datasets with multiple clusters with different densities and shapes. Good identification of local outliers with consideration of some global features in a dataset. Satisfactorily ranking of the top listed outliers.

Some Comparative Results



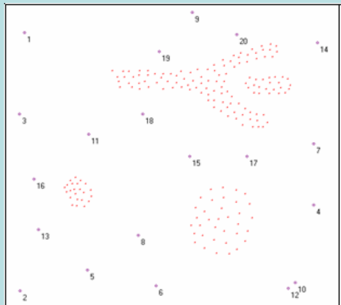
Top-10 outliers



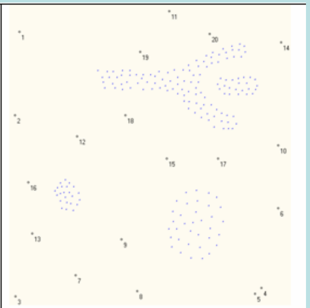
Top-20 outliers

Synthetic 200-tuple Dataset

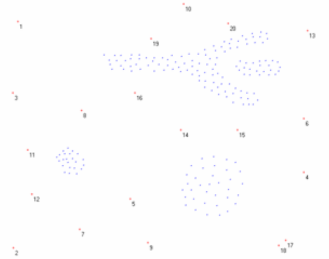
Comparison of results from DB-outlier, LOF-outlier, and RB-outlier on the 200-tuple dataset



Top-10 and Top-20 outliers identified by DB-outlier



Top-10 and Top-20 outliers identified by LOF-outlier



Top-10 and Top-20 outliers identified by RB-outlier

Notes:

1. DB-outlier: p=97/200;1-p=3/200, ranked by descending order of D_k;
2. LOF-outlier: MinPts=3, ranked by descending order of LOF;
3. RB-outlier: resolution changing ratio = 10%, ranked by ascending order of ROF.

Challenges For Distance and Density Based Approaches

- Both distance and density-based methods require KNN search, resulting in quadratic complexity
- Efficient algorithms
  - Reducing the number of computations
  - Scalable to large, multidimensional datasets
  - Detecting both global and local outliers
- Synthetic benchmark test datasets for evaluation

Reference-Based Outliers

- Approximation of distance-based outliers, yet able to identify local outliers
- Neighborhood density of a data point *x* is defined w.r.t. a set of reference points
- For a reference point *p*, computing the distances from each data point to *p*

$$X_p = \{d(x_i, p), 1 \le i \le n\}$$

- The above vector can be viewed as one dimensional representation of the original dataset *X*

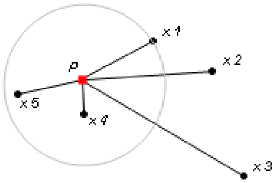
Reference-Based Nearest Neighbor

- For a given data point *x*, its reference-based nearest neighbor w.r.t. the vector X^p is the closest point to it in X^p

$$|d(x,p)-d(y,p)| = \min_{1\le i\le n} |d(x,p)-d(x_i,p)|$$

- Example: $X=\{x_1, x_2, x_3, x_4, x_5\}$

NN of x_1 is x_5,
NN of x_2 is x_1,
NN of x_3 is x_2
…



Reference-Based Neighborhood Density

- The density for *x* w.r.t. *p* is defined as the reciprocal of the average distance to its reference-based KNNs

$$D(x,k,p) = \frac{1}{\frac{1}{k}\sum_{j=1}^{k}|d(x_j,p)-d(x,p)|}$$

- The neighborhood density of *x* w.r.t. a reference set *P* is defined as the minimum of the density over all the reference points.

$$D^P(x,k) = \min_{1\le r\le R} D(x,k,p_r)$$

Reference-based Outlier Score (ROS)

- Data points with low density have high outlier scores

$$ROS(x) = 1 - \frac{D^P(x,k)}{\max_{1\le i\le n} D^P(x_i,k)}$$

- Data points are ranked according to ROS
- Outliers are those with high scores

Worst case complexity is *O(Rnlog(n))*, where *R* is the number of reference points and *n* is the data size.

Algorithm

- For each reference point *p*, sort the original dataset *X* in the vector X^p
- For each data point *x*, find the k reference-based nearest neighbors and compute the average neighborhood density
- Reference-based neighborhood density of *x* is the minimum of all neighborhood densities w.r.t the reference set *P*
- Compute Reference-based Outlier Score (ROS)

Determination of Reference Points

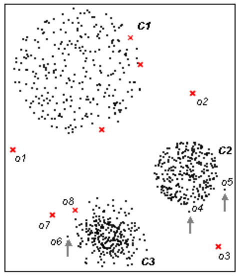
- Reference points are not necessarily data points
- Partition the data space into a grid and reference points are the vertices on the grid.
- To determine the number of reference points
  - Partition the space incrementally from coarse resolution to fine resolution
  - In the next round, only calculate the additional reference points since previous calculation can be reused
- For high dimensional data, it suffices to use only a few dimensions to partition the data space due to data sparsity

Compatible with Distance-Based Method

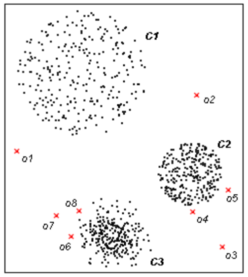
- Lower bounded by the neighborhood density computed using traditional KNN approach
- Sufficient to use one reference point (say 0) with one-dimensional data
- Equivalent to distance-based approach when all data points are used as reference points

Detecting global and local outliers in Complex data

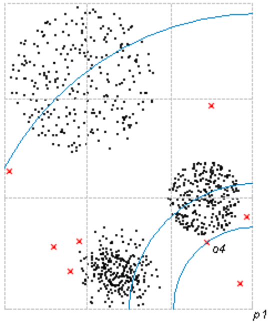
- Reference-based method is dynamic and able to see the whole dataset from different viewpoint
- Tradeoff between the number of reference points and the ability to detect global and local outliers



KNN result



Our result



Finding local outliers

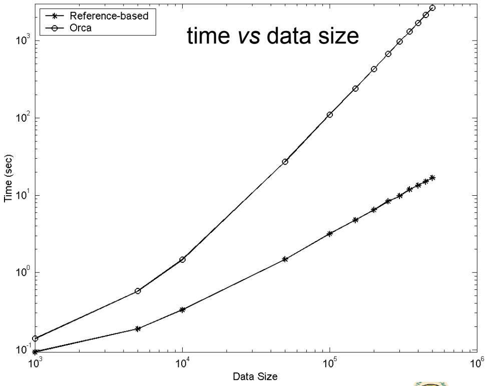
Experiment with Synthetic Data

- Developed a synthetic data generation system to facilitate the evaluation of different outlier detection methods
- Tested with two sets of 2D datasets
  - Each dataset in 1st set contains a normally distributed cluster
  - 2nd set has one complex dataset that contains several clusters with both global and local outliers

Experimental Results on Efficiency

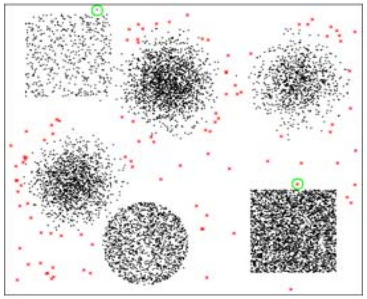
- Data size ranges from 1,000 to 500,000

- Orca is the C implementation of the near linear distance-based approach [Bay02].
- ROS is in Java.
- Mine top 1% outliers
- Logarithm scale

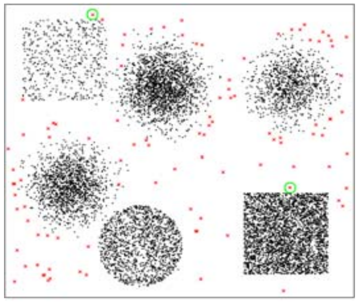


Experimental Results on Effectiveness

- ROS is effective in finding global and local outliers
- LOF is well known for detecting local outliers
- Data size = 10,000, top 1% outliers



LOF result



ROS result