

Outlier Detection: Principles, Techniques and Applications

Sanjay Chawla and Pei Sun

School of Information Technologies

University of Sydney

NSW, Australia

chawla|psun2712@it.usyd.edu.au

Outline

- Outlier Detection - a core data mining paradigm.
- Methods for Outlier Detection.
- Outlier Detection as Unsupervised Learning.
- Classical and Modern Statistical Approaches.
- Lunch Break
- Data Mining/Database Approaches.
- A contemporary Outlier Detection System.
- A historical example of Outlier Detection and consequences.

Outlier Detection - a core data mining paradigm

- At equilibrium, the data mining paradigms are
 1. Classification
 2. Clustering
 3. Association Rule Mining
 4. Outlier Detection (Anomaly Detection)
- Outlier Detection comes closest to the metaphor of discovering “nuggets” of useful, actionable information in large databases.
- Science evolves and “moves ahead” by proposing theories to “explain” outliers.

Methods for Outlier Detection

Definition 1 (Hawkins[5]) *An outlier is an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism*

- The goal of outlier detection is to uncover the “different mechanism”.
- If samples of the “different mechanism” exist, then build a classifier to learn from samples.
- Often called the “Imbalanced Classification Problem” - size of outlier sample is small vis-a-vis normal samples.
- In most practical real-life settings, samples of the outlier generating mechanism are non-existent.

Outlier Detection as Unsupervised Learning

- We will exclusively focus on “unsupervised learning” techniques.
- Reduces to finding sparse regions in large multidimensional data sets.
- We will first overview statistical methods and then provide an overview of methods that have emerged from within the DM community.

Statistical Approaches

Lets begin with the probability density function of the Normal distribution



$$f(x) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} e^{-[(x-\mu)/\sigma]^2/2}$$



$$f(x) > 0$$

and



$$\int_{-\infty}^{\infty} f(x) dx = 1$$

Univariate Normal Distribution

- What is the probability that a point will lie within k standard deviations of the mean (for a Normal distribution) ?
- This is equivalent to asking what is area under the Normal probability density function between $\mu - k\sigma$ and $\mu + k\sigma$.

k	Area
1	0.6827
2	0.9545
3	0.9973

- Be on the alert for fat tail distributions! (Katrina and LTCM).

Tails bounds for arbitrary distributions

- **Theorem 1 Markov Inequality:**

Let X be a non-negative random variable. For any $a > 0$

$$P(X > aE[X]) < \frac{1}{a}$$

Theorem 2 Chebyshev's Inequality: *let X be a random variable with expectation $E[X]$ and finite standard deviation σ . Then for any $t > 1$*

$$P(|X - E[X]| > t\sigma) < \frac{1}{t^2}$$

- For a Normal distribution, probability of finding a point which is at more than three standard deviations away from the mean is less 0.0027, for an arbitrary distribution the probability is less than $\frac{1}{9} = 0.11$.

Other Tail Bounds

Theorem 3 (Chernoff (Theory)) *Let X_1, X_2, \dots, X_n be a sequence of independent 0-1 random variables where $P(X_i = 1) = p$. Let $X = \sum X_i$ and $\mu \geq E[X]$. Then for any $\delta > 0$*

$$Pr(X > (1 + \delta)\mu) < \left[\frac{e^\delta}{(1 + \delta)^{1+\delta}} \right]^\mu$$

The data mining version makes the dependence on n explicit -often used for measuring the accuracy of sampling.

$$P(X \leq (1 - \epsilon)np) \leq e^{-\frac{\epsilon^2 np}{2}}$$

$$P(X \geq (1 + \epsilon)np) \leq e^{-\frac{\epsilon^2 np}{3}}$$

Notice the bounds are not symmetric.

Other Tail Bounds

Theorem 4 Hoeffding Like Weak Inequality[12]: *Let $\{x_1, \dots, x_l\}$ be an independent and identically distributed set of instances from a random variable X whose spread is bounded by R . Let $\mu = E[X]$ and $s = \frac{1}{l} \sum_{i=1}^l x_i$ be the sample mean. Then for any $0 < \delta < 1$, with probability at least $1 - \delta$*

$$|s - \mu| < \frac{R}{\sqrt{l}} \left(2 + \sqrt{2 + \ln \frac{1}{\delta}} \right) \equiv f(R, l, \delta)$$

- *f is an increasing function of R - the spread.*
- *f is a decreasing function of l - the number of points*
- *f is an increasing function of $1 - \delta$.*

Note: This holds for any X . X can be infinite-dimensional (thus can be used with kernel methods).

A Simple Outlier Detection Algorithm

As before, we have a set of iid points $\{x_1, \dots, x_l\}$ with sample mean s and population mean μ . A new point x_{l+1} arrives. Is x_{l+1} an outlier?

Conceptual: Let $d_i = |x_i - \mu|$. Let B be a ball centered at μ of radius $\max_{1 \leq i \leq l} d_i$. If d_{l+1} lies outside the ball then declare it as an outlier.

Problem: We don't know μ and therefore neither the d_i 's.

A Solution: Use the sample mean s and use Hoeffding Inequality to estimate the threshold.

Weakness: Why center on the mean? Choose a point which will minimize the radius. Furthermore *the mean* is not robust.

Analysis[12]

$$P\left(\max_{1 \leq i \leq l+1} d_i \neq \max_{1 \leq i \leq l} d_i\right) =$$

$$P(d_{l+1} > \max_{1 \leq i \leq l} d_i) \leq \frac{1}{l+1} \text{ (by symmetry and iid)}$$

Now,

$$d_{l+1} = |x_{l+1} - \mu| \geq |x_{l+1} - s| - |s - \mu|$$

and for all $i = 1..n$

$$d_i = |x_i - \mu| \leq |x_i - s| + |s - \mu|$$

Rearrange and Combine,

$$P(|x_{l+1} - s| > \max_{1 \leq i \leq l} |x_i - s| + 2f(R, \delta, l)) < \frac{1}{l+1}$$

From Univariate to Multivariate

- Again, examine the exponent of the univariate Normal distribution

$$\left(\frac{x - \mu}{\sigma}\right)^2 = (x - \mu)(\sigma^2)^{-1}(x - \mu)$$

From Univariate to Multivariate

- Again, examine the exponent of the univariate Normal distribution

$$\left(\frac{x - \mu}{\sigma}\right)^2 = (x - \mu)(\sigma^2)^{-1}(x - \mu)$$

- This is the square of the distance from the point x to the mean μ in units of the standard deviation.

From Univariate to Multivariate

- Again, examine the exponent of the univariate Normal distribution

$$\left(\frac{x - \mu}{\sigma}\right)^2 = (x - \mu)(\sigma^2)^{-1}(x - \mu)$$

- This is the square of the distance from the point x to the mean μ in units of the standard deviation.
- This distance can be generalized to higher dimension and is called the (square of the) Mahalanobis distance.

Multivariate Normal Distribution

- In one dimension, the exponent of the Normal distribution is

$$\left(\frac{x - \mu}{\sigma}\right)^2 = (x - \mu)(\sigma^2)^{-1}(x - \mu)$$

- For d dimension, the exponent is

$$(x - \mu)' \Sigma^{-1} (x - \mu)$$

- Σ is the $d \times d$ variance-covariance matrix. Σ is a positive-definite matrix.

Mahalanobis Distance

- Given an $N \times D$ data set (N rows, D columns), the (square of) Mahalanobis Distance between two points x and y is

$$Mahal(x, y) = (x - y)\Sigma^{-1}(x - y)'$$

- where Σ is the $D \times D$, variance-covariance matrix.

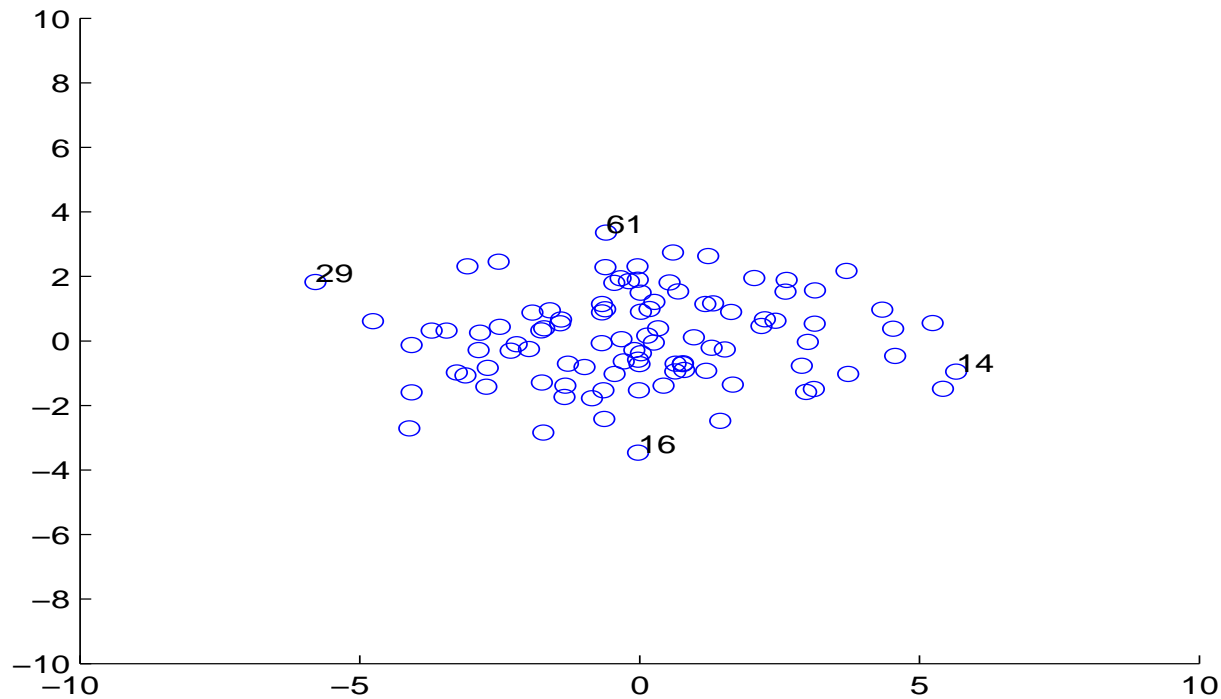
$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1d} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{d1} & \sigma_{d2} & \dots & \sigma_{dd} \end{bmatrix}$$

The Variance-Covariance Matrix

- Let X be an $N \times D$ data set, where N represents the number of entities and D represents the number of variables.
- The variance-covariance matrix is calculated as follows
 1. Center the data by subtracting the mean vector from each row.
 2. Calculate the dot product between the columns.
 3. Multiply the matrix by the constant $\frac{1}{N-1}$.

$$\begin{bmatrix} 2 & 0 \\ 1 & 2 \\ 0 & 7 \end{bmatrix}; m = [1, 3]; \begin{bmatrix} 1 & -3 \\ 0 & -1 \\ -1 & 4 \end{bmatrix} \xrightarrow{\text{dotp}/2} \begin{bmatrix} 1.0 & -3.5 \\ -3.5 & 13 \end{bmatrix}$$

Mahalanobis vs. Euclidean Distance



Point Pairs	Mahalanobis	Euclidean
(14,29)	5.07	11.78
(16,61)	4.83	6.84

Distribution of Mahalanobis Distance

- Given a set S of N d -dimensional points from a Normal distribution, what is the distribution of the Mahalanobis Distance to the mean of the set?

Distribution of Mahalanobis Distance

- Given a set S of N d -dimensional points from a Normal distribution, what is the distribution of the Mahalanobis Distance to the mean of the set?



$$(x - \mu)' \Sigma^{-1} (x - \mu)$$

Distribution of Mahalanobis Distance

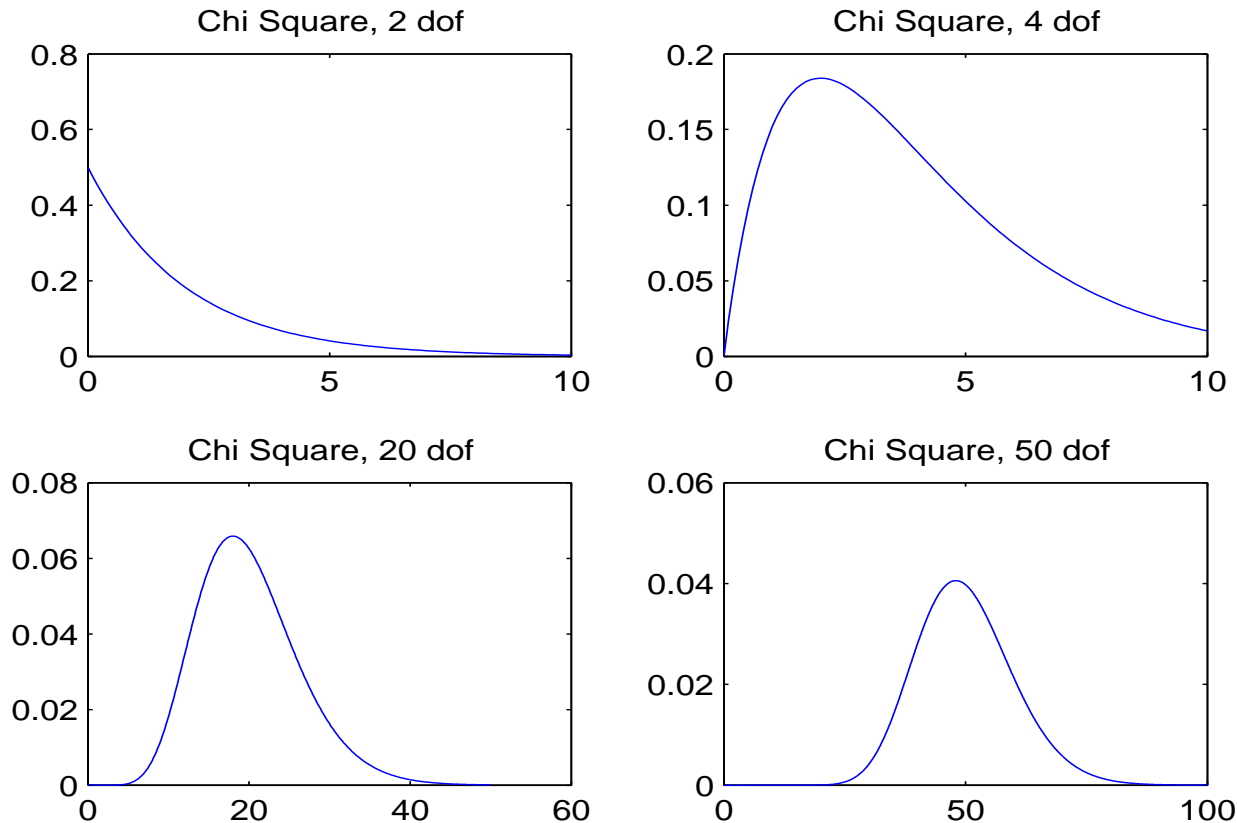
- Given a set S of N d -dimensional points from a Normal distribution, what is the distribution of the Mahalanobis Distance to the mean of the set?



$$(x - \mu)' \Sigma^{-1} (x - \mu)$$

- Answer: The Square of the Mahalanobis distance is distributed as a χ_d^2 distribution, i.e., a chi-square distribution with d degrees of freedom.

Examples of Chi-Square Distribution



Notice the concentration of square of the distance as the degree of freedom (dimension) increases. This is the famous “curse of dimensionality”

Algorithm For Finding Multivariate Outliers

- Input: An $N \times d$ data set S .
 - Output: Candidate Outliers
1. Calculate the mean μ and variance-covariance matrix Σ .
 2. Let M be $N \times 1$ vector consisting of square of the Mahalanobis distance to μ .
 3. Find points O in M whose value is greater than $inv(\sqrt{\chi_d^2(.975)})$.
 4. Return O

Robustification of Estimators

- It is well known that both the mean and standard deviation are extremely sensitive to outliers

Robustification of Estimators

- It is well known that both the mean and standard deviation are extremely sensitive to outliers
- Effectively one “bad point” can completely skew the mean

$$\text{Mean}(1, 2, 3, 4, 5) = 3 \text{ and } \text{Mean}(1, 2, 3, 4, 1000) = 202$$

Robustification of Estimators

- It is well known that both the mean and standard deviation are extremely sensitive to outliers
- Effectively one “bad point” can completely skew the mean

$$\text{Mean}(1, 2, 3, 4, 5) = 3 \text{ and } \text{Mean}(1, 2, 3, 4, 1000) = 202$$

- We are using the Mahalanobis distance to find outliers and yet it itself is being effected by the outliers!

Robustification of Estimators

- It is well known that both the mean and standard deviation are extremely sensitive to outliers
- Effectively one “bad point” can completely skew the mean

$$\text{Mean}(1, 2, 3, 4, 5) = 3 \text{ and } \text{Mean}(1, 2, 3, 4, 1000) = 202$$

- We are using the Mahalanobis distance to find outliers and yet it itself is being effected by the outliers!
- In the Statistics literature, P.J. Rousseeuw has done pioneering work to “robustify” estimators.

Robustification of Σ

- “Robustification” means making the statistical estimator less sensitive to outliers.
- The most famous method for the robustification is due to Rousseeuw called the Minimum Covariance Determinant (MCD)[11].
- The MCD estimator is determined by a subset of points (of size h) which minimizes the determinant of the variance-covariance matrix over all subsets of size h .



$$R^* = \operatorname{argmin}\{\det(\Sigma_R) \mid R \text{ is a subset of } D \text{ of size } h\}$$

- Compute μ and Σ and then Mahalanobis based on R^*

Practitioner's Guide to MCD

- For a $N \times d$ matrix, make sure $h > d$.
- Typical value of $h = 0.75N$.
- Assumes data set contains at most 25% “outliers”.
- The MCD estimators can resist $N - h$ outliers.

Fast-MCD Algorithm

- The MCD algorithm is based on the following observation:

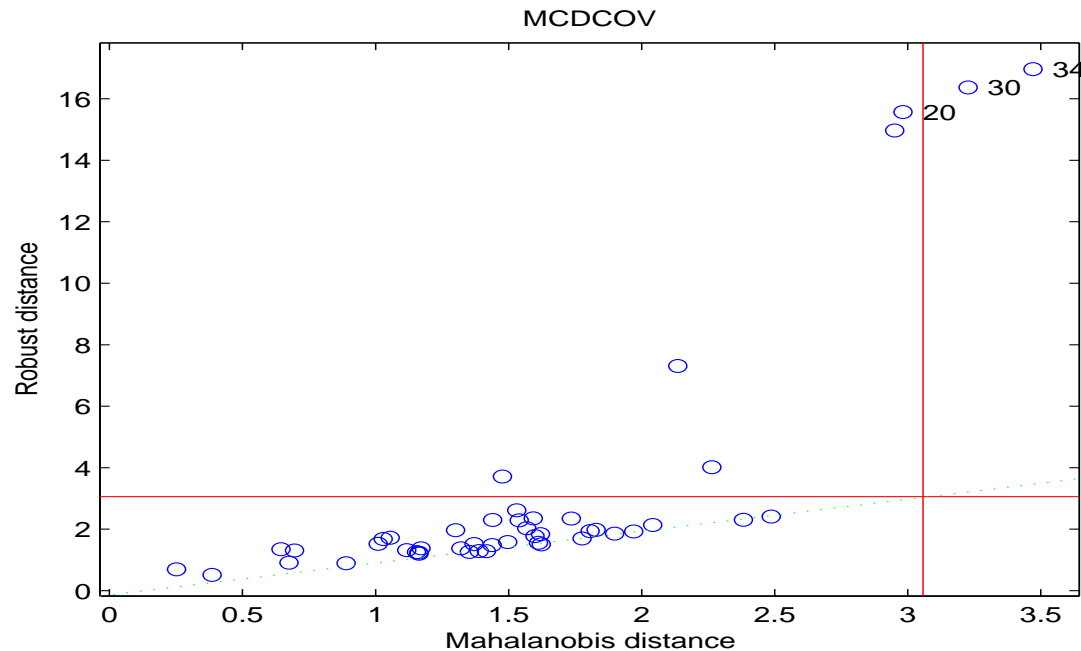
Theorem 5 *Let S be the original $N \times d$ data set. Let S_1 be a size h subset of D . Compute μ_{S_1} and Σ_{S_1} based on S_1 . Compute, the Mahalanobis distance of ALL points of S based on μ_{S_1} and Σ_{S_1} . Sort the Mahalanobis distance and select h points with smallest distance. Let the set be S_2 . Then*

$$\det(S_2) \leq \det(S_1)$$

- Thus starting from a random configuration, the next iteration will not increase the determinant.
- K-means type of algorithm.

D-D plots

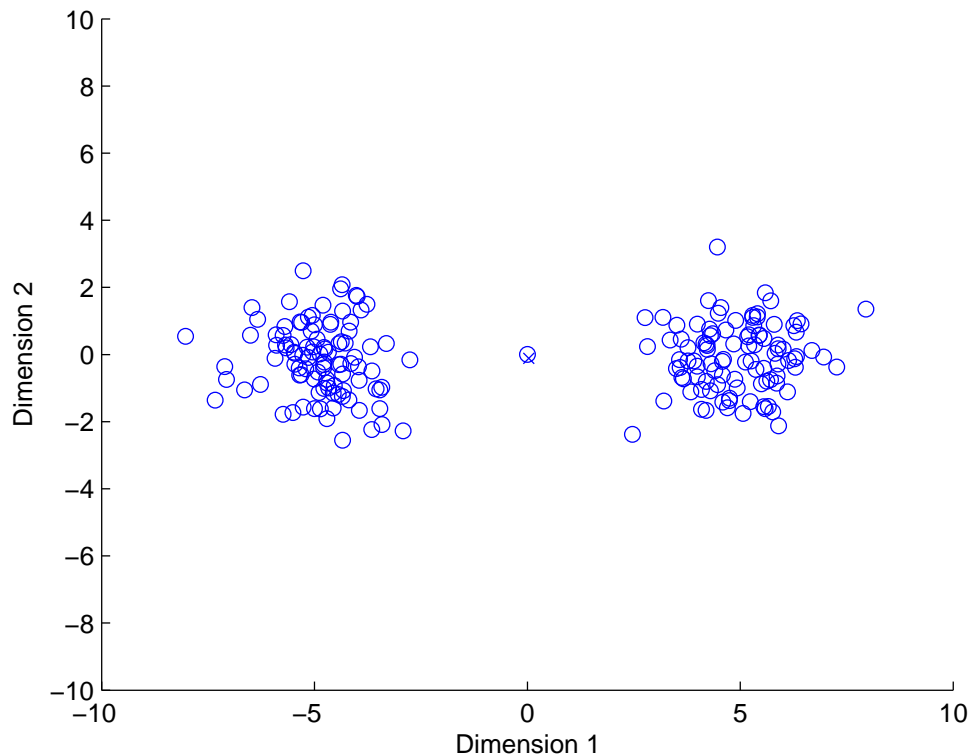
- D-D plots are useful for comparing the Robust Mahalanobis and the full Mahalanobis distance.
- Example[Libra Package (Verboven and Hubert)]:



- Robust estimates can “detect” some more candidate outliers.

Weakness of Statistical Methods

- Methods tied to known distributions.
- Consider for example:



- The arithmetic mean of the data set is the outlier !!!

Distance-based Methods

- Distance-based Methods were introduced by Knorr and Ng.
- **Definition 2** *An object O in a dataset T is a $DB(p,D)$ -outlier if at least fraction p of the objects in T are at greater distance than D from O .*
- **Theorem 6** *Let T be observations from a univariate Normal distribution $N(\mu, \sigma)$ and O is a point from T . Then the z-score of O is greater than three iff O is a $DB(0.9988, .13\sigma)$ outlier.*
- Knorr and Ng[6] proposed several other generalization of outliers based on other distributions.

Distance-based Outliers and K-Nearest Neighbors

- **Definition 3 (Kollios[7])** *An object o in a dataset T is a $DB(k,D)$ outlier if at most k objects in T lie at distance at most D from o .*
- **Definition 4 (Ramaswamy[10])** *Outliers are the top n data elements whose distance to the k th nearest neighbour is greatest.*

Simple Nested Loop Algorithm ($O(N^2)$)

For each object $o \in T$, compute distance to each $q \neq o \in T$ until $k + 1$ neighbors are found with distance less than or equal to D .

If $|Neighbors(o)| \leq k$, Report o as $DB(k, D)$ outlier.

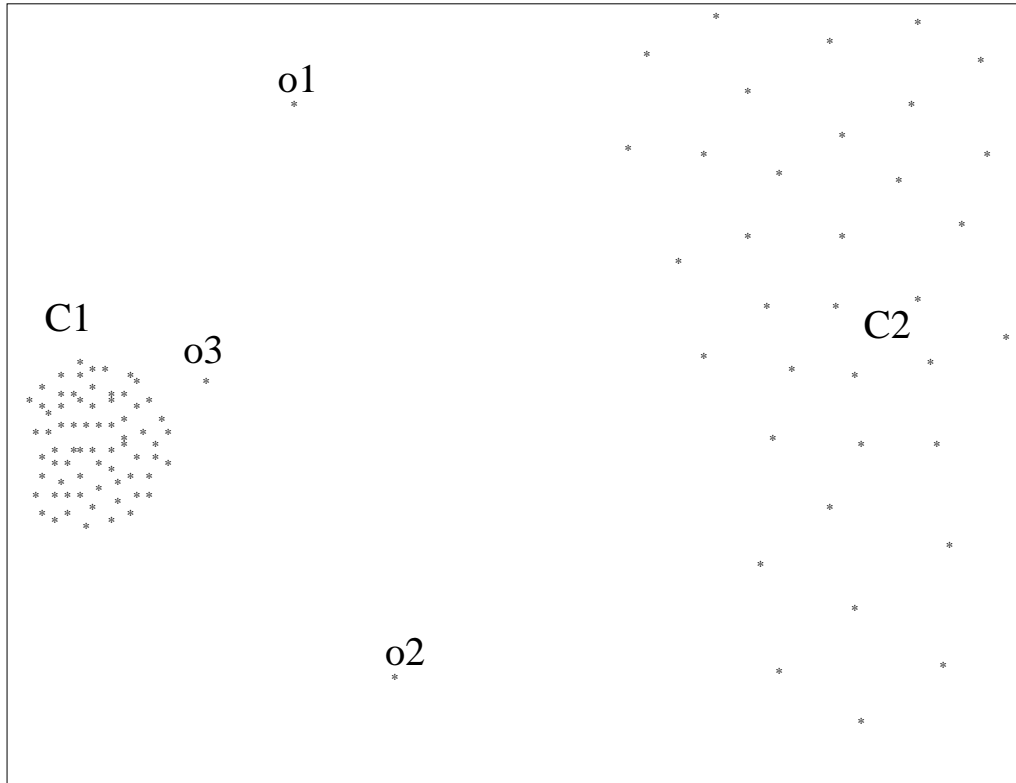
Nested Loop with Randomization and Pruning

- The fastest known distance-based algorithm for DB_n^k is due to Bay and Schwabacher[2] (Expected Running Time of $O(N)$)
- It is variation on the Nested-Loop algorithm.
 1. Randomize the data.
 2. Partition the data D into blocks.
 3. Compare each point in the block to every point in D.
 4. Keep track of the Top n outliers, $Topn$ and the weakest outlier (i.e., the point in $Topn$ which has the smallest k nearest neighbor, during block processing).
 5. Prune points as soon as they become non-outliers.
 6. As more blocks are processed, the weakest score keeps increasing and more points get pruned sooner.

Average Case Analysis of Algorithm

- Model point processing as Bernoulli trials until k successes are achieved, i.e, found k neighbors with distance $< d$.
- The number of trials $Y = y$ needed for k successes is given by the negative binomial distribution.
- Then it can be shown that
$$E[Y] \leq \frac{k}{\pi(x)} + (1 - \sum_{y=k}^N P(Y = y))N$$
- $\pi(x)$ is the probability of a random point being within distance d of x .
- The first term is independent of N and the second term is small because we are only interested in few outliers.
- Thus $E[Y]$ can be seen as being independent of N .

Limitations of Distance-based Outliers



- Cluster C1 and C2 have different densities
- For O_3 to be a distance-based outliers, all points of C_2 will have to be outliers

Density-based Outliers

The Local Outlier Factor (LOF) method[3] is based on scoring outliers on the basis of the density in the neighborhood. Our discussion is from Tan, Steinbach and Kumar.

- **Definition 5** *The outlier score of an object is the reciprocal of the density in the object's neighborhood.*
- Density can be defined as the average distance to the k nearest neighbors.

$$density(x, k) = \left(\frac{\sum_{y \in N(x, k)} distance(x, y)}{|N(x, k)|} \right)^{-1}$$

Relative Density of Points

- **Definition 6** *Relative density of a point x is the ratio of the density of a point and the average density of its neighbors.*

$$\mathit{reldensity}(x, k) = \frac{\mathit{density}(x, k)}{\sum_{y \in N(x, k)} \mathit{density}(y, k) / |N(x, k)|}$$

- The relative density of points “deep” inside a cluster is 1.
- Define outlier score of a point x as its relative density.
- Weakness: Complexity is $O(N^2)$ and selecting the right k is not obvious.

Outliers in High Dimensional Data

- We saw earlier that as the dimension of the space increases, the Mahalanobis distance starts clustering around a single value.
- Thus the notion of outliers (based on distance) may become meaningless in high dimensional space.
- Even for LOF, the density is defined in terms of distance so the LOF values will tend to cluster in high dimensional space.
- However, projection of points to a lower dimension space may yield meaningful outliers.

Subspace Outliers

- We present the algorithm, due to Yu and Aggarwal[1] to find outliers in subspaces of a high dimension space.
- Partition each d -dimension space into equi-depth ϕ intervals.
- Let X be a Bernoulli rv, where $X = 1$ if an object is present in a d' cube C and 0 otherwise.
- Then $E(X) = N f^{d'}$ where $f = \frac{1}{\phi}$, N is the number of points. The standard deviation is $\sqrt{(N f^{d'} (1 - f^{d'}))}$.
- Report low sparsity coefficient (SC) cubes:

$$SC(C) = \frac{N(C) - N f^{d'}}{\sqrt{(N f^{d'} (1 - f^{d'}))}}$$

Spatial Outliers

- Outlier Detection techniques are often used in GIS, climate studies, public health, etc. The spread and detection of “bird flu” can be cast as an outlier detection problem.
- The distinguishing characteristics of spatial data is the presence of “spatial attributes” and the neighborhood relationship.



Definition 7 (Shekhar et. al.[13]) *A spatial outlier is a spatial referenced object whose non-spatial attribute values are significantly different from those of other spatially referenced objects in its spatial neighborhood.*

Finding Spatial Outliers

- Let O be a set of spatially-referenced objects and $N \subset O \times O$ a neighborhood relationship. $N(o)$ are all the neighbors of o .
- **Theorem 7 (Shekhar et. al.)** *Let f be function which takes values from a Normally distributed rv on O . Let $g(x) = \frac{1}{|N(x)|} \sum_{y \in N(x)} f(y)$. Then $h = f - g$ is Normally distributed.*
- h effectively captures the deviation in the neighborhood. Because of spatial autocorrelation, a large value of $|h|$ would be considered unusual.
- For multivariate f , use the Chi-Square test to determine outliers.

Sequential Outliers

In many applications, data is presented as a set of symbolic sequences

Proteomics Sequence of amino acids

Computer Virus Sequence of register calls

Climate Data Sequence of discretized SOI readings

Surveillance Airline travel patterns of individuals

Health Sequence of diagnosis codes

Example Sequence Data

1. SRHPAZBGKPBFLBCYVSGFHPXZIZIBLLKB
2. IXCXNCXKEGHSARQFRA
3. MGVRNSVLSGKKADELEKIRLRPGGKKKYYML

Determine the outlier sequences?

Properties of Sequence Data

- Not of same length; Sequence of symbols.
- Unknown distribution; No standard notion of similarity.
- Is there an invariant?

Short Memory Property of Sequences

- Given a sequence $s = s_1 s_2 \dots s_l$, there exists an $L < l$ such that the conditional probabilities
$$P(s_l | s_{l-k} \dots s_{l-1}) \approx P(s_l | s_{l-L} \dots s_{l-1}) \quad \forall k > L.$$
- This is essentially a higher-order Markov condition.
- The size of the state is exponential in L .
- Key Insight by Ron, Singer and Tshiby: Use variable length markov chains.
- Led to the invention of the Probabilistic Suffix Tree data structure.

Basic Probability Identity

- $P(A, B) = P(A|B)P(B) = P(B|A)P(A)$.

- The above can be generalized to

$$P(s_1, \dots, s_l) = P(s_1)P(s_2|s_1)P(s_3|s_1s_2) \dots P(s_l|s_1s_2 \dots s_{l-1})$$

- Note that all the probabilities are being conditioned on the prefix of the the sequence s .
- The probabilistic suffix tree is a data structure which be used to efficiently calculate each term in the above formula.

Similarity Metric

- The similarity between a sequence and a set T is computed using the PST

$$Sim(s, T) = \log P(s_1) + \sum_{i=2}^l P(s_i | s_1 \dots s_{i-1})$$

- Report sequence with low similarity scores as outliers.
- One helpful insight is that outliers lie close to the root - can drastically truncate the PST and still recover outliers [14].

Advanced Topics

1. Finding “Heavy Hitters” in Streaming data
2. Using the “Kernel Trick” to find outliers in Hilbert Spaces

“Heavy Hitters” in streaming data

- Suppose in an Internet router we want to keep a counter of the traffic flowing at the source-destination granularity.
- Have n source-destination pairs $a = (a_1, a_2, \dots, a_n)$.
- At the end of the day we want to find all i such that $a_i > t$ where t is a user-defined threshold. These i 's are called the “Heavy Hitters”.
- We cannot explicitly store the vector a but only a synopsis or sketch \hat{a} which consumes sub-linear space.
- Instead of $a(i_t)$, $\hat{a}(i_t)$ is updated when $(i_t, c_t > 0)$ arrives.
- We also want to guarantee that $|a(i) - \hat{a}(i)| < \epsilon$ with high probability $1 - \delta$.

The Count-Min Sketch[4]

- We will use the Count-Min sketch though several others are available.
- The Count-Min sketch is a $d \times w$ matrix *COUNT* where d is the number of pair-wise independent hash functions each of size w .

$$\forall j = 1..d, h_j : \{1, 2, \dots, n\} \rightarrow \{1, 2, \dots, w\}$$

- As a new update (i_t, c_t) arrives, for each j

$$COUNT(j, h_j(i_t)) \leftarrow COUNT(j, h_j(i_t)) + c_t$$

- The estimate of a_i is

$$\hat{a}_i = \min_{1 \leq j \leq d} COUNT(j, h_j(i))$$

Main Result of Count-Min Sketch

Let $w = \lceil \frac{e}{\epsilon} \rceil$ and $d = \lceil \log(\frac{1}{\delta}) \rceil$. Then

Theorem 8 $a_i \leq \hat{a}_i \leq a_i + \epsilon \|a\|_1$ with probability at least $1 - \delta$.

- The space complexity of the sketch is therefore $O(\log(\frac{1}{\delta}) \frac{e}{\epsilon}) + O(2 * \log \frac{1}{\delta})$. Update time is $O(\log \frac{1}{\delta})$.
- The hash function $h_{a,b}$ are of the form $h_{a,b}(x) = (a + bx) \bmod(p)$ where p is a prime number bigger than n .
- For $\epsilon = 0.01$ and $\delta = 0.01$ the size of the sketch is approximately 1300 words.
- In practice this sketch works well when the distribution of the a'_i s is skewed (because then L_1 norm is small).

Analysis of the Count-Min Sketch

- Let

$$X_{ijk} = \begin{cases} 1 & h_j(i) = h_j(k) \text{ for } i \neq k \\ 0 & \text{otherwise} \end{cases}$$

- One can design simple hash functions which guarantee that

$$E(X_{ijk}) = P(X_{ijk} = 1) \leq \frac{1}{w}$$

Let $X_{ij} = \sum_{k=1}^n X_{ijk} a_k$. X_{ij} is non-negative.

Also by construction, $COUNT[j, h_j(i)] \geq a_i + X_{ij}$.

Now,

$$E[X_{ij}] = \sum_{k=1}^n E(X_{ijk} a_k) \leq \frac{1}{w} \|a\|_1$$

Analysis of the Count-Min Sketch

$$\begin{aligned}\forall j, P(\hat{a}_i > a_i + \epsilon \|a\|_1) &= \forall j, P(\text{COUNT}(j, h_j(i)) > a_i + \epsilon \|a\|_1) \\ &= \forall j, P(a_i + X_{ij} > a_i + \epsilon \|a\|_1) \\ &= \forall j, P(X_{ij} > \epsilon \|a\|_1) \\ &\leq \frac{E[X_{ij}]}{\epsilon \|a\|_1} \quad \forall j \\ &\leq \left(\frac{1}{w\epsilon}\right)^d \equiv \delta\end{aligned}$$

Now we want to choose w and d based on ϵ and δ . First select w as $\lceil \frac{e}{\epsilon} \rceil$. Then select d as $\lceil \log \frac{1}{\delta} \rceil$.

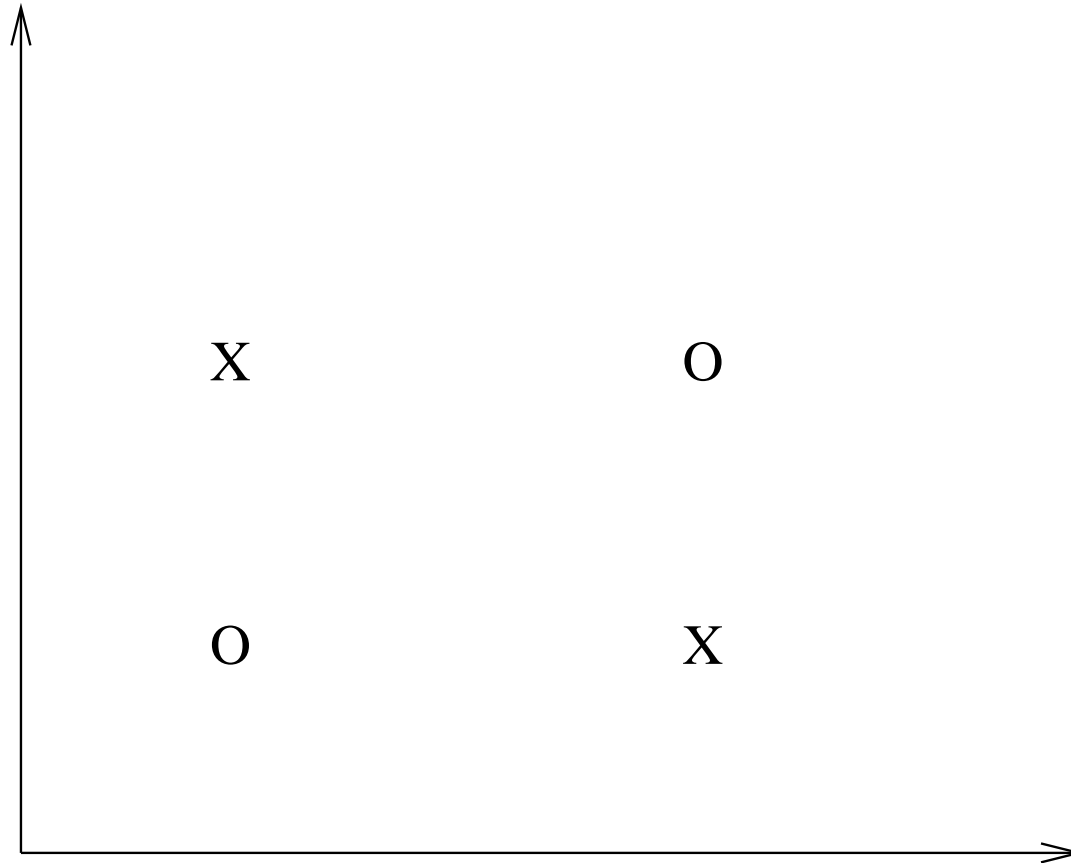
Kernel Methods for Outlier Detection

Kernel Methods are based on three core ideas

1. Data which is not linearly separable (in the original space) can be made so by “lifting” it into a higher dimensional space.
2. The inner product between two points in the higher dimensional space can be computed using a Kernel function K in the original lower dimensional space.
3. Many machine learning algorithms only require inner product information rather than the “data coordinates”.

XOR Example

Consider the XOR Function



This data is not linearly separable (i.e., there does not exist a straight line separating the Crosses(x) and the Circles(o).

XOR Example

- Let $y = \begin{cases} 1 & \text{if } (x_1 = 0 \wedge x_2 = 0) \vee (x_1 = 1 \wedge x_2 = 1) \\ -1 & \text{otherwise} \end{cases}$
- This is the XOR function.
- Define a function ϕ on the space $X = (x_1, x_2)$ as $\phi(x_1, x_2) = (x_1, x_2, x_1x_2)$
- Can check that image of X under ϕ is linearly separable

Kernel Functions

- Let X be an input space. Let $\phi(X)$ be a mapping to a higher dimensional space Hilbert space. If there exists a function $K : X \times X \rightarrow R$ such that

$$K(x, y) = \langle \phi(x), \phi(y) \rangle$$

then roughly speaking K is kernel function.

- All calculation which depends on the inner product (in the ϕ space) can be done using K .
- X does not have to be a vector space (could be set of graphs, strings, sequences etc.).
- Example: Let D be a set. Let A_1 and A_2 be subsets of D . Then

$$K(A_1, A_2) = 2^{|A_1 \cap A_2|}$$

is a valid kernel function.

Kernel Properties

- For example the norm of $\phi(x)$

$$\|\phi(x)\|_2 = \langle \phi(x), \phi(x) \rangle^{\frac{1}{2}} = K(x, x)^{\frac{1}{2}}$$

- The distance between $\phi(x)$ and $\phi(y)$

$$\|\phi(x) - \phi(z)\|^2 = k(x, x) - 2k(x, z) + k(z, z)$$

- Most importantly, let $\phi_S = \frac{1}{l} \sum_{i=1}^l \phi(x_i)$. Then

$$\|\phi(x) - \phi_S\|^2 = k(x, x) + \frac{1}{l^2} \sum_{i,j=1}^l k(x_i, x_j) - \frac{2}{l} \sum_{i=1}^l k(x, x_i)$$

Outlier Detection with Kernels

Recall from Slide 12 that

$$P(|x_{l+1} - s| > \max |x_i - s| + 2f(R, \delta, l)) < \frac{1}{l+1}$$

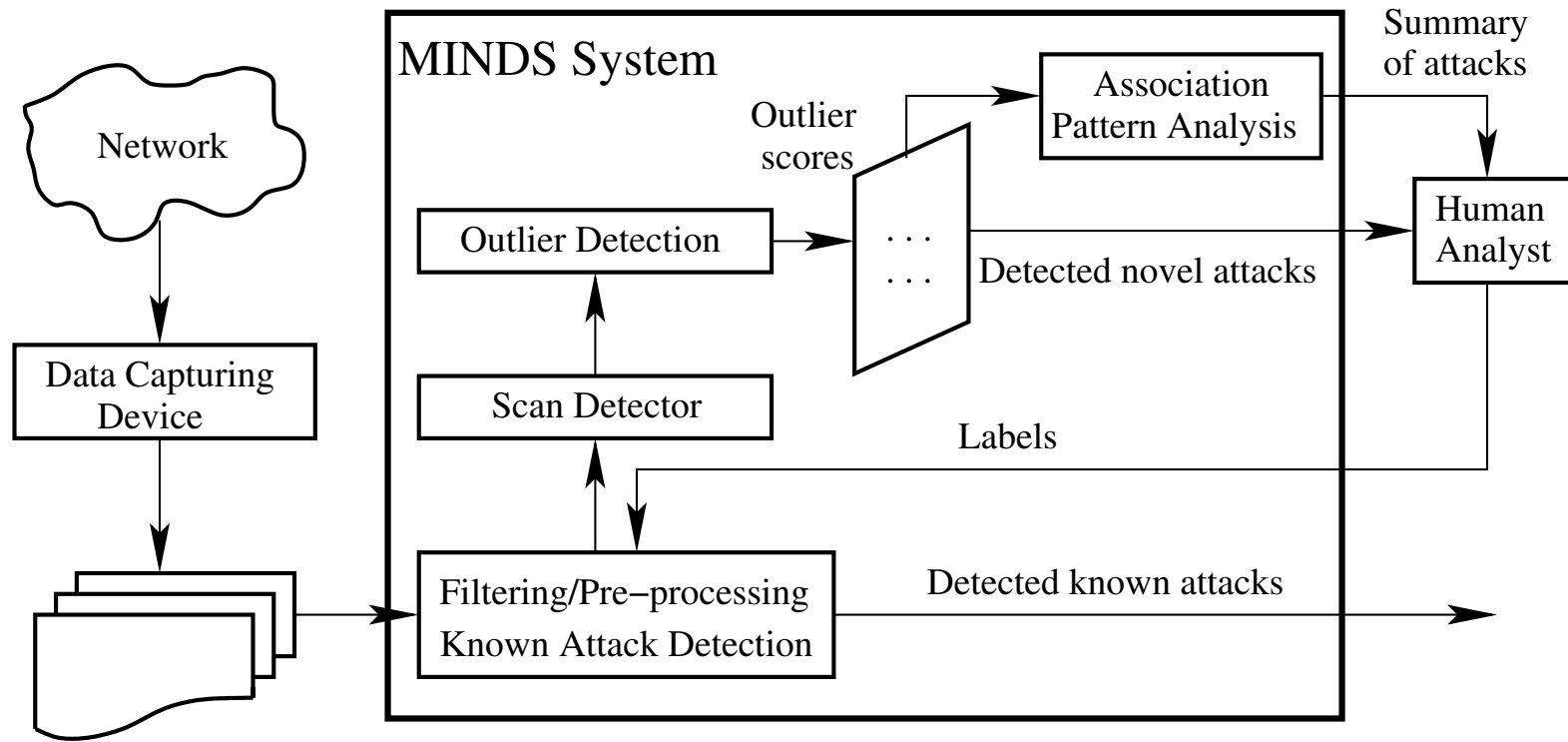
This can be easily “kernelized”

$$P(|\phi(x)_{l+1} - \phi_s| > \max |\phi(x)_i - \phi_s| + 2f(R, \delta, l)) < \frac{1}{l+1}$$

We now have an outlier detection algorithm for arbitrary spaces!!

The challenge now becomes to design a “domain specific” kernel which captures the right properties.

MINDS: Minnesota Intrusion Detection System[8]



MINDS uses LOF for Outlier Detection

Southern Oscillation Index (SOI)[9]

- SOI is the normalized sea surface pressure difference between Tahiti and Darwin Australia.
- Sir Gilbert Walker was the first to notice the relationship between SOI and global climate.
- SOI values of 2 standard deviation below the mean are related to El Nino.
- Helped explain drought in Asia and unusually wet weather in the Americas.
- *I cannot help believing that we shall gradually find out the physical mechanism by which these [relationships] are maintained ...*
Sir Gilbert T. Walker, 1918

References

- [1] C. C. Aggarwal and P. S. Yu. Outlier detection for high dimensional data. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, Santa Barbara, California, USA, 2001*.
- [2] S. D. Bay and M. Schwabacher. Mining distance-based outliers in near linear time with randomization and a simple pruning rule. In *Proceedings of Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, D.C. USA, pages 29–38, 2003*.
- [3] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander. Lof: Identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, Dallas, Texas, USA, pages 93–104. ACM, 2000*.
- [4] G. Cormode and S. Muthukrishnan. An improved data stream summary: The count-min sketch and its applications. In *LATIN, pages 29–38, 2004*.
- [5] D. Hawkins. *Identification of Outliers*. Chapman and Hall, London, 1980.

- [6] E. M. Knorr and R. T. Ng. Finding intensional knowledge of distance-based outliers. In *Proceedings of 25th International Conference on Very Large Data Bases*, pages 211–222. Morgan Kaufmann, 1999.
- [7] G. Kollios, D. Gunopulos, N. Koudas, and S. Berchtold. Efficient biased sampling for approximate clustering and outlier detection in large datasets. *IEEE Transactions on Knowledge and Data Engineering*, 2003.
- [8] A. Lazarevic, L. Ertöz, V. Kumar, A. Ozgur, and J. Srivastava. A comparative study of anomaly detection schemes in network intrusion detection. In *SDM*, 2003.
- [9] M. McPhadden. El nino and la nina: Causes and global consequences. In *Encyclopedia of Global Environmental Change*, pages 353–370, 2002.
- [10] S. Ramaswamy, R. Rastogi, and K. Shim. Efficient algorithms for mining outliers from large data sets. In *SIGMOD Conference*, pages 427–438, 2000.
- [11] P. J. Rousseeuw and K. V. Driessen. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41:212–223, 1999.
- [12] J. Shawne-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge, 2005.

- [13] S. Shekhar, C.-T. Lu, and P. Zhang. Detecting graph-based spatial outliers: algorithms and applications (a summary of results). In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, San Francisco, CA, USA. ACM, 2001*, pages 371–376, 2001.
- [14] P. Sun, S. Chawla, and B. Arunasalam. Outlier detection in sequential databases. In *Proceedings of SIAM International Conference on Data Mining, 2006*.