Data Mining for Anomaly Detection

Aleksandar Lazarevic

United Technologies Research Center

Arindam Banerjee, Varun Chandola,
Vipin Kumar, Jaideep Srivastava

University of Minnesota

Tutorial at the European Conference on Principles
and Practice of Knowledge Discovery in Databases

Antwerp, Belgium, September 19, 2008

Outline

- Introduction
- Aspects of Anomaly Detection Problem
- Applications
- Different Types of Anomaly Detection Techniques
- Case Study
- Discussion and Conclusions

Introduction

- We are drowning in the deluge of data that are being collected world-wide, while starving for knowledge at the same time*
- Anomalous events occur relatively infrequently
- However, when they do occur, their consequences can be quite dramatic and quite often in a negative sense

"Mining needle in a haystack. So much hay and so little time"

* - J. Naisbitt, Megatrends: Ten New Directions Transforming Our Lives. New York: Warner Books, 1982.

What are Anomalies?

- Anomaly is a pattern in the data that does not conform to the expected behavior
- Also referred to as outliers, exceptions, peculiarities, surprise, etc.
- Anomalies translate to significant (often critical) real life entities
  - Cyber intrusions
  - Credit card fraud
  - Faults in mechanical systems

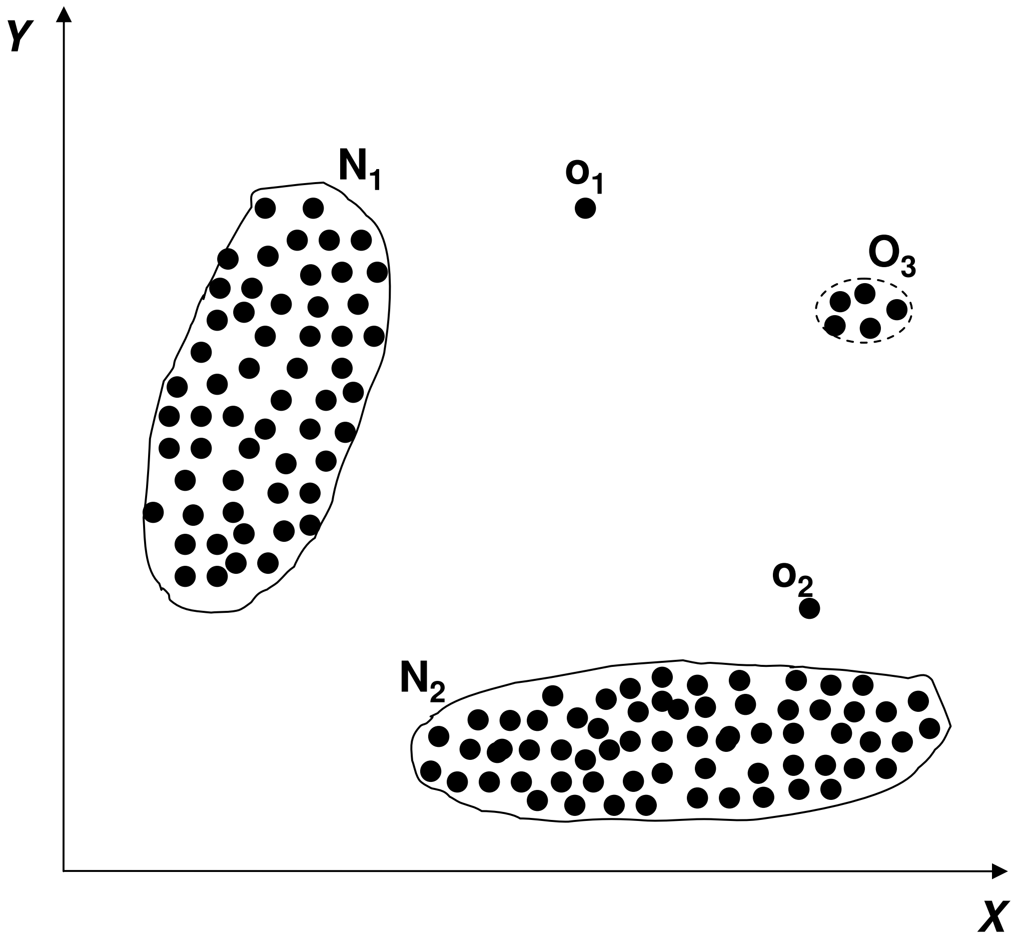
Real World Anomalies

- Credit Card Fraud
  - An abnormally high purchase made on a credit card

- Cyber Intrusions
  - A web server involved in *ftp* traffic

Simple Examples

- N_1 and N_2 are regions of normal behavior
- Points o_1 and o_2 are anomalies
- Points in region O_3 are anomalies



Related problems

- Rare Class Mining
- Chance discovery
- Novelty Detection
- Exception Mining
- Noise Removal
- Black Swan*

* N. Taleb, The Black Swan: The Impact of the Highly Probable?, 2007

Key Challenges

- Defining a representative normal region is challenging
- The boundary between normal and outlying behavior is often not precise
- Availability of labeled data for training/validation
- The exact notion of an outlier is different for different application domains
- Malicious adversaries
- Data might contain noise
- Normal behavior keeps evolving

Aspects of Anomaly Detection Problem

- Nature of input data
- Availability of supervision
- Type of anomaly: point, contextual, structural
- Output of anomaly detection
- Evaluation of anomaly detection techniques

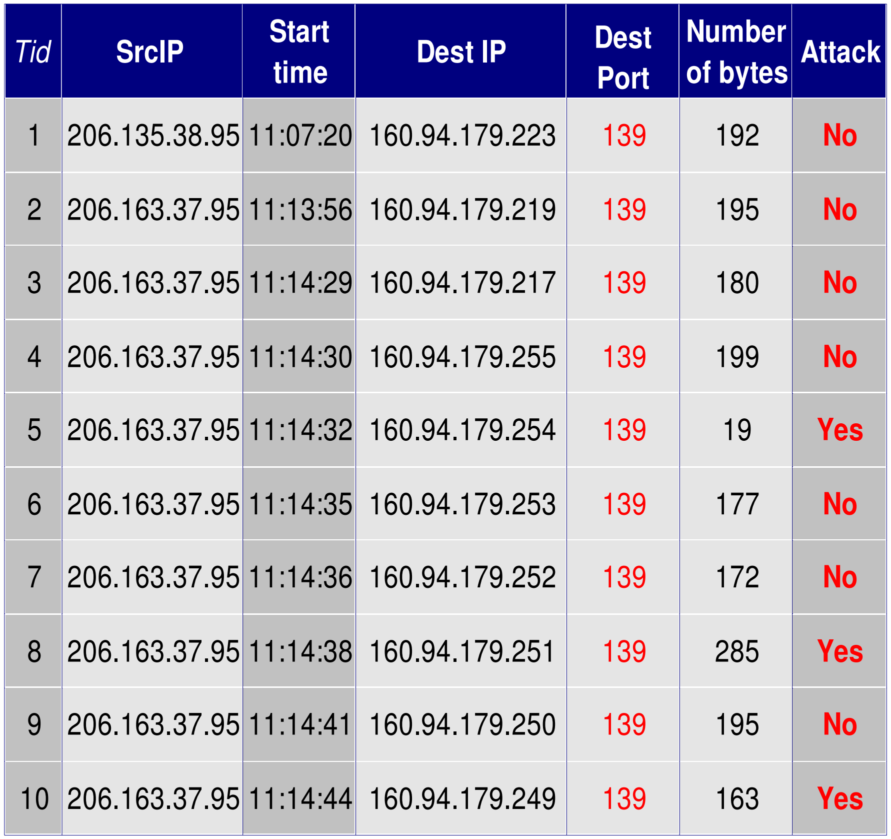
Input Data

- Most common form of data handled by anomaly detection techniques is *Record Data*
  - Univariate
  - Multivariate

Engine Temperature
192
195
180
199
19
177
172
285
195
163

Input Data

- Most common form of data handled by anomaly detection techniques is *Record Data*
  - Univariate
  - Multivariate



Tid	SrcIP	Start time	Dest IP	Dest Port	Number of bytes	Attack
1	206.135.38.95	11:07:20	160.94.179.223	139	192	No
2	206.163.37.95	11:13:56	160.94.179.219	139	195	No
3	206.163.37.95	11:14:29	160.94.179.217	139	180	No
4	206.163.37.95	11:14:30	160.94.179.255	139	199	No
5	206.163.37.95	11:14:32	160.94.179.254	139	19	Yes
6	206.163.37.95	11:14:35	160.94.179.253	139	177	No
7	206.163.37.95	11:14:36	160.94.179.252	139	172	No
8	206.163.37.95	11:14:38	160.94.179.251	139	285	Yes
9	206.163.37.95	11:14:41	160.94.179.250	139	195	No
10	206.163.37.95	11:14:44	160.94.179.249	139	163	Yes

Input Data – *Nature of Attributes*

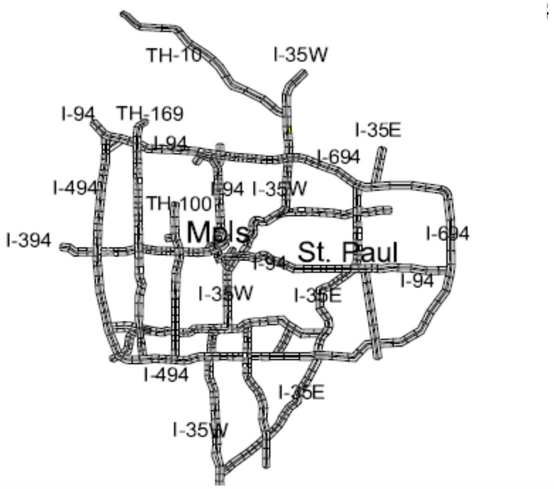
- Nature of attributes
  - Binary
  - Categorical
  - Continuous
  - Hybrid

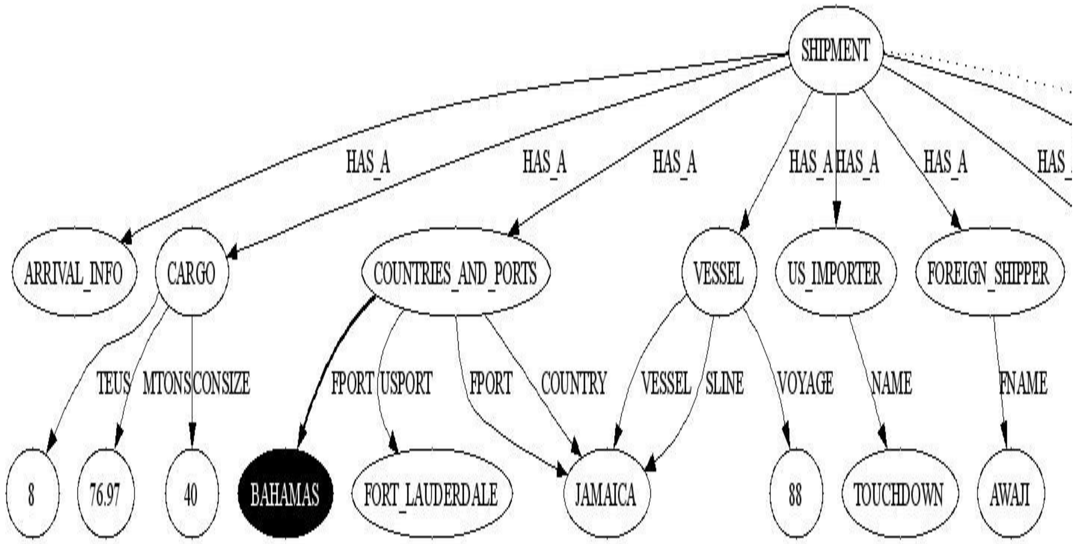
	categorical	continuous	categorical	continuous	binary
Tid	**SrcIP**	**Duration**	**Dest IP**	**Number of bytes**	**Internal**
1	206.163.37.81	0.10	160.94.179.208	150	**No**
2	206.163.37.99	0.27	160.94.179.235	208	**No**
3	160.94.123.45	1.23	160.94.179.221	195	**Yes**
4	206.163.37.37	112.03	160.94.179.253	199	**No**
5	206.163.37.41	0.32	160.94.179.244	181	**No**

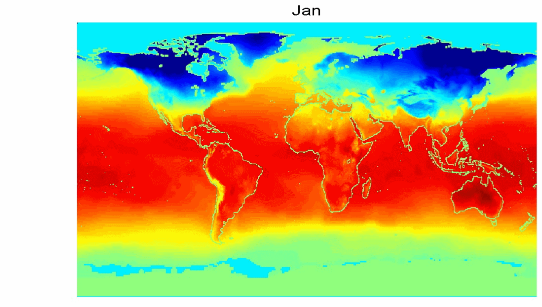
Input Data – *Complex Data Types*

- Relationship among data instances
  - Sequential
    - Temporal
  - Spatial
  - Spatio-temporal
  - Graph

```
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
```







United Technologies

University of Minnesota

Data Labels

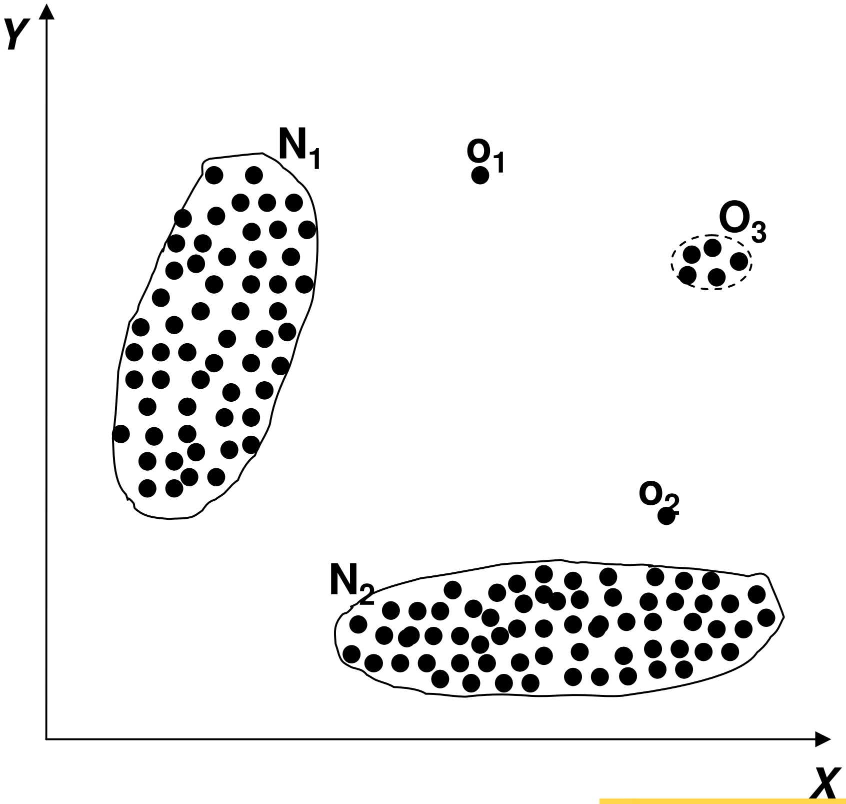
- Supervised Anomaly Detection
  - Labels available for both normal data and anomalies
  - Similar to rare class mining
- Semi-supervised Anomaly Detection
  - Labels available only for normal data
- Unsupervised Anomaly Detection
  - No labels assumed
  - Based on the assumption that anomalies are very rare compared to normal data

Type of Anomaly

- Point Anomalies
- Contextual Anomalies
- Collective Anomalies

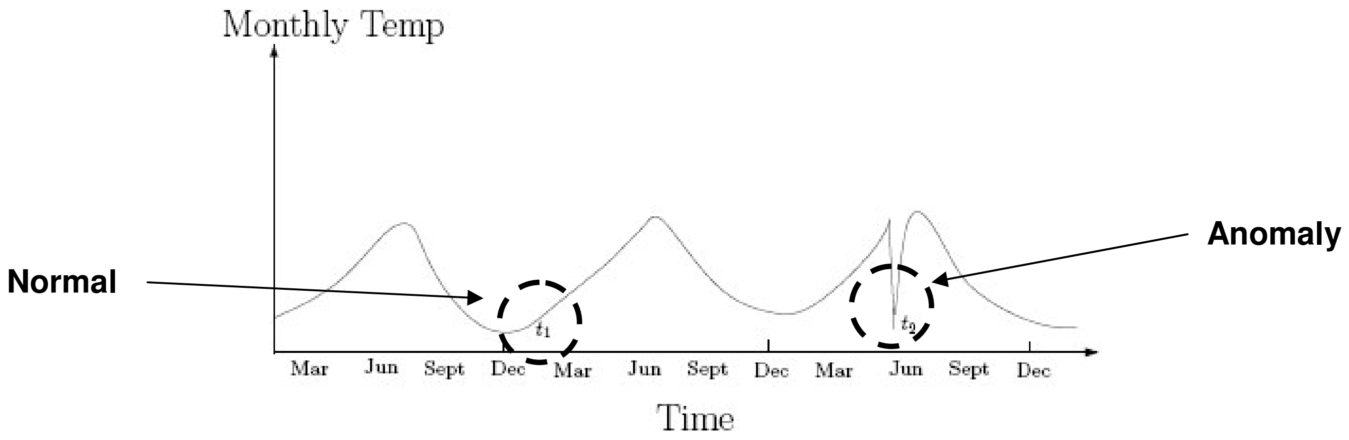
Point Anomalies

- An individual data instance is anomalous w.r.t. the data



Contextual Anomalies

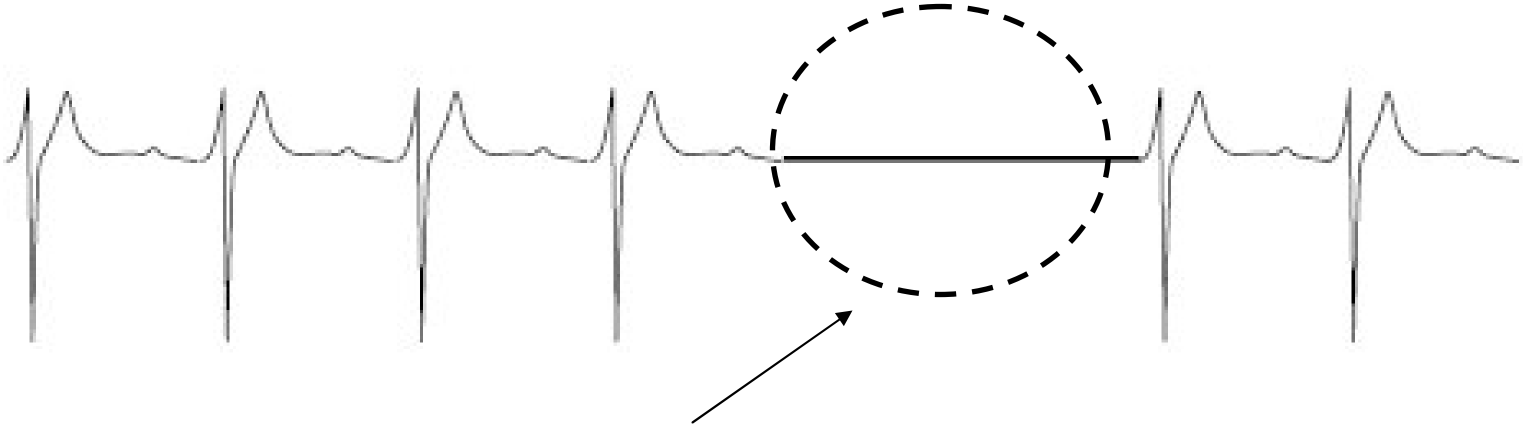
- An individual data instance is anomalous within a context
- Requires a notion of context
- Also referred to as conditional anomalies*



* Xiuyao Song, Mingxi Wu, Christopher Jermaine, Sanjay Ranka, Conditional Anomaly Detection, IEEE Transactions on Data and Knowledge Engineering, 2006.

Collective Anomalies

- A collection of related data instances is anomalous
- Requires a relationship among data instances
  - Sequential Data
  - Spatial Data
  - Graph Data
- The individual instances within a collective anomaly are not anomalous by themselves



Anomalous Subsequence

Output of Anomaly Detection

- Label
  - Each test instance is given a *normal* or *anomaly* label
  - This is especially true of classification-based approaches
- Score
  - Each test instance is assigned an anomaly score
    - Allows the output to be ranked
    - Requires an additional threshold parameter

Evaluation of Anomaly Detection – F-value

- Accuracy is not sufficient metric for evaluation
  - Example: network traffic data set with 99.9% of normal data and 0.1% of intrusions
  - Trivial classifier that labels everything with the normal class can achieve 99.9% accuracy !!!!!

Confusion matrix		Predicted class	
		NC	C
Actual class	NC	TN	FP
	C	FN	TP

anomaly class – C

normal class – NC

- **Focus on both recall and precision**
  - **Recall (R) = TP/(TP + FN)**
  - **Precision (P) = TP/(TP + FP)**
- **F – measure = 2*R*P/(R+P) =** $\frac{(1+\beta^2)\cdot R\cdot P}{\beta^2\cdot P+R}$

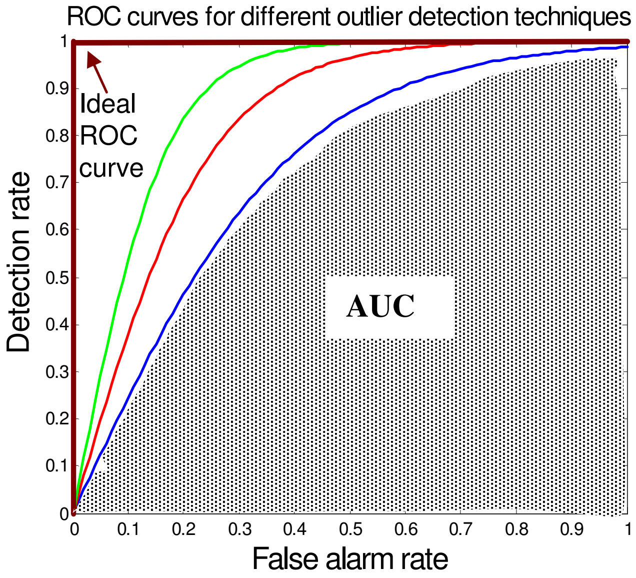
Evaluation of Outlier Detection – ROC & AUC

Confusion matrix		Predicted class	
		NC	C
Actual class	NC	TN	FP
	C	FN	TP

anomaly class – C

normal class – NC

- Standard measures for evaluating anomaly detection problems:
  - *Recall (Detection rate)* - ratio between the number of correctly detected anomalies and the total number of anomalies
  - *False alarm (false positive) rate* – ratio between the number of data records from normal class that are misclassified as anomalies and the total number of data records from normal class
  - *ROC Curve* is a trade-off between detection rate and false alarm rate
  - *Area under the ROC curve (AUC)* is computed using a trapezoid rule



Applications of Anomaly Detection

- Network intrusion detection
- Insurance / Credit card fraud detection
- Healthcare Informatics / Medical diagnostics
- Industrial Damage Detection
- Image Processing / Video surveillance
- Novel Topic Detection in Text Mining
- …

Intrusion Detection

- Intrusion Detection:
  - Process of monitoring the events occurring in a computer system or network and analyzing them for intrusions
  - Intrusions are defined as attempts to bypass the security mechanisms of a computer or network
- Challenges
  - Traditional signature-based intrusion detection systems are based on signatures of known attacks and cannot detect emerging cyber threats
  - Substantial latency in deployment of newly created signatures across the computer system
- Anomaly detection can alleviate these limitations

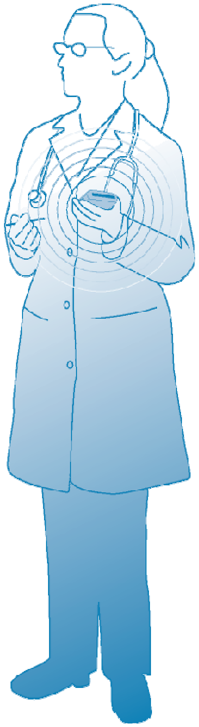
Fraud Detection

- Fraud detection refers to detection of criminal activities occurring in commercial organizations
  - Malicious users might be the actual customers of the organization or might be posing as a customer (also known as identity theft).
- Types of fraud
  - Credit card fraud
  - Insurance claim fraud
  - Mobile / cell phone fraud
  - Insider trading

- Challenges
  - Fast and accurate real-time detection
  - Misclassification cost is very high

Healthcare Informatics

- Detect anomalous patient records
  - Indicate disease outbreaks, instrumentation errors, etc.
- Key Challenges
  - Only normal labels available
  - Misclassification cost is very high
  - Data can be complex: spatio-temporal



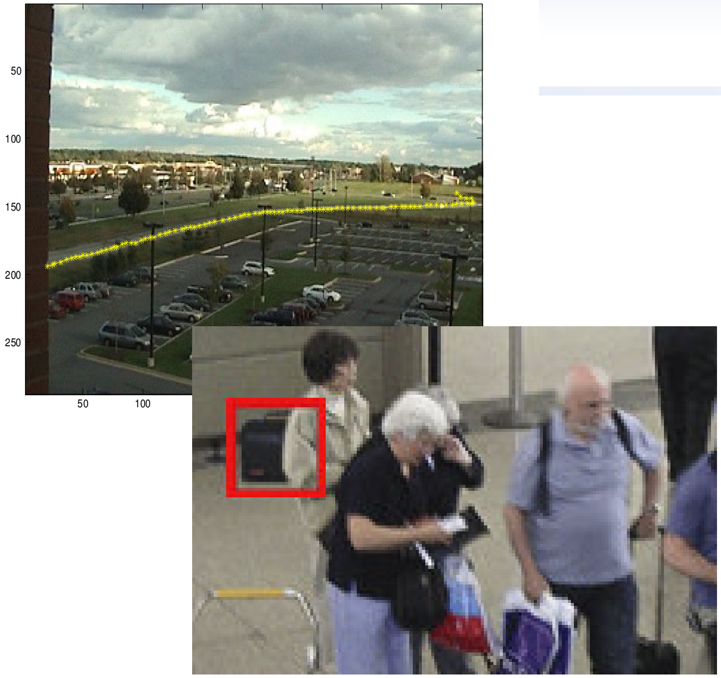
Industrial Damage Detection

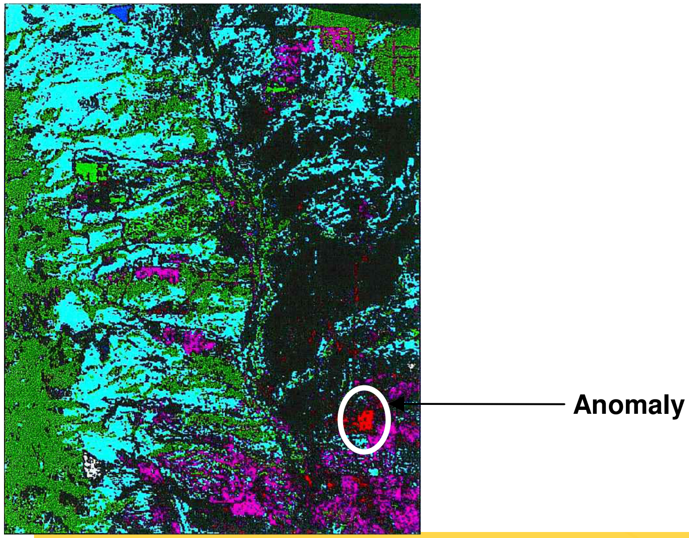
- Industrial damage detection refers to detection of different faults and failures in complex industrial systems, structural damages, intrusions in electronic security systems, suspicious events in video surveillance, abnormal energy consumption, etc.
  - Example: Aircraft Safety
    - Anomalous Aircraft (Engine) / Fleet Usage
    - Anomalies in engine combustion data
    - Total aircraft health and usage management

- Key Challenges
  - Data is extremely huge, noisy and unlabelled
  - Most of applications exhibit temporal behavior
  - Detecting anomalous events typically require immediate intervention

Image Processing

- Detecting outliers in a image monitored over time
- Detecting anomalous regions within an image
- Used in
  - mammography image analysis
  - video surveillance
  - satellite image analysis
- Key Challenges
  - Detecting collective anomalies
  - Data sets are very large

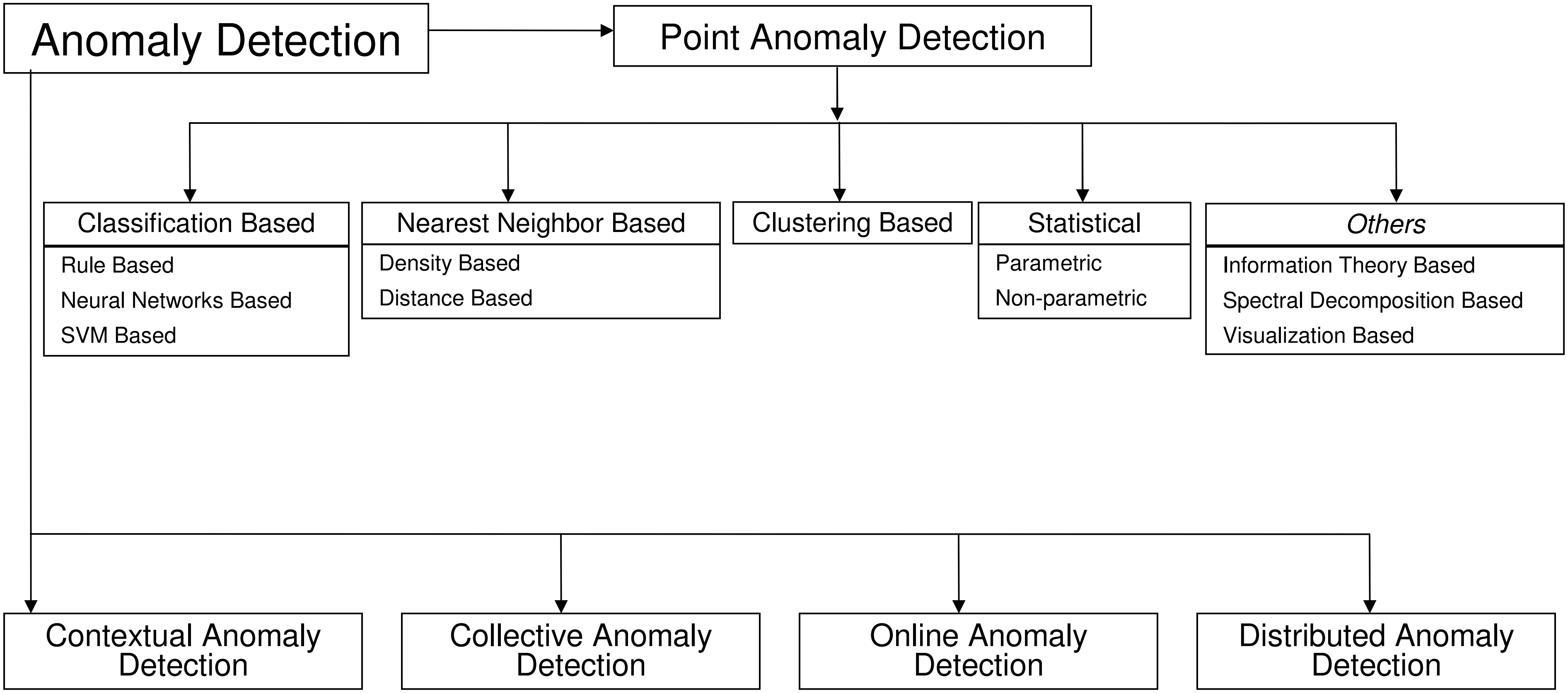




United Technologies

University of Minnesota

Taxonomy*



Anomaly Detection

Point Anomaly Detection

Classification Based
- Rule Based
- Neural Networks Based
- SVM Based

Nearest Neighbor Based
- Density Based
- Distance Based

Clustering Based

Statistical
- Parametric
- Non-parametric

Others
- Information Theory Based
- Spectral Decomposition Based
- Visualization Based

Contextual Anomaly Detection

Collective Anomaly Detection

Online Anomaly Detection

Distributed Anomaly Detection

* Anomaly Detection – A Survey, Varun Chandola, Arindam Banerjee, and Vipin Kumar, To Appear in ACM Computing Surveys 2008.

Classification Based Techniques

- Main idea: build a classification model for normal (and anomalous (rare)) events based on labeled training data, and use it to classify each new unseen event
- Classification models must be able to handle skewed (imbalanced) class distributions
- Categories:
  - *Supervised classification techniques*
    - Require knowledge of both ***normal*** and ***anomaly*** class
    - Build classifier to distinguish between normal and known anomalies
  - *Semi-supervised classification techniques*
    - Require knowledge of ***normal*** class only!
    - Use modified classification model to learn the normal behavior and then detect any deviations from normal behavior as anomalous

Classification Based Techniques

- Advantages:
  - *Supervised classification techniques*
    - Models that can be easily understood
    - High accuracy in detecting many kinds of known anomalies
  - *Semi-supervised classification techniques*
    - Models that can be easily understood
    - Normal behavior can be accurately learned
- Drawbacks:
  - *Supervised classification techniques*
    - Require both labels from both normal and anomaly class
    - Cannot detect unknown and emerging anomalies
  - *Semi-supervised classification techniques*
    - Require labels from normal class
    - Possible high false alarm rate - previously unseen (yet legitimate) data records may be recognized as anomalies

Supervised Classification Techniques

- Manipulating data records (oversampling / undersampling / generating artificial examples)
- Rule based techniques
- Model based techniques
  - Neural network based approaches
  - Support Vector machines (SVM) based approaches
  - Bayesian networks based approaches
- Cost-sensitive classification techniques
- Ensemble based algorithms (SMOTEBoost, RareBoost

Manipulating Data Records

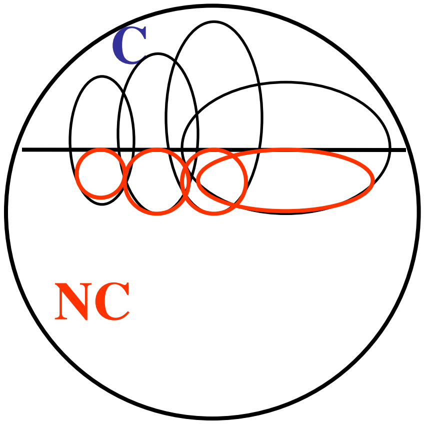
- **Over-sampling the rare class** [Ling98]
  - Make the duplicates of the rare events until the data set contains as many examples as the majority class => balance the classes
  - Does not increase information but increase misclassification cost
- **Down-sizing (undersampling) the majority class** [Kubat97]
  - Sample the data records from majority class (Randomly, Near miss examples, Examples far from minority class examples (far from decision boundaries)
  - Introduce sampled data records into the original data set instead of original data records from the majority class
  - Usually results in a general loss of information and overly general rules
- **Generating artificial anomalies**
  - SMOTE (Synthetic Minority Over-sampling TEchnique) [Chawla02] - new rare class examples are generated inside the regions of existing rare class examples
  - Artificial anomalies are generated around the edges of the sparsely populated data regions [Fan01]
  - Classify synthetic outliers vs. real normal data using active learning [Abe06]

Rule Based Techniques

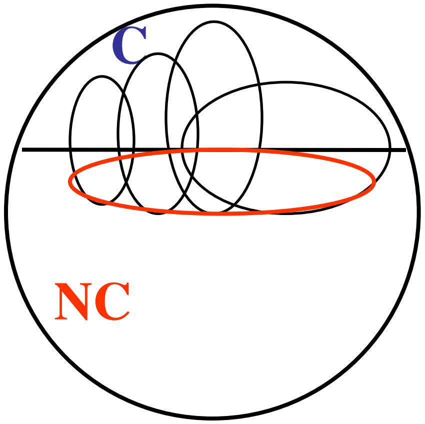
- **Creating new rule based algorithms (PN-rule, CREDOS)**
- **Adapting existing rule based techniques**
  - Robust C4.5 algorithm [John95]
  - Adapting multi-class classification methods to single-class classification problem
- **Association rules**
  - Rules with support higher than pre specified threshold may characterize normal behavior [Barbara01, Otey03]
  - Anomalous data record occurs in fewer frequent itemsets compared to normal data record [He04]
  - Frequent episodes for describing temporal normal behavior [Lee00,Qin04]
- **Case specific feature/rule weighting**
  - Case specific feature weighting [Cardey97] - Decision tree learning, where for each rare class test example replace global weight vector with dynamically generated weight vector that depends on the path taken by that example
  - Case specific rule weighting [Grzymala00] - LERS (Learning from Examples based on Rough Sets) algorithm increases the rule strength for all rules describing the rare class

New Rule-based Algorithms: PN-rule Learning*

- ***P-phase*:**
  - **cover most of the positive examples with high support**
  - **seek good recall**
- ***N-phase*:**
  - **remove FP from examples covered in P-phase**
  - **N-rules give high accuracy and significant support**



Existing techniques can possibly learn erroneous small signatures for absence of C

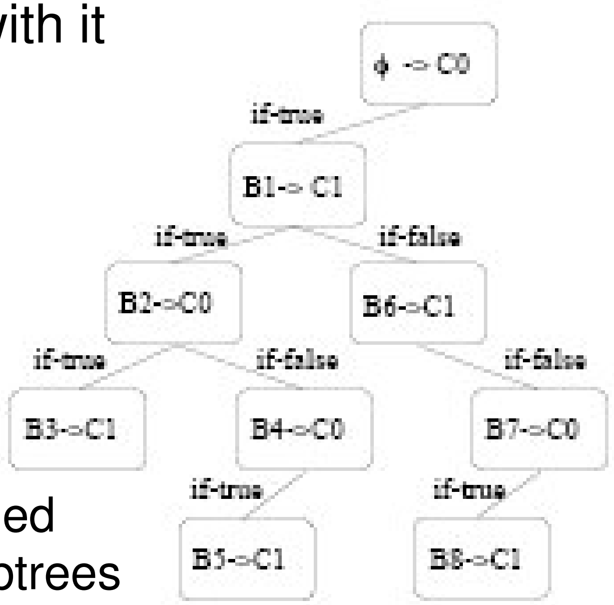


PNrule can learn strong signatures for presence of NC in *N-phase*

* M. Joshi, et al., PNrule, Mining Needles in a Haystack: Classifying Rare Classes via Two-Phase Rule Induction, ACM SIGMOD 2001

New Rule-based Algorithms: CREDOS*

- Ripple Down Rules (RDRs) can be represented as a decision tree where each node has a predictive rule associated with it
- RDRs specialize a generic form of multi-phase PNrule model
- Two phases: growth and pruning
- Growth phase:
  - Use RDRs to overfit the training data
  - Generate a binary tree where each node is characterized by the rule R_h, a default class and links to two child subtrees
  - Grow the RDS structure in a recursive manner
- Prune the structure to improve generalization
  - Different mechanism from decision trees



* M. Joshi, et al., CREDOS: Classification Using Ripple Down Structure (A Case for Rare Classes), SIAM International Conference on Data Mining, (SDM'04), 2004.

Using Neural Networks

- Multi-layer Perceptrons
  - Measuring the activation of output nodes [Augusteijn02]
  - Extending the learning beyond decision boundaries
    - Equivalent error bars as a measure of confidence for classification [Sykacek97]
    - Creating hyper-planes for separating between various classes, but also to have flexible boundaries where points far from them are outliers [Vasconcelos95]
- Auto-associative neural networks
  - Replicator NNs [Hawkins02]
  - Hopfield networks [Jagota91, Crook01]
- Adaptive Resonance Theory based [Dasgupta00, Caudel93]
- Radial Basis Functions based
  - Adding reverse connections from output to central layer allows each neuron to have associated normal distribution, and any new instance that does not fit any of these distributions is an anomaly [Albrecht00, Li02]
- Oscillatory networks
  - Relaxation time of oscillatory NNs is used as a criterion for novelty detection when a new instance is presented [Ho98, Borisyuk00]

Using Support Vector Machines

- SVM Classifiers [Steinwart05,Mukkamala02]
- Main idea [Steinwart05] :
  - Normal data records belong to high density data regions
  - Anomalies belong to low density data regions
  - Use unsupervised approach to learn high density and low density data regions
  - Use SVM to classify data density level
- Main idea: [Mukkamala02]
  - Data records are labeled (normal network behavior vs. intrusive)
  - Use standard SVM for classification

* A. Lazarevic, et al., A Comparative Study of Anomaly Detection Schemes in Network Intrusion Detection, SIAM 2003

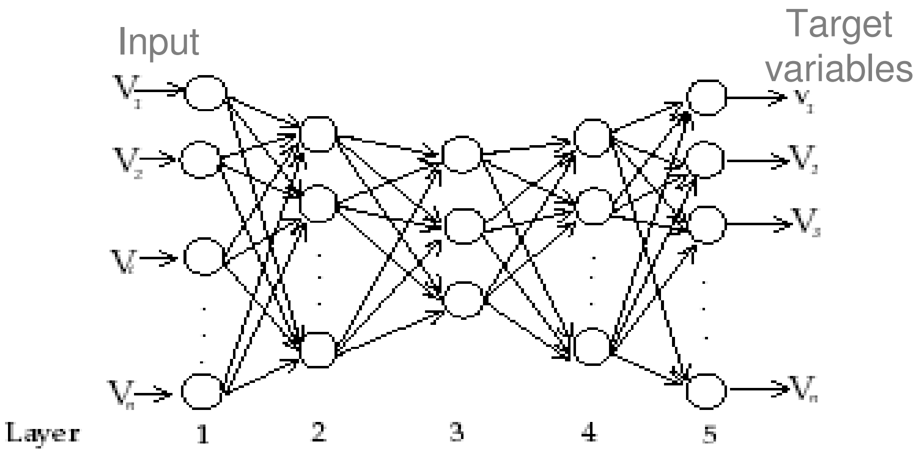
Semi-supervised Classification Techniques

- Use modified classification model to learn the normal behavior and then detect any deviations from normal behavior as anomalous

- Recent approaches:
  - Neural network based approaches
  - Support Vector machines (SVM) based approaches
  - Markov model based approaches
  - Rule-based approaches

Using Replicator Neural Networks*

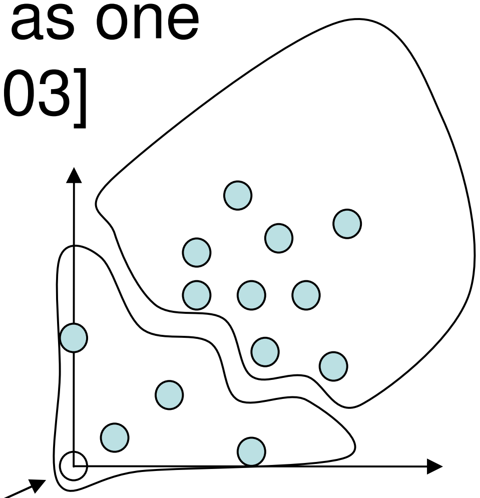
- Use a replicator 4-layer feed-forward neural network (RNN) with the same number of input and output nodes
- Input variables are the output variables so that RNN forms a compressed model of the data during training
- A measure of outlyingness is the reconstruction error of individual data points.



* S. Hawkins, et al. Outlier detection using replicator neural networks, DaWaK02 2002.

Using Support Vector Machines

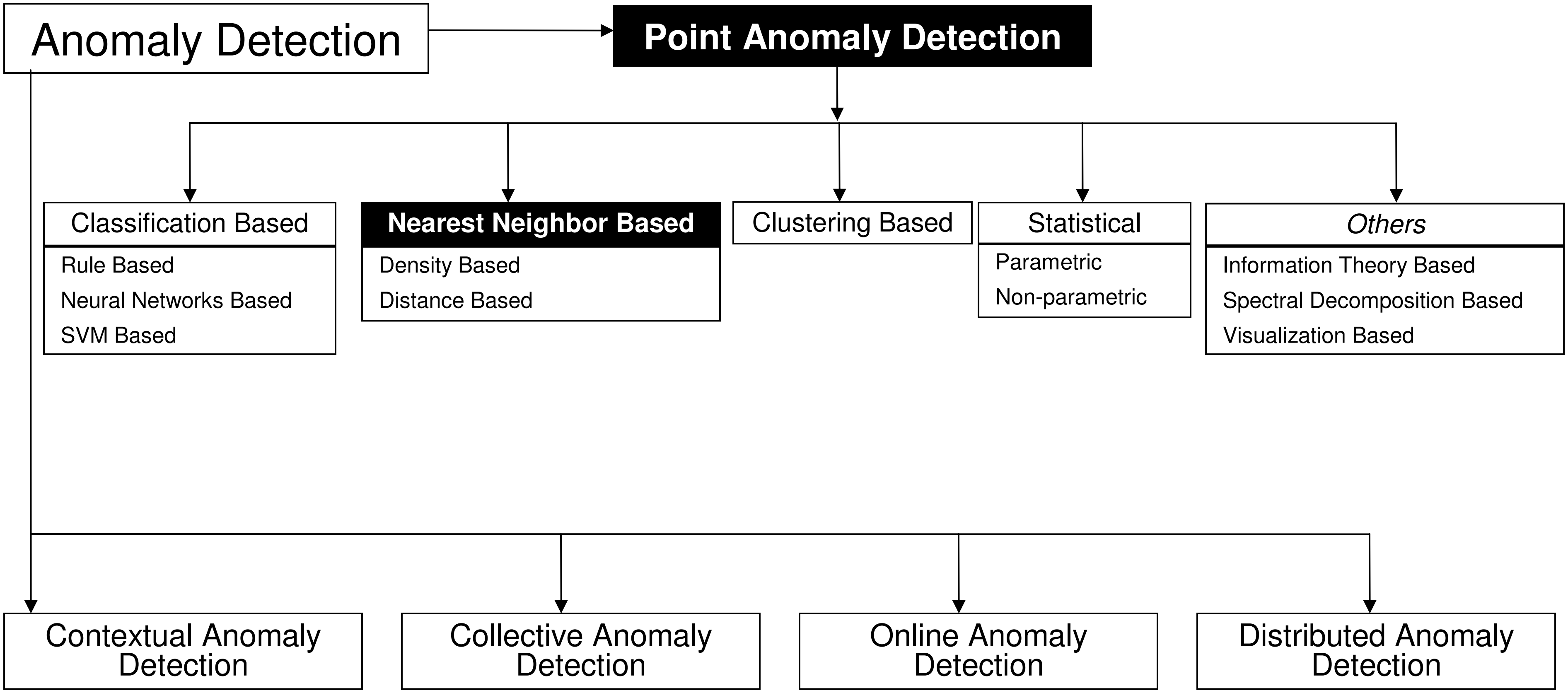
- Converting into one class classification problem
  - Separate the entire set of training data from the origin, i.e. to find a small region where most of the data lies and label data points in this region as one class [Ratsch02, Tax01, Eskin02, Lazarevic03]
    - Parameters
      - Expected number of outliers
      - Variance of rbf kernel (As the variance of the rbf kernel gets smaller, the number of support vectors is larger and the separating surface gets more complex)
  - Separate regions containing data from the regions containing no data [Scholkopf99]



origin

push the hyper plane away from origin as much as possible

Taxonomy



Anomaly Detection

Point Anomaly Detection

Classification Based
- Rule Based
- Neural Networks Based
- SVM Based

Nearest Neighbor Based
- Density Based
- Distance Based

Clustering Based

Statistical
- Parametric
- Non-parametric

Others
- Information Theory Based
- Spectral Decomposition Based
- Visualization Based

Contextual Anomaly Detection

Collective Anomaly Detection

Online Anomaly Detection

Distributed Anomaly Detection

Nearest Neighbor Based Techniques

- *Key assumption*: normal points have close neighbors while anomalies are located far from other points
- General two-step approach
  1. Compute neighborhood for each data record
  2. Analyze the neighborhood to determine whether data record is anomaly or not
- Categories:
  - Distance based methods
    - Anomalies are data points most distant from other points
  - Density based methods
    - Anomalies are data points in low density regions

Nearest Neighbor Based Techniques

- Advantage
  - Can be used in unsupervised or semi-supervised setting (do not make any assumptions about data distribution)
- Drawbacks
  - If normal points do not have sufficient number of neighbors the techniques may fail
  - Computationally expensive
  - In high dimensional spaces, data is sparse and the concept of similarity may not be meaningful anymore. Due to the sparseness, distances between any two data records may become quite similar => Each data record may be considered as potential outlier!

Nearest Neighbor Based Techniques

- Distance based approaches
  - A point *O* in a dataset is an *DB*(*p*, *d*) outlier if at least fraction *p* of the points in the data set lies greater than distance *d* from the point *O**
- Density based approaches
  - Compute local densities of particular regions and declare instances in low density regions as potential anomalies
  - Approaches
    - Local Outlier Factor (LOF)
    - Connectivity Outlier Factor (COF↵
    - Multi-Granularity Deviation Factor (MDEF)

*Knorr, Ng,Algorithms for Mining Distance-Based Outliers in Large Datasets, VLDB98

Distance based Outlier Detection

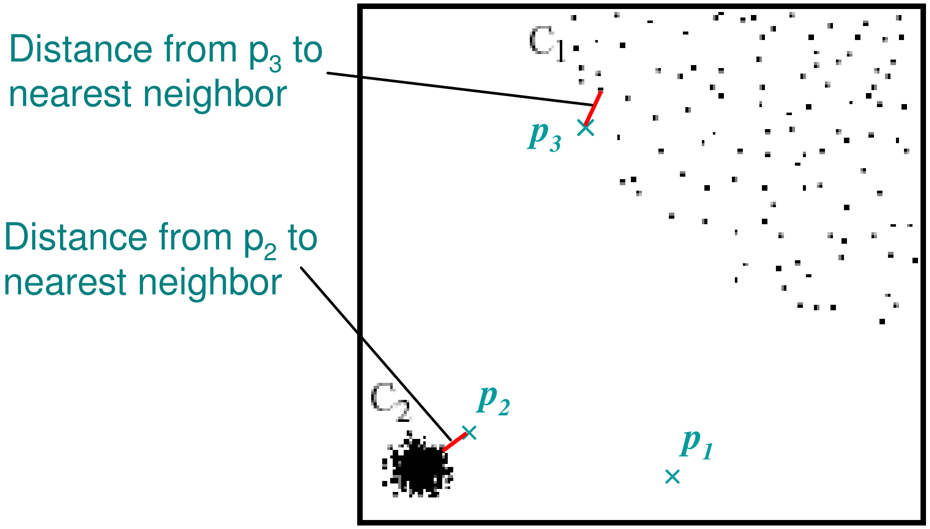
- *Nearest Neighbor (NN) approach*[*,**]
  - For each data point *d* compute the distance to the *k-th* nearest neighbor d_k
  - Sort all data points according to the distance d_k
  - Outliers are points that have the largest distance d_k and therefore are located in the more sparse neighborhoods
  - Usually data points that have top *n*% distance d_k are identified as outliers
    - *n* – user parameter
  - Not suitable for datasets that have modes with varying density

[*] Knorr, Ng,Algorithms for Mining Distance-Based Outliers in Large Datasets, VLDB98

[**] S. Ramaswamy, R. Rastogi, S. Kyuseok: Efficient Algorithms for Mining Outliers from Large Data Sets, ACM SIGMOD Conf. On Management of Data, 2000.

Advantages of Density based Techniques

- *Local Outlier Factor (LOF) approach*
  - Example:



Distance from p_3 to nearest neighbor

Distance from p_2 to nearest neighbor

In the *NN* approach, p_2 is not considered as outlier, while the *LOF* approach find both p_1 and p_2 as outliers

NN approach may consider p_3 as outlier, but LOF approach does not

Local Outlier Factor (LOF)*

- For each data point *q* compute the distance to the *k*-th nearest neighbor (*k-distance*)
- Compute *reachability distance* (*reach-dist*) for each data example *q* with respect to data example *p* as:

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

- Compute *local reachability density* (*lrd*) of data example *q* as inverse of the average reachabaility distance based on the *MinPts* nearest neighbors of data example *q*

$$lrd(q) = \frac{MinPts}{\sum_p reach\_dist_{MinPts}(q,p)}$$

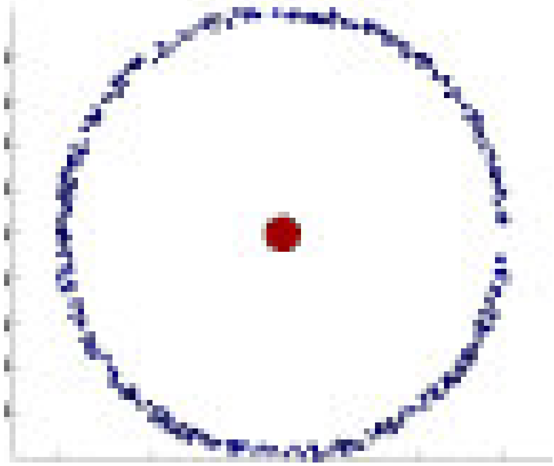
- Compaute *LOF*(*q*) as ratio of average local reachability density of *q*'s *k*-nearest neighbors and local reachability density of the data record *q*

$$LOF(q) = \frac{1}{MinPts} \cdot \sum_p \frac{lrd(p)}{lrd(q)}$$

* - Breunig, et al, LOF: Identifying Density-Based Local Outliers, KDD 2000.

Connectivity Outlier Factor (COF)*

- Outliers are points *p* where average chaining distance $ac\text{-}dist_{kNN(p)}(p)$ is larger than the average chaining distance (*ac-dist*) of their k-nearest neighborhood kNN(p)



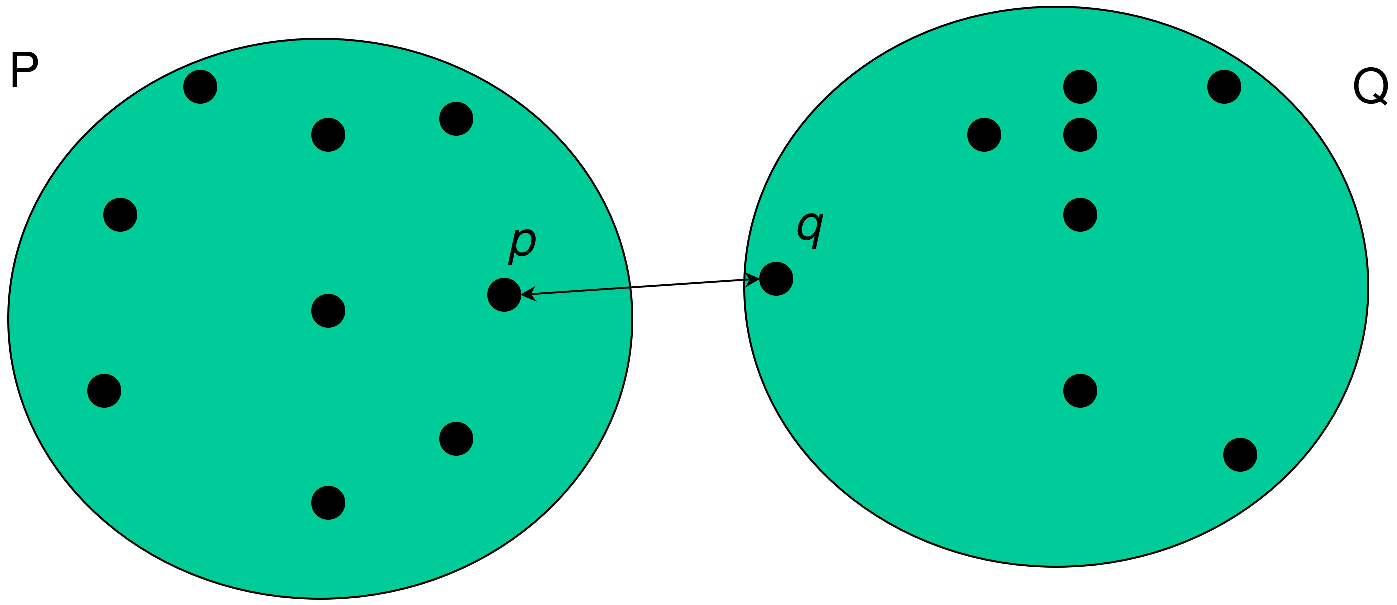
- COF identifies outliers as points whose neighborhoods is sparser than the neighborhoods of their neighbors

* J. Tang, Z. Chen, A. W. Fu, D. Cheung, "A robust outlier detection scheme for large data sets," Proc. Pacific-Asia Conf. Knowledge Discovery and Data Mining, Taïpeh, Taiwan, 2002.

Couple of Definitions

- Distance Between Two Sets

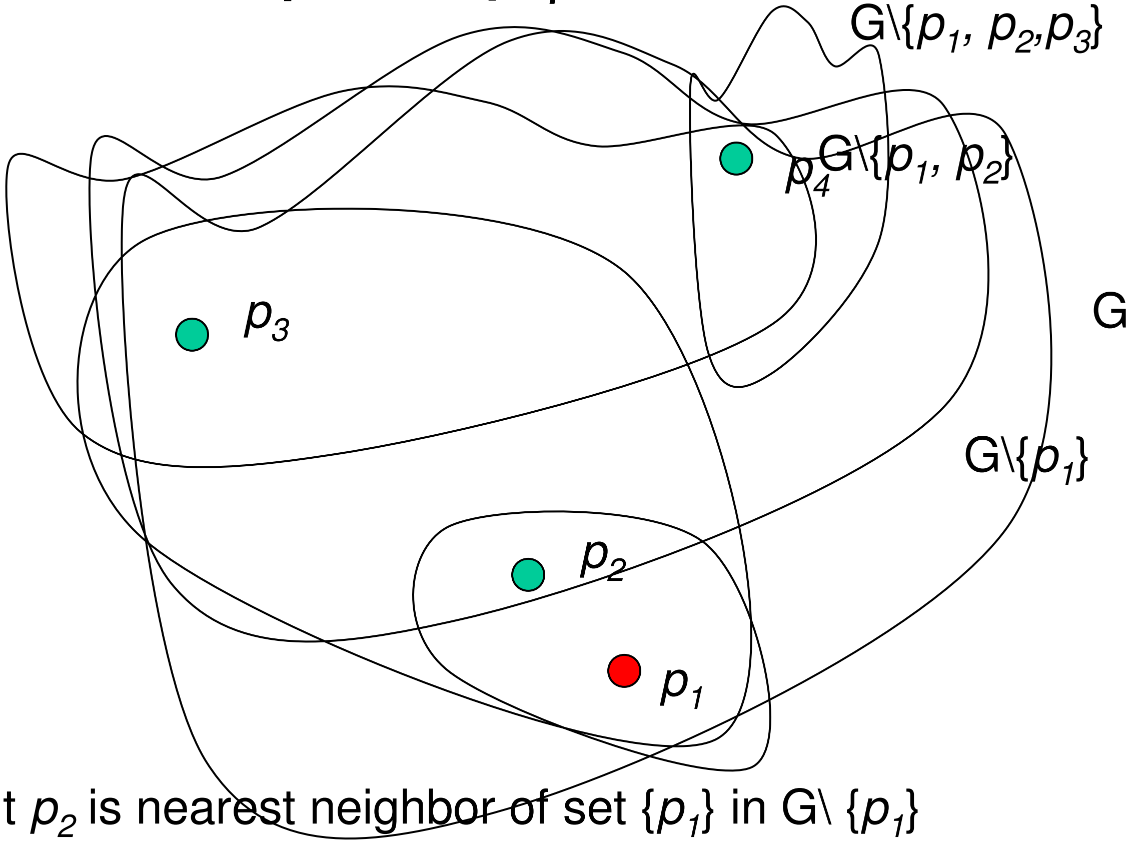
=Distance Between Nearest Points in Two Sets



Point *p* is nearest neighbor of set Q in P

Set-Based Path

- Consider point p_1 from set G



Point p_2 is nearest neighbor of set {p_1} in G\ {p_1}

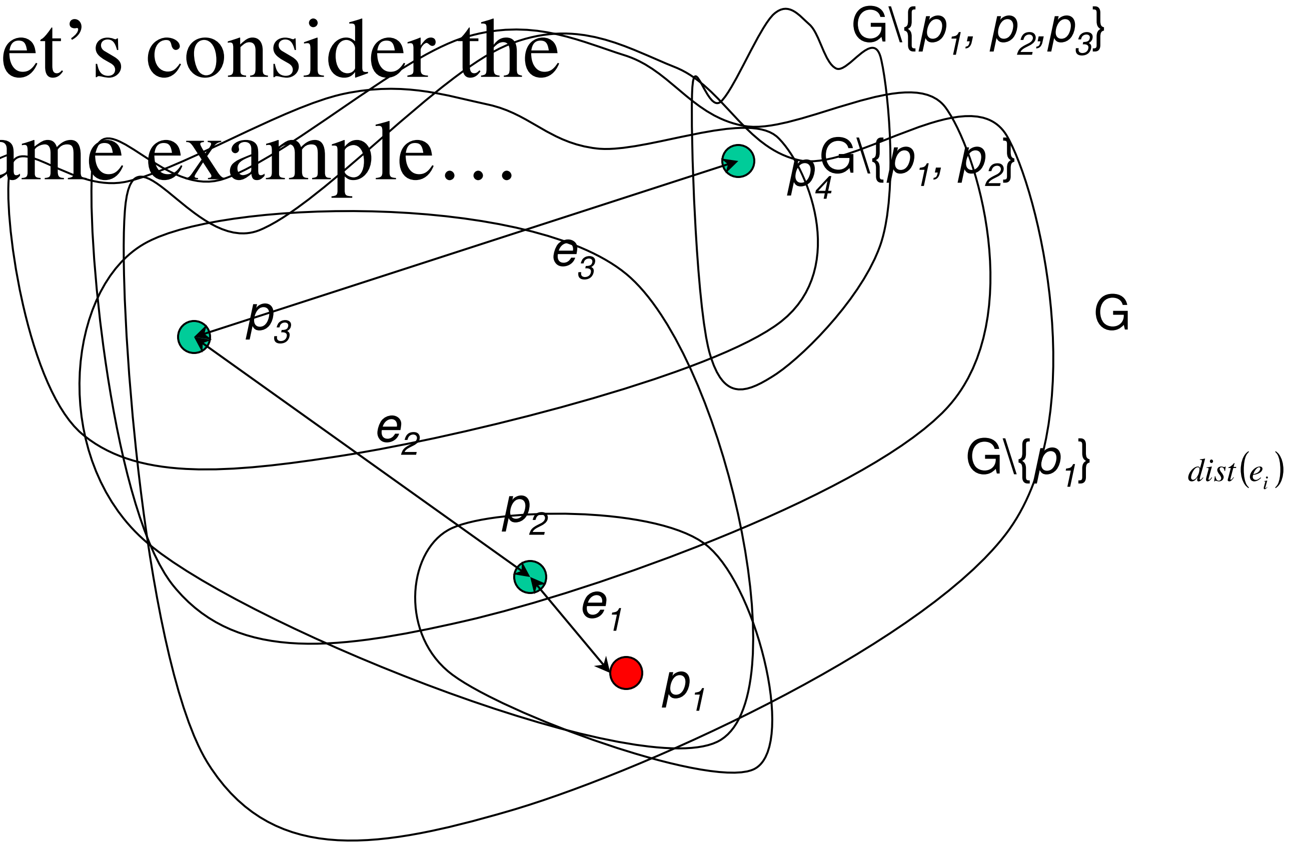
Point p_3 is nearest neighbor of set {p_1, p_2} in G\ {p_1,p_2}

Point p_4 is nearest neighbor of set {p_1, p_2 , p_3} in G\ {p_1,p_2 , p_3}

Sequence {p_1, p_2 , p_3 , p_4} is called Set based Nearest Path (SBN) from p_1 on G

Cost Descriptions

- Let’s consider the same example…



Distances dist(e_i) between two sets {p_1,…, p_i} and G\{p_1,…, p_i} for each *i* are called COST DESCRIPTIONS

Edges e_i for each *i* are called SBN trail
SBN trail may not be a connected graph!

Average Chaining Distance (ac-dist)

- We average *cost descriptions!*
- We would like to give more weights to points closer to the point p_1
- This leads to the following formula:

$$ac - dist_G(p) \equiv \sum_{i=1}^{r} \frac{2(r-i)}{r(r-1)} dist(e_i)$$

- The smaller *ac-dist*, the more compact is the neighborhood *G* of *p*

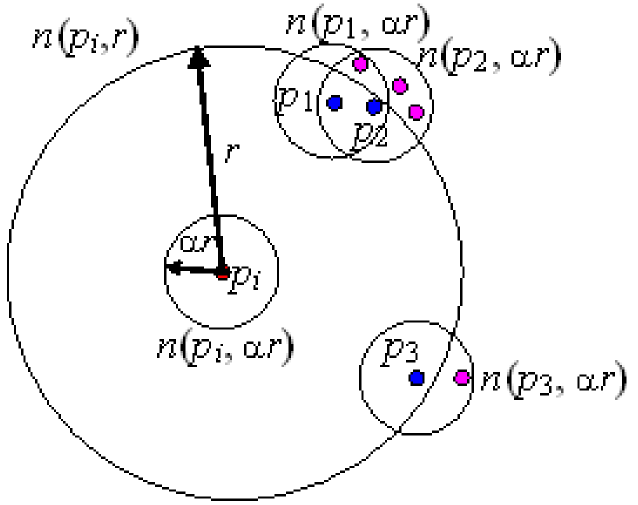
Connectivity Outlier Factor (COF)

- COF is computed as the ratio of the ac-dist (average chaining distance) at the point and the mean ac-dist at the point’s neighborhood
- Similar idea as LOF approach:
  - A point is an outlier if its neighborhood is less compact than the neighborhood of its neighbors

$$COF_k(p) \equiv \frac{ac-dist_{N_k(p)\cup p}(p)}{\frac{1}{k}\sum_{o\in N_k(p)} ac-dist_{N_k(o)\cup o}(o)}$$

Multi-Granularity Deviation Factor - LOCI*

- LOCI computes the neighborhood size (the number of neighbors) for each point and identifies as outliers points whose neighborhood size significantly vary with respect to the neighborhood size of their neighbors
- This approach not only finds outlying points but also outlying micro-clusters.
- LOCI algorithm provides LOCI plot which contains information such as inter cluster distance and cluster diameter
- *r*-neighbors p_j of a data sample p_i are all the samples such that d(p_i, p_j) ≤ *r*
- $n(p_i, r)$ *denotes the number of r neighbors of the point pi.*



Outliers are samples p_i where for any $r \in [r_{min}, r_{max}]$, $n(p_i, \alpha \cdot r)$ significantly deviates from the distribution of values $n(p_j, \alpha \cdot r)$ associated with samples p_j from the *r*-neighborhood of p_i. Sample is outlier if:

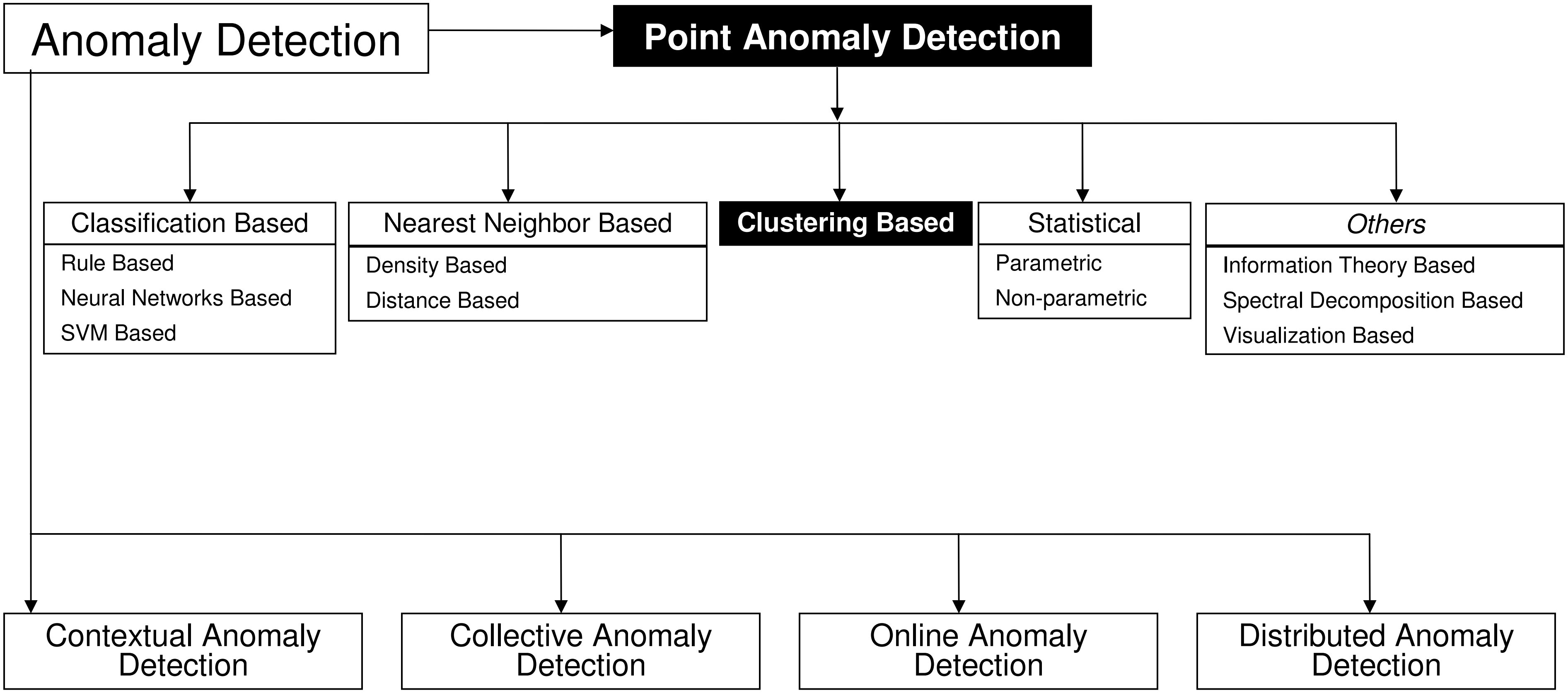
$$n(p_i, \alpha r) < \hat{n}(p_i, r, \alpha) - k_\sigma \sigma_{\hat{n}}(p_i, r, \alpha)$$

Example:

$n(p_i,r)=4$, $n(p_i,\alpha \cdot r)=1$, $n(p_1,\alpha \cdot r)=3$, $n(p_2,\alpha \cdot r)=5$, $n(p_3,\alpha \cdot r)=2$, $\hat{n}(p_i, r, \alpha)$ = (1+3+5+2) / 4 = 2.75, $\sigma_{\hat{n}}(p_i, r, \alpha) \approx 1.479$; α = 1/4.

*- S. Papadimitriou, et al, “LOCI: Fast outlier detection using the local correlation integral,” *Proc. 19th ICDE'03*, Bangalore, India, March 2003.

Taxonomy



Anomaly Detection

Point Anomaly Detection

Classification Based
- Rule Based
- Neural Networks Based
- SVM Based

Nearest Neighbor Based
- Density Based
- Distance Based

Clustering Based

Statistical
- Parametric
- Non-parametric

Others
- Information Theory Based
- Spectral Decomposition Based
- Visualization Based

Contextual Anomaly Detection

Collective Anomaly Detection

Online Anomaly Detection

Distributed Anomaly Detection

Clustering Based Techniques

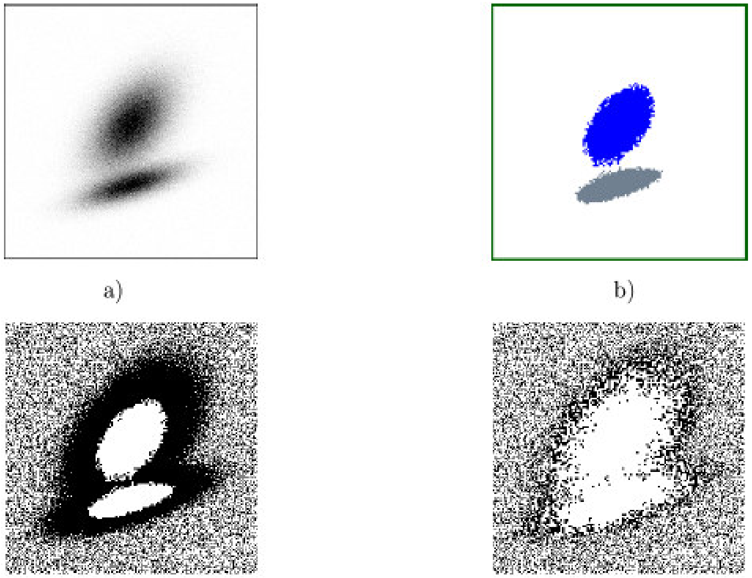
- *Key Assumption:* Normal data instances belong to large and dense clusters, while anomalies do not belong to any significant cluster.
- *General Approach:*
  - Cluster data into a finite number of clusters.
  - Analyze each data instance with respect to its closest cluster.
  - Anomalous Instances
    - Data instances that do not fit into any cluster (residuals from clustering).
    - Data instances in small clusters.
    - Data instances in low density clusters.
    - Data instances that are far from other points within the same cluster.

Clustering Based Techniques

- Advantages
  - Unsupervised.
  - Existing clustering algorithms can be plugged in.
- Drawbacks
  - If the data does not have a natural clustering or the clustering algorithm is not able to detect the natural clusters, the techniques may fail.
  - Computationally expensive
    - Using indexing structures (k-d tree, R* tree) may alleviate this problem.
  - In high dimensional spaces, data is sparse and distances between any two data records may become quite similar.

FindOut*

- FindOut algorithm as a by-product of *WaveCluster.*
- Transform data into multidimensional signals using wavelet transformation
  - High frequency of the signals correspond to regions where is the rapid change of distribution – boundaries of the clusters.
  - Low frequency parts correspond to the regions where the data is concentrated.
- Remove these high and low frequency parts and all remaining points will be outliers.



* D. Yu, G. Sheikholeslami, A. Zhang, FindOut: Finding Outliers in Very Large Datasets, 1999.

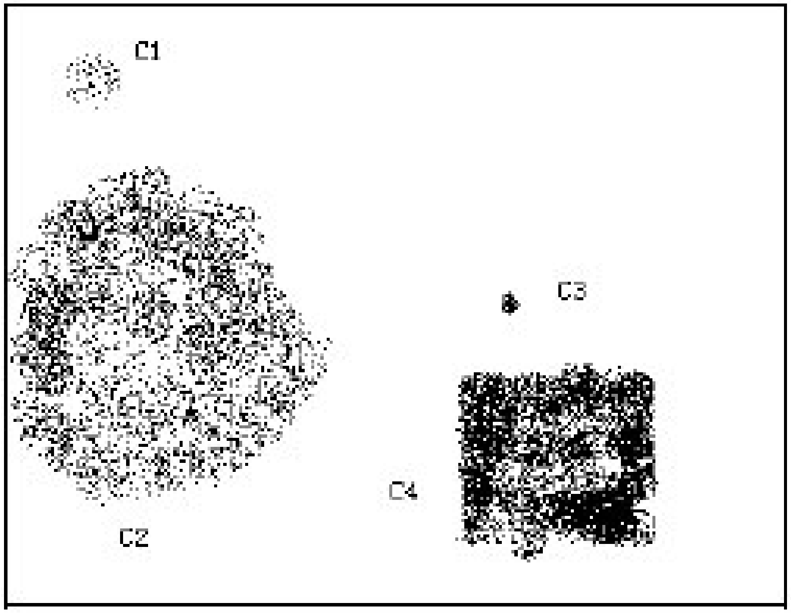
Clustering for Anomaly Detection*

- Fixed-width clustering is first applied
  - The first point is the center of first cluster.
  - Two points x_1 and x_2 are "near" if $d(x_1, x_2) \leq \omega$.
    - ω is a user defined parameter.
  - If every subsequent point is "near", add to a cluster
    - Otherwise create a new cluster.
- Points in small clusters are anomalies.

* E. Eskin et al., A Geometric Framework for Unsupervised Anomaly Detection: Detecting Intrusions in Unlabeled Data, 2002.

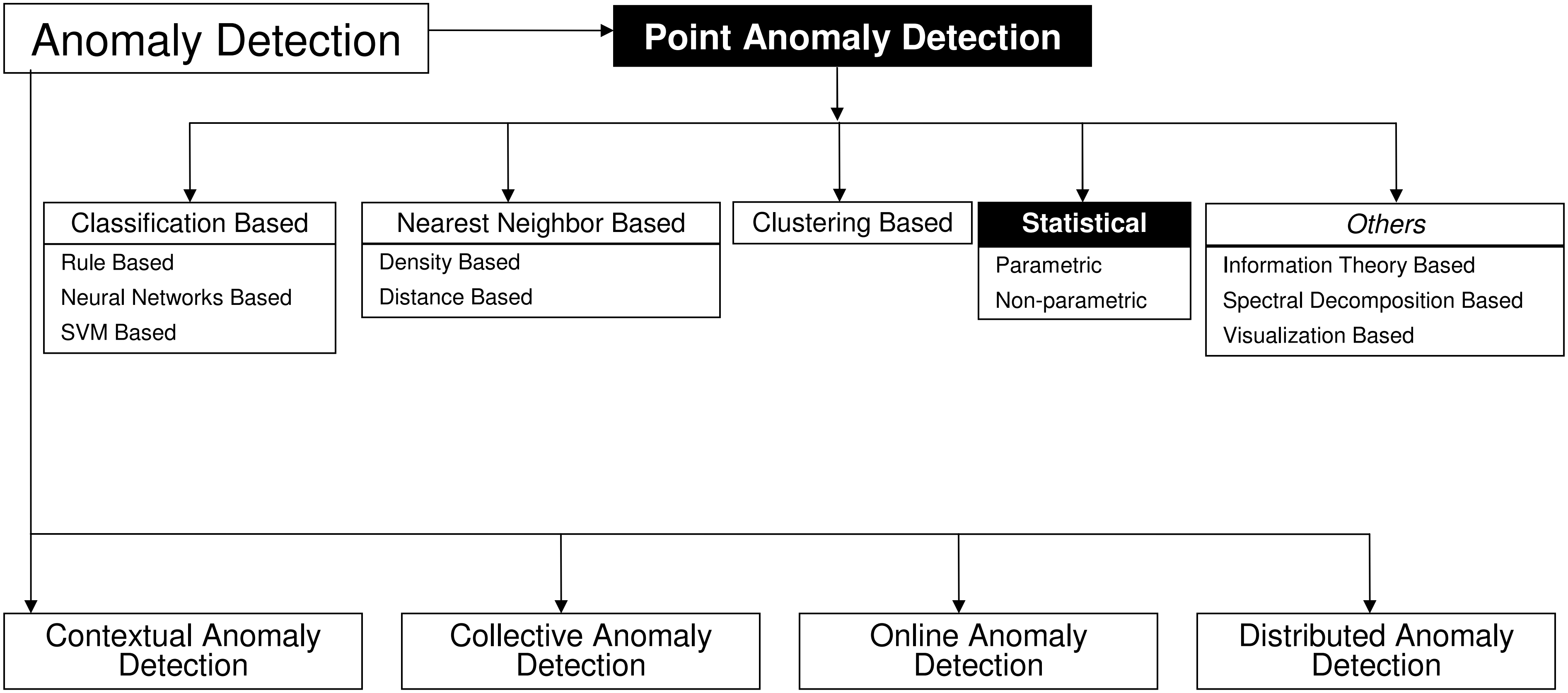
Cluster based Local Outlier Factor*-CBLOF

- Use squeezer clustering algorithm to perform clustering.
- Determine CBLOF for each data instance
  - if the data record lies in a ***small*** cluster, CBLOF = (size of cluster) X (distance between the data instance and the closest larger cluster).
  - if the object belongs to a ***large*** cluster, CBLOF = (size of cluster) X (distance between the data instance and the cluster it belongs to).



*He, Z., Xu, X. i Deng, S. (2003). Discovering cluster based local outliers, Pattern Recognition Letters, 24 (9-10), str. 1651-1660

Taxonomy


Anomaly Detection
Point Anomaly Detection
Classification Based
Rule Based
Neural Networks Based
SVM Based
Nearest Neighbor Based
Density Based
Distance Based
Clustering Based
Statistical
Parametric
Non-parametric
Others
Information Theory Based
Spectral Decomposition Based
Visualization Based
Contextual Anomaly Detection
Collective Anomaly Detection
Online Anomaly Detection
Distributed Anomaly Detection

Statistics Based Techniques

- *Key Assumption*: Normal data instances occur in high probability regions of a statistical distribution, while anomalies occur in the low probability regions of the statistical distribution.
- *General Approach:* Estimate a statistical distribution using given data, and then apply a statistical inference test to determine if a test instance belongs to this distribution or not.
  - *If an observation is more than 3 standard deviations away from the sample mean, it is an anomaly.*
  - *Anomalies have large value for* $T^2 = \frac{n}{n+1}(\mathbf{X}-\overline{\mathbf{X}})'\mathbf{S}^{-1}(\mathbf{X}-\overline{\mathbf{X}})$

Statistics Based Techniques

- Advantages
  - Utilize existing statistical modeling techniques to model various type of distributions.
  - Provide a statistically justifiable solution to detect anomalies.
- Drawbacks
  - With high dimensions, difficult to estimate parameters, and to construct hypothesis tests.
  - Parametric assumptions might not hold true for real data sets.

Types of Statistical Techniques

- Parametric Techniques
  - Assume that the normal (and possibly anomalous) data is generated from an underlying parametric distribution.
  - Learn the parameters from the training sample.
- Non-parametric Techniques
  - Do not assume any knowledge of parameters.
  - Use non-parametric techniques to estimate the density of the distribution – *e.g., histograms, parzen window estimation.*

Using Chi-square Statistic*

- Normal data is assumed to have a multivariate normal distribution.
- Sample mean is estimated from the normal sample.
- Anomaly score of a test instance is

$$\sum_{i=1}^{n} \frac{(X_i - \overline{X}_i)^2}{\overline{X}_i}$$

Ye, N. and Chen, Q. 2001. An anomaly detection technique based on a chi-square statistic for detecting intrusions into information systems. *Quality and Reliability Engineering International 17, 105-112.*

SmartSifter (SS)*

- Statistical modeling of data with continuous and categorical attributes.
  - Histogram density used to represent a probability density for categorical attributes.
  - Finite mixture model used to represent a probability density for continuous attributes.
- For a test instance, SS estimates the probability of the test instance to be generated by the learnt statistical model – p_{t-1}
- The test instance is then added to the sample, and the model is re-estimated.
- The probability of the test instance to be generated from the new model is estimated – p_t.
- Anomaly score for the test instance is the difference $|p_t – p_{t-1}|$.

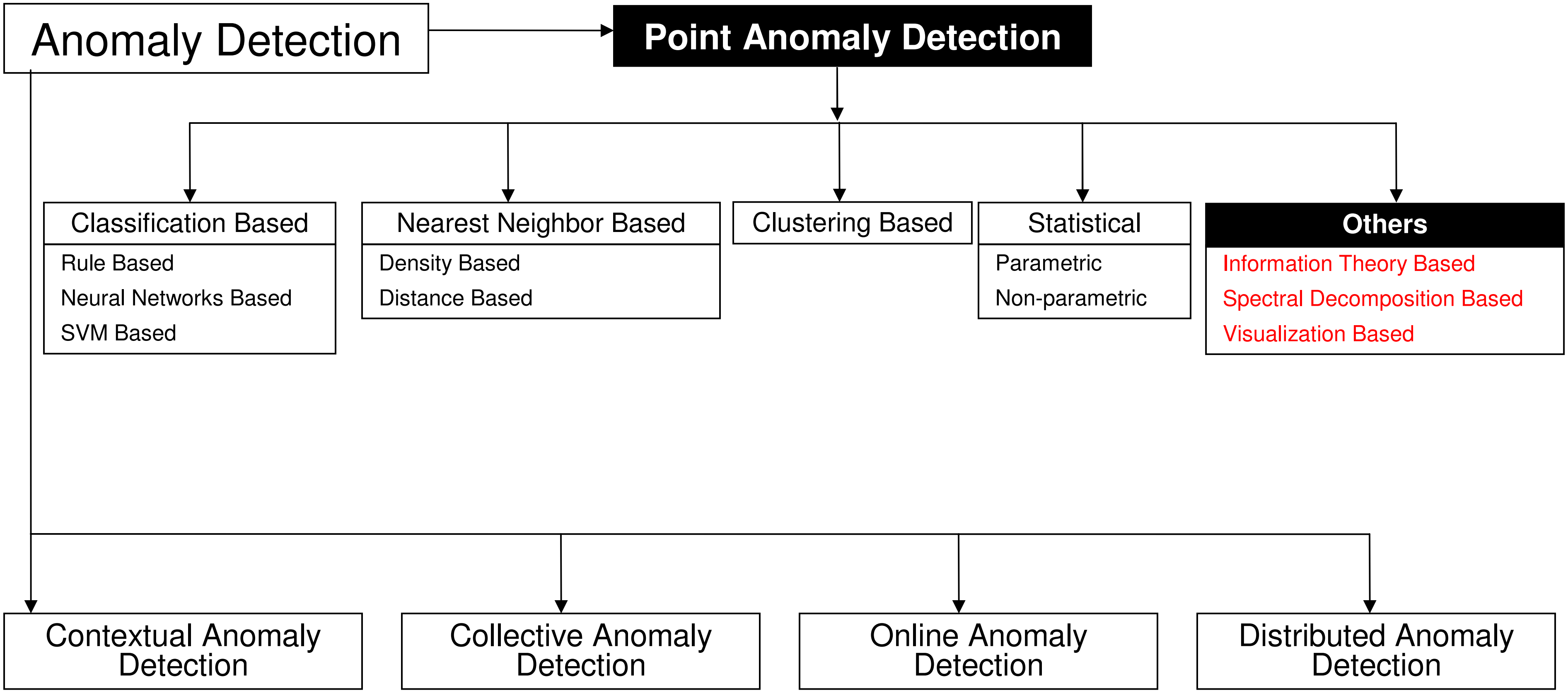
* K. Yamanishi, On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms, KDD 2000

Modeling Normal and Anomalous Data*

- Distribution for the data *D* is given by:
  - $D = (1-\lambda)\cdot\mathbf{M} + \lambda\cdot\mathbf{A}$
    M - majority distribution, **A** - anomalous distribution.
  - *M*, *A* : sets of normal, anomalous elements respectively.
  - Step 1 : Assign all instances to *M*, *A* is initially empty.
  - Step 2 : For each instance x_i in *M,*
    - Step 2.1 : Estimate parameters for **M** and **A**.
    - Step 2.2 : Compute log-likelihood *L* of distribution **D.**
    - Step 2.3 : Remove *x* from *M* and insert in *A*.
    - Step 2.4 : Re-estimate parameters for **M** and **A**.
    - Step 2.5 : Compute the log-likelihood L' of distribution **D**.
    - Step 2.6 : If $L' - L > \delta$, *x* is an anomaly, otherwise *x* is moved back to *M*.
  - Step 3 : Go back to Step 2.

* E. Eskin, Anomaly Detection over Noisy Data using Learned Probability Distributions, ICML 2000

Taxonomy



Anomaly Detection

Point Anomaly Detection

Classification Based

- Rule Based
- Neural Networks Based
- SVM Based

Nearest Neighbor Based

- Density Based
- Distance Based

Clustering Based

Statistical

- Parametric
- Non-parametric

Others

- Information Theory Based
- Spectral Decomposition Based
- Visualization Based

Contextual Anomaly Detection

Collective Anomaly Detection

Online Anomaly Detection

Distributed Anomaly Detection

Information Theory Based Techniques

- *Key Assumption*: Outliers significantly alter the information content in a dataset.
- *General Approach*: Detect data instances that significantly alter the information content
  - Require an information theoretic measure.

Information Theory Based Techniques

- *Advantages*
  - Can operate in an unsupervised mode.
- *Drawbacks*
  - Require an information theoretic measure sensitive enough to detect irregularity induced by very few anomalies.

Using Entropy*

- Find a k-sized subset whose removal leads to the maximal decrease in entropy of the data set.
- Uses an approximate search algorithm LSA to search for the k-sized subsets in linear fashion.
- Other information theoretic measures have been investigated such as conditional entropy, relative conditional entropy, information gain, etc.

He, Z., Xu, X., and Deng, S. 2005. An optimization model for outlier detection in categorical data. In Proceedings of International Conference on Intelligent Computing. Vol. 3644. Springer.

Spectral Techniques

- Analysis based on Eigen decomposition of data.
- Key Idea
  - Find combination of attributes that capture bulk of variability.
  - Reduced set of attributes can explain normal data well, but not necessarily the anomalies.
- Advantage
  - Can operate in an unsupervised mode.
- Drawback
  - Based on the assumption that anomalies and normal instances are distinguishable in the reduced space.

Using Robust PCA*

- Compute the principal components of the dataset
- For each test point, compute its projection on these components
- If y_i denotes the *i*th component, then the following has a chi-squared distribution

$$\sum_{i=1}^{q} \frac{y_i^2}{\lambda_i} = \frac{y_1^2}{\lambda_1} + \frac{y_2^2}{\lambda_2} + \ldots + \frac{y_q^2}{\lambda_q}, q \leq p$$

  – An observation is anomalous, if for a given significance level

$$\sum_{i=1}^{q} \frac{y_i^2}{\lambda_i} > \chi_q^2(\alpha)$$

- Another measure is to observe last few principal components

$$\sum_{i=p-r+1}^{p} \frac{y_i^2}{\lambda_i}$$

- Anomalies have high value for the above quantity.

* Shyu, M.-L., Chen, S.-C., Sarinnapakorn, K., and Chang, L. 2003. A novel anomaly detection scheme based on principal component classifier, In Proceedings of the IEEE Foundations and New Directions of Data Mining Workshop.

PCA for Anomaly Detection*

- Top few principal components capture variability in normal data.
- Smallest principal component should have constant values for normal data.
- Outliers have variability in the smallest component.
- Network intrusion detection using PCA
  - For each time *t*, compute the principal component.
  - Stack all principal components over time to form a matrix.
  - Left singular vector of the matrix captures normal behavior.
  - For any *t*, angle between principal component and the singular vector gives degree of anomaly.

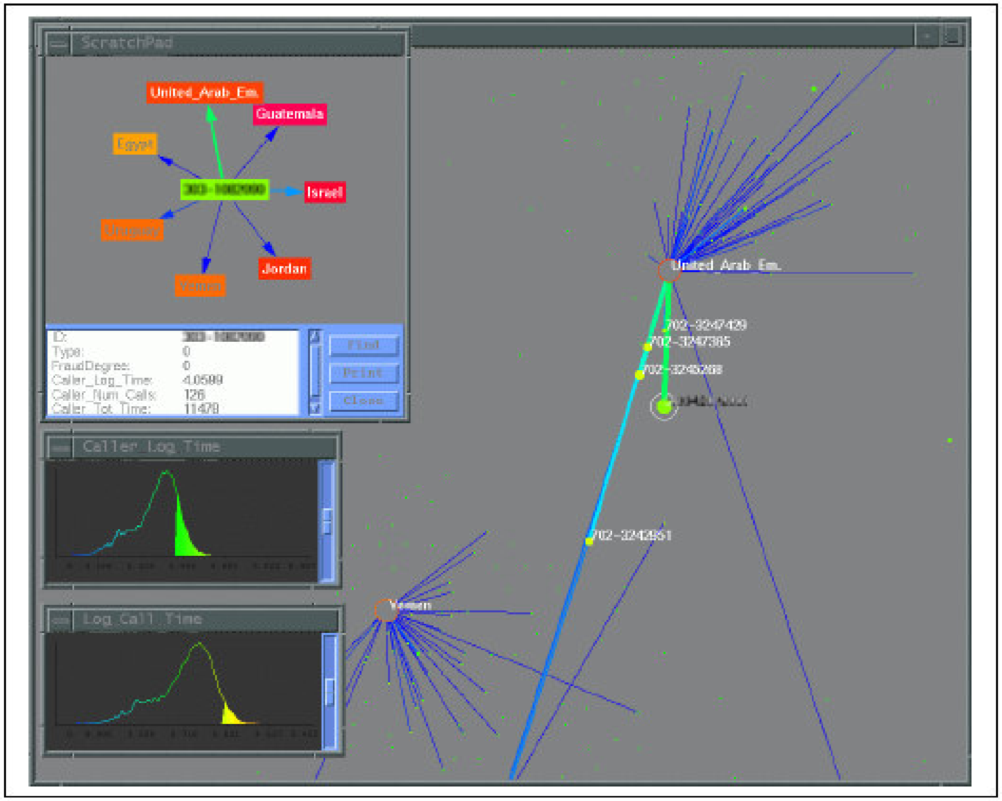
* Ide, T. and Kashima, H. Eigenspace-based anomaly detection in computer systems. KDD, 2004

Visualization Based Techniques

- Use visualization tools to observe the data.
- Provide alternate views of data for manual inspection.
- Anomalies are detected visually.
- Advantages
  - Keeps a human in the loop.
- Drawbacks
  - Works well for low dimensional data.
  - Anomalies might be not identifiable in the aggregated or partial views for high dimension data.
  - Not suitable for real-time anomaly detection.

Visual Data Mining*

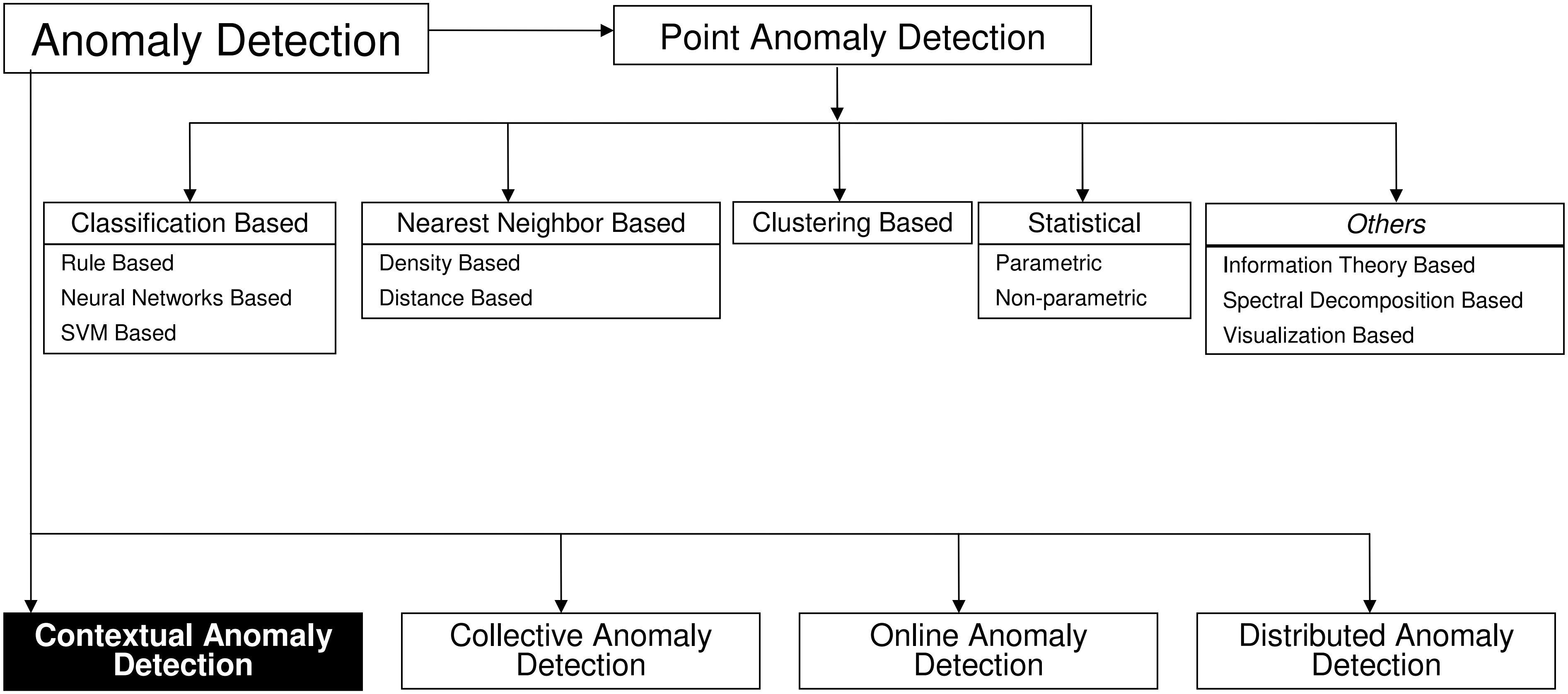
- Detecting Tele-communication fraud.
- Display telephone call patterns as a graph.
- Use colors to identify fraudulent telephone calls (anomalies).



ID: [illegible]
Type: 0
FraudDegree: 0
Caller_Log_Time: 4.0589
Caller_Num_Calls: 126
Caller_Tot_Time: 11478

[*] Cox et al 1997. Visual data mining: Recognizing telephone calling fraud. *Journal of Data Mining and Knowledge Discovery.*

Taxonomy



Anomaly Detection
Point Anomaly Detection
Classification Based
Rule Based
Neural Networks Based
SVM Based
Nearest Neighbor Based
Density Based
Distance Based
Clustering Based
Statistical
Parametric
Non-parametric
Others
Information Theory Based
Spectral Decomposition Based
Visualization Based
Contextual Anomaly Detection
Collective Anomaly Detection
Online Anomaly Detection
Distributed Anomaly Detection

Contextual Anomaly Detection

- Detect contextual anomalies.
- *Key Assumption :* All normal instances within a context will be similar (in terms of behavioral attributes), while the anomalies will be different from other instances within the context.
- *General Approach :*
  - Identify a context around a data instance (using a set of *contextual attributes*).
  - Determine if the test data instance is anomalous within the context (using a set of *behavioral attributes*).

Contextual Anomaly Detection

- Advantages
  - Detect anomalies that are hard to detect when analyzed in the global perspective.
- Challenges
  - Identifying a set of good contextual attributes.
  - Determining a context using the contextual attributes.

Contextual Attributes

- Contextual attributes define a neighborhood (context) for each instance
- For example:
  - Spatial Context
    - *Latitude, Longitude*
  - Graph Context
    - *Edges, Weights*
  - Sequential Context
    - *Position, Time*
  - Profile Context
    - *User demographics*

Contextual Anomaly Detection Techniques

- Reduction to point anomaly detection
  - Segment data using contextual attributes.
  - Apply a traditional anomaly outlier within each context using behavioral attributes.
  - Often, contextual attributes cannot be segmented easily.
- Utilizing structure in data
  - Build models from the data using contextual attributes.
    - E.g. – Time series models (ARIMA, etc.)
  - The model automatically analyzes data instances with respect to their context.

Conditional Anomaly Detection*

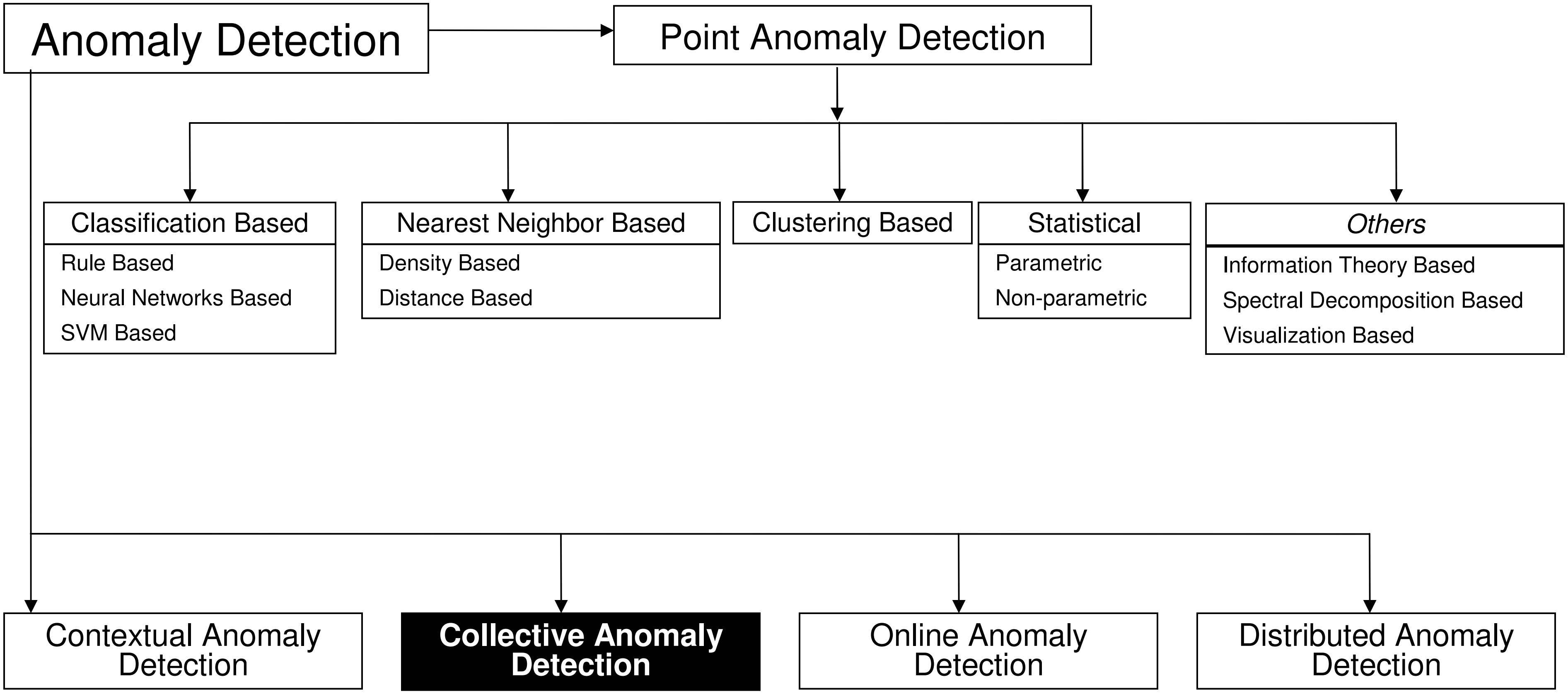
- Each data point is represented as [x,y], where x denotes the *contextual attributes* and y denotes the *behavioral attributes.*
- A mixture of n_U Gaussian models, **U** is learnt from the contextual data.
- A mixture of n_V Gaussian models, **V** is learn from the behavioral data.
- A mapping $p(V_j/U_i)$ is learnt that indicates the probability of the behavioral part to be generated by component V_j when the contextual part is generated by component U_i.
- Anomaly Score of a data instance ([x,y]):

$$\sum_{i=1}^{n_U} p(x \in U_i) \sum_{j=1}^{n_V} p(y \in V_j) p(V_j | U_i)$$

  - How likely is the contextual part to be generated by a component U_i of **U**?
  - Given U_i, what is the most likely component V_j of **V** that will generate the behavioral part?
  - What is the probability of the behavioral part to be generated by V_j.

* Xiuyao Song, Mingxi Wu, Christopher Jermaine, Sanjay Ranka, Conditional Anomaly Detection, IEEE Transactions on Data and Knowledge Engineering, 2006.

Taxonomy



Anomaly Detection

Point Anomaly Detection

Classification Based
- Rule Based
- Neural Networks Based
- SVM Based

Nearest Neighbor Based
- Density Based
- Distance Based

Clustering Based

Statistical
- Parametric
- Non-parametric

Others
- Information Theory Based
- Spectral Decomposition Based
- Visualization Based

Contextual Anomaly Detection

Collective Anomaly Detection

Online Anomaly Detection

Distributed Anomaly Detection

Collective Anomaly Detection

- Detect collective anomalies.
- Exploit the relationship among data instances.
- Sequential anomaly detection
  - Detect anomalous sequences.
- Spatial anomaly detection
  - Detect anomalous sub-regions within a spatial data set.
- Graph anomaly detection
  - Detect anomalous sub-graphs in graph data.

Sequential Anomaly Detection

- Multiple sub-formulations
  - Detect anomalous sequences in a database of sequences, or
  - Detect anomalous subsequence within a sequence.

Sequence Time Delay Embedding (STIDE)*

- Assumes a training data containing normal sequences
- Training
  - Extracts fixed length (k) subsequences by sliding a window over the training data.
  - Maintain counts for all subsequences observed in the training data.
- Testing
  - Extract fixed length subsequences from the test sequence.
  - Find empirical probability of each test subsequence from the above counts.
  - If probability for a subsequence is below a threshold, the subsequence is declared as anomalous.
  - Number of anomalous subsequences in a test sequence is its anomaly score.
- Applied for system call intrusion detection.

* Warrender, Christina, Stephanie Forrest, and Barak Pearlmutter. Detecting Intrusions Using System Calls: Alternative Data Models. To appear, 1999 IEEE Symposium on Security and Privacy. 1999.

Sequential Anomaly Detection – Current State of Art

		State Based				Model Based	Kernel Based	
Data/Applications		FSA	PST	SMT	HMM	Ripper	Clustering	kNN
Univariate Symbolic Sequences	Operating System Call Data	[4][7] [10] [12]		[3]	[4][5] [11]	[4][8]		
	Protein Data		[9]					
	Flight Safety Data				[14]		[13]	
Multivariate Symbolic Sequences								
Univariate Continuous Sequences		[2][7]					[1]	[15]
Multivariate Continuous Sequences								

- [1] – Blender et al 1997
- [2] – Bu et al 2007
- [3] – Eskin and Stolfo 2001
- [4] – Forrest et al 1999
- [5] – Gao et al 2002
- [6] – Hofmeyr et al 1998
- [7] – Keogh et al 2006
- [8] – Lee and Stolfo 1998
- [9] – Sun et al 2006
- [10] – Nong Ye 2004
- [11] – Zhang et al 2003
- [12] – Michael and Ghosh 2000
- [13] – Budalakoti et al 2006
- [14] – A. Srivastava 2005
- [15] – Chan and Mahoney 2005

Anomaly Detection for Symbolic Sequences – A Comparative Evaluation[*]

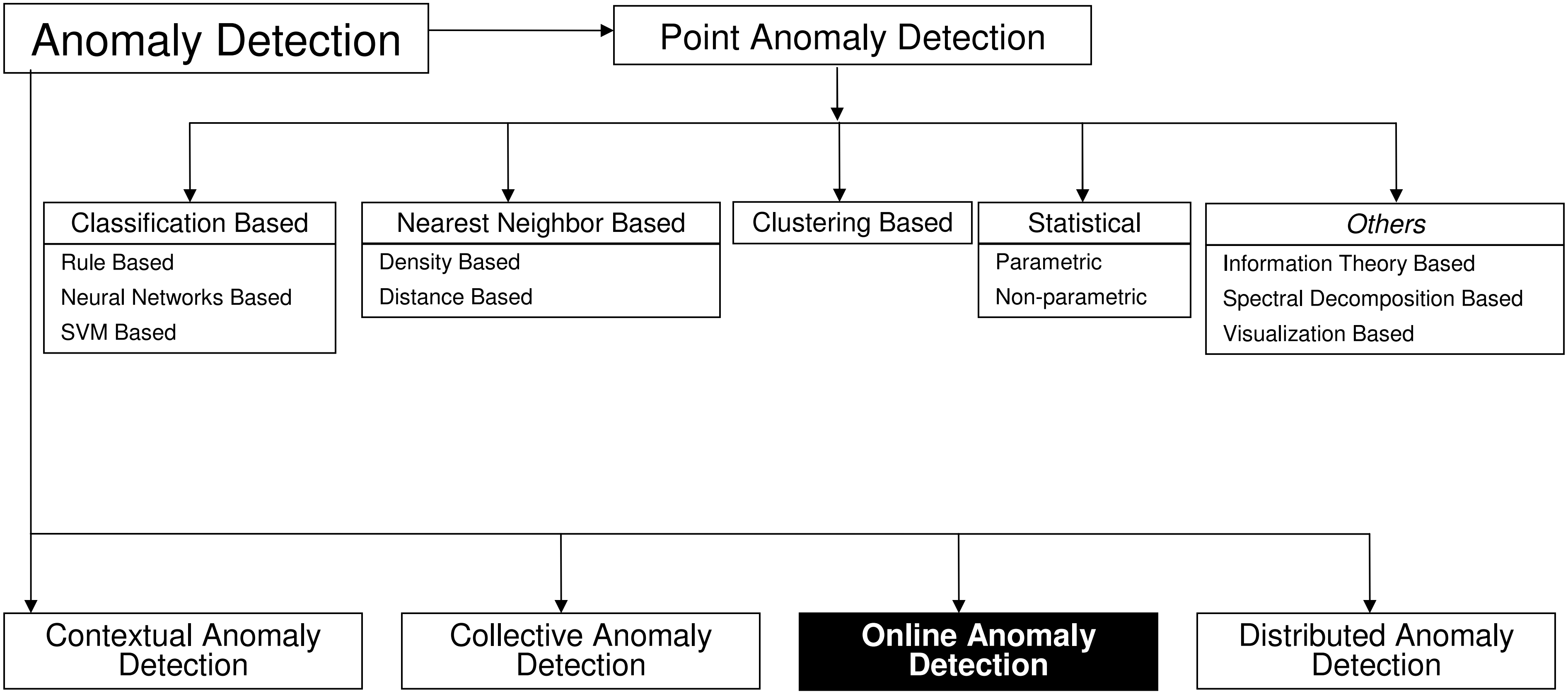
- Test data contains 1000 normal sequences and 100 anomalous sequences.
- Values in table show the percentage of “true” anomalies in top 100 “predicted” anomalies.

Techniques**	**Protein Data**					**System Call Data**	
	HCV	NAD	TET	RUB	RVP	Stide	Sendmail
Clustering	0.88	0.68	0.90	0.96	0.92	0.99	0.72
KNN	0.97	0.79	0.90	0.98	0.94	0.99	0.48
k-MM	1.00	1.00	1.00	1.00	1.00	0.99	0.64
HMM	0.14	0.07	0.28	0.23	0.00	0.98	0.00
PST	0.64	0.13	0.74	0.71	0.07	0.99	0.00
Ripper	0.14	0.16	0.00	0.90	0.82	0.97	0.48

[*] Chandola and Kumar, *Work in Progress.*

[**] Different parameter settings and combination methods (for sequence modeling techniques) were investigated. Best results for each technique are reported here.

Taxonomy



Anomaly Detection

Point Anomaly Detection

Classification Based
- Rule Based
- Neural Networks Based
- SVM Based

Nearest Neighbor Based
- Density Based
- Distance Based

Clustering Based

Statistical
- Parametric
- Non-parametric

Others
- Information Theory Based
- Spectral Decomposition Based
- Visualization Based

Contextual Anomaly Detection

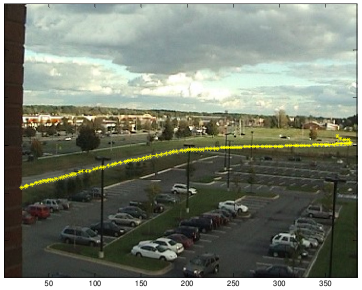
Collective Anomaly Detection

Online Anomaly Detection

Distributed Anomaly Detection

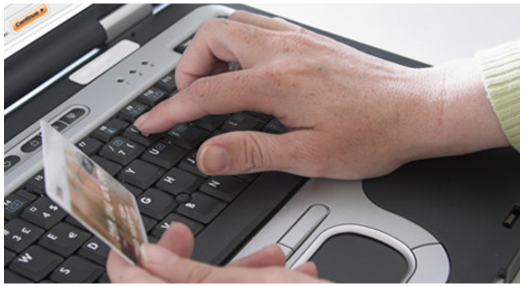
On-line Anomaly Detection

- Often data arrives in a streaming mode.
- Applications
  - Video analysis



  - Network traffic monitoring
  - Aircraft safety

  - Credit card fraudulent transactions

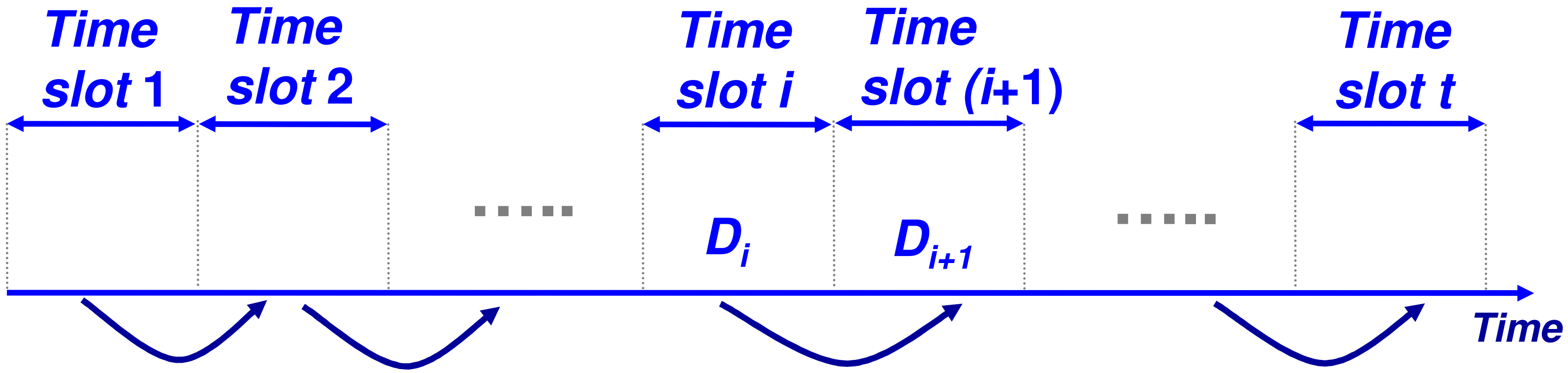


Challenges

- Anomalies need to be detected in real time.
- When to *reject*?
- When to *update*?
  - Require incremental model update techniques as retraining models can be quite expensive.

On-line Anomaly Detection – Simple Idea

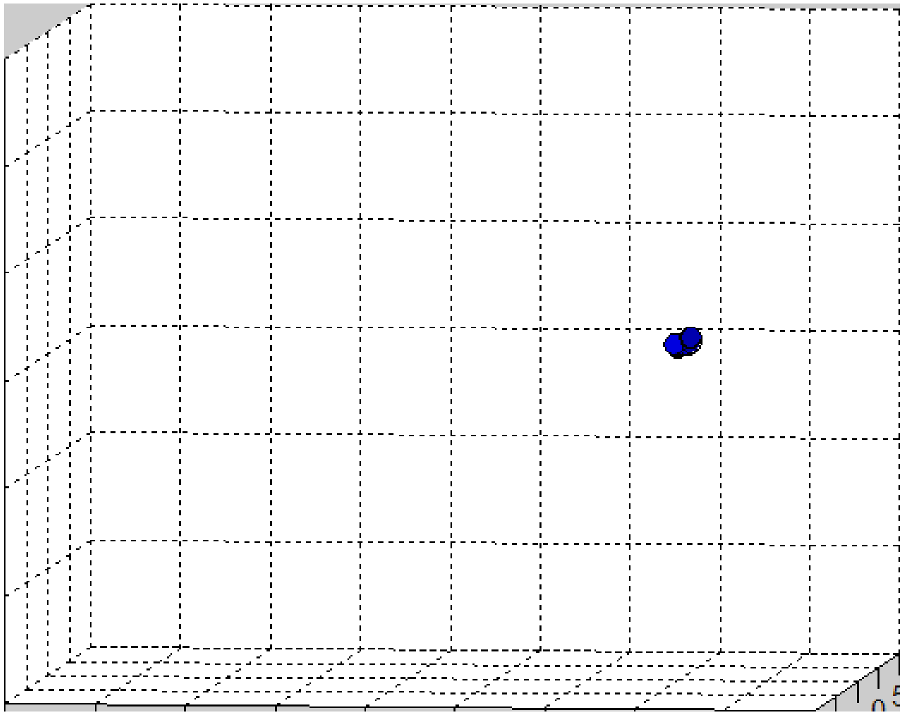
- The normal behavior is changing through time
- Need to update the “normal behavior” profile dynamically
  - Key idea: Update the normal profile with the data records that are “probably” normal, i.e. have very low anomaly score



  - Time slot *i* – Data block D_i – model of normal behavior M_i
  - Anomaly detection algorithm in time slot (*i+1*) is based on the profile computed in time slot *i*

Motivation for Model Updating

- If arriving data points start to create a new data cluster, this method will not be able to detect these points as anomalies.

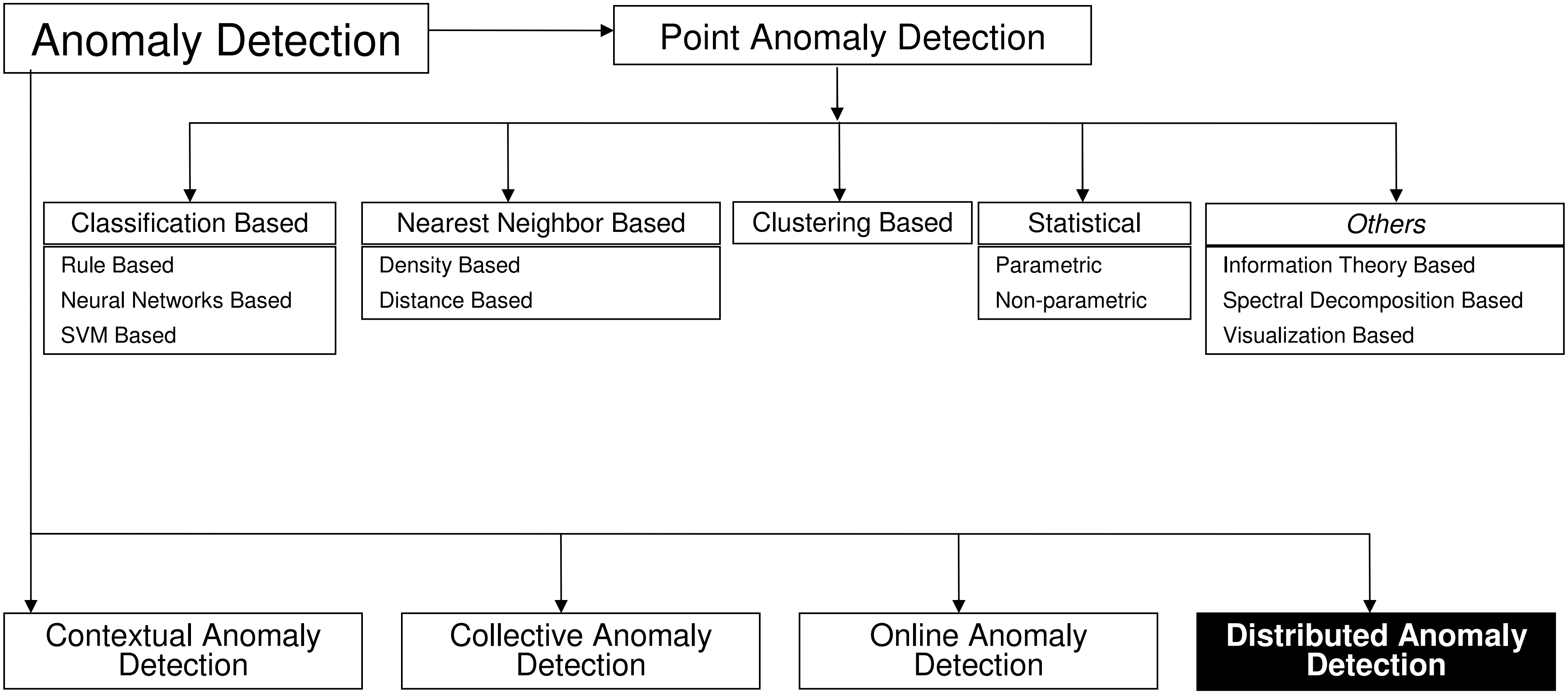


Incremental LOF*

- Incremental *LOF* algorithm computes *LOF* value for each inserted data record and instantly determines whether that data instance is an anomaly.
- *LOF* values for existing data records are updated if necessary.

* D. Pokrajac, A. Lazarevic, and L. J. Latecki. Incremental local outlier detection for data streams. In *Proceedings of IEEE Symposium on Computational Intelligence and Data Mining, 2007.*

Taxonomy



Anomaly Detection

Point Anomaly Detection

Classification Based

- Rule Based
- Neural Networks Based
- SVM Based

Nearest Neighbor Based

- Density Based
- Distance Based

Clustering Based

Statistical

- Parametric
- Non-parametric

Others

- Information Theory Based
- Spectral Decomposition Based
- Visualization Based

Contextual Anomaly Detection

Collective Anomaly Detection

Online Anomaly Detection

Distributed Anomaly Detection

Need for Distributed Anomaly Detection

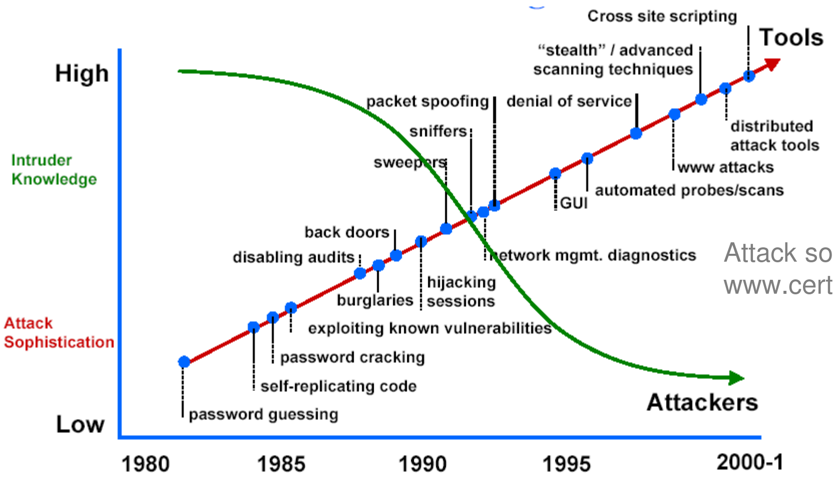
- Data in many anomaly detection applications may come from many different sources
  - Network intrusion detection
  - Credit card fraud
  - Aviation safety
- Failures that occur at multiple locations simultaneously may be undetected by analyzing only data from a single location
  - Detecting anomalies in such complex systems may require integration of information about detected anomalies from single locations in order to detect anomalies at the global level of a complex system
- There is a need for the high performance and distributed algorithms for correlation and integration of anomalies

Distributed Anomaly Detection Techniques

- Simple data exchange approaches
  - Merging data at a single location
  - Exchanging data between distributed locations
- Distributed nearest neighboring approaches
  - Exchanging one data record per distance computation – computationally inefficient
  - privacy preserving anomaly detection algorithms based on computing distances across the sites [Vaidya and Clifton 2004].
- Methods based on exchange of models
  - explore exchange of appropriate statistical / data mining models that characterize normal / anomalous behavior
    - identifying modes of normal behavior;
    - describing these modes with statistical / data mining learning models; and
    - exchanging models across multiple locations and combing them at each location in order to detect global anomalies

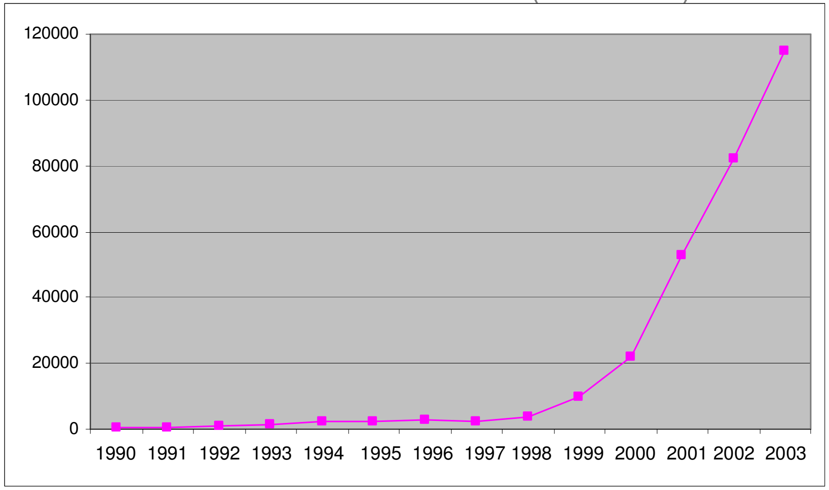
Case Study: Data Mining in Intrusion Detection

- **Due to the proliferation of Internet, more and more organizations are becoming vulnerable to cyber attacks**
- **Sophistication of cyber attacks as well as their severity is also increasing**

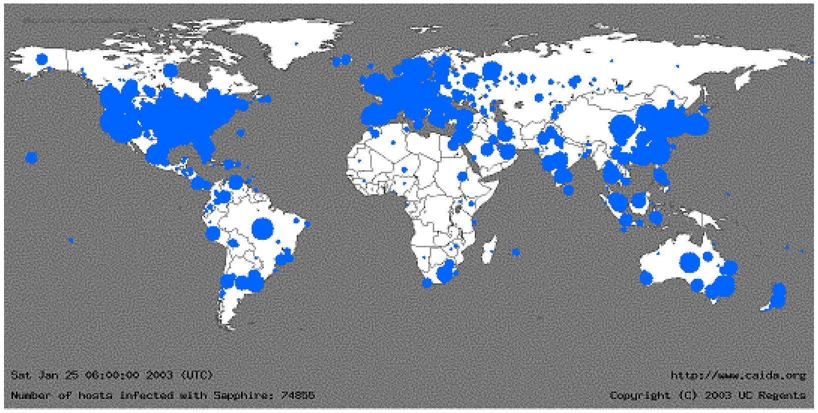


Attack sophistication vs. Intruder technical knowledge, source: www.cert.org/archive/ppt/cyberterror.ppt

Incidents Reported to Computer Emergency Response Team/Coordination Center (CERT/CC)



- **Security mechanisms always have inevitable vulnerabilities**
  - **Firewalls are not sufficient to ensure security in computer networks**
  - **Insider attacks**


Sat Jan 25 06:00:00 2003 (UTC)
Number of hosts infected with Sapphire: 74855
http://www.caida.org
Copyright (C) 2003 UC Regents

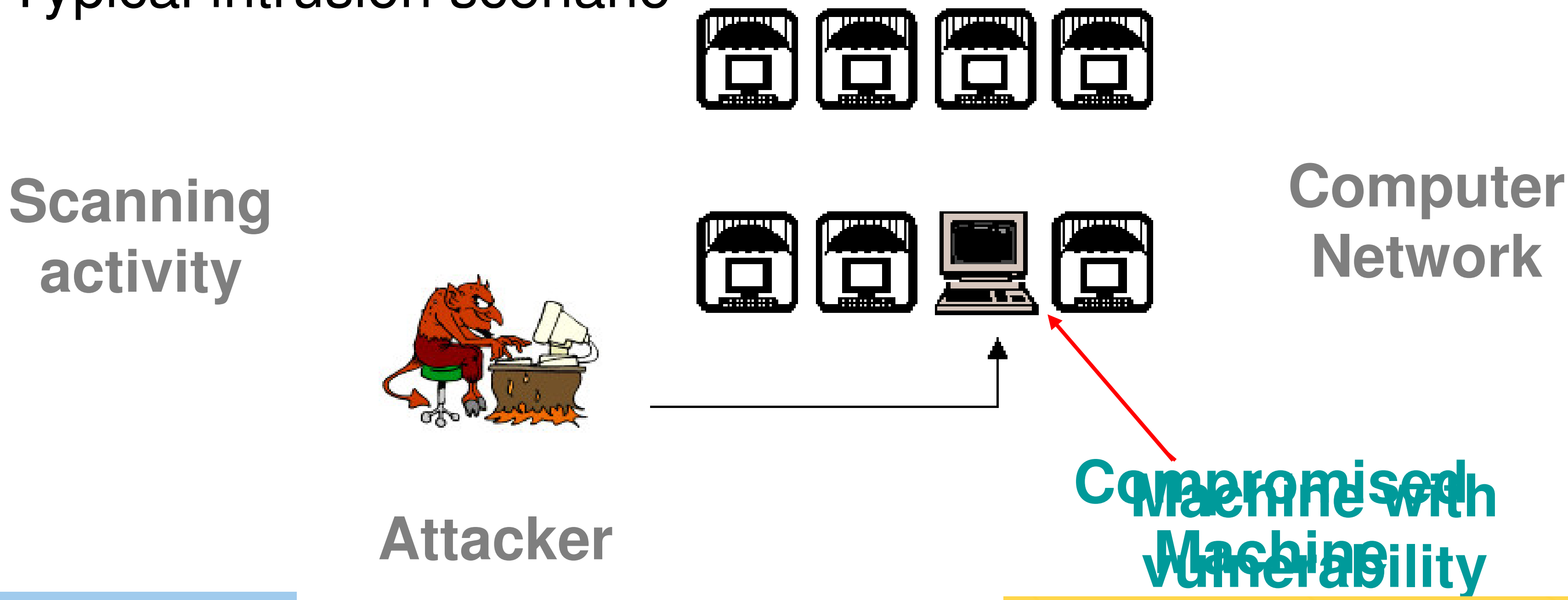
The geographic spread of Sapphire/Slammer Worm 30 minutes after release (**Source:** www.caida.org)

United Technologies

University of Minnesota

What are Intrusions?

- Intrusions are actions that attempt to bypass security mechanisms of computer systems. They are usually caused by:
  - Attackers accessing the system from Internet
  - Insider attackers - authorized users attempting to gain and misuse non-authorized privileges
- Typical intrusion scenario



Scanning activity

Attacker

Computer Network

Compromised Machine

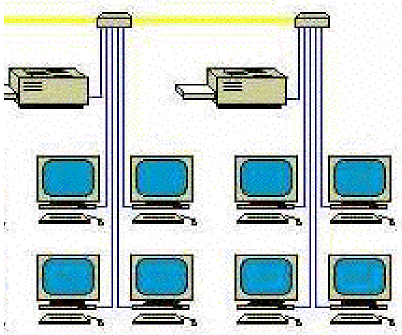
Machine with Vulnerability

IDS - Analysis Strategy

- *Misuse detection* is based on extensive knowledge of patterns associated with known attacks provided by human experts
  - Existing approaches: pattern (signature) matching, expert systems, state transition analysis, data mining
  - Major limitations:
    - Unable to detect novel & unanticipated attacks
    - Signature database has to be revised for each new type of discovered attack
- *Anomaly detection* is based on profiles that represent normal behavior of users, hosts, or networks, and detecting attacks as significant deviations from this profile
  - Major benefit - potentially able to recognize unforeseen attacks.
  - Major limitation - possible high false alarm rate, since detected deviations do not necessarily represent actual attacks
  - Major approaches: statistical methods, expert systems, clustering, neural networks, support vector machines, outlier detection schemes

Intrusion Detection

- Intrusion Detection System
  - combination of software and hardware that attempts to perform intrusion detection
  - raises the alarm when possible intrusion happens



- Traditional intrusion detection system IDS tools (e.g. SNORT) are based on signatures of known attacks
  - Example of SNORT rule (MS-SQL “Slammer” worm)

    any -> udp port 1434 (content:"|81 F1 03 01 04 9B 81 F1 01|"; content:"sock"; content:"send")

www.snort.org

- Limitations
  - Signature database has to be manually revised for each new type of discovered intrusion
  - They cannot detect emerging cyber threats
  - Substantial latency in deployment of newly created signatures across the computer system
- Data Mining can alleviate these limitations

Data Mining for Intrusion Detection

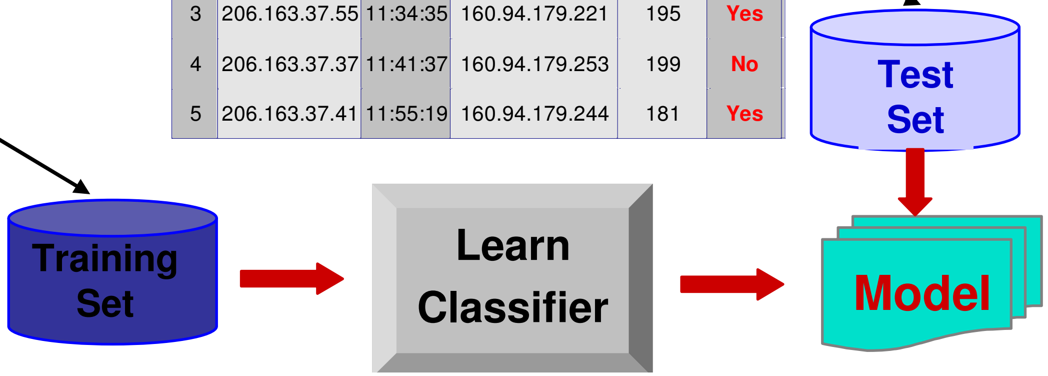
- Increased interest in data mining based intrusion detection
  - Attacks for which it is difficult to build signatures
  - Attack stealthiness
  - Unforeseen/Unknown/Emerging attacks
  - Distributed/coordinated attacks
- Data mining approaches for intrusion detection
  - *Misuse detection*
    - Building predictive models from labeled labeled data sets (instances are labeled as "normal" or "intrusive") to identify known intrusions
    - High accuracy in detecting many kinds of known attacks
    - Cannot detect unknown and emerging attacks
  - *Anomaly detection*
    - Detect novel attacks as deviations from "normal" behavior
    - Potential high false alarm rate - previously unseen (yet legitimate) system behaviors may also be recognized as anomalies
  - *Summarization of network traffic*

Data Mining for Intrusion Detection

Misuse Detection – Building Predictive Models

Tid	SrcIP	Start time	Dest IP	Dest Port	Number of bytes	Attack
1	206.135.38.95	11:07:20	160.94.179.223	139	192	No
2	206.163.37.95	11:13:56	160.94.179.219	139	195	No
3	206.163.37.95	11:14:29	160.94.179.217	139	180	No
4	206.163.37.95	11:14:30	160.94.179.255	139	199	No
5	206.163.37.95	11:14:32	160.94.179.254	139	19	Yes
6	206.163.37.95	11:14:35	160.94.179.253	139	177	No
7	206.163.37.95	11:14:36	160.94.179.252	139	172	No
8	206.163.37.95	11:14:38	160.94.179.251	139	285	Yes
9	206.163.37.95	11:14:41	160.94.179.250	139	195	No
10	206.163.37.95	11:14:44	160.94.179.249	139	163	Yes

	categorical	temporal	categorical	continuous	class
Tid	SrcIP	Start time	Dest IP	Number of bytes	Attack
1	206.163.37.81	11:17:51	160.94.179.208	150	No
2	206.163.37.99	11:18:10	160.94.179.235	208	No
3	206.163.37.55	11:34:35	160.94.179.221	195	Yes
4	206.163.37.37	11:41:37	160.94.179.253	199	No
5	206.163.37.41	11:55:19	160.94.179.244	181	Yes

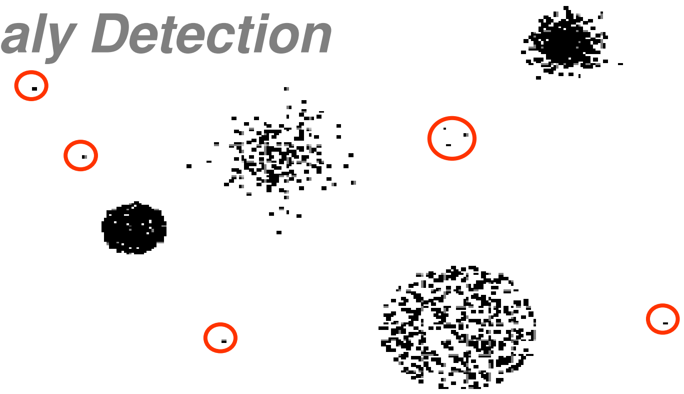


Summarization of attacks using association rules

Rules Discovered:

{Src IP = 206.163.37.95,
Dest Port = 139,
Bytes ∈ [150, 200]} --> {ATTACK}

Anomaly Detection

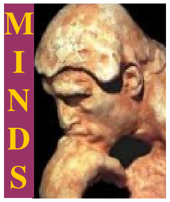


United Technologies

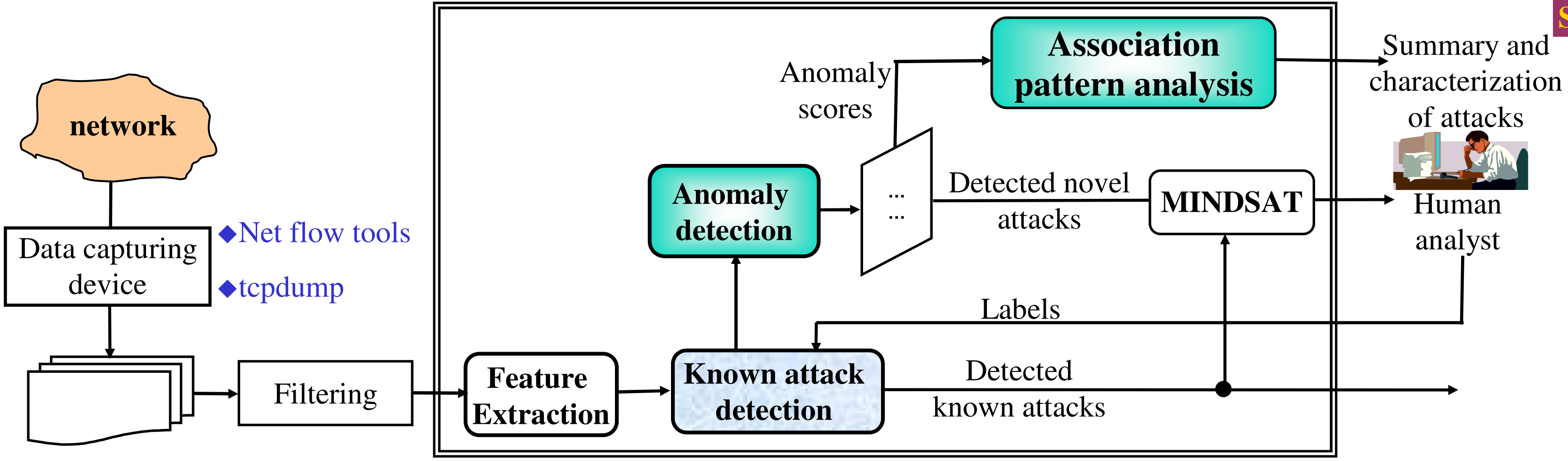
University of Minnesota

Anomaly Detection on Real Network Data

- Anomaly detection was used at U of Minnesota and Army Research Lab to detect various intrusive/suspicious activities
- Many of these could not be detected using widely used intrusion detection tools like SNORT
- Anomalies/attacks picked by *MINDS*
  - Scanning activities
  - Non-standard behavior
    - Policy violations
    - Worms

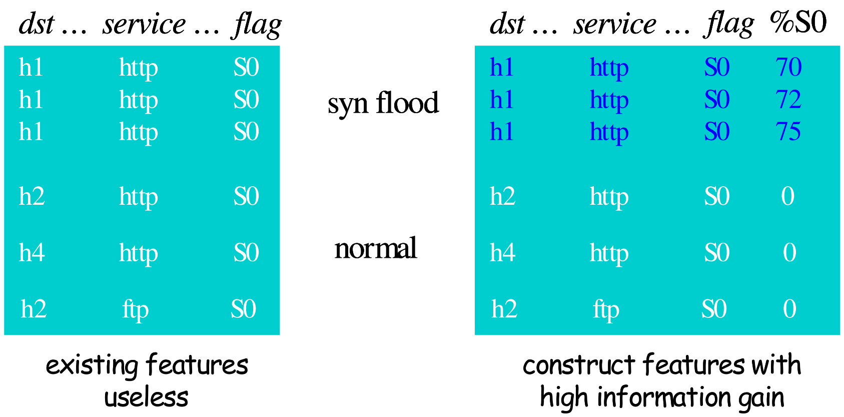


MINDS – Minnesota Intrusion Detection System



Feature Extraction

- Three groups of features
  - –Basic features of individual TCP connections
    - source & destination IP *Features 1 & 2*
    - source & destination port *Features 3 & 4*
    - Protocol *Feature 5*
    - Duration *Feature 6*
    - Bytes per packets *Feature 7*
    - number of bytes *Feature 8*



dst …	service …	flag	
h1	http	S0	syn flood
h1	http	S0	
h1	http	S0	
h2	http	S0	normal
h4	http	S0	
h2	ftp	S0	

existing features useless

dst …	service …	flag	%S0
h1	http	S0	70
h1	http	S0	72
h1	http	S0	75
h2	http	S0	0
h4	http	S0	0
h2	ftp	S0	0

construct features with high information gain

  - –**Time based features**
    - For the same source (destination) IP address, number of unique destination (source) IP addresses inside the network *in last T seconds – Features 9 (13)*
    - Number of connections from source (destination) IP to the same destination (source) port *in last T seconds – Features 11 (15)*
  - –Connection based features
    - For the same source (destination) IP address, number of unique destination (source) IP addresses inside the network *in last N connections - Features 10 (14)*
    - Number of connections from source (destination) IP to the same destination (source) port *in last N connections - Features 12 (16)*

Typical Anomaly Detection Output

– 48 hours after the “slammer” worm

score	srcIP	sPort	dstIP	dPort	protoco	flags	packets	bytes	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
37674.69	63.150.X.253	1161	128.101.X.29	1434	17	16	[0,2)	[0,1829)	0	0	0	0	0	0	0	0	0.81	0	0.59	0	0	0	0	0
26676.62	63.150.X.253	1161	160.94.X.134	1434	17	16	[0,2)	[0,1829)	0	0	0	0	0	0	0	0	0.81	0	0.59	0	0	0	0	0
24323.55	63.150.X.253	1161	128.101.X.185	1434	17	16	[0,2)	[0,1829)	0	0	0	0	0	0	0	0	0.81	0	0.58	0	0	0	0	0
21169.49	63.150.X.253	1161	160.94.X.71	1434	17	16	[0,2)	[0,1829)	0	0	0	0	0	0	0	0	0.81	0	0.58	0	0	0	0	0
19525.31	63.150.X.253	1161	160.94.X.19	1434	17	16	[0,2)	[0,1829)	0	0	0	0	0	0	0	0	0.81	0	0.58	0	0	0	0	0
19235.39	63.150.X.253	1161	160.94.X.80	1434	17	16	[0,2)	[0,1829)	0	0	0	0	0	0	0	0	0.81	0	0.58	0	0	0	0	0
17679.1	63.150.X.253	1161	160.94.X.220	1434	17	16	[0,2)	[0,1829)	0	0	0	0	0	0	0	0	0.81	0	0.58	0	0	0	0	0
8183.58	63.150.X.253	1161	128.101.X.108	1434	17	16	[0,2)	[0,1829)	0	0	0	0	0	0	0	0	0.82	0	0.58	0	0	0	0	0
7142.98	63.150.X.253	1161	128.101.X.223	1434	17	16	[0,2)	[0,1829)	0	0	0	0	0	0	0	0	0.82	0	0.57	0	0	0	0	0
5139.01	63.150.X.253	1161	128.101.X.142	1434	17	16	[0,2)	[0,1829)	0	0	0	0	0	0	0	0	0.82	0	0.57	0	0	0	0	0
4048.49	142.150.Y.101	0	128.101.X.127	2048	1	16	[2,4)	[0,1829)	0	0	0	0	0	0	0	0	0.83	0	0.56	0	0	0	0	0
4008.35	200.250.Z.20	27016	128.101.X.116	4629	17	16	[2,4)	[0,1829)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
3657.23	202.175.Z.237	27016	128.101.X.116	4148	17	16	[2,4)	[0,1829)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
3450.9	63.150.X.253	1161	128.101.X.62	1434	17	16	[0,2)	[0,1829)	0	0	0	0	0	0	0	0	0.82	0	0.57	0	0	0	0	0
3327.98	63.150.X.253	1161	160.94.X.223	1434	17	16	[0,2)	[0,1829)	0	0	0	0	0	0	0	0	0.82	0	0.57	0	0	0	0	0
2796.13	63.150.X.253	1161	128.101.X.241	1434	17	16	[0,2)	[0,1829)	0	0	0	0	0	0	0	0	0.82	0	0.57	0	0	0	0	0
2693.88	142.150.Y.101	0	128.101.X.168	2048	1	16	[2,4)	[0,1829)	0	0	0	0	0	0	0	0	0.83	0	0.56	0	0	0	0	0
2683.05	63.150.X.253	1161	160.94.X.43	1434	17	16	[0,2)	[0,1829)	0	0	0	0	0	0	0	0	0.82	0	0.57	0	0	0	0	0
2444.16	142.150.Y.236	0	128.101.X.240	2048	1	16	[2,4)	[0,1829)	0	0	0	0	0	0	0	0	0.83	0	0.56	0	0	0	0	0
2385.42	142.150.Y.101	0	128.101.X.45	2048	1	16	[0,2)	[0,1829)	0	0	0	0	0	0	0	0	0.83	0	0.56	0	0	0	0	0
2114.41	63.150.X.253	1161	160.94.X.183	1434	17	16	[0,2)	[0,1829)	0	0	0	0	0	0	0	0	0.82	0	0.57	0	0	0	0	0
2057.15	142.150.Y.101	0	128.101.X.161	2048	1	16	[0,2)	[0,1829)	0	0	0	0	0	0	0	0	0.83	0	0.56	0	0	0	0	0
1919.54	142.150.Y.101	0	128.101.X.99	2048	1	16	[2,4)	[0,1829)	0	0	0	0	0	0	0	0	0.83	0	0.56	0	0	0	0	0
1634.38	142.150.Y.101	0	128.101.X.219	2048	1	16	[2,4)	[0,1829)	0	0	0	0	0	0	0	0	0.83	0	0.56	0	0	0	0	0
1596.26	63.150.X.253	1161	128.101.X.160	1434	17	16	[0,2)	[0,1829)	0	0	0	0	0	0	0	0	0.82	0	0.57	0	0	0	0	0
1513.96	142.150.Y.107	0	128.101.X.2	2048	1	16	[0,2)	[0,1829)	0	0	0	0	0	0	0	0	0.83	0	0.56	0	0	0	0	0
1389.09	63.150.X.253	1161	128.101.X.30	1434	17	16	[0,2)	[0,1829)	0	0	0	0	0	0	0	0	0.82	0	0.57	0	0	0	0	0
1315.88	63.150.X.253	1161	128.101.X.40	1434	17	16	[0,2)	[0,1829)	0	0	0	0	0	0	0	0	0.82	0	0.57	0	0	0	0	0
1279.75	142.150.Y.103	0	128.101.X.202	2048	1	16	[0,2)	[0,1829)	0	0	0	0	0	0	0	0	0.83	0	0.56	0	0	0	0	0
1237.97	63.150.X.253	1161	160.94.X.32	1434	17	16	[0,2)	[0,1829)	0	0	0	0	0	0	0	0	0.83	0	0.56	0	0	0	0	0
1180.82	63.150.X.253	1161	128.101.X.61	1434	17	16	[0,2)	[0,1829)	0	0	0	0	0	0	0	0	0.83	0	0.56	0	0	0	0	0

- Anomalous connections that correspond to the “slammer” worm
- Anomalous connections that correspond to the ping scan
- Connections corresponding to UM machines connecting to “half-life” game servers

United Technologies

University of Minnesota

Detection of Anomalies on Real Network Data

- Anomalies/attacks picked by MINDS include scanning activities, worms, and non-standard behavior such as policy violations and insider attacks. Many of these attacks detected by MINDS, have already been on the CERT/CC list of recent advisories and incident notes.
- Some illustrative examples of intrusive behavior detected using MINDS at U of M

- Scans
  - August 13, 2004, **Detected scanning for Microsoft DS service on port 445/TCP (Ranked#1)**
    - Reported by CERT as recent DoS attacks that needs further analysis (CERT August 9, 2004)
    - Undetected by SNORT since the scanning was non-sequential (very slow). Rule added to SNORT in September 2004
  - August 13, 2004, Detected scanning for Oracle server (Ranked #2), Reported by CERT, June 13, 2004
    - Undetected by SNORT because the scanning was hidden within another Web scanning
  - October 10, 2005, Detected a distributed windows networking scan from multiple source locations (Ranked #1)
- Policy Violations
  - August 8, 2005, Identified machine running Microsoft PPTP VPN server on non-standard ports (Ranked #1)
    - Undetected by SNORT since the collected GRE traffic was part of the normal traffic
  - August 10 2005 & October 30, 2005, Identified compromised machines running FTP servers on non-standard ports, which is a policy violation (Ranked #1)
    - Example of anomalous behavior following a successful Trojan horse attack
  - February 6, 2006, The IP address 128.101.X.0 (not a real computer, but a network itself) has been targeted with IP Protocol 0 traffic from Korea (61.84.X.97) (bad since IP Protocol 0 is not legitimate)
  - February 6, 2006, Detected a computer on the network apparently communicating with a computer in California over a VPN or on IPv6
- Worms
  - October 10, 2005, Detected several instances of slapper worm that were not identified by SNORT since they were variations of existing worm code
  - February 6, 2006, Detected unsolicited ICMP ECHOREPLY messages to a computer previously infected with Stacheldract worm (a DDos agent)

Conclusions

- Anomaly detection can detect critical information in data.
- Highly applicable in various application domains.
- Nature of anomaly detection problem is dependent on the application domain.
- Need different approaches to solve a particular problem formulation.

Thanks!!!

- Questions?

References

- Ling, C., Li, C. Data mining for direct marketing: Problems and solutions, KDD, 1998.
- Kubat M., Matwin, S., Addressing the Curse of Imbalanced Training Sets: One-Sided Selection, ICML 97.
- N. Chawla et al., SMOTE: Synthetic Minority Over-Sampling Technique, JAIR, 2002.
- W. Fan et al, Using Artificial Anomalies to Detect Unknown and Known Network Intrusions, ICDM 2001
- N. Abe, et al, Outlier Detection by Active Learning, KDD 2006
- C. Cardie, N. Howe, Improving Minority Class Prediction Using Case specific feature weighting, ICML 97.
- J. Grzymala et al, An Approach to Imbalanced Data Sets Based on Changing Rule Strength, AAAI Workshop on Learning from Imbalanced Data Sets, 2000.
- George H. John. Robust linear discriminant trees. AI&Statistics, 1995
- Barbara, D., Couto, J., Jajodia, S., and Wu, N. Adam: a testbed for exploring the use of data mining in intrusion detection. SIGMOD Rec., 2001
- Otey, M., Parthasarathy, S., Ghoting, A., Li, G., Narravula, S., and Panda, D. Towards nic-based intrusion detection. KDD 2003
- He, Z., Xu, X., and Deng, S. 2005. An optimization model for outlier detection in categorical data. In Proceedings of International Conference on Intelligent Computing. Vol. 3644. Springer.
- Lee, W., Stolfo, S. J., and Mok, K. W. Adaptive intrusion detection: A data mining approach. Artificial Intelligence Review, 2000
- Qin, M. and Hwang, K. Frequent episode rules for internet anomaly detection. In Proceedings of the 3rd IEEE International Symposium on Network Computing and Applications, 2004
- Ide, T. and Kashima, H. Eigenspace-based anomaly detection in computer systems. KDD, 2004
- Sun, J. et al., Less is more: Compact matrix representation of large sparse graphs. SDM 2007

References

- Lee, W. and Xiang, D. Information-theoretic measures for anomaly detection. In Proceedings of the IEEE Symposium on Security and Privacy. IEEE Computer Society, 2001
- Ratsch, G., Mika, S., Scholkopf, B., and Muller, K.-R. Constructing boosting algorithms from SVMs: An application to one-class classification. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2002
- Tax, D. M. J. One-class classification; concept-learning in the absence of counter-examples. Ph.D. thesis, Delft University of Technology, 2001
- Eskin, E., Arnold, A., Prerau, M., Portnoy, L., and Stolfo, S. A geometric framework for unsupervised anomaly detection. In Proceedings of Applications of Data Mining in Computer Security, 2002
- A. Lazarevic, et al., A Comparative Study of Anomaly Detection Schemes in Network Intrusion Detection, SDM 2003
- Scholkopf, B., Platt, O., Shawe-Taylor, J., Smola, A., and Williamson, R. Estimating the support of a high-dimensional distribution. Tech. Rep. 99-87, Microsoft Research, 1999
- Baker, D. et al., A hierarchical probabilistic model for novelty detection in text. ICML 1999
- Das, K. and Schneider, J. Detecting anomalous records in categorical datasets. KDD 2007
- Augusteijn, M. and Folkert, B. Neural network classification and novelty detection. International Journal on Remote Sensing, 2002
- Sykacek, P. Equivalent error bars for neural network classifiers trained by Bayesian inference. In Proceedings of the European Symposium on Artificial Neural Networks. 121–126, 1997
- Vasconcelos, G. C., Fairhurst, M. C., and Bisset, D. L. Investigating feedforward neural networks with respect to the rejection of spurious patterns. Pattern Recognition Letter, 1995

References

- S. Hawkins, et al. Outlier detection using Replicator neural networks, DaWaK02 2002.
- Jagota, A. Novelty detection on a very large number of memories stored in a Hopfield-style network. In Proceedings of the International Joint Conference on Neural Networks, 1991
- Crook, P. and Hayes, G. A robot implementation of a biologically inspired method for novelty detection. In Proceedings of Towards Intelligent Mobile Robots Conference, 2001
- Dasgupta, D. and Nino, F. 2000. A comparison of negative and positive selection algorithms in novel pattern detection. IEEE International Conference on Systems, Man, and Cybernetics, 2000
- Caudell, T. and Newman, D. An adaptive resonance architecture to define normality and detect novelties in time series and databases. World Congress on Neural Networks, 1993
- Albrecht, S. et al. Generalized radial basis function networks for classification and novelty detection: self-organization of optional Bayesian decision. Neural Networks, 2000
- Steinwart, I., Hush, D., and Scovel, C. A classification framework for anomaly detection. JMLR, 2005
- Srinivas Mukkamala et al. Intrusion Detection Systems Using Adaptive Regression Splines. ICEIS 2004
- Li, Y., Pont et al. Improving the performance of radial basis function classifiers in condition monitoring and fault diagnosis applications where unknown faults may occur. Pattern Recognition Letters, 2002
- Borisyuk, R. et al. An oscillatory neural network model of sparse distributed memory and novelty detection. Biosystems, 2000
- Ho, T. V. and Rouat, J. Novelty detection based on relaxation time of a network of integrate-and-fire neurons. Proceedings of Second IEEE World Congress on Computational Intelligence, 1998
- J. Vaidya and C. Clifton. Privacy-preserving outlier detection. In Proceedings of the 4th IEEE International Conference on Data Mining, pages 233–240, 2004.

References

- R. Blender, K. Fraedrich, and F. Lunkeit. Identification of cyclone track regimes in the north atlantic. *Quarterly Journal of the Royal Meteorological Society, 123(539):727–741, 1997.*
- Y. Bu, T.-W. Leung, A. Fu, E. Keogh, J. Pei, and S. Meshkin. Wat: Finding top-k discords in time series database. In *Proceedings of 7th SIAM International Conference on Data Mining, 2007.*
- E. Eskin and S. Stolfo. Modeling system call for intrusion detection using dynamic window sizes. In *Proceedings of DARPA Information Survivability Conference and Exposition, 2001.*
- S. Forrest, C. Warrender, and B. Pearlmutter. Detecting intrusions using system calls: Alternate data models. In *Proceedings of the 1999 IEEE Symposium on Security and Privacy, pages 133–145, Washington, DC,* USA, 1999. IEEE Computer Society.
- B. Gao, H.-Y. Ma, and Y.-H. Yang. Hmms (hidden markov models) based on anomaly intrusion detection method. In *Proceedings of International Conference on Machine Learning and Cybernetics, pages* 381–385. IEEE Computer Society, 2002.
- R. Gwadera, M. J. Atallah, and W. Szpankowski. Detection of significant sets of episodes in event sequences. In *Proceedings of the Fourth IEEE International Conference on Data Mining, pages 3–10, Washington, DC,* USA, 2004. IEEE Computer Society.
- S. A. Hofmeyr, S. Forrest, and A. Somayaji. Intrusion detection using sequences of system calls. *Journal of Computer Security, 6(3):151–180,* 1998.
- E. Keogh, J. Lin, S.-H. Lee, and H. V. Herle. Finding the most unusual time series subsequence: algorithms and applications. *Knowledge and Information Systems, 11(1):1–27, 2006.*

References

- W. Lee and S. Stolfo. Data mining approaches for intrusion detection. In *Proceedings of the 7th USENIX Security Symposium, San Antonio,* TX, 1998.
- P. Sun, S. Chawla, and B. Arunasalam. Mining for outliers in sequential databases. In *Proceedings of SIAM Conference on Data Mining, 2006.*
- N. Ye. A markov chain model of temporal behavior for anomaly detection. In *Proceedings of the 5th Annual IEEE Information Assurance Workshop. IEEE, 2004.*
- X. Zhang, P. Fan, and Z. Zhu. A new anomaly detection method based on hierarchical hmm. In *Proceedings of the 4th International Conference on Parallel and Distributed Computing, Applications and Technologies,* pages 249–252, 2003.
- C. C. Michael and A. Ghosh. Two state-based approaches to program based anomaly detection. In *Proceedings of the 16th Annual Computer Security Applications Conference, page 21,* 2000.
- S. Budalakoti, A. Srivastava, R. Akella, and E. Turkov. Anomaly detection in large sets of high-dimensional symbol sequences. Technical Report NASA TM-2006-214553, NASA Ames Research Center, 2006.
- A. N. Srivastava. Discovering system health anomalies using data mining techniques. In *Proceedings of 2005 Joint Army Navy NASA Airforce Conference on Propulsion, 2005.*
- P. K. Chan and M. V. Mahoney. Modeling multiple time series for anomaly detection. In *Proceedings of the Fifth IEEE International Conference on Data Mining, pages 90–97, Washington, USA, 2005.*

Backup Slides

- Anomaly Detection Techniques

Using Bayesian Networks

- Typical Bayesian networks
  - Aggregates information from different variables and provide an estimate of the expectancy that event belong to one of normal or anomalous classes [Baker99, Das07]
- Naïve Bayesian classifiers
  - Incorporate prior probabilities into a reasoning model that classifies an event as normal or anomalous based on observed properties of the event and prior probabilities [Sebyala02, Kruegel03]
- Pseudo-Bayes estimators [Barbara01]
  - I stage: learn prior and posterior of unseen anomalies from the training data
  - II stage: use Naive Bayes classifier to classify the instances into normal instances, known anomalies and new anomalies