Anomaly Detectoin: A Survey

by Varun Chandola, Arindam Banerjee, Vipin Kumar

Presenter: Jaehyeong Ahn

Department of Applied Statistics, Konkuk University

jayahn0104@gmail.com

Table of Contents

Introduction

Rapid Introduction to Anomaly Detection Problem

What are Anomalies?

- Anomalies are patterns in data that do not conform to a well defined notion of normal behavior

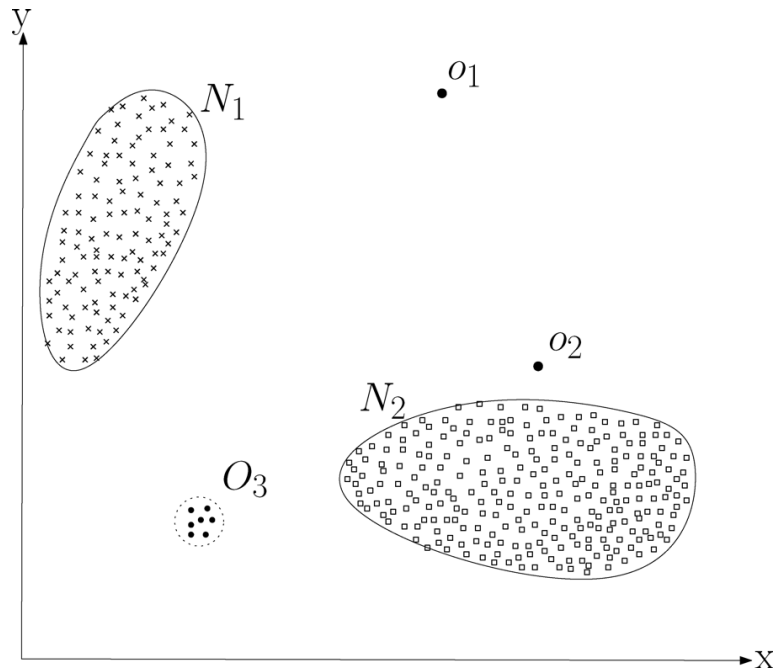


Figure: Simple illustration of anomalies in a two-dimensional data set

- N_1, N_2: Normal regions, o_1, o_2, O_3: Anomalies

Rapid Introduction to Anomaly Detection Problem

How are they called?

- Anomalies, Outliers, Discordant observations, Exceptions, Aberrations, Surprises, Peculiarities, or Contaminants in different application domains

What is an Anomaly Detection?

- Anomaly Detection refers to the problem of finding patterns in data that do not conform to expected behavior

Rapid Introduction to Anomaly Detection Problem

Various applications of Anomaly Detection

- **fraud detection** for credit card, insurance, or health care
- **intrusion detection** for cyber-security
- **fault detection** in safety critical systems
- **military surveillance** for enemy activities

Why Detecting Anomaly is important?

- The importance of anomaly detection is due to the fact that in spite of the small numbers, anomalies cause significant and critical issues
- For example, credit card fraud, cyber-intrusion, terrorist activity or break down a system, etc

Different Aspects of An Anomaly Detection Problem

Anomaly Detection problem can be characterized as a specific formulation by categorizing several different factors:

- **Nature of Input Data**
- **Type of Anomaly**
- **Availability of Data Labels**
- **Output Type of Anomaly Detection Technique**

Nature of Input Data

Input data can be categorized based on the nature of attribute type and the relationship among the data instances

- Attribute type
  - Univariate vs Multivariate
  - Categorical vs Continuous
  - Same type vs Mixture of different types
- Relationship among data instances
  - No relationship vs Related (e.g. sequence data, spatial data, graph data)

Type of Anomaly

An important aspect of an anomaly detection technique is the nature of the desired anomaly.

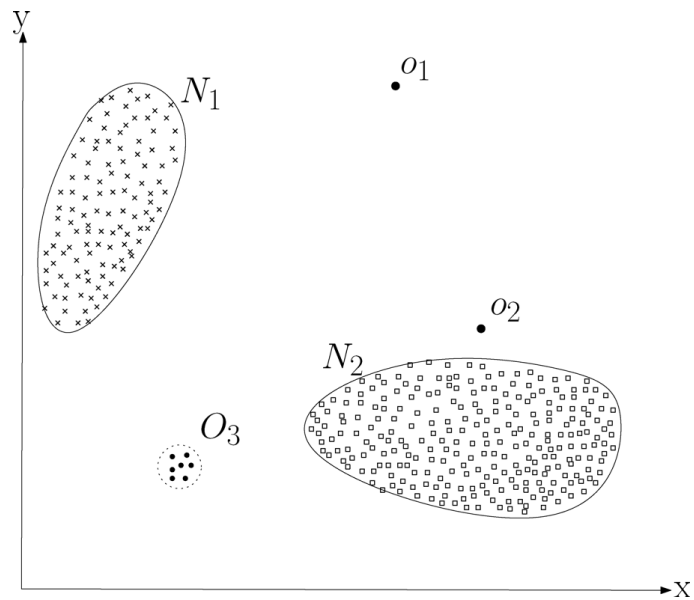
Anomalies can be classified into three categories

- *Point Anomalies*
- *Contextual Anomalies*
- *Collective Anomalies*

Different Aspects of An Anomaly Detection Problem

Point Anomalies

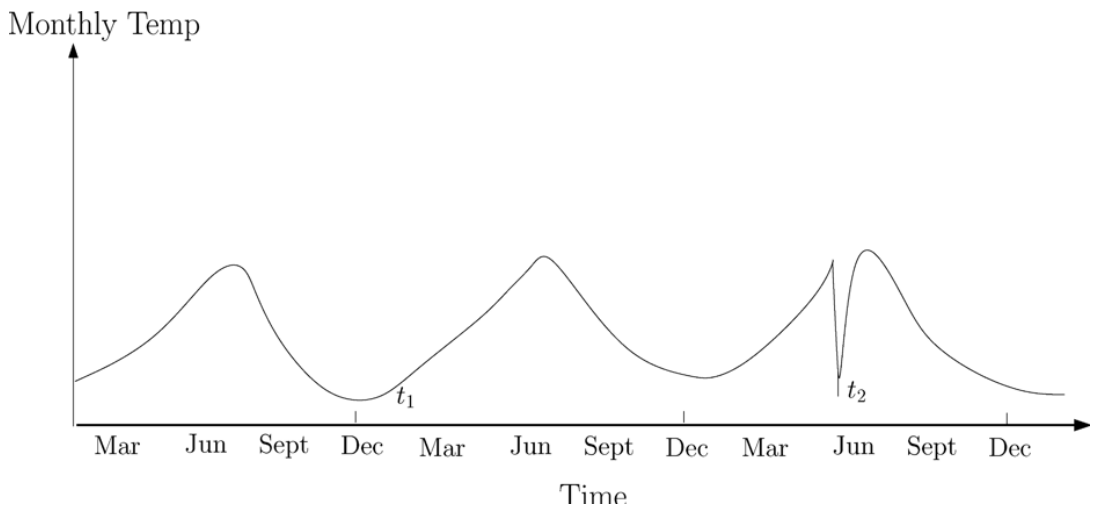
- If an **individual** data instance can be considered as anomalous with respect to the rest of the data, then the instance is termed a point anomaly
- This is the simplest type of anomaly and is the focus of majority of research on anomaly detection



Different Aspects of An Anomaly Detection Problem

Contextual Anomalies

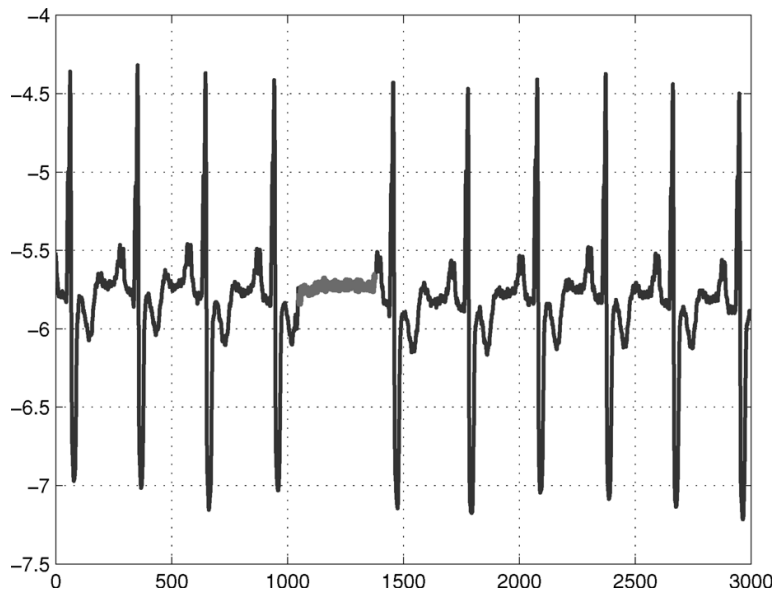
- If a data instance is anomalous **in a specific context**, but not otherwise, then it is termed a contextual anomaly
- Thus when a data instance is a contextual anomaly in a given context, but an identical data instance could be considered normal in a different context
- Most commonly explored in time-series data and spatial data



Different Aspects of An Anomaly Detection Problem

Collective Anomalies

- If a **collection of related data instances** is anomalous with respect to the entire data set, it is termed a collective anomaly
- Thus the individual instances in a collective anomaly may not be anomalies by themselves, but their occurrence together as a collection is anomalous



Data Labels

- Obtaining labeled data that is accurate is often prohibitively expensive
- Furthermore, getting a labeled set of anomalous data instances that covers all possible type of anomalous behavior is more difficult than getting labels for normal behavior (Because Anomalous behaviors are not in standardized manner)
- Based on the extent to which the labels are available, anomaly detection techniques can operate in one of the three modes:
  - *Supervised Anomaly Detection*
  - *Semi-Supervised Anomaly Detection*
  - *Unsupervised Anomaly Detection*

Different Aspects of An Anomaly Detection Problem

Supervised Anomaly Detection

- Require full accurate class labels
- Extremely imbalanced data set
- Similar to solve an imbalanced classification problem

Semi-Supervised Anomaly Detection

- Require labels only for the normal instances
- More widely applicable than supervised techniques
- Typical approach: to build a model for the normal behavior, and use the model to identify anomalies in the test data

Unsupervised Anomaly Detection

- Require no label information. Thus it is most widely applicable
- Requires implicit assumption which is "Anomalies are few and different from the normal points" (If this assumption is not true then such techniques suffer from high false alarm rate)

Output of Anomaly Detection

The outputs produced by anomaly detection techniques are one of the following two types:

Scores

- Scoring techniques report anomaly score to each instance in the test data depending of the degree to which that instance is considered an anomaly
- Thus the outputs can be represented in a ranked form

Labels

- These techniques assign a label (normal or anomalous) to each test instances

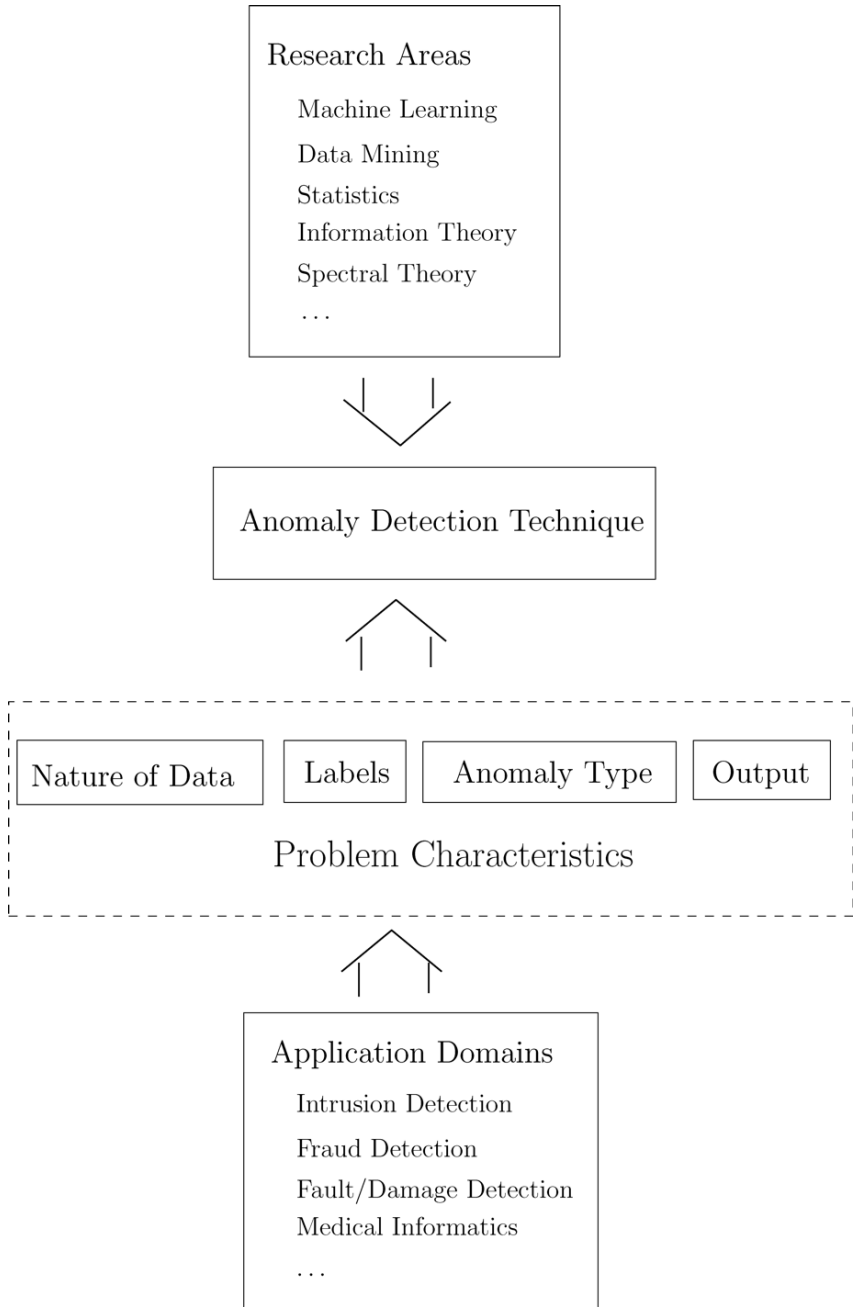


Figure: General approach for solving Anomaly Detection problem

Anomaly Detection Techniques

General Assumption

A classifier that can distinguish between normal and anomalous classes can be learned in the given feature space

Basic Idea

It works similar to classification problem in supervised learning except for that it only obtains *normal* class instances for training

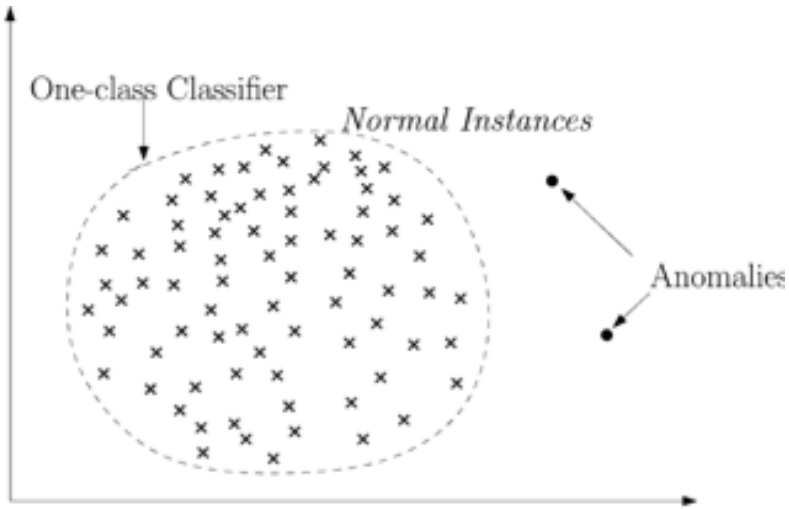
- Training Phase
  - Learns a classifier using the available labeled training data
- Testing Phase
  - Classifies a test instance as normal or anomalous, using the classifier

Two broad categories based on the availability of labels

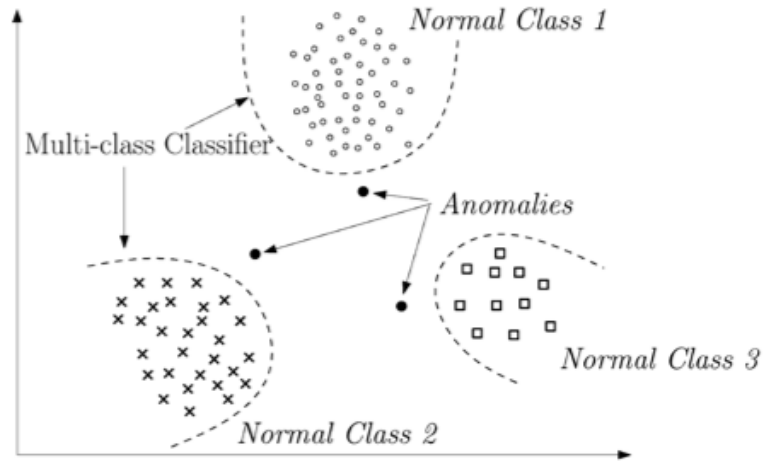
- One-Class Classification based anomaly detection
  - Assumption: all training instances have only one class label
  - Learns a discriminative boundary around the normal instances using a one-class classification algorithm
- Multi-Class Classification based anomaly detection
  - Assumption: the training data contains labeled instances belonging to multiple normal classes
  - Learns discriminative boundaries around the normal class instances

- Both techniques identify test instance as anomalous if it is not classified as normal by classifier

Two broad categories based on the availability of labels



(a) One-Class Classification



(b) Multi-Class Classification

Various Classification-Based Anomaly Detection Techniques

- Neural Networks-Based
- Bayesian Networks-Based
- Support Vector Machines-Based
- ...

Computational Complexity

- It mainly depends on the classification algorithm being used
- The testing phase is usually very fast since the testing phase uses a learned model for classification

Advantages and Disadvantages

(+) It can make use of powerful algorithms that can distinguish between instances belonging to different classes

(+) The testing phase is fast, since each test instance needs to be compared against the precomputed model

(−) Require the availability of accurate labels which is often not possible

(−) Return binary output, normal or anomalous, which can be a disadvantage when a meaningful anomaly score (or ranking) is desired

(some techniques obtain a probabilistic prediction score to overcome this problem)

General Assumption

Normal data instances occur in dense neighborhoods, while anomalies occur far from their closest neighbors

Basic Idea

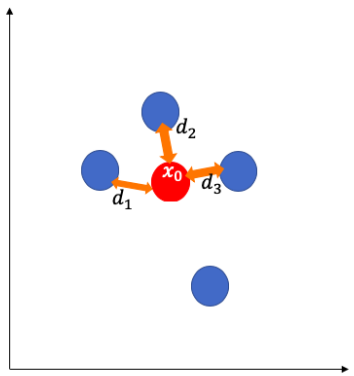
- Lazy learning algorithm (Training stage does just storing the training data)
- Training Phase
  - Store training data
- Testing Phase
  - Compute specified distance between test instance and all training instances
- For a test instance, if computed distance is high, then it is identified as anomalous

Two Broad Categories of Nearest Neighbor Techniques

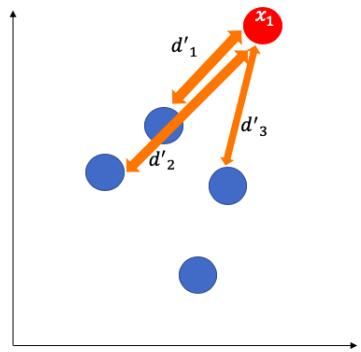
- Using Distance to k^{th} Nearest Neighbor
  - Use the distance of a data instance to its k^{th} nearest neighbor as the anomaly score
- Using Local Density of test instance
  - use the inverse of local density of each data instance to compute its anomaly score

Using Distance to k^{th} Nearest Neighbor

- The anomaly score of a data instance is defined as its distance to its k^{th} nearest neighbor in a given data set



(a) Normal point

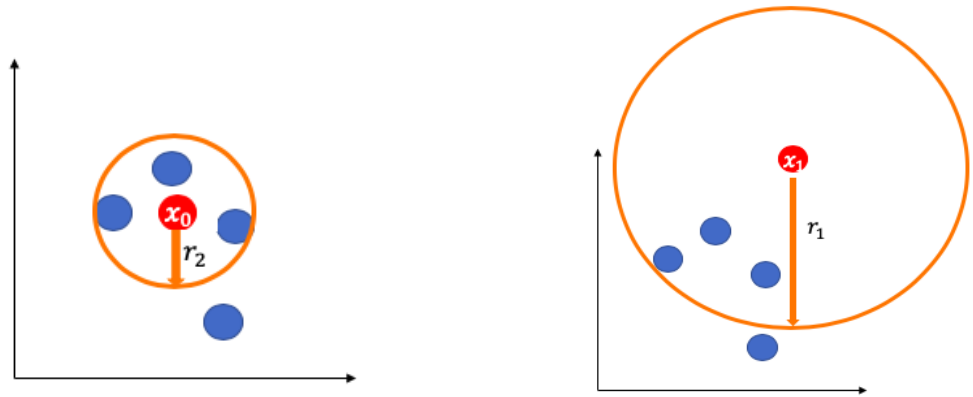


(b) Anomaly point

- Basic Extensions
  - Modifies the definition to obtain anomaly score (e.g. average,max, $\cdots$)
  - Use different distance measure to handle complex data type
  - Improve the computational efficiency

Density-Based

- An instance that lies in a neighborhood with low density is declared to be anomalous while an instance that lies in a dense neighborhood is declared to be normal



(a) Normal point (b) Anomaly point

- Density of the instance is usually estimated as k divided by the radius of hypershpere, centered at the given data instance, which contains k other instances

Problem of Basic Nearest Neighbor Method

- If the data has regions of varying densities, the basic nearest neighbor method fails to detect anomalies

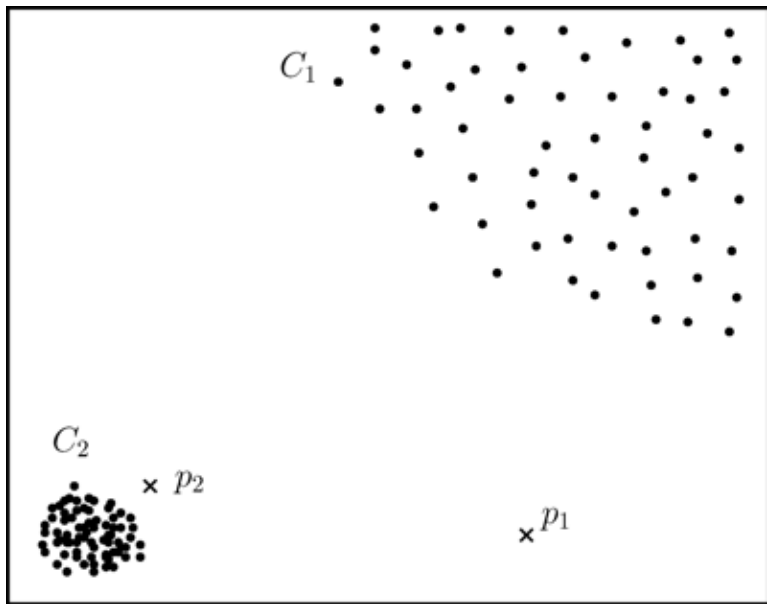


Figure: Problem of Basic Nearest Neighbor method, retrieved from [2]

- To handle this issue, Relative Density method has been proposed

Using Relative Density

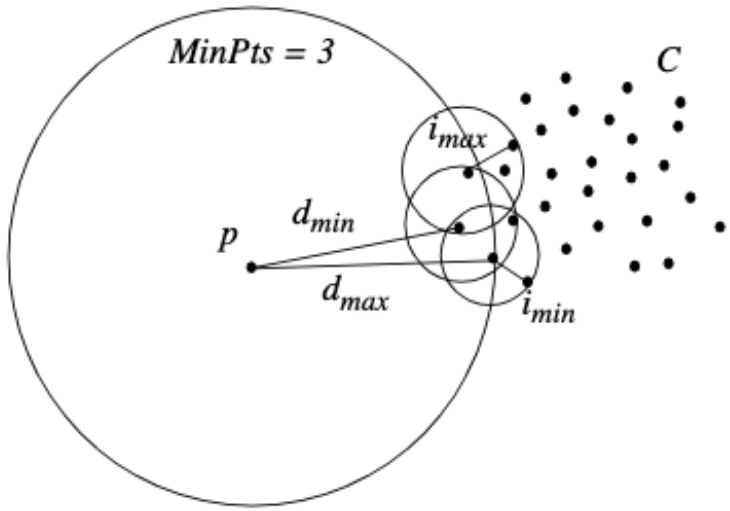


Figure: Local Outlier Factor

- Basic Extensions
  - Estimate the local density in a different way
  - Compute more complex data types
  - Improve the computational efficiency

Advantages and Disadvantages

(+) Unsupervised in nature

(+) Adapting nearest neighbor-based techniques to a different data type is straightforward, and primarily requires defining a appropriate distance measure for the given data

(−) If the general assumption does not conform to the given data set, the detector fails to catch anomalies

(−) The computational complexity of the testing phase is a challenge $O(N^2)$. Since it has to compute all distances between each test instance to all training data points

(−) Performance greatly relies on a distance measure

Clustering Based

- Clustering is used to group similar data instances into clusters
- Clustering-Based anomaly detection techniques can be grouped into three categories based on its assumption
  - Normal data instances belong to a cluster in the data, while anomalies do not belong to any cluster
  - Normal data instances lie close to their closest cluster centroid, while anomalies are far away from their closest cluster centroid
  - Normal data instances belong to large and dense clusters, while anomalies either belong to small or sparse clusters

First Category

- Assumption
  - Normal data instances belong to a cluster in the data, while anomalies do not belong to any cluster
- Several clustering algorithms that do not force every data instance to belong to a cluster can be used (e.g. DBSCAN, ROCK, SNN, ...)
- Disadvantage
  - They are not optimized to find anomalies, since the main aim of the underlying clustering algorithm is to find clusters

First Category

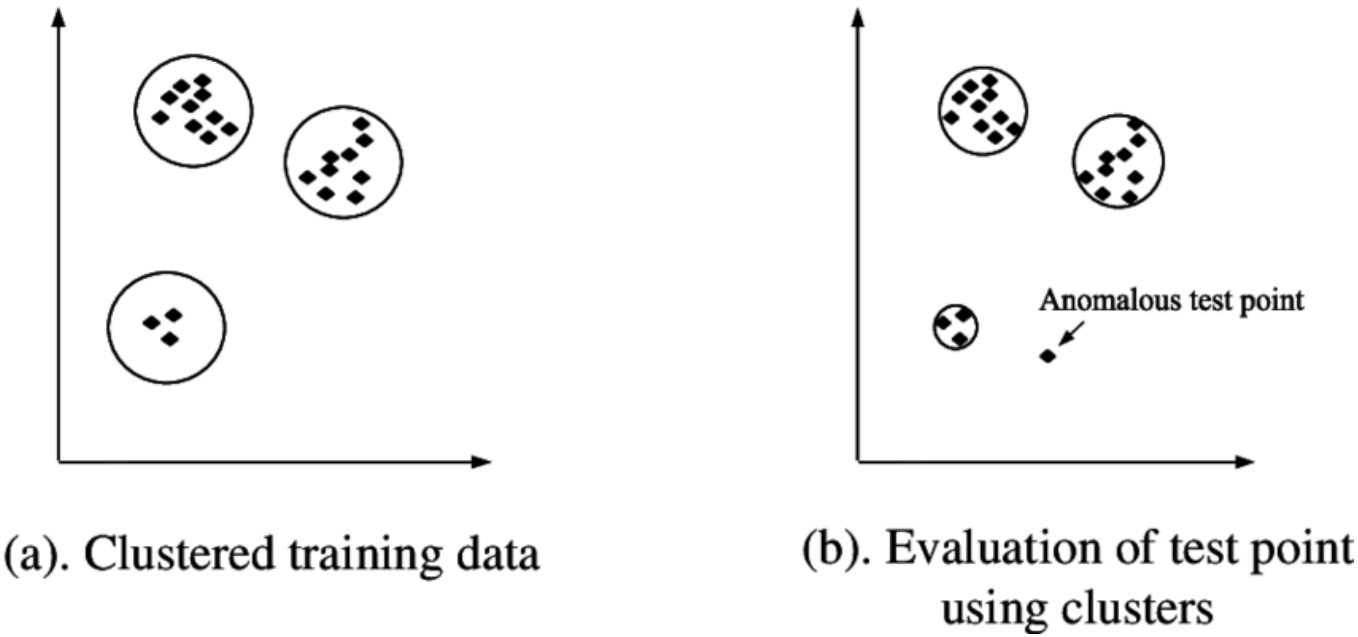


Figure: Clustering-based anomaly detection of first category, retrieved from [3]

Second Category

- Assumption
  - Normal data instances lie close to their closest cluster centroid, while anomalies are far away from their closest cluster centroid
- These kind of techniques are operated in two-steps
  1. The data is clustered using a clustering algorithm
  2. For each data instance, its distance to its closest cluster centroid is calculated as its anomaly score
- Various clustering algorithms can be obtained (e.g. Self- Organizing Maps (SOM), K-means Clustering, $\cdots$)
- Disadvantage
  - If the anomalies in the data form clusters by themselves, these techniques will not be able to detect such anomalies

Second Category

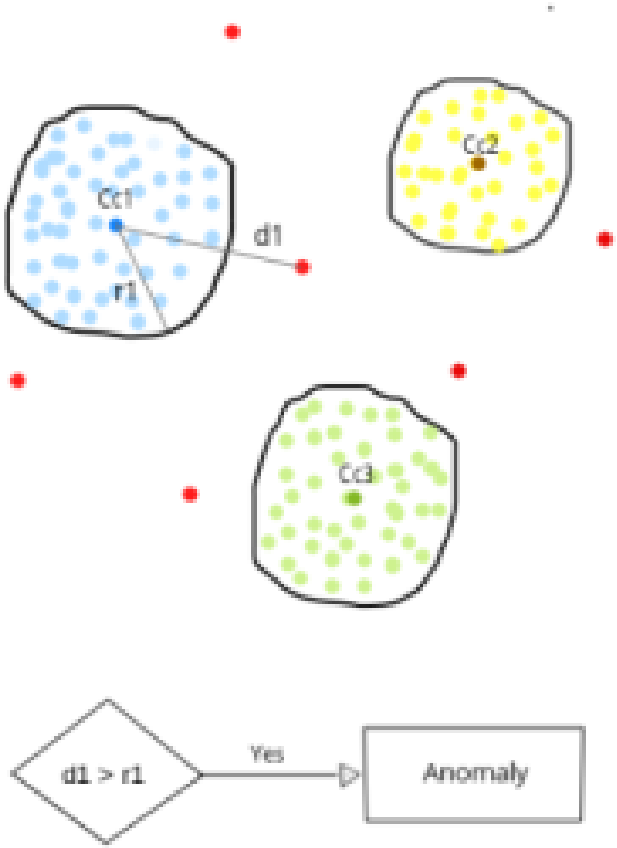


Figure: Clustering-based anomaly detection of second category, retrieved from [4]

Third Category

- Assumption
  - Normal data instances belong to large and dense clusters, while anomalies either belong to small or sparse clusters
- Techniques based on this assumption declare instances belonging to clusters whose size and/or density is below a certain threshold, as anomalous
- Several variations have been proposed for this categorry
  - For example, Cluster-Bassed Local Outlier Factor ($CBLOF$) captures the size of the cluser to which the data instance belongs, as well as the distance of the data instance to its cluster centroid

Advantages and Disadvantages

(+) Can operate in an unsupervised mode

(+) Can be adapted to other complex data types by simply plugging in a clustering algorithm that can handle the particular data type

(+) The testing phase is fast since the number of clusters against which every test instance needs to be compared is a small constant

(−) Performance is highly dependent on the effectiveness of clustering algorithms

(−) They are not optimized for anomaly detection

(−) The computational complexity for clustering the data is often a bottleneck

General Assumption

Normal data instances occur in high probability regions of a stochastic model, while anomalies occur in the low probability regions of the stochastic model

Basic Idea

- Fit a statistical model to the given data (usually for normal behavior) and then apply a statistical inference test to determine if an unseen instance belongs to this model or not

Two broad categories of statistical anomaly detection

- *Parametric Techniques*
  - which assume the knowledge of the underlying distribution and estimate the parameters from the given data
- *Nonparametric Techniques*

Parametric Techniques

- Assume that the normal data is generated by a parametric distribution with parameters Θ and probability density function $f(\mathbf{x}, \Theta)$
- The anomaly score of a test instance $\mathbf{x}$ is the inverse of the probability density function, $f(\mathbf{x}, \Theta)$. Θ is estimated from the given data
- Alternatively, a statistical hypothesis test may be used
  - The null hypothesis H_{o}: the data instance $\mathbf{x}$ has been generated using the estimated distribution (with parameters Θ)
  - If the statistical test rejects H_{o}, $\mathbf{x}$ is declared to be anomaly
  - Since a statistical hypothesis test is associated with a test statistic, the test statistic value can be obtained as an anomaly score for $\mathbf{x}$

Parametric Techniques

- Basic Example: Gaussian Model-Based
  - Assume that the data is generated from a Gaussian distribution
  - The parameters are estimated using $Maximum\ Likelihood\ Estimates$(MLE)
  - The distance of a data instance to the estimated mean is the anomaly score for that instance
  - A certain threshold is applied to the anomaly scores to determine the anomalies (e.g. $\mu \pm 3\sigma$, since this region contains 99.7% of the data)

Nonparametric Techniques

- The model structure is not defined a priori, but is instead determined from given data
- Basic Example: Histogram-Based
  1. Build a histogram based on the different values in the training data
  2. Check if a test instance falls in any one of the bins of the histogram
  - The size of the bin used when building the histogram is a key for anomaly detection
  - Alternatively the height of the bin that contains the instance can be obtained as an anomaly score

Advantages and Disadvantages

(+) If the underlying data distribution assumption holds true, they provide a statistically justifiable solution for anomaly detection

(+) The anomaly score is associated with a confidence interval, which can be used as additional information

(−) They rely on the assumption that the data is generated from a particular distribution.

(−) Choosing the best statistic is often not a straightforward task

General Assumption

Data can be embedded into a lower dimensional subspace in which normal instances and anomalies appear significantly different

Basic Idea

- Spectral techniques try to find an approximation of the data using a combination of attributes that capture the bulk of the variability in the data
- Thus these techniques seek to determine the subspaces in which the anomalous instances can be easily identified
- Basic Example: Principal Component Analysis (PCA)
  - Projecting the data into lower dimensional space while maximizing the variability
  - Compute the distance between a data instance and the subspace, generated by PCA as an anomaly score

Spectral Anomaly Detection

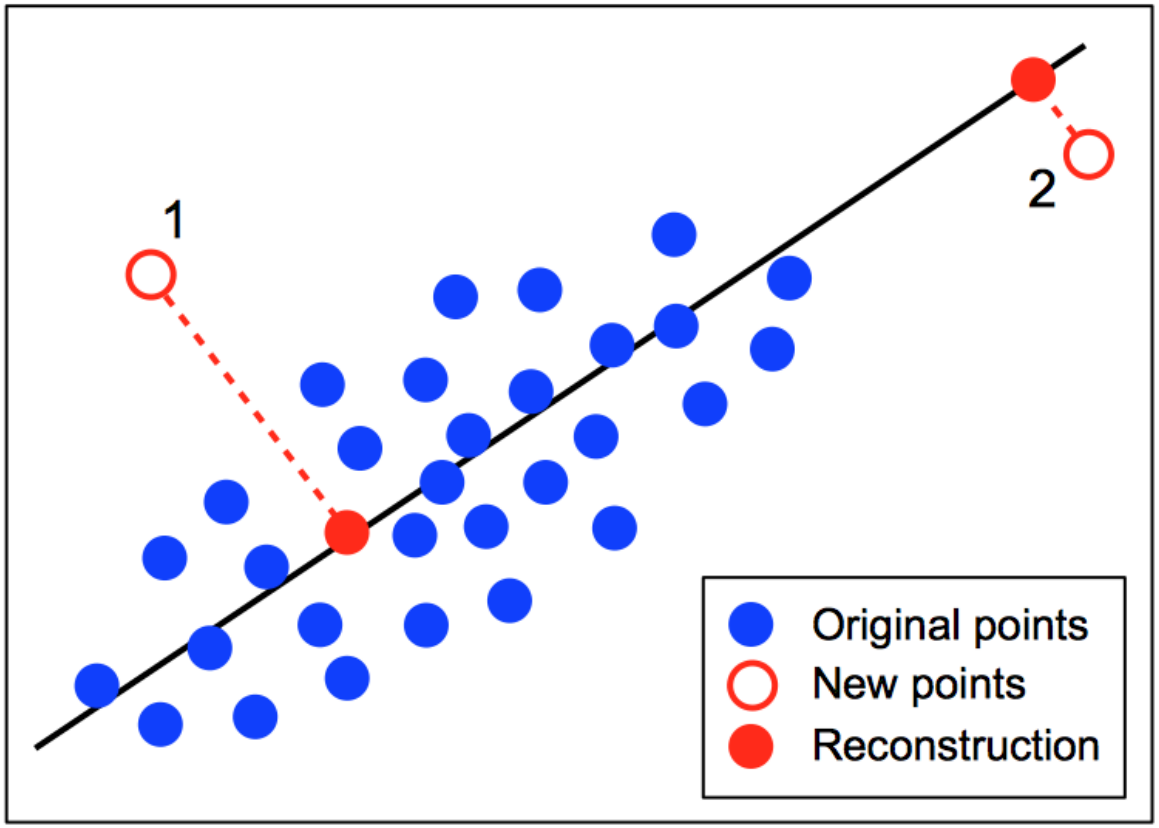


Figure: Simple illustration of PCA-based anomaly detection, retrieved from [5]

Advantages and Disadvantages

(+) These techniques automatically perform dimensionality reduction and hence are suitable for handling high dimensional data

(+) Thus they can also be used as a preprocessing step before applying any existing anomaly detection technique

(+) Can be used in unsupervised setting

(−) Useful only if the normal and anomalous instances are separable in the lower dimensional embedding of the data

(−) High computational complexity

Conclusion

Relative Strengths and Weaknesses of Anomaly Detection Techniques

- Each of the anomaly detection techniques discussed in the previous sections have their unique strengths and weaknesses
- Thus it is important to know which anomaly detection technique is best suited for a given problem

Concluding Remarks

- We have discussed the way that anomaly detection problem can be characterized and have provided a overview of the various techniques
- For each category, we have identified a unique assumption regarding the notion of normal and anomalous data
- When applying a given technique to a particular domain, these assumptions can be used as a guidelines for application

The Contribution of this paper

Table I. Comparison of our Survey to Other Related Survey Articles. 1—Our Survey, 2—Hodge and Austin [2004], 3—Agyemang et al. [2006], 4—Markou and Singh [2003a], 5—Markou and Singh [2003b], 6—Patcha and Park [2007], 7—Beckman and Cook [1983], 8—Bakar et al. [2006]

		1	2	3	4	5	6	7	8
Techniques	Classification Based	√	√	√	√		√		
	Clustering Based	√	√	√			√		
	Nearest Neighbor Based	√	√	√			√		√
	Statistical	√	√	√		√	√	√	√
	Information Theoretic	√							
	Spectral	√							
Applications	Cyber-Intrusion Detection	√					√		
	Fraud Detection	√							
	Medical Anomaly Detection	√							
	Industrial Damage Detection	√							
	Image Processing	√							
	Textual Anomaly Detection	√							
	Sensor Networks	√							

- This survey is an attempt to provide a structured and broad overview of extensive research on anomaly detection techniques spanning multiple research areas and application domains

References

1 Chandola, Varun, Arindam Banerjee, and Vipin Kumar. "Anomaly detection: A survey." ACM computing surveys (CSUR) 41, no. 3 (2009): 1-58.

2 Breunig, Markus M., Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander. "LOF: identifying density-based local outliers." In Proceedings of the 2000 ACM SIGMOD international conference on Management of data, pp. 93-104. 2000.

3 Eik Loo, Chong, Mun Yong Ng, Christopher Leckie, and Marimuthu Palaniswami. "Intrusion detection for routing attacks in sensor networks." International Journal of Distributed Sensor Networks 2, no. 4 (2006): 313-332.

4 Anomaly Detection Using K means Clustering, Ashen Weerathunga, https://ashenweerathunga.wordpress.com/2016/01/05/anomaly-detection-using-k-means-clustering/

5 Anomaly detection using PCA reconstruction error, https://stats.stackexchange.com/questions/259806/anomaly-detection-using-pca-reconstruction-error