UNIVERSITÀ
DEGLI STUDI
FIRENZE

ANOMALY DETECTION: A SURVEY

VARUN CHANDOLA, ARINDAM BANERJEE AND VIPIN KUMAR

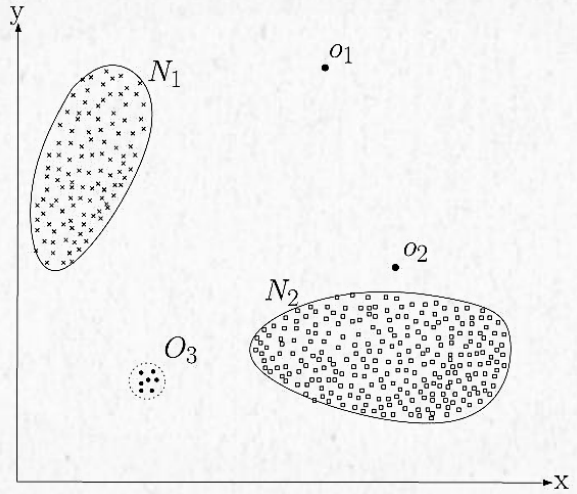
ACM COMPUTING SURVEYS, VOL. 41, N°3, ARTICLE 15, PUBLICATION DATE: JULY 2009

INTRODUCTION

- *Anomaly detection* refers to the problem of finding patterns in data that do not conform to expected behavior *(anomalies, outliers, exception, peculiarities, ecc.).*

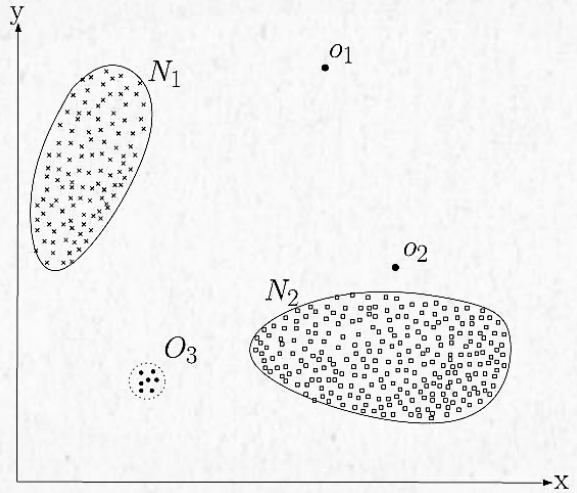
INTRODUCTION

- *Anomaly detection* refers to the problem of finding patterns in data that do not conform to expected behavior *(anomalies, outliers, exception, peculiarities, ecc.).*



INTRODUCTION

- *Anomaly detection* refers to the problem of finding patterns in data that do not conform to expected behavior *(anomalies, outliers, exception, peculiarities, ecc.).*



- Wide variety of applications (fraud detection, intrusion detection, fault detection)

ANOMALY DETECTION

- What is anomaly detection?
  - Anomalies are patterns in data that do not conform to a well defined notion of normal behavior
  - Anomalies are interesting to the analyst
  - Anomaly detection aims finding these patterns

ANOMALY DETECTION

- What is anomaly detection?
  - Anomalies are patterns in data that do not conform to a well defined notion of normal behavior
  - Anomalies are interesting to the analyst
  - Anomaly detection aims finding these patterns

- What is not anomaly detection?
  - Anomaly detection is distinct from *noise removal* or *noise accomodation*
  - Noise can be defined as a phenomenon in data that is not of interest to the analyst but act as a hindrance to data analysis.

CHALLENGES

- Define a region representing a normal behavior and declare any observation in the data that does not belong to this normal region as an anomaly:
  - Defining a region that encompasses every possible normal behavior is very difficult
  - Boundary between normal and anomalous behavior is often not precise
  - Malicious adversaries make the anomalous observations appear normal
  - Normal behavior can keeps evolving
  - Exact notion of anomaly is different for different application domains
  - Data contains noise

NATURE OF INPUT DATA

- Each data instance can be described using a set of attributes
- Attributes can be of different types
- Univariate/multivariate (in the second case attributes might be a mixture of types)

NATURE OF INPUT DATA

- Each data instance can be described using a set of attributes
- Attributes can be of different types
- Univariate/multivariate (in the second case attributes might be a mixture of types)
- *Sequence data:*
  - Instances are linearly ordered (time-series, genome, ecc.)

NATURE OF INPUT DATA

- Each data instance can be described using a set of attributes
- Attributes can be of different types
- Univariate/multivariate (in the second case attributes might be a mixture of types)
- *Sequence data:*
  - Instances are linearly ordered (time-series, genome, ecc.)
- *Spatial data:*
  - Instances are related to their neighboring instances (vehicular traffic data, ecological, ecc.)

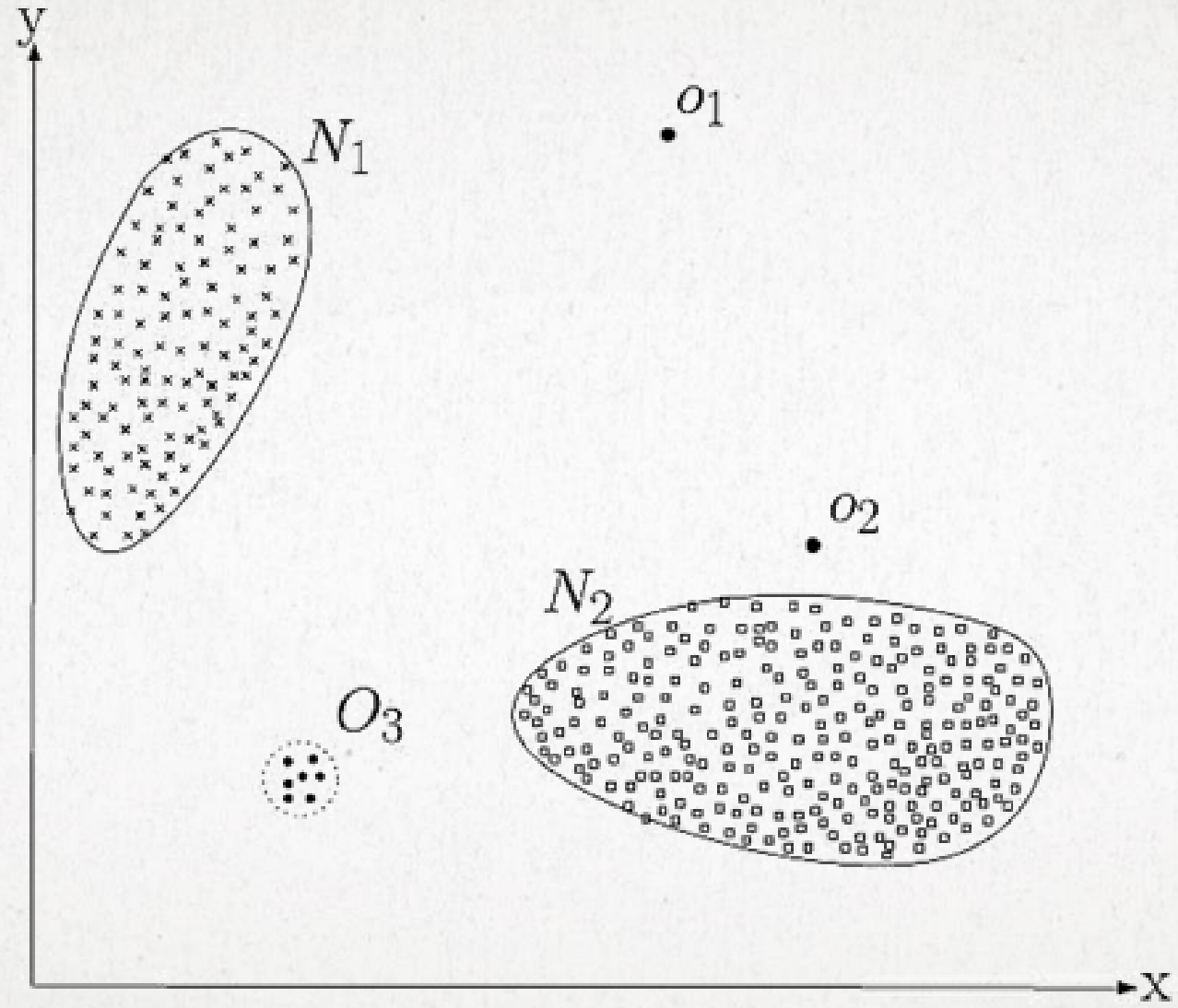
NATURE OF INPUT DATA

- Each data instance can be described using a set of attributes
- Attributes can be of different types
- Univariate/multivariate (in the second case attributes might be a mixture of types)
- *Sequence data:*
  - Instances are linearly ordered (time-series, genome, ecc.)
- *Spatial data:*
  - Instances are related to their neighboring instances (vehicular traffic data, ecological, ecc.)
- *Graph data:*
  - Instances are represented as vertices in a graph connected with other vertices with edges

TYPE OF ANOMALY

- *Point anomalies*
  - An individual data instance considered as anomalous with respect to the rest of data

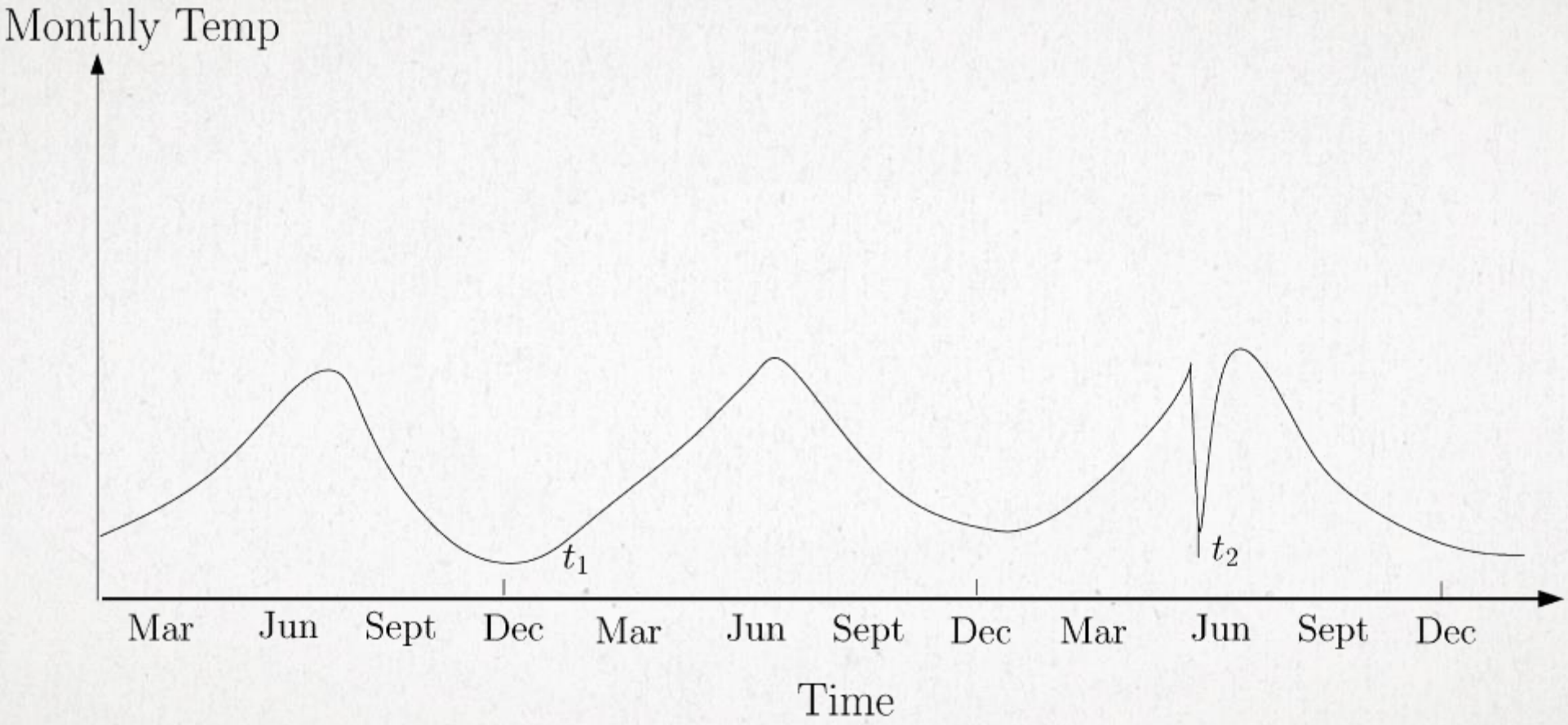
TYPE OF ANOMALY



TYPE OF ANOMALY

- *Point anomalies*
  - An individual data instance considered as anomalous with respect to the rest of data
- *Contextual anomalies*
  - A data instance can be anomalous in a specific context but not otherwise

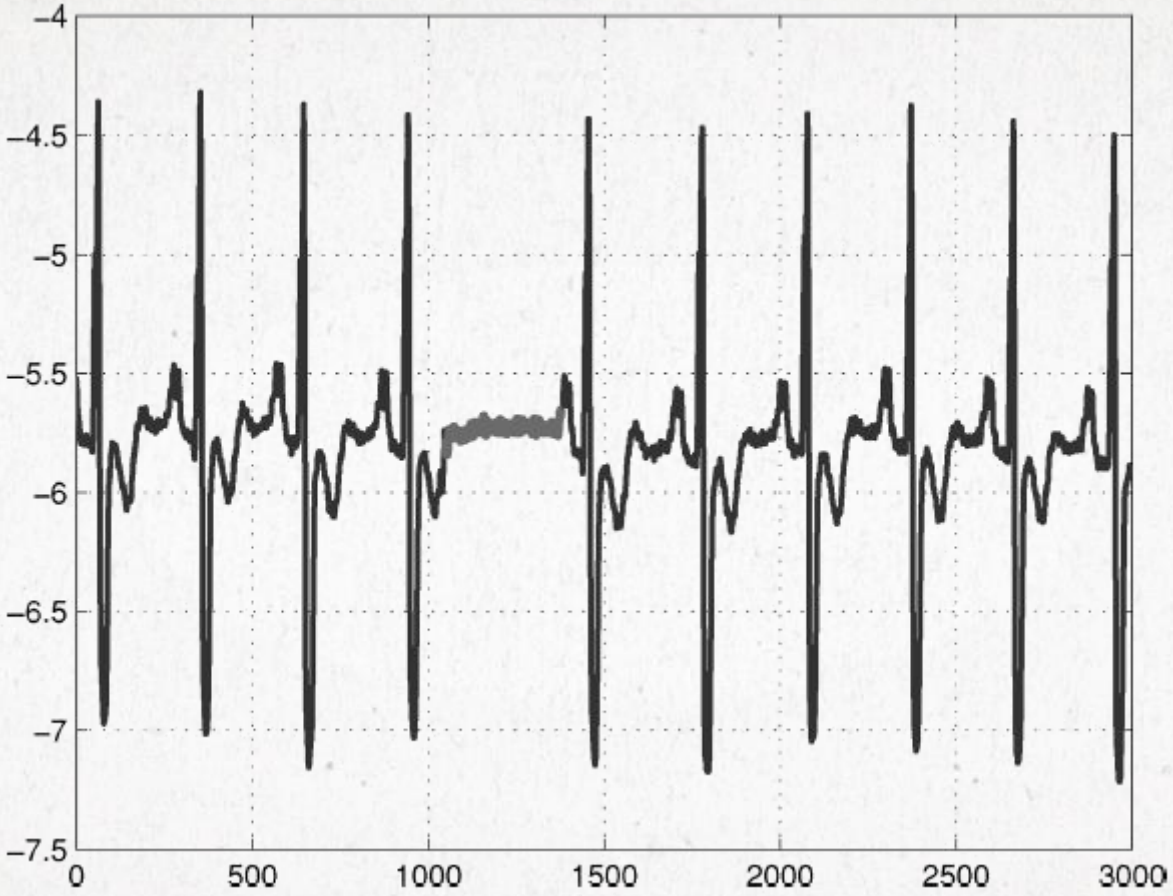
TYPE OF ANOMALY



TYPE OF ANOMALY

- *Point anomalies*
  - An individual data instance considered as anomalous with respect to the rest of data
- *Contextual anomalies*
  - A data instance can be anomalous in a specific context but not otherwise
- *Collective anomalies*
  - A collection of related data instances considered as anomalous with respect to the entire data set

TYPE OF ANOMALY



DATA LABELS

- The labels associated with a data instance denote wheter that instance is *normal* or *anomalous*
- Labelling is often done manually by a human expert
- Getting a labeled set of anomalous data instances that covers all possible type of anomalous behavior is more difficult than getting labels for normal behaviors

DATA LABELS

- The labels associated with a data instance denote wheter that instance is *normal* or *anomalous*
- Labelling is often done manually by a human expert
- Getting a labeled set of anomalous data instances that covers all possible type of anomalous behavior is more difficult than getting labels for normal behaviors
  - *Supervised anomaly detection techniques*
    - Availability of a training data set that has labeled instances for normal as well as anomaly classes

DATA LABELS

- The labels associated with a data instance denote wheter that instance is *normal* or *anomalous*
- Labelling is often done manually by a human expert
- Getting a labeled set of anomalous data instances that covers all possible type of anomalous behavior is more difficult than getting labels for normal behaviors
  - *Supervised anomaly detection techniques*
    - Availability of a training data set that has labeled instances for normal as well as anomaly classes
  - *Semisupervised anomaly detection techniques*
    - Training data has labeled instances only for the normal class

DATA LABELS

- The labels associated with a data instance denote wheter that instance is *normal* or *anomalous*
- Labelling is often done manually by a human expert
- Getting a labeled set of anomalous data instances that covers all possible type of anomalous behavior is more difficult than getting labels for normal behaviors
  - *Supervised anomaly detection techniques*
    - Availability of a training data set that has labeled instances for normal as well as anomaly classes
  - *Semisupervised anomaly detection techniques*
    - Training data has labeled instances only for the normal class
  - *Unsupervised anomaly detection techniques*
    - Do not require training data

OUTPUT OF ANOMALY DETECTION

- *Labels*
  - Techniques in this category assign a label to each test instance *(normal* or *anomalous)*
  - Binary labels

OUTPUT OF ANOMALY DETECTION

- *Labels*
  - Techniques in this category assign a label to each test instance *(normal* or *anomalous)*
  - Binary labels

- *Scores*
  - Scoring techniques assign an anomaly score to each instance in the test data depending on the degree to wich that instance is considered anomaly
  - Ranked list

APPLICATIONS

- *Intrusion detection*
  - Detection of malicious activity (break-ins, penetrations, and other form of computer abuse)
  - Huge volume of data
  - Data tipically comes in streaming fashion, thereby requiring online analysis
  - False alarm rate due to large sized input
  - Semisupervised and unsupervised techniques are preferred
  - *Host-based intrusion detection systems*
  - *Network-based intrusion detection system*

APPLICATIONS

- *Host-based intrusion detection systems*
  - Anomalous subsequences in operating system call traces (collective anomalies)

APPLICATIONS

open,	read,	mmap,	mmap,	open,	read,	mmap	...
open,	mmap,	mmap,	read,	open,	close	...	
open,	close,	open,	close,	open,	mmap,	close	...

APPLICATIONS

- *Host-based intrusion detection systems*
  - Anomalous subsequences in operating system call traces (collective anomalies)
  - Co-occurence of events is the key factor in differentiating between normal or anomalous
  - Alphabet is usually large (183 system calls for SunOS 4.1x OS)
  - The length of the sequence for each program varies
  - Sequencial data
  - Point anomaly detection techniques are not applicable in this domain

APPLICATIONS

- *Host-based intrusion detection systems*
  - Anomalous subsequences in operating system call traces (collective anomalies)
  - Co-occurence of events is the key factor in differentiating between normal or anomalous
  - Alphabet is usually large (183 system calls for SunOS 4.1x OS)
  - The length of the sequence for each program varies
  - Sequencial data
  - Point anomaly detection techniques are not applicable in this domain
- *Network-based intrusion detection systems*
  - Detecting intrusion in network data (point anomalies)
  - Data available can be at different levels of granularity (packet level, CISCO net-flows data)
  - The nature of anomalies keeps changing over time

APPLICATIONS

- *Fraud detection*
  - Detection of criminal activities occurring in commercial organizations (banks, insurance ecc.)
  - Users consume the resources provided by the organization in an unauthorized way
  - Needs an immediate detection
  - Maintain usage profile for each customer and detect any deviations
  - Credit card, mobile phone, insurance claim, insider trading fraud detection

APPLICATIONS

- *Fraud detection*
  - Detection of criminal activities occurring in commercial organizations (banks, insurance ecc.)
  - Users consume the resources provided by the organization in an unauthorized way
  - Needs an immediate detection
  - Maintain usage profile for each customer and detect any deviations
  - Credit card, mobile phone, insurance claim, insider trading fraud detection
  - *Credit card fraud detection*
    - Data is typically comprised of records defined over several dimensions
    - The frauds are typically reflected in transactional records (point anomalies)
    - Profiling and clustering based techniques are typically used in this domain
    - Contextual anomalies

APPLICATIONS

- *Medical and public health anomaly detection*
  - Patient records (age, blood group, weight,…)
  - Abnormal patient condition, instrumentation errors, recording errors…
  - Is a very critical problem in this domain and requires a high degree of accuracy
  - Data might have a temporal as well as spatial aspect too
  - Detecting anomalous records (point anomalies)
  - Semisupervised approach
  - Cost of classifying an anomaly as normal can be very high

APPLICATIONS

- *Image processing* (satellite imagery, mammographic image analysis, video surveillance..)
  - Motion or insertion of a foreign object
  - Data has spatial and temporal aspects
  - Continuous attribute(color, lightness, texture…)
  - Key challenge in this domain is the large size of the input

APPLICATIONS

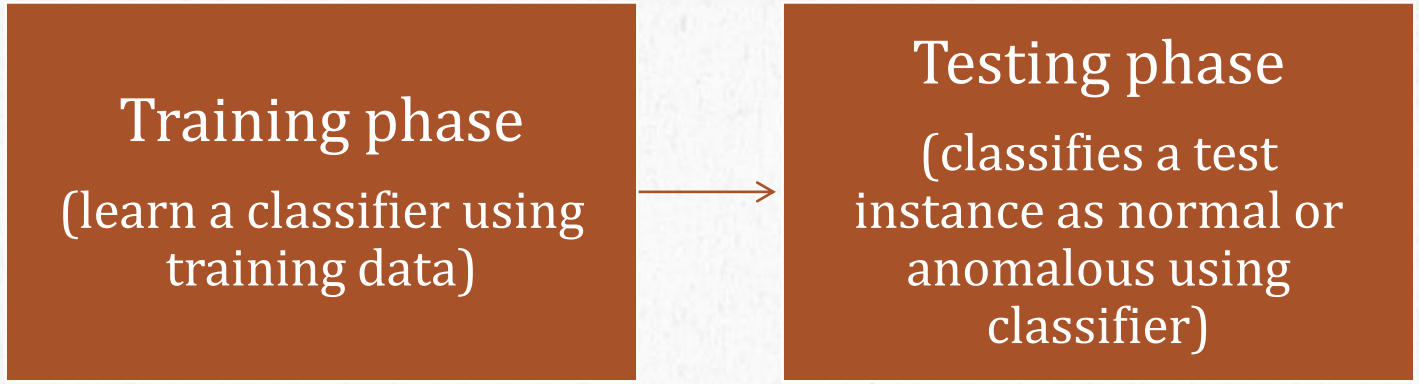
- *Sensor networks*
- *Speech recognition*
- *Traffic monitoring*
- *Anomalous biological data*
- *Associations among criminal activities*
- *Anomalies in astronomical data*
- *Ecosystem disturbance*
- *Text data*
- *……*

CLASSIFICATION BASED TECHNIQUES

- Classification is used to learn a model *(classifier)* from a set of labeled data instances *(training)* and then, classify a test instance into one of the classes using the model *(testing)*

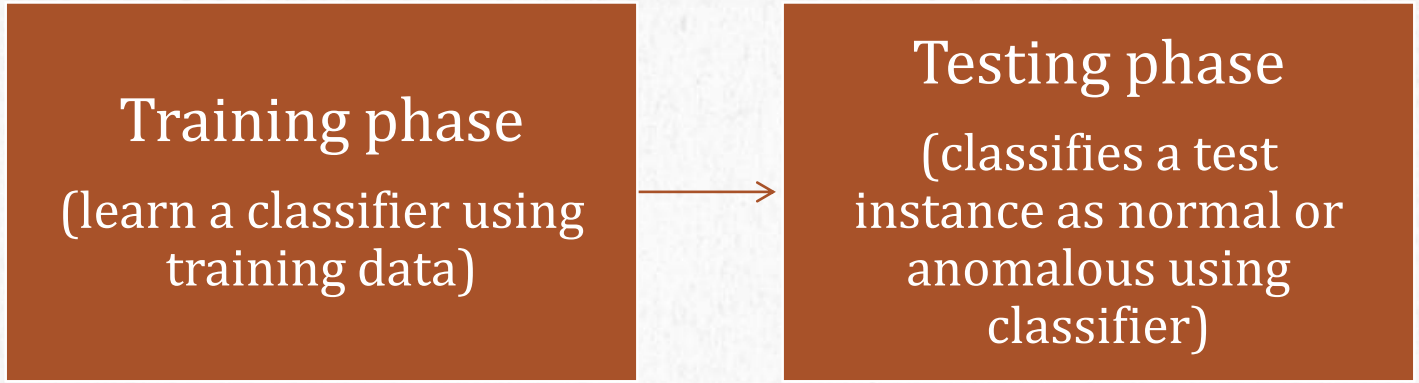
CLASSIFICATION BASED TECHNIQUES

- Classification is used to learn a model *(classifier)* from a set of labeled data instances *(training)* and then, classify a test instance into one of the classes using the model *(testing)*



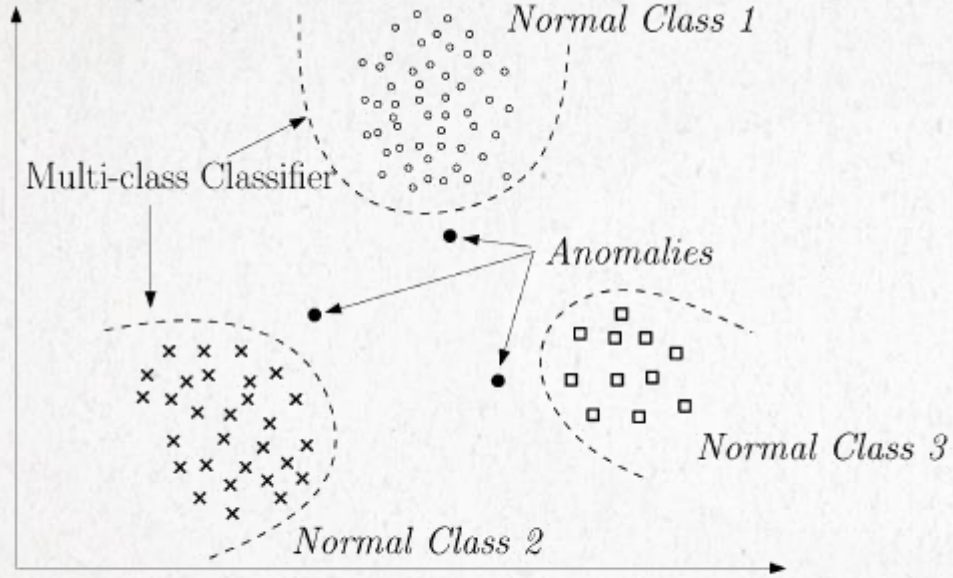
CLASSIFICATION BASED TECHNIQUES

- Classification is used to learn a model *(classifier)* from a set of labeled data instances *(training)* and then, classify a test instance into one of the classes using the model *(testing)*

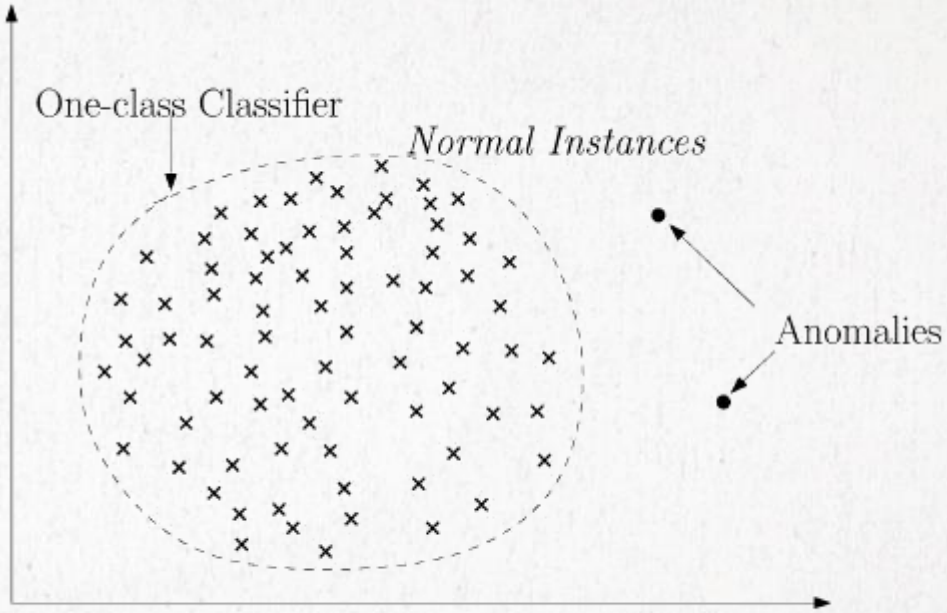


- *Multi-class* or *one-class* anomaly detection techniques

CLASSIFICATION BASED TECHNIQUES



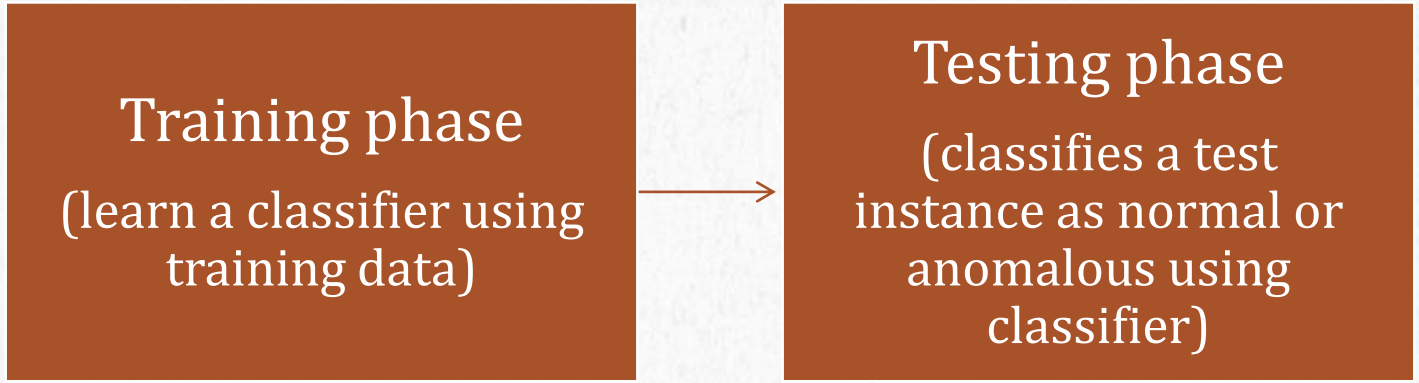
(a) Multi-class Anomaly Detection



(b) One-class Anomaly Detection

CLASSIFICATION BASED TECHNIQUES

- Classification is used to learn a model *(classifier)* from a set of labeled data instances *(training)* and then, classify a test instance into one of the classes using the model *(testing)*



- *Multi-class* or *one-class* anomaly detection techniques
- Neural network
- Rule based

NEURAL NETWORK BASED TECHNIQUES

- Multi-class as well as one-class settings
- A neural network is trained on the normal training data to learn the different normal classes.
- Each test instance is provided as an input to the neural network
- If the network accepts the test input, it is normal, otherwise, it is an anomaly
- Several variant of basic neural network have been proposed

RULE-BASED TECHNIQUES

- Rule-based anomaly detection techniques learn rules that capture the normal behavior of a system
- Multi-class as well as one-class settings
- Each rule has an associated confidence value that is proportional to ratio between the number of training instances correctly classified by the rule and the total number of training instances covered by the rule.
- The inverse of the confidence associated with the best rule is the anomaly score of the test instance

RULE-BASED TECHNIQUES

Learn rules from the trainig data using a learning algorithm (RIPPER, decision trees,…)

RULE-BASED TECHNIQUES

Learn rules from the trainig data using a learning algorithm (RIPPER, decision trees,…)

Find, for each test instance, the rule that best captures the test instance

COMPUTATIONAL COMPLEXITY

- The computational complexity of classification based techniques depends on the classification algorithm being used
- Generally, training decision trees tend to be faster, while techniques that involve quadratic optimization, such *support-vector machines based,* are more expensive
- Testing phase is usually very fast since the testing phase uses a lerned model for classification

ADVANTAGES AND DISADVANTAGES

- Advantages
  - Classification-based techniques, especially the multi-class techniques, can make use of powerful algorithms
  - The testing phase is fast, since each test instance needs to be compared against the precomputed model

ADVANTAGES AND DISADVANTAGES

- Advantages
  - Classification-based techniques, especially the multi-class techniques, can make use of powerful algorithms
  - The testing phase is fast, since each test instance needs to be compared against the precomputed model
- Disadvantages
  - Multi-class classification based techniques rely on the availability of accurate labels fo various normal classes, wich is often not possible
  - Classification-based techniques assign a binary label to each test instance (become a disadvantage when a meaningful anomaly score is desired)

NEAREST NEIGHBOUR-BASED TECHNIQUES

- *Assumption:* normal data instances occur in dense neighborhoods, while anomalies occur far from their closest neighbors
- Require a distance or similarity measure defined between two data instances
- Two categories:
  - Using distance of a data instance to its k^{th} nearest neighbor as the anomaly score
  - Computing the relative density of each data instance to compute the anomaly score

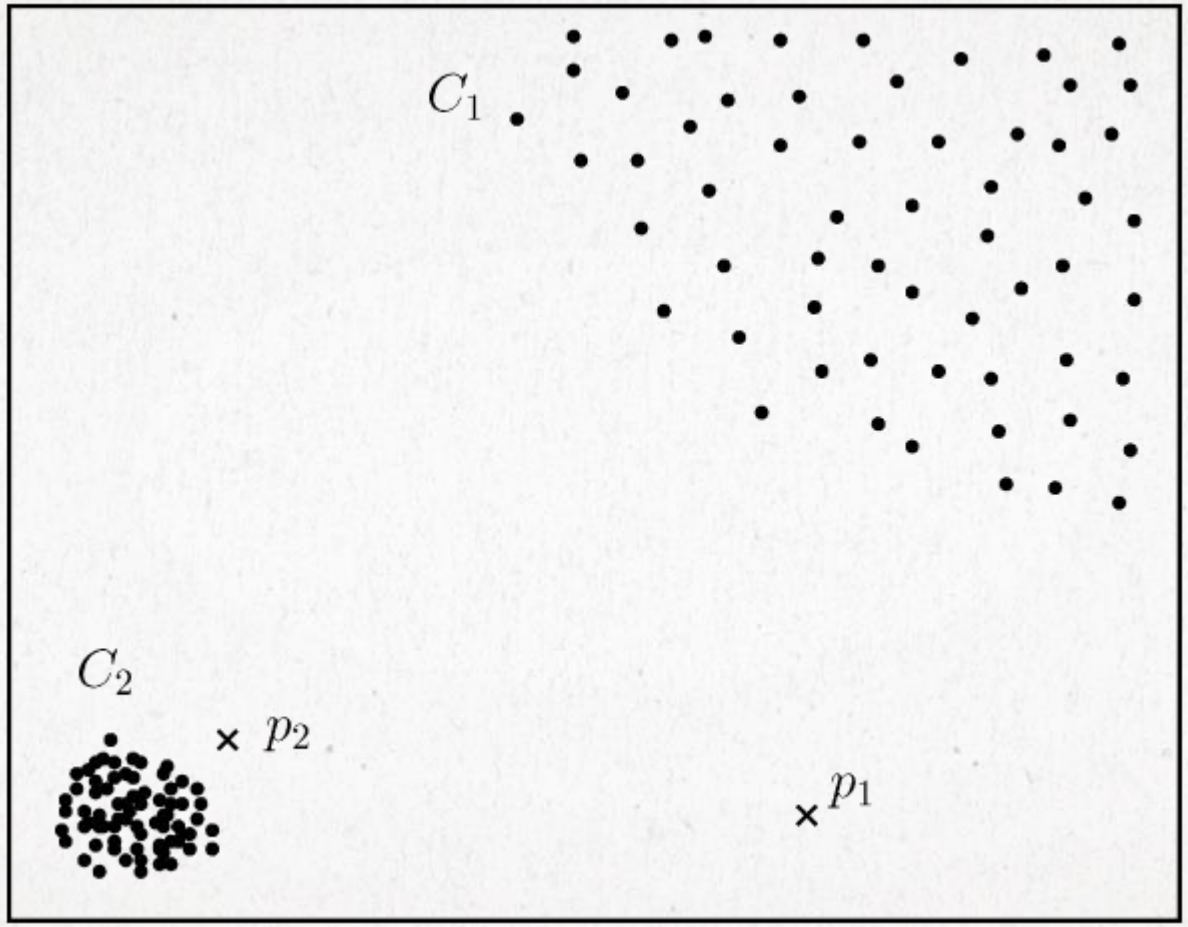
K^{th} NEAREST NEIGHBOR

- First way: the anomaly score of a data instance is defined as its distance to its K^{th} nearest neighbor in a given data set (applied to detect land mines from satellite ground images)
- Second way: the anomaly score of a data instance is defined as the number of nearest neighbor (n) that are not more than *d* distance apart from the given data instance (credit card fraud)
- Different distances between categorical and continuous attributes
- Variants to improve efficiency:
  - Pruning search space by either ignoring instances that cannot be anomalous or by focussing on instances that are most likely to be anomalous
  - Cluster-based pruning

RELATIVE DENSITY

- 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
- These techniques perform poorly if the data has regions of varying densities

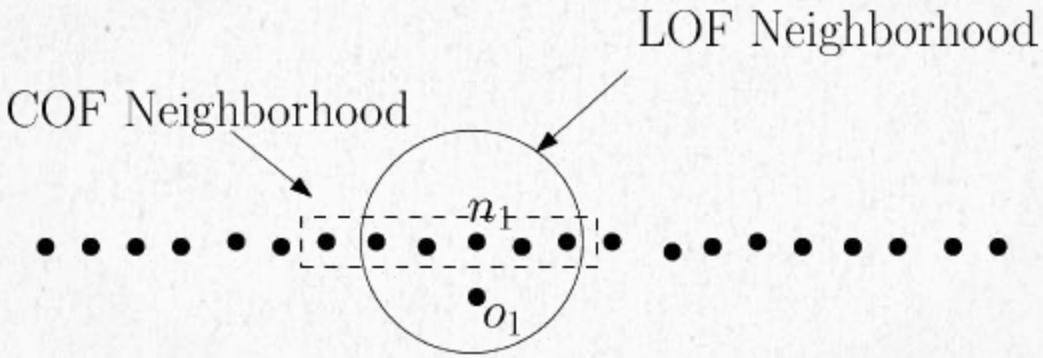
RELATIVE DENSITY



RELATIVE DENSITY

- An instance that lies in a neighborhood with low density is declared to be anomalou while an instance that lies in a dense neighborhood is declared to be normal
- These techniques perform poorly if the data has regions of varying densities
  - LOF (Local Outlier Factor): for any given data instance, the LOF score is equal to ratio of average local density (in 2D $\frac{k}{\pi d^2}$) of the k nearest neighbors of the instance and the local density of the data instance itself
  - COF(Connectivity–based Outlier Factor): the neighborhood for an instance is computed in an incremental mode

RELATIVE DENSITY



COMPUTATIONAL COMPLEXITY

- Basic techniques has $O(N^2)$ complexity
- More efficient data structure can be used
- Several techniques has directly optimized the anomaly detection under the assumption that only the top few anomalies are insteresting

ADVANTAGES AND DISADVANTAGES

- Advantages
  - Unsupervised (purely data driven)
  - Semisupervised techniques perform better than unsupervised in term of missed anomalies
  - To adapt nearest neighbor-based techniques to a different type is straightforward (only need distance)
- Disadvantages
  - For unsupervised techniques we can have missed anomalies (normal instances without enough close neighbors or anomalies with enough close neighbors)
  - Computational complexity
  - Defining distance measures between instances can be challenging

CLUSTERING-BASED ANOMALY DETECTION

- Unsupervised techniques
- Two categories:
  - Normal data instances belong to a cluster in the data, while anomalies do not belong to any cluster
    - Apply a known clustering-based algorithm to the data set and declare any data instance that does not belong to any cluster as anomalous
  - Normal data instance lie close to their closest cluster centroid, while anomalies are far away from their closest cluster centroid
    - The data is clustered using a clustering algorithm, then, for each data instance, its distance to its closest cluster centroid is calculated as its anomaly score
  - If the anomalies in the data form clusters by themselves, these techniaues will not be able to detect such anomalies.
    - Normal data instances belong to large and dense clusters, while anomalies belong to small or sparse cluster

COMPUTATIONAL COMPLEXITY

- Depends on the clustering algorithm used to generate clusters from the data
- Clustering algorithms have generally quadratic complexity
- Heuristic-based techniques have linear complexity
- Testing phase is really fast since it envolves comparing a test instance with a small number of clusters

ADVANTAGES AND DISADVANTAGES

- Advantages
  - Unsupervised (purely data driven)
  - To adapt clustering techniques to a different type is straightforward (clustering algorithm)
  - Testing phase is very fast
- Disadvantages
  - Performance depends on the effectiveness of clustering algorithm
  - Many techniques detect anomalies as a byproduct of clustering, and hence are not optimized for anomaly detection
  - Several clustering algorithms force every instance to be assigned to some cluster
  - Several clustering based techniques are effective only when the anomalies do not form significant clusters among themselves
  - Computational complexity for clustering the data

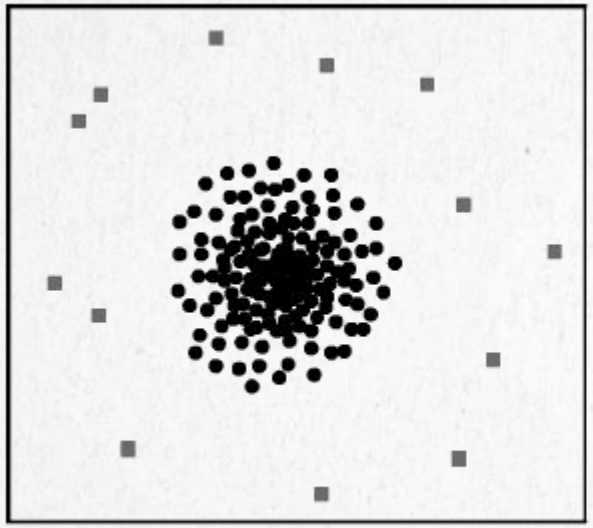
OTHER TECHNIQUES

- Statistical anomaly detection techniques
  - Parametric techniques
  - Non parametric techniques
- Information theoretic anomaly detection techniques
- Spectral anomaly detection techniques
- ……

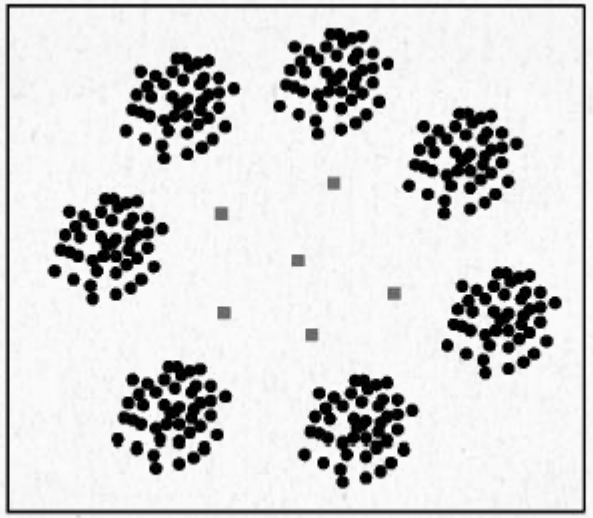
CONCLUSIONS

- Anomalies in data translate to significant and often critical information in a wide variety of application domains
- Unlimited applications
- Very challenging problem
- Each problem is different from any other

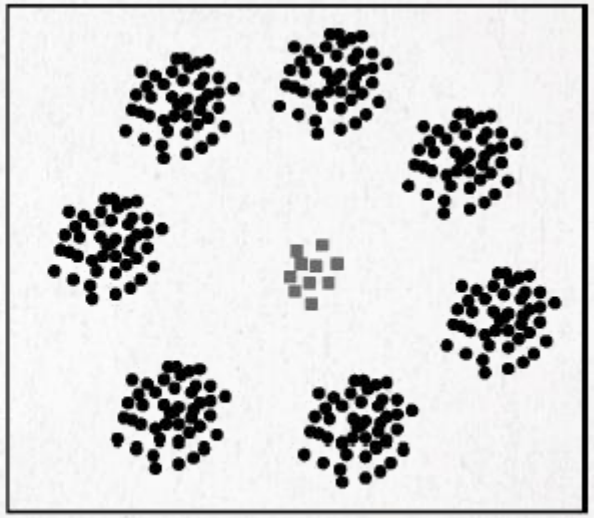
CONCLUSIONS



(a) Data Set 1



(b) Data Set 2



(c) Data Set 3

THE END

Questions?

VARUN CHANDOLA
University at Buffalo

ARINDAM BANERJEE
University of Minnesota

VIPIN KUMAR
University of Minnesota

THE END

Thank you!

VARUN CHANDOLA
University at Buffalo

ARINDAM BANERJEE
University of Minnesota

VIPIN KUMAR
University of Minnesota