IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No 3, January 2012
ISSN (Online): 1694-0814
www.IJCSI.org

Outlier Detection: Applications And Techniques

Karanjit Singh and Dr. Shuchita Upadhyaya

HQ Base Workshop Group EME
Meerut Cantt, UP, India

Department of Computer Science and Applications, Kurukshetra University
Kurukshetra, Haryana, India

Abstract

Outliers once upon a time regarded as noisy data in statistics, has turned out to be an important problem which is being researched in diverse fields of research and application domains. Many outlier detection techniques have been developed specific to certain application domains, while some techniques are more generic. Some application domains are being researched in strict confidentiality such as research on crime and terrorist activities. The techniques and results of such techniques are not readily forthcoming. A number of surveys, research and review articles and books cover outlier detection techniques in machine learning and statistical domains individually in great details. In this paper we make an attempt to bring together various outlier detection techniques, in a structured and generic description. With this exercise, we hope to attain a better understanding of the different directions of research on outlier analysis for ourselves as well as for beginners in this research field who could then pick up the links to different areas of applications in details.
Keywords: *Outlier Applications, Outliers, Outlier Detection.*

1. Introduction

Outlier detection aims to find patterns in data that do not conform to expected behavior. It has extensive use in a wide variety of applications such as military surveillance for enemy activities, intrusion detection in cyber security, fraud detection for credit cards, insurance or health care and fault detection in safety critical systems. Their importance in data is due to the fact that they can translate into actionable information in a wide variety of applications. An anomalous traffic pattern in a computer network could mean that a hacked computer is sending out sensitive data to an unauthorized destination [1]. An abnormal MRI image may indicate presence of malignant tumors [2]. Outliers in credit card transaction data could indicate credit card or identity theft [3] or abnormal readings from a space craft sensor could signify a fault in some component of the space craft [4]. In statistical data study of outliers dates as early as the 19th century [5]. Since then several research communities have developed a variety of outlier detection techniques with many of these specifically meant for certain applications and others being generic in nature. With this exercise, we hope to get a better understanding of the different directions of research on outlier analysis and think of applying techniques in different areas to our areas of interest of crime detection and counter terrorism, even if they were they were not intended, to begin with.

2. Defining Outliers

Outliers are patterns in data that do not conform to a well defined notion of normal behavior. Figure 1 illustrates outliers in a simple 2-dimensional data set. The data has two normal regions, N1 and N2, since most observations lie in these two regions. Points that are sufficiently far away from the regions, e.g., points o1 and o2, and points in region O3, are outliers. x y N1 N2 o1 o2 O3

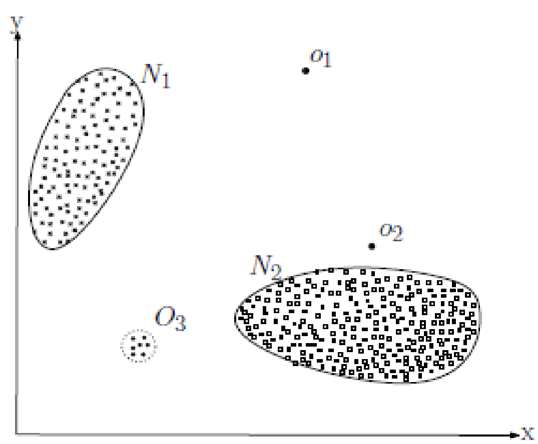

Fig. 1. A simple example of outliers in a 2-dimensional data set.

outliers might be induced in the data for a variety of reasons, such as malicious activity, e.g., credit card fraud, cyber-intrusion, terrorist activity or breakdown of a system, but the common point of all is that they are interesting to the analyst. The "interestingness" or real life relevance of outliers is a key feature of outlier detection.

Outlier detection is related to, but distinct from noise removal [6] and noise accommodation [7], both of which deal with unwanted noise in the data. Noise can be defined as a phenomenon in data which is not of interest to the analyst, but acts as a hindrance to data analysis. Noise removal is driven by the need to remove the unwanted objects before any data analysis is performed on the data. Noise accommodation refers to immunizing a statistical model estimation against anomalous observations [8].

Another topic related to outlier detection is novelty detection [9,10,11] which aims at detecting previously unobserved (emergent, novel) patterns in the data, e.g., a new topic of discussion in a news group. The distinction between novel patterns and outliers is that the novel patterns are typically incorporated into the normal model after being detected. Another topic related to outlier detection is novelty detection [9,10,11]. The distinction between novel patterns and outliers is that the novel patterns are previously unobserved and get typically incorporated into the normal model after being detected e.g., a new topic of discussion in a news group.

We have discussed above mentioned related problems because their solutions are often used for outlier detection and vice-versa.

3. Difficulties in Outlier Detection

Abstractly speaking outliers are patterns that deviate from expected normal behavior, which in its simplest form could be represented by a region and visualize all normal observations to belong to this normal region and consider the rest as outliers This approach looks simple but is highly challenging due to following reasons.

It is very difficult to define the normal behavior or a normal region. The difficulties are as under.

- Encompassing of every possible normal behavior in the region.
- Imprecise boundary between normal and outlier behavior since at times outlier observation lying close to the boundary could actually be normal, and vice-versa.
- Adaptation of malicious adversaries to make the outlier observations appear like normal when outliers result from malicious actions.
- In many domains normal behavior keeps evolving and may not be current to be a representative in the future.
- Differing notion of outliers in different application domains makes it difficult to apply technique developed in one domain to another. For example, in the medical domain a small deviation from normal body temperature might be an outlier, while similar value deviation in the stock market domain might be considered as normal. Even within same domain say crime detection there could be situations where use of foreign make weapons may be considered normal in crimes committed in metro cities but an outlier for murders of commoners in tribal regions.
- Availability of labeled data for training/ validation of models used by outlier detection techniques.
- Noise in the data which tends to be similar to the actual outliers and hence difficult to distinguish and remove.

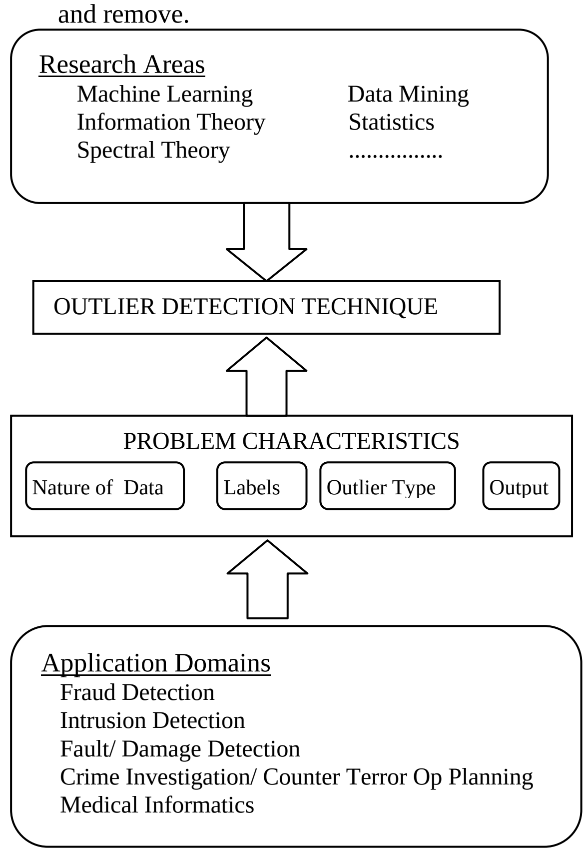

Fig. 2. Key components associated with outlier detection technique.

Due to the above challenges, the outlier detection problem, in its most general form, is not easy to solve. In fact, most of the existing outlier detection techniques solve a specific problem formulation which is induced by various factors such as nature of the data, availability of labeled data, type of outliers to be detected, etc. Often, these factors are determined by the application domain in which the outliers need to be detected.

Researchers adopt concepts from diverse disciplines such as statistics, machine learning, data mining, information theory, spectral theory, and apply them to specific problem formulations. Figure 2 shows the above mentioned key components associated with any outlier detection technique.

4. Previous Work

A number of surveys, review articles and books especially Hodge and Austin [12], cover outlier detection techniques in machine learning and statistical domains. Numeric and symbolic data approaches [13], neural networks and statistical approaches [9, 10, 11] have been presented by various researchers. Cyber-intrusion detection surveys [14, 15] and research and review books on outlier detection techniques [16,17,18] are excellent sources of literature on the subject.

5. Our Contribution

The above literature on outlier detection set focus on individual applications or on a particular research area. We have attempted to structure and present a broad overview of the detailed research on outlier detection techniques in multifarious research areas and applications also trying to highlight the richness and complexity associated with each application domain. We distinguish simple outliers from complex outliers and define two types of complex outliers, viz., contextual and collective outliers.

6. Aspects Determining the Formulation of Problem

As mentioned earlier, a specific formulation of the problem is determined by several different factors some of which are discussed below. These are also depicted in Figure 2 above. Broadly speaking they are:-

- Nature of Input Data
- Type of Outlier – Point, Contextual, Collective
- Data Labels
- Output of Outlier Detection.

7. Nature of Input Data

This is a key aspect of any outlier detection technique. Input is generally a collection of data instances (also referred as object, record, point, vector, pattern, event, case, sample, observation, entity) [20] . Each data instance can be described using a set of attributes (also referred to as variable, characteristic, feature, field, dimension). The attributes can be of different types such as binary, categorical or continuous. Each data instance might consist of only one attribute (univariate) or multiple attributes (multivariate). In the case of multivariate data instances, all attributes might be of same type or might be a mixture of different data types. The nature of attributes determines the applicability of outlier detection techniques. For example, for statistical techniques different statistical models have to be used for continuous and categorical data. Similarly, for nearest neighbor based techniques, the nature of attributes would determine the distance measure to be used. Often, instead of the actual data, the pair-wise distance between instances might be provided in the form of a distance (or similarity) matrix. In such cases, techniques that require original data instances are not applicable, e.g., many statistical and classification based techniques. In case these statistical methods are applied to OLAP cubes for data mining then the distance between dimensional data can be found out by applying score functions.
Input data can also be categorized based on the relationship present among data instances [20]. Most of the existing outlier detection techniques deal with record data (or point data), in which no relationship is assumed among the data instances. In case these statistical methods are applied to OLAP cubes for data mining then the distance between dimensional data can be found out by applying some sort of score functions and then determining the outliers.
In general, data instances can be related to each other. Some examples are sequence data, spatial data, and graph data. In sequence data, the data instances are linearly ordered, e.g., time-series data, genome sequences, protein sequences. In spatial data, each data instance is related to its neighboring instances, e.g., vehicular traffic data, ecological data. When the spatial data has a temporal (sequential) component it is referred to as spatio-temporal data, e.g., climate data. In graph data, data instances are represented as vertices in a graph and are connected to other vertices with edges. Later we will discuss situations where such relationship among data instances becomes relevant for outlier detection.

8. Types of Outliers

An important aspect of an outlier detection technique is the nature of the desired outlier. Outliers can be classified into following three categories:

- Point Outliers
- Contextual Outliers
- Collective Outliers.

9. Point Outliers

If an individual data instance can be considered as anomalous with respect to the rest of data, then the instance is termed as a point outlier. This is the simplest type of outlier and is the focus of majority of research on outlier detection. For example, in Figure 1, points o1 and o2 as well as points in region O3 lie outside the boundary of the normal regions, and hence are point outliers since they are different from normal data points. As a real life example, if we consider credit card fraud detection with data set corresponding to an individual's credit card transactions assuming data definition by only one feature: amount spent. A transaction for which the amount spent is very high compared to the normal range of expenditure for that person will be a point outlier.

10. Contextual Outliers

If a data instance is anomalous in a specific con-text (but not otherwise), then it is termed as a contextual outlier (also referred to as conditional outlier [21]). The notion of a context is induced by the structure in the data set and has to be specified as a part of the problem formulation. Each data instance is defined using two sets of attributes:

- **Contextual attributes.** The contextual attributes are used to determine the context (or neighborhood) for that instance. For example, in spatial data sets, the longitude and latitude of a location are the contextual attributes. In time-series data, time is a contextual attribute which determines the position of an instance on the entire sequence.

- **Behavioral attributes.** The behavioral attributes define the non-contextual characteristics of an instance. For example, in a spatial data set describing the average rainfall of the entire world, the amount of rainfall at any location is a behavioral attribute.

The anomalous behavior is determined using the values for the behavioral attributes within a specific context. A data instance might be a contextual outlier in a given context, but an identical data instance (in terms of behavioral attributes) could be considered normal in a different context. This property is key in identifying contextual and behavioral attributes for a contextual outlier detection technique.

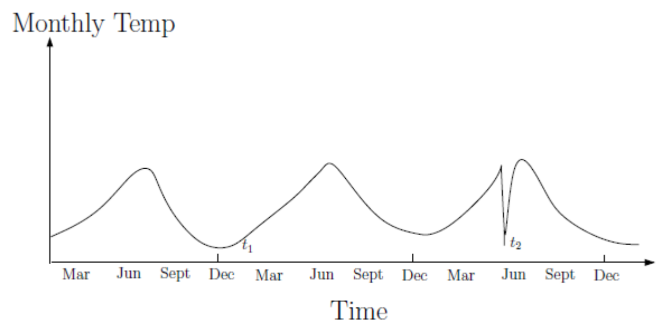

Fig. 3. Contextual outlier t2 in a temperature time series. Temperature at time t1 is same as that at time t2 but occurs in a different context and hence is not considered as an outlier.

Contextual outliers have been most commonly explored in time-series data [22] and spatial data [23]. Figure 3 shows one such example for a temperature time series which shows the monthly temperature of an area over last few years. A temperature of 35F might be normal during the winter (at time t_1) at that place, but the same value during summer (at time t_2) would be an outlier. A six ft tall adult may be a normal person but if viewed in *context of age* a *six feet* tall *kid* will definitely be an outlier.

A similar example can be found in the credit card fraud detection with contextual as *time* of purchase. Suppose an individual usually has a weekly shopping bill of $100 except during the Christmas week, when it reaches $1000. A new purchase of $1000 in a week in July will be considered a contextual outlier, since it does not conform to the normal behavior of the individual in the context of time (even though the same amount spent during Christmas week will be considered normal).

The choice of applying a contextual outlier detection technique is determined by the meaningfulness of the contextual outliers in the target application domain.
Applying a contextual outlier detection technique makes sense if contextual attributes are readily available and therefore defining a context is straightforward. But it becomes difficult to apply such techniques if defining a context is not easy.

11. Collective Outliers

If a collection of related data instances is anomalous with respect to the entire data set, it is termed as a collective outlier. The individual data instances in a collective outlier may not be outliers by themselves, but their occurrence together as a collection is anomalous. Figure 4 illustrates an example which shows a human electrocardiogram output [24]. The highlighted region denotes an outlier because the same low value exists for an abnormally long time (corresponding to an Atrial Premature Contraction). It may be noted that low value by itself is not an outlier but its successive occurrence for long time is an outlier.

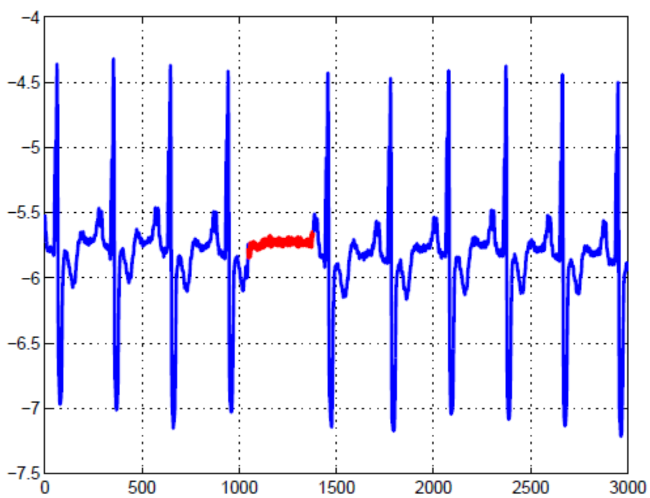

Fig. 4. Collective outlier in an human ECG output corresponding to an Atrial Premature Contraction.

As an another illustrative example, consider a sequence of actions occurring in a computer as shown below:
……..http-web, buffer-overflow, http-web, http-web, smtp-mail, ftp, http-web, ssh, smtp-mail, http-web, ssh, buffer-overflow, ftp, http-web, ftp, smtp-mail, http-web……
The highlighted sequence of events (buffer-overflow, ssh, ftp) correspond to a typical web based attack by a remote machine followed by copying of data from the host computer to remote destination via ftp. It should be noted that this collection of events is an outlier but the individual events are not outliers when they occur in other locations in the sequence.

Collective outliers have been explored for sequence data [25,26], graph data [27], and spatial data [28]. It should be noted that while point outliers can occur in any data set, collective outliers can occur only in data sets in which data instances are related. In contrast, occurrence of contextual outliers depends on the availability of context attributes in the data. A point outlier or a collective outlier can also be a contextual outlier if analyzed with respect to a context. Thus a point outlier detection problem or collective outlier detection problem can be transformed to a contextual outlier detection problem by incorporating the context information

12. Data Labels

The labels associated with a data instance denote if that instance is normal or anomalous. It should be noted that obtaining labeled data which is accurate as well as representative of all types of behaviors, is often prohibitively expensive. Labeling is often done manually by a human expert and hence requires substantial effort to obtain the labeled training data set. Typically, getting a labeled set of anomalous data instances which cover all possible type of anomalous behavior is more difficult than getting labels for normal behavior. Moreover, the outlier behavior is often dynamic in nature, e.g., new types of outliers might arise, for which there is no labeled training data. In certain cases, such as air traffic safety, outlier instances would translate to catastrophic events, and hence will be very rare. Based on the extent to which the labels are available, outlier detection techniques can operate in one of the following three modes:

- **Supervised outlier detection:** Techniques trained in supervised mode assume the availability of a training data set which has labeled instances for normal as well as outlier class. Typical approach in such cases is to build a predictive model for normal vs. outlier classes. Any unseen data instance is compared against the model to determine which class it belongs to. There are two major issues that arise in supervised outlier detection. First, the anomalous instances are few, as compared to the normal instances in the training data. Second, obtaining accurate and representative labels, especially for the outlier class is usually challenging. A number of techniques have been proposed [29, 30, 31] that inject artificial outliers in a normal data set to obtain a labeled training data set. Other than these two issues, the supervised outlier detection problem is similar to building predictive models. Hence we will not address this category of techniques in this survey.

- **Semi-Supervised outlier detection:** Techniques that operate in a semi-supervised mode, assume that the training data has labeled instances for only the normal class. Since they do not require labels for the outlier class, they are more widely applicable than supervised techniques. For example, in space craft fault detection [32], an outlier scenario would signify an accident, which is not easy to model. The typical approach used in such techniques is to build a model for the class corresponding to normal behavior, and use the model to identify outliers in the test data. A limited set of outlier detection techniques exist that assume availability of only the outlier instances for training [25, 33, 34]. Such techniques are not commonly used, primarily because it is difficult to obtain a training data set which covers every possible anomalous behavior that can occur in the data.

- **Unsupervised outlier detection:** Techniques that operate in unsupervised mode do not require training data, and thus are most widely applicable. The techniques in this category make the implicit assumption that normal instances are far more frequent than outliers in the test data. If this assumption is not true then such techniques suffer from high false alarm rate.

Many semi-supervised techniques can be adapted to operate in an unsupervised mode by using a sample of the unlabeled data set as training data. Such adaptation assumes that the test data contains very few outliers and the model learnt during training is robust to these few outliers.

13. Output of Outlier Detection

An important aspect for any outlier detection technique is the manner in which the outliers are reported. Typically, the outputs produced by outlier detection techniques are one of the following two types:

- **Scores:** Scoring techniques assign an outlier score to each instance in the test data depending on the degree to which that instance is considered an outlier. Thus the output of such techniques is a ranked list of outliers. An analyst may choose to either analyze top few outliers or use a cut-off threshold to select the outliers.

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

Scoring based outlier detection techniques allow the analyst to use a domain-specific threshold to select the most relevant outliers. Techniques that provide binary labels to the test instances do not directly allow the analysts to make such a choice, though this can be controlled indirectly through parameter choices within each technique.

14. Applications of Outlier Detection

We shall highlight several applications of outlier detection. For each application we shall discuss following aspects:

- The notion of outlier.
- Nature of the data.
- Challenges associated with detecting outliers.
- Existing outlier detection techniques.

15. Intrusion Detection

Intrusion detection refers to detection of malicious activity (break-ins, penetrations, and other forms of computer abuse) in a computer related system [35] interesting from a computer security perspective. Being different from normal system behavior, intrusion detection is a perfect candidate for applying outlier detection techniques. The key challenges for outlier detection are :-

- **Huge Data Volume:** This calls for computationally efficient techniques.
- **Streaming Data:** This requires on-line analysis.
- **False alarm rate:** Smallest percentage of false alarms among millions of data objects can make be overwhelming for an analyst.
- **Labeled data not usually available for Intrusions:** This gives preference to semi-supervised and unsupervised outlier detection techniques.

Intrusion detection systems have been classified into host based and network based intrusion detection systems [36].

The major differences being tabled as under:-

Table 1: Differences in Nature of Host Based and Network Intrusion Systems

Aspect	*Host Based*	*Network Based*
Outliers In	OS Calls	Network Data.
Translates to	Malicious Code Unusual Behaviour Policy Violations	Denial of Network Services
Nature of Data analysis	Sequential	Point, Sequential, Collective
Granularity/ Profiling	User / Program	Packet Level/ NetFlows

The examples of outlier detection techniques for Intrusion Detection are tabled below:-

Table 2: Some outlier detection techniques used in Host Based and Network Intrusion Systems

Technique Used	*References*
Host Based Intrusion Detection Systems	
Statistical Profiling Using Histograms	[37- 45]
Mixture of Models	[46]
Neural Networks	[47]
Support Vector Machines	[2]
Rule Based Systems	[48 - 50]
Network Based Intrusion Detection Systems	
Statistical Profiling using Histograms	[51 - 54]
Parametric Statistical Modeling	[55]
Non-parametric Statistical Modeling	[56]
Bayesian Networks	[57 - 60]
Support Vector Machines	[46]
Rule Based Systems	[61]
Neural Networks	[46, 62 - 68]

16. Fraud Detection

Fraud refers to criminal activities occurring in commercial organizations such as banks, credit card companies, insurance agencies, cell phone companies, stock market, etc. Malicious users could be actual customers of the organization or resorting to identity theft (posing as customers). The detection activity aims at detection of unauthorized consumption of resources provided by the organization to prevent economic losses.

A general approach to outlier detection here would involve maintaining a usage profile for each customer and monitor the profiles to detect any deviations termed as activity monitoring [73]. Some specific applications of fraud detection are discussed below.

Credit Card Fraud Detection: Outlier detection techniques are applied to detect :-

- **Fraudulent Applications for Credit Card:** This is similar to detecting insurance fraud [69]
- **Fraudulent Usage of Credit Card:** Associated with credit card thefts.

The data records are defined over several dimensions such as the user ID, spent amount, time between consecutive card usage, etc. The frauds are typically reflected in transactional records (point outliers) and correspond to high payments, high rate of purchase, purchase of items never purchased by the user before, etc. Availability of labeled records is no problem since credit companies have complete data available. Moreover, the data falls into distinct profiles based on the credit card user. Hence profiling and clustering based techniques are typically used in this domain.

Online detection of fraud as soon as fraudulent transaction occurs is a challenge in detecting unauthorized credit card usage. This problem is addressed in two different ways.

Table 3: Approaches in detecting fraudulent transactions.

Approach	*By-Owner*	*By-Operation*
Context	User	Geographic Location
Cost	Expensive; querying a central data repository with every transaction.	

Some outlier detection techniques used in fraud detection are listed in Table IV.

Table IV: Some outlier detection techniques used in fraud detection

Technique Used	***References***
Neural Networks	[3, 69 - 71]
Rule-based Systems	[70]
Clustering	[72]

17. Mobile Phone Fraud Detection.

In this activity monitoring problem the calling behavior of each account is scanned to issue an alarm when an account appears to have been misused.

Calling activity is usually represented with call records. Each call record is a vector of continuous (e.g., Call-Duration) and discrete (e.g., Calling-City) features. However, there is no inherent primitive representation in this domain. Calls are aggregated by time, for example into call-hours or call-days or user or area depending on the granularity desired. The outliers correspond to high volume of calls or calls made to unlikely destinations.

Some techniques applied to cell phone fraud detection are listed in Table V.

Table V: Examples of different outlier detection techniques used for cell phone fraud detection.

Technique Used	***References***
Statistical Profiling using Histograms	[73, 74]
Parametric Statistical Modelling	[75, 76]
Neural Networks	[77, 78]
Rule based Systems	[78,79]

18. Insurance Claim Fraud Detection

An important problem in the property-casualty insurance industry is claims fraud, e.g. automobile insurance fraud. Individuals and conspiratorial rings of claimants and providers manipulate the claim processing system for unauthorized and illegal claims.

The data in this domain for fraud detection comes from the documents submitted by the claimants. The techniques extract different features (both categorical as well as continuous) from these documents. Typically, claim adjusters and investigators assess these claims for frauds. These manually investigated cases are used as labeled instances by supervised and semi-supervised techniques for insurance fraud detection.

Insurance claim fraud detection is quite often handled as a generic activity monitoring problem [73]. Neural network based techniques have also been applied to identify anomalous insurance claims [80, 81].

19. Insider Trading Detection

Insider trading is a phenomenon found in stock markets, where people make illegal profits by acting on (or leaking) inside information before the information is made public.

The inside information can be of different forms [82] generally referring to any information which would affect the stock prices in a particular industry. It could be knowledge about a pending merger/acquisition, a terrorist attack affecting a particular industry, a pending legislation affecting a particular industry.

Fraud has to be detected in an online manner and as early as possible, to prevent people/organizations from making illegal profits. The available data comes from heterogeneous sources such as option trading data, stock trading data, news. The data has temporal associations since the data is collected continuously. The temporal and streaming nature has also been exploited in certain techniques [75].

Some outlier detection techniques used in this domain are listed in Table VI.

Table VI: Examples of different outlier detection techniques used for insider trading detection.

Technique Used	***References***
Statistical profiling using Histograms	[75, 82]
Information Theoretic	[83]

20. Medical and Public Health Outlier Detection

The data typically consists of patient records which may have several different types of features such as patient age, blood group, weight. The data might also have temporal as well as spatial aspect to it. The data can have outliers due to several reasons such as abnormal patient condition or instrumentation errors or recording errors. Most of the current outlier detection techniques in this domain aim at detecting anomalous records (point outliers). Typically the labeled data belongs to the healthy patients, hence most of the techniques adopt semi-supervised approach. Another form of data handled by outlier detection techniques in this domain is time series data, such as Electrocardiograms (ECG) and Electroencephalograms (EEG). Collective outlier detection techniques have been applied to detect outliers in such data [91]. Several techniques have also focussed on detecting disease outbreaks in a specific area [90]. Thus the outlier detection is a very critical problem in this domain and requires high degree of accuracy.

The most challenging aspect of the outlier detection problem in this domain is that the cost of classifying an outlier as normal can be very high.

Some outlier detection techniques used in this domain are listed in Table VII.

Table VII: Examples of different outlier detection techniques used in medical and public health domain.

Technique Used	***References***
Parametric Statistical Modelling	[84 - 88]
Neural Networks	[89]
Bayesian Networks	[90]
Rule-based Systems	[75]
Nearest Neighbor based techniques	[91]

21. Industrial Damage Detection

Industrial units suffer damage due to continuous usage and the normal wear and tear. Such damages need to be detected early to prevent further escalation and losses. The data in this domain is usually sensor data recorded using different sensors and collected for analysis.

Outlier detection in this domain is classified into two fields as tabulated below.

Table VIII: Characteristics of Fault Detection in Mechanical Units and Structural Damage Domain.

Aspect	***System Health Management***	***By-Operation***
Defects Dealt pertaining to	Mech components such as motors, engines, turbines, oil flow in pipelines etc.	Structures,
Cause of Defects	Wear and Tear or other unforeseen circumstances.	Cracks in beams, strains in airframes . Unforeseen data.
Data Aspect	Temporal	Temporal
Analysis	Time Series	Time series with special corelations
Types of Outliers	Contextual or Collective outliers	Novelty detection or change point detections
Normal data	Readily Available	Is learnt and typically static over time.
Supervision	Semi-supervised	Semi-supervised
Literature	[94, 95]	[108, 111, 112, 115]
Techniques	Table IX.	Table X.

Table IX: Examples of outlier detection techniques used for fault detection in mechanical units.

Technique Used	***References***
Parametric Statistical Modelling	[92, 93, 94, 95]
Non-Parametric Statistical Modelling	[96]
Neural Networks	[97, 89, 98-105]
Spectral	[4, 106]
Rule Based Systems	[107]

Table X: Examples of outlier detection techniques used for structural damage detection.

Technique Used	***References***
Statistical profiling using Histograms	[108, 109, 110]
Parametric Statistical Modelling	[111]
Mixture of Models	[112, 113, 114]
Neural Networks	[115 to 122]

22. Image Processing

Outlier detection here aims to detect changes in an image over time (motion detection) or in regions which appear abnormal on the static image. This domain includes satellite imagery, digit recognition, spectroscopy, mammographic image, and video surveillance. The outliers are caused by motion or insertion of foreign object or instrumentation errors. The data has spatial as well as temporal characteristics. Each data point has a few continuous attributes such as color, lightness, texture, etc. The interesting outliers are either anomalous points or regions in the images (point and contextual outliers).

One of the key challenges in this domain is the large size of the input. The challenge is greater when dealing with video data and, online detection techniques are required.

Some references on various applications are tabulated below:-

Table XI: Examples of outlier detection techniques used in image processing domain.

Application Domain	*References*
Satellite Imagery	[123, 124, 125, 126, 127]
Digit Recognition	[128]
Mammographic Image Analysis	[129, 130]
Spectroscopy	[131, 132, 133, 134]
Video Surveillance	[135, 136, 137].

Some outlier detection techniques used in this domain are listed in Table XII.

Table XII: Examples of outlier detection techniques used in image processing domain.

Technique Used	*References*
Mixture of Models	[124, 129, 130]
Regression	[126, 131]
Bayesian Networks	[135]
Support Vector Machines	[132, 138]
Neural Networks	[123, 125, 128, 133, 136]
Clustering	[134]
Nearest Neighbour Techniques	[124, 137]

23. Outlier Detection in Text Data

Outlier detection techniques in this domain primarily detect novel topics or events or news stories in a collection of documents or news articles. The outliers are caused due to a new interesting event or an anomalous topic. The data in this domain is typically high dimensional and very sparse. The data also has a temporal aspect since the documents are collected over time.

A challenge for outlier detection techniques in this domain is to handle the large variations in documents belonging to one category or topic. Some outlier detection techniques used in this domain are listed in Table XIII.

Table XIII: Examples of techniques used for outlier topic detection in text data.

Technique Used	*References*
Statistical Profiling using Histograms	[73]
Mixture of Models	[139]
Neural Networks	[140]
Support Vector Machines	[141]
Clustering Based	[142, 143, 144]

24. Sensor Networks

Sensor networks have lately become an important topic of research from data analysis perspective, since the data collected from various wireless sensors has several unique characteristics. Outliers in such data collected can either imply one or more faulty sensors (sensor fault detection applications), or the sensors are detecting events (intrusion detection applications) that are interesting for analysts.

A single sensor network might comprise a mix of sensors that collecting different types of data, such as binary, discrete, continuous, audio, video, etc. The data is generated in a streaming mode and the collected data often contains noise and missing values due to limitations imposed by deployment environment and communication channel.

This poses a set of unique challenges. The streaming data calls for outlier detection techniques to operate in an online approach. The severe resource constraints call for light-weight detection techniques. The data collected in a distributed fashion calls for a distributed data mining approach to analyze the data [145]. Lastly the presence of noise in sensor data makes outlier detection more challenging, since it has to now distinguish between interesting outliers and the unwanted values (noise/missing values).

Table XIV lists some outlier detection techniques used in this domain.

Table XIV: Some outlier detection techniques used for outlier detection in sensor networks.

Technique Used	***References***
Bayesian Networks	[146]
Rule-based Systems	[147]
Parametric Statistical Modelling	[148, 149]
Nearest Neighbor Based Techniques	[150, 151, 152]
Spectral Techniques	[145]

25. Other Domains

Some other domains where outlier detection has also been applied are as tabulated below.

Table XV: Examples of outlier detection techniques used in other application domains.

Technique Used	***References***
Speech Recognition	[153, 154]
Novelty Detection in Robot Behavior	[155, 156, 157, 158, 159]
Traffic Monitoring	[160]
Click Through Protection	[161]
Detecting Faults in Web Applications	[162, 163]
Detecting Outliers in Biological Data	[55, 164, 165, 166, 167, 176]
Detecting Outliers in Census Data	[168]
Detecting Associations among Criminal Activities	[169]
Detecting Outliers in Customer Relationship Management (CRM) Data	[170]
Detecting Outliers in Astronomical Data	[171, 172, 173]
Detecting Ecosystem Disturbances	[23,174, 175]

26. Conclusion

In this paper we have brought together various outlier detection techniques, in a structured and generic description. With this exercise, we have attained a better understanding of the different directions of research on outlier analysis for ourselves as well as for beginners in this research field who can pick up the links to different areas of applications in details.

Acknowledgments

Karanjit Singh and Dr Shuchita Upadhyaya thanks the staff of CS and IT department of Kurukshetra University for their wholehearted support in referencing the study material. The authors sincerely thank the computer centre and library staff for their unstinted support as well.

References

[1] Kumar, V. 2005. Parallel and Distributed Computing for Cybersecurity. Distributed Systems Online, IEEE 6, 10.
[2] Spence, C., Parra, L., and Sajda, P. 2001. Detection, synthesis and compression in mammographic image analysis with a hierarchical image probability model. In Proceedings of the IEEE Workshop on Mathematical Methods in Biomedical Image Analysis. IEEE Computer Society, Washington, DC, USA, 3.
[3] Aleskerov, E., Freisleben, B., and Rao, B. 1997. Cardwatch: A neural network based database mining system for credit card fraud detection. In Proceedings of IEEE Computational Intelligence for Financial Engineering. 220-226.
[4] Fujimaki, R., Yairi, T., and Machida, K. 2005. An approach to spacecraft outlier detection problem using kernel feature space. In Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining. ACM Press, New York, NY, USA, 401-410
[5] Edgeworth, F. Y. 1887. On discordant observations. Philosophical Magazine 23, 5, 364 -375.
[6] Teng, H., Chen, K., and Lu, S. 1990. Adaptive real-time outlier detection using inductively generated sequential patterns. In Proceedings of IEEE Computer Society Symposium on Re-search in Security and Privacy. IEEE Computer Society Press, 278-284.

[7] Rousseeuw, P. J. and Leroy, A. M. 1987. Robust regression and outlier detection. John Wiley & Sons, Inc., New York, NY, USA.
[8] Huber, P. 1974. Robust Statistics. Wiley, New York.
[9] Markou, M. and Singh, S. 2003a. Novelty detection: a review-part 1: statistical approaches. Signal Processing 83, 12, 2481-2497.
[10] Markou, M. and Singh, S. 2003b. Novelty detection: a review-part 2: neural network based approaches. Signal Processing 83, 12, 2499-2521.
[11] Saunders, R. and Gero, J. 2000. The importance of being emergent. In Proceedings of Artificial Intelligence in Design.
[12] Hodge, V. and Austin, J. 2004. A survey of outlier detection methodologies. Artificial Intelligence Review 22, 2, 85-126
[13] Agyemang, M., Barker, K., and Alhajj, R. 2006. A comprehensive survey of numeric and symbolic outlier mining techniques. Intelligent Data Analysis 10, 6, 521-538.
[14] Patcha, A. and Park, J.-M. 2007. An overview of outlier detection techniques: Existing solutions and latest technological trends. Comput. Networks 51, 12, 3448-3470
[15] Snyder, D. 2001. Online intrusion detection using sequences of system calls. M.S. thesis, Department of Computer Science, Florida State University.
[16] Rousseeuw, P. J. and Leroy, A. M. 1987. Robust regression and outlier detection. John Wiley & Sons, Inc., New York, NY, USA.
[17] Barnett, V. and Lewis, T. 1994. Outliers in statistical data. John Wiley and sons.
[18] Bakar, Z., Mohemad, R., Ahmad, A., and Deris, M. 2006. A comparative study for outlier detection techniques in data mining. Cybernetics and Intelligent Systems, 2006 IEEE Conference, 1- 6
[19] Varun Chandola, Arindam Banerjee, Vipin Kumar 2009, Outlier Detection, University of Miniesota
[20] Tan, P.-N., Steinbach, M., and Kumar, V. 2005. Introduction to Data Mining. Addison-Wesley.
[21] Song, X., Wu, M., Jermaine, C., and Ranka, S. 2007. Conditional outlier detection. IEEE Transactions on Knowledge and Data Engineering 19, 5, 631-645.
[22] Weigend, A. S., Mangeas, M., and Srivastava, A. N. 1995. Nonlinear gated experts for time-series - discovering regimes and avoiding overfitting. International Journal of Neural Systems 6, 4, 373-399.
[23] Kou, Y., Lu, C.-T., and Chen, D. 2006. Spatial weighted outlier detection. In Proceedings of SIAM Conference on Data Mining.
[24] Goldberger, A. L., Amaral, L. A. N., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C.-K., and Stanley, H. E. 2000. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for com-plex physiologic signals. Circulation 101, 23, e215 - e220. Circulation Electronic Pages: http://circ.ahajournals.org /cgi/content/full/101/23/e215
[25] Forrest, S., Warrender, C., and Pearlmutter, B. 1999. Detecting intrusions using system calls: Alternate data models. In Proceedings of the 1999 IEEE ISRSP. IEEE Computer Society, Washington, DC, USA, 133 - 145.
[26] Sun, P., Chawla, S., and Arunasalam, B. 2006. Mining for outliers in sequential databases. In SIAM International Conference on Data Mining.
[27] Noble, C. C. and Cook, D. J. 2003. Graph-based outlier detection. In Proceedings of the 9th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM Press, 631 - 636.
[28] Sekar, R., Bendre, M., Dhurjati, D., and Bollineni, P. 2001. A fast automaton-based method for detecting anomalous program behaviors. In Proceedings of the IEEE Symposium on Security and Privacy. IEEE Computer Society, 144.
[29] Theiler, J. and Cai, D. M. 2003. Resampling approach for outlier detection in multispectral images. In Proceedings of SPIE 5093, 230-240, Ed.
[30] Abe, N., Zadrozny, B., and Langford, J. 2006. Outlier detection by active learning. In Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM Press, New York, NY, USA, 504 - 509.
[31] Steinwart, I., Hush, D., and Scovel, C. 2005. A classification framework for outlier detection. Journal of Machine Learning Research 6, 211 – 232
[32] Fujimaki, R., Yairi, T., and Machida, K. 2005. An approach to spacecraft outlier detection problem using kernel feature space. In Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining. ACM Press, New York, NY, USA, 401 – 410
[33] Dasgupta, D. and Nino, F. 2000. A comparison of negative and positive selection algorithms in novel pattern detection. In Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics. Vol. 1. Nashville, TN, 125 - 130.
[34] Dasgupta, D. and Majumdar, N. 2002. Outlier detection in multidimensional data using negative selection algorithm. In Proceedings of the IEEE Conference on Evolutionary Computation. Hawaii, 1039 - 1044.
[35] Phoha, V. V. 2002. The Springer Internet Security Dictionary. Springer-Verlag.
[36] Denning, D. E. 1987. An intrusion detection model. IEEE Transactions of Software Engineering 13, 2, 222 - 232.
[37] Forrest, S., D'haeseleer, P., and Helman, P. 1996. An immunological approach to change detection: Algorithms, analysis and implications. In Proceedings of the 1996 IEEE Symposium on Security and Privacy. IEEE Computer Society, 110.
[38] Forrest, S., Esponda, F., and Helman, P. 2004. A formal framework for positive and negative detection schemes. In IEEE Transactions on Systems, Man and Cybernetics, Part B. IEEE, 357 - 373.
[39] Forrest, S., Hofmeyr, S. A., Somayaji, A., and Longstaff, T. A. 1996. A sense of self for Unix processes. In Proceedinges of the ISRSP96. 120 - 128.
[40] Forrest, S., Perelson, A. S., Allen, L., and Cherukuri, R. 1994. Self nonself discrimination in a computer. In Proceedings of the 1994 IEEE Symposium on Security and

Privacy. IEEE Computer Society, Washington, DC, USA, 202.

[41] Forrest, S., Warrender, C., and Pearlmutter, B. 1999. Detecting intrusions using system calls: Alternate data models. In Proceedings of the 1999 IEEE ISRSP. IEEE Computer Society, Washington, DC, USA, 133 - 145.

[42] Hofmeyr, S. A., Forrest, S., and Somayaji, A. 1998. Intrusion detection using sequences of system calls. Journal of Computer Security 6, 3, 151 - 180.

[43] Jagadish, H. V., Koudas, N., and Muthukrishnan, S. 1999. Mining deviants in a time series database. In Proceedings of the 25th International Conference on Very Large Data Bases. Morgan Kaufmann Publishers Inc., 102 - 113.

[44] Cabrera, J. B. D., Lewis, L., and Mehra, R. K. 2001. Detection and classification of intrusions and faults using sequences of system calls. SIGMOD Records 30, 4, 25 - 34.

[45] Gonzalez, F. A. and Dasgupta, D. 2003. Outlier detection using real-valued negative selection. Genetic Programming and Evolvable Machines 4, 4, 383- 403.

[46] Eskin, E. 2000. Outlier detection over noisy data using learned probability distributions. In Proceedings of the Seventeenth International Conference on Machine Learning. Morgan Kaufmann Publishers Inc., 255 - 262.

[47] Ghosh, A. K., Wanken, J., and Charron, F. 1998. Detecting anomalous and unknown intrusions against programs. In Proceedings of the 14th Annual Computer Security Applications Conference. IEEE Computer Society, 259

[48] Lee, W. and Stolfo, S. 1998. Data mining approaches for intrusion detection. In Proceedings of the 7th USENIX Security Symposium. San Antonio, TX.

[49] Lee, W., Stolfo, S., and Chan, P. 1997. Learning patterns from Unix process execution traces for intrusion detection. In Proceedings of the AAAI 97 workshop on AI methods in Fraud and risk management.

[50] Lee, W., Stolfo, S. J., and Mok, K. W. 2000. Adaptive intrusion detection: A data mining approach. Artificial Intelligence Review 14, 6, 533 - 567.

[51] Anderson, Lunt, Javitz, Tamaru, A., and Valdes, A. 1995. Detecting unusual program behavior using the statistical components of NIDES. Tech. Rep. SRI - CSL - 95 - 06, Computer Science Laboratory, SRI International.

[52] Anderson, D., Frivold, T., Tamaru, A., and Valdes, A. 1994. Next-generation intrusion detection expert system (NIDES), software users manual, beta-update release. Tech. Rep. SRI CSL - 95 - 07, Computer Science Laboratory, SRI International.

[53] Porras, P. A. and Neumann, P. G. 1997. EMERALD: Event monitoring enabling responses to anomalous live disturbances. In Proceedings of 20th NIST-NCSC National Information Systems

[54] Yamanishi, K. and Ichi Takeuchi, J. 2001. Discovering outlier filtering rules from unlabeled data: combining a supervised learner with an unsupervised learner. In Proceedings of the 7th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM Press, 389 - 394.

[55] Gwadera, R., Atallah, M. J., and Szpankowski, W. 2005b. Reliable detection of episodes in event sequences. Knowledge and Information Systems 7, 4, 415 - 437.

[56] Chow, C. and Yeung, D.-Y. 2002. Parzen-window network intrusion detectors. In Proceedings of the 16th International Conference on Pattern Recognition. Vol. 4. IEEE Computer Society, Washington, DC, USA, 40385.

[57] Siaterlis, C. and Maglaris, B. 2004. Towards multisensor data fusion for DoS detection. In Proceedings of the 2004 ACM symposium on Applied computing. ACM Press, 439 - 446.

[58] Sebyala, A. A., Olukemi, T., and Sacks, L. 2002. Active platform security through intrusion detection using naive Bayesian network for outlier detection. In Proceedings of the 2002 London Communications Symposium.

[59] Valdes, A. and Skinner, K. 2000. Adaptive, model-based monitoring for cyber attack detection. In Proceedings of the 3rd International Workshop on Recent Advances in Intrusion Detection. Springer-Verlag, 80 - 92.

[60] Bronstein, A., Das, J., Duro, M., Friedrich, R., Kleyner, G., Mueller, M., Singhal, S., and Cohen, I. 2001. Bayesian networks for detecting anomalies in internet based services. In International Symposium on Integrated Network Management

[61] Barbara, D., Couto, J., Jajodia, S., and Wu, N. 2001a. Adam: a testbed for exploring the use of data mining in intrusion detection. SIGMOD Rec. 30, 4, 15 - 24.

[62] Zhang, Z., Li, J., Manikopoulos, C., Jorgenson, J., and Ucles, J. 2001. HIDE: a hierarchical network intrusion detection system using statistical preprocessing and neural network classification. In Proceedings of IEEE Workshop on Information Assurance and Security. West Point, 85 - 90.

[63] Labib, K. and Vemuri, R. 2002. NSOM: A real-time network-based intrusion detection using self-organizing maps. Networks and Security.

[64] Smith, R., Bivens, A., Embrechts, M., Palagiri, C., and Szymanski, B. 2002. Clustering approaches for outlier based intrusion detection. In Proceedings of Intelligent Engineering Systems through Artificial Neural Networks. ASME Press, 579 - 584.

[65] Williams, G., Baxter, R., He, H., Hawkins, S., and Gu, L. 2002. A comparative study of rnn for outlier detection in data mining. In Proceedings of the 2002 IEEE International Conference on Data Mining. IEEE Computer Society, Washington, DC, USA, 709.

[66] Kruegel, C., Mutz, D., Robertson, W., and Valeur, F. 2003. Bayesian event classification for intrusion detection. In Proceedings of the 19th Annual Computer Security Applications Conference. IEEE Computer Society, 14.

[67] Manikopoulos, C. and Papavassiliou, S. 2002. Network intrusion and fault detection: a statistical outlier approach. IEEE Communication Magazine 40.

[68] Ramadas, M., Ostermann, S., and Tjaden, B. C. 2003. Detecting anomalous network traffic with self-organizing maps. In Proceedings of Recent Advances in Intrusion Detection. 36 - 54.

[69] Ghosh, S. and Reilly, D. L. 1994. Credit card fraud detection with a neural-network. In Proceedings

of the 27th Annual Hawaii International Conference on System Science. Vol. 3. Los Alamitos, CA.

[70] Brause, R., Langsdorf, T., and Hepp, M. 1999. Neural data mining for credit card fraud detection. In Proceedings of IEEE International Conference on Tools with Artificial Intelligence. 103 - 106.

[71] Dorronsoro, J. R., Ginel, F., Sanchez, C., and Cruz, C. S. 1997. Neural fraud detection in credit card operations. IEEE Transactions On Neural Networks 8, 4 (July), 827 - 834.

[72] Bolton, R. and Hand, D. 1999. Unsupervised profiling methods for fraud detection. In Credit Scoring and Credit Control VII.

[73] Fawcett, T. and Provost, F. 1999. Activity monitoring: noticing interesting changes in behavior. In Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM Press, 53 - 62.

[74] Cox, K. C., Eick, S. G., Wills, G. J., and Brachman, R. J. 1997. Visual data mining: Recognizing telephone calling fraud. Journal of Data Mining and Knowledge Discovery 1, 2, 225 - 231.

[75] Aggarwal, C. 2005. On abnormality detection in spuriously populated data streams. In Proceedings of 5th SIAM Data Mining. 80 - 91.

[76] Scott, S. L. 2001. Detecting network intrusion using a Markov modulated non homogeneous Poisson process. Submitted to the Journal of the American Statistical Association.

[77] Barson, P., Davey, N., Field, S. D. H., Frank, R. J., and McAskie, G. 1996. The detection of fraud in mobile phone networks. Neural Network World 6, 4.

[78] Taniguchi, M., Haft, M., Hollmn, J., and Tresp, V. 1998. Fraud detection in communications networks using neural and probabilistic methods. In Proceedings of IEEE International Conference in Acoustics, Speech and Signal Processing. Vol. 2. IEEE Computer Society, 1241 - 1244.

[79] Phua, C., Alahakoon, D., and Lee, V. 2004. Minority report in fraud detection: classification of skewed data. SIGKDD Explorer Newsletter 6, 1, 50 - 59.

[80] He, Z., Xu, X., and Deng, S. 2003. Discovering Cluster-based local outliers. Pattern Recognition Letters 24, 9-10, 1641 - 1650.

[81] Brockett, P. L., Xia, X., and Derrig, R. A. 1998. Using Kohonen's self-organizing feature map to uncover automobile bodily injury claims fraud. Journal of Risk and Insurance 65, 2 (June), 245 - 274.

[82] Donoho, S. 2004. Early detection of insider trading in option markets. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM Press, New York, NY, USA, 420 - 429.

[83] Anscombe, F. J. and Guttman, I. 1960. Rejection of outliers. Technometrics 2, 2, 123 - 147. Arning, A., Agrawal, R., and Raghavan, P. 1996. A linear method for deviation detection in large databases. In Proceedings of 2nd International Conference of Knowledge Discovery and Data Mining. 164 - 169.

[84] Horn, P. S., Feng, L., Li, Y., and Pesce, A. J. 2001. Effect of outliers and non healthy individuals on reference interval estimation. Clinical Chemistry 47, 12, 2137 - 2145.

[85] Laurikkala, J., Juhola1, M., and Kentala., E. 2000. Informal identification of outliers in medical data. In Fifth International Workshop on Intelligent Data Analysis in Medicine and Pharmacology. 20 - 24.

[86] Solberg, H. E. and Lahti, A. 2005. Detection of outliers in reference distributions: Performance of Horn's algorithm. Clinical Chemistry 51, 12, 2326 - 2332.

[87] Roberts, S. 1999. Novelty detection using extreme value statistics. In Proceedings of IEEE - Vision, Image and Signal processing. Vol. 146. 124 - 129.

[88] Suzuki, E., Watanabe, T., Yokoi, H., and Takabayashi, K. 2003. Detecting interesting exceptions from medical test data with visual summarization. In Proceedings of the 3rd IEEE International Conference on Data Mining. 315 - 322.

[89] Campbell, C. and Bennett, K. 2001. A linear programming approach to novelty detection. In Proceedings of Advances in Neural Information Processing. Vol. 14. Cambridge Press.

[90] Wong, W.-K., Moore, A., Cooper, G., and Wagner, M. 2003. Bayesian network outlier pattern detection for disease outbreaks. In Proceedings of the 20th International Conference on Machine Learning. AAAI Press, Menlo Park, California, 808 - 815.

[91] Lin, J., Keogh, E., Fu, A., and Herle, H. V. 2005. Approximations to magic: Finding unusual medical time series. In Proceedings of the 18th IEEE Symposium on Computer-Based Medical Systems. IEEE Computer Society, Washington, DC, USA, 329 - 334.

[92] Guttormsson, S., II, R. M., and El Sharkawi, M. 1999. Elliptical novelty grouping for on-line short-turn detection of excited running rotors. IEEE Transactions on Energy Conversion 14, 1 (March).

[93] Lin, J., Keogh, E., Fu, A., and Herle, H. V. 2005. Approximations to magic: Finding unusual medical time series. In Proceedings of the 18th IEEE Symposium on Computer-Based Medical Systems. IEEE Computer Society, Washington, DC, USA, 329 - 334.

[94] Keogh, E., Lonardi, S., and Chi' Chiu, B. Y. 2002. Finding surprising patterns in a time series database in linear time and space. In Proceedings of the eighth ACM SIGKDD International conference on Knowledge discovery and data mining. ACM Press, New York, NY, USA, 550 - 556.

[95] Keogh, E., Lin, J., Lee, S.-H., and Herle, H. V. 2006. Finding the most unusual time series subsequence: algorithms and applications. Knowledge and Information Systems 11, 1, 1 - 27.

[96] Desforges, M., Jacob, P., and Cooper, J. 1998. Applications of probability density estimation to the detection of abnormal conditions in engineering. In Proceedings of Institute of Mechanical Engineers. Vol. 212. 687 - 703.

[97] Bishop, C. 1994. Novelty detection and neural network validation. In Proceedings of IEEE Vision, Image and Signal Processing. Vol. 141. 217 - 222.

[98] Diaz, I. and Hollmen, J. 2002. Residual generation and visualization for understanding novel process conditions.

In Proceedings of IEEE International Joint Conference on Neural Networks. IEEE, Honolulu, HI, 2070 - 2075.

[99] Harris, T. 1993. Neural network in machine health monitoring. Professional Engineering. Hartigan, J. A. and Wong, M. A. 1979. A k-means clustering algorithm. Applied Statistics 28, 100 - 108.

[100] Jakubek, S. and Strasser, T. 2002. Fault-diagnosis using neural networks with ellipsoidal basis functions. In Proceedings of the American Control Conference. Vol. 5. 3846 - 3851.

[101] King, S., King, D., P. Anuzis, K. A., Tarassenko, L., Hayton, P., and Utete, S. 2002. The use of novelty detection techniques for monitoring high-integrity plant. In Proceedings of the 2002 International Conference on Control Applications. Vol. 1. Cancun, Mexico, 221 - 226.

[102] Li, Y., Pont, M. J., and Jones, N. B. 2002. Improving the performance of radial basis function classifiers in condition monitoring and fault diagnosis applications where unknown faults may occur. Pattern Recognition Letters 23, 5, 569 - 577.

[103] Petsche, T., Marcantonio, A., Darken, C., Hanson, S., Kuhn, G., and Santoso, I. 1996. A neural network auto associator for induction motor failure prediction. In Proceedings of Advances in Neural Information Processing. Vol. 8. 924 - 930.

[104] Streifel, R., Maks, R., and El-Sharkawi, M. 1996. Detection of shorted-turns in the field of turbine-generator rotors using novelty detectors - development and field tests. IEEE Transactions on Energy Conversations 11, 2, 312 - 317.

[105] Whitehead, B. and Hoyt, W. 1993. A function approximation approach to outlier detection in propulsion system test data In Proceedings of 29th AIAA/SAE/ASME/ASEE Joint Propulsion Conference. IEEE Computer Society, Monterey, CA, USA.

[106] Parra, L., Deco, G., and Miesbach, S. 1996. Statistical independence and novelty detection with information preserving nonlinear maps. Neural Computing 8, 2, 260 - 269.

[107] Yairi, T., Kato, Y., and Hori, K. 2001. Fault detection by mining association rules from house-keeping data. In In Proceedings of International Symposium on Artificial Intelligence, Robotics and Automation in Space.

[108] Manson, G. 2002. Identifying damage sensitive, environment insensitive features for damage detection. In Proceedings of the IES Conference. Swansea, UK.

[109] Manson, G., Pierce, G., and Worden, K. 2001. On the long-term stability of normal condition for damage detection in a composite panel. In Proceedings of the 4th International Conference on Damage Assessment of Structures. Cardiff, UK.

[110] Manson, G., Pierce, S. G., Worden, K., Monnier, T., Guy, P., and Atherton, K. 2000. Long-term stability of normal condition data for novelty detection. In Proceedings of Smart Structures and Integrated Systems. 323 - 334.

[111] Ruotolo, R. and Surace, C. 1997. A statistical approach to damage detection through vibration monitoring. In Proceedings of the 5th Pan American Congress of Applied Mechanics. Puerto Rico.

[112] Hickinbotham, S. J. and Austin, J. 2000a. Novelty detection in airframe strain data. In Proceedings of 15th International Conference on Pattern Recognition. Vol. 2. 536 - 539.

[113] Hickinbotham, S. J. and Austin, J. 2000b. Novelty detection in airframe strain data. In Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. Vol. 6. 24 - 27.

[114] Hollier, G. and Austin, J. 2002. Novelty detection for strain-gauge degradation using maximally correlated components. In Proceedings of the European Symposium on Artificial Neural Networks. 257 - 262 - 539.

[115] Brotherton, T. and Johnson, T. 2001. Outlier detection for advance military aircraft using neural networks. In Proceedings of 2001 IEEE Aerospace Conference.

[116] Brotherton, T., Johnson, T., and Chadderdon, G. 1998. Classification and novelty detection using linear models and a class dependent - elliptical basis function neural network. In Proceedings of the IJCNN Conference. Anchorage AL.

[117] Nairac, A., Corbett-Clark, T., Ripley, R., Townsend, N., and Tarassenko, L. 1997. Choosing an appropriate model for novelty detection. In Proceedings of the 5th IEEE International Conference on Artificial Neural Networks. 227 - 232.

[118] Nairac, A., Townsend, N., Carr, R., King, S., Cowley, P., and Tarassenko, L. 1999. A system for the analysis of jet engine vibration data. Integrated Computer-Aided Engineering 6, 1, 53 - 56.

[119] Surace, C. and Worden, K. 1998. A novelty detection method to diagnose damage in structures: an application to an offshore platform. In Proceedings of Eighth International Conference of Off-shore and Polar Engineering. Vol. 4. Colorado, USA, 64 - 70.

[120] Surace, C., Worden, K., and Tomlinson, G. 1997. A novelty detection approach to diagnose damage in a cracked beam. In Proceedings of SPIE. Vol. 3089. 947 - 953.

[121] Sohn, H., Worden, K., and Farrar, C. 2001. Novelty detection under changing environmental conditions. In Proceedings of Eighth Annual SPIE International Symposium on Smart Structures and Materials. Newport Beach, CA.

[122] Worden, K. 1997. Structural fault detection using a novelty measure. Journal of Sound Vibration 201, 1, 85 - 101.

[123] Augusteijn, M. and Folkert, B. 2002. Neural network classification and novelty detection. International Journal on Remote Sensing 23, 14, 2891 - 2902.

[124] Byers, S. D. and Raftery, A. E. 1998. Nearest neighbor clutter removal for estimating features in spatial point processes. Journal of the American Statistical Association 93, 577 - 584.

[125] Moya, M., Koch, M., and Hostetler, L. 1993. One-class classifier networks for target recognition applications. In Proceedings on World Congress on Neural Networks, International Neural Network Society. Portland, OR, 797 - 801.

[126] Torr, P. and Murray, D. 1993. Outlier detection and motion segmentation. In Proceedings of SPIE, Sensor Fusion VI, Paul S. Schenker; Ed. Vol. 2059. 432 - 443.
[127] Theiler, J. and Cai, D. M. 2003. Re-sampling approach for outlier detection in multispectral images. In Proceedings of SPIE 5093, 230-240, Ed.
[128] Cun, Y. L., Boser, B., Denker, J. S., Howard, R. E., Habbard, W., Jackel, L. D., and Henderson, D. 1990. Handwritten digit recognition with a back-propagation network. Advances in neural information processing systems, 396 - 404.
[129] Spence, C., Parra, L., and Sajda, P. 2001. Detection, synthesis and compression in mammographic image analysis with a hierarchical image probability model. In Proceedings of the IEEE Workshop on Mathematical Methods in Biomedical Image Analysis. IEEE Computer Society, Washington, DC, USA, 3.
[130] Tarassenko, L. 1995. Novelty detection for the identification of masses in mammograms. In Proceedings of the 4th IEEE International Conference on Artificial Neural Networks. Vol. 4. Cambridge, UK, 442 - 447.
[131] Chen, D., Shao, X., Hu, B., and Su, Q. 2005. Simultaneous wavelength selection and outlier detection in multivariate regression of near-infrared spectra. Analytical Sciences 21, 2, 161 - 167.
[132] Davy, M. and Godsill, S. 2002. Detection of abrupt spectral changes using support vector machines. an application to audio signal segmentation. In Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing. Orlando, USA.
[133] Hazel, G. G. 2000. Multivariate Gaussian MRF for multispectral scene segmentation and outlier detection. GeoRS 38, 3 (May), 1199 - 1211.
[134] Scarth, G., McIntyre, M., Wowk, B., and Somorjai, R. 1995. Detection of novelty in functional images using fuzzy clustering. In Proceedings of the 3rd Meeting of International Society for Magnetic Resonance in Medicine. Nice, France, 238.
[135] Diehl, C. and Hampshire, J. 2002. Real-time object classification and novelty detection for collaborative video surveillance. In Proceedings of IEEE International Joint Conference on Neural Networks. IEEE, Honolulu, HI.
[136] Singh, S. and Markou, M. 2004. An approach to novelty detection applied to the classification of image regions. IEEE Transactions on Knowledge and Data Engineering 16, 4, 396 - 407.To Appear in ACM Computing Surveys, 09 2009.
[137] Pokrajac, D., Lazarevic, A., and Latecki, L. J. 2007. Incremental local outlier detection for data streams. In Proceedings of IEEE Symposium on Computational Intelligence and Data Mining.
[138] Song, Q., Hu, W., and Xie, W. 2002. Robust support vector machine with bullet hole image classification. IEEE Transactions on Systems, Man, and Cybernetics - Part C:Applications and Reviews 32, 4.
[139] Baker, D., Hofmann, T., McCallum, A., and Yang, Y. 1999. A hierarchical probabilistic model for novelty detection in text. In Proceedings of International Conference on Machine Learning.
[140] Manevitz, L. M. and Yousef, M. 2000. Learning from positive data for document classification using neural networks. In Proceedings of Second Bar-Ilan Workshop on Knowledge Discovery and Learning. Jerusalem.
[141] Manevitz, L. M. and Yousef, M. 2002. One-class SVMS for document classification. Journal of Machine Learning Research 2, 139 - 154.
[142] Allan, J., Carbonell, J., Doddington, G., Yamron, J., and Yang, Y. 1998. Topic detecion and tracking pilot study. In Proceedings of DARPA Broadcast News Transcription and Understanding Workshop. 194 - 218.
[143] Srivastava, A. and Zane-Ulman, B. 2005. Discovering recurring outliers in text reports regarding complex space systems. Aerospace Conference, 2005 IEEE, 3853 - 3862.
[144] Srivastava, A. 2006. Enabling the discovery of recurring outliers in aerospace problem reports using high-dimensional clustering techniques. Aerospace Conference, 2006 IEEE, 17 - 34.
[145] Chatzigiannakis, V., Papavassiliou, S., Grammatikou, M., and Maglaris, B. 2006. Hierarchical outlier detection in distributed large-scale sensor networks. In ISCC '06: Proceedings of the 11th IEEE Symposium on Computers and Communications. IEEE Computer Society, Washington, DC, USA, 761 - 767.
[146] Janakiram, D., Reddy, V., and Kumar, A. 2006. Outlier detection in wireless sensor networks using Bayesian belief networks. In First International Conference on Communication System Software and Middleware. 1 - 6.
[147] Branch, J., Szymanski, B., Giannella, C., Wolff, R., and Kargupta, H. 2006. In-network outlier detection in wireless sensor networks. In 26th IEEE International Conference on Distributed Computing Systems.
[148] Phuong, T. V., Hung, L. X., Cho, S. J., Lee, Y., and Lee, S. 2006. An outlier detection algorithm for detecting attacks in wireless sensor networks. Intelligence and Security Informatics 3975, 735 - 736.
[149] Du, W., Fang, L., and Peng, N. 2006. Lad: localization outlier detection for wireless sensor networks. J. Parallel Distrib. Comput. 66, 7, 874 - 886.
[150] Subramaniam, S., Palpanas, T., Papadopoulos, D., Kalogeraki, V., and Gunopulos, D.2006. Online outlier detection in sensor data using non-parametric models. In VLDB '06: Proceedings of the 32nd international conference on Very large data bases. VLDB Endowment, 187 - 198.
[151] Kejia Zhang, Shengfei Shi, H. G. and Li, J. 2007. Unsupervised outlier detection in sensor networks using aggregation tree. Advanced Data Mining and Applications 4632, 158 - 169.
[152] Ide,T., Papadimitriou, S., and Vlachos, M. 2007. Computing correlation outlier scores using stochastic nearest neighbors. In Proceedings of International Conference Data Mining. 523 - 528.
[153] Albrecht, S., Busch, J., Kloppenburg, M., Metze, F., and Tavan, P. 2000. Generalized radial basis function networks for classification and novelty detection: self-organization of optional Bayesian decision. Neural Networks 13, 10, 1075 - 1093.
[154] Emamian, V., Kaveh, M., and Tewfik, A. 2000. Robust clustering of acoustic emission signals using the Kohonen

network. In Proceedings of the IEEE International Conference of Acoustics, Speech and Signal Processing. IEEE Computer Society.

[155] Crook, P. and Hayes, G. 2001. A robot implementation of a biologically inspired method for novelty detection. In Proceedings of Towards Intelligent Mobile Robots Conference. Manchester, UK.

[156] Crook, P. A., Marsland, S., Hayes, G., and Nehmzow, U. 2002. A tale of two filters - on-line novelty detection. In Proceedings of International Conference on Robotics and Automation. 3894 - 3899.

[157] Marsland, S., Nehmzow, U., and Shapiro, J. 1999. A model of habituation applied to mobile robots. In Proceedings of Towards Intelligent Mobile Robots. Department of Computer Science, Manchester University, Technical Report Series, ISSN 1361-6161, Report UMCS-99-3-1.

[158] Marsland, S., Nehmzow, U., and Shapiro, J. 2000b. A real-time novelty detector for a mobile robot. In Proceedings of the EUREL Conference on Advanced Robotics Systems.

[159] Marsland, S., Nehmzow, U., and Shapiro, J. 2000a. Novelty detection for robot neotaxis. In Proceedings of the 2nd International Symposium on Neural Compuatation. 554 - 559.

[160] Shekhar, S., Lu, C.-T., and Zhang, P. 2001. Detecting graph-based spatial outliers: algorithms and applications (a summary of results). In Proceedings of the 7th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM Press, New York, NY, USA, 371 - 376.

[161] Ihler, A., Hutchins, J., and Smyth, P. 2006. Adaptive event detection with time-varying poisson processes. In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM Press, New York, NY, USA, 207 - 216.

[162] Ide, T. and Kashima, H. 2004. Eigenspace-based outlier detection in computer systems. In Proceedings of the 10th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM Press, New York, NY, USA, 440 - 449.

[163] Sun, J., Qu, H., Chakrabarti, D., and Faloutsos, C. 2005. Neighborhood formation and outlier detection in bipartite graphs. In Proceedings of the 5th IEEE International Conference on Data Mining. IEEE Computer Society, Washington, DC, USA, 418 - 425.To Appear in ACM Computing Surveys, 09 2009.

[164] MacDonald, J. W. and Ghosh, D. 2007. Copa - cancer outlier profile analysis. Bioinformatics 22, 23, 2950 - 2951.

[165] Kadota, K., Tominaga, D., Akiyama, Y., and Takahashi, K. 2003. Detecting outlying samples in microarray data: A critical assessment of the effect of outliers on sample classification. Chem-Bio Informatics 3, 1, 30 - 45.

[166] Tomlins, S. A., Rhodes, D. R., Perner, S., Dhanasekaran, S. M., Mehra, R., Sun, X. W., Varambally, S., Cao, X., Tchinda, J., Kuefer, R., Lee, C., Montie, J. E., Shah, R., Pienta, K. J., Rubin, M., and Chinnaiyan, A. M. 2005. Recurrent fusion of tmprss2 and ETS transcription factor genes in prostate cancer. Science 310, 5748, 603 - 611.

[167] Tibshirani, R. and Hastie, T. 2007. Outlier sums for differential gene expression analysis. Biostatistics 8, 1, 2 - 8.

[168] Lu, C.-T., Chen, D., and Kou, Y. 2003. Algorithms for spatial outlier detection. In Proceedings of 3rd International Conference on Data Mining. 597 - 600.

[169] Lin, S. and Brown, D. E. 2003. An outlier-based data association method for linking criminal incidents. In Proceedings of 3rd SIAM Data Mining Conference.

[170] He, Z., Xu, X., Huang, J. Z., and Deng, S. 2004b. Mining class outliers: Concepts, algorithms and applications. 588 - 589.

[171] Dutta, H., Giannella, C., Borne, K., and Kargupta, H. 2007. Distributed top-k outlier detection in astronomy catalogs using the DEMAC system. In Proceedings of 7th SIAM International Conference on Data Mining.

[172] Escalante, H. J. 2005. A comparison of outlier detection algorithms for machine learning. In Proceedings of the International Conference on Communications in Computing.

[173] Protopapas, P., Giammarco, J. M., Faccioli, L., Struble, M. F., Dave, R., and Alcock, C.2006. Finding outlier light curves in catalogues of periodic variable stars. Monthly Notices of the Royal Astronomical Society 369, 2, 677 - 696.

[174] Blender, R., Fraedrich, K., and Lunkeit, F. 1997. Identification of cyclone-track regimes in the North Atlantic. Quarterly Journal of the Royal Meteorological Society 123, 539, 727 - 741.

[175] Sun, P. and Chawla, S. 2004. On local spatial outliers. In Proceedings of 4th IEEE International Conference on Data Mining. 209 - 216.

[176] Sun, P., Chawla, S., and Arunasalam, B. 2006. Mining for outliers in sequential databases. In In SIAM International Conference on Data Mining.