The Journal of
Mathematics and Computer Science

Available online at
http://www.TJMCS.com

The Journal of Mathematics and Computer Science Vol. 4 No.2 (2012) 129 - 138

Surveying Different Aspects of Anomaly Detection and Its Applications

Neda Noori[1], Leila Boti[2], Ebrahim Nowzarpoor Shami[3]

[1]Islamic Azad University, Zanjan branch, Zanjan, Iran,
nedanoori_827@yahoo.com
[2] Islamic Azad University, Zanjan branch, Zanjan, Iran,
boti_leila@yahoo.com
[3] Islamic Azad University, Zanjan branch, Zanjan, Iran,
enowsl@aol.com

Received: January 2012, Revised: April 2012
Online Publication: June 2012

ABSTRACT:

Detecting anomalies is a significant issue which is being investigated in different levels of research and application. Many techniques of anomaly detection are widely and specifically developed for special applied domains while other techniques are mostly general. The purpose of this article is to provide a concise and comprehensive summary of surveys and researches related to anomaly detection. We have categorized the available techniques based on the certified methods; and to distinguish between normal and abnormal behaviors, we have defined some key hypotheses which are used by techniques. When a given technique turns out to be efficient for a specific domain, its hypotheses can provide strategies for accessing technique-efficiency in that domain. For each category, we have provided a basic technique for anomaly detection, and then we have shown how different basic techniques which are available in any category are distinguished from the defined basic technique. This procedure gives us a brief and simple understanding of the techniques which belong to any category. In addition, we will define the pros and cons of the techniques of each category. We hope that this article provides a better understanding of different aspects of the investigated issues, and that how the techniques which are developed in one area can turn out to be efficient in other areas which have not been part of the presuppositions at first.

Keywords: Anomaly detection, Outlier detection.

1. INTRODUCTION

"Anomaly detection" refers to the problem of finding patterns in the data which do not conform to expected behaviors. Such non-conforming patterns usually include "anomalies", "outliers" and "discordant observations". Of these, "anomalies" and "outliers" are the terms which are more prevalent in detecting the exceptions, and sometimes they are used interchangeably. Anomaly detection is practical in a wide range of applications including Credit Card Fraud Detection, Insurance, Network and Cyber-security Intrusion Detection, Troubleshooting of the Safety Critical System, and military surveillance enemy activities. The significance of anomaly detection is due to the fact that anomalies in data are rendered into essential and sensitive information in various application domains. For example, 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 anomalous MRI image may indicate presence of malignant tumors (2). Anomalies in credit card transaction data could indicate credit card or identity theft (3), or anomalous readings from a space craft sensor could signify a fault in some component of the space craft (4).

Detecting anomalies in data has been studied in the statistics community as early as the 19th century. Over time, a variety of anomaly detection techniques have been developed in several research communities. Many of these techniques have been specifically developed for certain application domains, while others are more generic. This survey tries to provide a structured and comprehensive overview of the research on anomaly detection.

In 2004, Hodge and Austin provide an extensive survey of anomaly detection techniques developed in machine learning and statistical domains. In 2006, a broad review of anomaly detection techniques for numerous data is presented by Agyemang. In 2003, an extensive review of anomaly detection techniques using neural networks and statistical approaches has been presented. In 2007, Patch and Park presented a survey of anomaly detection techniques used specifically for cyber-intrusion detection.

This survey that is an attempt to provide a structured and a broad overview of extensive research on anomaly detection is organized in six parts. In the second part we will deal with the concept of anomaly. The third section introduces the challenges which are on the way of anomaly detection, and different aspects of the problem of anomaly detection are taken into account in part four. In the fifth section we will examine the applications of anomaly detection, and finally, the overall conclusion of the study will be presented in part six.

2. What Are the Anomalies?

Anomalies are patterns in data that do not conform to a well defined notion of normal data. Figure 1 illustrates anomalies in a simple 2-dimensional data set. The data have 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 anomalies.

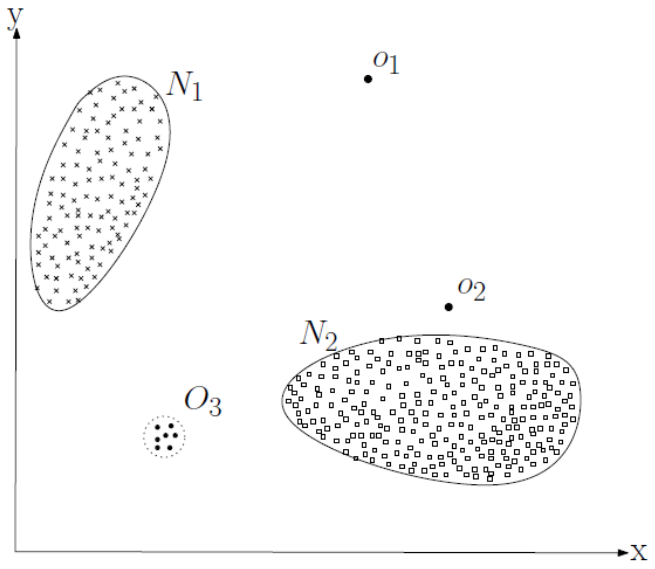

Fig. 1. A simple example of anomalies in a 2-dimensional data set.

Anomalies 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 all of the reasons have a common characteristic that they are interesting to the analyst. The real life relevance of anomalies is a key feature of anomaly detection.

Anomaly detection depends on noise removal and noise accommodation. Noise can be a phenomenon in data which is of interest to the analyst, but acts as a hindrance to data analysis. Noise removal means removing the unwanted objects before any data analysis is performed on the data (5). Noise accommodation encompasses designing a statistical model for observing anomalies.

Another topic related to anomaly detection is novelty detection which aims at detecting previously unobserved patterns in the data. The distinction between novel patterns and anomalies is that the novel patterns are typically incorporated into the normal model after being detected. It should be noted that solutions for above mentioned problems are often used for anomaly detection and vice-versa. Table (1) shows the set of techniques and application domains.

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

Table 1. Techniques and application domains of anomaly detection

3. Challenges of Anomaly Detection

A straightforward anomaly detection approach is to define a region representing normal behavior and declare any observation in the data which does not belong to this normal region as an anomaly. But several factors make this apparently simple approach very challenging:

- Defining a normal region which encompasses every possible normal behavior is very difficult. In addition, the boundary between normal and anomalous behavior is often not precise. Thus an anomalous observation which lies close to the boundary can actually be normal, and vice-versa.
- When anomalies are the result of malicious actions, the malicious adversaries often adapt themselves to make the anomalous observations appear like normal, thereby making the task of defining normal behavior more difficult.

- In many domains normal behavior keeps evolving and a current notion of normal behavior might not be sufficiently representative in the future.
- The exact notion of an anomaly is different for various application domains. For example, in the medical domain a small deviation from normal (e.g., fluctuations in body temperature) might be an anomaly, while similar deviation in the store of a market might be considered as normal. Thus applying a technique developed in one domain to another is not straightforward.
- Availability of labeled data for training and validation of models used by anomaly detection techniques is usually a major issue.
- Often the data contains noise which tends to be similar to the actual anomalies and hence is difficult to distinguish and remove.

Due to the above challenges, the anomaly detection problem, in its most general form, is not easy to solve. In fact, most of the existing anomaly detection techniques solve a specific formulation of the problem. The formulation is induced by various factors such as nature of the

data, availability of labeled data, type of anomalies to be detected, etc. Often, these factors are determined by the application domain in which the anomalies need to be detected. Researchers have adopted concepts from diverse disciplines such as statistics, machine learning, data mining, and information theory, and they have applied them to specific problem formulations.

4. Different Aspects of the Anomaly Detection Problem

This section identifies and discusses the different aspects of anomaly detection. As mentioned earlier, a specific formulation of the problem is determined by several different factors such as the structure and nature of the input data, and the availability or unavailability of labels.

4.1. Structure and Nature of the Input Data

A key aspect of any anomaly detection technique is the structure and nature of the input data. Input is generally a collection of data instances (also referred as object, record, point, vector, pattern, event, case, sample, observation, entity). Each data instance can be described using a set of attributes. 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 and structure of attributes determine the applicability of anomaly 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 pairwise distance between instances might be provided in the form of a similarity matrix. In such cases, techniques that require original data instances are not applicable, e.g., many statistical and classification based techniques. Input data can also be categorized based on the relationship present among data instances. Most of the existing anomaly detection techniques deal with record data, in which no relationship is assumed among the data instances. In general, data instances can be related to each other. Some examples are sequence data, multidimensional data, and graph data. In sequence data, the data instances are linearly ordered, e.g., time-series data, genome sequences, protein sequences. In multidimensional data, each data instance is related to its neighboring instances, e.g., traffic data in media. When the spatial data have a temporal 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 in this section we will discuss situations where such relationship among data instances becomes relevant for anomaly detection.

4.2. Types of Anomaly

An important aspect of an anomaly detection technique is the nature and structure of the desired anomaly. Anomalies can be classified into following categories:

4.2.1. Point Anomalies: 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 anomaly. This is the simplest type of anomaly and is the focus of majority of research on anomaly 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 anomalies since they are different from normal data points. As a real life example, consider credit card fraud detection. Let the data set correspond to an individual's credit card transactions. For the sake of simplicity, let us assume that the data is defined using 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 anomaly.

4.2.2. Contextual Anomalies: If a data instance is anomalous in a specific context and concept, then it is termed as a contextual anomaly (7). Each data instance is defined using following 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 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 anomaly in a given context, but an identical data instance could be considered normal in a different context. This property is key in identifying contextual and behavioral attributes for a contextual anomaly detection technique. Contextual anomalies have been most commonly used in time-series data and multidimensional data. Figure 2 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 anomaly.

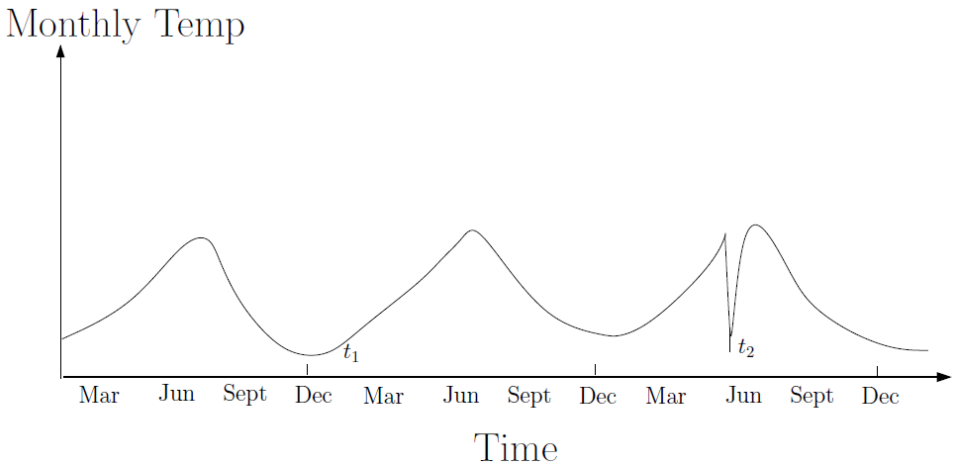

Fig. 1. Contextual anomaly at time t_2 in a temperature time series

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

anomalous. Figure 3 illustrates an example which shows a human electrocardiogram output. The highlighted region denotes an anomaly because the same low value exists for an abnormally long time. Note that that low value by itself is not an anomaly.

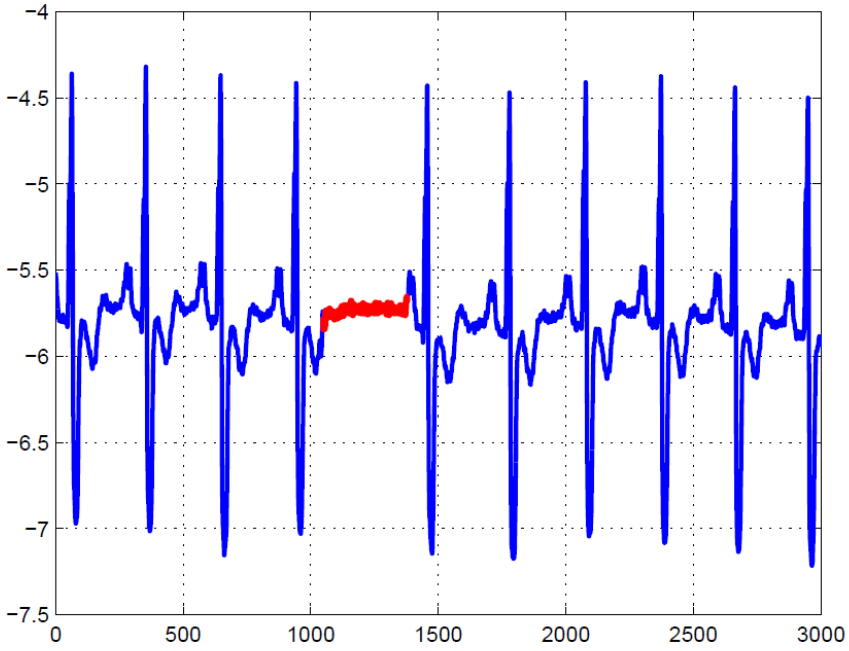

Fig. 3. Collective anomaly in a human electrocardiogram output

4.3. 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 difficult. 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 anomalous behavior is often dynamic in nature, e.g., new types of anomalies might arise, for which there is no labeled training data. In certain cases, such as air traffic safety, anomalous instances would translate to catastrophic events, and hence will be very rare. Based on the extent to which the labels are available, anomaly detection techniques can operate in one of the following three modes:

4.3.1. Supervised anomaly detection: Techniques trained in supervised mode assume the availability of a training data set which has labeled instances for normal as well as anomaly class. Typical approach in such cases is to build a predictive model for normal vs. anomaly 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 anomaly detection. First, the anomalous instances are far fewer compared to the normal instances in the training data. Second, obtaining accurate and representative labels, especially for the anomaly class is usually challenging. A number of techniques have been proposed that inject artificial anomalies in a

normal data set to obtain a labeled training data set (8). Other than these two issues, the supervised anomaly detection problem is similar to building predictive models.

4.3.2. Semi-Supervised anomaly 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 anomaly class, they are more widely applicable than supervised techniques. For example, in space craft fault detection, an anomaly 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 anomalies in the test data. A limited set of anomaly detection techniques exist that assumes availability of only the anomaly instances for training. 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.

4.3.3. Unsupervised anomaly 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 anomalies 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 anomalies and the model learnt during training is robust to these few anomalies.

4.3.4. Output of anomaly detection: An important aspect for any anomaly detection technique is the output of anomaly. Typically, the outputs produced by anomaly detection techniques are one of the following two types:

4.3.4.1. Scores: Scoring techniques assign an anomaly score to each instance in the test data depending on the degree to which that instance is considered an anomaly. Thus the output of such techniques is a list of anomalies. An analyst may choose to use a cut-off threshold to select the anomalies.

4.3.4.2. Labels: Techniques in this category assign a label (normal or anomalous) to each test instance. Scoring based anomaly detection techniques allow the analyst to use a specific threshold to select the most relevant anomalies. Techniques that provide binary labels to the test instances do not directly allow the analysts to make such a choice, and this can be controlled indirectly through parameter choices within each technique.

5. Applications of Anomaly Detection

In this section we discuss several applications of anomaly detection.

5.1. Intrusion Detection: Intrusion detection refers to detection of malicious activity in a computer related system (9). These malicious activities or intrusions are interesting and significant from a computer security perspective. An intrusion is different from the normal behavior of the

system, and hence anomaly detection techniques are applicable in intrusion detection domain. The key challenge for anomaly detection in this domain is the huge volume of data. The anomaly detection techniques need to be computationally efficient to handle these large sized inputs. Moreover the data typically comes in a streaming fashion, thereby requiring online analysis. Another issue which arises because of the large sized input is the false alarm rate. Since the data amounts to millions of data objects, a few percent of false alarms can make analysis overwhelming for an analyst. Labeled data corresponding to normal behavior is usually available, while labels for intrusions are not. Thus, semi-supervised and unsupervised anomaly detection techniques are preferred in this domain.

5.2. Fraud Detection: Fraud detection refers to detection of criminal activities occurring in commercial organizations such as banks, credit card companies, insurance agencies, stock market, etc. The malicious users might be the actual customers of the organization or might be posing as a customer that is known as identity theft. The fraud occurs when these users consume the resources provided by the organization in an unauthorized way. The organizations are interested in immediate detection of such frauds to prevent economic losses. Fawcett and Provost in 1999 introduce the term activity monitoring as a general approach to fraud detection. The typical approach of anomaly detection techniques is to maintain a usage profile for each customer and monitor the profiles to detect any deviations.

5.3. Medical and Public Health Anomaly Detection: Anomaly detection in the medical and public health domains typically work with patient records. The data can have anomalies due to several reasons such as abnormal patient condition or instrumentation errors. Several techniques have also focused on detecting disease outbreaks in a specific area (10). Thus the anomaly detection is a very critical problem in this domain and requires high degree of accuracy. The data typically consist of records which may have several different types of features such as patient age, blood group, weight. The most prevalent anomaly detection technique in this domain is the point anomaly detection. Typically the labeled data belong to the healthy patients; hence most of the techniques adopt semi-supervised approach.

6. CONCLUSION

The current survey attempts to provide a structured and extensive revision for the vast researches that have been done on anomaly detection techniques. Most of the available articles and surveys are focusing either on a specific application domain or on a specific research domain. Agyeman, Hodge, and Austin have classified anomaly detection into several categories and examine the techniques for each category. This survey follows such an arrangement, too, and does a significant and exhaustive expansion on the discussed issues. For the four categories, not only have we examined the techniques, but also we have defined unified assumptions for the nature and structure of the anomalies. These assumptions have a critical role in determining when the techniques in that core can detect the anomalies and when they fail. In addition, for each category, we defined a basic anomaly detection technique which shows how much the available techniques in that category are different from the basic techniques. Eventually, the applications of anomaly detection in various domains were investigated.

REFERENCE

[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 anomaly 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] Teng, H., Chen, K., and Lu, S. 1990. Adaptive real-time anomaly detection using inductively generated sequential patterns. In Proceedings of IEEE Computer Society Symposium on Research in Security and Privacy. IEEE Computer Society Press, 278-284.

[6] Rousseeuw, P. J. and Leroy, A. M. 1987. Robust regression and outlier detection. John Wiley & Sons, Inc., New York, NY, USA.

[7] Song, X., Wu, M., Jermaine, C., and Ranka, S. 2007. Conditional anomaly detection. IEEE Transactions on Knowledge and Data Engineering 19, 5, 631-645.

[8] Theiler, J. and Cai, D. M. 2003. Resampling approach for anomaly detection in multispectral images. In Proceedings of SPIE 5093, 230-240, Ed.

[9] Phoha, V. V. 2002. The Springer Internet Security Dictionary. Springer-Verlag.

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