CHAPTER 1

INTRODUCTION

Outlier detection has been a very important concept in the realm of data analysis. Recently, several application domains realized the direct mapping between outliers in data and real world anomalies that are of great interest to an analyst. Outlier detection refers to the problem of finding patterns in data that do not conform to expected normal behavior. These anomalous patterns are often referred to as outliers, anomalies, discordant observations, exceptions, faults, defects, aberrations, noise, errors, damage, surprise, novelty, peculiarities or contaminants in different application domains. Outlier detection has been a widely researched problem and immense use in a wide variety of application domains such as credit card, insurance, tax fraud detection, intrusion detection for cyber security, fault detection in safety conforming patterns in data; that is, they are patterns that do not exhibit normal behavior. Data mining is the principle of sorting through large amounts of data and picking out relevant information. Outlier detection is one of the data mining techniques that detects rare events, deviant objects, and exceptions. Most previous studies focused on finding outliers that are hidden in numerical datasets (Vries and Chawla 2010).

In this chapter, an overview of outlier detection is discussed. Section 1.1 discusses about outlier detection. Section 1.2 lists out the various types of outliers and section 1.3 briefs the various outlier detection methods. Section 1.4 elaborates the challenging issues in outlier detection. Section 1.5 elaborates the role of data mining

in outlier detection. The applications in outlier detection methods are discussed in section 1.6.

1.1 OUTLIER DETECTION

In many data processing tasks, a large amount of data is being collected and processed. One prime step in obtaining a coherent analysis is the detection of anomalous observations. Outlier detection refers to the problem of finding patterns in data that do not conform to expected normal behavior. Outlier detection involves the process of identifying data objects that do not compromise with the remaining objects in the data set. These anomalous patterns are often referred to as outliers, anomalies, exceptions, discordant observations, faults, aberrations, defects, noise, damage, errors, surprise, novelty, contaminants or peculiarities in different application domains. In spite of the terms noise or error, these are also considered to carry important information. Outlier detection methods have been suggested for numerous applications, such as credit card fraud detection, clinical trials, voting irregularity analysis, data cleansing, network intrusion, severe weather prediction, geographic information systems, athlete performance analysis, and other data-mining tasks. Detecting outliers is of utmost importance as they may lead to model misspecification, biased parameter estimation and incorrect results. This process of detecting outliers should be done prior to analysis and modeling. Outlier detection has been studied in the context of many research areas like statistics, data mining, sensor networks, environmental science, distributed systems, spatio-temporal mining, etc. Outlier detection has been studied on a large variety of data types including high-dimensional data, uncertain data, stream data, graph data, time series data, spatial data, and spatio-temporal data. Consequently, outlier detection is a quite

active research area with many new methods proposed every year, based on different underlying methodologies like statistical reasoning (Vries and Chawla 2010), distances (Knorr et al. 2000, Orair et al. 2009, Ramaswamy et al. 2000, Vu and Gopalkrishnan 2009, Zhang et al. 2009), or densities (Breunig et al. 2000, Ren et al. 2004, Kriegel et al. 2009).

The significance of outlier detection lies in the fact that outlier in data can be improved to useful information in an extensive array of application domain. For example, an abnormality in the traffic pattern in a network indicates that the computer is hacked and is sending out sensitive data to an unauthorized destination. In public health data, outlier detection techniques are widely used to detect anomalous patterns in patient medical records which may represent symptoms of a new disease. In general wellbeing information, exception discovery systems are broadly used to identify strange examples in patient healing report which may speak to manifestation of another disorder (Slezak et al. 2011, Zimek et al. 2014). Similarly, discordant observation in credit card transaction data could indicate misuse or theft of credit card. Outliers can also translate to vital entities such as in military scrutiny, where the presence of an unusual region in a satellite image of enemy area could indicate movement of enemy troop. An anomalous reading from a space craft would signify a fault in some component of the craft.

Outlier detection is the process of identifying those observations, which deviate substantially from the remaining data. Most of the definitions of outliers that exist in the statistics literature, are usually tied to certain assumptions on the underlying data distribution. There exists many classic examples for the definition of an outlier. A few of which are mentioned as follows. “An observation which deviates so much

from other observations as to arouse suspicions that it was generated by a different mechanism" (Hawkins 1980). "An observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data"(Vries and Chawla 2010). The varied definition of outliers indicate that, independent of the application, whether it is a traffic network, web server traffic, credit card data, sensor data in some scientific experiment, or the human metabolism, offers characteristic observations that could be predicted if the process was well-understood. Any unpredicted observation indicates a lack of understanding of the particular process, or is produced by a different process and therefore requires further investigation. It is an important data mining task with broad applications, such as credit card fraud detection, insurance claim fraud detection, medical diagnosis, image processing, intrusion detection, and event detection. Most information sets contain anomalies that have abnormally vast or little values when contrasted with others in the information set. The existence of anomalies can lead to expanded error rates and significant distortions of parameter and measurement gauges when utilizing either parametric or nonparametric tests. In order to deal with the problem of processing voluminous data efficient outlier detection method need to be used.

In Fig 1.1, point labeled O1and points labeled O2 deviate significantly from regions labeled G1and G2. Outlier detection has been studied extensively in the data mining research community. However, as the emergence of huge data sets in real-life practice nowadays, outlier detection faces a series of new challenges.

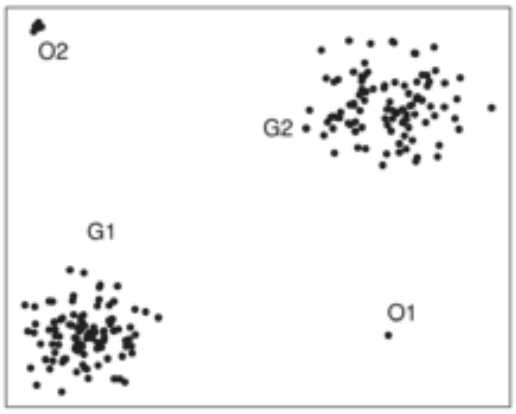

Fig.1.1 Example of two-dimensional outliers

Many traditional outlier detection methods do not work well in such an environment. Therefore, developing up-to-date outlier detection methods becomes urgent tasks. Despite the tremendous improvements in healthcare practice and systems nowadays, having sustainable healthcare is still one of the most important and urgent issues. From the individual point of view, healthcare is important to everyone, even for those who are currently in good health. No one wants to get sick, but everyone will get sick at some point of his life. From the global point of view, healthcare has a notable impact on a country's economy. Healthcare has many outlier detection applications.

1.2 OUTLIER TYPES

An important aspect of an outlier detection technique is the nature of the desired outlier. Three main types of outliers studied in literature are point outliers, contextual outliers, and collective outliers. Each of which is discussed as follows.

1.2.1 Point Outliers

If an individual data instance is 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. Point outliers are also referred to as global outliers. A point outlier remains distinct from other data points by representing its outlierness. They are detected by analyzing this outlierness metrics. This outlierness metric indicates the extent to which an individual data gets deviated from the other data in the data set. Consider a real life example dataset of a credit card fraud detection. The dataset contains the transaction details of an individual using a credit card. If one of the attributes of the dataset is the amount spent by the individual for a particular period, an anomaly can be identified as a transaction that is very high when compared to the normal range of the individual's expenditure.

1.2.2 Contextual Outliers

If a data instance is anomalous in a specific context, then it is termed as a contextual outlier. Contextual outliers is also sometimes referred to as conditional outliers. Naturally, a contextual outlier represents a small group of objects that are similar to some attributes with a significantly larger reference group of objects. This type of outliers deviates vividly from the reference group on some other attributes. 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 by using two sets of attributes. The contextual attributes determine the context or the neighboring object for that instance. In contrast, the behavioral attributes define the non-contextual characteristics of an instance. The choice of applying a contextual outlier

detection technique is determined by the meaningfulness of the contextual outliers in the target application domain. Application of a contextual outlier detection technique would make sense if the attributes are available readily. This makes the process of defining a context a straightforward one. On the contrary, if the context is not readily available or not defined easily, it becomes difficult to apply any such outlier detection techniques.

1.2.3 Collective Outliers

Any group of data instances related to each other and anomalous with respect to the entire data set, is termed as a collective outlier. Each data instance in a collective outlier may not be an outlier by itself. Instead, the occurrence of a group of such data instances together is said to be anomalous. Collective outliers have been explored from various types of data like sequence data, graph data, and spatial data. The occurrence of point outliers can happen in any type of dataset. But collective outliers have a restriction that it occurs only in the data sets where the data instances are related to each other. Contextual outliers have a necessity that the data instances should have context attributes in them. Any outlier, whether it is a point outlier or a collective outlier can also be a contextual outlier when analyzed with respect to a context. By incorporating a context information, an outlier detection problem to detect a point outlier or a collective outlier can be transformed to an outlier detection problem for detecting contextual outliers.

The output of an outlier detection algorithm can be one of two types. Most outlier detection algorithms give as output a score that describes the level of “outlierness” of that particular data point. This score value gives a ranking to each of the data points which describes their tendency to be an outlier. This type of output is very

general and retains all the information given by a particular algorithm. These types of algorithms do not provide a summary of the small number of data points which are normally considered as outliers. The second type of output is a binary value. It is actually a label which indicates that the data point is outlier or not. While some algorithms have the tendency to produce binary labels as direct output, some other algorithms have a special quality of converting the outlier scores what they get as output to equivalent binary labels. This can be achieved by applying a threshold on outlier score values, depending upon their statistical distribution. A scoring mechanism provides more information than a binary labeling information. Only the final outcome of the outlier detection algorithm which tells us whether the data point is an outlier or not is of more importance for decision making in practical applications rather than the information provided by the algorithm.

1.3 OUTLIER DETECTION METHODS

Outlier detection methods may be classified into three categories, namely supervised, unsupervised methods and semi-supervised methods. A brief introduction of all the methods is as follows.

1.3.1 Supervised Methods

Supervised techniques assumes the availability of a training data set. The training dataset has labeled instances for both normal as well as outlier class. The typical approach in such case is to build predictive models for both normal and outlier classes. Any unseen data instance is compared against the two models to determine which class it belongs to. Supervised outlier detection techniques have an explicit notion of the normal and outlier behavior and hence accurate models can be built.

One drawback here is that accurately labeled training data might be prohibitively expensive to obtain. The number of anomalous instances is very little when compared to the normal instances in the training data. Labeling process can be carried out manually by a human expert and hence requires a lot of effort to obtain the labeled training data set. Obtaining accurate and representative labels, especially for the outlier class is usually challenging. The supervised outlier detection problem is similar to building predictive models.

1.3.2 Semi-Supervised Methods

In the case of semi-supervised mode, the data objects in the training dataset are assigned with normal class label. Objects that are identified as anomalous does not require any labels, and hence are widely applicable when compared to supervised techniques. The characteristic approach used in such techniques is to build a model for the class objects corresponding to normal behavior. This model can be used 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. 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.

1.3.3 Unsupervised Methods

Unsupervised methods are widely applicable and do not require training data. These techniques do not require any knowledge of class labels. They do not make any assumption about the availability of labeled training data. The techniques in this category make other assumptions about the data. A frequently occurring pattern is typically considered normal while a rare occurrence is an outlier. The techniques

under this classification make an assumption that data objects with normal instances are far more frequent than outliers in the test data. Unsupervised techniques suffer from high false alarm rate, as the underlying assumption does not hold true often. Based on assumptions about normal data, the unsupervised outlier detection method can be further classified into

Statistical methods

A traditional approach to solve the outlier detection problem is based on the construction of a probabilistic data model and the use of statistical methods and probability theory. A probabilistic model can be either a priori given or automatically constructed by given data. Constructing a probabilistic model solves the problem of determining whether a particular object of the dataset belongs to the model or generated in accordance with some other distribution law. If the object does not suit the probabilistic model, it is considered an outlier. Probabilistic models are constructed with the use of standard probability distributions and combinations of such distributions. Sometimes, models include unknown parameters, which are determined in the course of data mining. Along with a priori given probability distributions, there exist algorithms for estimating probability distributions by empirical data.

Proximity-based methods

These methods examine the spatial proximity of each object in the data space, if the proximity of an object considerably deviates from the proximity of other objects it is considered an outlier. For deviation based approach given a set of data points, outliers are points that do not fit to the general characteristics of that set. Outliers are

the outermost points of the data set. For distance based approach, it judges a point based on the distance to its neighbors. Normal data objects are assumed to have a dense neighborhood and those objects identified as outliers will have less dense neighborhood. Outliers are normally identified as objects that are far apart from their neighbors and they are said to have less dense neighborhood. This is classified into two types, namely distance based and density based. Distance based method judge a point based on the distance(s) to its neighbors. Density based determines the degree of outlierness of each data instance based on its local density. In this algorithm, outliers are data objects with high local outlier factor (LOF) values whereas data objects with low LOF values are likely to be normal with respect to their neighborhood. High LOF is an indication of low-density neighborhood and hence high potential of being an outlier (Mansur et al. 2005).This approach identifies the data structure via density estimation.

Clustering-based methods

Clustering based approach (Breunig and Kriegel 2000) always apply a clustering based method on sample of data to characterize the local behavior of the data. The sub-clusters contain significantly less data points than the remaining clusters and are termed as outliers. Most of the earlier clustering based outlier detection methods found outliers as the byproduct of clustering. Hence, any data point that does not comply with any cluster is called an outlier. As the main aim is to find clusters, these approaches are not optimized to find outliers. The advantage of the cluster based technique is that they do not have to be supervised. Moreover, clustering based techniques is capable of being used in an incremental mode, i.e. after learning the clusters, new points can be inserted into the system and tested for the outliers.

Clustering based approaches are computationally expensive as they compromise huge computation of pairwise distances. Clustering algorithms are optimized to find clusters rather than outliers. Accuracy of outlier detection depends on how good the clustering algorithm captures the structure of clusters and set of many abnormal data objects that are similar to each other would be recognized as a cluster rather than as noise or outliers. The performance of outlier detection is limited because, the clustering based approaches are unsupervised and do not require labeled training data. Examples of this method are k-means clustering algorithm and fuzzy c-means (FCM).

1.4 CHALLENGES OF OUTLIER DETECTION

Conceptually speaking outliers are patterns that deviate from expected normal behavior, which in its simplest form could be represented by a region. At this point all normal observations are visualized as normal objects and the rest are considered as outliers. Even though this approach looks simple, it is certainly a highly challenging task due to following reasons. It is very difficult to define the normal behavior or a normal region. The difficulties are as under.

- The definition of an outlier is highly dependent on the domain and the application.
- It is difficult to evaluate or compare outlier detection algorithms, as there is no ground truth available.
- It is very hard to enumerate every possible normal behavior in a dataset for any particular application. The reason is that the border between normal and abnormal objects is usually a gray area. The boundary will be imprecise, such

that at times an observation lying close to the border as outlier could actually be a normal object and vice-versa.

- The choice of distance measure between objects and modeling the relationship among them are purely dependent on the type of application being used. For example, in a health care dataset, a small deviation can be termed as an outlier whereas in marketing analysis, larger fluctuations are identified as outliers.
- When outliers result from malicious action adaptation of such malicious adversary to make the outlier observations appear like normal is highly challenging.
- The notion of outliers differs for different application domains. This creates a difficulty in applying the technique developed in one domain to another domain. For example, in the medical domain a small difference in normal body temperature might be an outlier, while similar difference in the stock market domain might be considered as normal.
- Availability of labeled data for training process and validation of models carried out by outlier detection techniques.
- Noise in the data tends to be similar to the actual outliers and hence is difficult to distinguish and remove them from malicious outliers. Presence of noise in the data set distorts normal objects and obscures the distinction between normal data objects and outlier objects. Noise hides outlier objects, thus dropping the effectiveness of the outlier detection algorithm.

Due to the above challenges, the outlier detection problem, in its most general form, is difficult to comprehend. In fact, a large number of the current outlier detection procedures solve a definite problem formulation which is impelled by

different components, for example, the nature of the information, accessibility of named information, kind of anomalies to be recognized, and so forth. Often, these factors are determined by the application domain in which the outliers need to be detected.

1.5 OUTLIER DETECTION IN DATA MINING

With the advancement of data innovations, the quantity of databases, and also their measurement and multifaceted nature, grow rapidly, bringing about the need of automated analysis of enormous amount of heterogeneous organized data. For this reason, data mining frameworks are utilized. The main aim of these systems is to identify and reveal hidden dependencies in databases. The analysis results are then utilized for making a choice by a human or system, with the end goal that the nature of the choice made clearly relies upon the nature of the information mining. One of the basic problems of data mining is outlier detection. Outlier detection technique scans for items in the database that do not obey laws substantial for the major part of the information. The identification of an item as an outlier is influenced by different elements, a large portion of which are of interest in practical applications. For instance, an uncommon stream of system bundles, uncovered by analyzing the framework log, may be classified as an outlier, because it may be a virus attack or an attempt of an intrusion. The recognition of an object as an outlier might be an evidence that there appeared new readiness in information. For instance, a data mining framework can distinguish changes in the business sector sooner than a human expert. The outlier detection problem is similar to the classification problem. A specific feature of the former, is that the considerable greater majority of the

database objects being analyzed are not outliers. Moreover, in many cases, it is not a priori known what objects are outliers.

1.6 MOTIVATION OF OUTLIER DETECTION

Outlier detection also known as anomaly detection or deviation detection, is one of the fundamental tasks of data mining along with predictive modelling, cluster analysis and association analysis (Tan et al. 2006). Compared with these other three tasks, outlier detection is the closest to the initial motivation behind data mining, i.e., mining useful and interesting information from a large amount of data (Han and Kamber, 2006). Outlier detection has been widely researched in various disciplines such as statistics, data mining, machine learning, information theory, and spectral decomposition (Chandola et al. 2007). Also, it has been widely applied to numerous application domains such as fraud detection, network intrusion, performance analysis, weather prediction, etc (Chandola et al. 2007).

Despite numerous improvements in healthcare practice, the occurrence of medical errors remains a persistent and serious problem (Kohn LT et al. 2000). The urgency and the scope of the medical error problem have prompted the development of solutions to aid clinicians in eliminating such mistakes. Current computer tools for monitoring patients are primarily knowledge-based; the ability to monitor depends on the knowledge represented in the computer and extracted a priori from clinical experts. Unfortunately, these systems are time consuming to build and their clinical coverage is quite limited. Thus motivated us to investigate on outlier detection in health care domain in general and on diabetics in particular.

1.7 APPLICATIONS OF OUTLIER DETECTION

An outlier often contains useful information about abnormal characteristics of the systems and entities, which impact the data generation process. The recognition of such unusual characteristics provides useful application-specific insights. Some examples are as follows:

- Intrusion Detection Systems: In many host-based or networked computer systems, different kinds of data are collected about the operating system calls, network traffic, or other activity in the system. This data may show unusual behavior because of malicious activity. The detection of such activity is referred to as intrusion detection.
- Credit Card Fraud: Credit card fraud is quite prevalent, because of the ease with which sensitive information such as a credit card number may be compromised. This typically leads to unauthorized use of the credit card. In many cases, unauthorized use may show different patterns, such as a buying spree from geographically obscure locations. Such patterns can be used to detect outliers in credit card transaction data (Thornton et al. 2014).
- Interesting Sensor Events: Sensors are often used to track various environmental and location parameters in many real applications. The sudden changes in the underlying patterns may represent events of interest. Event detection is one of the primary motivating applications in the field of sensor networks.
- Medical Diagnosis: In many medical applications the data is collected from a variety of devices such as MRI scans, PET scans or ECG time-series. Unusual patterns in such data typically reflect disease conditions.

- Law Enforcement: Outlier detection finds numerous applications to law enforcement, especially in cases, where unusual patterns can only be discovered over time through multiple actions of an entity. Determining fraud in financial transactions, trading activity, or insurance claims typically requires the determination of unusual patterns in the data generated by the actions of the criminal entity.
- Earth Science: A significant amount of spatiotemporal data about weather patterns, climate changes, or land cover patterns are collected through a variety of mechanisms such as satellites or remote sensing. Anomalies in such data provide significant insights about hidden human or environmental trends, which may have caused such anomalies.

In all these applications, the data has a "normal" model, and anomalies are recognized as deviations from this normal model.

1.8 ORGANIZATION OF THE THESIS

Chapter 2 is organized as follows. Section 2.1 presents a general overview of data mining in outlier detection. Section 2.2 explains the perspectives of outlier detection methods. An overview of various application areas of outlier detection is discussed in Section 2.3. Section 2.4 provides a summary of the literature review done. Sections 2.5 and 2.6 lists the objectives and contributions of the research work.

In chapter 3, the design of the research work is proposed in section 3.1. Section 3.2 provides an overview of the various individual outlier detection methods used in the experimental approach. In Section 3.3, details about proposed meta outlier detection method is provided. Sections 3.4 provide details of various classification methods applied.

In chapter 4, the implementation details of the outlier detection techniques employed and the results of these techniques to handle outliers are discussed. Discussion is also made on the type of classifier that is employed. In section 4.1, discussion about the datasets used is carried out. Section 4.2 describes the various outlier detection methods employed. Section 4.3 discusses about the performance of individual outlier detection methods. Feature bagging, the meta outlier technique is discussed in 4.4. The performance of meta outlier detection technique is discussed in section 4.5. Chapter 5 serves as the conclusion to the thesis and describes future work involving outlier detection techniques.