Chapter 9
Data Mining Techniques for Outlier Detection

N N R Ranga Suri
C V Raman Nagar, India

M Narasimha Murty
Indian Institute of Sceince, India

G Athithan
C V Raman Nagar, India

ABSTRACT

Among the growing number of data mining techniques in various application areas, outlier detection has gained importance in recent times. Detecting the objects in a data set with unusual properties is important as such outlier objects often contain useful information on abnormal behavior of the system described by the data set. Outlier detection has been popularly used for detection of anomalies in computer networks, fraud detection and such applications. Though a number of research efforts address the problem of detecting outliers in data sets, there are still many challenges faced by the research community in terms of identifying a suitable technique for addressing specific applications of interest. These challenges are primarily due to the large volume of high dimensional data associated with most data mining applications and also due to the performance requirements. This chapter highlights some of the important research issues that determine the nature of the outlier detection algorithm required for a typical data mining application. The research issues discussed include the method of outlier detection, size and dimensionality of the data set, and nature of the target application. Thus this chapter attempts to cover the challenges and possible research directions along with a survey of various data mining techniques dealing with the outlier detection problem.

DOI: 10.4018/978-1-4666-2455-9.ch009

INTRODUCTION

The recent developments in the field of data mining have lead to the outlier detection process mature as one of the popular data mining tasks. Due to its significance in the data mining process, outlier detection is also known as outlier mining. Typically, outliers are data objects that are significantly different from the rest of the data. Outlier detection or outlier mining refers to the process of identifying such rare objects in a given data set. Although rare objects are known to be fewer in number, their significance is high compared to other objects, making their detection an important task. The general requirement of this task is to identify and remove the contaminating effect of the outlying objects on the data and as such to purify the data for further processing. More formally, the outlier detection problem can be defined as follows: given a set of data objects, find a specific number of objects that are considerably dissimilar, exceptional and inconsistent with respect to the remaining data (Han, 2000). A number of new techniques have been proposed recently in the field of data mining to solve this problem. This chapter mainly deals with these techniques for outlier detection and highlights their relative merits and demerits.

Many data mining algorithms try to minimize the influence of outliers or eliminate them all together. However, this could result in the loss of important hidden information since one person's noise could be another person's signal (Knorr, 2000). Thus, the outliers themselves may be of particular interest, as in the case of fraud detection, where they may indicate some fraudulent activity. Besides fraud detection, financial applications and niche marketing and network intrusion detection are other applications of outlier detection, making it an interesting and important data-mining task. Depending on the application domain, outlier detection has been variously referred to as novelty detection (Markou, 2003a), chance discovery (McBurney, 2003), or exception mining (Suzuki, 2000), etc. A related field of research is activity monitoring with the purpose of detecting illegal access. This task consists of monitoring an online data source in the search for unusual behavior (Fawcett, 1999).

Much of the research related to outlier detection has evolved in the context of anomaly detection. An anomaly is something that is different from normal behavior. Though anomalies are often considered as noise, they could be deemed as the early indicators of a possible major adverse effect. Thus, detection of anomalies is important in its own right and also due to the increasing number of applications like computer network intrusion detection (Chandola, 2009; Lazarevic, 2003), fraud detection, astronomical data analysis (Chaudhary, 2002), etc. In most of the cases, anomaly detection is intended to understand evolving new phenomena that is not seen in the past data. A standard method for detecting anomalies is to create a model of the normal data and compare the future observations against the model. However, as the definition of normality differs across various problem domains, the problem of anomaly detection turns out to be a more challenging and involved process. A generic computational approach is to look for outliers in the given data set. Some research efforts in this direction can be found in (Lazarevic, 2003; Sithirasenan, 2008), in the context of network intrusion detection.

There have been various definitions in the literature for outliers that were proposed in different research contexts. A popular one among them, given by (Hawkins, 1980), is to define an outlier as an observation that deviates so much from other observations as to arouse suspicion that it is generated by a different mechanism. The presence of an outlying object in a data set shows itself in some form or the other. For example, data objects P and Q in Figure 1(a) are outliers, which is obvious from a visual examination. However, in cases where the objects like P_1 and P_2 in Figure 1(b) are present in a data set, identifying them as outliers requires some extra effort. Also, depending