Chapter 14
Performance Enhancement of Outlier Removal Using Extreme Value Analysis-Based Mahalonobis Distance

Joy Christy A
School of Computing, SASTRA University (Deemed), India

Umamakeswari A
School of Computing, SASTRA University (Deemed), India

ABSTRACT

Outlier detection is a part of data analytics that helps users to find discrepancies in working machines by applying outlier detection algorithm on the captured data for every fixed interval. An outlier is a data point that exhibits different properties from other points due to some external or internal forces. These outliers can be detected by clustering the data points. To detect outliers, optimal clustering of data points is important. The problem that arises quite frequently in statistics is identification of groups or clusters of data within a population or sample. The most widely used procedure to identify clusters in a set of observations is k-means using Euclidean distance. Euclidean distance is not so efficient for finding anomaly in multivariate space. This chapter uses k-means algorithm with Mahalanobis distance metric to capture the variance structure of the clusters followed by the application of extreme value analysis (EVA) algorithm to detect the outliers for detecting rare items, events, or observations that raise suspicions from the majority of the data.

INTRODUCTION

Outlier detection is a part of data analytics that helps user to find discrepancies in working machine by applying outlier detection algorithm on the captured data for every fixed interval. An outlier is a data point that exhibits different properties from other points that are due to some external or internal forces.

DOI: 10.4018/978-1-7998-2491-6.ch014

These outliers can be detected by clustering the data points. To detect outliers, optimal clustering of data points is important. Problem, which arises quite frequently in statistics, is identification of groups or clusters of data within a population or sample. The most widely used procedure to identify clusters in a set of observations is K-Means using Euclidean distance. However, Euclidean distance is not so efficient for finding anomaly in multivariate space. To remedy this shortfall in the K-Means algorithm, Mahalanobis distance metric is used to capture the variance structure of the clusters that is followed by the application of Extreme Value Analysis (EVA) algorithm to detect the outliers. This method serves as a significant improvement over its competitors and will provide a useful tool for detecting rare items, events or observations which raise suspicions by differing significantly from the majority of the data.

In this Information era, it is believed that information leads to power and success (Alberts, 2003). Future of many companies and government organizations relies on the information what they have with them. With the improvement in the storage techniques, now it is possible to collect and store a tremendous volume of information. Organizations have been collecting an immeasurable data from simple text documents to more complex information such as Medical data, Satellite data, spatial data and multimedia data. Mining of these data, using sophisticated mathematical algorithms, provides much useful information regarding the probability of future events, unusual events that might be interesting or data errors that require further investigation. Data mining is the process of uncovering patterns and finding anomalies and relationships in large datasets that can be used to make predictions about future trends. The main purpose of data mining is extracting valuable information from available data. It is also popularly known as Knowledge Discovery in Databases (KDD) (Tembhurne, 2019) (Krochmal, 2018). Data Mining comprises of few steps starting from preliminary raw data collections to some form of identifying new knowledge It is an iterative process and uses the following steps such as Data cleaning, Data integration, Data selection, Data transformation, Data mining, Pattern evaluation and Knowledge Representation. Once the extracted information is offered to the user, the assessment measures can be improved and further refined to get more fitting results.

One of the important applications of data mining is outlier detection. Outlier detection is the process of detecting and subsequently excluding inappropriate data from the given set of data. An outlier is a piece of data that deviates drastically from the standard norm or average of the data set. Outlier detection has two-steps viz., Clustering and detecting deviated data among the clustered sets. Therefore, the process of grouping observations into cluster is a foremost problem in analyzing data sets. So far, the most widely used algorithm to identify clusters in a set of observations is K-Means. But, the main constraint of this algorithm is that it uses Euclidean distance metric, which is prone to noisy data and outliers, which in turn give a non-spherical cluster. Also, this distance suites well only for univariate datasets. Hence, this book chapter introduces the technique of Mahalanobis distance (MD) to detect an observation having an unusual pattern. The MD measures the relative distance between two variables with respect to the mean of the multivariate data. These calculated distance values are used by Extreme Value Analysis (EVA) algorithm to find outliers, and thereby, eliminating the need of deciding threshold value manually.

For detecting outliers in a set of observations, it is important to cluster the points accurately. Clusters are characterized by groups of data points which are in "close" proximity to one another. While it is much easier to visually detect clusters in univariate or bivariate data, the task becomes increasingly difficult as the dimensionality of the data increases. One of the largest used clustering algorithms is K-Means using Euclidean distance. Euclidean distance suffers from a scaling effect that describes a situation where the variability of one parameter masks the variability of another parameter and it happens when the measurement ranges or scales of two parameters are different; thus, makes it difficult to find the