Abstract

An outlier is defined as a data point which is very different from the rest of the data set based on some measure. Such point often contains useful information on abnormal behaviour of the system described by data. The detection of outliers has gained considerable interest in data mining, with the realization that outliers can be the key discovery to be made from very large databases. Outliers arise due to various reasons such as mechanical faults, changes in system behavior, fraudulent behavior, human error and instrument error. Indeed, for many applications the discovery of outliers leads to more interesting and useful results than the discovery of inliers. Detection of outliers can lead to identification of system faults so that administrator can take preventive measures before they escalate. Anomaly detection may enable detection of new attacks. It is becoming a hot issue in the data mining research and lots of literatures have been conducted on outlier mining for large datasets.

Outlier detection is an important anomaly detection approach. There are various approaches for outlier detection, namely, distance-based approach where objects that are at a considerable distance from any other cluster are considered outliers, density-based approach where the degree of "outlying" of an object is measured with regard to its surrounding neighbourhood and soft computing approach.

Keywords: outlier, anomaly, anomaly identification, outlier detection.