Research Journal of Applied Sciences, Engineering and Technology 10(9): 1021-1028, 2015
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: February 27, 2015 Accepted: March 25, 2015 Published: July 25, 2015

A Review of Outlier Prediction Techniques in Data Mining

[1]S. Kannan and [2]K. Somasundaram
[1]Department of Computer Science and Engineering, Karpagam University, Coimbatore, Tamilnadu 641021, India
[2]Department of CSE, Vel Tech High Tech Dr RR and Dr SR Engineering College, Avadi, Chennai-60062, India

Abstract: The main objective of this review is that to predict the outliers in data mining. In general, the data mining is a process of applying various techniques to extract useful patterns or models from the available data. It plays a vital role to choose, explore and model high dimensional data. Outlier detection refers a substantial research problem in the domain of data mining those objectives to uncover objects which exhibit significantly different, exceptional and inconsistent from rest of the data. The outlier potential sources can be noise and errors, events and malicious attack in the network. The main challenges involved in the outlier detection with high complexity, size and different types of datasets, are how to catch similar outliers as a group by using clustering-based approach. The outlier or noise available in the clustered data is accurately removed and retrieves an efficient high dimensional data. Nowadays, the classification and clustering techniques for outlier prediction are applied in various fields like bioinformatics, natural language processing, military application, geographical domains etc. This study surveys various data classification and data clustering techniques in order to identify the optimal techniques, which provides better outlier predicted data detection. Moreover, the comparison between the various classification and clustering techniques for outlier prediction are illustrated.

Keywords: Data classification, data clustering, data mining, high dimensional data, outlier detection

INTRODUCTION

Data mining is defined as the process that includes the extraction of interesting, interpretable, useful and novel information form of data. It has been used since several years in businesses, scientists and governments to sift through volumes of data like airline passenger records, census data and the supermarket scanner data that produces market research reports (Romero and Ventura, 2010). The major tasks involved in the data mining are.

Classification: A prediction learning function is discovered to classify a data item into multiple pre-defined classes.

Regression: The prediction learning function is discovered to map the data item to a real-value prediction variable.

Clustering: A common descriptive task seeks to identify a finite set of categories or clusters to describe the data.

Summarization: An additional descriptive task is involved in the process of finding a compact description for a subset of data.

Dependency modelling: A local model is found to describe significant dependencies between variables or between the values of a feature in the dataset.

Change and deviation detection: The most significant changes in the dataset are discovered.

The major objective of the data mining is further reduced into prediction and the classification. The prediction process predicts unknown or future values of interest by utilizing some variables or fields in the dataset (Koteeswaran and Janet, 2012). The classification process is used to find patterns by describing the data. In order to execute these processes data mining requires clustering and outlier analysis for reducing and identifying the useful dataset.

Outlier detection is also named as anomaly detection or deviation detection. It is one type of the fundamental tasks of data mining along with a predictive modelling, a cluster analysis and association analysis. These outlier detection is the nearest task to the initial aim behind the data mining. The detection of outliers has regained considerable interest in data mining with the realization of outliers from very large databases. The outlier detection is divided into different univariate methods (Williams *et al*., 2002). A Replicator Neural Network (RNN) is one type of

Corresponding Author: S. Kannan, Department of Computer Science and Engineering, Karpagam University, Coimbatore, Tamilnadu 641021, India, Tel.: +91422 647 1115