Abstract

Outlier detection is one of data mining functionalities that aims to find data that are different from other majority data. Although it has a different behavior with the other majority data, outliers often contain very useful information. There are many methods to detect outliers, but most are designed for numeric data and not appropriate for categorical data. Moreover, many algorithms take time to process increasing amounts of data. CBLOF (Cluster Based Local Outlier Factor) is a method for detecting outlier for categorical data based on clusters. A CBLOF value calculated for each data, is based on the condition that data are included in large clusters or small clusters, whether outlier data or not. Tests carried out with several scenarios to find out the accuracy based on the detection rate, false positive rate and false negative rate, influence the percentage of rare class on accuracy and influence the amount of data on processing time. CBLOF can detect outliers with relatively good accuracy, based on detection rate, false positive rate and false negative rate. In addition, the process is also faster because CBLOF will only read once the dataset for a data that is considered as an outlier or otherwise.

Keywords : outlier, cluster, categorical, CBLOF