OUTLIER PREDICTION
USING BORDER DATA OF CLUSTERS

By : Barry Nuqoba
Student Identity Number : 5107201009
Supervisor : Prof. Ir. Arif Djunaidy, M.Sc., Ph.D.

ABSTRACT

Most of the algorithm involving outlier published by researchers related with outlier detection, but not many discussed about outlier prediction. Outlier prediction is important to keep data validity. Conventional algorithm of outlier prediction has a weakness in a matter of efficiency, because it has to compare query data with the entire data in data set. A new concept involving solving set arised as a solution to the efficiency problem on outlier prediction. However, using solving set, although the prediction time becomes faster but the accuracy becomes worse.

In this research, a new outlier prediction algorithm that can sufficiently satisfy both the efficiency and accuracy on predicting new incoming data, is developed. This new algorithm is an innovation to the existing concept of solving set. In this regard, the solving set was defined as a subset of data set containing the border data of clusters and its centers to represent data set instead of defining the solving set as a subset of data set containing the top-n outlier to represent the data set as defined in the previous research work. The border data is detected using the BORDER algorithm, since it was proven capable of detecting the border data accurately and efficiently, while a hierarchical clustering algorithm is used to cluster the border data. The median values of the border data of each cluster are then computed to get the center of each cluster. In this new outlier prediction algorithm, which is called as APOTEK, the outliers are predicted by comparing the following two distances; these are, the distance of the query data to the center of cluster and the distance of the query data to the nearest border data.

After performing a series of experiments using both data sets with normal and uniform distributions, APOTEK was proven capable of improving the existing algorithm. On the aspect of prediction accuracy, APOTEK succeeded to make improvement for about 5% in compared to the Angiulli’s outlier prediction algorithm both for both data sets with normal and uniform distributions. Moreover, on the aspect of prediction speed, APOTEK succeeded to make improvement for about 40% and 73% for both data sets with normal and uniform distributions, respectively.

Keywords: outlier analysis, outlier prediction, cluster border point, solving set, data mining.