Abstract

In text categorization task, there usually exist a lot of outliers in the training data, for example, documents mislabeled or lying on the border between different categories, and documents that are out of the defined categories, etc. Therefore, outlier detection must be done to increase the performance of a text document.

One of the methods in outlier detection is Distance Based Outlier Detection, which is searching outlier(s) based on the distance between data in dataset. Distance Based Outlier Detection often use the k-Nearest Neighbor method, that have three meanings of outlier(s), that are: outliers are the examples for which there are fewer or more then p other examples within distance d, outlier are the top n examples whose distance to the kth Nearest Neighbor is greatest, and outlier are the top n examples whose average distance to the k Nearest Neighbor is greatest.

Outputs of this sistem are comparison of accuration or performance of the sistem in detecting outlier(s) with three definitions of outliers. Besides, this sistem also show the performance of categorization in a document, before and after outlier(s) are eliminated.

Keywords: *data mining, outlier detection, k- nearest neigbhor*