Abstract

Finding outlier or anomaly in large dataset is an important problem in areas such as electronic commerce, credit card fraud, and even the analysis of performance statistics of professional athletes. The identification of outliers can lead to the discovery of truly unexpected knowledge. LOF (*local outlier factor*) is a classical density based outlier detection method, which is successfully used for detecting outliers in fields of machine learning, pattern recognition, and data mining. LOF has two steps. In the first step, it calculates k-nearest neighbors of each data point. In the next step, it assigns a score to each point called *local outlier factor (lof)* using k-nearest neighbors information. However, LOF computes a large number of distance computations to calculate k-nearest neighbors of each point in the dataset. Therefore, it cannot be applied to large datasets. In this report, an approach called TI-LOF is proposed to reduce the number of distance computations in classical LOF method. TI-LOF utilizes triangle inequality based indexing scheme to find k-nearest neighbors of each point. In the same line of classical LOF, TI-LOF assigns a score to each point using earlier computed information. Proposed approach performs significantly less number of distance computations compared to the classical LOF. We perform experiments with synthetic and real world datasets to show the effectiveness of our proposed method in large datasets.

Keywords: Large dataset, LOF(local outlier factor), TI-LOF.