Local Outlier Detection in Big Data by a Density-based Method

By: Sayyed-Ahmad Naghavi-Nozad

Abstract

According to the fact that nowadays, the volume of data collected and stored in different ways is increasing so rapidly, thus commonly used software methods for processing and managing this very large amount of data are not efficient and, as such, this type of data is categorized as Big Data. The huge size and high dimensionality of Big Data make it more necessary to use novel techniques and technologies that can deal with such amount of data. One of the major challenges in the field of Big Data is Outlier detection. Detecting and removing outliers from data could be considered as a preprocessing step for data analysis, which attempts to find those rare data points that are behaving in an unacceptable manner than the rest of the common data points. In this thesis, we are trying to provide a method for detecting local outliers in Big Data, which is based on scalable density-based clustering. Considering the fact that a Big Data does not have the capability to be fit into the RAM at once, so we have to process it chunk by chunk, as each chunk can be both inserted and processed in the main memory at the same time. Then, for each chunk, we update the information of the clustering model. Throughout the clustering procedure, our effort is to not let the outliers play any role in formation and updating clusters. At the end of scalable clustering, we obtain the structure of the final clusters. Subsequently, using a suitable criterion, we will give each data point an outlierness score. Our evaluations on real-life and synthetic datasets exhibit that our proposed method has a low linear time complexity and is effective in comparison with those traditional methods that need to observe the whole data at once, not chunk by chunk, and also is efficient in severe conditions.

***Keywords*—** Big Data, Outlier detection, Scalable, Density-based.

Link to Master's Thesis