INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI
SHORT ABSTRACT OF THESIS

Name of the Student	:	Rajendra Pamula
Roll Number	:	05610105
Programme of Study	:	Ph.D.
Thesis Title	:	Data Pruning Based Outlier Detection
Name of Thesis Supervisor(s)	:	Prof Jatindra Kr Deka & Prof Sukumar Nandi
Thesis Submitted to the Department/ Center	:	Department of Computer Science & Engineering
Date of completion of Thesis Viva-Voce Exam	:	20-11-2015
Key words for description of Thesis Work	:	outlier, cluster, centriod, radius, pruning, distance, summarization, correlation

SHORT ABSTRACT

Due to the advancement of the data storage and processing capabilities of computers, most of the real life applications are shifted to digital domains and many of them are data intensive. In general, most of the applications deal with similar type of data items, but due to variety of reasons some data points are present in the data set which are deviating from the normal behaviors of common data points. Such type of data points are referred as outliers and in general the number of outliers in a data set is less in number. Identifying the outliers from a reasonably big data set is a challenging task. Several methods have been proposed in the literature to identify the outliers, but most of the methods are computation intensive. Due to the diverse nature of data sets, a particular outlier detection method may not be effective for all types of data set. The main focus of this work is to develop algorithms for outlier detection with an emphasis to reduce the number of computations. The number of computations can be reduced if the data set is reduced by removing some data points which are obviously not outliers. The number of computations again depends on the number of attributes of data points. While detecting outliers it may be possible to work with less number of attributes by considering only one attributes from a set of similar or correlated attributes. The objective of this work is to reduce the number of computations while detecting outliers and study the suitability of the method for a particular class of data set. Our methods are based on the clustering techniques and divide the whole data set into several clusters at the beginning. Depending on the nature of the clusters we propose methods to reduce the size of the data sets, and then apply outliers detection method to find the outliers. We propose three methods based on the characteristics of the clusters to identify the clusters that may not contain outliers and such clusters are pruned from the data set. We also propose a method to identify the inlier points from each cluster and prune those points from clusters. We use the principle of data summarization and propose a method that involves both cluster pruning and point pruning. For high dimensional data set, we propose a method that involves attributes pruning to reduce the number of computations while detecting outliers. Once we perform the pruning step, a reduced data set is resulted and then outlier detection techniques are used to detect the outliers. For each method we demonstrate the effectiveness of our proposed methods by performing experiments.