ABSTRACT

Data mining for knowledge discovery is one of the important area of research in the present era of information world. Though, there are several techniques existing, there is a need and demand for exploration of new and better techniques. Outlier detection has received significant attention in many applications, such as network intrusion detection and credit card fraud detection. Much of the existing research focuses on numerical data and cannot be directly applied to categorical and mixed type of data sets where there is little sense in ordering the data and calculating distances among data points. Furthermore, a number of the current outlier detection methods require quadratic time and memory space to generate the combinations of item sets with respect to the dataset size and usually need multiple scans of the data. These requirements are undesirable when the data size is large in records, large in Attributes, large in values and distributed over multiple geographically locations. In this thesis, we focus and evaluate, experimentally, a few representative current outlier detection approaches for both categorical and mixed attribute datasets. In addition, we introduce two simple, scalable and efficient outlier detection algorithms those have the advantages of discovering outliers in categorical and mixed attribute datasets by performing a single scan of the dataset. This newly introduced outlier detection algorithms are compared with the existing approaches like
Attribute Value Frequency (AVF), Frequent Pattern Outlier Factor (FPOF) and Fast Distributed Outlier Detection (FDOD) methods and aforementioned outlier detection strategies. The conclusion from this comparison is that the simple outlier detection algorithms that we introduce are more accurate than the existing strategies.

Similarly the classifiers built by proposed systems are more accurate than existing systems. Frequent Pattern Outlier Factor (FPOF) and Fast Distributed Outlier Detection (FDOD) methods are derived based on frequent patterns which take more time and complexity of these methods are more. Even the existing method Attribute Value Frequency (AVF) takes less complexity, it finds similar scores for different objects. Proposed Mudimbi-Aliseri-Dirisinapu (MAD) method finds different scores for different records exactly. Comparison of MAD and the other existing systems shows that MAD is good for categorical dataset. Another proposed method Multi Attribute Outlier Factor (MAOF) shows better performance in finding outliers in mixed attribute datasets. This proposed method takes only one scan of the entire database for mixed attribute datasets and works better than existing systems.