ABSTRACT

Nowadays, Peer to Peer (P2P) network traffic experiences a major portion of internet traffic, and becomes the elementary data source while considering high volume data streams. Among the different types of internet applications in today’s scenario P2P network is one of the high traffic network characterized by, which allows a large number of users to establish communication with each other which in turn directly access and download files from the peers machine, and share the resources which are considered to be inevitable. The basic requirement in the internet traffic is the classification of network to be performed for high volume data streams in P2P network. In this research work, the use of clustering techniques by dividing into hierarchical and partition clustering to identify interesting traffic patterns based on the volume of data from network traffic data in an efficient manner is categorized.

Earlier models dealt with clustering of data stream which handles network traffic that comprises of numerical data set by way of using K-means algorithm whereas this proposed model works well for both numerical and categorical data set. This research work has extended from numerical data set to categorical data set using K-modes partition cluster algorithm for high volume data streams. The failure of traditional model demands to design adaptive hierarchical partition cluster approach in P2P network. A structural architecture to deal with both types of attributes including numerical and
categorical attributes by means of two-stage hierarchical clustering algorithm is being developed and presented in this thesis.

Cluster validity of hierarchical and partition clustering algorithm for both types of data sets such as numerical and categorical data mainly depends upon the selection of initial data points in order to instigate the clustering process. The traditional way of clustering is to choose the initial data points randomly. As a result of this, clustering results cannot be generated and repeated in a consistent manner. In this proposed work, a methodology to figure out the initial data points for K-mode clustering algorithm to cluster both numerical and categorical data sets is presented.

The objective of this research work is to observe the internet traffic characteristics. Based on the observation, different streams of data can be clustered according to the nature of data. This work presents a new approach using clustering techniques and in particular two stages, an adaptive architecture for classifying traffic in peer to peer network. In the first stage, traffic streams are divided into two types based on the stream of data arriving in network. In the second stage, if the incoming traffic stream is normal, it is characterized by hierarchical clustering approach and in case of high volume of data streams, it is characterized into partition clustering approach.

The experimental results are obtained with analysis made based on the network traffic data extracted from Ednet-Internet Service Provider (ISP) traffic data repository logs. Initially, only half of the data from the repository log of Ednet-ISP server are evaluated. At uniform intervals, Ednet-ISP
gradually introduced the distribution of network traffic across its peer nodes with effective load balance of their network server for large volume of data stream. Experimentally it was found that by adopting this method, Adaptive Hierarchical and Partition Clustering (AHPC) approach on Ednet-ISP traffic data repository logs, data streams were characterized between Hierarchical cluster and Partition Cluster based on the nature of streams of data. The present work demonstrates the cluster validity in terms of purity, F-measure and entropy to cluster traffic data streams based on the nature of the volume of data stream in P2P network in comparison to the previous work on clustering of data streams. Empirical results confirm with the superiority of the proposed model Adaptive Hierarchical and Partition Clustering using K-Modes algorithm. Results showed that the proposed model AHPC using K-Modes algorithm improved the cluster validity in terms of purity, F-measure and entropy, clustered high volume of data streams that occurred in network traffic in P2P environment.