ABSTRACT

Data clustering is an unsupervised learning method which aims at creating groups of objects or clusters in such a way that objects in the same group are similar to one another and are dissimilar to the objects of other groups. Clustering is one of the methods that can be used on large datasets for knowledge discovery. Several attempts have been made in the past to study the behavior of the existing system of $k$ clusters when a new point enters the system. This research work attempts to develop estimation models on these lines to forecast the occurrence of Inter Cluster Data Migration (ICDM) on data ingress.

The term Inter Cluster Data Migration (ICDM) denotes the movement of data points between the clusters which is initiated by the arrival of a new data point into the existing clustering system. The objective of this research is to develop estimation models to forecast the occurrence of ICDM on data ingress. A distance threshold estimation model attempts to find distance threshold $D$ for the new points. If the distance of the new data point to the center of its closest cluster (also known as reference cluster) is below the value $D$, there is no occurrence of ICDM. There is an occurrence of ICDM if the distance of the new data point is greater than or equal to $D$. The probability of occurrence of ICDM is also calculated based on the distance of the incoming data point to the center of its reference cluster. An estimation model is built to identify the sectors of the reference cluster, from where the data points move to its nearest cluster. Another estimation model to count the number of data points that moves out of the reference cluster to its nearest cluster is also developed.

Real datasets from UCI data repository and multidimensional synthetic datasets are used to generate compact and well-separated cluster partitions. Error percent is used as a performance measure. The estimated values are compared with observed experimental results. A comparison of the results shows a lower error percent which indicates the efficiency of these estimation models. Further proposals on future research directions are also enumerated.