6. CONCLUSIONS

Clustering algorithms mine the static database and generate a set of patterns in the form of clusters with high cohesion and separation. New data may be added periodically either on a daily or weekly basis to the existing databases which may grow dynamically. So the patterns extracted from the older snapshot of database become obsolete for dynamically growing databases. This problem is handled by repeating the process of clustering on the entire database whenever a significant set of data items are added. The process of re-running the clustering algorithm on the entire dataset is inefficient and time-consuming. Thus most of the clustering algorithms are not suitable for incremental databases.

Conventional algorithms aim at formation of clusters with an inherent characteristic of nearly uniform distribution within a cluster. However, applications like growing patterns of urban habitats and growing memberships in communities in social networking, protection mechanism against infections due to injuries in a human body etc. are better modeled as clusters with non- uniformly distributed entities. This research work has investigated the applicability of existing clustering algorithms for such cluster formation and incremental maintenance. It was found that the proximity metrics used by conventional clustering algorithms have limited capability in determining the membership of new entities into non- uniformly distributed clusters as the shape / dispersion of such clusters is irregular due to combination of dense and sparse regions constituting such clusters.

6.1 Significance of the Research Work

This research work contributes by devising a new proximity metric namely Inverse Proximity Estimate (IPE) specifically suitable for non- uniformly distributed clusters by considering the spread of entities within a cluster in addition to the cluster prototype. An incremental clustering algorithm called Cluster Feature-Based Incremental Clustering Approach for numerical data (CFICA) which makes use of Inverse Proximity Estimate to handle entities described in terms of only numerical attributes was developed and evaluated. Cluster Feature while being compact includes all essential information required for maintenance and expansion of clusters. Thus CFICA avoids redundant processing which is the essential feature of an incremental algorithm. The performance of this algorithm, in terms of purity is compared with the state of art incremental clustering algorithm namely BIRCH on different bench mark datasets. The results were presented in chapter 5 and are found to be consistently better for the proposed CFICA algorithm.

Since datasets with mixed types of attributes are common in many real life data mining applications, clustering such mixed data involves the combination of a good distance measure to adequately capture similarities between mixed data objects and an efficient clustering algorithm to perform clustering effectively. Therefore the Inverse Proximity Estimate (IPE) devised previously was extended to deal with mixed distances. Mixed distance measures the proximity between two data points by combining the distance estimated in terms of numeric attributes & dissimilarity estimated in terms of categorical attributes giving due importance to weightage. An information theoretic approach to estimate the Dissimilarity (DS) between two data points based on their categorical attributes has been designed. The specificity of an attribute-value pair is estimated to discriminate the clusters and the most prominent attribute value pairs were selected to represent the Weight vector. Thus the Weight vector represents the cluster prototype from the perspective of categorical attributes while the mean vector represents the prototype in the perspective of numerical attributes. Another algorithm named Cluster Feature-Based Incremental Clustering Approach to Mixed Data (M-CFICA) was devised as an extension of CFICA to handle entities described in terms of all types of attributes.

The performance of M-CFICA is compared with K-PROTOTYPES algorithm which is the most preferred algorithm to handle mixed datasets and has been evaluated on bench mark datasets as well as hypothetical datasets. The results were presented in chapter 5 and were found to be on par with K-PROTOTYPES algorithm for census income dataset and better for hypothetical datasets. The comparative statement in support of the authors claim was presented.

6.2 Future Extensions

The proposed incremental clustering algorithms can be extended to handle scalability problem for dealing with very large databases which are non-memory resident. The database may be processed incrementally chunk by chunk, the size of the chunk being decided based on the availability of RAM.

Both CFICA and M-CFICA can be extended to deal with data streams like time-series data for trend analysis to capture concept - drift as the time progresses. The Inverse Proximity Estimate (IPE) along with the growth rate of newly formed very small clusters can be used to estimate the outlier score of data points for detecting the outliers in the context of dynamically growing datasets with non-uniformly distributed clusters.

This research work may be extended to explore the customer Relationship Management (CRM) domain for modeling the changing interests of the customers with respect to changing lifestyles through generations and urbanization.