Synopsis

1. Introduction

With the recent advances of data generation and acquisition systems and the success of several projects such as Human Genome, a large number of databases, especially in biological field is now available worldwide. The growing rate of such databases is also exponential. There is a need to explore and analyze such massive data to infer some inherent information. Clustering has been recognized as one of the widely used data mining techniques which is essential for data analysis to reveal natural structures and to identify interesting patterns in the underlying data. The main goal of clustering analysis is to partition a given set of objects (data) into homogeneous groups based on given features, such that objects within a group are more similar to each other and more different from objects that are in other groups [1], [2]. Clustering has become one of the most popular classification techniques and has been applied in a wide variety of fields that ranges from machine learning, artificial intelligence, pattern recognition, web mining, spatial database analysis, textual document collection, image segmentation, genetics and biology to remote sensing and economics [2], [3], [4], [5], [6], [7], [8], [9]. Indeed, clustering the complex biological data has been paid enormous attention over the recent years and several algorithms [10], [11], [12], [13], [14], [15], [16] have been developed for the same.

The major criterion that differentiates the clustering algorithms is the selection of suitable similarity measure. Based on the similarity measure and nature of the given data the clustering methods can be categorized into various types, such as, hierarchical [17], partitional [18], density based [19] and grid-based [20] clustering. In spite of various types, the main goal of any clustering algorithm is to maximize both the homogeneity within each cluster and the heterogeneity among different clusters.

In the last decade, significant amount of research works have been carried out on cluster analysis and a large number of algorithms have been developed, reviews of which
can be found in [8], [9]. However, there is no universal clustering technique that can deal with all clustering problems due to arbitrary shapes, sizes, varied densities and high dimensionality, particularly for biological data [21], [22]. Therefore, designing of new clustering algorithms has become an emerging area of research. Recently, much attention has been paid to develop various clustering algorithms based on neighborhood graphs such as minimum spanning tree (MST), Voronoi diagram and $kd$-trees. In this thesis, we mainly report on the hierarchical, partitional, density-based and graph-based clustering algorithms which are developed using such neighborhood graphs.

The graph-based clustering techniques [23] have widely been applied due to their efficiency in clustering high-dimensional complex data. Most of the graph-based clustering algorithms are able to deal with the noisy data points. In addition to the graph-based methods, the density and grid-based algorithms are also known to be very effective in clustering arbitrary data of various dimensions. The density-based methods have extensively used to form the clusters of varied densities where as the grid-based methods are well-known in producing the clusters with less computational cost. With this context, a large number of clustering algorithms have been developed, a brief review of which is as follows.

Zahn [24] presented a graph-based clustering algorithm using the minimum spanning tree (MST). This method eliminates the inconsistent edges depending on the given limit on the weights of the edges in the MST. Wang et al. [25] developed a novel divide-and-conquer algorithm for MST-based clustering by using an efficient implementation of the cut and the cycle property of the MST. Zhong et al. [26] designed an efficient hierarchical clustering algorithm based on the MST and MST-based graph which have been used for the split-and-merge process. Yan and Weibel [27] proposed an algorithm for point cluster generation using the Voronoi diagram [28]. Kao et al. [29] described an efficient algorithm to cluster the uncertain data using the Voronoi diagram and the $R$-tree index [30]. Redmond and Heneghan [31] proposed a $kd$-tree based algorithm that uses the density information of the leaf buckets to locate the initial cluster centers. Kanungo
et al. [32] designed a novel initialization scheme for $K$-means clustering using $kd$-tree. A fast hybrid density based clustering method called Rough-DBSCAN [33] has proposed to cluster the large data sets. Ren et al. [34] used Mahalanobis distance metric to propose a new density-based clustering method called DBCAMM. Yue et al. [35] developed a general grid clustering approach (GGCA) based on the divide-and-conquer strategy of hierarchical clustering. Pilevar et al. [36] presented a novel grid-based technique called GCHL to cluster the multi-dimensional large spatial data.

Although, most of the above clustering algorithms are efficient, they are vulnerable in some situations due to various characteristics of the complex real world data, especially in biological field. In this thesis, we propose various new or improved clustering algorithms using minimum spanning tree, Voronoi diagram, $kd$-tree, DBSCAN and grid-based approaches. The experimental results demonstrate that the proposed algorithms are able to cluster both the complex synthetic and biological data efficiently. A description of our research contribution is as follows.

2. Research Contribution

We first develop two hierarchical clustering algorithms using minimum spanning tree. Then we propose two novel clustering approaches using Voronoi diagram, a well-known geometrical structure for solving the nearest neighbor problems. Next, we present two partitional clustering algorithms which are based on the $kd$-tree that removes the drawbacks of $K$-means clustering. A density-based clustering algorithm is also proposed to improve the computational cost of DBSCAN. Finally, we propose, a grid-based clustering algorithm. The proposed algorithms are briefly discussed as follows.

2.1 Minimum Spanning Tree Based Clustering

The minimum spanning tree (MST) based clustering algorithms [24] are well-known to be capable of producing clusters of arbitrary (complex) shapes among all the graph-based
methods. Here, we propose two novel MST based clustering algorithms depending on the variation between the edge weights. In the first algorithm, the concept of coefficient of variation [37] is used to find the appropriate number of edges to be eliminated from the MST. The second method uses the dynamic validity index [38] which is computed for \( k + 1 \) number of times to find the optimal case of MST-based clustering, where \( k \) is the number of clusters. Both the algorithms are experimented on a wide variety of synthetic as well as biological data including Iris, Spect-Heart, Wine, Ecoli, Statlog Heart, Pima-India-Diabetes, Soybean, Breast Tissue, Yeast Cell-Cycle, Sporulation, Lymphoma, Diauxic and Fibroblasts [21], [22]. The results are compared with several existing clustering techniques namely, \( K \)-means [39], SFMST [40], SC [41], MSDR [42], IMST [43], SAM [26] and MinClue [44] which shows the effectiveness of the proposed algorithms in terms of dynamic validity index [38] and computational cost.

2.2 Voronoi Diagram Based Clustering

The Voronoi diagram [45] is a versatile geometric structure that has been used to design quite a few clustering algorithms [27], [29], [46] recently. We propose here, three Voronoi diagram based algorithms that exploit the properties of the Voronoi edges, circles and vertices. In the first algorithm, we produce the clusters using the Voronoi edges by resolving the problems of redundancy and infinite edges. The second approach uses a function defined by the radii of the Voronoi circles to merge the sub-clusters. Similarly, the third algorithm produces desired clusters by reconstructing the Voronoi diagram. The algorithms are tested on various synthetic as well as biological data. The clustering results which are evaluated using the Dunn validity index [47] depict that the proposed algorithm consistently outperforms the existing methods \( K \)-means [39], FCM [48], SC [41], CTVN [49] and VDAC [46].
2.3 Enhanced K-means Clustering

*K*-means [39] algorithm is one of the most popular iterative partitioning clustering algorithms as it is simple and robust for large data sets. However, *K*-means has few limitations. Firstly, it does not produce unique results because of the random selection of initial cluster centers. Secondly, it requires the users to input the number of clusters in advance which may not be feasible for many real world data sets. In this thesis, we contribute two algorithms to boost the *K*-means clustering as follows.

To solve the problem of random initialization of the cluster centers, we use the concept of *kd*-tree. The algorithm is shown to be insensitive for the outlier points during the initialization of cluster centers. Then, we exploit the joint distance function (JDF) [50] to handle the second problem of *K*-means. The turning point in the trajectory of the JDF supplies the number of clusters of the given data. The experimental analysis of a large number of synthetic and biological data illustrates the efficiency of the proposed scheme over several existing algorithms such as *K*-means [39], *IK*-means [51], CCIA [52] and FCM [48]. For the sake of comparison, we use Intra-Inter ratio validity index [53] that describes the compactness and the separateness of the underlying clusters.

2.4 Modified DBSCAN Clustering

DBSCAN [54] is one of the standard density based clustering methods that identifies the clusters of complex shapes by differentiating the densely populated regions. However, the notion of DBSCAN is directed by two parameters, namely, neighborhood and the minimum size. The better results are possible if and only if these two parameters are chosen efficiently. In this method, all the given points are visited multiple times to find the distances which increase the computational cost unnecessarily.

We propose a novel modified DBSCAN algorithm using the partitional approach to overcome the above challenges. Here, we use the prototypes generated by a squared error
clustering method such as $K$-means. The proposed modified DBSCAN algorithm is able to produce the clusters of various synthetic data of arbitrary shapes and dimensions. It is shown to be faster than the various existing DBSCAN-based algorithms namely, I-DBSCANS [55], DBCAMM [34], VDBSCAN [56] and KFWDBSCAN [57] in case of several gene expression data. The proposed method is also efficient in clustering different real world data as evaluated by the error rate [52].

2.5 Grid-Based Clustering

Grid-based clustering methods [58] deal with the cells/grids of the segmented data space or grid structure of the given data objects. The grid clustering algorithms do not require any input parameters and usually have linear time complexity. However, the main problem of the grid-based methods lies in finding the optimal size of the grid in order to produce the desired clusters. Several attempts [59], [60], [61] have made to determine the optimal size of the grids. Although, these methods are close to compute optimal grid size, a feasible solution is yet due in this direction.

Here, we make an effort to design a new grid-based clustering algorithm. This scheme finds the optimal grid-size using the novel concept of boundary grids. The proposed algorithm is insensitive to the outlier points by exploiting the local outlier factor ($LOF$) [62]. We perform wide range of experiments on diverse data sets of both synthetic as well as biological. The comparison results demonstrate the efficiency of the proposed clustering approach over the algorithms $K$-means [39], IGDCA [61] and GGCA [35] using the normalized information gain [31].
References


<table>
<thead>
<tr>
<th>Reference</th>
<th>Details</th>
</tr>
</thead>
</table>
References


References


References


