2.1 Introduction

In this chapter, we present a comprehensive review of several existing clustering algorithms related to our research work. First we focus on the existing MST-based clustering techniques and describe the various methodologies to handle different problems faced by them. Next, we pay the attention on detailed study of the Voronoi diagram based clustering algorithms. We then provide a report on the initialization and automation problems of the $K$-means. After this, we present an extensive survey on density-based clustering techniques and their associated problems with their solutions. Finally, we furnish the information of several grid-based clustering algorithms with their merits and demerits.

2.2 MST-Based Clustering Algorithms

Among various graph-based methods following the hierarchical model, minimum spanning tree (MST) based clustering is very attractive for its intuitive and effective data representation. In this approach, minimum spanning tree of the given data points is constructed and then a criterion is proposed to remove the longer edges (inconsistent edges) that result into different clusters.

MST-based clustering has been initialed by Zahn [27]. Then, a large number of algorithms [46], [47], [48], [49], [50], [51], [52], [53], [54], [55] have been developed. Wang et al. [56] proposed a divide-and-conquer approach for MST-based clustering by using the cut and the cycle property of the MST. The algorithm runs in $O (n \log n)$ time. Zhong et al. [57] developed a two-round MST-based clustering technique which also discovers the number of clusters. This method mainly deals
with separated and touching cluster problems. It is not robust for the outliers and overlapped clusters. Grygorash et al. [58] proposed a clustering method called MSDR that does not require any input parameters. The method runs in $O(n^2)$ time. It is unable to detect the overlapped clusters and outliers. Paivinen [59] proposed a scale-free minimum spanning tree-based clustering algorithm. This algorithm requires high computation time and unable to detect the outliers. He and Chen [60] developed an approach named “MinClue” to automate the threshold value for eliminating the inconsistent edges of the MST using $k$-nearest-neighbor graph. The method mainly deals with the single linkage clusters and runs in $O(kn \log kn)$ time. Maria et al. [61] presented an algorithm using the comparison between the $k$ nearest neighbors graph and the MST. Chowdhury et al. [62] developed a clustering algorithm based on the MST and Bayes classifier [63]. This method requires large number of computations for high dimensional data and limited to the bounded object classes only. Zhong et al. [64] designed a hierarchical clustering algorithm using the split-and-merge process in MST-based graph. This method cannot be applicable for large scale data. Forina et al. [65] designed an MST-based algorithm called OETICS which is similar to OPTICS [66]. The OETICS is not very flexible compared to that of OPTICS. It is also very sensitive to the local differences in the edge lengths. Xu et al. [67] developed three algorithms for gene expression data out of which two algorithms are guaranteed for the global optimality, but not scalable and unable to detect the outliers. Victor et al. [68] contributed with an MST-based which gradually generates $k$ clusters with a segment for each of the cluster. This algorithm runs in quadratic time. Laszlo et al. [69] developed a novel clustering approach to partition the MST by adding a constraint on the minimum group size. This algorithm is motivated by micro aggregation. It can efficiently be applied on large data sets.

### 2.3 Voronoi Diagram Based Clustering Algorithms

The Voronoi diagram [34], [70] is another well-known neighborhood graph that has attracted the researchers due to its efficient geometrical structure. The Voronoi
2.3 Voronoi Diagram Based Clustering Algorithms

diagram based clustering algorithms usually run in $O(n \log n)$ time [34]. There are some challenges in dealing with the Voronoi diagram such as the redundancy and infinite edges. Despite these drawbacks, we usually prefer Voronoi diagram based techniques over the MST-based methods which have quadratic computational complexity. A few algorithms have developed using this approach. A concise tour into Voronoi diagram based approaches is as follows.

Yan and Weibel [71] proposed a Voronoi diagram based algorithm for point cluster generation. This method is mostly applicable in thematic dot maps and cartographic maps. Kao et al. [72] developed an algorithm to cluster the uncertain data with the help of Voronoi diagram and a probability density function (PDF). Kao et al. have used Voronoi diagram to propose pruning techniques which are helpful during the cluster formation. Jana et al. [73] proposed a novel clustering algorithm using the Voronoi diagram and the cluster density measure defined by Jiang et al. [74]. This is not applicable for finding the clusters of arbitrary shapes. Koivistoinen et al. [75] designed an agglomerative algorithm for autonomous clustering. Here, the Voronoi diagram has been used to access the density information of the input samples. Bishnu and Bhattacherjee [76] developed a method called CTVN with the help of $K$-means algorithm and Voronoi diagram. This scheme depends on two parameters $k$ and a threshold value. It is also unable to deal with scalability issues. Lee and Yang [77] have proposed a fast and efficient hybrid clustering algorithm that uses topological information gained by the Voronoi diagram. Liu et al. [78] designed a density-based algorithm named VDC that is designed with the help of Semi Delaunay Diagram. The VDC is not scalable and unable to locate the noise. Wu et al. [79] presented an algorithm with the help of Voronoi diagram and Vector Quantization. Kao et al. [80] proposed a new algorithm to cluster the uncertain data using the Voronoi diagram and $R$-tree index [81]. This method mainly deals with the problems faced by $UK$-means [82]. Liu et al. [83] proposed a spatial clustering algorithm known as TRICLUST with the help of Delaunay triangulation. The time complexity of this algorithm is $O(n \log n)$. 
2.4 Enhanced K-means Clustering Algorithms

Although hierarchical clustering techniques are efficient for arbitrary data, partitional approaches are also researched extensively due to their simplicity. K-means is the most popular clustering technique of this model developed by MacQueen [26] in 1967. However, as discussed in previous chapter, it is sensitive to the random selection of initial cluster centers. In addition to that, a prior knowledge of the number of clusters is necessity to input to K-means. Many researches proposed various methods [84], [85] to overcome these problems.

Kanungo et al. [86] proposed a novel initialization method for K-means using kd-tree. This scheme does not pass information from one stage to its next. Du et al. [87] developed an initialization scheme for K-means clustering called PK-means to cluster the gene expression data. The convergence rate of this technique is fast and the computational load is less. A novel clustering algorithm called modified filtering algorithm (MFA) has been proposed in [88]. It is the improvement of the algorithm in [86]. A fast K-means clustering algorithm named FKMCUCD was proposed in [89] using cluster center displacement. This method is significant for high-dimensional large data. Zalik [90] proposed an efficient algorithm named K’-means to enhance the K-means algorithm by exploiting a cost function. This scheme fails when the clusters are of various shapes such as elliptical. Redmond et al. [91] proposed a novel seed selection algorithm using kd-tree [30]. This scheme is unable to deal with the noise. Cao et al. [92] proposed an algorithm by defining the cohesion degree of the neighborhood of a given point and the coupling degree between neighborhoods of the points. This algorithm has quadratic time complexity. Khan et al. [93] designed an algorithm called CCIA. This method first develops k’(>k) cluster centers from which the desired k centers are chosen. Lu et al. [94] contributed with a hierarchical initialization approach in which the clustering problem has treated as a weighted clustering problem. A genetic clustering algorithm named GAGR [95] has been proposed to cluster the genome data using K-means. It uses the genetic algorithm
with gene rearrangement process. Ahmad et al. [96] proposed an enhanced $K$-means clustering algorithm for mixed numeric and categorical data based on co-occurrence of the values. An algorithm called KGA [97] was proposed using the genetic algorithm. This method may not produce fine results whenever the number of clusters is unknown. An improved version of $K$-means called $K^*$-means has been developed in [98]. It is unable to deal with the noisy data. Likas et al. [99] proposed a global $K$-means clustering algorithm in which the clusters are formed using a global search procedure. A recursive method is proposed by Duda and Hart [100]. Milligan [101] developed an enhanced algorithm based on Ward’s hierarchical method [102] that helps in finding the initial cluster centers. The algorithm proposed by Fisher [103] generates good seeds by constructing initial hierarchical clustering based on [104]. Both Higgs et al., [105] and Snarey et al. [106] developed a method using MaxMin algorithm to choose a subset of the original database as initial cluster centers. Bradley et al., [107] formed the initial clusters based on the bilinear program. Tou and Gonzales [108] presented a method which entirely depends on the order of the points and the threshold value. Linde et al., [109] proposed a method based on Binary Splitting (BS). Here, the clusters quality depends on the selection of a random vector. Kaufman and Rousseeuw [110] developed a method based on the reduction in the Distortion. Babu and Murty [111] proposed a technique for the near optimal seed selection based on genetic programming. This is not robust for large data bases. Huang and Harris [112] projected a method called Direct Search Binary Splitting (DSBS) based on the Principal Component Analysis (PCA) and the vector of Linde et al., [109]. Thiesson et al., [113] designed an algorithm that depends on the mean value of the given data. Bradley and Fayyad [114] proposed an initialization approach for $K$-means using the Forgy’s method [115].

2.5 Improved DBSCAN Algorithms

The hierarchical and partitional clustering methods face difficulties to deal with the data that has clusters of variable densities and noise. Density-based approaches have
been applied in this direction efficiently. Ester et al. [29] has developed a method called Density-Based Spatial Clustering and Application with Noise (DBSCAN) to produce the clusters of such data. The crucial task of this scheme is to select the appropriate parameters so as to reduce the computational cost. An extensive research has been carried out in this direction and different algorithms [116], [117] have been proposed. A brief discussion of such techniques is as follows.

Bicici et al. [118] proposed a density-based clustering algorithm known as LSDBC that introduces the concept of local scaling to find out the density threshold. The time complexity of LSDBC is $O(n \log n)$. A top down approach has been presented in [119] with the help of the density information stored in index nodes of a multidimensional index. The method provides branch-and-bound pruning mechanism to prune unnecessary search. A fast hybrid density based clustering method called Rough-DBSCAN [120] was proposed using the leaders approach. Ren et al. [121] used Mahalanobis distance [122] to propose a new density-based method called DBCAMM which is motivated by the DBSCAN and few fuzzy clustering techniques [123]. This method has quadratic computational complexity. Tu et al. [124] presented a density-based hierarchical clustering technique to cluster the streaming data. This technique is based on an agglomerative clustering framework. Tran et al. [125] developed an algorithm to cluster the high-dimensional multivariate data. The algorithm called KNNCLUST has been designed by combining K-nearest-neighbor (KNN) and KNN-kernel density estimation techniques [126]. This technique automatically finds the correct number of clusters. It runs in $O(n \log n)$ time. A novel density-based algorithm was designed by Zhao et al. [127] to cluster the high dimensional data. It uses the density information in the grid structure and by so doing this, its computational cost is almost linear. It is also shown that this method deals with the noisy data. Duan et al. [128] proposed a new algorithm named LDBSCAN that forms local-density-based clusters.
2.6 Grid-Based Clustering Algorithms

The grid-based clustering methods are introduced recently to obtain the results with lesser computational complexity. In grid-based clustering, the dense regions of the multi-dimensional data space are approximated by quantizing it (data space) into a finite number of cells to form a grid structure. The most crucial task of the grid-based algorithms is to identify the appropriate grid size. Similar to the methods of the other models, the grid-based methods have the problem of locating the outliers. Several attempts have been made towards these challenges.

Yue et al. [129] developed a general grid clustering approach (GGCA) based on the divide-and-conquer strategy of hierarchical clustering algorithms. This algorithm is able to deal with the noise. Pilevar et al. [130] was proposed a novel technique called GCHL to cluster the multi-dimensional large spatial data. The GCHL can filter the noise and locate the outliers. This algorithm has linear computational complexity. An algorithm has been proposed by Huang et al. [131] based on least clustering cell (LCC). Although, this method is fast, it is unable to produce the clusters of variable densities and complex shapes. A non-parametric density and grid-based clustering algorithm has been proposed in [132] with the help of shifting grid. It needs additional computational efforts for shifting grid scheme. Zhao et al. [133] proposed an enhanced algorithm to cluster high-dimensional data in which the objects are considered as atomic units. The computational cost of this method is $O(n)$. Sarmah and Bhattacharya [134] have recently developed a grid-based algorithm to find the clusters in satellite image. Darong and Peng [135] developed an algorithm named GRPDBSCAN which is able to deal with the noise. Wang et al. [136] was proposed an algorithm namely, STING for clustering the spatial data. But, only two-dimensional spatial data has been considered here. This scheme is also not scalable. The improved version of the above method is known as STING+ [137] developed using the active database systems along with STING. Similar to STING, the STING+ is also designed for low-dimensional data. Hinneburg
2.6 Grid-Based Clustering Algorithms

and Keim [138] proposed an algorithm for clustering high-dimensional data with noise. The computational cost of this scheme is $O(dn \log n)$. Schikuta [139] developed a new clustering approach named GRIDCLUS to cluster very large data. An extended version of GRIDCLUS, known as BANG can be seen in [140]. Zhao and Song [141] designed an algorithm using the density-isoline figure. A method called WaveCluster [142] has been presented by Sheikholeslami et al. to cluster the spatial data using wavelet transforms. The method is able to produce the clusters of complex shapes and eliminate the noise. Other contributions of grid-based clustering can be found in [143], [144], [145], [146], [147].