A Clustering Algorithm Based on the Combination of MST and Cluster Centers

Xiao-bo LV¹, Yan MA¹,*, Xiao-fu HE² and Hui HUANG¹

¹College of Information and Electrical Engineering, Shanghai Normal University, Shanghai, 200234, China
²College of Physicians & Surgeons, Columbia University, New York, USA
*Corresponding author

Keywords: Minimum spanning tree, K-means, Cluster center, Geodesic distance.

Abstract. Most of traditional MST-based (Minimum spanning tree) clustering algorithms cluster by removing the inconsistent edge. The performances of these algorithms are influenced by the shape of clusters. To address this issue, we proposed a novel cluster algorithm based on the combination of MST and cluster centers (iGMST). Firstly, the cluster centers are determined by the Geodesic distance between vertex pair in the MST. Then, the inconsistent edge is defined along the path between cluster center pair. The experimental results on the synthetic and real data sets show that iGMST is better than k-means++, hierarchical clustering, and spectral clustering. Besides, iGMST can discover clusters of different shape steadily.

Introduction

Clustering is the process of discovering object categories, which makes the objects in the same group more similar, and objects belonging to different groups are not similar. Clustering analysis has very broad applications on data analysis, such as statistics, machine learning and data mining. A wide variety of clustering algorithms have been proposed for different applications [1].

Because of the huge variety of the problems and data distributions, no clustering algorithm is completely satisfactory for all the cases. For example, k-means is one of the most well-known algorithms yet it suffers major shortcomings like convergence to local minima. The drawback of hierarchical-based clustering algorithm is that it is computationally prohibitive when they construct a dendrogram for a large data set. DBSCAN (density-based spatial clustering of applications with noise) [14] algorithm can identify the cluster structure accurately yet it relies on the parameter selection. The performance of spectral clustering algorithm is affected by the similarity matrix. Different similarity matrices will lead to different clustering results.

Sufficient empirical evidences have shown that a minimum spanning tree representation is quite invariant to the detailed geometric changes in clusters’ boundaries. Therefore, the shape of a cluster has little impact on the performance of minimum spanning tree (MST)-based clustering algorithms, which allows us to overcome many of the problems faced by the classical clustering algorithms [2].

Many systems in Science and Engineering can be modeled as graph in which every vertex corresponds to the data point and the edge represents the relationship between points. In MST-based clustering, the weight for each edge is considered as the Euclidean distance between the end points forming that edge. As a result, any edge that connects two sub-trees in the MST must be the shortest. In such clustering methods, inconsistent edges which are unusually longer are removed from the MST. The connected components of the MST obtained by removing these edges are treated as the clusters. Under the ideal condition, that is, the clusters are well separated and there exist no outliers, the inconsistent edges are just the longest edges. However, in real-world tasks, outliers often exist, which makes the longest edges an unreliable indication of cluster separations. To tackle this problem, a local density factor for each data point is taken into consideration in the MST-based clustering algorithm. The degree of link of a vertex [3] or the number of vertices in its k-nearest neighbor [13] is

238
defined as the density of data point. In addition, some MST-based clustering methods are combined with other methods, such as information theory [4], k-means [5], multivariable Gaussian [6], etc.

For the MST-based clustering method, how to find proper inconsistent edges is an important issue. Two vertices forming an inconsistent edge belong to two different clusters. Consequently, the path between two cluster centers of the two clusters must contain the inconsistent edge. Based on this finding, we propose a novel clustering algorithm based on the combination of MST and cluster centers named iGMST. In the procedure of determining the cluster centers, we improve density peaks clustering (DPC) algorithm [12] by introducing Geodetic distance and Gaussian density. With a predefined cluster number K, K-1 inconsistent edges are found on the path between two cluster centers by considering both of the weight of edge and the densities of two vertices forming the edge.

Related Work
Since Zahn et al. [7] firstly proposed MST-based clustering algorithm, researchers have proposed many solutions for the definition of the inconsistent edges. Xu et al. [15] believed that different MST-based clustering problems may need different objective functions. They defined three objective functions. The first objective function is defined to minimize the change of the total weight of the current clusters from the previous one. The objective function for the second and third clustering algorithms is defined to minimize the total distance between a cluster center and each data point in that cluster, which is restricted to the shape of the clusters. Laszlo et al. [8] proposed an MST-based clustering algorithm that puts a constraint on the minimum cluster size rather than on the number of cluster.

Only the information of the edge is used in the traditional MST-based clustering algorithms, which makes the results prone to outliers. Recently, several researchers have developed some improved algorithms. Zhong et al. [3] proposed density oriented MST-based clustering technique assumes that the boundary between any two clusters must belong to a valley region and the inconsistency measure is based on the finding such valley regions. Luo et al. [9] proposed a MST-based clustering algorithm with neighborhood density difference estimation. Wang et al. [10] proposed to find a local density factor for each data point during the construction of an MST. Zhong et al. [11] proposed a graph-theoretical clustering method based on two rounds of minimum spanning tree, which classifies cluster problem into two groups, i.e. separated cluster problems and touching cluster problems, and identifies the two groups of cluster problems automatically.

The Clustering Method
Geodesic Distance
Given a data set \( D \) in \( E^n \) and the desired number of cluster \( K \), the MST-based method starts by constructing an MST \( T_D = (V, E) \) from the points in \( D \). For an edge \( e = (u, v) \in E \), the weight of \( e \) is the Euclidean distance between the two end points \( u \) and \( v \).

Most of MST-based clustering algorithm employs the Euclidean distance as the distance measure between data point pair. However, this kind of distance measure is not suitable for non-spherical clusters. To tackle this problem, we introduce the Geodesic distance measure.

Definition 1 (Geodesic distance): The path \( P \) between two vertices \( u \) and \( v \) in the MST is unique. The geodesic distance between two vertices \( u \) and \( v \) is the sum of Euclidean distance of edges contained in the path \( P \), denoted by \( \text{Geod}(u, v) \).

Selection of Cluster Centers
The DPC algorithm [12] is based on the assumptions that the density of the cluster center is larger than its neighbors and the cluster center is at larger distance from the points with higher density. In the DPC algorithm, the local density of a point \( x_i \), denoted by \( \rho_i \), is defined as
\[ \rho_i = \sum_j \exp \left( -\frac{d(x_i, x_j)^2}{2\sigma^2} \right) \]  

(1)

where \( d(x_i, x_j) \) is the Euclidean distance between \( x_i \) and \( x_j \), \( \sigma \) is the standard deviation, controlling the weight degradation rate. \( \sigma \) is set to 0.02 in the literature [12]. The Euclidean distance measure is only suitable to spherical clusters. In fact, the shape of clusters may be spherical, non-spherical, linear, elongated, drawn-out, etc. To address this problem, we apply Geodesic distance as the distance measure to develop the density calculation. We rewrite \( \rho_i \) as

\[ \rho_i = \sum_j \exp \left( -\frac{\text{Geod}(x_i, x_j)^2}{2\sigma^2} \right) \]  

(2)

where \( \text{Geod}(u, v) \) is the Geodesic distance between \( x_i \) and \( x_j \).

The distance \( \delta_i \) of each point \( x_i \) from points of higher density is defined as

\[ \delta_i = \begin{cases} 
\min_{j: \rho_j < \rho_i} (\text{Geod}(x_i, x_j)), & \text{if } \rho_i < \rho_j \\
\max_j (\text{Geod}(x_i, x_j)), & \text{otherwise}
\end{cases} \]  

(3)

Let \( \rho_{\text{max}} \) be the maximum value of the \( \rho_i \) and \( \delta_{\text{max}} \) be the maximum value of \( \delta_i \). Then, \( \rho_i \) and \( \delta_i \) are normalized as

\[ \rho_i = \frac{\rho_i}{\rho_{\text{max}}} \]  

(4)

\[ \delta_i = \frac{\delta_i}{\delta_{\text{max}}} \]  

(5)

It is pointed out in the literature [12] that the only points of high \( \delta_i \) and relatively high \( \rho_i \) are the cluster centers. Hence we define \( \gamma_i \) as

\[ \gamma_i = \rho_i \times \delta_i \]  

(6)

Finally, the points with larger \( \gamma_i \) are chosen as the cluster center.

**Definition of Inconsistent Edge**

Most of MST-based clustering methods attempt to partition the MST \( T \) into \( K \) subtrees, \( \{ T_i \}_{i=1}^K \), by removing the \( K-1 \) inconsistent edges. It is due to the fact that two end point vertices of the inconsistent edge belong to two different clusters and there exists only a unique path between any two vertices in the MST. We conclude that the inconsistent edge must be on the path connecting two cluster centers. We construct the MST \( T_c \) for the \( K \) cluster centers with the Geodesic distance between cluster center pair. Correspondingly, we obtain \( K \)-1 paths from \( T_c \). It is noted that the \( K \)-1 paths corresponds to the paths in the \( T_D \). The next task is to find the \( K \)-1 inconsistent edges on each of the \( K \)-1 paths. Let \( P_i = \{u_i, u_{i+1}, \ldots, u_{n_i}, u_{i+1}\} \) denote the \( i \)th path. \( u_i \) and \( u_{i+1} \) represent two end vertices of the edge \( u_i u_{i+1} \). Both of the local densities of two end points and the weight of the edge should be taken into the consideration of the determination of the inconsistent edges. Based on this idea, we define a new quantity to give a measurement of the inconsistent edge. For the edge \( u_i u_{i+1} \), this new quantity is defined as

\[ \xi_{i2} = \frac{d(u_i, u_{i+1})}{\rho_i + \rho_{i+1}} \]  

(7)
where \( d(u_i, u_{i+1}) \) is the Euclidean distance between \( u_i \) and \( u_{i+1} \), and \( \rho_i \) and \( \rho_{i+1} \) are the local density of \( u_i \) and \( u_{i+1} \), respectively. The edges with the larger value of \( \varsigma \) are taken as the inconsistent edges and removed from the MST.

**Algorithm Flow**

The proposed algorithm iGMST is illustrated as follows:

- **Input:** Dataset \( D \) to be partitioned, number of cluster \( K \).
- **Output:** The partitioned sub datasets.

**Step 1:** Construct the MST of dataset \( D \) in Euclidean space;

**Step 2:** Calculate \( \rho_i \) and \( \delta_i \) according to Eq. 2-Eq. 5;

**Step 3:** Calculate \( \gamma_i \) according to Eq. 6. Take \( K \) points with the largest \( \gamma_i \) as the cluster centers;

**Step 4:** Construct the MST of the cluster centers;

**Step 5:** Calculate \( \varsigma \) according to Eq. 7. Remove the \( K-1 \) edges with the larger value of \( \varsigma \) from the MST.

**Experiments**

In order to demonstrate the effectiveness of iGMST, we compare the performance of iGMST, k-means++, hierarchical clustering method and spectral clustering on the datasets with two-dimensional attributes and UCI datasets [16] [17], including Iris, Wine, Haberman, Segment, Soybean, and Zoo, as list in Table 1.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Number of Instances</th>
<th>Number of Attributes</th>
<th>Number of Classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iris</td>
<td>150</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>Wine</td>
<td>178</td>
<td>13</td>
<td>3</td>
</tr>
<tr>
<td>Haberman</td>
<td>306</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Segment</td>
<td>2100</td>
<td>19</td>
<td>7</td>
</tr>
<tr>
<td>Soybean</td>
<td>47</td>
<td>35</td>
<td>4</td>
</tr>
<tr>
<td>Zoo</td>
<td>101</td>
<td>16</td>
<td>7</td>
</tr>
</tbody>
</table>

It can be seen from Fig.1 that iGMST algorithm can identify the spherical or non-spherical clusters. The performance of iGMST is also robust to the outliers.

For the UCI Datasets, we used F1-Measure to evaluate the clustering results:

\[
F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}
\]  

(8)

Table 2 shows the clustering results of iGMST, k-means++, hierarchical clustering and spectral clustering with F1-Measure. The higher value of F1-measure indicates a better clustering result. The optimal result for the F1-Measure is denoted in bold in Table 2. As shown in Table 2, except for Haberman, the clustering performance of iGMST is the best among four methods. The performance of iGMST on Haberman is slightly weakened than that of hierarchical clustering.
Figure 1. Clustering resulting of iGMST.

Table 2. Results of the four methods.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>iGMST</th>
<th>K-means++</th>
<th>Hierarchical Clustering</th>
<th>Spectral Clustering</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iris</td>
<td>0.95478</td>
<td>0.8981</td>
<td>0.9168</td>
<td>0.8903</td>
</tr>
<tr>
<td>Wine</td>
<td>0.73384</td>
<td>0.5689</td>
<td>0.6716</td>
<td>0.7163</td>
</tr>
<tr>
<td>Haberman</td>
<td>0.57881</td>
<td>0.5776</td>
<td><strong>0.5887</strong></td>
<td>0.5171</td>
</tr>
<tr>
<td>Segment</td>
<td><strong>0.628</strong></td>
<td>0.2082</td>
<td>0.5101</td>
<td>0.5742</td>
</tr>
<tr>
<td>Soybean</td>
<td><strong>0.90781</strong></td>
<td>0.7667</td>
<td>0.9078</td>
<td>0.9078</td>
</tr>
<tr>
<td>Zoo</td>
<td><strong>0.75618</strong></td>
<td>0.7059</td>
<td>0.6872</td>
<td>0.5060</td>
</tr>
</tbody>
</table>

Conclusions

This paper presents a clustering algorithm iGMST based on the combination of MST and cluster centers. Firstly, our algorithm constructs the MST of the dataset. Next, we search the K cluster centers according to the modified DPC method. Finally, we determine the inconsistent edge along the path between cluster center pair. The experimental results show that our algorithm can identify various shapes of clusters and can obtain better clustering results. In the future work, we plan to determine the correct number of clusters.

Acknowledgment

This research was financially supported by the National Natural Science Foundation of China (No.61373004, 61501297).
Reference


