Tech Review: K-Means vs AGNES

Fan Zang

Introduction

Text clustering is a common exploratory analysis technique used for text mining. The main task is to group similar text objects such as documents, passages and websites in the same cluster. The reason why text clustering is a necessary and crucial preprocess step is because most of the text objects are represented using high-dimensional features such as bag of word model. As the number of features increases, data points become sparse and bigger data set is required in order to get accurate estimations and analysis. Hence, grouping subset of data into clusters allow us to have basic understanding about content of text objects and also apply additional techniques such as principle component analysis on each cluster. This article will cover two common clustering algorithms: K-Means and Agglomerative clustering (AGNE), and compare the drawback between them.

Method Overview

K-Means algorithm tries to partition the data set into K non-overlapping clusters and optimize the distance between each object and its cluster center. It is an iterative method that follows the Expectation-Maximization to solve the problem. After specifying the number of clusters K, K randomly selected data points become the initial centroids. Then the EM iteration starts until convergence: The E step aims to make data points inside the same cluster as close as possible, thus all points are assigned to closest centroids. The M step recalculates all the centroids by taking the average of all points inside each cluster. This iteration stops after the location of centroids and assignment of data points remain the same.

Agglomerative clustering(AGNE) groups data points in “bottom up” manner: the whole data set starts with a cluster per point and then repetitively merges clusters based on inter-cluster distance. The main challenge for this method is to get a efficient way to compute the distance between clusters while merging. By computing the distances between pairs of points in two different clusters, three common clustering algorithms could be used: single-link clustering uses the minimum distance between pairs as inter-cluster distance, complete link clustering uses maximum distance and average clustering uses average distance. At each stage of the merging process, only two cluster that have the smallest inter-cluster distance are merged together and a hierarchical tree Dendrogram representing the relationship between each pair of cluster is built up. The whole process iterate until there is only one cluster and Dendrogram reach the root.

Comparison

K-Means and AGNE are both similarity based approaches which require a objective function to calculate the similarity between two objects. Therefore, most of the computation cost of both methods relies on an efficient similarity function. One optimization used in both methods is feature standardization, which reduce the error in similarity function like Euclidean distance. “Standardizing feature to have a mean of zero and standard deviation of one removes the discrepancy in unit of feature measurement.” (Dabbura, Imad 2018)

The main drawback of K-Means algorithm is inherited from EM iteration: it may only converge to local optimum and the final result relies on the initialization of centroids.

Therefore, running the same algorithm with different K and initial points is crucial to get a satisfying result. Since there is no ground truth of final clusters, two common methods are used to evaluate the output: the Elbow method calculate the sum of squared distance between points and help determine the most suitable K value for a data set, and the Silhouette Analysis determine the degree of separations by computing the average distance between data points. However, even with these methods, K-Means will still have poor performance when number of clusters is huge or the shape of cluster is complicated. Since SSE is the objective function, K-Means doesn't allow points far-away from each other to stay in the same cluster. Thus "K-Means algorithm is good in capturing structure of the data if clusters have a spherical-like shape."(Dabbura, Imad 2018) Moreover, K-Means doesn't learn from data and it always require K as input. Sometimes domain knowledge about data set is needed.

Agglomerative clustering(AGNE) on the other hand is good at identifying large number of small clusters. Unlike K-Means that uses global objective function, AGNE forms hierarchical clustering tree based on local proximity. Moreover, since no predefined variable is needed, the final output is only affected by the local clustering algorithm. Different clustering algorithm could lead to different merging decisions: Single linkage would generate loose clusters, but it's sensitive to outliers and "groups with close pairs can merge sooner even if those groups have overall dissimilarity". (Aggarwal, C. ,2013) Complete linkage is also sensitive to outliers, but would generate tight clusters. Average linkage makes decision based on groups of pairs, it is insensitive to outliers and produce more rounded clusters. Generally it is recommended to run different linkage algorithms in order to get best output. However, the major disadvantage of AGNE is the time and space complexity. "This complex algorithm are about four times the size of K-Means algorithm."(Aggarwal, C. ,2013) Additionally, local proximity merging process is irreversible, which may easily create a problem if data set contains noisy high-dimensional data.

Conclusion

In conclusion, both K-Means and Agglomerative clustering(AGNE) are useful and popular clustering technique that can link similar text objects and preprocess the data points to have a general sense about contents. They are both similarity based methods which rely on efficient similarity function and have their own strengths and weakness. For different applications, if domain knowledge about data set is known and the shape of the clusters is generally spherical, then K-Means will be a better choice considering the time and space complexity. If high-dimensional data points form irregular shape of clusters and size of clusters is small, then choosing AGNE with different linkage methods could form a hierarchical cluster tree and thus have better understanding about the data set.

References

Aggarwal, C. (2013). Data Clustering: Algorithms and Applications (Chapman & Hall/CRC Data Mining and Knowledge Discovery Series Book 31). Chapman & Hall/CRC.

Dabbura, Imad. “K-Means Clustering: Algorithm, Applications, Evaluation Methods, and ...” *K-Means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks*, 17 Sept. 2018, https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a.