Compare K-Means, kernel K-means, K-Medians, K-Medoids

Introduction

In this article, I will go through the algorithm of each cluster method and compare them. In addition, I will rank these four methods to see which one is the most popular nowadays.

Text Clustering

Text clustering is used to understand and categorize unstructured text data. It is an important method used to get insights from text data. Text clustering is the application of different cluster analysis to text-based documents [1]. Text clustering algorithm includes five stages shown below [2]:

1. Transformations on raw text data
2. Transformations on the text into a vector of numbers using TF-IDF
3. Different clustering methods
4. Auto-Tagging based on Cluster Centers

In stage one, we can do some text data cleaning such as removing punctuations, transforming to lower case and so on. In stage two, we need to transform the cleaned text data into a vector of numbers using TF-IDF. In the third stage, we can apply different text clustering methods such as K-Means, kernel K-means, K-Medians, and K-Medoids. I will explore each clustering method in detail in the next paragraph. In the final stage, the algorithm will generate cluster centers, which represent the documents contained in these clusters [2].

K-Means

K-Means is a very popular and straight forward machine learning algorithms. The idea is that there will be K clusters with K centroids, and each centroid is a point that represents the center of a cluster. The algorithm runs iteratively. In this first step, each centroid is generated randomly. Then they move to the center of the points which are closer to them. In each new run the centroids move again to the center of the closest points. When the distance in which the centroids change doesn't surpass a pre-defined threshold, the algorithm is finished [3].

K-Means clustering is very straight forward and easy to implement. It is also fast and efficient in terms of computational cost. However, it is hard to determine value of K. More importantly, this algorithm is sensitive to outliers [4].

Kernel K-Means

K-Means can only detect clusters that are linearly separable, and Kernel K-Means can be used to detect non-convex clusters. Thus, Kernel K-Means suggests that we need to project data onto the high-dimensional kernel space first using the kernel function, and

then perform K-Means clustering [5]. There are several Kernel functions, such as polynomial kernel function, Gaussian radial basis function, and sigmoid kernel.

Kernel-K-Means has higher computational cost compared to K-Means clustering since we need to calculate n by n kernel matrix and store them [5].

K-Medians

K-Medians can handle outliers by computing medians. This algorithm is very similar to K-Means clustering. Instead of calculating means of each cluster, we need to calculate medians [6]. In detail, at the first step, we select K points as initial K medians. Then we get into loop, assigning each point to its nearest medians, then recalculate the median using the median of each individual feature [6]. K-Means minimizes within-cluster variance by calculating the squared Euclidean distances, while K-Medians minimizes absolute deviations [7].

K-Medoids

The term medoid is an object within a cluster for which average dissimilarity between it and all the other the members of the cluster is minimal. It represents the most centrally located point in the cluster [7]. There are many disadvantages for K-Medoids. One of the drawbacks is that this algorithm cannot handle arbitrary shaped groups of objects, and reason is that it needs to calculate and minimize the distance between non-medoid objects and the medoid [8]. Moreover, it is not very stable since it generates different results for different runs on the same dataset, because the first k medoids are chosen randomly [7].

Conclusion

For different text clustering methods such as K-Means, kernel K-means, K-Medians, and K-Medoids, they all need to determine the value of k first, and then generate first k objects randomly. Computation time may differ based on the first k objects. K-Means clustering is the simplest among these clustering algorithms, but it is not very robust to outliers. K-Medians is very similar to K-Means clustering, and it solves the problem of sensitive to outliers. Kernel K-Means and K-Medoids requires more computation time and they are more complex. For Kernel K-Means, it requires some experience of what kind of kernel functions work well in what kinds of situations. Many practitioners use Gaussian Kernels since they believe it usually works well. The most popular clustering algorithm is K-Means clustering. Most people prefer to use this algorithm since it is easy to interpret and implement.

Reference

[1] What is text clustering. July 26, 2018. Retrieved from: https://insidebigdata.com/2018/07/26/what-is-text-clustering/

[2] V,K. Text Clustering: Get quick insights from Unstructured Data. 2017. Retrieved from: https://www.kdnuggets.com/2017/06/text-clustering-unstructured-data.html

[3] L,D. Text Clustering with K-Means. Retrieved from: https://medium.com/bexs-test/text-clustering-with-k-means-a039d84a941b#:~:text=K%2DMeans%20is%20one%20of,a%20K%20number%20of%20clusters.

[4] Advantage of K-Means Clustering. Retrieved from: https://www.quora.com/What-are-the-advantages-of-K-Means-clustering

[5] Kernel K-Means Clustering, Retrieved from: https://www.coursera.org/lecture/cluster-analysis/3-6-kernel-k-means-clustering-g3twJ

[6] The K-Medians and K-Modes Clustering Methods. Retrieved from: https://www.coursera.org/lecture/cluster-analysis/3-5-the-k-medians-and-k-modes-clustering-methods-pShI2

[7] K-Medoids in Retrieved from: https://www.datanovia.com/en/lessons/k-medoids-in-r-algorithm-and-practical-examples/

[8] Disadvantages of K-Medoids. Retrieved from: https://stackoverflow.com/questions/46514123/drawbacks-of-k-medoid-pam-algorithm