DATA MINING ALGORITHMS: K-MEANS CLUSTERING PROBLEM LITERATURE SURVEY PAPER

Clustering Algorithms for Unsupervised Learning

https://youtu.be/lOgKB2bSGbw

Ananya Agrawal, Sharan Sai Reddy Konda
COT5405 - Analysis of Algorithms
University of Florida, CISE Department

***Abstract* -** This paper aims to provide a summarized and organized review of the usage of the Data Mining Algorithm - K-means Clustering in various applications of quantization in Pulse Code Modulation, Information Theory, Machine Learning, etcetera. It emphasizes developing a general perspective of the topic through the usage of certain algorithms, claims, and theorems in use.

I. Introduction

Clustering is a powerful technique for grouping similar items based on their features or characteristics. In the literature, various clustering algorithms have been proposed and studied. One of the most widely used algorithms for clustering is k-means. K-means clustering involves partitioning a dataset into k clusters, with each cluster represented by a centroid. The goal of the k-means clustering is to minimize the distance between the data points and their assigned centroids, resulting in tight and well-separated clusters. It is a method of vector quantization, originally from signal processing, that aims to partition several observations into k clusters concerning the closeness of that observation with its mean or cluster centre of cluster centroid.[2] The problem is computationally difficult (NP-hard); however, efficient heuristic algorithms converge this problem quickly to a local optimum[16].

One key factor that affects the accuracy and efficiency of the k-means clustering is the initialization step. The traditional approach to initialize k clusters is by randomly selecting k data points as the initial centroids however, yielding poor results, especially if the data points are highly correlated or if the clusters have different sizes and shapes. To address this issue, Kanungo et al. (2002) proposed an efficient initialization algorithm called k-means++. The k-means++ algorithm selects initial centroids more smartly by iteratively choosing data points that are farthest from the existing centroids. This approach leads to better initial centroids and faster convergence to the optimal solution.[5]

Another challenge in k-means clustering is determining the optimal value of k or the number of clusters. This problem is often referred to as the "elbow" problem, as the optimal value of k can be identified by observing the point at which the within-cluster sum of squares starts to level off. However, this method can be subjective and difficult to apply in practice. Elkan and Nigam (2003) proposed a new metric, called the gap statistic, for estimating the optimal value of k by comparing the within-cluster dispersion of a k-means solution to that of a null reference distribution. The authors demonstrated that this metric outperforms existing methods for selecting k.

In addition to k-means clustering, other clustering algorithms have been proposed and studied in the literature. For example, Lloyd (1957) proposed a clustering method based on the similarity index, which calculates the similarity between any two samples based on the number of shared features.

II. Body

There are four main types of clustering algorithms: k-means clustering, hierarchical clustering, spectral clustering, and density-based clustering.

K-means clustering: K-means clustering is a simple and efficient clustering algorithm that involves iteratively partitioning the dataset into k clusters based on the distance between data points and the centroids of the clusters. One of the main drawbacks of the k-means clustering is that it is sensitive to the choice of initial centroids and can get stuck in local optima.

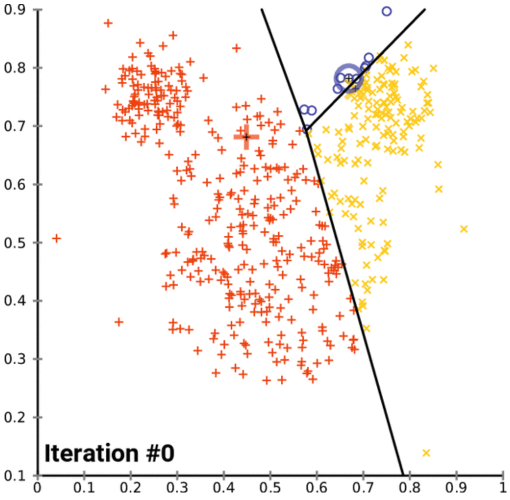
[1]Given a set of observations (x_1, x_2, ..., x_n), where each observation is a d-dimensional real vector, k-means clustering aims to partition the n observations into k (≤ n) sets S = {S1, S2, ..., S_k} so as to minimize the within-cluster sum of squares (WCSS) (i.e. variance). Formally, the objective is to find:

$$\underset{\mathbf{S}}{\arg\min} \sum_{i=1}^{k} \sum_{\mathbf{x}\in S_i} \|\mathbf{x}-\boldsymbol{\mu}_i\|^2 = \underset{\mathbf{S}}{\arg\min} \sum_{i=1}^{k} |S_i| \operatorname{Var} S_i$$

where μ_i is the mean of points in S_i.

The classical k-means algorithm and its variations are known to only converge to local minima of the minimum-sum-of-squares clustering problem defined as

$$\underset{\mathbf{S}}{\arg\min} \sum_{i=1}^{k} \sum_{\mathbf{x}\in S_i} \|\mathbf{x}-\boldsymbol{\mu}_i\|^2.$$



Convergence of k-means

Lloyd's Algorithm uses an iterative refinement technique for the K-means algorithm, also referred to as "naive k-means".

Given an initial set of k means m1,...,mk, the algorithm proceeds by alternating between two steps:

1. ASSIGNMENT STEP: Each observation is assigned to the cluster with the nearest mean or least squared Euclidean Distance,

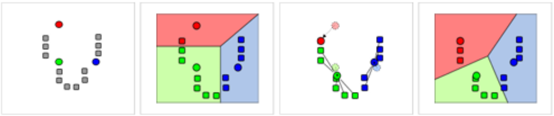
$$S_i^{(t)} = \{x_p : \|x_p - m_i^{(t)}\|^2 \le \|x_p - m_j^{(t)}\|^2 \ \forall j, 1 \le j \le k\},$$

Here, each xp is assigned to exactly one S(t).

2. UPDATE STEP: Recalculate the centroid observations and re-assign points to each cluster based on the assignment condition.

$$m_i^{(t+1)} = \frac{1}{\left|S_i^{(t)}\right|} \sum_{x_j \in S_i^{(t)}} x_j$$

This algorithm converges when the assignments become static.
It does not guarantee an optimal solution, only aims to assign objects to the nearest cluster by distance.



1. *k* initial "means" (in this case *k*=3) are randomly generated within the data domain (shown in color).

2. *k* clusters are created by associating every observation with the nearest mean. The partitions here represent the Voronoi diagram generated by the means.

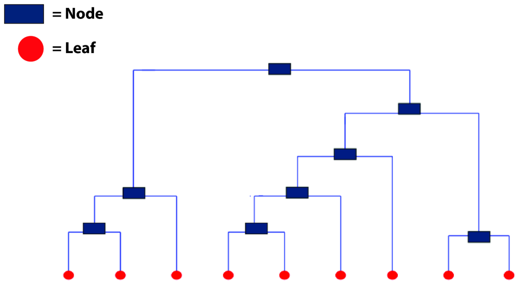
3. The centroid of each of the *k* clusters becomes the new mean.

4. Steps 2 and 3 are repeated until convergence has been reached.

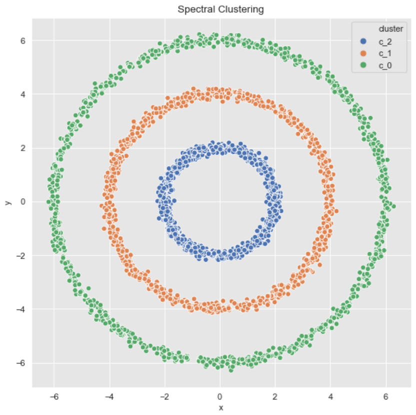
Demonstration[1]

Hierarchical clustering: Hierarchical clustering is a clustering algorithm that involves grouping items into a tree-like structure. Hierarchical clustering can be agglomerative, where each data point starts in its cluster and is successively merged with the closest clusters, or divisive, where all data points start in one cluster and are successively split into smaller clusters. Hierarchical clustering can handle datasets of arbitrary shapes and does not require the number of clusters to be specified in advance, but is computationally expensive and does not scale well to large datasets.

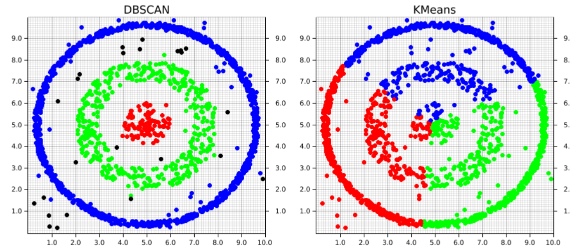
D(r,s) = Trs / (Nr * Ns); Trs is the sum of all pairwise distances between cluster r and cluster s. Nr and Ns are the sizes of the clusters r and s, respectively. At each stage of hierarchical clustering, the minimum D(r,s) is merged.



Spectral clustering: Spectral clustering is a clustering algorithm that involves transforming the data into a lower-dimensional space and then clustering the transformed data using k-means clustering or another clustering algorithm. Spectral clustering can handle non-linearly separable data, but is computationally expensive and can be sensitive to the choice of similarity metric.



Density-based clustering: Density-based clustering is a clustering algorithm that involves partitioning the dataset based on the density of the data points. The algorithm identifies core points and assigns each core point and its neighbouring points to the same cluster. Density-based clustering can handle datasets of arbitrary shapes and can automatically determine the number of clusters, but is computationally expensive.



Comparison of Clustering Algorithms:

Each of the clustering algorithms reviewed in this paper has its strengths and weaknesses. K-means clustering is simple and efficient, but is sensitive to the choice of initial centroids and can get stuck in local optima. Hierarchical clustering can handle datasets of arbitrary shapes and does not require the number of clusters to be specified in advance, but is computationally expensive and does not scale well to large datasets. Spectral clustering can handle non-linearly separable data and Density-based clustering can handle datasets of arbitrary shapes, but both are computationally expensive and can be sensitive to the choice of parameters.

Lloyd’s algorithm does not guarantee convergence to the global optimum. The result may depend on the initial clusters. As the algorithm is usually fast, it is common to run it multiple times with different starting conditions. However, worst-case performance can be slow: in particular certain point sets, even in two dimensions, converge in exponential time, that is 2Ω(n).[16]

Complexity:

K - means clustering in d dimension is an NP-hard problem in general Euclidean space for 2 or more clusters.[11][12][13]
Thus, a variety of heuristic algorithms such as Lloyd's algorithm given above are generally used. The running time of Lloyd's algorithm (and most variants) is, O(nkdi) where n is the number of d-dimensional vectors (to be clustered), k is the number of clusters, and i is the number of iterations needed until convergence.[8]

Evaluation of Trends:

One of the main trends in clustering research is the development of clustering algorithms that can handle large datasets with high dimensionality. Another trend is the development of clustering algorithms that are robust to noise and outliers. In addition, there has been a significant amount of research on clustering algorithms that can handle non-linearly separable data, such as spectral clustering. There has also been a trend towards developing clustering algorithms that can handle datasets with arbitrary shapes, such as hierarchical clustering and density-based clustering.

Challenges:

One of the main challenges in clustering research is the choice of the appropriate clustering algorithm for a particular application. The choice of clustering algorithm depends on the characteristics of the dataset and the specific requirements of the application. Another challenge is the evaluation of clustering algorithms since there is no universally accepted measure of clustering quality. In addition, there is a need for clustering algorithms that can handle streaming data, where the data is continuously arriving and changing over time.

One important factor to consider when using clustering algorithms is the quality of the data being clustered. In real-world datasets, missing or inconsistent data can affect the accuracy and reliability of clustering results. One approach to addressing this issue is imputation, where missing data is filled in with estimated values based on the available data. For example, Hamerly and Elkan (2010) proposed a k-means algorithm called k-means-- which includes a novel imputation step that estimates missing values using a Bayesian network. The authors demonstrated that k-means-- outperforms existing k-means algorithms on several datasets with missing values.[4]

Another approach is to combine them with other techniques, such as dimensionality reduction or feature selection. Dimensionality reduction techniques, such as principal component analysis (PCA) and t-distributed stochastic neighbour embedding (t-SNE)[17], can be used to reduce the dimensionality of high-dimensional datasets, making it easier to visualize and cluster the data. Feature selection techniques, such as mutual information or recursive feature elimination, can be used to identify the most important features for clustering, reducing the computational complexity and improving the accuracy of clustering results.

While clustering algorithms are powerful tools for grouping similar items, they can be limited in their ability to handle complex datasets with multiple conflicting objectives. Multi-objective genetic algorithms (MOGAs) have been proposed to address this challenge. MOGAs involve optimizing multiple objectives simultaneously, where each objective corresponds to a different aspect of the problem. Deb et al. (2002) proposed a fast and elitist MOGA called NSGA-II, which includes several new features such as a fast non-dominated sorting algorithm, a crowding distance metric for selection, and a diversity-preserving mechanism. The authors demonstrated that NSGA-II outperforms existing MOGAs on several benchmark problems.[15]

In the paper, "Least Squares Quantization in PCM" by Stuart P. Lloyd, published in 1957, Lloyd introduces a new method for quantizing signals in Pulse Code Modulation (PCM) systems, which he calls "Least Squares Quantization." [2]

Lloyd begins by discussing the problem of quantizing signals in PCM systems, which involves approximating a continuous signal with a finite number of discrete values. He notes that traditional methods for quantization can result in significant distortion of the signal, particularly when the number of quantization levels is low and proposes a new method based on the principle of least squares where the optimal quantization levels can be found by minimizing the sum of the squared errors between the original signal and its quantized approximation. He derives an algorithm for finding these optimal levels and shows that it converges to a stable solution.

Lloyd concludes by discussing some practical considerations for implementing his method, including how to choose an appropriate number of quantization levels and how to handle non-uniform probability distributions of signal amplitudes.

Dan Pelleg and Andrew Moore propose a new algorithm for the K-means clustering method in their research paper "Accelerating exact K-means algorithms with geometric reasoning," which was published in the Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98) in 1998. The algorithm makes use of geometric reasoning to minimize the number of distance calculations needed.

Hae-Sang Park and Chi-Hyuck Jun propose a new algorithm for the K-means clustering method in their 2009 article, "A Fast Algorithm for the K-Means Clustering Method Using Triangular Inequality", which makes use of triangular inequality to minimize the number of distance calculations needed. They show that their algorithm can significantly reduce the computational complexity of the K-means clustering method, particularly for large datasets.

The authors provide a detailed description of their algorithm and analyze its computational complexity. They also compare its performance to other existing algorithms for the K-means clustering method and show that it is faster in many cases.

III. Conclusion

Clustering is an important technique in unsupervised learning that has a wide range of applications in various fields. In this literature survey paper, we provided a general overview of the field of clustering and presented a classification of the existing literature, a perspective on the area, and an evaluation of trends. We identified various clustering algorithms and discussed their strengths and weaknesses along with the challenges they pose.

Clustering algorithms are powerful techniques for grouping similar items based on their features or characteristics. K-means clustering is a very popular and widely used clustering algorithm, and its accuracy and efficiency can be improved by using smarter initialization algorithms and estimating the optimal value of k using methods such as the gap statistic. Other clustering algorithms, such as hierarchical clustering and Lloyd's method, have also been proposed and applied to various domains. However, clustering algorithms can face challenges in handling complex datasets with multiple conflicting objectives, which can be addressed by using multi-objective genetic algorithms, imputation techniques, or combining clustering with other techniques such as dimensionality reduction or deep learning.

It is important to note that the performance of clustering algorithms heavily depends on the quality of the data being clustered. Missing or inconsistent data can affect the accuracy and reliability of clustering results, and imputation techniques such as k-means-- can be used to estimate missing values and improve clustering performance. Additionally, dimensionality reduction techniques such as PCA and t-SNE can be used to reduce the dimensionality of high-dimensional datasets, making it easier to visualize and cluster the data.

Feature selection techniques like Mutual information and recursive feature elimination can also be used to identify the most important features for clustering, reducing the computational complexity and improving the accuracy of clustering results.

In addition to traditional clustering algorithms, there are also several variants of k-means clustering that have been proposed to address its limitations. For example, k-means++ is an improvement over the traditional k-means algorithm that provides better initialization of the cluster centroids, while k-medoids is a variant that uses medoids instead of centroids. K-means clustering is a popular and widely used algorithm, but its limitations can be addressed by using smarter initialization algorithms, estimating the optimal value of k, or using variants such as k-means++, k-medoids, k-means--, or kernel k-means. They include an imputation step, and kernel k-means, which uses a kernel function to map the data to a higher-dimensional space before clustering.

Furthermore, multi-objective genetic algorithms (MOGAs) have been proposed as a means of optimizing multiple objectives simultaneously. MOGAs have been applied to clustering in various domains, such as image and text data, and have shown promising results.

Other clustering algorithms such as hierarchical clustering and Lloyd's method have also been proposed and applied to various domains. Additionally, imputation techniques, feature selection techniques, dimensionality reduction techniques, and deep clustering can be used to improve the performance of clustering algorithms. Clustering continues to be an important and active area of study, with many challenges and opportunities for new research.

IV. References

1. k-means clustering - Wikipedia. (2009, April 7). K-means Clustering - Wikipedia. https://en.wikipedia.org/wiki/K-means_clustering
2. Lloyd, S. (1957). Least squares quantization in PCM. IEEE Transactions on Information Theory, 28(2), 129-137.
3. MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability (Vol. 1, pp. 281-297).
4. Hamerly, G., & Elkan, C. (2003). Learning the k in k-means. In Proceedings of the 17th International Conference on Neural Information Processing Systems (pp. 281-288).
5. Arthur, D., & Vassilvitskii, S. (2007). K-means++: The advantages of careful seeding. In Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms (pp. 1027-1035).
6. Pelleg, Dan; Moore, Andrew (1999). "Accelerating exact k -means algorithms with geometric reasoning" (http://portal.acm.org/citation.cfm?doid=312129.312248). Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining KDD '99. San Diego, California, United States: ACM Press: 277–281.
7. Hartigan, J. A.; Wong, M. A. (1979). "Algorithm AS 136: A k-Means Clustering Algorithm". Journal of the Royal Statistical Society, Series C.
8. Hamerly, Greg; Elkan, Charles (2002). "Alternatives to the k-means algorithm that find better clusterings" (http://people.csail.mit.edu/tieu/notebook/kmeans/15_p600-hamerly.pdf)
9. Celebi, M. E.; Kingravi, H. A.; Vela, P. A. (2013). "A comparative study of efficient initialization methods for the k-means clustering algorithm". Expert Systems with Applications.
10. Aloise, D.; Deshpande, A.; Hansen, P.; Popat, P. (2009). "NPhardness of Euclidean sum-of-squares clustering" (https://doi.org/10.1007/s10994-009-5103-0). Machine Learning.
11. Kleinberg, Jon; Papadimitriou, Christos; Raghavan, Prabhakar (199812-01). "A Microeconomic View of Data Mining". Data Mining and Knowledge Discovery.
12. Dasgupta, S.; Freund, Y. (July 2009). "Random Projection Trees for Vector Quantization". IEEE Transactions on Information Theory.
13. Hamerly, Greg; Drake, Jonathan (2015). Accelerating Lloyd's algorithm for k-means clustering. Partitional Clustering Algorithms. pp. 41–78.

14. Deb, Kalyan & Pratap, Amrit & Agarwal, Sameer & Meyarivan, T.. (2002). A fast and elitist multiobjective genetic algorithm: NSGA-II. Evolutionary Computation, IEEE Transactions on. 6. 182 - 197. 10.1109/4235.996017.
15. Vattani, A. (2011). "k-means requires exponentially many iterations even in the plane" (http://cseweb.ucsd.edu/users/avattani/papers/kme ans-journal.pdf)
16. Mahajan, Meena; Nimbhorkar, Prajakta; Varadarajan, Kasturi (2009). The Planar k-Means Problem is NP-Hard. Lecture Notes in Computer Science.
17. Ding, Chris; He, Xiaofeng (July 2004). "K-means Clustering via Principal Component Analysis" (http://ranger.uta.edu/~chqding/papers/KmeansPCA1.pdf)