A Comparative Analysis of the Study of Optimization Schemes for K-Means Algorithm Clustering Centers in High-Dimensional Data

Jinglu Tian*

School of Communication Engineering, Xidian University, Xi'an, 710000, China
**Corresponding author. Email: jt197@hw.ac.uk*

ABSTRACT
As an important concept of artificial intelligence in the field of information mining and in the broader field of deep learning, clustering analysis has attracted a large number of researchers to think about and improve its research methods, application areas and disadvantage optimization to different degrees. The traditional K-means clustering algorithm suffers from the fact that the number of clusters required needs to be determined artificially, and therefore the clustering results can be influenced by the different initial cluster centers, and the computational complexity of the clustering iteration process. Especially when processing multi-dimensional or high-dimensional data, the number of iterations, the computational complexity and the long running time can affect the effectiveness and accuracy of the algorithm. Different researchers have proposed optimization solutions for this drawback based on different priorities. This paper provides a comparative analysis of these schemes to explore their feasibility and advantages and disadvantages.

Keywords: *Clustering algorithms, K-means algorithm, initial clustering centres, centre optimization, high-dimensional data*

1. INTRODUCTION

Clustering analysis is an important part of research in artificial intelligence and one of the most widely researched areas at present. It has been widely used in different research areas such as machine learning, statistical analysis of data, pattern recognition, database data mining, etc. In particular, the field of data mining, which extracts potentially useful information from a large number of fuzzy and irregular data samples, will have a broader research prospect in the future.

As one of the main data mining methods, clustering algorithms use the characteristics of samples to compare the similarity of samples and place the similar samples into the same category, while the samples with greater differences are stored in different groups. The aim is to maximise the similarity of data objects in the same cluster and minimise the similarity of data objects in different classes. Unlike classification algorithms, clustering analysis algorithms do not require artificially set classification criteria for different classes, but are automatically classified by the program according to the characteristics of the data group.

Clustering has a wide range of applications and has given rise to many clustering algorithms. This paper discusses the k-means algorithm, which has a wider range of applications and a simpler principle.

The k-means algorithm is based on partitioning, determining the initial cluster centres and classifying the data by taking the mean. In fact, since MacQueen proposed the k-means algorithm, a large body of literature has been produced on his algorithm, including studies on the A-value of the number of clusters, similarity measures and clustering evaluation criteria, and the elimination of noise points and isolated points. For the selection of the initial clustering centre, the simplicity of the principle of the algorithm gives rise to many problems accordingly. For example, the initial clustering centres are chosen randomly by the program without any manual operation, which will lead to problems of locally optimal solutions and unstable clustering results.

In response, different researchers have proposed solutions to optimise the clustering centres, including ideas such as the K -means++ algorithm, reference distance density values, outlier detection, the use of

outlier factors and the maximum-minimum algorithm. As different optimisation solutions have different focuses, the aim of this paper is to investigate the operational steps, feasibility as well as advantages and disadvantages of different optimisation algorithms through a comparative study, and to propose further optimisation directions and prospects for future research. At present, research on k-means algorithm also includes its application on clustering in business, industry, science and other fields, such as k-means algorithm in banking and telecommunication customer information clustering and data extraction.

2. THEORETICAL BASE

2.1 Clustering algorithms

Clustering is an unsupervised learning method for machine learning in the field of artificial intelligence that divides a collection of physical or abstract objects into multiple classes consisting of similar objects. It is characterised by the fact that the clusters generated by clustering are a collection of data objects that are similar to each other and dissimilar to each other in the same cluster as the objects in other clusters. Cluster analysis, also known as cluster analysis, requires that the classes classified are unknown and has a wide range of applications. In business, clustering helps market analysts to analyse a base pool of customers to discover different customer groups; in biology it can be used to classify genes and derive classifications of plants and animals; and even in data mining clustering plays an important role.

Traditional clustering analysis algorithms are generally divided into methods based on probabilistic models and non-parametric methods. For non-parametric methods, clustering is mostly based on the objective function of similarity or dissimilarity measures, and can be divided into hierarchical and partitioned methods, with the partitioned method being more commonly used. Usually the partitioning method is based on the assumption that the data set can be represented by a number of cluster prototypes with their respective objective functions. It is therefore more important to determine the dissimilarity (or distance) between a point and a clustering prototype, of which the K-means algorithm is more popular and widely known, not only triggering research and extensions by many scholars, but also being used in various fields.

2.2 K-means algorithm

The K-means algorithm is a division-based clustering algorithm, which is based on Euclidean clustering and has the advantages of simplicity, ease of implementation, interpretability and fast convergence, making it a classic algorithm in data mining. There are three important parameters in the algorithm that need to be specified artificially, namely the number of clusters, the initial cluster centres and the similarity measure. In this case, the initial cluster centres are randomly generated from the dataset based on a user-defined number of K clusters. The choice of the initial cluster centres plays a crucial role in this algorithm. However, the simple principle also brings some shortcomings to the K-means algorithm. Firstly, the number of clusters, K, is generally unknown and needs to be determined artificially, and too large or too small a K value can affect the clustering effect. Secondly, the initial clustering centres are chosen randomly by the program without human control, and different initial clustering centres will produce different clustering results and different accuracy, which is prone to the problem of local optimal solutions and unstable clustering results. Thirdly, the K-means algorithm is an iterative method that eventually converges to a certain result, and when it converges, the distribution of data points to the cluster centre no longer changes, the distance is called the "fitting error", the fitting error is plotted as a function of the location of the cluster centre, this function has many local minima. In the program, it takes a lot of time and luck to reach the global minimum, so isolated points in the data have a greater impact on the results and the algorithm tends to get stuck on the local minimum.

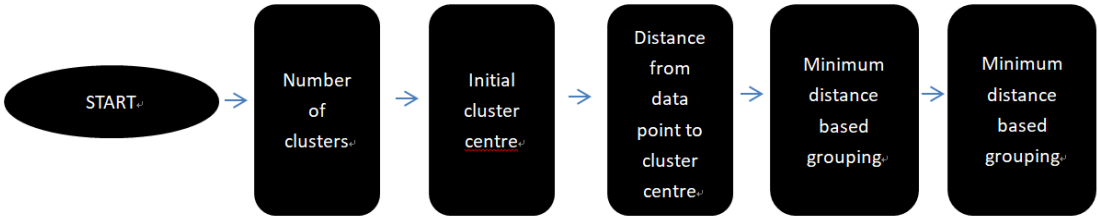

Figure 1 The diagram of K-means algorithm

2.3 Subspace clustering algorithms

Subspace clustering algorithm, as one of the key techniques of cluster analysis in the field of data mining, refers to dividing the original feature space of data into different feature subsets, studying the significance of observing the clustering of each data cluster from different subspace perspectives, while finding the

corresponding feature subspace for each data cluster in the process. In the field of data analysis, as high-dimensional data clustering is a major challenge in cluster analysis techniques, the subspace clustering algorithm is an effective way to achieve clustering of high-dimensional data sets based on an extension of the traditional clustering algorithm, the central idea of which is to localise the search in the relevant dimension.

In general, there are two main tasks in subspace clustering: discovering subspaces that can be clustered and clustering on the corresponding subspaces. This algorithm actually combines traditional feature selection techniques with clustering algorithms to obtain the subset of features or weights corresponding to each data cluster in the process of clustering and dividing the data samples. Research to date has shown that subspace clustering can be divided into two forms: hard subspace clustering and soft subspace clustering. The difference is that hard subspace clustering algorithms identify the exact subspace in which different classes are located, while soft subspace clustering finds a soft subspace for each class. In simple terms, in hard subspace clustering, a subspace has one and only one attribute and the clustering process takes place in these subspaces, whereas soft subspace clustering clusters the entire data set in a full dimensional space with all attributes in each subspace and each attribute is given a different weight from 0 to 1. The higher the weight the more important the attribute is and the stronger the association with that subspace.

2.4 Pearson's correlation coefficient

Pearson's correlation coefficient is a measure of the similarity between two variables, with a value between -1 and 1. A correlation coefficient greater than 0 indicates a positive correlation, while the opposite is true, and a value equal to 0 indicates no correlation. There are constraints on the Pearson correlation coefficient, which requires that both variables are continuous and independent, and that there is a linear relationship between the variables, which is normally distributed, and that their binary distribution is also normally distributed. In practice, the correlation coefficient (i.e. the calculated value) and the independent sample test coefficient (which tests the consistency of the sample) are generally output.

3. ANALYSIS

The K-means algorithm has been developed to date, but due to limitations such as the initial clustering centres not being set artificially, the problems of stability, speed and accuracy of the results still exist and have not been completely solved. As the choice of the initial clustering centre is completely random, it may lead to a slow convergence of the algorithm, which has been studied by different researchers based on different algorithms and considerations for algorithm improvement when processing high-dimensional data.

First of all, the main idea of the most widely known K-Means++ algorithm is to optimize the selection of the initial clustering centre points, so that the initial clustering centre points are spread out as much as possible, and its steps are mainly to select the initial clustering centre A1 randomly for the input data set, and then calculate the distance from each data point Xi to the nearest clustering center in the set according to the formula $D(X_i) = \mathrm{argmin}(|X_i - A_r|)^2, (r = 1,2, \ldots$ selected $k)$. The distance from each data point Xi in the set to the nearest cluster centre of the selected cluster centres is calculated, after which a new cluster centre is selected based on the principle that the point with the higher D(X) has a higher probability of being selected as the new cluster centre, and these two steps are repeated until k cluster centres are selected. Numerous experimental results have demonstrated that this algorithm can effectively improve the overall computing speed due to the reduced number of iterative operations.

In addition, Ping Zong, Junyan Jiang and Jun Qin from Nanjing University of Posts and Telecommunications, China, proposed an EDK-means algorithm based on subspace clustering algorithm and traditional k-means algorithm, combined with distance optimization method and density method, to improve the effectiveness and feasibility when dealing with high-dimensional data. The basic idea of the algorithm is as follows.

The density parameters of all data points in the data set according to the distance density feature model formula are calculated. Given a set of m-dimensional data sets containing n data objects, the Euclidean distance between data objects and the Euclidean distance between any two data objects in this space can be defined by the following formula.

$$d(x_i, x_j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{im} - x_{jm})^2} \quad (1)$$

At this point the average distance $\bar{d}$ of the data for the entire data set X can be defined as follows, where C_n^2 is the number of pairwise combinations selected from the n data objects of the data set.

$$\bar{d} = \frac{1}{C_n^2 \sum d(x_i, x_j)} \quad (2)$$

For a given data object, the density parameter can be defined as the number of data objects within a circle with centre x_i and radius $\bar{d}$. The following equations can be obtained from equations (1) and (2).

$$D(x_i) = \sum_{j=1}^{n} f\left(\bar{d} - d(x_i, x_j)\right) \quad (3)$$

At this point, f(z) satisfies $f(z) = \begin{cases} 0, & z < 0 \\ 1, & z \geq 0 \end{cases}$. The average density $\overline{D}(X)$ of the set X is obtained as follows.

$$\overline{D}(X) = \frac{1}{n}\sum_{i=1}^{n} D(x_i) \quad (4)$$

According to the formula $D(x_i) < \alpha\overline{D}(X)$,the isolated data points are judged and removed to obtain a new data set B1 with high-density parameters.

10%-15% of the data samples in B1 are taken to form the sample set C1, the average distance of adjacent data objects is calculated in C1 and it is defined as the interval length LENGTH, and the relevant interval Ii according to the formula $I_i = (i * LENGTH, (i + 1) * LENGTH)$ is defined.

The object is selected with the highest density parameter from B1 as the first cluster centre and it is removed from B1 and added to the empty set B2.

Afterwards, selecting the data object furthest from B2 in B1 as the next cluster centre, again removing it in B1 and adding it to B2. Repeating this step until k cluster centres are found.

According to the Euclidean distance formula, each data object in the boundary set is assigned to the nearest cluster and the safe distance e_p for each data object in the boundary set is calculated according to the formula$e_p = \min(d_{PC_j} - d_{PC_i})$, which is mapped to the corresponding interval.

The centre of each cluster is recalculated and the maximum offset of the centre Δd, is obtained according to $\Delta d = \max\left(d_{C_l C_{l'}}\right), (l = 1,2, \ldots, k)$. If the maximum offset is 0 or the maximum number of iterations is reached, the algorithm ends.

The boundaries of the relevant interval are updated by reducing the size of 2*Δd, taking out all data objects corresponding to the lower edge of the relevant interval that are less than or equal to 0 and mark them as the new set of boundaries, and returning to step v.

Experimental results on accuracy and recall analysis show that the algorithm improves accuracy and completeness over the k-means algorithm, and reduces the extra computational process of the iterative process, with a corresponding reduction in computation time.

In addition, a novel algorithm based on the ant algorithm, Pearson's correlation and density ideas was proposed by Qingqing Xie and He Jiang from Qilu University, improving the problem of randomness in the selection of initial clustering centres in the traditional k-means algorithm.

The principle is to select the clustering centres uniformly while avoiding the selection of initial clustering centres as far as possible. The basic steps are shown below.

For a set D containing n data objects, selecting the smallest data object min(D) and calculating the Pearson correlation coefficient r between all data objects and the smallest data object, sorting r in descending order from the largest to smallest.

It is grouped according to the above arrangement at a length of m/k, where m denotes the total number of data objects and k denotes the total number of clusters.

For data objects that are not specified to be grouped, the similarity between them and the grouped data objects are calculated and grouped into the group with the greatest similarity. The initial cluster centroids are calculated based on the groupings. At this point the initial clustering centroids are represented as follows.

$$= \frac{\sum_{j-1}^{n} T_j}{n}, T_j M_i \quad (5)$$

At this point M_i denotes the initial cluster centre, T_j denotes the data belonging to cluster M_i, and n denotes the number of data in M_i.

The experimental data show that this algorithm has improved in varying degrees over the original k-means algorithm in terms of precision, accuracy and accuracy.

4. CONCLUSION

From the previous analysis, it can be seen that the upgraded k-means++ algorithm, despite the extra time required for initial selection, can effectively reduce the number of iterations and converge quickly at a later stage, resulting in faster operations overall, but the algorithm only samples one sample per traversal, which means it is difficult to achieve parallelism. It does not change the clustering criterion function $E = \sum_{i=1}^{k}\sum_{p\in C_i} \text{dist}(p, C_i)^2$, (p is the point in the space that represents the specified data object, and C_i represents the cluster centre) so it is still a convergent algorithm.

For the EDK-means algorithm, the advantages of the algorithm are, obviously, faster convergence, optimised execution efficiency and reduced time to select the initial clustering centres. However, despite its many optimally tuned advantages in terms of accuracy, completeness and iteration time, the algorithm does not show an advantage over the traditional K-means algorithm as it focuses on reducing the time to select the initial clustering centres when the clustering centres of the data points are better found. Therefore, this algorithm is only useful for more complex data sets where local optimisation exists.

As for the third algorithm discussed, the one based on the "Pearson correlation coefficient", it takes into account that the random selection of the initial clustering centres can have a significant impact on the data results and therefore chooses to select the clustering centres evenly rather than circumventing the selection of the initial

clustering centres. The result is a significant improvement over the traditional k-means algorithm in terms of accuracy and stability, avoiding the randomness and unevenness found in traditional algorithms and improving the quality of the clusters. The algorithm circumvents the problem of randomly selecting initial clustering centres, reducing the sensitivity of the input sequence and the possibility of local optima to a greater extent. However, there is no significant improvement or advantage in reducing the number of iterations to reduce computation time for all databases, but it does not show unique advantages for high-dimensional data.

Therefore, in summary, for different algorithms, how to improve accuracy while eliminating the effect of randomly selected clustering centres on the stability of the results, reduce the number of iterations to save more running time and improve the efficiency of the algorithm will be the focus and centre of future research to further explore optimisation options and retain as many advantages as possible.

ACKNOWLEDGMENT

I would like to conclude this article by expressing my gratitude to Professor Lio of the University of Cambridge, who can be considered as the leader of my research in this field. I am honored to have had the opportunity to participate in his classes and learn about the field of artificial intelligence and deep learning, which has greatly expanded my knowledge. I can hardly forget his deep concern for the progress and outcomes of his students during my study period with him, while I was also impressed by the wealth of knowledge in the classroom. It is fair to say that this research paper would not have been produced without his leadership and I would like to thank him again profusely and hope to have the opportunity to learn more deeply with him in the future.

REFERENCES

[1] FAHIM A M, SALEM A M, TO R KEY F A, et al. An efficient en-hanced k-means clustering algorithm [J]. Journal of Zhejiang University, Science A, 2006, 7 (10): 1626-1633.

[2] P. Zong, J. Jiang and J. Qin, "Study of High-Dimensional Data Analysis based on Clustering Algorithm," 2020 15th International Conference on Computer Science & Education (ICCSE), 2020, pp. 638-641, doi: 10.1109/ICCSE49874.2020.9201656.

[3] V. Divya, R. Deepika, C. Yamini and P. Sobiyaa, "An Efficient K-Means Clustering Initialization Using Optimization Algorithm," 2019 International Conference on Advances in Computing and Communication Engineering (ICACCE), 2019, pp. 1-7, doi: 10.1109/ICACCE46606.2019.9079998.

[4] Q. Xie, H. Jiang, B. Han and D. Wang, "Improved Initial Clustering Center Selection Method for k-Means Algorithm," 2018 Eighth International Conference on Instrumentation & Measurement, Computer, Communication and Control (IMCCC), 2018, pp. 1092-1095, doi: 10.1109/IMCCC.2018.00227.

[5] S. Paul, S. De and S. Dey, "A Novel Approach of Data Clustering Using An Improved Particle Swarm Optimization Based K–Means Clustering Algorithm," 2020 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT), 2020, pp. 1-6, doi: 10.1109/CONECCT50063.2020.9198685.

[6] T. Lei and S. Li, "Improved K-means Clustering Algorithm by Combining with Multiple Factors," 2021 3rd International Conference on Advances in Computer Technology, Information Science and Communication (CTISC), 2021, pp. 258-263, doi: 10.1109/CTISC52352.2021.00054.

[7] L. Bai, Z. Song, H. Bao and J. Jiang, "K-means Clustering Based on Improved Quantum Particle Swarm Optimization Algorithm," 2021 13th International Conference on Advanced Computational Intelligence (ICACI), 2021, pp. 140-145, doi: 10.1109/ICACI52617.2021.9435862.

[8] Hui Ai and Wei Li, "K-means initial clustering center optimal algorithm based on estimating density and refining initial," 2012 6th International Conference on New Trends in Information Science, Service Science and Data Mining (ISSDM2012), 2012, pp. 603-606.

[9] X. Chen and Y. Xu, "K-Means Clustering Algorithm with Refined Initial Center," 2009 2nd International Conference on Biomedical Engineering and Informatics, 2009, pp. 1-4, doi: 10.1109/BMEI.2009.5304749.

[10] J. Tian, L. Zhu, S. Zhang and L. Liu, "Improvement and parallelism of k-means clustering algorithm," in Tsinghua Science and Technology, vol. 10, no. 3, pp. 277-281, June 2005, doi: 10.1016/S1007-0214(05)70069-9.

[11] K. P. Sinaga and M. Yang, "Unsupervised K-Means Clustering Algorithm," in IEEE Access, vol. 8, pp. 80716-80727, 2020, doi: 10.1109/ACCESS.2020.2988796.