

Analysis of Research Progress on Large-Scale Datasets Based on Clustering Algorithms

Chuanjian Jiang^{1*}

¹ Chongqing College of International Business and Economics, Chongqing, China

Email Address

317387646@qq.com (Chuanjian Jiang)

*Correspondence: 317387646@qq.com

Received: 21 October 2024; **Accepted:** 14 November 2024; **Published:** 2 December 2024

Abstract:

The article analyses the research progress of clustering algorithms for large-scale datasets as an example. It includes clustering algorithm and its development, research progress of serial clustering algorithm in large-scale datasets, and research progress of parallel clustering algorithm in large-scale datasets. It is hoped that this analysis can provide a reference for the reasonable application of clustering algorithms to achieve the effective processing of large-scale datasets.

Keywords:

Large-scale Datasets, Clustering Algorithms, Serial Clustering Algorithms, Parallel Clustering Algorithms

1. Introduction

Clustering algorithms in the field of machine learning play an indispensable application advantage in large-scale data set processing. Therefore, in the study of large-scale data sets, researchers need to use clustering algorithms as a supportive technology, according to the actual situation and the needs of large-scale data processing, through the serial clustering algorithm or parallel clustering algorithm for processing. In this way, the clustering algorithm can give full play to the application advantages of the clustering algorithm, and the large-scale data can be effectively processed to meet the practical application requirements.

2. Clustering Algorithms and Their Development

2.1. Basic clustering algorithms

The clustering algorithm belongs to a machine learning algorithm in an unsupervised mode, whose basic role is to mine the similarities inherent in large-scale data. In specific processing, the algorithm will divide the large-scale data into a number of subsets, and treat each subset as a cluster, where the samples in each cluster are similar, but the samples within different clusters are different. When processing, the algorithm describes its sample similarity by probabilistic similarity,

Pearson's similarity, Manhattan distance, Mars distance, Chebyshev distance, cosine similarity, and Euclidean distance [1]. With the advantages of fast processing speed and high degree of intelligence, clustering algorithms have been widely used in modern data processing. Especially in large-scale dataset processing, clustering algorithms show more significant applicability and have gained faster research progress in it. In terms of the current address book customer grouping, address book friend grouping, gene sequence classification, biological species classification, library book classification and other large-scale data set processing, through the reasonable application of clustering algorithms, in order to ensure the quality of large-scale data set processing on the basis of the data processing time will be significantly shortened, and the processing system operation will be more fluent.

2.2. The development of clustering algorithms

With the development of modern intelligent technologies, the accuracy of clustering algorithms and their applicability have been increasing. At first, the clustering algorithm was only applicable in the processing of various small and medium-sized data sets. Through the reasonable application of the algorithm, the convenience of small and medium-sized data collection and storage can be significantly improved. In the current era of big data, with the emergence of GB-level datasets, TB-level datasets and larger datasets, the clustering algorithms applied in them have also been continuously developed. After the comprehensive consideration of the quality and efficiency of large-scale data sets in various fields, more and more new clustering algorithm models have begun to appear, so that clustering algorithms have been updated on the traditional basis and have shown good adaptability in large-scale data set processing. When processing large-scale data sets, serial clustering algorithms and parallel clustering algorithms are the main clustering algorithms. Through the reasonable application of the above clustering algorithms, while ensuring the quality of large-scale data set processing, it can effectively solve the problem of poor system smoothness caused by long-time computing, and provide strong support for the analysis, processing and application of large-scale data sets.

3. Advances in Serial Clustering Algorithms for Large Scale Datasets

3.1. Introduction to Serial Clustering Algorithms

The so-called serial clustering algorithm is to consider each data point in a large-scale dataset one by one, and take the similarity between each data point and the existing data class as the basis for deciding to assign it to the existing data class or create another data class separately. When applied specifically, the essence is to implement compression processing of the sample space or feature space on the basis of the original large-scale dataset, and to improve the processing speed of large-scale data computation by clustering the compressed dataset. In terms of the current large-scale data set computational processing, there are two most commonly used serial clustering algorithms, one is the serial clustering algorithm for large-scale data sets in the incremental learning mode, and the other is the serial clustering algorithm for large-scale data sets in the feature transformation mode.

3.2. Serial clustering algorithm for large-scale datasets in incremental learning mode

The incremental learning mode is to take the incremental learning framework as the basis and complete the large-scale dataset operation within it through the clustering algorithm. In this process, the serial clustering algorithm will first run on certain large-scale data sets, and then according to the influence of large-scale data in the algorithm, the serial clustering results are dynamically updated to achieve the addition, deletion, query or modification of various types of data and other operations. In this mode, the DBSCAN serial incremental clustering algorithm is the most representative, which enables consistent clustering of incremental and non-incremental calculated results. Compared with other serial incremental clustering algorithms, this algorithm can directly eliminate the adverse effects on the clustering results through data addition and data deletion [2]. At the same time, the large-scale data processing speed can also be further improved by batch processing of incremental data.

Besides, K-means incremental mean serial clustering algorithm is also a typical serial clustering algorithm in incremental learning mode. When applied specifically, the algorithm can be divided into two stages, the first stage is to train the traditional K-means algorithm; the second stage is to implement incremental training of K-means algorithm, that is, through the incremental training to achieve the organic combination of the old and new K-means results. After the above two stages of processing, the large-scale data will obtain a good serial clustering effect under the incremental learning mode, in order to significantly improve its computational processing speed.

3.3. Serial clustering algorithm for large-scale datasets in feature transformation mode

In the feature variable mode, the main processing characteristic of serial clustering algorithm for large-scale data sets is to implement the original spatial change processing for large-scale data sets. In specific processing, the typical method is to transform the data dimensions in the large-scale dataset, including the transformation from high dimension to low dimension or from low dimension to high latitude, and then process the large-scale dataset which has completed the dimensional transformation through the parallel clustering algorithm. In this process, reducing the dimensionality of data in large-scale datasets is one of the most commonly used processing methods. The ANIC serial clustering algorithm supported by AutoEncoder (Artificial Neural Network) is the most suitable for the dimensionality reduction of large-scale datasets [3]. The algorithm can first learn the sample features in the large-scale dataset through artificial neural networks to achieve effective integration of low-dimensional sample features, read the data over and over again and dynamically update the way to achieve the centroid clustering, and then re-implement the clustering process on the newly generated centroid clustering samples. In this way, the amount of data processing in large-scale datasets can be significantly reduced, so that they can be processed quickly.

4. Advances in Parallel Clustering Algorithms for Large Scale Datasets

4.1. Introduction to Parallel Clustering Algorithms

The so-called parallel clustering algorithm is a clustering algorithm that uses parallel computing technology to process large-scale data in a multi-processor or multi-core processing environment to enhance its computational processing speed. In specific applications, the main purpose of this algorithm is to shorten the large-scale

data computation time and improve the efficiency of the clustering algorithm through the reasonable application of multi-processor or multi-core computing resources. In terms of the current large-scale data set computational processing, there are three most commonly used parallel clustering algorithms, one of which is the large-scale data parallel clustering algorithm under the adaptive similarity matrix mode, the second is the large-scale data parallel clustering algorithm under the approximate eigenvector mode, and the third is the large-scale data parallel clustering algorithm under the K-means mode.

4.2. Parallel clustering algorithm for large-scale data in adaptive similarity matrix mode

Parallel clustering algorithm in adaptive similarity matrix mode is a typical parallel clustering algorithm in current large-scale data set computational processing. In the process of vectorless graph partitioning of large-scale data, the spacing of data in the large-scale data set should be measured with the help of Gaussian kernel function, so that the result of the measurement will form a connection graph with data similarity characteristics, i.e., similarity graph. Each similarity graph corresponds to a similarity matrix in the large-scale data set [4]. Generally, the process of large-scale data mining needs to process the diverse data, so the similarity graph applied in it should be in the fully connected mode. In this mode, the distance value of the data in the large-scale data set can be used to calculate the edge weights of the undirected graph, and its calculation formula is as follows:

$$A_{ij} = \exp\left(-\frac{d(x_i, x_j)^2}{2\sigma^2}\right) \quad (1)$$

Where, A_{ij} representative data x_i and x_j connected edge weights for both, σ Scale parameter representing the extent of the neighbourhood of the control data, whose basic function is to adjust the degree of dissimilarity between the two data, $d(x_i, x_j)$ representative data x_i and x_j distance between the two.

By the parallel clustering idea, For the data x_i to data x_j dissimilarity, it can be calculated according to the following formula

$$s_{ij} = d(x_i, x_j) / \sigma_i \quad (2)$$

Among, s_{ij} representative data x_i to data x_j dissimilarity, σ_i represents the average of the distances between the current data and its neighbours, its main function is to implement a neighbourhood density description of the current data.

And for the data x_i to data x_j dissimilarity, then it can also be calculated according to formula (2).

From this, the corresponding adaptive form can be obtained for its similar function parameters, and the calculation formula is as follows:

$$A_{ij} = \exp\left[-\frac{d(x_i, x_j)_2}{\sigma^2} \left(1 + \frac{|\sigma_i - \sigma_j|}{\sigma_{\max}}\right)\right] \quad (3)$$

Among, $|\sigma_i - \sigma_j|$ representing the difference in data density, $\sigma_{\max}$ represents the maximum difference in data density, $\frac{|\sigma_i - \sigma_j|}{\sigma_{\max}}$ representing weight factors, its main function is to tune the Gaussian kernel function.

When performing computational processing on large-scale datasets, if the above similarity matrix is used for calculation, the computational workload will be enormous. To solve this problem, we can use adaptive matrices as the basic pattern and introduce a parallel clustering algorithm framework to distribute different data blocks to parallel computing nodes in the framework through data partitioning, in order to achieve parallel computing of data similarity among each node. Afterwards, using the node number as a basis, the similarity between the internal data of the two nodes is calculated sequentially. To reduce the probability of duplicate calculations, for the data in the current numbered node, similarity calculation is only performed with the data from nodes with larger numbers, thus obtaining a similarity matrix $A = \{A_{ij}\}$. Through this matrix, Can establish a regularization matrix L , the formula is as follows

$$L = 1 - D^{-1/2} \cdot A \cdot D^{-1/2} \quad (4)$$

Among, D represents the diagonal matrix, and there are $D = \{D_{ij}\}$, regarding D_{ij} , the calculation formula is:

$$D_{ij} = \sum_j^N A_{i,j} \quad (5)$$

Among, N represents the current node number. If $A_{i,j}$ values of 0 and above, its regularisation matrix will take on a semi-positive definite symmetric form, which suggests that there will be non-zero block-symmetric matrices present within the regularisation matrix in the case where the samples within the local clusters and other local clusters are uncorrelated. In such a case, the eigenvectors corresponding to the smallest eigenvalue with number k in front of the regularisation matrix form a matrix V . V by means of K-means clustering algorithm, the final spectral clustering can then be obtained. In this way, the calculation process of large-scale data sets can be effectively simplified, and its calculation efficiency can be further improved on the basis of ensuring its processing quality.

4.3. Parallel clustering algorithms for large-scale data in approximate eigenvector mode

In the calculation of large-scale data sets, the traditional spectral clustering algorithm is through the first k feature vectors composed of a dimensionality reduction matrix for the calculation of processing, this mode of data computation is still very huge, and can not meet the current large-scale data set calculation and processing needs [5]. In order to solve the above problems effectively, we can implement the low-rank approximation process to the feature vectors by Nystrom sampling method. Assuming that the size of the massive dataset to be processed is n . The number of points randomly sampled during processing is m , in the entire large-scale dataset, the number of data that were not sampled was $(n-m)$. For large-scale dataset similarity matrix, the algorithm is formulated as follows:

$$K = \begin{bmatrix} A & B \\ B^T & C \end{bmatrix} \quad (6)$$

Among, K representing large-scale dataset similarity; A represents the similarity matrix of the individual data sampled; B representing the matrix of similarity between the sampled data and other data; T represents the sampling moment; C similarity matrix representing unsampled data. Since the amount of unsampled data in it is much larger than the amount of sampled data, Nystrom's algorithm can be approximated by implementing an approximation to K through A and E in the actual computation of Eq:

$$E = \begin{bmatrix} A \\ B^T \end{bmatrix} \quad (7)$$

$$K \approx EA^{-1}E^T = \begin{bmatrix} A & |B| \\ B^T & B^T A^{-1}B \end{bmatrix} \quad (8)$$

In this process, the algorithm mainly implements low-rank approximation processing in the form of matrices, which can complete the computational processing of large-scale data with only a small amount of sampled data. This prevents the use of unsampled data in large-scale datasets and simplifies the overall data processing algorithm. However, because the traditional sampling method does not fully consider the importance of different sampled data, it is difficult to ensure the contribution of sampled data to the final clustering results. In order to solve this problem, we can implement data sampling by weighted sampling method, and with the support of regular matrix, complete the eigenvalues of the sampled data and the eigenvectors corresponding to their eigenvalues. In this way, the quality of large-scale data processing can be ensured to the maximum extent possible, while achieving a significant simplification of the process.

4.4. Parallel clustering algorithm for large-scale data in K-means mode

In the process of implementing parallel clustering processing for large-scale datasets, K-means is currently the most simplified parallel computing processing model. Typically, we need to introduce the KD tree (i.e., a data organisation structure used in K-dimensional space) in the parallel clustering computation of large-scale data through this model. With the help of KD tree, the distance between the node and its neighbours can be queried and compared at that time, so that the process of calculating the spacing of each data node can be omitted, thus achieving fast matching of data nodes [6]. The KD tree is used as a structure for storing feature vectors, and the establishment of the KD tree structure for the large-scale data to be processed is completed at the clustering centre, after which each data point and the clustering centre point with the minimum distance from it are grouped by the nearest-neighbour node query function that comes with the KD tree. With the reasonable application of the parallel clustering algorithm, the iterative distance calculation process in large-scale data processing can be maximally omitted to achieve significant optimisation of

the overall clustering process, thus further saving the processing time of large-scale data sets and making the processing system run more smoothly.

5. Conclusions

In summary, clustering algorithm is an important branch algorithm in modern artificial intelligence machine learning techniques. In the computational processing of the current large-scale data sets, through the reasonable application of clustering algorithms, not only can effectively ensure the quality of its processing, but also can further improve its processing speed, to ensure the smoothness of the data processing system. In order to realise the good application of clustering algorithms and meet the actual processing needs of large-scale data sets, we need to take the traditional clustering algorithms as the basis, combined with the current development of big data and intelligent technology, and study the computational processing of large-scale data sets under the conditions of serial clustering algorithms and parallel clustering algorithms. Through the reasonable selection of different serial or parallel clustering algorithms, the speed of large-scale data processing can be further improved to adapt to the current development needs of the big data era.

Conflicts of Interest

The author declares that there is no conflict of interest regarding the publication of this article.

Funding

This work was supported by the Chongqing Education Commission “Research on incremental network machine learning algorithms for large-scale datasets”(NO.KJZD-K202402002).

References

- [1] Beijing Faber Hongye Technology Development Co. A large-scale data key feature extraction method, electronic device and medium: CN202411261799.X[P]. 2024-10-18.
- [2] Shandong Lusoftware Digital Technology Co. A parallel chunking method for power grid vector slicing based on clustering algorithm: CN202410854096.1[P]. 2024-10-11.
- [3] Guangxi Normal University. Blockchain-based heterogeneous clustering method for federated data: CN202410694355.9[P]. 2024-10-11.
- [4] Qian, Y.H.; Wang, J.X.; Zheng, X.T. An airborne LiDAR point cloud monowood segmentation method combining spectral clustering and particle swarm improved K-means clustering. *Journal of Geo-Information Science*, 2024, 9, 2177-2191.
- [5] Wang, K.T. Screening of anomalous data points of wind turbines based on density clustering algorithm. *Electronic Design Engineering*, 2024, 18, 127-131.
- [6] Southeast University. Data lake large-scale table clustering method, storage medium, and electronic device based on data relationship graph: CN202410529360.4[P]. 2024-07-26.



© 2024 by the author(s); licensee International Technology and Science Publications (ITS), this work for open access publication is under the Creative Commons Attribution International License (CC BY 4.0). (<http://creativecommons.org/licenses/by/4.0/>)