Mathematical Modeling of Data Similarity Recognition Based on Ordered Clustering Equations

Yajing Pan

School of Mathematics, Hangzhou Normal University, Hangzhou, China

Keywords: Ordered clustering equation; Data similarity recognition; mathematical modeling

Abstract: In the information age, data has permeated every aspect of our lives, forming a vast ocean of data. How to discover potential valuable information from these data is an important research direction in the field of data science. Especially in the network environment, the complexity of data is becoming increasingly prominent. The mixed existence of structured, semi-structured, and unstructured data makes data processing and analysis exceptionally complex. Meanwhile, the possible similarity between data blocks also increases the difficulty of identifying data similarity. To effectively address this challenge, this paper proposes a mathematical model for data similarity recognition based on ordered clustering equations. This model aims to discover the potential similarity between data through in-depth analysis and processing of data, and then conduct orderly clustering. Through this model, we can further improve the utilization of data, reduce data redundancy, and better mine the potential value of data. At the same time, this model can also provide strong support for other data science research and promote the development of data science.

1. Introduction

In the information age, every corner of human life is surrounded by data [1]. Every day, countless amounts of data emerge from various sources, whether it's interactions on social media, transaction records on e-commerce platforms, or sensor data collected by smart devices [2]. These data contain enormous value, however, at the same time, the explosive growth of data also brings many challenges [3]. The massive amount of data not only increases the redundancy of the network's central space, making the system's operational burden heavier, but also prolongs the access time to network target information [4]. How to discover and extract valuable information from massive amounts of data has become a powerful guarantee for the continuous progress of modern society. Clustering, as an important data analysis method, can divide data within a dataset into multiple groups based on fixed similarity metrics. The data within these groups have high similarity, while there are significant differences in data between different groups. The application range of clustering methods is very wide, from image segmentation, text clustering to recommendation systems, and its presence can be seen [5].

However, traditional clustering algorithms mostly focus on detecting and recognizing salient features, while processing non salient features is relatively rare [6]. This makes it difficult for traditional algorithms to meet practical application needs when facing the increasingly complex application scenarios [7]. The study of data features can be divided into two categories: salient features and non salient features. Significant features are usually more intuitive, easy to recognize and extract, and therefore have been widely studied and applied in traditional data processing and analysis. However, although non salient features are not as obvious as salient features, they also contain important information and may even be a key source of information for certain application scenarios. Therefore, how to effectively extract and utilize non salient features has become a research hotspot in the field of data science. The method of identifying similar data is an important means of measuring the degree of similarity between data. It calculates the similarity between a given pair of data sequences. This method has significant application value in the field of information science, as it can help us quickly and accurately find other data similar to the given data, thereby extracting valuable information [8].

Traditional methods for identifying similar data often have low efficiency when dealing with

  DOI: 10.25236/icmmct.2024.028

massive amounts of data, making it difficult to meet the needs of practical applications. Therefore, how to design an efficient and accurate method for identifying similar data has become an urgent problem in the current field of data science. To address the aforementioned issues, this paper proposes a mathematical modeling method for data similarity recognition based on ordered clustering equations. This method first uses the ordered clustering equation to divide the data in the dataset into multiple groups, and then calculates the similarity of the data in each group. Through this method, not only can the redundancy of data be effectively reduced and the utilization rate of data be improved, but other data similar to the given data can also be quickly found. In addition, this method can be flexibly adjusted and optimized according to actual application needs to adapt to different application scenarios.

2. Data Preprocessing

In the digital information age, there is a massive amount of data stored in the central space of the network, which is often disturbed by external factors during transmission and storage, resulting in a large amount of Gaussian white noise mixed in [9]. Gaussian white noise, as a random signal, follows a uniform distribution of power spectral density and exhibits white characteristics in both time and frequency domains, i.e. no correlation [10]. The impact of this noise on data is multifaceted. It not only reduces the convergence speed of the data, but also generates distorted signals by randomly adding or eliminating frequency domain information, thereby affecting the accuracy and credibility of the data. For data similarity recognition, the presence of Gaussian white noise undoubtedly increases the difficulty of recognition. Because similarity recognition typically requires comparing subtle differences between data, the interference of Gaussian white noise may mask these differences, leading to biased recognition results. Therefore, before conducting data similarity recognition, it is necessary to preprocess the data to remove Gaussian white noise from it.

Among numerous denoising methods, wavelet technology is widely used in the field of data denoising due to its excellent time-frequency analysis characteristics. Wavelet transform can decompose signals into a series of superposition of wavelet basis functions, which have localization characteristics in both time and frequency domains, and can effectively capture local features of signals. By selecting appropriate wavelet basis functions and threshold processing strategies, Gaussian white noise can be separated from useful signals, thereby achieving data denoising processing. The use of wavelet technology to remove data noise and then perform data similarity recognition can effectively improve the accuracy and credibility of recognition. Because the denoised data is closer to the true value and the differences between them are more obvious, it is beneficial for similarity recognition algorithms to accurately capture these differences. Meanwhile, denoising can also reduce data redundancy, improve data utilization, and provide a more reliable data foundation for subsequent data analysis and mining.

When processing layered and neatly sorted raw data containing Gaussian white noise, a method combining discrete robust filters and improved wavelet packet decomposition can be used to effectively suppress noise and reconstruct the original signal. Add a filter sufficient to reconstruct Gaussian white noise to the original data using a discrete robust filter. Discrete robust filters can provide effective estimation of signals in the presence of noise. By adding appropriate filtering, it is possible to reduce the interference of noise on the data while maintaining the main characteristics of the signal. The mathematical expression for adding filtering to a discrete robust filter is as follows:

$$K = f \times 2|\omega|^{-1} \tag{1}$$

In the formula, f represents the filtering frequency; ω represents the amount of original data involved in noise reduction.

3. Data Similarity Recognition Model

In a multi-dimensional ordered space, the clustering effect of data is influenced by the

granularity of grid partitioning. Grid granularity, which refers to the size or thickness of grid partitioning, plays a crucial role in the quality and efficiency of clustering results. Choosing the appropriate grid partitioning granularity is an important step in optimizing clustering performance. When there are too many grid divisions, each grid cell may contain less data, which may lead to data loss or information omission. Due to insufficient data volume, clustering algorithms may find it difficult to accurately capture the intrinsic structure and features of data, thereby affecting the accuracy of clustering. On the contrary, if the number of grid divisions is too small, the amount of data contained in each grid cell may be too large. In this case, the data within the grid may have high similarity with the data in the original data space, making it difficult for clustering algorithms to distinguish between different categories of data. In order to determine the most suitable grid partitioning granularity, it is necessary to comprehensively consider the quantity and distribution of data in a multivariate ordered space, as well as the characteristics of clustering algorithms. Divide the multi-dimensional ordered data space grid using the scale parameter ξ, and the scale parameter expression is as follows:

$$\xi = N / k \tag{2}$$

In the formula, N and k represent the number of samples and the number of clusters, respectively.

The data similarity recognition model is composed of a point cloud classification network and an ordered clustering equation. By adding temporal and spatial similarity judgments to the equation, the model can more accurately capture the internal connections and changes between data points. This dual dimensional similarity determination can provide more comprehensive data feature information, which helps to more accurately divide data point groups. The formula for clustering feature vectors in the model using the ordered clustering equation is as follows:

$$W = \frac{X \cdot \sigma^2}{\sqrt{E_n}\sqrt{E_m}} \tag{3}$$

In the formula, E_n represents the number of neurons in the point cloud classification network; E_m represents the clustering fault tolerance of similarity data; σ^2 represents the Euclidean distance of neighboring neurons.

4. Experimental Result

The clustering convergence curve is an important indicator used to evaluate the convergence speed and effectiveness of clustering algorithms at different iterations. An ideal clustering convergence curve should show that the algorithm reaches a stable state with fewer iterations and the internal consistency of the clustering results is high. Figure 1 shows a comparison of clustering convergence curves for different models. It can be seen from the figure that our model achieved rapid convergence within 20 iterations. Compared to traditional models, our model achieved stable clustering results in a shorter amount of time. Fast convergence not only improves the efficiency of the algorithm, but also reduces the consumption of computing resources and time. The model in this article achieved fast convergence in a few iterations and achieved a high intra class distance in the lower approximation, demonstrating good clustering convergence performance. These advantages make the model in this article highly practical and performance in clustering tasks.

Figure 2 shows a comparison of the time consumption of different models in identifying experimental objects of different scales. It can be clearly seen from the graph that the model in this article takes no more than 0.8 seconds to identify data of different scales. This result indicates that the model proposed in this paper has high efficiency in data similarity recognition, and its performance is not affected by the size of the data. For large-scale datasets, traditional similarity recognition models have significantly increased time consumption due to increased computational

complexity. However, the model proposed in this paper can complete the similarity recognition of all data in a relatively short time, indicating that the model has high efficiency and good adaptability to storage environments in data similarity recognition. It can complete similarity recognition of data of different scales in a relatively short time, and is less affected by the size of the data. This makes the model in this article have significant advantages and potential in practical applications.

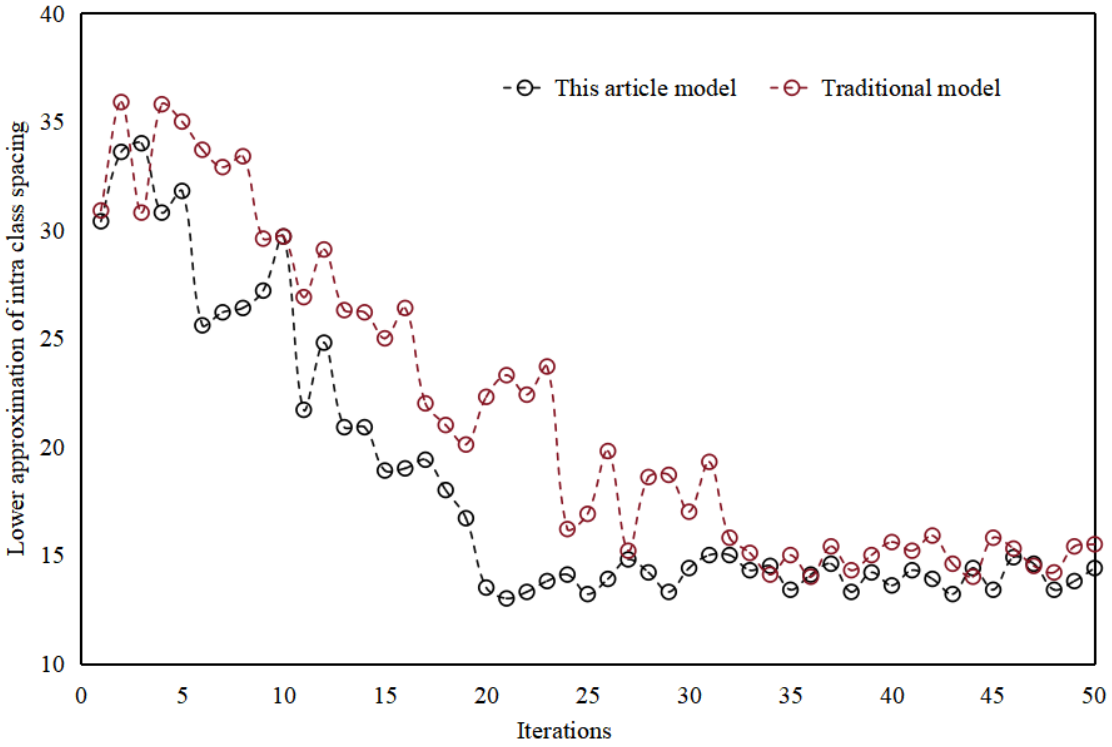

Figure 1 Comparison of clustering curves

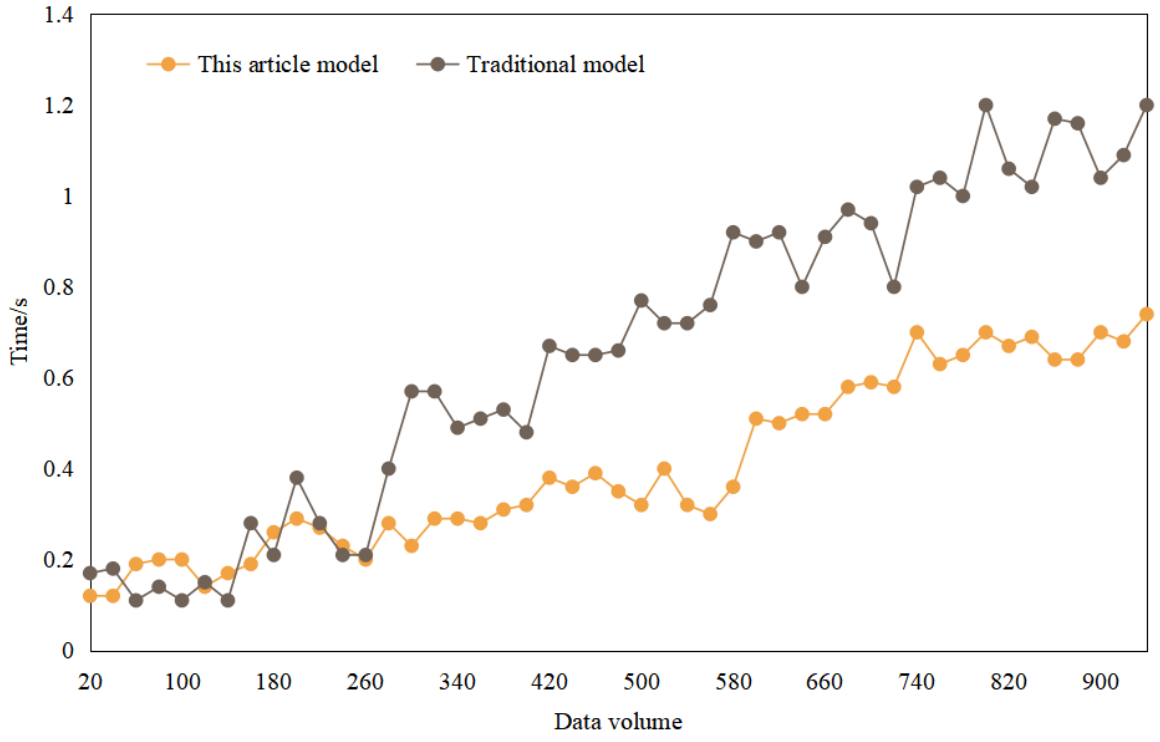

Figure 2 Recognition time of different models

5. Conclusions

Ordered data, due to its complex structural characteristics, increases the difficulty of data clustering. In practical applications, ordered data often has multi-level and multi-dimensional characteristics, which makes traditional clustering methods difficult to directly apply. In addition, with the rapid growth of network data, the burden of processing large amounts of data by computers continues to increase. How to effectively process and analyze this data has become an urgent problem to be solved. In order to avoid the problems caused by high data redundancy, this paper proposes a mathematical model for data similarity recognition based on ordered clustering equations. This model combines ordered clustering equations and data similarity recognition techniques, aiming to improve clustering effectiveness and efficiency. By using the ordered clustering equation, the model can more accurately capture the ordered relationships and similarities between data, thereby achieving more accurate data clustering. At the same time, the model also considers the temporal and spatial dimensions of the data, making the clustering results more comprehensive and accurate. The experimental results show that the model proposed in this article has good performance in clustering ordered data. This model can accurately capture the ordered relationships and similarities between data, achieve fast convergence and efficient clustering, and has important practical value and significance for processing and analyzing large-scale ordered data.

References

[1] Yu Chunyan, Zhang Yumei. Mathematical modeling of data similarity recognition based on ordered clustering equation [J]. Computer Simulation, 2023, 40(7):514-518.

[2] Yang Shiwen, Yang Mingjing, Liu Weijing. Research on similarity action recognition method based on unbalanced data [J]. Information Communication, 2018, 000(005):11-14.

[3] Ye Baoxuan, Zhang Juan, Liu Heng, et al. Method of identifying the phase sequence of substation based on voltage similarity clustering [J]. Electromechanical Engineering Technology, 2022(008):051.

[4] Li Lei. Similarity identification method of power industry training projects based on data association analysis [J]. China Science and Technology, 2023(20):117-119.

[5] Zhang Yuan, Zhang Huijun. Simulation of accurate identification of data similarity based on ordered clustering equation [J]. Computer Simulation, 2023, 40(4):402-406.

[6] Han Fucai, Xu Xun, Ma Wei. Spatial and temporal characteristics of urban air pollutant diffusion based on data mining [J]. Environmental Science and Management, 2023, 48(4):50-54.

[7] Xu Lijuan, Ye Shitong. Optimization of SOM clustering algorithm in non-salient feature data mining [J]. Computer Simulation, 2023, 40(9):497-501.

[8] Kang Zeyu, Kang Fangyuan. Application of additive product risk rate model in clustering failure time data [J]. Applied advances in mathematics, 2022, 11(2):9.

[9] Li Senping, Feng Jianxing. Study on characteristics identification of construction dust pollution based on multi-source data fusion [J]. Environmental Science and Management, 2022, 47(4):5.

[10] Kang Yaolong, Zhang Jingan, Feng Lilu. Simulation of scheduling algorithm for big data clustering center based on constraint satisfaction [J]. Computer Simulation, 2020, 37(3):5.