Hindawi
Security and Communication Networks
Volume 2023, Article ID 9809630, 1 page
https://doi.org/10.1155/2023/9809630

Retraction

Retracted: Impact of the Validity Analysis Model and Multirelational Data Clustering Based on the Trust Probability

Security and Communication Networks

Received 5 December 2023; Accepted 5 December 2023; Published 6 December 2023

This article has been retracted by Hindawi, as publisher, following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of systematic manipulation of the publication and peer-review process. We cannot, therefore, vouch for the reliability or integrity of this article.

Please note that this notice is intended solely to alert readers that the peer-review process of this article has been compromised.

Wiley and Hindawi regret that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] Z. Zhang and M. Talha, "Impact of the Validity Analysis Model and Multirelational Data Clustering Based on the Trust Probability," *Security and Communication Networks*, vol. 2022, Article ID 4852736, 9 pages, 2022.

Hindawi
Security and Communication Networks
Volume 2022, Article ID 4852736, 9 pages
https://doi.org/10.1155/2022/4852736

Research Article

Impact of the Validity Analysis Model and Multirelational Data Clustering Based on the Trust Probability

Zhongzhen Zhang[1] **and Muhammad Talha**[2]

[1]*Hunan Biological and Electromechanical Polytechnic, Changsha 410127, China*
[2]*Department of Computer Science, Superior University, Lahore, Pakistan*

Correspondence should be addressed to Muhammad Talha; talhashoaibt@yahoo.com

Received 20 April 2022; Revised 16 May 2022; Accepted 19 May 2022; Published 31 May 2022

Academic Editor: Muhammad Arif

In view of the present poor impact of multi relational data clustering, a multirelational data clustering effectiveness analysis model based on trusted probability is created. The construction of the multirelational data clustering trusted probability evaluation model, standardization of the multirelational data clustering trusted probability evaluation index, optimization of the multirelational data clustering trusted probability analysis process, and data clustering processing quality and finally, investigations show that the validity analysis model of multirelational data clustering based on credible probability is more effective in practice and satisfies the research goals entirely.

1. Introduction

With the advancement of information technology, people from all occupations have generated vast amounts of data that may be used in practical applications. As a crucial process of knowledge discovery, data mining technology has gotten a lot of attention in order to identify useable information and knowledge in this area in this huge data [1]. Cluster analysis technology has naturally become a highly active study issue in the area of data mining as a way of data mining. Cluster analysis is an unsupervised machine learning approach that classifies an object collection into numerous categories (clusters) based on similarity. The similarity between items in the same category is high and the similarity between objects in other categories is low. Clustering analysis may be used to find the data's underlying structure and to further study select particular clusters by monitoring their characteristics [2]. It may also be used as a preprocessing step before other algorithms process the formed clusters. Although structured data are often stored in several relational tables in relational databases in many practical applications, most clustering approaches are only applicable to data contained in a single relational table. Although multiple relational tables can be linked or aggregated into a single table, this processing method will not only result in high-dimensional data, but the data points may also be distributed in subspaces of different dimensions after integration, resulting in equal distances between data objects in different dimensions and the significance of distance measurement being lost. It is also challenging to account for the impact of intertable linkages on clustering [3]. For this application, multirelational data clustering is created. However, research into the clustering algorithm for multirelational data has not yielded a viable solution to issues such as the existence of one-to-many relationships between objects, incomplete correspondence information between tables, resulting in each target object being described by information of varying orders, and a loop in the relationship between relational tables in a multirelational data set. Furthermore, a full cluster analysis procedure must still assess the quality of clustering findings after clustering to establish if the results are consistent with the data's internal distribution features, i.e., to confirm the efficacy of clustering results. Also, we employ acceptable and effective approaches to assess and explain the findings to assist data analysts in making decisions. The method of categorizing related targets

is known as cluster analysis [4]. The idea is to uncover the data's fundamental structure and look for patterns. Cluster analysis is an unsupervised classification procedure, which means there is no pre-determined class identification, which is the most significant distinction between that and the classification. The following is a quick overview of clustering analysis algorithm classification, focusing on clustering effectiveness indicators.

2. Clustering Validity Analysis Model of Multirelational Data

2.1. Reliability Probability Evaluation Model of Multirelational Data Clustering. Cluster analysis is an unsupervised classification process. Its purpose is to divide the target objects into a series of meaningful groups, so that the targets in each group are "similar" or "close" as much as possible, and the targets in different groups are "different" or "far away" as much as possible. Thus, it is helpful to find the distribution pattern of targets and the relationship between targets. The problem of evaluating cluster quality is called cluster validity analysis [5]. Some typical classification algorithms are introduced, and a new clustering effectiveness analysis index is proposed. For the research of credible probability, the focus of work has always been the acquisition of electronic data and the analysis of the obtained data. The standardized operation process is generally divided into the following: identifying the types of available information and the acquisition methods [6]. Multirelational data mining integrates inductive logic programming, relational database, KDD, machine learning mining and other technologies, studies the representation method of relational data, uncertain reasoning algorithm and learning algorithm, and solves the complex problems in various fields in the real world. The content of trusted probability mining for multirelational data is shown in Figure 1.

Multirelational data mining methods can be divided into two categories: converting multirelational data into single relational data and applying traditional data mining methods. There are two ways to convert multirelational data into single relational (single table) data: establish a full name relationship, add all data to a table, and then form single relational data. The other way is to create new attributes in the central relationship [7]. This technique has the benefit of being able to use current data mining tools directly. However, difficulties such as data size expansion, probable data loss, data duplication, and so on may arise during the shift from a multirelationship to single relationship. The content of this paper is organized according to the practical application steps of cluster analysis. The reliability probability evaluation process of cluster analysis relationship is shown in Figure 2.

For the clustering process, two multirelational clustering algorithms are studied to address several problems existing in multirelational clustering. Then, a formal result verification method is studied, aiming at the problems existing in the existing evaluation indexes of clustering results. Finally, in order to make the clustering results effectively provide corresponding support for decision analysis, a clustering

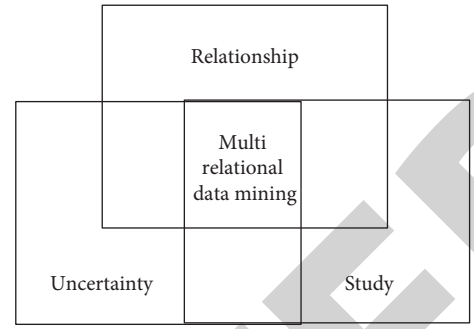

Figure 1: Content of trusted probability mining for multirelational data.

influence factor analysis method based on one-way ANOVA is studied [8]. Based on the research background and significance of this topic, this paper deeply analyzes the research status of clustering algorithm, focuses on the research status of multirelational clustering algorithm, and also analyzes the evaluation methods and analysis interpretation methods of clustering results in detail. Finally, the paper introduces the research content and organizational structure arrangement. Chapter 2 studies the constructing hierarchical multirelationship clustering algorithm based on the IDEFIX model [9]. First, the problems in multirelational clustering are attributed to the characteristics of different connections in the database physical model. On this basis, an association hierarchy model of multirelational data set based on the def X model is proposed. Then, the influence of each connection in the model on the transmission of clustering results and the method of clustering result transmission are defined. Then, a new multirelational clustering algorithm is proposed, so that the algorithm can make full use of the original information embodied in each table: at the same time, there may be multiple association paths between any two tables, or the relationship between tables may form a directed loop, and a path selection algorithm for transmitting clustering results is proposed, which can obtain a clustering result transmission path without loop and keep the relationship between tables as much as possible.

2.2. Trust Probability Evaluation Algorithm for Cluster Analysis of Multirelational Data. The process of judging the clustering results is called clustering effectiveness analysis. Generally speaking, the clustering that minimizes the intraclass distance and maximizes the interclass distance is the optimal clustering [10]. There are usually three categories of criteria:

(1) External criteria: Based on the known data structure, test the consistency between the clustering results and the known classification.

(2) Internal standard: Test the clustering results only according to the amount of data and the internal characteristics of data.

(3) Relative standards: The above two types of standards are based on statistical tests, requiring a large amount of calculation. The relative standard does not require

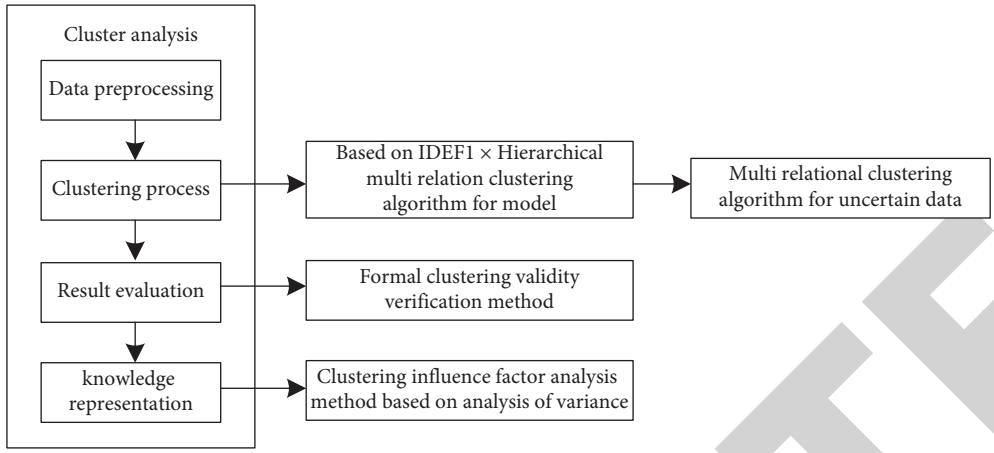

Figure 2: Evaluation process of reliability probability of the cluster analysis relationship.

statistical test, and its basic idea is to find the best aggregation method according to the predefined standard [11]. This paper presents an effectiveness index based on the K-means algorithm. The following formula can describe the main idea of K-means clustering algorithm.

$$\text{Minimize } J(X,U,V) = \sum_{i=1}^{k}\sum_{i}^{n} u_{ij}\left\|x_j - v_i\right\|^2, \tag{1}$$

where n is the number of data in a given dataset and K is the number of clusters. $X\ \{X_1, X_2, X_3, \ldots X_n\}$ Cr is the given data set. P and Cr are the center points of the class. $X_1, X_2, X_3, \ldots, X_n$ represent K classes, n represents the number of data in X, $u = (U)$ is the clustering matrix, which is composed of the membership relationship between X and class k, $d_{xy} = \|XY\|\ XY \in R$ is a distance function (e.g., Euclidean distance). To minimize JX, u, 1), class center point ($w_i = 1,2, \ldots, K$). Membership matrix U, the following iterative formula shall be used to calculate step by step:

$$u_i = \begin{cases} 1; \left\|x_j - v_i\right\| \leqslant \left\|x_j - v_h\right\|, h = 1,2,\ldots,k, h \neq i, \\ 0. \end{cases} \tag{2}$$

Further available:

$$v = \frac{\sum_{j=1}^{n} u_{ij}x_j}{\sum_{j=1}^{n} u_{ij}} = \frac{\sum_{x\in X} x_j}{n_i}. \tag{3}$$

In this paper, the following effectiveness indicators are proposed to define the clustering effectiveness function:

$$V_{\text{km}}(k) = \frac{\text{Intra}(k) + \text{Inter}(k)}{\text{Inter}(k_{\max})}, \tag{4}$$

where $k_{\max}$ is the maximum number that can be clustered.

$$\text{Intra}(k) = \frac{\left(1/k_{i=1}\right)^k \|\sigma(v)\|}{\|\sigma(X)\|}, \tag{5}$$

where

$$\begin{aligned} \|x\| &= \left(x^x x\right)^{1/2}, \\ \overline{x} &= \frac{1}{n}\sum_{F=1}^{n} x_j, \end{aligned} \tag{6}$$

$$\begin{aligned} \sigma(X) &= \left\{\sigma(X)', \sigma(X)^2, \ldots, \sigma(X)^r\right\}, \\ \sigma(X)' &= \frac{1}{n}\sum_{j=1}^{n}\left(x_j^h - \overrightarrow{x}^n\right)^2, \\ \sigma(v_i) &= \left\{\sigma(v_i)', \sigma(v_i)^2, \ldots, \sigma(v)^s\right\}, \\ \sigma(v_i)^k &= \frac{1}{n_i x \in X}\left(x_k^h - v_i\right)^2. \end{aligned} \tag{7}$$

Signal collection equipment collects a huge number of transmission signals, and the normal disturbance signals are extracted using modal decomposition and disturbance signal feature extraction. Signal detection may be made more precise by using the Fourier transform processing of the signals. The identification issue of many disturbance signals may be solved using the constrained fuzzy clustering approach. In order to illustrate the effectiveness of the new index, this paper makes a practical test using the data in the Iris data set. This set of data is 150 biometric data on three kinds of flowers. Input parameters: $K_{\min} = 2$, $k_{\max} = 10$, $n = 150$ (number of data), $s = 4$ (data dimension), termination condition = 0 (= 0 means that the data at the center point will not change as the termination condition). The calculation results are shown in Table 1.

We determine the trust probability according to the interaction relationship and function definition, establish the trust probability session, define the interaction protocol, create its internal functions, and instantiate it. The specific cooperation process is not described in detail in this paper. Information entropy is a measurement method used to measure system uncertainty in information theory [12]. The more unclear an attribute value's value is in a system, the more chaotic the system becomes. The higher the system's information entropy under this characteristic, the less

Table 1: Calculation results of $V_{\mathrm{km}}(K)$.

Number of clusters (K)	Effectiveness index $V_{\mathrm{in}}(K)$
2	0.09732
3	0.07498
4	0.12658
5	0.23598
6	0.42352
7	0.56582
8	0.63528
9	0.49853
10	0.46852

information it offers and the less valuable the attribute becomes. On the contrary, the more ordered the system is, the less the uncertainty programme of the value of an attribute. The lower the information entropy under the property, the more information is delivered, and the more important the characteristic becomes. Each signal is classified and processed using constrained fuzzy clustering based on its fuzzy similarity. Throughout the clustering procedure, the signal transitivity is assumed and conveyed from the fuzzy matrix to the fuzzy equivalent matrix. Despite its low similarity, this transitivity has more limitations than standard recognition techniques, allowing the disrupted signal to be detected well under fuzzy constraints. Information entropy has been extensively employed in cluster analysis, outlier identification, uncertainty measurement, and so on as an efficient measurement technique [13]. The information entropy within and between classes is used to measure the importance of each attribute in the clustering process. Due to different calculation methods of description entropy, it is described below, respectively [14]. An information entropy that can measure continuous random variables is proposed, which is called trusted probability entropy. Assuming that the probability density function of continuous random variable x is $f(x)$, the trusted probability entropy of the random variable is defined as follows:

$$H_R(x) = \frac{1}{1-\alpha}\ln \int f^{\alpha}(x)\mathrm{d}x, \quad \alpha > 0,\ \alpha \neq 1. \tag{8}$$

The intraclass entropy given by the above definition reflects the uncertainty of data distribution of a class under different attributes in the clustering division results, that is, in a class, if the intraclass entropy under an attribute is smaller, the uncertainty of the attribute in the class is smaller, and the weight of the attribute is larger in the clustering process [15].

2.3. Implementation of Clustering Validity Analysis of Multirelational Data. The benefits of using trustworthy probability correlation analysis to examine the degree of correlation between many data series are numerous. By examining the similarity and closeness of plane geometry between system behavior characteristic data series and related factor data series curves, credible probability association analysis determines if the link between system behavior and related factors is close [16]. It does not require large sample data, and the data variables do not require to conform to typical probability distribution characteristics. Credible probability correlation analysis makes up for the shortcomings of classical statistics and mathematical statistics. In practical application, the system's behavior characteristic data series and related factor data series mostly appear in the form of panel data. For example, when analyzing the science and technology input-output system, the fund input factors can select the government financial science and technology fund input index, enterprise science and technology fund input index, and social fund-raising fund input index to describe their behavior characteristics. Moreover, each index has different observation values at different times, which involves the correlation analysis between the system behavior characteristics and related factors based on panel data [17]. The usual analytic approach computes the correlation degree between each data sequence of connected factors and the data sequence of behavior factors separately, then calculates the average correlation degree as the correlation degree between relevant factors and behavior components. This technique overlooks the data's underlying relationships, resulting in a huge inaccuracy that cannot represent the overall connection between system behavior characteristics and important variables. For the correlation analysis of data sequence described by panel data, first, an index system reflecting the system behavior characteristics and relevant factors should be established. Generally, multiple data sequences are selected as the mapping quantity reflecting the system behavior and related factors, and the mapping quantity is used to indirectly represent the system behavior and related factors. Then, the multidata sequence describing the system behavior and related factors is divided into principal components. The multi data sequence is transformed into a single data sequence, and the original multidata sequence information is saved [18]. Finally, the trusted probability correlation analysis method is used to analyze the correlation between the system behavior characteristics and the data sequence of related factors, and the trusted probability correlation order is obtained. To achieve the trusted probability correlation analysis of the data sequence of system behavior characteristics and related variables under panel data, the trusted probability correlation order is utilized to indicate the effect of related factors on the system behavior. After the dimensionality reduction index processing of principal component analysis, the multi-index data mapping sequence of factors is transformed into a single component component comprehensive score data sequence. The principal component comprehensive score data sequence saves the information of the original multi index data sequence [19]. The method of credible probability correlation degree analysis can correlate and divide the factors of a single data sequence. The principal component comprehensive score sequence of the system behavior sequence can be selected as the parent sequence, which is recorded as follows.

$$Z_0 = \left(z_0(t_1), z_0(t_2), \ldots, z_0\left(t_p\right)\right). \tag{9}$$

The principal component comprehensive score sequence of relevant factors is selected as the subsequence and recorded as follows.

$$Z_i = \left(z_i(t_1), z_i(t_2), \ldots, z_i\left(t_p\right)\right). \tag{10}$$

The correlation coefficient at time is as follows.

$$r\left(z_0(t_k), z_i(t_k)\right) = \frac{\min_{i\le i\le n}\min_{1\le k\le p}\left|z_0(t_k) - z_i(t_k)\right| + \rho\max_{1\le i\le n}\max_{1\le k\le p}\left|z_0(t_k) - z_i(t_k)\right|}{\left|z_0(t_k) - z_i(t_k)\right| + \rho\max_{1\le i\le n}\max_{1\le k\le p}\left|z_0(t_k) - z_i(t_k)\right|}. \tag{11}$$

The above formula reflects the system behavior data sequence and related factors as (I = 1,2, . . ., n) the greater the correlation degree of the data sequence, r (z_0, z_i), the greater the correlation degree between the relevant factor s and the system behavior, and the greater the impact on the system [20]. Therefore, the order relationship between the correlation degree of the relevant factor and the behavior factor is determined. In the research of evolutionary data clustering, the smoothness of clustering results is an important index to judge whether a clustering algorithm is excellent. The smaller the degree of clustering change, the smoother the clustering results. The algorithm under the ed-pcm framework is still valid because its definition is independent of the time regular term. The definition is repeated below:

$$\mathrm{DCC}_{t,t-1} = k - \sum_{i=1}^{k}\sum_{j=1}^{k}\frac{\left|V_{ij}\right|}{\left|V_{i,t}\right|\cdot\left|V_{j,t-1}\right|}. \tag{12}$$

Given the computational difficulty, the present clustering division must be applied to the data at the historical period by the trusted probability. Each historical time element should be handled to represent the data's details at all historical periods. Ed-pcm just needs to compare the present clustering with the history clustering division. The amount of calculation is small, discarding the details and focusing on the big picture. Therefore, compared with the ed-pcm framework, the trusted probability framework is slightly lower in efficiency, but generally can reflect more data details. If the formally processed data DS = formalize (d) {D_1, D_2, . . ., D}, then DS is the starting set of the reasoning process. The intermediate node set of the reasoning process is represented by D_1 (where it represents the step in the reasoning process and j represents the node in the step. But I is not strict, because the node may be used many times in reasoning), terminate the node with DT = {D_1, D_2, . . ., D}. The schematic diagram of reasoning process is shown in the figure. The square point representing the transition node is omitted in the figure, and the transition is represented above the connecting line. A few nodes and steps only represent it, and the actual process may be much more complex than the diagram. The message analysis flow of clustering data relationship features is shown in Figure 3.

Clustering integration uses ensemble learning technology to obtain a better and more robust clustering result by learning multiple base clustering divisions of the fusion data set. Given a set of clustering results, the purpose of clustering integration is to find a final cluster and make it as consistent as possible relative to all input clustering results. Clustering integration is to use multiple clustering results to find a new data partition, which shares the clustering information of all input clustering results on the data set to the greatest extent. As a result, clustering integration gives the following benefits over a single clustering algorithm: (1) The final clustering results are less affected by noise and outliers and have good robustness due to the integration of information from multiple clusters; (2) the processing of irregular shape data has good performance and can deal with nonlinear clustering problems; (3) for large-scale data sets, appropriate base clustering is used, which has good scalability; and (4) clustering integration is now extensively employed in disciplines such as biological data, medical diagnostics, computer vision, and network data analysis. The schematic diagram of clustering integration process is shown in Figure 4.

In cluster integration, the accuracy and difference of base clustering are the key factors affecting the integration results. If there is no difference between base clusters in cluster integration, the effect of cluster integration is not obvious. Therefore, the difference between base clustering members is an important factor in determining the final integration effect. It is actually the first step of cluster integration that helps to produce the multiple base clustering that results in high accuracy [21].

3. Analysis of Experimental Results

To comprehensively evaluate different algorithms, the Friedman test is applied to the experimental results. Under different evaluation indexes, the clustering results of the proposed new algorithm and K-prototypes and K-centers clustering algorithm on numerical data are shown in Table 2.

The information of classified data sets used in the experiment is shown in the table. All these data sets contain supervised label information. However, class label information is not used in the process of clustering or cluster integration selection and is only used to evaluate the final clustering results. Because the bsefcm algorithm does not get experimental results on some data sets, the other algorithms are compared and analyzed. Therefore, there are $a = 5$ algorithms and $B = 160$ combined experimental results. Suppose R represents the ranking of the advantages and disadvantages of the j algorithm in case. For Ca, ARI, and Nm3 evaluation indexes, the larger the index value, the smaller the order value.

On the other hand, the lower the sorting value is, the faster the algorithm runs. The Friedman test calculates the

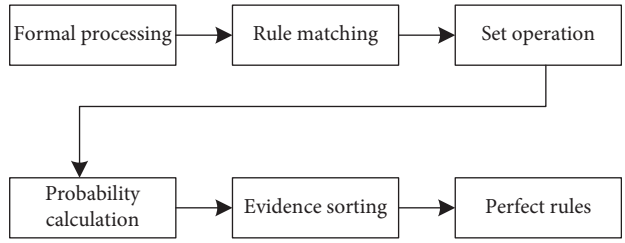

Figure 3: Message analysis flow of clustering data relationship features.

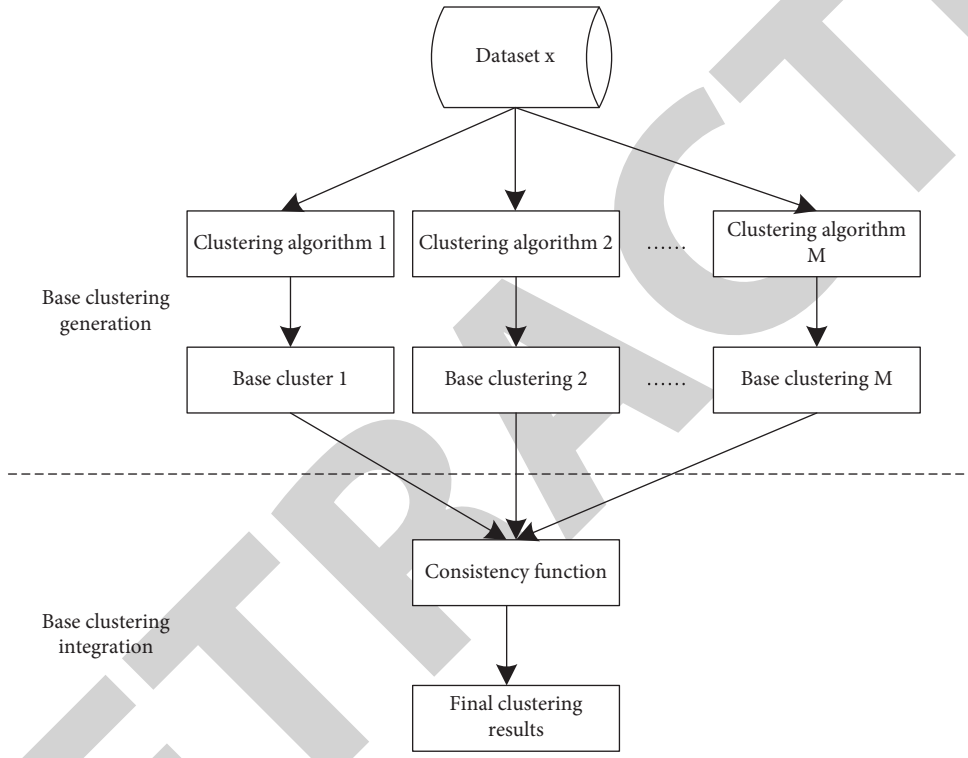

Figure 4: Schematic diagram of data clustering integration process.

Table 2: Comparison of clustering results of numerical data: CA value (mean ± standard deviation).

Data set	K-centers algorithm	Improved K-prototypes algorithm	Paper method
Segment	0.5307 ± 0.0003	0.6035 ± 0.0368	0.5685 ± 0.0043
Waveform	0.5198 ± 0.004	0.5368 ± 0.0001	0.5921 ± 0.0007
Waveform + Noise	0.6432 ± 0.0352	0.5265 ± 0.0000	0.6988 ± 0.0365

average sorting of each algorithm based on the algorithm's average performance in each data set in the table. To further test the performance of the system algorithm proposed in this chapter, the scalability is evaluated by testing the algorithm's running time concerning the number of data objects and the number of features. For this test, we use the composite data generator 9 to generate a set of composite data sets with different numbers of data objects and attributes. The number of data objects ranges from 100,000 to 500,000, and the dimensions vary between 10, 20, 30, 40, and 50. The comparison of analysis effectiveness and running time of data clustering algorithm is shown in Figure 5.

Experiments are carried out on five data sets in the UCI machine learning database using the feature weighted clustering model, FCM and DBSCAN clustering algorithms, and Ig and ReliefF feature weight learning methods. The feature descriptions of each data set are shown in Table 3.

There are no predefined topic categories in clustering. Its purpose is to organize the data into numerous groups. It demands that data in the same category have as much similarity as feasible, while data from separate categories have as little similarity. It is a fully automated data grouping processing procedure. There are several approaches for evaluating clustering findings. At present, clustering validity,

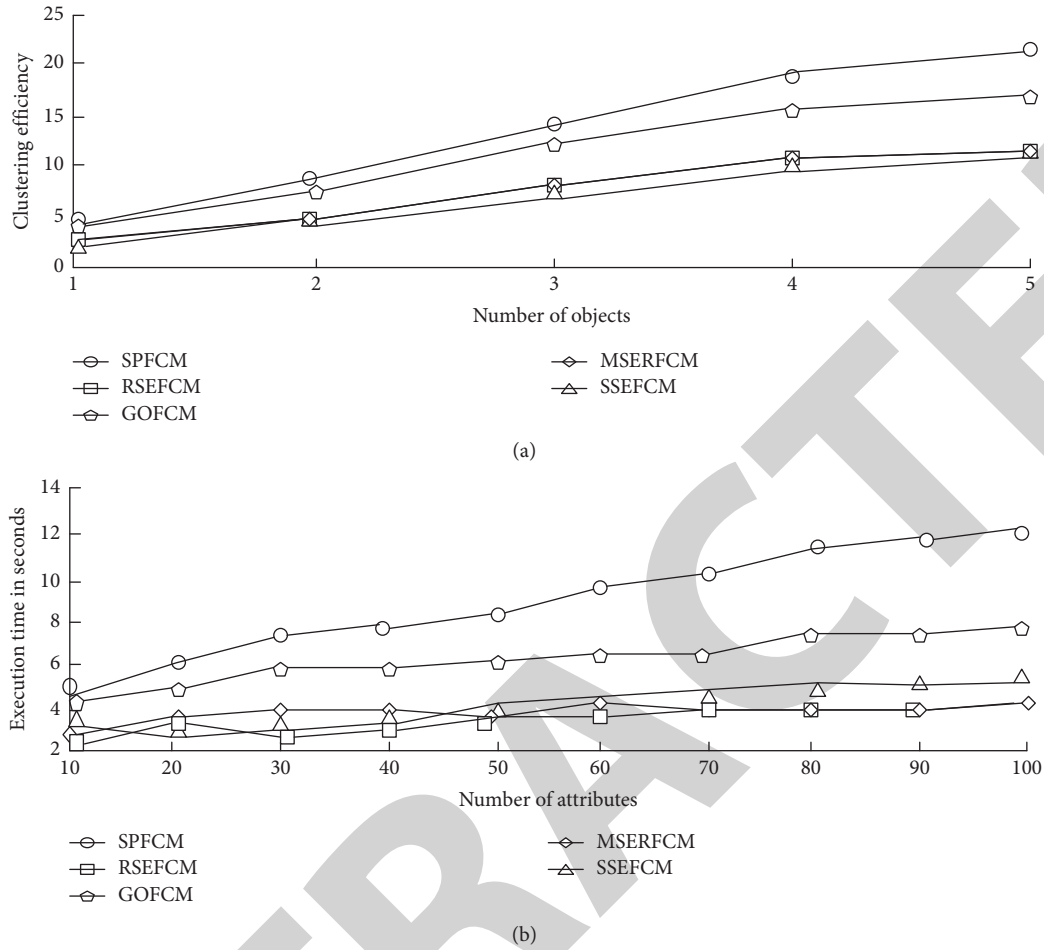

Figure 5: Comparison of analysis effectiveness and running time of the data clustering algorithm. (a) Analysis of validity of data clustering algorithm. (b) Execution time comparison with the increasing number of attributes.

Table 3: Data set characteristics.

Data set name	Number of data	Number of attributes	Number of clusters
Glass	346	7	7
Ionosphere	352	35	3
Iris	151	5	4
Pima	769	9	3
Wine	180	15	4

the function is often used to evaluate the results of different clustering algorithms and the clustering results obtained by the same algorithm under different parameters. In the experiment, first, the data were standardized by formula. Then, the weighted Euclidean distance is used to measure the dissimilarity between samples and between samples and cluster center. In the first clustering, the weight of each Viterbi sign is one.

$$x = \frac{x - \min(X)}{\max(X) - \min(X)},$$
$$d(v_i, x_j) = \sqrt{\sum_{l=1}^{M} w_l \left| x_{jl} - v_{il} \right|^2}. \quad (13)$$

The created random interval data set is subjected to this method's two fuzzy c-means clustering analysis techniques and the interval mean method, with the final findings separated into three groups. The clustering results are compared to the known previous partition, and the difference is measured using the CR index (derived using a formula). Calculate the average value of 60 sets of experimental findings for each instance of random simulation test, and the final average CR index value is displayed in Table 4.

In general, the effectiveness of fuzzy clustering for data set a is significantly better than that for data set *B*. In other words, when the original partition boundary is obvious and there is no overlap between various types, the effectiveness of the trusted probability clustering analysis method is more prominent. While when the original partition boundary is

Table 4: Average CR index value of FCM cluster analysis of interval data.

R1 and R2	Data set A		Data set B	
	Hausdorff distance double matrix method	Interval midpoint method	Hausdorff distance double matrix method	Interval midpoint method
[1, 4]	0.868	0.655	0.735	0.435
[1, 8]	0.878	0.635	0.728	0.455
[1, 12]	0.865	0.603	0.655	0.429
[1, 16]	0.792	0.598	0.539	0.421
[1, 20]	0.768	0.568	0.435	0.386

fuzzy and there is overlap between various types, the effectiveness of the trusted probability clustering analysis method is relatively poor. This is determined by the nature of the sample data set itself and has nothing to do with the specific FCM algorithm. In all cases, the clustering effect of this method is better than the interval mean method.

4. Conclusion

This article begins with an overview of the theoretical foundations of evolutionary data clustering, and then moves on to the research concepts, research content, research techniques, and an analysis and summary of the present research state. The second chapter begins with an overview of data mining and traditional clustering analysis, followed by an examination of traditional clustering algorithms and comparing the benefits and drawbacks of various common clustering algorithms. Finally, an introduction to evolutionary data clustering and an explanation of common smoothing regularization methods and explicit modelling methods. Finally, the experiment's data sets, distributed data set and kdd-cup99 data set, are described. The experiments suggest that the method described in this research is successful, has a broad application potential, and is adaptable.

Data Availability

The data used to support the findings of this study are available from the corresponding author on request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] W. Fan and N. Bouguila, "Spherical data clustering and feature selection through nonparametric Bayesian mixture models with von Mises distributions," *Engineering Applications of Artificial Intelligence*, vol. 94, no. 4, Article ID 103781, 2020.

[2] L. F. Zhu, J. S. Wang, H. Y. Wang, S. S. Guo, and W. Xie, "Data clustering method based on improved bat algorithm with six convergence factors and local search operators," *IEEE Access*, vol. 8, 2020.

[3] B. Nza and A. Nb, "High-dimensional count data clustering based on an exponential approximation to the multinomial Beta-Liouville distribution - ScienceDirect," *Information Sciences*, vol. 524, no. 5, pp. 116–135, 2020.

[4] W. Liang, K. C. Li, J. Long, X. Kui, and A. Y. Zomaya, "An industrial network intrusion detection algorithm based on multifeature data clustering optimization model," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 3, pp. 2063–2071, 2020.

[5] L. T. H. Nguyen, T. Wada, I. Masubuchi, T. Asai, and Y. Fujisaki, "Bounded confidence gossip algorithms for opinion formation and data clustering," *IEEE Transactions on Automatic Control*, vol. 64, no. 3, pp. 1150–1155, 2019.

[6] C. Wang and Z. Zhang, "Rapid identification and simulation of user demand information in wireless network," *Computer Simulation*, vol. 36, no. 4, pp. 392–395, 2019.

[7] A. Zigomitros, F. Casino, A. Solanas, and C. Patsakis, "A survey on privacy properties for data publishing of relational data," *IEEE Access*, vol. 8, Article ID 51071, 2020.

[8] M. Gort, C. Feregrino-Uribe, A. Cortesi, and F. Fernández-Pea, "HQR-Scheme: a High Quality and resilient virtual primary key generation approach for watermarking relational data," *Expert Systems with Applications*, vol. 138, no. 7, Article ID 112770, 2019.

[9] S. Bou, T. Amagasa, and H. Kitagawa, "Scalable keyword search over relational data streams by aggressive candidate network consolidation," *Information Systems*, vol. 81, pp. 117–135, 2019.

[10] S. Link and H. Prade, "Relational database schema design for uncertain data," *Information Systems*, vol. 84, pp. 88–110, 2019.

[11] M. S. Ramada, J. C. da Silva, and P. de Sá Leitão-Júnior, "From keywords to relational database content: a semantic mapping method," *Information Systems*, vol. 88, no. 4, Article ID 101460, 2020.

[12] J. Ktters and P. W. Eklund, "Conjunctive query pattern structures: a relational database model for Formal Concept Analysis - ScienceDirect," *Discrete Applied Mathematics*, vol. 273, no. 8, pp. 144–171, 2020.

[13] A. Andrejev, K. Orsborn, and T. Risch, "Strategies for array data retrieval from a relational back-end based on access patterns," *Computing*, vol. 102, no. 5, pp. 1139–1158, 2020.

[14] F. Grandia, M. Federica, M. Riccardo, and P. Wilma, "Unleashing the power of querying streaming data in a temporal database world: a relational algebra approach," *Information Systems*, vol. 103, no. 13, Article ID 101872, 2021.

[15] V. Puri, S. Sachdeva, and P. Kaur, "Privacy preserving publication of relational and transaction data: survey on the anonymization of patient data," *Computer Science Review*, vol. 32, no. 9, pp. 45–61, 2019.

[16] C. K. Tsung, H. Y. Hsieh, and C. T. Yang, "An implementation of scalable high throughput data platform for logging semiconductor testing results," *IEEE Access*, vol. 7, Article ID 26497, 2019.

[17] M. Vucetic, M. Hudec, and B. Bozilovic, "Fuzzy functional dependencies and linguistic interpretations employed in

knowledge discovery tasks from relational databases," *Engineering Applications of Artificial Intelligence*, vol. 88, Article ID 103395, 2020.

[18] Q. Xu, Q. Zhang, J. Liu, and B. Luo, "Efficient synthetical clustering validity indexes for hierarchical clustering," *Expert Systems with Applications*, vol. 151, no. 10, Article ID 113367, 2020.

[19] M. Kargar, A. Isazadeh, and H. Izadkhah, "New internal metric for software clustering algorithms validity," *IET Software*, vol. 14, no. 4, pp. 402–410, 2020.

[20] J. Fan, W. Xie, and H. Du, "A robust multi-sensor data fusion clustering algorithm based on density peaks," *Sensors*, vol. 20, no. 1, p. 238, 2019.

[21] S. Dong, Z. Sha, and C. Zhu, "Disturbance signal recognition method of power system based on constrained fuzzy clustering," in *Proceedings of the 2019 IEEE 3rd International Electrical and Energy Conference (CIEEC)*, pp. 276–280, Beijing, China, September 2019.