Research Article

Economic Management Data Envelopes Based on the Clustering of Incomplete Data

Shuo Dong ¹ and Sang-Bing Tsai ²

¹*School of Economics, Central University of Finance and Economics, Beijing 10081, China*

²*Regional Green Economy Development Research Center, School of Business, WUYI University, Nanping, China*

Correspondence should be addressed to Shuo Dong; shuodong@email.cufe.edu.cn

Received 27 October 2021; Revised 17 November 2021; Accepted 22 November 2021; Published 3 December 2021

Academic Editor: Zhen Song Chen

Copyright © 2021 Shuo Dong and Sang-Bing Tsai. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

In this paper, the economic management data envelope is analyzed by an algorithm for clustering incomplete data, a local search method based on reference vectors is designed in the algorithm to improve the accuracy of the algorithm, and a final solution selection method based on integrated clustering is proposed to obtain the final clustering results from the last generation of the solution set. The proposed algorithm and various aspects of it are tested in comparison using benchmark datasets and other comparison algorithms. A time-series domain partitioning method based on fuzzy mean clustering and information granulation is proposed, and a time series prediction method is proposed based on the domain partitioning results. Firstly, the fuzzy mean clustering method is applied to initially divide the theoretical domain of the time series, and then, the optimization algorithm of the theoretical domain division based on information granulation is proposed. It combines the clustering algorithm and the information granulation method to divide the theoretical domain and improves the accuracy and interpretability of sample data division. This article builds an overview of data warehouse, data integration, and rule engine. It introduces the business data integration of the economic management information system data warehouse and the data warehouse model design, taking tax as an example. The fuzzy prediction method of time series is given for the results of the theoretical domain division after the granulation of time-series information, which transforms the precise time-series data into a time series composed of semantic values conforming to human cognitive forms. It describes the dynamic evolution process of time series by constructing the fuzzy logical relations to these semantic values to obtain their fuzzy change rules and make predictions, which improves the comprehensibility of prediction results. Finally, the prediction experiments are conducted on the weighted stock price index dataset, and the experimental results show that applying the proposed time-series information granulation method for time series prediction can improve the accuracy of the prediction results.

1. Introduction

Clustering is an unsupervised data mining method. The basic idea is to measure the similarity between the data based on the intrinsic properties of the data and classify the samples with greater similarity into the same class and those with less similarity into different classes [1]. In the field of finance, clustering is widely used in problems, such as personal credit collection and risk identification of listed companies. However, in specific applications, traditional clustering algorithms cannot handle high-dimensional data well. The presence of a large amount of noise and redundant

features makes it very unlikely that clusters exist in all dimensions. As the sample dimensionality increases, the distance difference between the samples becomes smaller, and the data becomes sparse in the high-dimensional space [2]. It is shown that in the processing of high-dimensional data, low-dimensional feature subspaces can approximate the high-dimensional data features. The subspace clustering methods follow this idea and seek to identify the different classes of clusters in different feature subspaces in the same dataset. Since the features of different classes of data may correspond to different feature subspaces, and the feature dimensions composing these feature subspaces

may also be different, it is more difficult to identify the class clusters in the original feature space. The subspace clustering algorithm divides the original feature space to obtain several different feature subspaces and identifies the possible class clusters from the feature subspaces. Many subspace clustering algorithms have been proposed in the existing research for clustering high-dimensional data [3]. However, these algorithms are optimized for a single objective function and adopt a greedy search strategy, and thus, they have the disadvantages of being sensitive to the initial points and easily falling into local optima. Moreover, optimizing multiple objective functions simultaneously can improve the robustness of the algorithm to different data.

Clustering analysis is one of the important techniques in the field of data mining and machine learning and has been widely used in several fields, including information granulation, image processing, bioinformatics, security assurance, web search, etc. The so-called clustering is to divide the sample objects in a dataset into different class clusters, where the sample objects in similar clusters are highly similar, while those in different class clusters are less similar. The role of clustering as an unsupervised learning technique in identifying the unlabeled data structures cannot be ignored [4]. For different division methods of samples, the existing clustering methods can be classified as hierarchical clustering methods, divisive clustering methods, grid-based clustering methods, density-based clustering methods, and other clustering methods. The existing clustering algorithms can also be roughly classified into two main categories: hierarchical clustering methods and divisive clustering methods [5]. Besides, the clustering algorithms can be classified into hard clustering algorithms and soft clustering algorithms based on other classification criteria. After years of research and development, many clustering algorithms have been widely used. Although there are several clustering algorithms in the field of clustering analysis, each algorithm has its unique method for discovering the underlying data structure in a dataset. However, different algorithms processing the same dataset may produce different clustering results, and it is difficult for us to evaluate which clustering result is more consistent with the data structure of that dataset without supervised information [6]. Completely random missing means that the missing values are lost completely at random, and the tendency of the data points to be missing is independent of their hypothetical values and the values of other variables. Random missing means that the missing values are missing because of some observed data and the tendency of the data points to be missing is independent of the missing data, however, it is related to some observed data. Nonrandom missing means that the missing values are not lost randomly but for a reason. Usually, the reason is that the missing value depends on the assumed value or the value of another variable.

Because of the existence of data sets containing missing data, the traditional clustering algorithms are no longer able to deal with these data directly. Therefore, exploring how to solve the clustering problem of incomplete datasets

has become a pressing challenge in the study of cluster analysis. In summary, the study of clustering methods for incomplete data has very wide scientific research value and practical application value. It is an information management technology whose main purpose is to support the management decisions through smooth, rational, and comprehensive information management. A data warehouse is a new application of database technology and not a replacement for the database. Data warehouse and operational database, respectively, undertakes two different tasks of high-level decision analysis and daily operational processing and play different roles. There is a close connection between the data warehouses and real time databases, and the data warehouses require real time databases to provide large amounts of historical data to provide answers, analysis, and prediction results for the various topics required.

The second part of this paper is the research status, the third part is the introduction of the related research algorithm structure, the fourth part is the analysis and explanation of the results, and the fifth part is the conclusion of this paper.

2. Current status of Research

With the rapid development of the internet technology and the improved performance of data storage devices in recent years, a large amount of data is generated and stored in various industries. Among these data, a large portion is time-tagged, i.e., a series of observations recorded in a chronological order, known as time series. How to effectively analyze and process the time series data to uncover potential and valuable knowledge and information to support more efficient production, operation, management, and decision-making activities of the enterprises is one of the important tasks in today's big data era [7]. Traditional time series analysis mainly uses statistical models to analyze and process time series, and with the rapid development of artificial intelligence, time series analysis methods based on data mining and machine learning theory have gradually become mainstream, forming a research branch of time series data mining. In the system sense, time series refers to the response of a system at different times [8]. From the viewpoint of system operation, the definition points out that time series are arranged in a certain order. The "certain order" here can be either a time order or a physical quantity with various meanings, such as representing temperature, velocity, or other monotonically increasing values. The time series is an objective record of the historical behavior of the system under study, which contains the structural characteristics of the system and its operation laws [9]. In summary, the time series has the following characteristics: the data or the position of the data points in the series depends on time, that is, the values of the data depend on the change of time but not necessarily a strict function of time. The values or the position of the data points at each moment have certain randomness, and it is impossible to predict the historical values with complete accuracy. The values or the position of

the data points at the preceding and following moments (not necessarily adjacent moments) have a certain correlation, and this correlation is the dynamic regularity of the system. As a whole, the time series tend to show some kind of trend or cyclical variation [10].

In general, we cannot completely avoid the missing samples, similarly, in the association rule algorithm or decision tree algorithm, the missing data will directly affect the calculation of confidence, support of the frequent itemset, or the selection of the splitting attributes of the tree nodes. Therefore, the handling of missing values plays a crucial role in whether the clustering process can be carried out smoothly [11]. Therefore, the handling of missing values plays a crucial role in the success of the clustering process. The effective processing of incomplete data has become hot research in the field of pattern recognition. The modeling method generally analyzes the patterns of data in the dataset by finding them, establishing a suitable mathematical model, and calculating the corresponding missing attribute filler values by the established model, with the disadvantage that it is only suitable for the datasets of moderate size and with certain patterns. In addition to the above-mentioned methods, in recent years, with the rising popularity of machine learning, many missing value processing methods have been derived in combination with machine learning methods [12]. How credit card issuers specifically manage these risks concerning the probability of occurrence and the types of risks is also a focus of academic research. Recognizing these risks and improving the system of risk response appears to be crucial. From the point of view of cost reduction, credit card risk management needs to find the balance between the risk revenue and the expected cost of risk to maximize revenue with risk minimization, reduce costs in the balance, improve the efficiency of bank operations, and use digital models to assess the risk revenue of the cost of risk accompanying the credit card business based on theoretical research [13]. However, Yikin looks at the three perspectives of internal operational risk, external systemic risk, and technical operational risk for analysis and proposes the basic ideas of risk control under the network model of multiple risk interactions [5].

The populated incomplete dataset was processed using integrated clustering methods to obtain multiple clustering results. The consistent partitioning of each clustering result is performed using voting. Firstly, label matching is done for each class of clusters in different clustering results. Then, the intersection of the same labeled class clusters is obtained, and the samples in the intersection are divided into the core domain of the corresponding class clusters. The remaining samples decide whether they belong to the core domain or the boundary domain of the class clusters according to the relationship between the number of votes obtained by voting and the set threshold value. Determining the core-domain and boundary-domain samples of the class clusters results in a three-branch clustering result. The feasibility of the algorithm is demonstrated by evaluating the clustering results by clustering validity metrics.

3. Economic Management Data Envelopment Analysis for Incomplete Data Clustering

In this section, we give a detailed description of the main algorithms and structures. Although we have just proposed a high-quality clustering criterion, it is still difficult for us to quantify it. The final evaluation of the quality of the clustering is often the satisfaction of the stakeholders after the clustering. If the demander is satisfied with your clustering results and obtains valuable information from it, then the clustering is effective and high-quality.

3.1. Design of Clustering Algorithms for Incomplete Data.

The missing values in a dataset can be classified as completely random missing, random missing, and completely non-random missing in terms of the distribution of missingness. Completely random missing means that the missing data is random, and the missing data does not depend on any incomplete or complete variables. Random missing means that the missing data is not completely random, i.e., the missing data of that type depends on other complete variables. Completely nonrandom missing means that the missing of the data depends on the incomplete variables themselves [14]. The data objects studied in this chapter are incomplete information systems under completely random missingness, i.e., the missingness of the data is random and the missingness of the data does not depend on any incomplete or complete variables. In most cases, a high percentage of missingness is often accompanied by inefficient clustering results, and when the missingness rate of the dataset is high, the accuracy for filling the missing values of the sample objects decreases. It can also directly cause a decrease in the performance of the clustering algorithm. Therefore, we set the missing rate of the dataset between 5% and 30%, i.e., the missing attribute values of the sample objects need to satisfy two conditions [15]. Any sample object must retain at least one full attribute value. Each attribute has at least one corresponding complete value in the incomplete dataset. In other words, a sample cannot be missing all attribute values and all samples cannot be missing the same attribute. This chapter and the next chapter preprocess the dataset to the incomplete dataset following the two basic conditions for missing the attribute values and the missing rate range value requirement for the dataset.

For a sample point, in addition to the inherent characteristics of the sample itself, it will be influenced by other samples. If there are more samples of a certain category around the sample, then the denser the distribution, the more likely the sample belongs to this category. Conversely, if the number of samples of a certain category around the sample is sparse and the distribution is sparser, it is less likely the sample belongs to this category. Therefore, the effective use of sample distribution information can make the clustering results more accurate. Therefore, when calculating the distance between a sample point and the cluster center, the distance calculation can be improved by introducing the proximity category information in the form of a ratio. The distance formula with the sample

distribution information can make certain adjustments to the distance measurement process with the changes in the data set, thus obtaining a more accurate distance value [16]. In the process of filling incomplete datasets, the information of missing attributes can also be collected from their nearest neighbor samples. The denser the distribution of the nearest neighbor samples, the higher the possibility of finding valid attribute information and the closer the filled value is to the true value. Inspired by this idea, this paper proposes a fuzzy mean algorithm for the incomplete data based on the sample spatial distance. The algorithm uses the nearest neighbor rule to fill the missing attributes of the incomplete data and introduces the sample spatial distribution information into the clustering process from two aspects. One is to determine the clustering influence value of the sample based on the sample nearest neighbor density, which is added to the clustering objective function in the form of weights, and the other is to correct the class information based on the sample nearest distance between the sample and the clustering center so that the process of distance metric can be adjusted somewhat with different data sets, and different sample spatial information is further introduced into the clustering process as shown in Figure 1.

Autonomous motivation significantly and positively predicts creative thinking. Controlled motivation has no significant predictive effect on creative thinking. Autonomous motivation plays a complete mediating role between moderate control and creative thinking, and it partly plays a mediating role between moderate autonomy/high autonomy and creative thinking. However, determining the number of nearest neighbors of a sample becomes a new problem. The number of nearest neighbors needs to be specified artificially, and if the number of the selected nearest neighbor samples is too small, there is a possibility that not enough attribute information is obtained to fill the missing values, resulting in too large a gap between the filled values and the true values. However, if too many nearest neighbor samples are selected, the filled attribute features will be confused by too many sample subclasses. It has a certain degree of impact on the accuracy of the algorithm.

$$D_{ik} = \frac{1}{\sum_{j=1}^s I_{jk}} \sqrt{\sum_{j=1}^s (I_{jk} + x_{jk})^2 I_{jk}}, \quad (1)$$

$$x_{jk} = \frac{1}{N} \sum_{j=1}^s I_{jk},$$

D_{ik} represents the flux, and I_{jk} represents the economic value in the j^{th} column and k^{th} row. In this way, the attributes of the complete data and the information of the attributes that are not missing in the incomplete data are fully utilized. The nearest neighbor samples of the incomplete data are identified, and the missing part of the incomplete data is filled using the average of the information of the complete attributes of the N nearest neighbor samples. It makes the filling effect more reasonable and realistic.

$$H_{ik} = N_k^i \cdot N_k, \quad (2)$$

$$d_{ik}^* = d_{ik} - a(1 - H_{ik}),$$

H_{ik} represents the flow rate, and d_{ik} represents the year. After the information granulation operation on the time series, the original time series is transformed into granular time series. The next step is to measure the similarity between the granular time series. The commonly used time series similarity measures are Euclidean distance, dynamic time-bending distance, cosine similarity distance, etc. Since the number of information grains and the size of the time window contained in the granular time series obtained from the two-time series after the information granulation operation may be different, this section proposes a new similarity measure, i.e., the linear information granulation-based time series similarity measure. For many practical data classification problems, samples originating from different classes often partially overlap in the feature space [17]; see Figure 1.

Although the training samples in the overlapping regions have accurate category labels, these samples do not reflect the exact distribution of the categories, i.e., the information provided by the samples in these overlapping regions of the categories is inaccurate. Therefore, a reasonable modeling of the imprecise training data in these overlapping regions is needed to achieve the effective utilization of this part of the training data.

Incomplete training data usually refers to the fact that the obtained training data is not sufficient to provide an effective portrayal of the true conditional probability distribution. In general, fewer training samples and higher feature dimensions are the main factors that cause the incompleteness of training data. Therefore, how to obtain better classification performance based on incomplete training data is an important topic in the design of classification methods. Unreliable training data usually means that the obtained training data has large noise in terms of categories or features. Category noise refers to the training samples being labeled as the wrong category, while feature noise refers to the deviation of some feature values of the samples from the normal range. Therefore, to obtain better classification performance based on unreliable training data, robust classification methods need to be designed to suppress the data noise.

$$J_m(U, V) = \sum_{k=1}^n \sum_{i=1}^n w_k u_{ik}^m d_{ik}^2, \quad (3)$$

$$u_{ik} = \left(\frac{\sum_{l=1}^c d_{ik}^*}{\sum_{l=1}^c d_{lk}^*} \right)^{2/1+m},$$

$$v_i = \frac{\sum_{k=1}^n w_k w_k u_{ik}^m d_{ik}^2}{\left(\sum_{l=1}^c d_{ik}^* / d_{lk}^* \right)^{2/1+m}},$$

$J_m(U, V)$ represents the fixed flow, w is the weight, u represents the proportion of different positions, and v represents the corresponding rate. The rules configuration

FIGURE 1: The framework of complete data clustering algorithm.

management of the economic data reporting system is divided into two parts: one is the configuration management of the rules for the splitting of reporting documents, and the other is the configuration management of the rules for the verification of reporting documents. These two rules are managed and configured separately while performing normative constraints for a report. The splitting rules mainly describe the parsing rules of the business unit report, such as the interval symbol between data, the split symbol between each data item, the report description item, etc. The verification rules mainly describe some requirements for the corresponding data items of the report, such as the type, name, definition, and constraints of the data items; the rule configuration is where the user input is a specific application. When the business data reported by the business unit is verified by the splitting rules and data verification rules, the system will automatically store the economic data in the corresponding original economic database for future extraction to the data warehouse as shown in Figure 2.

Since the processed data are in large batches, when storing the data into the database, one can imagine how inefficient the system would be if it were simply inserted with one SQL statement. Therefore, in this part, it can be considered to improve the speed of data depositing in terms of performance. It may be useful to intervene in the concept of the data persistence layer in terms of technical implementation. The design goal of the data persistence layer is to provide a high-level, unified, secure, and concurrent data

persistence mechanism for the entire project. The users of the economic operation platform system are the municipal and district local taxation bureaus, national taxation bureaus, industrial and commercial bureaus, the development and reform commission, the bureau of statistics, and relevant leaders at all levels. From the viewpoint of the users of the system, the level of use and computer knowledge of the users of the economic operation platform is relatively high, and many business units, such as the state taxation bureau, local taxation bureau, and the industry and commerce bureau have established their professional business systems. Some have participated in the construction of the government portal system and the construction of the office automation system, which has promoted the improvement of the level of computer use. In terms of the frequency of use of the system, the most frequent use would be the statistical analysis part of the system, i.e., the frequency of operation of the system by NDRC and leaders at all levels would be greater.

3.2. Economic Management Data Envelope Design. The data integration rules of the original platform are cured in the program by the programmer. Although the current data integration is achieved, when the data of each department needs to be newly integrated according to the new economic rules, or when new departments are added to this system, the horizontal association requirements of the data between the

FIGURE 2: Economic platform architecture.

departments will be greatly increased. At this time, the update of the data integration module will face a lot of repetitive work, which will bring a lot of inconvenience to the operation and maintenance of the system. In the data integration, the data of each department is linked horizontally and new economic data is generated. It has certain economic data analysis functions [18,19]. However, with the continuous updating of economic rules and the addition of new departments and data, the system will gradually become huge and the data will become more complicated. In this case, the management of the database group in the original design will become difficult, and the analysis of economic data will put forward higher requirements. Hence, there is an urgent need for a technology that can effectively solve massive data storage and can effectively realize data mining and analysis. Therefore, we need to improve the platform to solve these problems [20].

If a unified economic database with regional attributes is to be established instead of a single economic vertical database, a horizontal correlation of multiple vertical data will be required to create a large integrated economic database. The core purpose of horizontal data processing is to eliminate the sectoral attributes of the data itself so that the data established with horizontal correlation has regional economic attributes. At present, several cities in China have already established preliminary economic data exchange systems. For example, Qingdao has established an economic data exchange system with four departments, including

industry and commerce, taxation, and quality inspection. A part of our data source comes from our own collation, and the other part is open-source data. The practical effects and operational results reflected by the initial economic data exchange systems established in several cities show that it is feasible to establish a large horizontal economic data exchange system and management system. At present, many economic theories can only be understood by professional economic experts. The conclusions of many theories have a reference role for the regional governments to manage the economy. The system needs to correlate some standard general economic theories with the regional economic state and deduce some reference opinions for the regional governments to use as shown in Figure 3.

In a market economy, people in a transaction will have different information, and the fact that some people have information that others do not will create information asymmetry. The two results of information asymmetry are moral hazard and adverse selection. With imperfect access to information, a credit card holder may refuse to disclose all personal information to the bank so that the bank cannot get accurate information to evaluate whether the cardholder can have a credit card. Thus, in the credit card market, the mixed information leaves banks with no way to determine which cardholder has higher integrity and better cash flow. On the other hand, information asymmetry can also cause potential problems in the ex-ante credit card segment. If, after a successful credit card application, the cardholder's

FIGURE 3: Schematic diagram of the feature subspace.

repayment ability fluctuates because of a combination of factors, such as job changes, cash flow turnover, and changes in income trends, then the bank is often unable to capture that part of the information and maintains the credit limit at a similar level [19]. The bank is often unable to capture this information, and the credit limit remains at a similar level as before, while the cardholder's repayment ability has changed. This information asymmetry also hides the potential risk of credit failure and default.

From the perspective of big data, banks, and other cardholding institutions, to achieve a penetrating supervision of the cardholders, they can obtain comprehensive information of cardholders from all aspects. It not only contains financial data directly affecting the repayment ability but also includes their consumption habits, work habits, social environment, moral risks, and many other data collections, thus constituting a comprehensive information judgment system. It is conducive to reducing the risk brought by information asymmetry to banks. With deeper application in the financial field, each financial institution has built its own big data platform one after another, using the computing power of the platform to standardize and centralize the data originally scattered in various business systems using a unified data platform or data warehouse. By a scenario-based design, the making of each business scenario can be described and applied by models using their existing data for a model test and complete the application of relevant business. The application of big data in the banking industry is mainly in various aspects such as accurate marketing, refined management, low-cost management, and centralized management of banks. Banks can make precise marketing strategies for individuals using information technology and have a valuable prediction and judgment for each customer's preference and ability. As per the big data model, banks can record credit card information while also providing feedback on these bearers' consumption behaviors. It is summarized and organized, especially in terms of risk control for loans, as shown in Figure 4.

Data integration in the regional economic management system is mainly to solve the problem of data ambiguity among economic data scattered among various economic management departments describing the same economic affairs and to horizontally associate the scattered

economic data with the same business meaning to generate new economic data, i.e., to perform data aggregation [20]. The form of the dispersed economic data is shown in Figure 4. The main work to achieve the disambiguation and data integration of a regional dispersed economic data is to identify and locate the original dispersed economic data and define and identify the business association relationships between the dispersed data. Among the more general solutions available, it is more practical and feasible to use the principle of rule engines to solve it.

Operational risk is the loss caused by the bank's internal systems, resulting in inadequate internal processes and external events, such as deficiencies in related information systems and staff errors. Operational risk can be manifested in the form of losses caused to customers by the design or implementation of a customized product or by the lack of training of the bank's internal staff, which makes its employees not aware of the risks they should have and the gaps or imperfections in internal processes, as well as the risks caused by errors in the authorization and approval of information systems and the technical environment.

4. Analysis of Results

4.1. Performance Results of Incomplete Data Clustering Algorithm. Figure 5 shows the experimental results of the algorithm KM-IMI and the algorithm KM-CD on the metrics DBI, AS, and ACC, where the experiments are done 100 times on each data set. The mean and best values of these three metrics are obtained. The underlined data in the figure indicates that the clustering effect of the KM-IMI algorithm is not as good as the clustering effect of the KM-CD algorithm. By observing the underlined data, we can directly see that the clustering results of the data sets Iris, Wine, WDBC, Pen digits, and Page Blocks on the mean of the ACC evaluation metrics on the KM-IMI algorithm are not as effective as on the algorithm KM-CD. Although the two datasets, Pen digits and Page Blocks, are less effective than the KM-IMI algorithm on the mean of the ACC metrics and on their best values, the difference is only between 0.01 and 0.02. It is easy to find that one of the reasons is that there is a positive relationship between the missing rate and the accuracy, i.e., the higher the missing rate, the lower the

FIGURE 4: Original and aggregated repository clusters for the economic platform.

FIGURE 5: Experimental results on the dataset.

accuracy of the clustering results. It directly leads to the degradation of the performance of the clustering algorithm. The mean and best values of the algorithm KM-IMI on the metrics DBI and AS outperformed the algorithm KM-CD on most of the datasets, except for the underlined data. It is worth mentioning that in the dataset Banknote, CMC outperforms the mean and best values on all three indicators. Based on the above analysis, we can conclude that the

improved clustering algorithm for the mean interpolation of incomplete data, i.e., KM-IMI can effectively solve the clustering problem of incomplete data as shown in Figure 5.

Observe Figure 6, the mean and best values of the experimental results for the data on the UCI dataset for the indicators ACC and FMI, where the bolded data indicate the better experimental results. The experiments were performed 100 times under different data set missing rates, and

FIGURE 6: Average ACC.

the mean and best values were found from these 100 experiments, i.e., the mean and best values of ACC and FMI were obtained. From the figure, we can directly observe that the bolded experimental results are basically from the algorithm. Thus, we can conclude that the experimental results of the algorithm on the metrics ACC and FMI are better than the algorithms OCS-FCM and NPS-FCM, both in terms of the mean and best values of the metrics. Also, we observe that the data distribution of the algorithm is relatively stable with the values of ACC and FMI gradually decreasing as the missing rate of the data set increases. The data of the NPS-FCM algorithm is not so stable, which means that using the nearest neighbor approach to fill incomplete data is very much variable, and the nearest neighbor objects of the missing data samples are not stable, especially in the Pen digits dataset. The two algorithms compared are implemented based on the FCM algorithm. Using this algorithm, it is difficult to obtain good clustering results on non-spherical datasets. The algorithms are based on integrated clustering and can effectively improve the robustness of the clustering results, stability, and quality of the clustering results; see Figure 6.

The three-branch decision clustering uses the core and boundary domains to describe the relationship between the sample objects and class clusters. It is more appropriate than using an ensemble to represent a class cluster. At the same time, integrated clustering is an effective approach in dealing with clustering problems. The chapter proposes a three-branch integrated clustering algorithm for incomplete data by combining three-branch decision clustering with integrated clustering. Firstly, the attribute values corresponding to the missing data objects are filled according to the incomplete data filling method proposed in Chapter 2, i.e., based on the mean values of the attributes of all the sample objects in the clustering results of the complete data set.

Then, the optimal estimates are obtained using the perturbation analysis of the clustering center. The three-branch integrated clustering method is used, i.e., if the class labels of the data objects agree after multiple clustering, then the object is classified into the core of the corresponding class cluster domain. Otherwise, the object is classified into the boundary domain.

4.2. Results of the Economic Management Data Envelope.

Firstly, the data in the data warehouse comes from various data sources, including various heterogeneous database systems, data file data, other data, etc. Using data extraction tools, by the process of data extraction, cleansing, conversion, and loading, the data is loaded into the data warehouse according to different business themes, i.e., different analysis needs, to achieve integrated storage and facilitate data sharing. Then, various analysis tools are applied, such as retrieval query tools, OLAP tools for multidimensional data, statistical analysis tools, and data mining tools, to present the analysis results in the form of intuitive charts. These analytical tools, such as data mining tools, are highly functional with the help of a human.

At present, there are few information management systems for managing regional economies, and some local governments have established their own regional economic management systems according to their actual situation and needs. These systems are derived from the actual business needs of a competent economic department. Some of them also integrate the work needs of other economic management departments, and the main method used is the aggregation and analysis of economic data utilizing large data centralization. There is no data warehouse design for these information systems, however, the national information systems such as the construction of the four major databases

FIGURE 7: Tests for each method at different characteristic noise levels.

FIGURE 8: Classification error as the weight coefficient changes from 0 to 1.

adopt the data warehouse design. However, the station is higher and generally adopted by provincial units and municipalities. This system is still a blank field because it is standing in the perspective of the local government as shown in Figure 7.

The results under the same class noise condition are similar and most of the datasets proposed in this paper obtain better classification results under any feature noise level. To show more clearly, the robustness of different methods to feature noise gives the relative accuracy loss of each classification method at different feature noise levels. Based on a similar statistical analysis approach, firstly, we

analyze whether there is a significant difference between the methods on the whole using the Friedman test. The running time of the training and classification phases depends mainly on the number of rules generated. More rules mean more time is needed to train the rule base and more time is needed when classifying an input sample. Therefore, we can analyze the impact of these factors on the time complexity from the perspective of the number of training samples, several features, and some fuzzy divisions of the dataset on the number of rules.

Figure 8 gives the classification errors during the change of weight coefficients from 0 to 1 for different data noise

levels. The optimal values of the weight coefficients for different data noise levels are taken differently. As the noise level increases, the optimal value of the weight coefficient tends to become smaller. This is because the reliability of DBRB generated based on the training data decreases at high noise levels, and therefore, the noise-independent KBRB is needed to play a greater role in determining the final classification results; see Figure 8.

Credit business is a complete system, and through a big data system, it can become an organic integration of the whole business system. Big data technology can run through the whole credit business process, which can realize effective information collection and analysis before the loan, information sharing and transmission during the loan, and information monitoring and feedback after the loan. It can greatly improve the management efficiency of the loan business. After combining data analysis and questionnaire survey to analyze the problems in the bank credit business risk management from both objective and subjective aspects, it is necessary to further analyze the causes of these problems and provide a more reference piece value basis for the subsequent countermeasure formulation.

5. Conclusion

In this paper, two new methods for clustering incomplete data are proposed based on the nearest neighbor correlation of samples. However, there are still difficult problems that need to be studied. Since the incomplete samples themselves have uncertainty in the attribute space distribution, how to put aside subjectivity to choose a suitable similarity measure to determine the nearest neighbor samples of the incomplete samples is an important issue that deserves further research. Based on the existing economic platform, this thesis investigates the feasibility of building a data warehouse on the regional government economic operation platform, discusses the methods and steps of integrating business data and building a data warehouse, designs the data model and architecture of the tax data warehouse, and investigates the technical difficulties in the design of data storage and gives specific implementation methods. The key technologies in the design and implementation of building the data warehouse system are analyzed, and the theory of data warehouse is used to guide the design and development of the regional government economic operation platform system. The strategy of business data integration of the regional government economic operation platform system is studied, and the construction of the data integration platform based on the rule engine technology is analyzed and demonstrated, designed, and implemented. The data warehouse design was completed, focusing on solving technical difficulties, such as topic analysis and dimension table design. Based on this research, a unified data storage structure of the regional government economic operation platform system will be formed to provide a standard and comprehensive data source for data analysis and utilization, as well as future government decision-making. Compared with other studies, the efficiency and accuracy of our research results showed, approximately, a 10% improvement.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

No competing interests exist concerning this study.

Acknowledgments

This paper was not funded by any organization.

References

- [1] Y. Zhang, W. Wu, R. T. Toll et al., "Identification of psychiatric disorder subtypes from functional connectivity patterns in resting-state electroencephalography," *Nature Biomedical Engineering*, vol. 5, no. 4, pp. 309–323, 2021.
- [2] I. A. Udugama, C. L. Gargalo, Y. Yamashita et al., "The role of big data in industrial (Bio)chemical process operations," *Industrial & Engineering Chemistry Research*, vol. 59, no. 34, pp. 15283–15297, 2020.
- [3] M. Balistrocchi, R. Metulini, M. Carpita, and R. Ranzi, "Dynamic maps of human exposure to floods based on mobile phone data," *Natural Hazards and Earth System Sciences*, vol. 20, no. 12, pp. 3485–3500, 2020.
- [4] J. Chen, X. Lian, H. Su, Z. Zhang, X. Ma, and B. Chang, "Analysis of China's carbon emission driving factors based on the perspective of eight major economic regions," *Environmental Science and Pollution Research*, vol. 28, no. 7, pp. 8181–8204, 2021.
- [5] W. Yu, Z. Zhang, and Q. Zhong, "Consensus reaching for MAGDM with multi-granular hesitant fuzzy linguistic term sets: a minimum adjustment-based approach," *Annals of Operations Research*, vol. 300, no. 2, pp. 443–466, 2021.
- [6] M. Hussein, R. R. Stewart, D. Sacrey, J. Wu, and R. Athale, "Unsupervised machine learning using 3D seismic data applied to reservoir evaluation and rock type identification," *Interpretation*, vol. 9, no. 2, pp. T549–T568, 2021.
- [7] F. Torti, A. Corbellini, and A. C. Atkinson, "fsdaSAS: a package for robust regression for very large datasets including the batch forward search," *Stats*, vol. 4, no. 2, pp. 327–347, 2021.
- [8] A. Sayghe, Y. Hu, I. Zografopoulos et al., "Survey of machine learning methods for detecting false data injection attacks in power systems," *IET Smart Grid*, vol. 3, no. 5, pp. 581–595, 2020.
- [9] R. Metulini and M. Carpita, "A spatio-temporal indicator for city users based on mobile phone signals and administrative data," *Social Indicators Research*, vol. 156, no. 2, pp. 761–781, 2021.
- [10] A. Ben Abdelkrim, T. Tribout, O. Martin, D. Boichard, V. Ducrocq, and N. C. Friggens, "Exploring simultaneous perturbation profiles in milk yield and body weight reveals a diversity of animal responses and new opportunities to identify resilience proxies," *Journal of Dairy Science*, vol. 104, no. 1, pp. 459–470, 2021.
- [11] X. Qinlin, T. Chao, W. Yanjun, X. Li, and L. Xiao, "Measurement and comparison of Urban Haze governance level and efficiency based on the DPSIR model: a case study of 31 cities in North China," *Journal of Resources and Ecology*, vol. 11, no. 6, pp. 549–561, 2020.

- [12] D. S. Davis, R. J. DiNapoli, M. C. Sanger, and C. P. Lipo, "The integration of lidar and legacy datasets provides improved explanations for the spatial patterning of shell rings in the American Southeast," *Advances in Archaeological Practice*, vol. 8, no. 4, pp. 361–375, 2020.
- [13] L. K. Chen, A. C. Hwang, W. J. Lee et al., "Efficacy of multidomain interventions to improve physical frailty, depression and cognition: data from cluster-randomized controlled trials," *Journal of cachexia, sarcopenia and muscle*, vol. 11, no. 3, pp. 650–662, 2020.
- [14] J. Cawley, D. Frisvold, and D. Jones, "The impact of sugar-sweetened beverage taxes on purchases: evidence from four city-level taxes in the United States," *Health Economics*, vol. 29, no. 10, pp. 1289–1306, 2020.
- [15] G. He, S. Wang, and B. Zhang, "Watering down environmental regulation in China," *Quarterly Journal of Economics*, vol. 135, no. 4, pp. 2135–2185, 2020.
- [16] A. Adhvaryu, N. Kala, and A. Nyshadham, "The light and the heat: productivity co-benefits of energy-saving technology," *The Review of Economics and Statistics*, vol. 102, no. 4, pp. 779–792, 2020.
- [17] G. F. Fan, Y. H. Guo, J. M. Zheng, and W. C. Hong, "A generalized regression model based on hybrid empirical mode decomposition and support vector regression with back-propagation neural network for mid-short-term load forecasting," *Journal of Forecasting*, vol. 39, no. 5, pp. 737–756, 2020.
- [18] P. C. Medina, "Side effects of nudging: evidence from a randomized intervention in the credit card market," *Review of Financial Studies*, vol. 34, no. 5, pp. 2580–2607, 2021.
- [19] S. D. Bennett, J. H. Cross, A. E. Coughtrey et al., "MICE—mental Health Intervention for Children with Epilepsy: a randomised controlled, multi-centre clinical trial evaluating the clinical and cost-effectiveness of MATCH-ADTC in addition to usual care compared to usual care alone for children and young people with common mental health disorders and epilepsy—study protocol," *Trials*, vol. 22, no. 1, pp. 1–16, 2021.
- [20] F. Wang, D. Wang, G. Guo, M. Zhang, J. Lang, and J. Wei, "Potential distributions of the invasive barnacle scale *ceroplastes cirripediformis* (Hemiptera: coccidae) under climate change and implications for its management," *Journal of Economic Entomology*, vol. 114, no. 1, pp. 82–89, 2021.