Fuzzy Clustering and Kernel PCA-Based High-Dimensional Imbalanced Data Integration with Octree Encoding

Qin Wang
Financial Teaching and Research Office, Zhongyuan University of Science and Technology, Zhengzhou 461100, China
E-mail: cgwac_wq@163.com

Keywords: fuzzy clustering, web crawler, high dimensional imbalance, national accounts, principal component analysis, data coding

Received: February 12, 2025

Due to the high-dimensional and unbalanced characteristics of national economic accounting data, there is a large amount of redundant information in the data, which will lead to problems such as boundary shift and integration overfitting shift when integrating the data, and will increase the difficulty of subsequent data integration. For this reason, a fuzzy clustering-based method for integrating high-dimensional unbalanced data of national accounts is proposed. Using the kernel principal component analysis method to reduce the dimensionality of high-dimensional imbalanced national economic accounting data, in order to reduce the complexity and sparsity of the data while preserving the main information of the original data as much as possible. Use fuzzy clustering algorithm for data clustering. Fuzzy clustering allows data points to belong to multiple clusters simultaneously, with each cluster having a membership measure that represents the strength of the relationship between data points and each cluster. Introducing deviation maximization for optimizing fuzzy clustering methods to ensure that the distance between each data point and its cluster center is as large as possible, while ensuring that the distance between data points within the same cluster is as small as possible. Based on text free grammar rules and conversion functions, convert national economic accounting data into hesitant fuzzy language data and obtain the optimal data attribute weight vector. Calculate the distance between different categories and the minimum distance, and determine the repulsion phenomenon between unknown and known classes through the objective function. Using Lagrange multipliers to solve the objective function and obtain the optimal clustering center. According to the optimal clustering center, complete the clustering of national economic accounting data and obtain different categories of national economic accounting data. According to the experimental results, the data integration imbalance of the proposed method ranges from 1.68% to 32.85%, and the total number of samples fluctuates between 139 and 5136. The three indicators of the integrated data are all greater than 0.88. Through actual coding cases, the coding ability of our method for high-dimensional imbalanced data in national economic accounting has been verified.

Povzetek: Predstavljena je vizualno-tekstualna klasifikacija sentimentov z uporabo računalniških metod. Prispevek izboljšuje analizo z integracijo večmodalnih podatkov in predlaga nov model.

1 Introduction

Integration of high-dimensional unbalanced data in national economic accounting refers to the process of systematically organizing, summarizing and integrating the high-dimensional and unbalanced data involved in the process of national economic accounting [1]. National economic accounting data mainly come from statistics, administrative data, accounting accounts, census and sample survey data and other aspects, covering agriculture, industry, services and other major industrial sectors, and covering the government, enterprises (including state-owned enterprises, private enterprises, foreign-funded enterprises, etc.), residents and other economic subjects, including gross domestic product (GDP), consumer price index (CPI), investment in fixed assets, total imports and exports and other economic indicators, so it has a significant high-dimensional characteristics; at the same time, there are differences in the proportion and growth rate of different industrial sectors in the national economy, the government and enterprises, as well as residents, and other economic agents in the distribution of income, consumption, investment and other aspects of the performance of the different regions of the level of economic development, industrial structure, resource endowments, etc., resulting in uneven economic performance between the regions. This leads to the imbalance of economic performance between regions, and there may be differences in economic performance in different time periods, which also leads to the significant imbalance of national economic accounting data [2-3]. The integrated high-dimensional unbalanced data provide richer materials for data mining and analysis, and researchers can use the integrated economic data to explore the deep economic laws and discover the potential economic problems and

trends, which helps to deepen the understanding of the national economy and provide strong support for economic development.

The research design and objectives of this paper focus on the integration of high-dimensional imbalanced data in national economic accounting. The research aims to address how to effectively manage and utilize these complex data to provide more accurate reflection of the national economic situation and policy formulation basis. The specific assumption is that by using web crawling technology to obtain comprehensive data and applying kernel principal component analysis for dimensionality reduction, combined with fuzzy clustering algorithms and data encoding methods, a more optimized data integration effect can be achieved than traditional methods. The measures of success include the degree of information retention after data dimensionality reduction, the accuracy of clustering results, and the convenience of data management. The evaluation indicators may involve the proportion of dimensionality reduction after data dimensionality reduction, the stability and interpretability of clustering results, as well as the storage efficiency and retrieval speed of encoded data. Through these measures, the research aims to provide new ideas and methods for the processing and analysis of national economic accounting data.

Fuzzy clustering is a clustering algorithm based on fuzzy set theory and fuzzy logic, which can take into account the situation that each data point belongs to multiple clustering centers, and can deal with uncertain or noisy data and improve the stability, reliability and flexibility of clustering analysis [4]. Applied to the integration of high-dimensional unbalanced data of national accounting, it can consider the ambiguity that each data point may belong to multiple categories, and realize the clustering division of national accounting data by calculating the affiliation degree of the data points to each category, so as to deal with high-dimensional and unbalanced data in a more flexible way. Therefore, we propose the integration of high-dimensional unbalanced data of national economic accounting based on fuzzy clustering to reflect the real structure of high-dimensional unbalanced data of national economic accounting more accurately, so as to improve the accuracy and flexibility of the high-dimensional unbalanced data of national economic accounting, solve the problem of unbalanced data of national economic accounting and improve the efficiency and automation degree of the integration of data, so that the overall operation status of the national economy can be reflected more accurately for the policy making. At the same time, through the integration and sharing of data resources, we can strengthen the international economic exchanges and cooperation, and jointly deal with the global economic problems and challenges [5].

The main contributions and limitations of different methods is shown in Table 1.

Table 1: Main contributions and limitations of different methods

Different methods	Main contributions	Limitation	Computing efficiency	Accuracy	Robustness
Ikoma et al [6]	Using web crawling technology to obtain Earth environmental data related to national life; Develop data integration layer and application layer	The diversity and format of data make API processing difficult; Insufficient data fusion	Medium	Medium	Low (data missing, incorrect, duplicate)
Yang et al [7]	Provide strategies for integrating big data analysis and neural network optimization design; Conduct predictive research and error analysis	Large error	High (big data processing and neural network training)	High (optimized through error analysis)	Medium (depending on data quality and neural network architecture)
Han et al [8]	Propose a data integration scheme based on k-anonymity and data privacy protection protocol; Introducing secure multi-party computation and ciphertext classification methods	There are significant differences in data format, structure, and quality; Encryption classification increases complexity	Low (ciphertext processing and secure multi-party computation)	Medium (depending on the effectiveness of privacy protection protocols)	High (Protecting Privacy)
Dong et al [9]	Propose an algorithm for clustering incomplete	The calculation process is quite complex	High (combining clustering and	High (time series prediction method)	Medium (depending on the performance of

	data; Combining clustering algorithms and information granulation methods; Propose a time series prediction method		information granulation to improve efficiency)		clustering algorithms)

2 Integration of high-dimensional imbalance data in national accounts

2.1 High-dimensional imbalance data acquisition for national accounts based on web crawler technology

Based on the characteristics of high-dimensional and unbalanced national economic accounting data, in order to effectively complete the integration of this data and realize the unified management of high-dimensional unbalanced data, web crawler technology is used to automate the data collection of national economic accounting data from different data sources, which mainly involves the selection of the Uniform Resource Locator System (URL) strategy and the extraction of page content [10]. First, the data collection specifies one or several starting URLs, downloads and interprets their corresponding web page source code. In order to ensure the accuracy of national accounts data extraction, regular expressions are utilized as the main extraction means to start data extraction, and after extraction, the extracted national accounts information is stored in the database for subsequent integration of high-dimensional imbalance data in national accounts. In addition, in the process of national economic accounting data collection, new URL addresses are continuously recognized and added to the pending list to ensure the comprehensiveness of national economic accounting data collection. The data collection based on web crawler will continue until it meets the preset termination criteria to stop crawling, and then the corpus will be constructed on this basis [11]. The specific process of crawling the high-dimensional imbalance data of national economic accounting is shown in Figure 1.

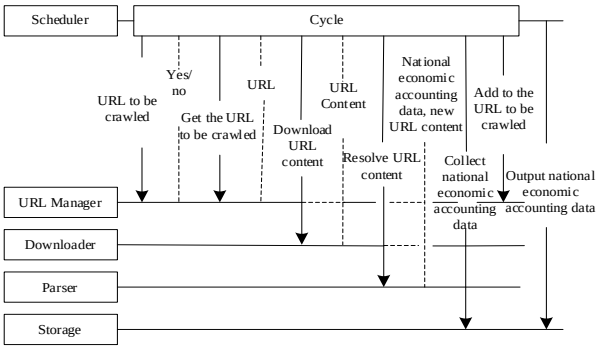

Figure 1: Crawling process of high-dimensional imbalanced data in national economic accounting

As shown in Figure 1, when crawling national accounting data, it mainly consists of five parts: scheduler, URL manager, downloader, parser and memory. Using the scheduler URL, query whether there are national accounting data URL resources to be crawled, and get the national accounting data URL resources to be crawled to locate the URL resources, and then transfer them into the URL manager; using the downloader, save the national accounting data resources URL resources; transfer the national accounting data resources URL resources to the parser, and use the parser to parse them, and get all the national accounting high-dimensional imbalance resources therein. data resource URL resource; transfer the national economic accounting data resource URL resource to the parser, and use the parser to parse it to obtain all the national economic accounting high-dimensional imbalance data and the new national economic accounting data resource URL locator; store the national economic accounting data in the memory and iteratively parse the new national economic accounting data resource URL locator to obtain it [12], after obtaining all the national economic accounting high-dimensional imbalance data, end the iteration, store all the national economic accounting high-dimensional imbalance data in the memory, so as to complete the acquisition of the national economic accounting high-dimensional imbalance data and obtain the national economic accounting data set X , this dataset contains imbalance data of different dimensions and attributes.

The reliability of using web crawlers to collect national economic data mainly depends on correct URL selection, accurate regular expression extraction, and continuous URL iteration parsing. By setting clear starting URLs and termination criteria, ensure the comprehensiveness and accuracy of data collection. To verify the accuracy of the data, data comparison and verification steps were implemented, comparing the crawled data with the officially released national economic accounting data. In addition, it also includes a data cleaning process, such as removing duplicate items, correcting erroneous data, etc., to ensure the quality of the final national economic accounting dataset obtained.

2.2 Kernel principal component analysis-based dimensionality reduction processing for high-dimensional unbalanced data

In high-dimensional unbalanced data for national economic accounting, "high-dimensional data" refers to data sets with multiple attributes or characteristics that

may be involved in many aspects of national economic accounting, such as time, region, industry, type of enterprise, economic indicators, etc. Unbalanced data refers to data sets where there are significant differences in sample sizes between categories. The unbalanced data refers to the data set in which there is a significant difference in the sample size between the categories. By integrating high-dimensional unbalanced data, redundancy and duplication in the data can be eliminated, and the accuracy and consistency of the data can be improved, which can help to reflect the real situation of the national economy more accurately, and provide a reliable basis for policy formulation and decision-making. In order to ensure the effect of subsequent data integration, the kernel principal component analysis is used to undersample the acquired high-dimensional imbalance data of national accounts, mapping the original data into a new low-dimensional space, while retaining the main information of the original data as much as possible. In this process, redundant information and noise will be effectively removed or weakened [13], which can realize the high-dimensional data dimensionality reduction, reduce the workload of the subsequent classifier training and testing, and synchronously improve the integration efficiency; and through the data dimensionality reduction can, to a certain extent, reduce the sparsity and complexity of the data distribution in the high-dimensional space, deal with the overfitting bias of the data. The preprocessed data can be maximized to achieve a relatively balanced state between the majority class and the minority class [14]. Kernel Principal Component Analysis (PCA) is a nonlinear dimensionality reduction technique that effectively addresses the problem of dimensionality reduction in high-dimensional imbalanced data by mapping raw data to a high-dimensional feature space and performing principal component analysis in that space. In the processing of high-dimensional imbalanced data in national economic accounting, kernel PCA first selects appropriate kernel functions and parameters, maps the original data to the kernel space, and forms a kernel matrix. By centralizing and decomposing the kernel matrix, the main components of the data, namely eigenvectors and eigenvalues, are extracted. Based on these main components, project the raw data into a low dimensional space to achieve dimensionality reduction of the data. This process not only preserves the main information of the data, but also helps balance the sample size between different categories, providing strong support for subsequent data integration and analysis. The steps to reduce and unbalance the high-dimensional unbalanced data of national accounts based on kernel principal component analysis are as follows.

(1) The initial sample matrix is established. Assuming that in the national economic accounting data set X, the value of the i th national economic accounting high-dimensional imbalance data under the j th attributes is x_{ij}, the initial matrix Z is established, and the matrix is normalized, and the processed matrix Z' is:

$$Z=\begin{pmatrix} x'_{11} & x'_{12} & \cdots & x'_{1n} \\ x'_{21} & x'_{22} & \cdots & x'_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ x'_{i1} & \cdots & x'_{ij} & \cdots \\ \vdots & \vdots & \vdots & \vdots \\ x'_{m1} & x'_{m2} & \cdots & x'_{mn} \end{pmatrix} \tag{1}$$

Among them, x'_{ij} represents initialized, arbitrary national accounts high-dimensional imbalance data; the m represents the number of high-dimensional imbalances in national accounts to be evaluated. n represents the number of high-dimensional imbalance data attribute features of the national accounts.

(2) Selection of the nuclear function and nuclear parameters based on the high and unbalanced characteristics of the national economic accounting data, selection of the nuclear function that is appropriate to the characteristics of the sample data and the setting of the relatively optimal parameter values, mapping the R space data into the H space gives the kernel matrix U as:

$$U_{ij}=\varphi\left(Z'_i\right)\cdot\varphi\left(Z'_j\right) \tag{2}$$

Among them, φ represents a mapping operation. Z'_i and Z'_j represent the number corresponding to row i and column j of the high-dimensional unbalanced data S' of the matrix national economic accounts.

(3) Centered on the kernel matrix, centered on the high-dimensional disequilibrium data of the national accounts, which is formulated as follows:

$$U'_{ij}=U_{ij}-\bar{U}_i-\bar{U}_j+\bar{U} \tag{3}$$

Among them, U'_{ij} represents the kernel matrix after centering the high-dimensional disequilibrium data from the national accounts. $\bar{U}$ represents all data averages in the kernel matrix.

(4) U'_{ij} is decomposed to obtain the eigenvalue λ_u and its eigenvector V_u of the matrix, and the decomposition formula is:

$$U'_{ij}V_u=\lambda_u V_u \tag{4}$$

(5) Determine the dimension of the feature space t, conditions for spatial dimension t should be fulfilled are is:

$$\frac{\sum_{u=1}^{t}\lambda_u}{\sum_{u=1}^{n}\lambda_u}\geq 0.8 \tag{5}$$

(6) Calculate the principal components of X, the national accounts were obtained by projecting the high-

dimensional data as follows:

$$\left(V\cdot\varphi\left(X\right)\right)=\sum_{j=1}^{n}\alpha_j^u U\left(X_j,X\right) \tag{6}$$

Where, α_j^u represents the j th data of the characteristic vector V_u of high-dimensional data of national economic accounting, $\left(V\cdot\varphi\left(X\right)\right)$ represents the u th principal component of the high-dimensional data point X in national economic accounting.

(7) The factor score $Y_{i\tilde{t}}$ of the i th national economic accounting high-dimensional data point in the $\tilde{t}$ th spatial principal component is calculated as:

$$Y_{i\tilde{t}}=X\cdot V_{i\tilde{t}} \tag{7}$$

(8) Calculate the score F_i for each sample of national accounts data as:

$$F_i=w_1Y_{i1}+w_2Y_{i2}+\cdots+w_iY_{i\tilde{t}} \tag{8}$$

Among them, w represents the weight coefficients of each principal component.

(9) According to the results of F_i, the national economic accounting data X is arranged in descending order, and the first r copies of multiple data are deleted, so as to realize the national economic accounting data and realize the undersampling [15], complete the under-sampling of high-dimensional data of national accounts, realize the balanced processing of the data downgrading and unbalanced data, and obtain the downgraded data set $\hat{X}$ of national accounts.

2.3 Integration of national accounts data

2.3.1 Clustering of national accounts data based on fuzzy clustering

According to the above subsection to complete the national economic accounting data dimensionality reduction processing, although through the data dimensionality reduction can be reduced to a certain extent the imbalance of the data [16], but the national economic accounting data is still characterized by imbalance, therefore, in order to ensure that the effective integration of the data, fuzzy clustering algorithm is used for the data clustering, and the introduction of the data coding method on the basis of the clustering, which is combined with the above two steps to realize the effective integration of the national economic accounting data. Fuzzy clustering method has significant advantages in dealing with high-dimensional unbalanced data, the method can establish the uncertain description of the sample to the category, can be well adapted to the complex structure of high-dimensional unbalanced data [17], at the time when it clustering, more objectively reflecting the ambiguities and uncertainties in the real world, allowing the data points to belong to multiple clusters at the same time, and each sub-cluster has a degree of subordination measure, which expresses the strength of the relationship between the data points and each cluster. Each sub-cluster has an affiliation measure, which indicates the strength of the relationship between the data point and each cluster. This flexibility allows fuzzy clustering to better capture subtle differences and overlaps in the data, thus providing more accurate clustering results [18].

When the fuzzy clustering method is used to cluster the national economic accounting data $\hat{X}$, in order to ensure that the method can be better adapted to the unbalanced characteristics of the data, the optimization of the fuzzy clustering method is introduced by deviation maximization, which ensures that the distance between each data point and the cluster center of the cluster to which it belongs is as large as possible, and at the same time ensures that the distance between the data points in the same cluster is as small as possible. In this way, not only can the clustering results be clearer and more accurate, but also can reduce the impact of noise and outliers on the clustering results; and through the maximization of the deviation, the optimal number of clusters and clustering centers can be automatically selected, making the clustering results more stable and reliable. Assuming the set of attributes of the national economic accounting data $\hat{X}$ is $N=\{n_1,n_2,\ldots,n_n\}$, the attribute weight vector is $w=\left(w_1,w_2,\cdots,w_m\right)^T$, the national economic accounting data $\hat{X}_i$ under each attribute n_j is represented using language expressions or statements, the steps for clustering national accounts data based on fuzzy clustering are as follows:

(1) According to the text free grammar rules, use the conversion function to transform the national economic accounting data into hesitant fuzzy language data, and obtain the optimal national economic accounting data attribute weight vector $w'=\left(w_1',w_2',\cdots,w_m'\right)^T$ based on the maximization of the deviation.

(2) Take each of the national accounts data $\hat{X}_i$ as a category, calculating the spacing $d\left(\hat{X}_i,\hat{X}_j\right)$ between the different categories as:

$$d\left(\hat{X}_i,\hat{X}_j\right)=\left[\frac{1}{L}\sum_{l=1}^{L}\left(\frac{\left|\hat{X}_i,\hat{X}_j\right|}{2D+1}\right)^{\lambda}\right]^{\frac{1}{\lambda}},\left(\hat{X}_i,\hat{X}_j\in A\right) \tag{9}$$

Among them, D represents the maximum distance, λ represents the degree of affiliation calculated with the category spacing.

(3) Calculate the minimum distance $d'\left(\hat{X}_i,\hat{X}_j\right)$ for national accounts data as:

$$d'\left(\hat{X}_i,\hat{X}_j\right)=\arg\min_{X_i,X_j\in\hat{X}} d\left(\hat{X}_i,\hat{X}_j\right),\left(0\le i,j\le n,i\ne j\right) \quad (10)$$

Among them, i and j represent the i th and j th attribute for national accounts data.

(4) Assuming that the center of clustering of the national accounts data is known to be B_κ, the unknown class clustering centers to be B_χ, calculating the objective function $\Im$ between the two, to determine the phenomenon of exclusion, the objective function between the unknown and known categories of national accounts data is formulated as follows:

$$\Im=\sum_{i=1}^{a}\sum_{k=1}^{n} w_k' u_{ik}^m\left(S_k-\hat{X}_i\right)+\sum_{j=1}^{b}\sum_{k=1}^{n} w_k' v_{jk}^m\left(S_k-\hat{X}_i\right) \quad (11)$$

Among them, u_{ik}^m and u_{jk}^m indicate that under dimension m, the degree for the k th national accounts data of the i th and the j th attributes, the relationship between the two is as follows:

$$\sum_{i=1}^{a} u_{ik}^m+\sum_{j=1}^{b} v_{jk}^m=1 \quad (12)$$

Among them, a,b represents the set of national economic accounting data, the range for values for u_{ik}^m and v_{jk}^m are both $[0,1]$.

(5) Using Lagrange multipliers, solving Eq. (11) yields:

$$\Im'=\Im-\Diamond\left[\sum_{i=1}^{a} u_{ik}^m+\sum_{j=1}^{b} v_{jk}^m-1\right] \quad (13)$$

Among them, $\Diamond$ represents the Lagrange multiplication operator.

(6) Calculate to obtain an optimal national accounting data center K_i as:

$$\begin{cases}\dfrac{\partial\Im'}{\partial K_i}=-2\sum_{i=1}^{a} w_k' u_{ik}^m\left(S_k-K_i\right)\\ K_i=\sum_{k=1}^{n}\omega_k u_{ik}^m S_k\end{cases} \quad (14)$$

The clustering of national accounts data can be accomplished by obtaining different categories of national accounts data $Y=\left(Y_1,Y_2,...,Y_k\right)$ based on the optimal data centers calculated in equation (14).

2.3.2 Integration of national accounts data

In order to improve the effect of national economic accounting data integration, make a large number of complex high-dimensional national economic accounting data become easy to manage, based on the above clustering, the clustered national economic accounting data are coded in the paper, and data coding is an important part of data integration. In national economic accounting, the data involved are huge and complex, and it is not only inefficient but also easy to make mistakes when dealing with these raw data directly. Through the coding process, each data point can be given a concise, unique identification, thus facilitating the rapid identification, retrieval and management of data, rapid identification of the clusters to which each data point belongs, so that the data has a clear identification, easy to understand and interpret [19]; at the same time, the coded data is more concise than the original data, which can save storage space, and realize the integration of high-dimensional unbalanced data.

In the paper, an octree model is used to code and store $Y=\left(Y_1,Y_2,...,Y_k\right)$ to complete the overall integration of high-dimensional unbalanced data of national economic accounting, we encode and store the data; fork tree is a special kind of tree data structure with order and hierarchy, which makes it has great advantages in the balance of display accuracy and speed, the elimination of hidden lines and hidden surfaces, etc., and it can efficiently deal with the sparse and dense data to complete the optimization of storage and integration of the data. Therefore, the octree model is constructed in the paper, and the clustering results are converted into binary codes and stored in the octree nodes to accomplish efficient data compression coding and storage [20]. The structure of the octree model is shown in Figure 2.

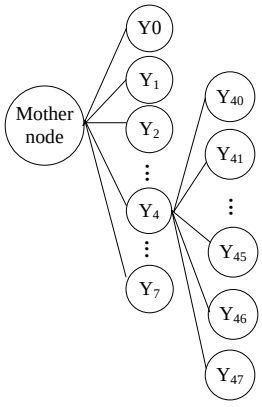

Figure 2: Octree model structure

The specific steps for octree-based coding of national accounts data are as follows.

(1) Using the national accounts data Y obtained in 2.3.1, the octree retrieval model is constructed to extend the quadtree in the two-dimensional space to the three-dimensional subdivision space.

$$\begin{cases}\tau=\tau\left(Y_k\right)+\hbar/4\\ \zeta=\zeta\left(Y_k\right)+\hbar/4\\ \xi=\xi\left(Y_k\right)+\hbar/4\end{cases} \quad (15)$$

Among them, τ, ζ and ξ represent the 3D coordinates of the node, $\hbar$ represents the cube length of the node, such that the acquired high-dimensional national accounting data Y_k as the octree parent node.

(2) Taking the parent node Y_k as the datum node, the positions of edge neighbor nodes and face neighbor nodes of the datum node are computed, respectively, using Y_5 and Y_6 to express as:

$$\begin{cases} \tau(Y_6) = \tau(Y_5) + 4 \\ \zeta(Y_6) = \zeta(Y_5) - 4 \\ \xi(Y_6) = \xi(Y_6) \end{cases} \tag{16}$$

(3) The expression for the octree retrieval model child nodes Y_4 are computed from the coordinates of the parent node:

$$\begin{cases} \tau(Y_4) \in \{\tau(Y), \tau(Y) + \hbar, \tau(Y) - \hbar\} \\ \zeta(Y_4) \in \{\zeta(Y), \zeta(Y) + \hbar, \zeta(Y) - \hbar\} \\ \xi(Y_4) \in \{\xi(Y), \xi(Y) + \hbar, \xi(Y) - \hbar\} \end{cases} \tag{17}$$

(4) Introducing judgmental encoding, judging the child nodes Y_{45} and child node Y_{46} whether the corresponding national economic accounting data is located in the code number is consistent. If it is consistent, the two child nodes are judged to be neighboring nodes to each other; if it is not consistent, the corresponding data Y_{46} is put into the neighborhood data sheet Y_{45}, calculated as follows:

$$\begin{cases} \tau(\Delta) = \tau(\nabla) \pm \hbar_{\nabla} \\ \zeta(\Delta) = \zeta(\nabla) \\ \xi(\Delta) = \xi(\nabla) \end{cases} \tag{18}$$

Among them, ∇ and Δ represent different judgment code.

(5) When the judgment code is the code of two nodes and only one bit is different, these two nodes are edge neighbor nodes to each other, this process is formulated as:

$$\begin{cases} \tau(\nabla) \pm \hbar_{\nabla} = \tau(\Delta) \\ \zeta(\nabla) \pm \hbar_{\nabla} = y(\Delta) \\ \xi(\Delta) = \xi(\nabla) \neq 0 \end{cases} \tag{19}$$

(6) When the judgment code is the code of two nodes, and only one bit is different but connected, these two nodes are point neighbor nodes to each other, this process is formulated as:

$$\begin{cases} \tau(\nabla) \pm \hbar_{\nabla} = \tau(\Delta) \\ \zeta(\nabla) \pm \hbar_{\nabla} = y(\Delta) \\ \xi(\nabla) \pm \hbar = \xi(\Delta) \end{cases} \tag{20}$$

To summarize the above steps, determine the relationship between all the child nodes, according to the node relationship can be expressed in the hierarchy and spatial relationship of the data block, according to the occupancy information and hierarchical relationship of the node, each node is assigned a unique address code, and finally, all the nodes of the address code and attribute bits are stored in order to form the linear octree coded data set, complete the national economic accounting data coding, so as to achieve the final integration of national economic accounting high-dimensional imbalance data. The final integration of the high-dimensional unbalanced data of national accounts is realized.

3 Test analysis

To verify this method can integrate high dimensional unbalanced data of national economic accounting, from labor dynamic survey in 2022, family financial survey, national health and nutrition survey, rural urban migration survey and family income five data sets, randomly selected the national economy unbalanced data, divided into regional unbalanced data, industry imbalance data, urban and rural imbalance data, income imbalance data and other unbalanced data (education, social welfare, etc.) five types of unbalanced data set, as an experimental data set. At the same time, the experiment utilizes the generalized web crawler technique to crawl the national economic accounting data, and the multi-threaded crawling object is expanded from the seed URL to the whole Web, so as to provide reliable data for the experiment. The specific coverage of the national economic accounting data set is shown in Table 2.

3.1 Analysis of data crawling results

In order to verify whether the method of this paper can effectively crawl the high-dimensional imbalance data of national economic accounting, use the method of this paper to crawl the high-dimensional imbalance data of national economic accounting from the experimental data set, in addition, part of the data is more than the categorized data set, this experiment uses a widely used transformation method to merge some of the categories of the multiclassified data set into a dichotomous data set, and use the method of this paper to crawl the data of national economic accounting to be shown in Table 3.

Table 2: Experimental dataset

Name of National Economic Accounting Dataset	Coverage
Labor force dynamic survey	Employment, unemployment, labor loss, etc
Family Financial Survey	Different provinces, cities, and rural areas
National Health and Nutrition Survey	Community, Family, Individual, Health
Survey on Rural Urban Migration	Rural residents, urban households, and migrant workers
Household Income Survey	Individual income distribution in urban and rural areas, as well as in rural areas

Table 3: High dimensional Imbalance data in national economic accounting

Name of National Economic Accounting Dataset	Total number of samples (pieces)	Attribute	Number of majority class samples (pcs)	Number of minority class samples (pcs)	Imbalance (°)
Labor force dynamic survey	2361	35	1298	210	8.5
Family sinancial survey	598	42	232	21	2.6
National Health and nutrition survey	1500	29	895	98	4.0
Survey on rural urban migration	139	28	59	9	1.5
Household income survey	5136	16	2150	413	32.5

Analysis of Table 2 shows that: using this paper's method to crawl to the national economic accounting high-dimensional imbalance data, the imbalance degree ranges from 1.68% to 32.85%, and the total number of samples fluctuates between 139 and 5136, the sample size and imbalance degree of the national economic accounting data is widely distributed, and the difference is as high as 4997bit; In addition, the high-dimensional imbalance data of national economic accounting crawled by the method of this paper contains five types of information in the experimental dataset, which indicates that the imbalance data of national economic accounting can be obtained by using the crawler method of this paper.

3.2 Analysis of the effect of data downscaling

In order to verify the effect of the method of this paper on the dimensionality reduction of high dimensional imbalance data of the national economy, the indicator pressure function o and descending masses ϕ are introduced, pressure function o denotes the loss value of the data before and after dimensionality reduction, the range of values is $[0,1]$, the smaller its value, the smaller the degradation loss, and vice versa; the degradation quality ϕ denotes the quality of the data after dimensionality reduction, and the range of values is $[0,1]$, the larger the value, the better the effect of data dimensionality reduction, and vice versa; both are calculated as follows:

$$o=\sqrt{\frac{\sum_{i=1,j=1,i\neq j}^{\vartheta}\left(d_{ij}-\bar{d}_{ij}\right)^2}{\sum_{i=1,j=1,i\neq j}^{\vartheta}d_{ij}^2}} \tag{21}$$

$$\phi=\frac{1}{\varpi\vartheta}\sum_{i,j\in 1,\cdots,\varpi}\Theta_{i,j}-\frac{\varpi}{\vartheta-1} \tag{22}$$

Where, d_{ij} represents the spatial distance between the i th data and the j th data after dimension reduction of high dimensional data, $\bar{d}_{ij}$ represents the mean spatial distance of d_{ij}, $\Theta_{i,j}$ represents the number of overlaps, ϑ represents the number of samples taken, ϖ represents the neighborhood size.

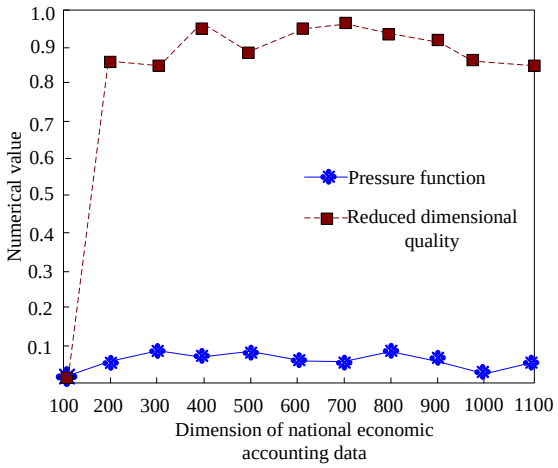

Figure 3: Calculation of dimensionality reduction results indicators for high dimensional imbalance data in national economic accounting

From the experimental data set, 10 national economic accounting high-dimensional imbalance data are randomly selected and numbered from 1 to 10, using the method of this paper for its dimensionality reduction processing, using the above formula to calculate the pressure function and quality of dimensionality reduction of the national economic accounting data after dimensionality reduction processing, and the results of the two calculations are shown in Figure 3.

The analysis of Figure 3 shows that: after using this paper's method to downsize the experimental national economic accounting high-dimensional imbalance data, the maximum value of the pressure function is 0.017, indicating that the loss value of the national economic accounting data after using this paper's method to downsize is smaller; and the minimum value of the quality of the downsizing is 0.86, indicating that the quality of the national accounting data after using this paper's method to downsize is retained to a higher degree, further proving that this paper's method is good for the downsizing of high-dimensional imbalance data of national economic accounting.

3.3 Analysis of the validity of the integration of high-dimensional imbalance data in national accounts

In order to verify whether the fuzzy clustering method in this paper can integrate the high-dimensional imbalance data of the experimental national accounts, and to introduce the adjust mutual information Q, adjust Rand factor ℓ indicators and guidelines for variance ratios σ, adjust mutual information Q is used to evaluate the correlation between the category to which the data belong and the experimental national accounts high-dimensional imbalance data; and the adjust Rand coefficients ℓ is used to evaluate the fit of distributions between different experimental national accounts high-dimensional imbalance data under the same category; the cubic difference ratio criterion σ represents the clustering effect, and all three take values in the range of $[0,1]$, the larger the value, the better the clustering effect of the high-dimensional imbalance data of the experimental national accounts, and the three formulas are:

$$\ell = \frac{\upsilon - E[\upsilon]}{\max(\upsilon) - E[\upsilon]} \tag{23}$$

$$Q(\iota_1, \iota_2) = \frac{M - E[M]}{\frac{1}{2}\left(H(\iota_1) + H(\iota_2) - E[M]\right)} \tag{24}$$

$$\sigma = \frac{tr(\Phi)}{tr(\Gamma)} \cdot \frac{\Xi - \mathrm{I}}{\mathrm{I} - 1} \tag{25}$$

Among them, M represents the mutual information between high-dimensional imbalance data in national accounts ι_1 and ι_2. E represents the expected value of both, υ represents the Rand coefficient, H represents the information entropy, φ denotes the number of pairs of elements belonging to the same category in the clustered and integrated high-dimensional imbalance data of the national accounts and the real category. ι represents the logarithm of the elements of the different classes of national accounts high-dimensional imbalance data after clustering, C represents the clustering completeness, Φ represents the covariance matrix between classes of high-dimensional unbalanced data from national accounts, Γ represents the covariance matrix within the class, Ξ represents the total number of high-dimensional imbalances in the national accounts. I represents the number of data categories.

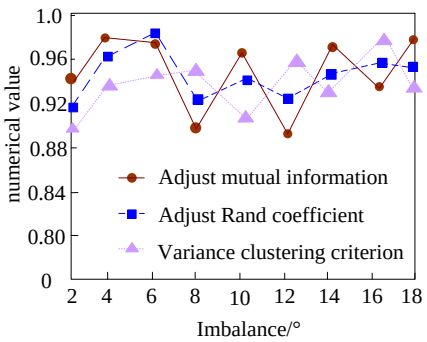

Figure 4: Calculation table of data clustering indicators

From the experimental data set, one data set was randomly selected, with the increase of the high-dimensional imbalance data of national economic accounting, calculate the adjust mutual information Q and the adjust Rand coefficient ℓ and the variance ratio criterion σ for the high-dimensional imbalance data of different national economic accounts, in order to analyze the data processing effect of the method in this paper, the test results are shown in Figure 4.

The analysis of Figure 4 shows that: after using the fuzzy clustering method of this paper to integrate the high-dimensional imbalance data of the experimental national economic accounting, with the continuous change of data imbalance, the adjusted mutual information, adjust Rand coefficient and variance ratio criterion of the integrated data are all greater than 0.88, which are in the larger value, indicating that the use of this paper's fuzzy clustering method can be used to integrate the high-dimensional imbalance data of the experimental national economic accounting effectively. integration.

3.4 Analysis of the effects of integrating high-dimensional imbalance data in national accounts

In order to verify whether the method of this paper can code the high-dimensional imbalance data of the experimental national accounts, the high-dimensional imbalance data of a city's national accounts are randomly selected from the experimental data set and integrated and coded using the method of this paper, and the coding results of the excerpted parts are shown in Table 4.

Table 4: Coding of high dimensional unbalanced data in national economic accounting (section table)

Data encoding	First level coding	Content	Secondary encoding	Content
2110205	2	National economy industry	5	Producer index
2110308	3	National economy retail enterprises	8	Operational index
2110401	4	Fixed assets in national economic accounting	1	Investment distribution data
2110502	5	Current assets in national economic accounting	2	Overall distribution data
2110602	6	Intangible assets in national economic accounting	2	Overall distribution data
2110701	7	Provincial National Economic Operation Status	1	Municipal level economic operation situation
21118Y1	18	2008 National Economy	Y1	Economic performance in January

The data encoding in Table 4 shows the implementation scheme of our method for encoding high-dimensional imbalanced data in national economic accounting. Specifically, the encoding scheme is divided into two levels: first level encoding and second level encoding. First level coding usually represents the major category or main classification of data, such as "2" representing "national economic industry", "3" representing "national economic retail enterprise", etc. These numbers are short and representative, making it easy to quickly identify and classify data.

The second level encoding further refines the specific content or attributes of the data, such as "5" representing "producer index" under the first level encoding "2", "8" representing "operation index" under the first level encoding "3", etc. This type of secondary encoding not only increases the accuracy of data description, but also helps to more accurately locate the required information during data analysis.

Analysis of Table 3 shows that: after using the method of this paper to code the high-dimensional imbalance data of national economic accounting, each expression of the data can be expressed in numbers or letters, for example, in Table 3, "2110205" stands for "national economic accounting industrial producer index", "21118Y1" stands for "January 2018 economic operation of the national economy", etc., where "211" stands for "January 2018 economic operation of the

national economy", and the last four digits represent a class code and a secondary code, respectively, so as to comprehensively describe the national economic accounting data, which further proves that the method of this paper can effectively encode the high-dimensional imbalance data of the national economic accounting, so as to improve the effect of the integration of the national economic accounting data.

In order to verify the superiority of the fuzzy clustering method proposed in this article in integrating high-dimensional imbalanced data of national economic accounting, fuzzy clustering was compared with K-means clustering method and hierarchical clustering method. The experimental dataset will still use the dataset described earlier, and one of the datasets will be randomly selected for the experiment. The following are the comparison results of three clustering methods.

Table 5: Comparison of clustering methods

Method	First level coding	Content	Secondary encoding	Content
Fuzzy Clustering	0.92	0.90	0.93	120
K-means Clustering	0.85	0.82	0.87	80
Hierarchical Clustering	0.88	0.86	0.89	180

Table 6: Comparison of dimensionality reduction methods

Method	Stress Function	Dimensionality Reduction Quality	Runtime (seconds)
Kernel PCA	0.017	0.86	150
PCA	0.035	0.78	100
t-SNE	0.022	0.80	240

According to Table 5, fuzzy clustering outperforms K-means clustering and hierarchical clustering in adjusting mutual information, adjusting Rand coefficient, and variance ratio criteria, indicating that the fuzzy clustering method has higher accuracy and effectiveness in integrating high-dimensional imbalanced data in national economic accounting.

To verify the superiority of the proposed kernel PCA method in dimensionality reduction of high-dimensional imbalanced data in national economic accounting, kernel PCA was compared with PCA and t-SNE dimensionality reduction methods. The experimental dataset will still use the dataset described earlier, and one of the datasets will be randomly selected for the experiment. Here are the comparison results of three dimensionality reduction methods.

Analysis of Table 6 shows that kernel PCA outperforms PCA and t-SNE in both pressure function and dimensionality reduction quality indicators, indicating that the kernel PCA method has higher accuracy and effectiveness in dimensionality reduction of high-dimensional imbalanced data in national economic accounting.

4 Conclusion

In order to accurately reflect the overall operation of the national economy and provide a scientific and reasonable basis for policy making, the integration of high-dimensional unbalanced data of national economic accounting based on fuzzy clustering is proposed. Firstly, we use the web crawler technology to obtain the high-dimensional unbalanced data of national economic accounting to make the basis for data integration; based on the principal component analysis, we balance the high-dimensional unbalanced data of national economy and complete the data under-sampling treatment, so that the pre-processed data can maximize the possibility of achieving the state of relative balance between the majority class and the minority class; based on the principal component analysis method, we downsize the high-dimensional balanced data of national economy, this completes the pre-processing of the high-dimensional imbalance data of national economy; using fuzzy clustering, the obtained national economic accounting data are combined into a cluster to realize the integration of the high-dimensional imbalance data of national economic accounting, so as to provide strong support for economic development.

This article uses clustering performance indicators such as Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI) to evaluate the integration effect. Compared with traditional clustering methods such as k-means, hierarchical clustering, or DBSCAN, fuzzy clustering exhibits higher performance in handling high-dimensional imbalanced data. This is mainly because fuzzy clustering can handle the ambiguity of data points belonging to multiple categories, thereby more accurately reflecting the inherent structure of the data. Specifically, fuzzy clustering is more effective in handling categories with overlapping or fuzzy boundaries, while traditional hard clustering methods such as k-means may perform poorly in these situations. Although hierarchical clustering can generate hierarchical clustering structures, it may face challenges in computational complexity and result interpretability when processing high-dimensional data. DBSCAN relies on density thresholds to identify clusters, which is not flexible enough for handling imbalanced data.

The experimental results show that when using the method proposed in this paper to reduce the dimensionality of these data, the maximum value of the pressure function is only 0.017, and the minimum value of the dimensionality reduction quality reaches 0.86, indicating that this method effectively reduces the dimensionality of the data while preserving its quality. The fuzzy clustering method was used to perform cluster analysis on the integrated high-dimensional imbalanced data. The results showed that the three indicators of adjusted mutual information, adjusted Rand coefficient, and variance ratio criterion all exceeded 0.90, and showed superiority compared to other clustering methods such as K-means clustering and hierarchical clustering. Meanwhile, the comparison results between the kernel PCA dimensionality reduction method and PCA and t-SNE dimensionality reduction methods show that kernel PCA outperforms the other two methods in terms of pressure function and dimensionality reduction quality, verifying the high accuracy and effectiveness of kernel PCA in dimensionality reduction of high-dimensional imbalanced data in national economic accounting.

In summary, the method proposed in this article for integrating high-dimensional imbalanced data in national economic accounting based on fuzzy clustering demonstrates higher performance in handling high-dimensional imbalanced data. By comparing with SOTA technology, we can find that fuzzy clustering is more effective in processing data with fuzzy boundaries and overlapping categories, while kernel PCA can more effectively extract nonlinear features from the data. These advantages make the method proposed in this article more accurate and reliable in handling complex national economic accounting data, providing a more scientific and reasonable basis for policy-making.

By using PCA dimensionality reduction to process high-dimensional imbalanced data in national economic accounting, key information can be effectively preserved, complexity and sparsity can be reduced, which helps to improve the accuracy and efficiency of policy decision-making and economic modeling. Fuzzy clustering methods can accurately reflect the economic characteristics of high-dimensional data, such as differences in labor market and economic conditions, and provide targeted recommendations for policy formulation. The accuracy of web crawling technology is crucial for the integrity of economic datasets. The method proposed in this article can accurately capture comprehensive economic data, providing reliable support for policy-making and economic research. Overall, the method proposed in this article has achieved significant results in data integration, improving data processing efficiency and accuracy, and providing strong data support for national economic decision-making.

Although kernel PCA has its advantages in processing high-dimensional data, it still faces challenges when dealing with high-dimensional spatial complexity data in the economic field. Economic datasets contain a large number of variables (such as labor dynamics, household finances, health and nutrition status, etc.) that may be interrelated and highly nonlinear. Kernel PCA transforms these variables into a new feature space through nonlinear mapping, which may reveal complex structures hidden in the original data. However, due to the limitations of time and space complexity, kernel PCA may not be directly applicable to very large economic datasets.

To overcome these limitations, researchers may adopt strategies such as using approximation algorithms to accelerate the computation and feature decomposition of kernel matrices, or using distributed computing frameworks to process large-scale datasets. In addition, feature selection or pre dimensionality reduction can be used to reduce the dimensionality of input data, thereby reducing the computational burden of kernel PCA.

References

[1] Rashid, A., Nakib, T. H., & Shahriar T. abib M.A. Hasanuzzaman M. (2024). Energy and economic analysis of an ocean thermal energy conversion plant for Bangladesh: A case study. Ocean engineering, 293(Feb.1):1.1-1.17. https://doi.org/10.1016/j.oceaneng.2023.116625.

[2] Dev, K., Chih-Lin I, & Khowaja, S. A. (2023). Guest editorial dense - data integrity, integration and security issues for consumer data in industry 5.0. IEEE Transactions on Consumer Electronics, 69(4):809-812. https://doi.org/10.1109/TVT.2024.3399470.

[3] Gallo-Bernal, S., Pea-Trujillo, V., Briggs, D., Machado-Rivas, F., Pianykh, O. S., & Flores, E. J., et al. (2024). A data science-based analysis of socioeconomic determinants impacting pediatric diagnostic radiology utilization during the COVID-19 pandemic. Pediatric radiology, 54(11):1831-

1841. https://doi.org/10.1007/s00247-024-06039-8.

[4] Liu R. Yang F., Wang. (2022). Incremental clustering algorithm for high dimensional data based on improved spark technology. Computer Simulation, 39(12), 383-386, 444. https://doi.org/10.3969/j.issn.1006-9348.2022.12.070.

[5] Cheng, Y., & Su, J. (2024). Economic data forecasting through interval data analysis. International Journal on Artificial Intelligence Tools, 33(07), 2440002. https://doi.org/10.1142/S0218213024400025.

[6] Ikoma, E., & Kitsuregawa, M. (2023). DIAS-earth environment data integration and analysis system. Communications of the ACM, 66(7):85-86. https://doi.org/10.1145/3589233.

[7] Yang, G., Li, X., Yu, T., Wu, S., & Liu, Y. (2022). A new model of environmental-economic coordination prediction using credible neural network integration and big data analysis. Security and Communication Networks, 2022(1), 3454821. https://doi.org/10.1155/2022/3454821.

[8] Han, S., Ma, H., Taherkordi, A., Lan, D., & Chen, Y. (2024). Privacy-preserving data integration scheme in industrial robot system based on fog computing and edge computing. IET communications, 18(7):461-476. https://doi.org/10.1049/cmu2.12749.

[9] Dong, S., & Tsai, S. B. (2021). Economic management data envelopes based on the clustering of incomplete data. Mathematical Problems in Engineering, 2021(1), 4312842. https://doi.org/10.1155/2021/431

[10] Silva, L., & Barbosa, L. (2024). Improving dense retrieval models with LLM augmented data for dataset search. Knowledge-based systems, 294(Jun.21):1.1-1.9. https://doi.org/10.1016/j.knosys.2024.111740.

[11] Stasenko, M., & Quinn, G. P. (2023). Stasenko, Marina, Quinn, Gwendolyn P. Improvements in sexual orientation and gender identity data collection through policy and education. American Journal of Public Health, 113(8):834-835. https://doi.org/10.2105/AJPH.2023.307344.

[12] Angaman, K. V., Mirzabaev, A., & Niang, B. B. (2024). Economic impacts of land degradation: Evidence from Côte d'Ivoire. Land Degradation and Development, 35(4):1541-1552. https://doi.org/10.1002/ldr.5004.

[13] Chatzimparmpas, A., Paulovich, F. V., & Kerren, A. (2023). HardVis: Visual analytics to handle instance hardness using undersampling and oversampling techniques. Computer Graphics Forum: Journal of the European Association for Computer Graphics, 42(1):135-154. https://doi.org/10.1111/cgf.14726.

[14] Lin, C., Tsai, C. F., & Lin, W. C. (2023). Towards hybrid over- and under-sampling combination methods for class imbalanced datasets: an experimental study. Artificial Intelligence Review: An International Science and Engineering Journal, 56(2):845-863. https://doi.org/10.1007/s10462-022-10186-5.

[15] Ephzibah, E. P., & Remya, L. (2022). Dimensionality reduction using principal component analysis and multi-label fuzzy classification for rice crop disease in rural areas of India. ECS transactions, 107(1):16451-16458. https://doi.org/10.1149/10701.16451ecst.

[16] Liu, Z., & Letchmunan, S. (2024). Enhanced fuzzy clustering for incomplete instance with evidence combination. ACM transactions on knowledge discovery from data,18(3):72.1-72.20. https://doi.org/10.1145/3638061.

[17] Madan, S., Komalavalli, C., Bhatia, M. K., Laroiya, C., & Arora, M. (2024). An optimized SVM? RFE based feature selection and weighted entropy Kmeans approach for big data clustering in MapReduce. Multimedia Tools and Applications, 83(30):74233-74254. https://doi.org/10.1007/s11042-023-18044-4

[18] Quintana-Orti, G., Hernando, F., & Igual, F. D. (2023). Algorithm 1033: Parallel Implementations for computing the minimum distance of a random linear code on distributed-memory architectures. ACM transactions on mathematical software,49(1):8.1-8.24. https://doi.org/10.1145/3573383

[19] Wang, L., Witherden, F., & Jameson, A. (2024). An efficient GPU-based h -adaptation framework via linear trees for the flux reconstruction method. Journal of Computational Physics, 502(3 Pt.1): ARTN 036108-036128. https://doi.org/10.1016/j.jcp.2024.112823.

[20] Richard D., L., Sabyasachi, S., & Ankani, Chattoraj, Ralf. M. Haefner. (2023). Bayesian encoding and decoding as distinct perspectives on neural coding. Nature neuroscience,26(12):2063-2072. https://doi.org/10.1038/s41593-023-01458-6.