Icon Visualization Application Based on PCA in Data Mining

Xiulin Jiang[1*]

[1]Computer Department, Bengbu Medical College, Bengbu 233030 Anhui, China

***Corresponding Author, Email: 20184714013@stu.qhnu.edu.cn, xiulin_jiang57@outlook.com**

Abstract

This paper explores the Principal Component Analysis (PCA) technology used for data analysis and dimensionality reduction to improve the shortcomings of current visualization technology and the mining ability of existing technologies for practical data. Firstly, PCA's fundamental principles and applications in data mining and its specific application modes in icon visualization are introduced. In terms of icon visualization, PCA-based methods can convert high-dimensional data into two- or three-dimensional graphical forms, making the structure and patterns of the dataset more intuitive and visible. This visualization helps discover features like clusters, outliers, and data distribution in the dataset, helping users better understand the intrinsic structure and potential associations. This paper proposes an active learning method based on uncertainty visualization, which aims to assist users in sample selection by visualizing the results. Experiments are conducted on WINE, Auto-MPJ, and WDBC datasets. The results show that in most cases, the result of multi-label optimization corresponds to a more minor mean square deviation of the data than the result of regular splitting. On the WDBC dataset, when the abstract cluster is 10, the mean square deviation of the data is only 0.094. Thus, multi-label optimization can automatically remove unnecessary labels based on energy from the initial label set. The research results have a specific reference value for applying and expanding the PCA method in icon visualization.

Key words: data mining; PCA; icon visualization; multi-label optimization; active learning

Introduction

In today's information age, the scale and complexity of data are increasing, making Data Mining (DM) a critical task [1-3]. Data from various sources, including time series, geographical, transactional, behavioural, scientific, social, and operational, may all be analyzed using data mining techniques. Understanding patterns and trends through this technique facilitates strategic decision-making in social media, business, finance, healthcare, and industrial processes. DM is discovering useful information and patterns in large amounts of data. It can help people uncover the hidden patterns behind data and provide valuable insights and guidance to decision-makers [4,5]. DM is discovering underlying patterns, associations, and information from large-scale data sets. In DM, data visualization is an essential tool that can display complex data in a graphical form to help users better understand and discover patterns, trends, and anomalies in the data. Intelligent data analytics automatically discovers patterns, correlations, and trends in data by applying artificial intelligence and machine learning techniques to provide insights and precise predictions. Among the many methods of DM, icon visualization is a powerful tool that can present data to users in an intuitive and easy-to-understand way so that they can better understand the internal structure and characteristics of data [6-8]. Using techniques like PCA, t-SNE, and UMAP, reducing dimensions, effectively aggregating data, enabling hierarchical visualization, managing data efficiently, utilizing parallel computing, incorporating interactive

visualization techniques, optimizing rendering, designing a responsive user interface, and continuously monitoring and optimizing performance are some ways to improve icon visualization for large datasets.

Principal Component Analysis (PCA) is a commonly used data dimensionality reduction technique in icon visualization. PCA uses a linear transformation to transform raw data into a new set of orthogonal variables called principal components. Principal components capture the direction of the most significant variance in the data and convert it into a new coordinate system. A mathematical technique called principal component analysis (PCA) determines which variables in a data set have the highest variation and computes their principal components. Minimizing dimensionality and extracting important characteristics entails standardization, covariance matrix computation, eigenvalue and eigenvector calculation, sorting, principal component selection, and transformation. As a result, PCA can reduce high-dimensional data to lower dimensions while retaining as much information as possible from the original data. A procedure that keeps the primary information while decreasing the dimensionality of high-dimensional data. Data must be standardized, the covariance matrix must be calculated, the covariance matrix must be broken down, the top k eigenvectors must be chosen, and the data must be transformed into a new k-dimensional space. In DM, high-dimensional data can be mapped to low-dimensional spaces while retaining the primary information in the data by applying PCA. The dimensionality of the data can be reduced by reducing dimensionality, and the complexity of the data can be simplified to help visual presentation. One of the goals of DM is to discover patterns and trends in data to reveal underlying information. Data can be mapped into two- or three-dimensional space to better observe the structure and distribution of data and discover hidden patterns and associations by combining PCA and icon visualization. PCA improves data mining visualization, especially for applications using icon visualization. It improves icon visualization, distils important patterns from complicated data, and simplifies it. As a result, data analysis and exploration have become more efficient and natural.

Icon visualization is an intuitive and effective way to visualize data. It makes data easier to understand and analyze using icons, graphics, and symbols. The application of icon visualization in DM can help users better identify patterns and trends in data, discover correlations between data, and spot abnormal values and outliers. Data tends to be high-dimensional in many fields, such as biology, finance, and social networks. It contains many features and attributes. Visualizing and analyzing high-dimensional data is a challenging task, and PCA can help reduce the dimensionality of the data and provide better visualizations. With the continuous advancement of computer technology and the development of visualization tools, more and more DM and analysis software have begun to integrate PCA-based icon visualization functions. These tools provide a user-friendly interface and interactive operation, making DM tasks more convenient and efficient. Data mining operations are automated using various techniques, including AutoML, scripting, template-based approaches, interactive platforms, parameterization scripts, MLOps tools, pre-built libraries, parameter tweaking tools, and interactive notebooks to ensure efficiency and user flexibility.

Visualization and visual analysis have become essential research directions in data visualization to support decision-makers in better understanding the uncertainty in data and enable them to use data more effectively to make decisions. Decision-makers can more clearly perceive bias and the degree of uncertainty in the data by presenting uncertainty in intuitive charts, graphs, or interactive visualization tools. This visualization helps them identify potential data issues, assess the risks of their decisions, and take appropriate action to mitigate the impact of uncertainty. Here, after sorting out the PCA-based icon visualization technology, an active learning method based on uncertainty visualization is proposed,

which aims to assist users in sample selection by visualizing the results. This paper expects to provide interactive data exploration support for large-scale data.

The rest of the paper is structured as follows: Section 2 describes the literature review; Section 3 illustrates the materials and methods, including the PCA method for DM dimensionality reduction processing; Section 4 shows the Results and discusses them; and Section 5 summarizes the study's conclusion along with its future scope.

Literature review

The research on PCA-based chart visualization has attracted widespread attention. Many researchers have analyzed it, especially in practical applications, and proposed many applicable methods and models.

Wang (2020) et al. proposed that PCA-based icon visualization could better reflect the intrinsic structure and essential characteristics of data, helping users understand data more comprehensively [9]. Intelligent data analysis technology can also combine dimensionality reduction data with other machine learning algorithms to achieve deeper DM and prediction. Using methods like PCA, dimensionality may be diffused to improve interpretability, decrease complexity, minimize noise, prevent overfitting, improve visualization, strengthen clustering and classification, conserve storage, and generate new features. This methodology streamlines data analysis, enhances efficiency, and facilitates improved data administration. Pouamoun (2021) et al. discovered more hidden patterns and associations by feeding PCA dimensionality reduction data into classification, clustering, or regression models to provide more accurate guidance for decision-makers [10-12]. Zhao (2022) et al. believed that through PCA dimensionality reduction, the dimensionality of the dataset could be reduced to simplify the analysis and interpretation of the data. High-dimensional data was often difficult to visualize, while reduced dimensionality data could be more easily presented in two- or three-dimensional space to help users better understand the data [13]. Ranjan et al. (2022) argued that PCA could extract the main features in the data and visualize these features. This was important for discovering patterns, clusters, and outliers in data. PCA-based icon visualization could also help users find potential relationships and interactions in the data, facilitating a deeper understanding of the data [14-16]. The study by Bovkir et al. (2021) compared multiple dimensionality reduction techniques, including PCA, for visualizing the effects of high-dimensional data. Through experimental comparison, the application effect of PCA in icon visualization was shown [17,18]. A paper by Zhang et al. proposed a method for data dimensionality reduction and icon visualization using PCA and applied it to interactive data visualization. The visualization of reduced dimensionality data was demonstrated using icon types, such as scatterplots and parallel coordinate plots, and interactive control and analysis functions were provided [19-21]. Yu's (2023) paper introduced a PCA-based dimensionality reduction technique called t-stochastic Neighbor Embedding (SNE).

T-SNE mapped high-dimensional data into two- or three-dimensional space by preserving the similarity relationship between data samples to achieve visual display [22,23]. To represent data points, manage heavy-tailed distributions, balance local and global structure, use the Barnes-Hut approximation for scalability, and perform iterative embedding optimization, t-SNE employs stochastic neighbour embedding with t-distribution. Piippo et al.(2022) proposed a PCA-based evolutionary computation method for visualizing high-dimensional data. They combined PCA and genetic algorithms to visually display data by optimizing the projected coordinates of the data [24-26]. Tuarob (2021) et al. proposed a projection-based PCA method for interactive visualization of high-dimensional data. Data was

mapped to two-dimensional space through projection and reprojection techniques, and interactive exploration and analysis capabilities were provided to help users discover patterns and trends in the data [27,28]. A paper by Fu (2020) et al. introduced an icon visualization method based on PCA and variable-level information to explore and analyze high-dimensional data. A visualization framework was proposed to visually present hierarchical details in data and provide interactive exploration capabilities [29-32]. The research offers an e-healthcare risk prediction system for health big data centred around licensed medical practitioners and powered by a heterogeneous network. It enhances prediction accuracy, monogenic score, density accuracy, execution time, and overall efficiency by 73.98% [33]. The usage of solar and wind energy to generate power is covered in the article, along with topics including cost, variability, and non-concentrated and diluted energy. It forecasts how plants react to temperature, light intensity, and humidity using crop production systems and artificial neural network-based expert systems [34]. Table 1 shows the critical components analysis of the literature review:

Table 1: Critical analysis component of the literature review

References	Method	Results	Limitations
[9]	In this paper, a comprehensive similarity life prediction method is proposed for analyzing rolling bearings using multi-dimensional feature fusion	The multi-dimensional feature prediction algorithm improves the accuracy and reliability of predicting rolling bearing life, supporting better predictive maintenance and health management.	This method's limitations include the need for extensive full-life tests to improve prediction accuracy and the focus on time-domain features, which restricts the depth of data analysis.
[13]	This paper developed a MATLAB® coding basis and integrated method, 'Ana', for data visualizations and statistical analysis	The code efficiently prints high-quality figures up to 150 or 300 dpi, providing enough contrast to differentiate the omics dataset. It is compatible with Windows and MacOS operating systems and is quick and efficient for publication and presentation.	However, this paper does not offer appropriate code annotations for undergraduate and postgraduate students to learn the coding basis of statistical data analysis.
[14]	The Sequence Graph Transform (SGT) is introduced as a feature embedding function that can capture a range of short- to long-term dependencies	In sequence clustering and classification, SGT outperforms existing methods such as sequence/string Kernels and LSTM, delivering higher accuracy with reduced	The proposed model is complex to implement, computationally intensive, sensitive to data variations, and can be prone to overfitting. It is also limited to non-

	without adding to the computational load.	computational demands.	sequential data and presents challenges in interpreting its features.
[17]	This study investigates big data visualization methods in smart cities by analyzing GIS-integrated dashboard examples and developing an open-source GIS-based dashboard with Apache Superset.	Data visualization uncovers the relationships between data, creating flexible and innovative connections between human perception and computer systems for enhanced understanding.	The primary challenges in visualizing big data in this study are missing data and visual noise.
[21]	CancerMIRNome was developed to support the mining of miRNA expression data from the Cancer Genome Atlas and extensive profiling studies.	The data analysis and visualization modules will significantly enhance the use of valuable resources and advance the practical application of miRNA biomarkers in cancer research.	This study has limitations, including potential integration challenges, user interface issues and technical constraints.
[22]	This study employs cheminformatics and machine learning techniques to explore the chemical space, scaffolds, structure-activity relationships, and landscape of human androgen receptor (AR) antagonists.	The study's findings offer new insights and recommendations for hit identification and lead optimization in the development of novel AR antagonists.	The study uses data from the ChEMBL database, which may be incomplete due to missing records of some assays and experiments. This leads to gaps in valuable data like additional scaffolds.

Research on PCA-based icon visualization applications includes data dimensionality reduction, anomaly detection, and data clustering and classification. It provides a rich theoretical and practical foundation that can be used to apply PCA and icon visualization techniques in DM tasks effectively.

Materials and Methods

PCA method for DM dimensionality reduction processing

Dimensionality reduction processing in data intelligence analysis is a commonly used technique to reduce the dimensionality of a high-dimensional data set while preserving the primary information in the data. In data intelligence analysis, dimensionality reduction reduces multicollinearity and identifies essential characteristics to simplify complicated datasets and enhance computing efficiency, model performance, visualization, and insights. PCA maps the original data to a new feature space through a

linear transformation that enables the latest features to explain most of the variance in the data [35]. PCA improves efficiency, simplifies visualization, and highlights important patterns to improve data intelligence compared to conventional techniques. It concentrates on high-variance directions, minimizes noise, and simplifies complicated data. PCA improves scalability, lowers computing burden, and facilitates interpretation for more extensive datasets.

Suppose a set of points $\{x_1, x_2 \ldots x_n\}$ is given in the space of R^d. PCA aims to find a projection matrix $W = [w_1, w_2 \ldots w_m]$ that can reflect the difference of the original spatial data to the greatest extent. Assuming the data has been decentralized, PCA can be expressed as a minimal reconstruction error model of Eq. (1).

$$\min_W e(x_i) \quad s.t. W^T W = I \tag{1}$$

$$e(x_i) = \sum_{i=1}^{n} \|x_i - WW^T x_i\|_2^2 \tag{2}$$

One approach to improve the PCA's robustness is to change the ℓ_2 norm in the PCA function to the ℓ_1 norm as a distance metric [36-38]. Numerous methods, such as sparse PCA, regularised PCA, incremental PCA, robust data preprocessing, kernel PCA, robust PCA (RPCA), and assembling PCA, can enhance principal component analysis (PCA). These techniques improve PCA's performance and reliability by tackling the data's outliers, noise, and non-normality. The idea of reconstruction error minimization is to calculate the reconstruction error using the ℓ_1 norm as a distance metric, and the model for solving the projection direction can be expressed as:

$$\min_W \sum_{i=1}^{n} \|x_i - WW^T x_i\|_1 \tag{3}$$

PCA with reconstruction error minimization improves data intelligence by simplifying complicated data, reducing noise and recognizing critical patterns, improving icon visualization and ease of understanding, and reducing computing burden. Furthermore, feature extraction, anomaly detection, noise reduction, data compression, and visualization are improvements. The idea of variance maximization is to change the ℓ_2 norm of mean squared to $the\ \ell_1$ norm based on PCA. The model of Eq. (4) can calculate the principal component direction.

$$\min_W \sum_{i=1}^{n} \|W^T x_i\|_1 \tag{4}$$

These two lines of thinking are often interrelated. The ℓ_1 norm-based reconstruction error minimization and variance maximization modeling in PCA have the problem of criterion inequality, while Angle PCA can take advantage of information hidden in the dataset. The equation for determining principal components by Angle PCA can be expressed as:

$$\max_W \sum_{i=1}^{n} \frac{\|W^T x_i\|_2^1}{\|x_i - WW^T x_i\|_2^1} \tag{5}$$

In Eq. (5), each term can be thought of as a cotangent term of the angle between the *i*th data projection variance and the reconstruction error, which is expressed as:

$$\frac{\|W^T x_i\|_2^1}{\|x_i - WW^T x_i\|_2^1} = cot\,\alpha_i \quad (6)$$

PCA based on the $\ell_{2,p}$ norm preserves the rotational invariance of PCA, and its objective function can be defined as:

$$\min_W \sum_{i=1}^n \|x_i - WW^T x_i\|_2^p \quad (7)$$

In Eq. (7), the value of p is (0,2), and the value of p can be changed according to the actual situation.

PCA is one of the representatives of global dimensionality reduction technology. Local preservation projection (LPP), as one of the most commonly used local dimension reduction techniques, can maintain the local structural characteristics of data [39,40]. LPP must first construct an initial adjacency graph $M = \{X, S\} M = \{X, S\}$. S refers to the similarity matrix, and its Gaussian kernel function is defined as:

$$s_{ij} = \begin{cases} exp\left(-\|x_i - x_j\|_2^2 / 2\sigma^2\right), x_i \in N_k(x_j) \vee x_j \in N_k(x_i) \\ 0 \qquad\qquad ,otherwise \end{cases} \quad (8)$$

In Eq. (8), $N_k(x_j)$ represents the k nearest neighbours of x_j. σ is a Gaussian kernel parameter. The objective function of LPP is:

$$\min_W \sum_{i,j=1}^n \left\|W^T x_i - W^T x_j\right\|_2^2 s_{ij} \quad (9)$$

From a simple mathematical derivation, Eq. (10) can be obtained.

$$\min_W \frac{Tr(W^T X L X^T W)}{Tr(W^T X D X^T W)} \quad (10)$$

In Eq. (10), D is the similarity matrix S degree matrix, and L is the Tulaplacian matrix.

In intelligent data analysis, PCA technology can reduce the dimensionality of data. Additionally, the importance of the newly solved "principal element" vector is sorted. The most important part of the front is taken, and the latter dimension is omitted, which can achieve dimensionality reduction to simplify the model or compress the data [41-43]. The PCA logic block diagram is shown in Figure 1.

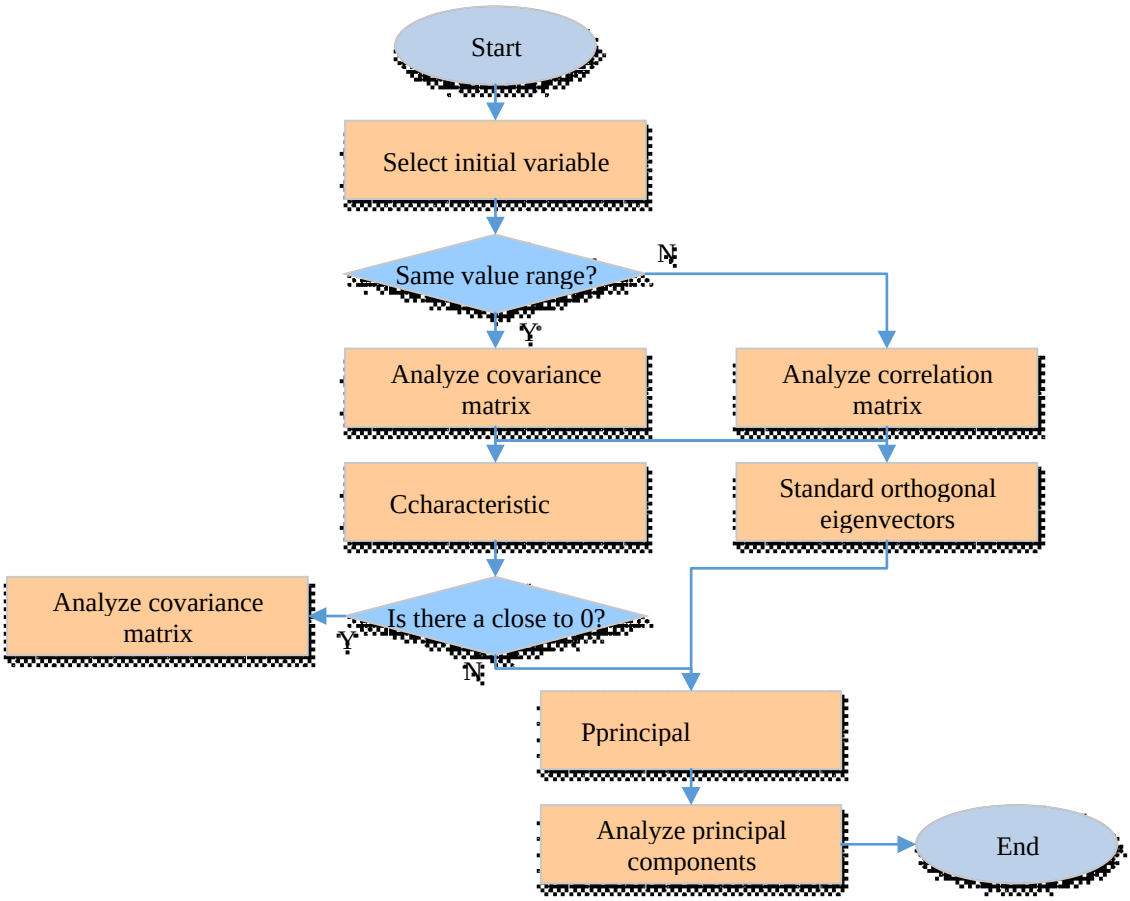

Figure 1: PCA logic block diagram

In Figure 1 above, PCA technology first standardizes the data to ensure each feature has the same scale in the dimensionality reduction of the data set. The core of this technology is to calculate the covariance matrix between features, then get the eigenvalues and corresponding eigenvectors by eigenvalue decomposition of the covariance matrix, arrange the eigenvalues in descending order, and discard the values with minor variance. Finally, the principal components are selected and projected according to the requirements, and the visualization of dimensionality reduction data is realized.

PCA-based icon visualization technology

In DM, icon-based multi-dimensional visualization techniques are a standard method. Its core idea is to use icons to represent multi-dimensional data and express multiple data dimensions through visual features, such as icons' size, length, shape, and colour [44]. The goals of icon-based multi-dimensional visualization techniques in research are to simplify complex structures, communicate meaningful data attributes, show data patterns intuitively, enable interactive exploration, adapt to particular application domains, prioritize visual appeal, and guarantee effective dataset handling. By facilitating generalization, strengthening interpretability, lowering noise, and boosting model resilience, dimensionality reduction approaches like PCA and t-SNE enhance machine learning models. Additionally, reducing the need for storage allows larger datasets to be processed and analyzed more quickly without depleting system capacity. Compared with other multi-dimensional data visualization methods, the icon-based visualization technique suits datasets with few dimensions and unique attributes [45-47]. In icon visualization, PCA decreases dimensionality using the following methods:

preprocessing high-dimensional data, computing covariance matrix, decomposing eigenvectors, choosing top eigenvectors, projecting data onto principle component subspace, and interpreting visualization for insights. For datasets with intricate structures and categorical features, icon-based visualization provides a visually appealing and semantically relevant data interpretation method. It enables interactive exploration, lowers dimensions, draws attention to patterns, strengthens narrative and communication, and makes qualitative and comparative analysis easier. The data visualization flowchart is shown in Figure 2. The icon visualization used in PCA technology has limitations, including subjectivity, loss of detail, interpretation challenges, restricted linear correlations, trouble with categorical data, outliers, limited scalability, variance overemphasis, preprocessing reliance, and missing values.

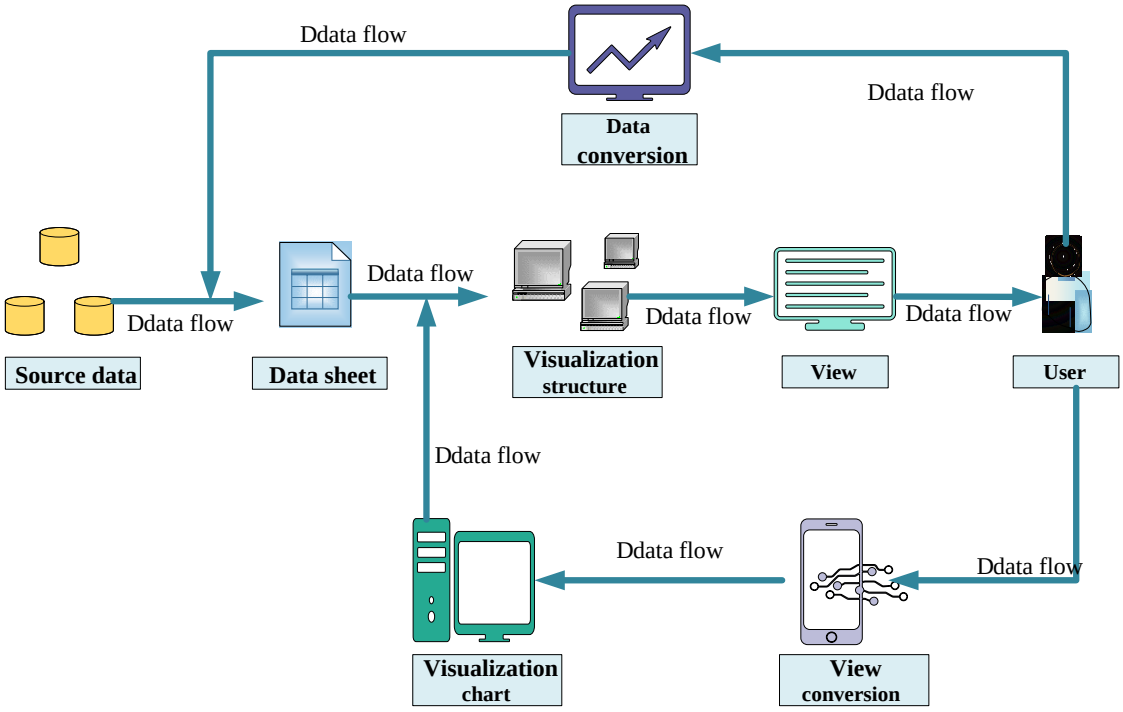

Figure 2: Data visualization process

In Figure 2 above, the PCA method realizes the dimensionality reduction operation of high-dimensional data based on PCA icon visualization, and the data dimensionality reduction results are presented in charts. Icon visualizations can represent multiple dimensions of gene expression data, revealing patterns of interactions between genes and biological processes [48]. This intuitive visualization makes the data analysis process easier and more efficient while facilitating the discovery of potential correlations and trends in the data.

In mathematics, PCA finds a projection direction so that the data projection in this direction has the most significant variance, called the first principal component. The explained variance ratio measures the variance explained by each principal component, and the variance explained by a group of principal components is represented by the cumulative variance explained, which helps determine how many trends are necessary to minimize dimensionality. Then, PCA proceeds to look for another projection direction orthogonal to the first principal component so that the projection of the data in that direction has the second largest variance, and this direction is called the second principal component. By analogy, PCA can find a set of orthogonal principal components irrelevant to the original data.

The original data matrix A is Z-score processed to obtain the normalized matrix *Z*.

$$Z_{ij} = \frac{x_{ij} - \overline{x_j}}{\sigma_j}, 1 \le i \le N, 1 \le j \le M \quad (11)$$

$$\overline{x_j} = \frac{\sum_{i=1}^{N} x_{ij}}{N} \quad (12)$$

$$\sigma_j = \sqrt{\frac{\sum_{i=1}^{N} (x_{ij} - \overline{x_j})^2}{N-1}} \quad (13)$$

In Eq. (11)-Eq. (13), x_{ij} represents the observation. $\overline{\square_{Xj}}$ and σ_j represent the mean and standard deviation at row *j*, respectively.

Find the correlation coefficient matrix *R* for the preprocessed matrix *Z*. PCA uses the correlation coefficient matrix R for several tasks, including measuring linear connections, standardizing variables, breaking down eigenvalues, determining maximum variance directions, decreasing dimensionality, and understanding data structures. It is essential to PCA since it guarantees equal analysis contribution.

$$R = [r_{ij}]_p xp = \begin{pmatrix} r_{11} & \dots & r_{1p} \\ & \cdots & \\ r_{p1} & \dots & r_{pp} \end{pmatrix} \quad (14)$$

$$r_{ij} = \frac{\sum Z_{ki} Z_{kj}}{N-1} \quad (15)$$

In Eq. (14)-Eq. (15), r_{ij} is the correlation coefficient.

The component of the characteristic variable is the weight, and the principal component is expressed as:

$$P_i = \gamma_{i1} Z_1 + \gamma_{i2} Z_2 + \gamma_{i3} Z_3 + \ldots + \gamma_{ip} Z_p \quad (16)$$

In Eq. (16), γ_i is a feature vector.

Active learning based on data uncertainty visualization

Active learning is an iterative learning process. The learning algorithm selects the most valuable data samples for labelling based on the current model performance to improve the model's performance in traditional active learning. Active learning chooses the most illuminating data samples for labelling, which maximizes the learning process. To iteratively improve model performance with the most miniature labelling work, it prioritizes samples when the model needs clarification or more confidence in its predictions. The selection of data samples is usually based on the predictive uncertainty of the model. However, in DM, the uncertainty of the model alone may not accurately reflect the uncertainty of the actual data. The active learning process based on data visualization is given in Figure 3.

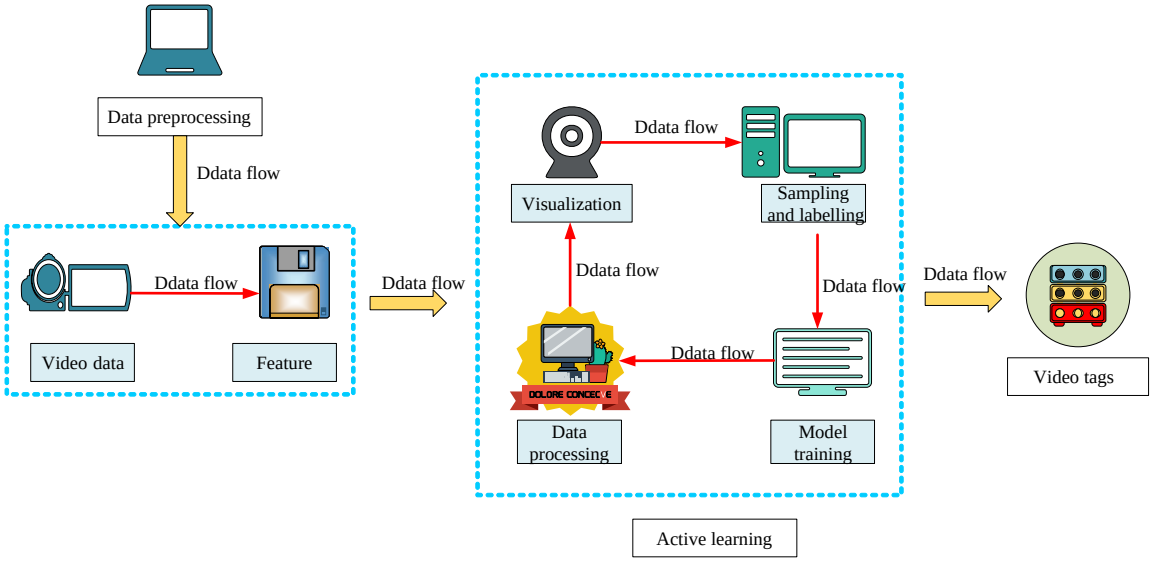

Figure 3: Active learning process based on data visualization

In Figure 3 above, active learning based on data uncertainty visualization firstly uses the initial limited labelled data set to train the model, predicts the unlabelled samples on the trained model, and calculates the uncertainty measure of each sample. Decision-making during sample selection is aided by uncertainty visualization, which helps comprehend data ambiguity. Error bars and heatmaps are techniques that prioritize high-uncertainty data and increase the trustworthiness of data-driven conclusions. Then, the uncertainty measurement is visualized together with the sample data. The uncertainty of the sample is usually displayed by scatter plot, heat map, histogram, and so on, and the sample is selected directionally. Finally, the samples are labelled, and the newly labelled samples are merged with the original limited data set. The model is retrained and iterated continuously to visualize the data uncertainty of active learning. However, for datasets with different data distributions, evaluating the appropriate parameters for each dataset is challenging in effectively assisting the active learning process. Sample imbalance, feature distribution, non-stationarity, complicated data structures, computational complexity, traditional evaluation metrics, generalization, and domain-specific concerns make it difficult to evaluate active learning parameters across various datasets. Intuitive visualization improves comprehension of complicated datasets by offering interactive exploration, drill-down capabilities, real-time manipulation, contextual information, user-centric design, and iterative analysis for well-informed decision-making. Scatterplots are an essential and indispensable technique for visualizing multi-dimensional data. When performing multi-dimensional data analysis, reducing the dimensionality or mapping the data to variables is often necessary. With a clear understanding of the data distribution, users can make intuitive and efficient sample selections. Diversity representation is ensured by active learning techniques such as reinforcement learning, density-based sampling, diversity sampling, QBC, expected model change, uncertainty sampling, margin sampling, hybrid methods, transfer learning, and meta-learning approaches. These techniques also increase the accuracy and efficiency of sample selection. To achieve this, the usual entropy method is used to estimate the uncertainty of each data point.

$$U_x = -\sum_i P(y_i|x)\, log\, P\,(y_i|x) \qquad (17)$$

In Eq. (17), x is a high-dimensional data point, and y is the possible label of x. $P(y_i|x)$ represents the probability that x will be given the label y_i.

The uncertainty of each pixel can be estimated according to Eq. (18).

$$U_p = \frac{\sum_{x \in N_P} K_H(x,p) U_x}{\sum_{x \in N_P} K_H(x,p) + \delta} \quad (18)$$

In Eq. (18), N_P represents the set of K-nearest data points for pixel p. K_H represents an adaptive Gaussian kernel function with an initial bandwidth of H.

Design of experiments

Qualitative and quantitative comparative verification methods are adopted to verify the effectiveness of the multi-label optimization clustering method in data abstraction in DM. In data mining, multi-label optimization clustering makes datasets with many features more accessible to understand, boosts abstraction, and promotes predictive modelling. It facilitates adaptable analysis and decision-making and provides insightful information for uses such as customized suggestions and client segmentation. The efficiency of clustering is assessed using both qualitative and quantitative techniques. While quantitative approaches employ statistical statistics and objective metrics, qualitative methods depend on subjective judgment and visual examination. While both approaches give information on the interpretability and coherence of clusters, quantitative approaches are more scalable and provide objective judgments. Several criteria must be considered when choosing the best technological stack for multi-label optimization clustering, including the algorithm's complexity, scalability, programming language, parallel processing, seamless integration, visualization, compatibility with the deployment environment, and community support.

For the experiment, datasets WINE, Auto-MPJ, and WDBC are frequently used in machine learning and statistical analysis due to their diverse characteristics and wide range of uses in various fields. Their significance is seen in Principal Component Analysis (PCA) and icon visualization technology.

The WINE dataset is a widely recognized dataset frequently utilized in classification and regression tasks, specifically within machine learning. It comprises 13 chemical characteristics, such as alcohol content, malic acid, and phenols, linked to three distinct wine types (varietals). By analyzing these chemical characteristics, wines can be classified into different varietals, which is crucial for ensuring quality and verifying authenticity. PCA can decrease the dimensions of a dataset, helping researchers pinpoint the key characteristics that impact wine categorization. PCA converts the data with many dimensions into a two- or three-dimensional space, making it easier to visualize and understand the connections between various types of wine.

Auto-MPJ Dataset: The Auto-MPJ dataset includes necessary measurements of car performance like horsepower, weight, acceleration, and other factors that assess the effectiveness and efficiency of different car models. Both manufacturers and consumers find value in this dataset, enabling them to evaluate and compare various car models. Using PCA helps simplify the dataset, allowing for easier detection of outliers and unique car models through performance metrics. This procedure provides car buyers and manufacturers to make educated choices by emphasizing the essential characteristics identified in PCA.

The WDBC Dataset, also known as the Wisconsin Diagnostic Breast Cancer Dataset, is a frequently used biomedical tool for diagnosing breast cancer. It includes 30 characteristics obtained from digital images of breast tumour fine needle aspirates (FNA), such as radius, texture, smoothness, and symmetry.

This dataset is essential for detecting and distinguishing between benign and malignant breast masses in the early stages. Because of its wide range of functions, utilizing Principal Component Analysis (PCA) is advantageous for decreasing dimensionality and streamlining data visualization and interpretation. PCA helps improve diagnostic models by focusing on the most significant features, resulting in increased efficiency and effectiveness.

PCA can boost the efficiency of machine learning models by decreasing the dimensionality, resulting in faster performance. Reducing dimensionality makes it faster and easier to analyze and interpret complex datasets. PCA maintains important data patterns, saving critical information and removing unwanted noise. It also allows for complex data to be transformed into 2D or 3D visualizations, making it easier to see the relationships within the data. Focusing on the most essential characteristics, PCA enhances the comprehension and insights gained from data, resulting in improved decision-making and the generation of valuable insights. In conclusion, PCA can improve performance, decrease dimensionality, and increase the comprehension and visualization of intricate multi-dimensional data, leading to better-informed decisions and insights in various industries like healthcare, automotive, and oenology. The experiment uses C++/Qt for related development, and the experimental environment is carried out on a single machine.

Table 2: Information on the dataset used for the experiment

Dataset	Data type	Number of records	Dimension	Projection method
WINE	Numerical	178	14	Multi-Dimensional Scaling (MDS)
Auto-MPJ	Numerical	290	8	t-SNE
WDBC	Numerical	569	31	MDS

Results and Discussion

Multi-dimensional data clustering results

The Wine dataset is selected and projected onto the scatterplot using the MDS method in this experiment. Figures 4 and 5 show the results of the projection. As the α value increases, the number of clusters decreases. When the α value is zero, many small clusters appear in the plot. However, as the α value rises, the small clusters fade away, and the result is more inclined to produce larger clusters. In practical applications, the selection of parameter α can be adjusted according to the actual needs of users.

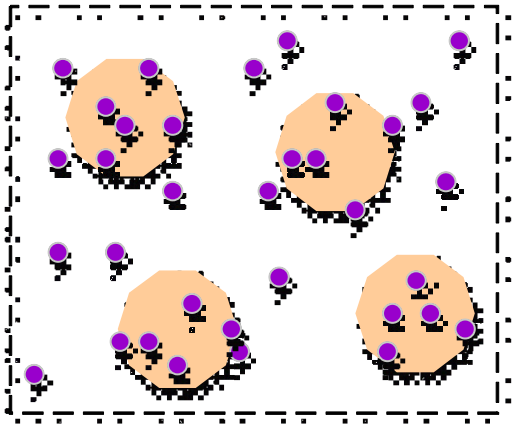

Figure 4: Abstract clustering results of data when α=0

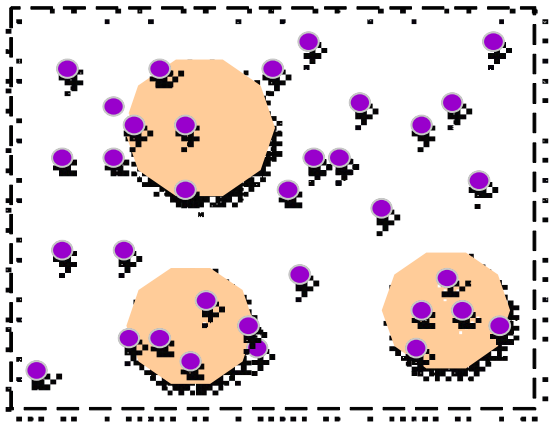

Figure 5: Abstract clustering results of data when α=10

In Figures 4 and 5 above, when α=0, four clusters appear on the projection scatter diagram, and each cluster has 4~6 data points, but with the adjustment of parameter α. When α=10, the number of clusters becomes three, each with 3~6 data points. Local linear embedding (LLE) minimizes distortion while maintaining local links between data points. It is resilient to noise, minimizes distortion, maximizes goal functions, promotes sparse representations, captures nonlinear structures, and requires parameter optimization for the best possible structure preservation. The data points and characteristics have not changed, but the number of clusters reflects the reduced data dimension. In this paper, the value of parameter α is finally determined to be 10 for subsequent analysis.

Data abstraction quality results

Data abstraction quality tests are performed on Auto-MPJ, WDBC, and WINE datasets. The data abstract quality results and the specific experimental results are shown in Figure 6 ~ Figure 8. In most cases, the result of multi-label optimization corresponds to a more minor mean square deviation of the data than the result of regular splitting. On the WDBC dataset, when the abstract cluster is 10, the mean square deviation of the data is only 0.094.

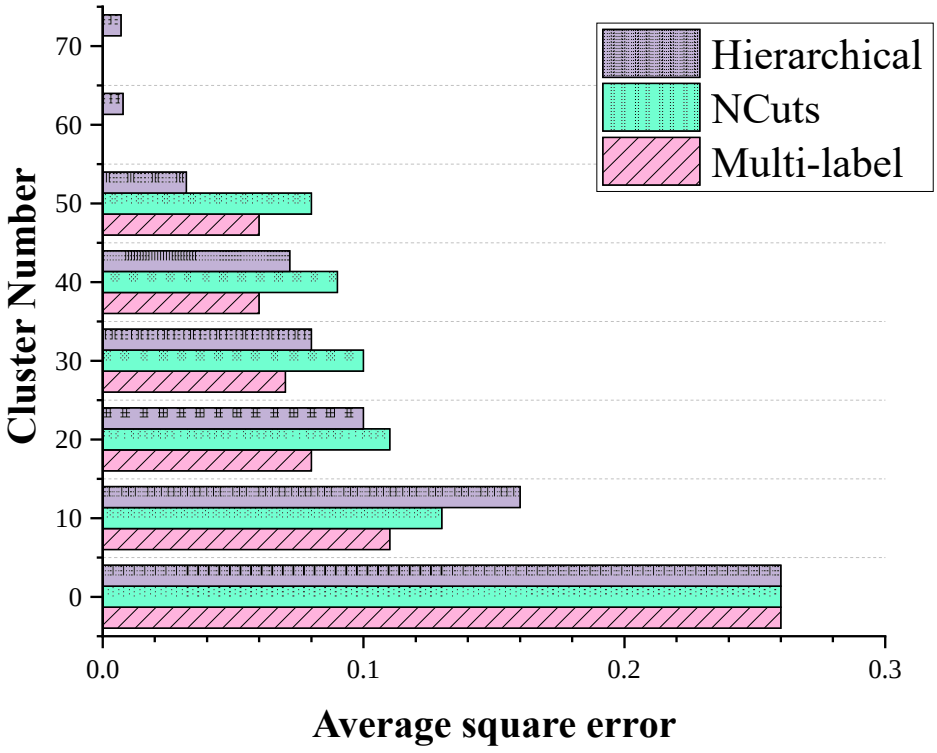

Figure 6: Data abstract quality halo results on the Auto-MPJ dataset

In Figure 6 above, on the Auto-MPJ dataset, the mean square error of multi-label optimization, regular segmentation, and hierarchical clustering all show a downward trend with increased clusters and stabilize at a small value. The mean square error of hierarchical clustering, regular segmentation, and multi-label optimization stabilizes at a minimal value as the number of clusters increases. The study offers icon visualization with a final error of 0.0637 that uses the PCA technique for precise data mining of car performance. On this data set, the icon visualization based on PCA technology proposed in this paper has the best performance and the most accurate results in data mining of automobile performance, and the final data mean square error is stable at 0.0637.

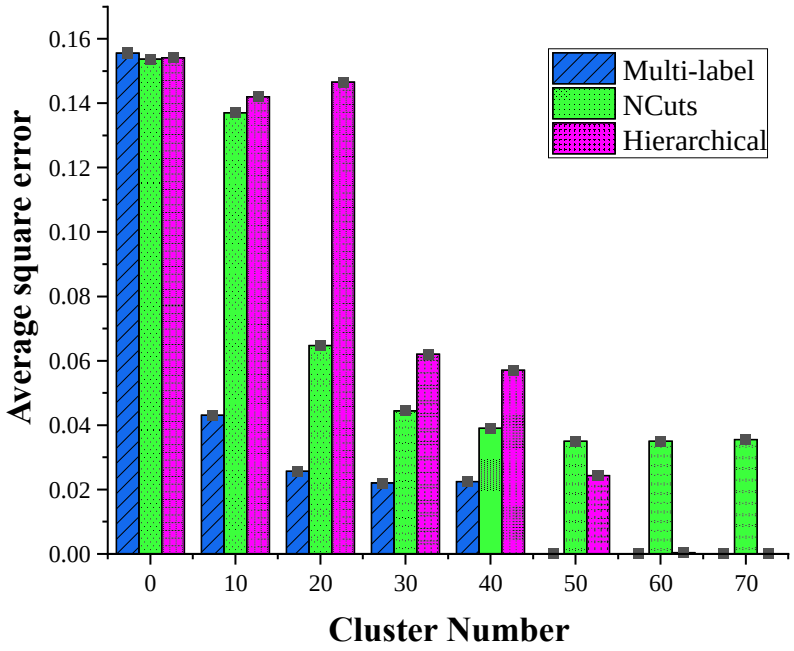

Figure 7: Data abstract quality halo results on the WDBC dataset

In Figure 7 above, among the 30 different feature recognition and visualization realization processes of breast tumour samples on the WDBC data set, the errors of regular segmentation and multi-label optimization are relatively small when the previous model training (the number of clusters is less than 30). Compared to traditional splitting, the multi-label optimization technique yields a decreased mean square deviation of the data. The WDBC dataset's mean square deviation is just 0.094 when the abstract cluster is 10. Only the hierarchical clustering method based on PCA technology designed in this paper can reduce the mean square error of data to zero with a limited number of clusters. Biomedical analysis, consumer segmentation, picture identification, environmental monitoring, and financial risk assessment are just a few sectors that employ PCA and hierarchical clustering technologies to analyze data and make decisions more effectively. This is mainly because the method designed in this paper can realize automatic learning and optimization iteration and has obvious advantages in error avoidance. The capacity of a machine learning system to learn and adapt on its own, picking up information from its surroundings and making judgments, is known as autonomous learning. Decision-making and dynamic behaviour in uncertain situations are made possible by unsupervised, self-supervised, and reinforcement learning methods. Automated Education and Enhancement Machine learning models and algorithms are optimized by iterative testing and improvement, known as iteration. It streamlines the model-building process and speeds up innovation by identifying the optimal configurations for a particular job using automated machine learning, neural architecture search, and hyperparameter tuning methods.

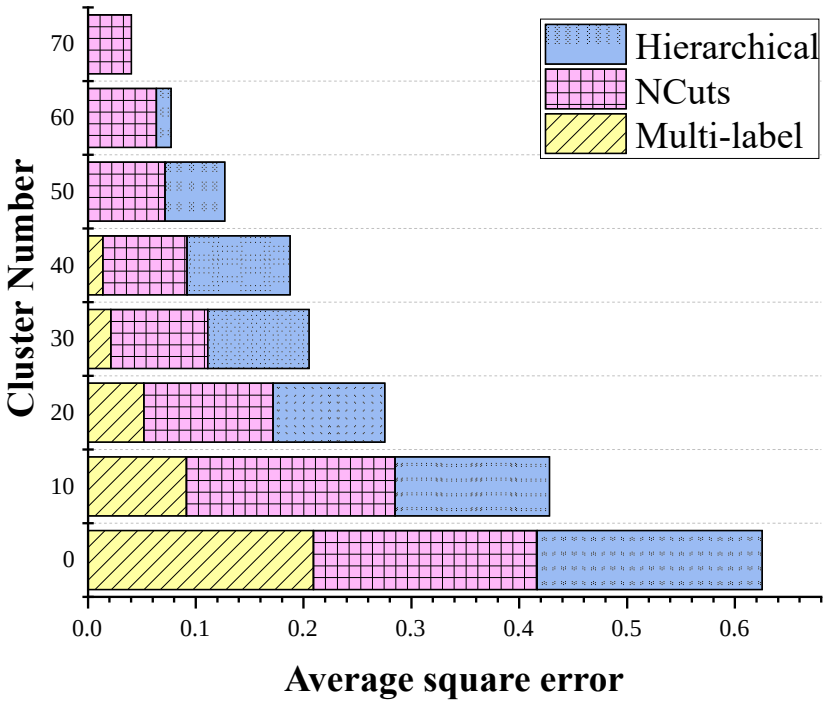

Figure 8: Data abstract quality halo results on the WINE dataset

In Figure 8 above, when α=10, both multi-label optimization and hierarchical clustering can eventually reduce the mean square error of data to about 0.7 in the training process on WINE data sets, but the hierarchical clustering method with autonomous learning ability can realize icon visualization faster and more stably in the training process. Multi-label optimization yields outcomes that are less than conventional splitting in terms of the data's mean square deviation. The suggested strategy improves the data abstraction quality for the WINE dataset, which comprises 13 distinct chemical properties.

Table 3 below shows the performance test data of the PCA-based algorithm proposed in this paper in the actual icon visualization application process:

Table 3: Performance test of PCA-based algorithm in an icon visualization application

index	Cluster number							
	10	20	30	40	50	60	70	80
Explanation variance percentage (%)	0.26	0.19	0.17	0.14	0.1	0.93	0.89	0.87
Page response time (ms)	3257	1987	1469	987	953	862	797	785
CPU utilization rate (%)	29%	35%	41%	37%	33%	36%	39%	40%

In Table 3 above, the icon visualization algorithm based on PCA proposed in this paper explains the differences. With the increase of clusters, under the protection of the autonomous learning function, the

percentage of disagreements gradually decreases and finally stabilizes at 0.87, which can effectively prevent data loss. The page response time finally stabilizes at around 800ms, with a faster response. The CPU utilization rate finally stabilizes within 40%, and the equipment occupation is regular. As shown in Table 4, five experts are invited to rate the icon visualization algorithm based on PCA in five aspects: the degree of function realization, the ability of autonomous learning, the response speed of the system, the retention of features, and the comprehensive experience of use:

Table 4: Evaluation score of icon visualization application based on PCA in the data mining field

Index	Score					Average
	Expert 1	Expert 2	Expert 3	Expert 4	Expert 5	
The degree of function realization	8	9	7	8	9	**8.2**
The ability of autonomous learning	9	7	9	8	7	8
The response speed of the system	8	8	8	7	7	7.6
The retention of features	7	8	7	9	8	7.8
The comprehensive experience of use	8	7	7	8	8	**7.6**

In Table 4 above, the score is designed to be 1-10, and the decimal point is not taken, indicating that the satisfaction degree is low to high. After the experts' scoring, the comprehensive score of icon visualization based on PCA proposed in this paper reaches 7.6, and the satisfaction degree is high. The score of overall function realization is the highest, at 8.2. In the field of data mining, the function of this algorithm is relatively complete in the application process, and it can realize data dimension reduction processing.

The results clearly show that multi-label optimization corresponds to a data mean square deviation that is smaller than the result of regular splitting in most cases. In the case of a small number of clusters, the results of multi-label optimization are also better than the results of hierarchical clustering. However, hierarchical clustering results provide a better-quality data abstraction when the number of clusters is large.

Discussion

Parsania et al. proposed an intuitive and user-friendly gene expression data analysis and visualization platform. It was designed to help laboratory scientists with little or no computer programming knowledge to achieve independent bioinformatics analysis and generate publication-ready data [49]. Yang et al. classified the astronomical literature and designed six sets of spectral datasets from data characteristics, quality, and volume to test the performance of the PCA visualization algorithm [50]. Zhou (2023) et al. proposed an origin-end-experience orthogonal function to discover important spatiotemporal flow patterns while maintaining a paired connection between the start and end points [51]. Here, the number of clusters gradually decreases with the increase of α values in multi-dimensional clustering. When the α value is zero, many tiny clusters occur. According to the needs, adjust the value of the parameter α. On the WDBC dataset, when the abstract cluster is 10, the mean square deviation

of the data is only 0.094. PCA-based chart visualization applications have significant advantages in evaluating data point uncertainty and the practical selection of samples.
In this field of data mining, compared with the traditional icon visualization method, the icon visualization based on PCA technology proposed in this paper can effectively achieve the goals of data dimension reduction, correlation elimination, visualization effect display, and noise filtering in the application process. It has effectively solved the problems of dimension limitation, information loss, visual deformation, and structural adaptation in the traditional methods. Moreover, regarding CPU occupation, data loss rate, and response time, icon visualization based on PCA has apparent advantages, among which the response time is finally stable at 800ms, which can achieve the goal of being fast and efficient, further illustrating the superiority of the research algorithm. The icon visualization based on PCA proposed in this paper is helpful to deeply understand the principle and characteristics of dimension reduction technology, to understand better and show the structure, pattern, and association of data, to help analysts intuitively understand the characteristics and structure of data, to improve the algorithm performance of data mining, to help realize data compression and storage, and to lay the foundation for the research of more complex dimension reduction and data visualization methods.

Conclusions

Here, the application of PCA-based icon visualization in DM is deeply explored, and an active learning method based on uncertainty visualization is proposed. PCA transforms high-dimensional data into low-dimensional data by projecting and transforming data while retaining as much important information as possible from the original data. This visualization method provides a better understanding of the dataset's intrinsic structure and potential associations, improving DM's effectiveness and accuracy. Users can visually assess the uncertainty of data points and make efficient sample selections by visualizing the results. Experimental results verify the effectiveness of the proposed method on WINE, Auto-MPJ, and WDBC datasets.

In conclusion, this study comprehensively studies and explores the application of PCA-based icon visualization, which provides valuable expansion and improvement ideas for visualization technology in DM. Future research can explore more datasets and application scenarios further and combine other algorithms and technologies to improve the application effect of PCA-based icon visualization in data intelligence analysis.

Declarations

Funding
No funds or grants were received by any of the authors.

Conflict of interest
There is no conflict of interest among the authors.

Data Availability
All data generated or analyzed during this study are included in the manuscript.

Code Availability
Not applicable.

Author’s contributions
All Authors contributed to the design and methodology of this study, the assessment of the outcomes, and the writing of the manuscript.

References

[1] Luo Y, Zhang X, Zhang Z, et al. Dual-principal component analysis of the Raman spectrum matrix to automatically identify and visualize microplastics and nanoplastics[J]. Analytical Chemistry, 2022, 94(7): 3150-3157.

[2] Zhang C. Feature Extraction of Color Symbol Elements in Interior Design Based on Extension Data Mining[J]. Mathematical Problems in Engineering, 2022, 2022.

[3] Mohamed A, Molendijk J, Hill M M. Lipidr: a software tool for data mining and analysis of lipidomics datasets[J]. Journal of proteome research, 2020, 19(7): 2890-2897.

[4] Johnson O V, Jinadu O T, Aladesote O I. On experimenting with large datasets for visualization using distributed learning and tree plotting techniques[J]. Scientific African, 2020, 8: e00466.

[5] Wolf J, Boneva S, Schlecht A, et al. The Human Eye Transcriptome Atlas: A searchable comparative transcriptome database for healthy and diseased human eye tissue[J]. Genomics, 2022, 114(2): 110286.

[6] Sandoval B, Barocio E, Korba P, et al. Three-way unsupervised data mining for power system applications based on tensor decomposition[J]. Electric Power Systems Research, 2020, 187: 106431.

[7] Cheng S, Li L, Yu X. Big data analysis revealed signalling activity and key regulators in human prostate cancer cell lines[J]. Cancer Research, 2023, 83(7_Supplement): 1452-1452.

[8] Roma G, Xambó A, Green O, et al. A General Framework for Visualization of Sound Collections in Musical Interfaces[J]. Applied Sciences, 2021, 11(24): 11926.

[9] Wang H, Ni G, Chen J, et al. Research on rolling bearing state health monitoring and life prediction based on PCA and Internet of things with multi-sensor[J]. Measurement, 2020, 157: 107657.

[10] Pouamoun A N. Visualization and clustering of text retrieval[J]. American Scientific Research Journal for Engineering, Technology, and Sciences, 2021, 76(1): 113-123.

[11] Chen N, Fu W, Zhao J, et al. BGVD: an integrated database for bovine sequencing variations and selective signatures[J]. Genomics, Proteomics & Bioinformatics, 2020, 18(2): 186-193.

[12] Taverna F, Goveia J, Karakach T K, et al. BIOMEX: an interactive workflow for (single cell) omics data interpretation and visualization [J]. Nucleic acids research, 2020, 48(W1): W385-W394.

[13] Zhao H, Wang S C. A Coding Basis and Three-in-One Integrated Data Visualization Method 'Ana'for the Rapid Analysis of Multi-dimensional Omics Dataset[J]. Life, 2022, 12(11): 1864.

[14] Ranjan C, Ebrahimi S, Paynabar K. Sequence graph transform (SGT): a feature embedding function for sequence data mining[J]. Data Mining and Knowledge Discovery, 2022, 36(2): 668-708.

[15] Ma Y, Wang P, Li B, et al. Research on Energy Consumption Generation Method of Fuel Cell Vehicles: Based on Naturalistic Driving Data Mining[J]. Machines, 2022, 10(11): 1047.

[16] Suatap C, Patanukhom K. Development of Convolutional Neural Networks for Analyzing Game Icon and Screenshot Images[J]. International Journal of Pattern Recognition and Artificial Intelligence, 2022, 36(14): 2254023.

[17] Bovkir R, Aydinoglu A C. Big Urban Data Visualization Approaches Within the Smart City: Gis-Based Open-Source Dashboard Example[J]. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2021, 46: 125-130.

[18] Parhusip H A, Trihandaru S, Heriadi A H, et al. Data Exploration Using Tableau and Principal Component Analysis[J]. JOIV: International Journal on Informatics Visualization, 2022, 6(4): 911-920.

[19] Zhang H, Tsai S B. An Empirical Study on Big Data Model and Visualization of Internet+ Teaching[J]. Mathematical Problems in Engineering, 2021, 2021: 1-10.

[20] Kerner H R, Wagstaff K L, Bue B D, et al. Comparison of novelty detection methods for multispectral images in rover-based planetary exploration missions[J]. Data Mining and Knowledge Discovery, 2020, 34: 1642-1675.
[21] Li R, Qu H, Wang S, et al. CancerMIRNome: an interactive analysis and visualization database for miRNome profiles of human cancer[J]. Nucleic acids research, 2022, 50(D1): D1139-D1146.
[22] Yu T, Nantasenamat C, Kachenton S, et al. Cheminformatic Analysis and Machine Learning Modeling to Investigate Androgen Receptor Antagonists to Combat Prostate Cancer[J]. ACS omega, 2023, 8(7): 6729-6742.
[23] Zhao L, Zhao X, Zhou H, et al. Prediction model for daily reference crop evapotranspiration based on hybrid algorithm and principal components analysis in Southwest China[J]. Computers and Electronics in Agriculture, 2021, 190: 106424.
[24] Piippo S, Sadeghi M, Koivisto E, et al. Semi-automated geological mapping and target generation from geochemical and magnetic data in Halkidiki region, Greece[J]. Ore Geology Reviews, 2022, 142: 104714.
[25] Gajjar S, Kulahci M, Palazoglu A. Least squares sparse principal component analysis and parallel coordinates for real-time process monitoring[J]. Industrial & Engineering Chemistry Research, 2020, 59(35): 15656-15670.
[26] Rostamzadeh N, Abdullah S S, Sedig K. Data-driven activities involving electronic health records: an activity and task analysis framework for interactive visualization tools[J]. Multimodal Technologies and Interaction, 2020, 4(1): 7.
[27] Tuarob S, Wettayakorn P, Phetchai P, et al. DAViS: a unified solution for data collection, analysis, and visualization in real-time stock market prediction[J]. Financial innovation, 2021, 7: 1-32.
[28] Shekhawat S S, Sharma H, Kumar S, et al. bSSA: binary salp swarm algorithm with hybrid data transformation for feature selection[J]. Ieee Access, 2021, 9: 14867-14882.
[29] Fu Y, Gao Z, Liu Y, et al. Actuator and sensor fault classification for wind turbine systems based on fast Fourier transform and uncorrelated multi-linear principal component analysis techniques[J]. Processes, 2020, 8(9): 1066.
[30] Wang Y, Huang H Y, Wang Y Z. Authentication of Dendrobium Officinale from similar species with infrared and ultraviolet-visible spectroscopies with data visualization and mining[J]. Analytical Letters, 2020, 53(11): 1774-1793.
[31] Jacques M A, Dobrzyński M, Gagliardi P A, et al. CODEX, a neural network approach to explore signalling dynamics landscapes[J]. Molecular systems biology, 2021, 17(4): e10026.
[32] Shibayama S, Marcou G, Horvath D, et al. Application of the mol2vec Technology to Large-size Data Visualization and Analysis[J]. Molecular Informatics, 2020, 39(6): 1900170.
[33] Sathyaprakash P, Alagarsundaram P, Devarajan M. V, Alkhayyat A, et al. Medical Practitioner-Centric Heterogeneous Network Powered Efficient E-Healthcare Risk Prediction On Health Big Data[J]. International Journal of Cooperative Information Systems, 2024.
[34] Kumar P. M, Kamruzzaman M. M, Alfurhood B. S, Hossain B, et al. Balanced Performance Merit On Wind and Solar Energy Contact With Clean Environment Enrichment[J]. IEEE Journal of the Electron Devices Society, 2024.
[35] Yang P, Yang G, Zhang F, et al. Spectral classification and particular spectra identification based on data mining[J]. Archives of Computational Methods in Engineering, 2021, 28: 917-935.
[36] Mehdizadeh A, Cai M, Hu Q, et al. A review of data analytic applications in road traffic safety. Part 1: Descriptive and predictive modelling [J]. Sensors, 2020, 20(4): 1107.
[37] Mohedano-Munoz M A, Soguero-Ruiz C, Mora-Jiménez I, et al. A streaming data visualization framework for supporting decision-making in the Intensive Care Unit[J]. Expert Systems with Applications, 2023, 227: 120252.

[38] Ciecierski L, Stolarek I, Figlerowicz M. Human AGEs: interactive spatiotemporal visualization and database of human archeogenomics[J]. Nucleic Acids Research, 2023: gkad428.
[39] Wang Q, Chen Z, Wang Y, et al. A survey on ML4VIS: Applying machine learning advances to data visualization [J]. IEEE transactions on visualization and computer graphics, 2021, 28(12): 5134-5153.
[40] Jiang L, Sullivan H, Wang B. Principal component analysis (PCA) loading and statistical tests for nuclear magnetic resonance (NMR) metabolomics involving multiple study groups[J]. Analytical Letters, 2022, 55(10): 1648-1662.
[41] Rodríguez A C, D'Aronco S, Schindler K, et al. Mapping oil palm density at country scale: An active learning approach[J]. Remote Sensing of Environment, 2021, 261: 112479.
[42] Bu Y, Gao R, Zhang B, et al. CoGT: Ensemble Machine Learning Method and Its Application on JAK Inhibitor Discovery[J]. ACS omega, 2023, 8(14): 13232-13242.
[43] Wang R, Zhong Y, Dong X, et al. Data Mining and Graph Network Deep Learning for Band Gap Prediction in Crystalline Borate Materials[J]. Inorganic Chemistry, 2023, 62(11): 4716-4726.
[44] Krog L S, Kirkensgaard J J K, Foderà V, et al. Application of Low-Frequency Raman Spectroscopy to Probe Dynamics of Lipid Mesophase Transformations upon Hydration[J]. The Journal of Physical Chemistry B, 2023, 127(14): 3223-3230.
[45] Shen I, Cherng F Y, Igarashi T, et al. EvIcon: Designing High-Usability Icon with Human-in-the-loop Exploration and IconCLIP[J]. arXiv preprint arXiv:2305.17609, 2023.
[46] Jetybayeva A, Borodinov N, Ievlev A V, et al. A review on recent machine learning applications for imaging mass spectrometry studies[J]. Journal of Applied Physics, 2023, 133(2): 020702.
[47] Younesy H, Pober J, Möller T, et al. ModEx: a general-purpose computer model exploration system[J]. Frontiers in Bioinformatics, 2023, 3: 1153800.
[48] Sun D, Ying Y. Cultural Heritage Applications of Laser-Induced Breakdown Spectroscopy[J]. Laser Induced Breakdown Spectroscopy (LIBS) Concepts, Instrumentation, Data Analysis and Applications, 2023, 2: 623-642.
[49] Parsania C, Chen R, Sethiya P, et al. FungiExpresZ: an intuitive package for fungal gene expression data analysis, visualization, and discovery[J]. Briefings in Bioinformatics, 2023, 24(2): bbad051.
[50] Yang H, Zhou L, Cai J, et al. [J]. Monthly Notices of the Royal Astronomical Society, 2023, 518(4): 5904-5928.
[51] Zhou M, Fu Q, Li Y, et al. Discovering spatiotemporal flow patterns: where the origin-destination map meets empirical orthogonal function decomposition[J]. Cartography and Geographic Information Science, 2023, 50(2): 113-129.