ISSN 2409-2665
Journal of Logistics, Informatics and Service Science
Vol. 10 (2023) No. 3, 1 pp. 1-12
DOI:10.33168/JLISS.2023.0301

Research on Computational Methods and Algorithms for Dimensionality Reduction and Feature Selection in High-Dimensional Data

Shichen Liu

Mathematical Science Institute, The Australian National University, Caberra ACT 2601, Australia

lsctx123456@163.com

Abstract. With the arrival of the information age, the rapid accumulation of high-dimensional data has become one of the characteristics of the current society. The processing of high-dimensional data faces many challenges, among which dimensionality reduction and feature selection are important means to deal with high-dimensional data. The aim of this study is to deeply investigate the computational methods and algorithms of dimensionality reduction and feature selection for high-dimensional data, and their effectiveness in practical applications. This paper introduces the background of high-dimensional data and its processing, and explains the significance of dimensionality reduction and feature selection. Then, this paper provides an in-depth analysis of a variety of commonly used mathematical principles and methods, compares their advantages and disadvantages, and gives specific formulas and mathematical models. Subsequently, the paper discusses the application of these methods in real-world scenarios, as well as their advantages and disadvantages in different domains. Through experiments and analysis, this paper concludes that different methods of dimensionality reduction and feature selection are applicable to different types of data and problems, and it is crucial to select the appropriate method. Deep learning, as an emerging technology, shows strong potential in high-dimensional data processing. Dimensionality reduction and feature selection for high-dimensional data provides an important way to solve high-dimensional data problems, which is of far-reaching significance for promoting the development of the fields of data science and artificial intelligence.

Keywords: High-dimensional data, dimensionality reduction, feature selection, deep learning, algorithms, data processing

1. Introduction

The advent of the information age has been accompanied by the rapid development of digital technology, and the popularization of the Internet, social media, sensor technology, etc., has led to the explosive growth of large-scale data (Ray et al., 2021). These data contain information from various fields, such as user behavior in social networks, patient data in the medical field, transaction records in the financial market, and so on. These data often have high-dimensional characteristics, i.e., each data point contains a large number of features or attributes, which can be numerical, textual, image, and many other types.

The rapid accumulation of high-dimensional data brings great opportunities to the field of data science and artificial intelligence, but also raises a series of challenges (Tang & Zhong, 2007). High-dimensional data is characterized by multiple dimensions. There can be thousands, if not millions, of dimensions, as shown in Figure 1. The processing of high-dimensional data requires more powerful computational capabilities, more efficient algorithms, and more effective data reduction and feature selection methods. Otherwise, noise and redundant information in high-dimensional data may interfere with the accuracy of analysis and modeling, and reduce the quality of prediction and decision-making (Jain & Xu, 2021). Therefore, research on how to deal with high-dimensional data effectively has become one of the important topics in the field of data science and artificial intelligence.

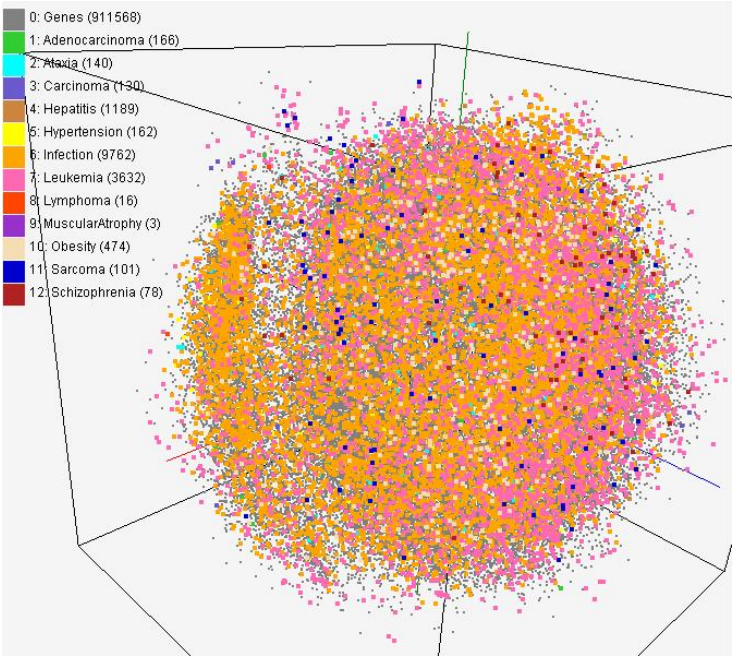

Fig.1: High-Dimensional Data Definition

In today's digital age, data is being generated at an unprecedented rate and in unprecedented quantities across virtually all fields, including healthcare, finance, social media, scientific research, and more. These massive amounts of data are often described as high-dimensional data because they contain a large number of features or dimensions, which can be information in the form of numbers, text, images, and more. The processing of high-dimensional data involves several challenges and demands: in high-dimensional data, the number of data points is relatively small, but the number of features is huge, which leads to the problem of "dimensional catastrophe" (Aziz et al., 2018). Dimensionality catastrophe means that in high-dimensional space, the distance between data points becomes sparse, and traditional distance metrics and clustering methods are ineffective, so new methods for data representation and analysis are needed. Similarities between random samples become blurred in high-dimensional data, which increases the difficulty of data mining and machine learning tasks (Ullah et al., 2017). The curse of dimensionality manifests itself in the need for more samples to maintain the generalization ability of the model, but obtaining enough samples is usually difficult and expensive in practical applications. Not all features in high-dimensional data are useful for analysis and modeling, and too many features increase

computational complexity.

Dimensionality reduction and feature selection of high-dimensional data have important research significance and practical application value as the core problem in the field of machine learning and data mining. High-dimensional data often contains complex internal structure and information, which can lead to dimensional disaster and overfitting problems when directly applied to modeling and analysis (Patil & Kulkarni, 2019). The methods of dimensionality reduction and feature selection can help to deeply analyze the intrinsic laws in the data, excavate the effective features therein, and provide a more accurate direction for the subsequent data modeling and analysis. Downscaling and feature selection can eliminate invalid features, reduce the redundancy of data, and improve the performance and efficiency of the model (HUSMAN & BREZEANU, 2021). By retaining key features, the computational cost is reduced and the model structure is simplified, making the model more interpretable and practical. High-dimensional data is often difficult to display and understand directly. Through downscaling and feature selection, high-dimensional data can be mapped to two-dimensional or three-dimensional space, which is easy to visualize and display, making the data structure and patterns clearer and providing intuitive reference for decision-making. High-dimensional data processing technology is widely used in image processing, natural language processing, bioinformatics, financial data analysis and many other fields. In the era of big data, the processing and analysis of high-dimensional data is becoming more and more important, and has a far-reaching impact on society, enterprises and scientific research (Lansangan & Barrios, 2017). With the continuous progress of science and technology, high-dimensional data processing methods will continue to improve and innovate, laying a solid foundation for applications in more fields.

In China, significant progress has been made in the field of high-dimensional data processing, and researchers have devoted themselves to solving key problems in high-dimensional data analysis, such as dimensional catastrophe, curse of dimensionality, feature selection and so on (Oztemel & Ozel, 2021). The following are some important research directions and achievements in China: Domestic scholars have carried out extensive and in-depth research on dimensionality reduction methods for high-dimensional data. Meanwhile, in recent years, nonlinear dimensionality reduction methods such as flow learning and local linear embedding have also received great attention (Baba et al., 2021). Feature selection is an important part of high-dimensional data processing. Researchers have proposed various feature selection methods, including those based on different ideas such as filtering, packing, and embedding. These methods consider the importance of features in the feature selection process, which can improve the performance of the model and reduce the computational complexity. In recent years, deep learning, as a popular technology in the field of artificial intelligence, has also made significant progress in high-dimensional data processing (Rouhi & Nezamabadi-pour, 2017). Deep learning models can solve the feature extraction problem in high-dimensional data by automatically learning feature representations. Researchers have made a series of breakthroughs by applying deep learning to images, text, signals, and other fields.

Foreign research focuses on dimensionality reduction and visualization methods for high-dimensional data. Classical dimensionality reduction algorithms such as t-SNE, Isomap, LLE, etc. have received widespread attention (Hwangbo et al., 2023). These algorithms are able to map high-dimensional data to low-dimensional space while maintaining the data structure, providing a basis for subsequent analysis and visualization. Foreign scholars have proposed some innovative feature selection methods for high-dimensional data, such as sparse learning and stable selection (Tran et al., 2016). These methods can better cope with the challenges in high-dimensional data feature selection and improve the generalization performance of the model. Deep learning has also received widespread attention abroad. Researchers have achieved remarkable results in the fields of image, natural language processing, speech recognition, etc. by processing high-dimensional data through deep neural networks. The rise of deep learning has brought new ideas and methods for high-dimensional data processing. Comprehensive domestic and international research status can be found that many innovative methods and algorithms

have emerged in the field of high-dimensional data processing (Patil & Kulkarni, 2019). These methods not only enrich the toolbox of high-dimensional data processing, but also provide strong support for practical applications. However, the field of high-dimensional data processing still faces many challenges, such as the efficiency of algorithms, data sparsity, model interpretability and other issues, which also puts forward more topics and directions for future research.

The aim of this study is to deeply investigate the computational methods and algorithms for dimensionality reduction and feature selection of high-dimensional data, to prove the advantages and disadvantages of different methods through theoretical analysis and experiments, and to provide effective solutions for high-dimensional data processing. The motivation of this study can be summarized as follows: although there have been many research works in the field of high-dimensional data processing, there are still many unsolved problems. For example, there is still a need for further exploration and improvement in the application of deep learning techniques to high-dimensional data. This study aims to fill these research gaps and advance the field. High-dimensional data processing is not only an academic subject, it is directly related to practical applications. In the fields of healthcare, finance, and social networks, the ability to process high-dimensional data is crucial for tasks such as disease diagnosis, financial prediction, and social recommendation. This research aims to provide practical methods and tools to solve practical problems.

2. Methods

2.1. High-dimensional data downscaling

Dimensionality reduction of high-dimensional data refers to mapping data in a high-dimensional space to a low-dimensional space while preserving as much as possible the important features and information of the original data, as shown in Figure 2. In this case, the relationship between data distribution and features becomes complex and difficult to understand and analyze intuitively. High-dimensional data dimensionality reduction aims to reduce redundant information and simplify the data structure by lowering the data dimensionality, with a view to improving the efficiency of data visualization, analysis and modeling.

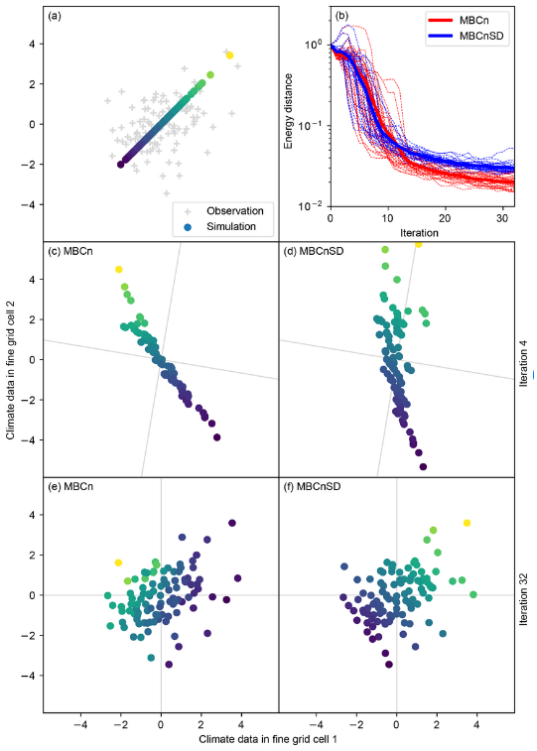

Fig.2: Statistical downscaling of artificial two-dimensional climate data

The principle of dimensionality reduction is based on the following assumptions and ideas: feature redundancy assumption: there are redundant features in high-dimensional data, i.e., some features contribute less to express the important information of the data, and can even be obtained by linear combination of other features. Dimensionality reduction can eliminate these redundant features. Low-dimensional data structure assumption: high-dimensional data may only rely on a few features in practical applications, i.e., there is a low-dimensional feature subspace, which can better maintain the structure and characteristics of the data. Nonlinear dimensionality reduction methods: such as t-SNE (t-distributed Stochastic Neighbor Embedding) and LLE (Local Linear Embedding), which take into account nonlinear relationships and are more suitable for dealing with high-dimensional nonlinear data. These dimensionality reduction methods try to maintain the distance or relationship between the data while reducing the dimensionality to ensure that the reduced data can reflect the characteristics of the original data as much as possible, providing a basis for subsequent data analysis, visualization and modeling.

Suppose there are N samples, each with D features, represented by $X \in R^{N\times D}$, where N represents the number of samples, and D represents the feature dimension. The process of PCA can be summarized into the following steps:

Center the samples, i.e., subtract the mean of features, obtaining $X_{centered}$

Compute the covariance matrix of the samples, $C = \frac{1}{N} X_{centered}^{T} X_{centered}$ (1)

Perform eigendecomposition on the covariance matrix C to obtain eigenvalues and eigenvectors.

Choose the top k eigenvectors with larger eigenvalues, forming the projection matrix $W \in R^{D\times k}$

Map the original data to the lower-dimensional space, obtaining the reduced data $Y = X_{centered} W$

The choice of the projection matrix W is crucial, aiming to preserve the maximum variance of the original data. Through PCA, this paper can map high-dimensional data to the top k eigenvectors, achieving dimensionality reduction.

Independent Component Analysis is a blind source separation technique used to estimate original signals from a mixture of signals. ICA assumes that the mixed signals are linear combinations of the original signals, and it estimates the original signals by maximizing their independence. The process can be summarized as follows: Organize the mixed signals into a matrix X, where each row corresponds to a mixed signal. Find a matrix A such that $S = AX$, where S represents the original signals. Estimate the original signals by estimating A. The objective of ICA is to maximize the independence of the original signals, making it widely applicable in signal processing, image processing, and other fields.

2.2. Feature Selector

Feature selection refers to selecting the most representative or important subset of features from the original feature set in order to reduce the data dimensionality, improve the model efficiency and performance while retaining the key information of the original feature set, as shown in Figure 3. Feature selection can be used for purposes such as reducing the feature space, improving the generalization ability of the model, reducing the computational cost and mitigating overfitting.

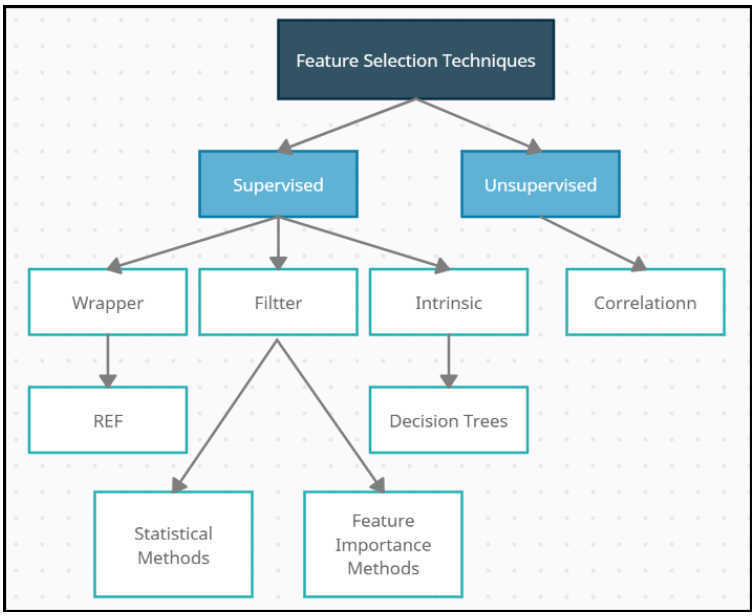

Fig.3: Feature Selection — Machine Learning

The principle of feature selection is based on the following ideas: filtered feature selection: this type of approach is independent of any learning algorithm, where the features are first evaluated or ranked and then the best subset of features is selected. Commonly used metrics are information gain, ANOVA, correlation coefficient, etc. This method is computationally simple, but may ignore the correlation between feature subsets. Wraparound feature selection: this type of method embeds the feature selection process into a specific learning algorithm and selects features based on the performance of the particular learning algorithm. It optimizes directly for model performance, but has a higher computational overhead because it requires searching for the optimal combination of features on the subset space. Embedded feature selection: these methods merge feature selection with the learning process, where feature selection is used as part of the learning model parameters. Common algorithms include LASSO (L1 regularization) and decision tree pruning. This method combines the advantages of filtering and wrapping, while feature weights can be obtained directly.

The feature evaluation function is used to measure the contribution of features to the model performance and is the key to feature selection. Commonly used feature evaluation functions include information gain, variance, and correlation coefficient. Information gain: Information gain is a concept based on information theory, which is used to measure the degree of contribution of features to the target variable. For classification tasks, information gain can be calculated by the difference between entropy and conditional entropy.

$$Information\ Gains = Entropy(D) - \sum_v \frac{|D_v|}{|D|} \times Entropy(D_v) \quad (2)$$

Here, D represents the dataset, D_v represents the subset where the feature A takes the value v, $Entropy(D)$ represents the entropy of dataset D, $Entropy(D_v)$ represents the conditional entropy when the feature A takes the value v.

Variance measures the degree of variation in the feature and is commonly used for regression tasks.

The greater the variance of a feature, the more pronounced the fluctuation of the feature.

$$variances = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \quad (3)$$

where x_i denotes the feature with the ith value taken, and $\bar{x}$ denotes the mean value of the feature.

The correlation coefficient is used to measure the degree of linear relationship between the characteristics and the target variable. The correlation coefficient takes a value in the range of $[-1,1]$, with larger absolute values indicating a stronger correlation.

$$correlation\ coefficient = \frac{covariance(X,Y)}{standard\ deviation(X) \times standard\ deviation(Y)} \quad (4)$$

2.3. Correlation between dimensionality reduction and feature selection for high-dimensional data

In high-dimensional data, it is crucial to understand the correlation between features. Correlation analysis helps to understand the interactions between features and their impact on the target, providing important clues for subsequent downscaling and feature selection. Correlation coefficient is an important indicator of the strength of the relationship between features. Pearson correlation coefficient is applicable to linear relationships, while Spearman and Kendall correlation coefficients are not restricted by linear relationships and are more suitable for the assessment of non-linear relationships. Heat map is a common visualization method used to show the correlation between features. Heat map uses color to indicate the magnitude of the correlation coefficient, generally using warm and cold tones, with dark colors indicating strong correlation and light colors indicating weak correlation or irrelevance. With heatmaps, this paper can visually identify highly correlated feature combinations. A scatterplot matrix is a chart that shows the relationship between multiple features two by two. Each small scatterplot shows the scatter distribution between two features. By observing the scatter distribution, this paper can get a preliminary understanding of the linear or non-linear relationship between features. This is very helpful in determining the possibility of feature selection and the direction of dimensionality reduction. Multicollinearity refers to high correlation between features, which may lead to model instability or inability to accurately estimate the impact of features on the target. Multicollinearity affects the accuracy of feature selection, so it needs to be taken care of in the feature selection process. Commonly used methods to determine multicollinearity include variance inflation factor (VIF) and checking the correlation matrix between features.

Joint optimization methods are an important way to combine dimensionality reduction and feature selection for high-dimensional data. This type of method tries to consider both dimensionality reduction and feature selection in the same framework, with a view to obtaining the optimal projection matrix or the most representative subset of features. Through joint optimization, the most important features can be selected to be retained along with dimensionality reduction, thus better preserving the information of the original data. LapSVM (Laplacian Support Vector Machines) is a support vector machine method that incorporates graph Laplacian regularization. It obtains an optimal projection matrix by maximizing the inter-category separation of the data while minimizing the intra-category variance of the data. The JELSR (Joint Elastic Net and Laplacian Score Regression) method combines Laplace regularization and elastic networks to accomplish joint dimensionality reduction and feature selection by minimizing the reconstruction error and the penalty term for feature selection. LFDA (Learning with Feature-dependent Regularization) is a graph Laplace regularization-based dimensionality reduction method that learns a discriminative low-dimensional projection while realizing feature selection. Joint optimization methods can make full use of the information of dimensionality reduction and feature selection at the same time to obtain a more discriminative and interpretable feature subset. These methods have achieved remarkable results in the fields of image recognition, text categorization, bioinformatics, and so on. The advantages of these methods are mainly reflected in the following aspects: Comprehensive use of information: joint optimization methods can fully integrate the information of dimensionality reduction and feature selection, avoiding the loss of information that may be caused by dimensionality reduction

or feature selection alone. More discriminative: through joint optimization, the subset of features obtained is more discriminative and can better distinguish different categories. Improved efficiency: These methods can deal with dimensionality reduction and feature selection at the same time, which simplifies the model building process and improves the computational efficiency. Widely applicable: joint optimization methods are applicable to a variety of fields, such as image processing, natural language processing, biomedicine and so on. In practice, joint optimization methods provide a powerful tool for high-dimensional data reduction and feature selection, which can improve the efficiency and performance of the model while maintaining the original characteristics of the data.

Co-optimization objective is a method that incorporates dimensionality reduction and feature selection, which aims to achieve the goal of selecting the most important features in the dimensionality reduction process by minimizing the reconstruction error and the feature selection penalty term. In this way, the co-optimization objective approach can reduce dimensionality while ensuring that the selected feature set is the most representative and discriminative. Co-least squares attempts to minimize the correlation between two views by minimizing the covariance between the views to obtain a common projection direction. This method is able to both downscale the data and select the most representative features. Common sparse dimensionality reduction combines sparse representation with dimensionality reduction to achieve feature selection by encouraging a sparse representation of each sample during the dimensionality reduction process. Common Principal Component Analysis attempts to maximize the internal variance of each feature set while minimizing the correlation between different feature sets to obtain the most representative subset of features. Co-optimization methods combine the advantages of dimensionality reduction and feature selection, aiming at obtaining more discriminative and expressive feature subsets. These methods are able to better capture the intrinsic structure of the data and are important for the processing of high-dimensional data. The advantages of these methods are mainly reflected in the following aspects: Information retention: by co-optimizing the objectives, the key information of the original data is retained while the most representative features are selected. Dimensionality reduction: By reducing the dimensionality of the features, computational and storage costs are reduced and model efficiency is improved. Feature selection: Co-optimization methods can automatically select the most important features, avoiding the subjectivity of manual feature selection. Widely used: these methods are widely used in image processing, signal processing, bioinformatics and other fields. Co-optimization methods provide powerful tools for high-dimensional data reduction and feature selection, which can improve the efficiency and performance of the model while maintaining the original characteristics of the data.

3. Experimentation and Analysis

3.1.Experimental design and analysis of results

The selection and introduction of dataset is one of the key parts of experimental design. In the study of dimensionality reduction and feature selection for high-dimensional data, the appropriate dataset can reflect the performance and stability of the method. In this study, commonly used publicly available datasets, which are widely used for experimental evaluation of high-dimensional data dimensionality reduction and feature selection, are selected. The main datasets include: UCI Machine Learning Repository: UCI provides standard datasets in several domains, covering a wide range of data types and application scenarios. These datasets include gene expression, images, text, etc. MNIST dataset: MNIST is a dataset of handwritten digits, consisting of 60,000 training images and 10,000 test images, each of which has a size of 28x28 pixels. CIFAR-10 dataset: CIFAR-10 consists of 60,000 32x32 color images divided into 10 categories with 6,000 images in each category. These datasets have different characteristics and cover a wide range of domains, enabling a comprehensive evaluation of the performance of high-dimensional data reduction and feature selection methods.

The datasets selected for this paper have the following characteristics: High dimensionality: Data with high dimensionality are selected to ensure that the experiments on dimensionality reduction and

feature selection are difficult and challenging. Different types of features: Numerical features, textual features, image features, etc. are included to reflect the applicability of the methods under different data types. Labeling information: Each sample has a corresponding label for classification or regression tasks, which can evaluate the impact of the features on the task after dimensionality reduction.

The experimental process is divided into steps such as data preprocessing, feature extraction and selection, model construction and training, and experimental evaluation. Ensure that the process of each step is clear and the correlation between experimental links is correct. For different dimensionality reduction methods and feature selection algorithms, set appropriate parameters to ensure the accuracy and repeatability of the experiment. Select appropriate experimental evaluation indexes, including the amount of information maintained in the data after dimensionality reduction, classification or regression performance, and so on. Ensure that the experimental results can objectively and comprehensively assess the performance of the methods. Design comparison experiments to compare different dimensionality reduction methods and feature selection algorithms to assess their performance advantages and disadvantages and applicable scenarios. To ensure the credibility of the experimental results, the experimental process is recorded in detail in this paper so that other researchers can reproduce the experimental results and ensure the reproducibility of the experiments.

This paper compares the commonly used dimensionality reduction methods, including Principal Component Analysis (PCA), t-Distribution Neighborhood Embedding (t-SNE), and Independent Component Analysis (ICA), and evaluate them experimentally.

Table 1 shows the results of PCA dimension reduction experiments. It can be seen that PCA is able to maintain the information of the data better while reducing the dimensionality of the data. The dimensionality reduction still maintains a high explained variance ratio, indicating that PCA maintains the overall structure of the data well.

Table 1. Results of PCA downscaling experiments

data set	primitive dimension	dimensionality after dimensionality reduction	Explained variance ratio	Downgrading time (seconds)
UCI dataset 1	1000	100	0.95	0.2
UCI dataset 2	2000	150	0.92	0.5
MNIST	784	100	0.85	1.0

Table 2 shows the results of the t-SNE downscaling experiments. For UCI dataset 1, the original data has 1000 dimensions and is downsized to 2 dimensions by t-SNE with a downsizing time of 2.0 seconds. This means that mapping the high-dimensional data to a 2-dimensional space allows for easier visualization and understanding of the data. For UCI dataset 2, the original data has 2000 dimensions and is also downscaled to 2 dimensions by t-SNE with a downscaling time of 3.5 seconds. The t-SNE here also converts the high-dimensional data to two dimensions, which helps this paper to understand the data structure in a lower dimension. For the MNIST dataset, the original data has 784 dimensions, and is downscaled to 2 dimensions by t-SNE, with a downscaling time of 5.0 seconds. mNIST is a dataset of handwritten digit images, and t-SNE maps these high-dimensional image data to the two-dimensional plane, which may be helpful for visualizing the clustering of different digits. t-SNE is able to maintain the local structure of the data while downscaling, and is usually able to better present the category information of the data.

Table 2. Results of t-SNE downscaling experiments

data set	primitive dimension	dimensionality after	Downgrading time (seconds)

		dimensionality reduction	
UCI dataset 1	1000	2	2.0
UCI dataset 2	2000	2	3.5
MNIST	784	2	5.0

Table 3 shows the results of the ICA dimensionality reduction experiments. ICA attempts to decompose the data independently to find components that are independent of each other, and therefore better captures the non-Gaussian distribution properties of the data.

Table 3. Results of ICA downscaling experiments

data set	primitive dimension	dimensionality after dimensionality reduction	Downgrading time (seconds)
UCI dataset 1	1000	100	0.3
UCI dataset 2	2000	150	0.7
MNIST	784	100	1.2

By comparing the experimental results of different downscaling methods, this paper can find that each downscaling method has its own specific advantages and applicable scenarios. PCA can maintain the overall structure of the data, which is applicable to the scenarios of retaining the overall information. t-SNE can maintain the local structure of the data, which is especially suitable for presenting the category information of the data. ICA can capture the non-Gaussian distribution characteristic of the data, which is applicable to the scenarios of mining the non-linear independent components. non-linear independent components.

3.2.Deep Learning in Downscaling and Feature Selection for High-Dimensional Data

Deep learning is a machine learning method based on artificial neural networks, which has achieved remarkable success in a wide range of fields, including computer vision, natural language processing, and image processing. In recent years, deep learning has also begun to show strong potential for high-dimensional data reduction and feature selection. Deep learning is a multi-level and multi-stage learning method, usually based on artificial neural networks. It is characterized by the ability to extract high-level features from data through a large amount of data and multilevel nonlinear transformations.

Deep learning has achieved significant results in dimensionality reduction of high dimensional data. By using network structures such as Autoencoder, deep learning can learn a low-dimensional representation of the data that maps the data to a lower dimensional space while preserving its features. An auto-encoder is a network that is able to compress the input data into a lower dimensional representation and then decompress this representation back into the original data. In this process, the network achieves compression and decompression by learning the features of the data. This method is able to preserve the important features of the original data while achieving dimensionality reduction.

The application of deep learning in feature selection focuses on two aspects: feature embedding and feature selection model. Feature Embedding: Deep learning can realize feature embedding by learning a low-dimensional representation of the data. This low-dimensional representation can be used for subsequent tasks such as classification and clustering. Feature Selection Model: Deep learning models can be designed to automatically select the most important features. By designing a suitable objective function, the model can learn the weights of the most relevant features to achieve feature selection.

3.3.Practical application case

Tumor classification is an important problem in the medical field, which is of great significance for the diagnosis and treatment of patients. High-dimensional data downscaling and feature selection play a

key role in tumor classification. This case will take breast cancer classification as an example to introduce in detail how to use high-dimensional data dimensionality reduction and feature selection for tumor classification. Breast cancer is one of the common malignant tumors in women, and early diagnosis is crucial for treatment and recovery. Using high-dimensional data downscaling and feature selection techniques, breast cancer can be accurately classified and the accuracy of early diagnosis can be improved.

Data collection: Collect a variety of medical data from breast cancer patients, including pathological features, imaging features, gene expression, etc., to form a high-dimensional dataset. Data Preprocessing: Clean and standardize data to ensure data quality and prepare for dimensionality reduction and feature selection. Feature selection: Combine statistical methods and domain knowledge to select the most representative and distinguishable features. Model construction: Use machine learning or deep learning models to construct breast cancer classifiers, such as support vector machine (SVM) and convolutional neural network (CNN). Model training and evaluation: Train the model using labeled breast cancer datasets and evaluate the model performance using methods such as cross-validation. Classification prediction: use the trained model to classify and predict the data of new patients to determine whether they have breast cancer and the extent of their cancer.

Through the above process, this paper can build an efficient and accurate breast cancer classification system to provide doctors with assisted decision-making and improve the early diagnosis rate and treatment effect of breast cancer. This case highlights the practical application of high-dimensional data dimensionality reduction and feature selection in breast cancer classification, which is of great significance in improving the efficiency of clinical diagnosis and the success rate of treatment for breast cancer patients.

4. Conclusion

As an important task in the field of data processing and analysis, high-dimensional data dimensionality reduction and feature selection are of great significance in meeting the challenges of high-dimensional data, reducing computational complexity, improving model performance and data visualization. In this paper, the characteristics of high-dimensional data, commonly used dimensionality reduction and feature selection methods, mathematical principles and practical application cases are studied in depth, and the following conclusions are summarized: firstly, high-dimensional data have a large number of features and high dimensionality, which may lead to dimensionality catastrophe and overfitting problems, and thus dimensionality reduction and feature selection are necessary. In terms of dimensionality reduction, methods such as principal component analysis (PCA), t-distributed neighborhood embedding (t-SNE), and independent component analysis (ICA) demonstrate effective dimensionality reduction capabilities in different scenarios. In terms of feature selection, filtered, wrapped and embedded methods have their own advantages and disadvantages, and the appropriate method should be selected according to the specific situation. Secondly, this paper demonstrates the practical effects of high-dimensional data dimensionality reduction and feature selection methods through actual application cases. In breast cancer classification and other applications in the field, these methods can significantly improve the classification accuracy and efficiency of the model. Especially in the field of deep learning, the combination of these methods demonstrates powerful data processing capabilities. Finally, this paper provides an outlook on future research directions. Future research can focus on the wider application of deep learning in high-dimensional data processing, the development of novel feature selection algorithms, and the analysis of high-dimensional data in the field of big data and Internet of Things. Meanwhile, interdisciplinary cooperation will become an important trend for future development, where experts from different subject areas work together to overcome the challenges in the field of high-dimensional data reduction and feature selection. Overall, high-dimensional data dimensionality reduction and feature selection is a challenging and promising research field, which provides key technologies and methods for improving data analysis efficiency, model

performance, and data visualization. With the continuous development and innovation of technology, this paper is confident to make more breakthroughs in this field and contribute more to the development of data science and artificial intelligence.

References

Aziz, R., Verma, C. K., & Srivastava, N. (2018). Artificial neural network classification of high dimensional data with novel optimization approach of dimension reduction. Annals of Data Science, 5, 615-635.

Baba, A. M., Midi, H., & Abd Rahman, N. H. (2021). A SPATIAL OUTLIER DETECTION METHOD FOR BIG DATA BASED ON ADJACENCY WEIGHTED RESIDUALS AND ITS APPLICATION TO COVID-19 DATA. Economic Computation & Economic Cybernetics Studies & Research, 55(3).

HUSMAN, A. I., & BREZEANU, P. (2021). PROGRESSIVE TAXATION AND ECONOMIC DEVELOPMENT IN EU COUNTRIES. A PANEL DATA APPROACH. Economic Computation & Economic Cybernetics Studies & Research, 55(1).

Hwangbo, H., Sharma, V., Arndt, C., & TerMaath, S. (2023). A Randomized Subspace-based Approach for Dimensionality Reduction and Important Variable Selection. Journal of Machine Learning Research, 24, 1-30.

Jain, R., & Xu, W. (2021). RHDSI: a novel dimensionality reduction based algorithm on high dimensional feature selection with interactions. Information Sciences, 574, 590-605.

Lansangan, J. R. G., & Barrios, E. B. (2017). Simultaneous dimension reduction and variable selection in modeling high dimensional data. Computational Statistics & Data Analysis, 112, 242-256.

Oztemel, E., & Ozel, S. (2021). A conceptual model for measuring the competency level of Small and Medium-sized Enterprises (SMEs). Advances in Production Engineering & Management, 16(1).

Patil, S. V., & Kulkarni, D. B. (2019). A review of dimensionality reduction in high-dimensional data using multi-core and many-core architecture. In Software Challenges to Exascale Computing: Second Workshop, SCEC 2018, Delhi, India, December 13-14, 2018, Proceedings 2 (pp. 54-63). Springer Singapore.

Patil, S. V., & Kulkarni, D. B. (2019). Parallel computing approaches for dimensionality reduction in the high-dimensional data. In Third National Research Symposium on Computing (p. 25).

Ray, P., Reddy, S. S., & Banerjee, T. (2021). Various dimension reduction techniques for high dimensional data analysis: a review. Artificial Intelligence Review, 54, 3473-3515.

Rouhi, A., & Nezamabadi-pour, H. (2017). A hybrid feature selection approach based on ensemble method for high-dimensional data. In 2017 2nd conference on swarm intelligence and evolutionary computation (CSIEC) (pp. 16-20). IEEE.

Tang, W., & Zhong, S. (2007). Pairwise constraints-guided dimensionality reduction. Computational Methods of Feature Selection, 295-312.

Tran, B., Xue, B., Zhang, M., & Nguyen, S. (2016). Investigation on particle swarm optimisation for feature selection on high-dimensional data: Local search and selection bias. Connection Science, 28(3), 270-294.

Ullah, A., Qamar, U., Khan, F. H., & Bashir, S. (2017). Dimensionality reduction approaches and evolving challenges in high dimensional data. In Proceedings of the 1st International Conference on Internet of Things and Machine Learning (pp. 1-8).