International Journal for Housing Science and Its Applications
Publish August 4, 2025. Volume 46, Issue 3 Pages 3372-3383

https://doi.org/10.70517/ijhsa463285

Research on Multidimensional Data Extraction and Mapping Methods for Computer Terminals Based on Principal Component Analysis Approach

Jun Han[1,*], Ke Liu[1], Yutong Liu[1], Wenqian Zhang[1] and Shaofei Wang[1]
[1] State Grid Qinghai Electric Power Company Electric Power Science Research Institute, Xining, Qinghai, 810008, China
Corresponding authors: (e-mail: gjdwgjdw@163.com).

Abstract With the advancement of Industry 4.0, computer terminals are increasingly integrated with the public Internet. While this integration improves operational efficiency and flexibility, it also brings new data management problems. Before starting the research, the theoretical knowledge of this research is defined and expressed. After that, the computer terminal data is obtained, and it is found that this data has problems such as redundancy and multidimensionality. In order to improve the efficiency of user data management, data extraction method based on PCA-ReliefF, data mapping method based on kernel principal components are designed and analyzed by numerical simulation using Matlab 7.1.The KPCA algorithm has a longer computing time than the PCA algorithm, which indicates that the kernel function is introduced on top of the original one in order to realize the nonlinear mapping from low-dimension to high-dimension, which results in the growth of the computing time. Although the computing time grows, the accuracy of KPCA is much higher than PCA, i.e., the introduction of kernel function in traditional PCA can improve the accuracy of computerized multidimensional data mapping, which facilitates the users to better manage the data of computer terminals.

Index Terms PCA, ReliefF, kernel function, computer terminal

I. Introduction

In recent years, big data has almost become the pronoun of the times, around the emergence of various scenes in people's daily life. The big data recommendation when shopping, the interest push of social software, the user profile of medical organizations, etc., these business scenarios based on big data generate very large data [1]-[3]. As the volume of data explodes and the dimensionality of data grows, the complexity of data also surges, and the dimensionality of multidimensional data becomes higher and higher [4]. Since people live in three-dimensional space, there are limitations in perceiving high-dimensional space, these multidimensional data are difficult to be analyzed directly in the form of observation, and the inter-data correlations hidden under the multidimensional data cannot be directly observed [5], [6]. How to make these multidimensional data can be quickly understood and analyzed has been a hot research topic in the field of data analysis.

Adopting visualization can make users more intuitively and efficiently obtain the latent, unknown and valuable information in the data stream, discover the intrinsic patterns in the data stream, and provide strong support for managers to make decisions [7]-[9]. However, due to the problem of “data dimensionality disaster”, direct visualization of multidimensional streaming data can cause a lot of visual confusion, and the internal structure of streaming data cannot be observed effectively [10]. Therefore, before visualization, it becomes necessary to pre-process the stream data using data mining techniques. Principal component analysis is a typical linear dimensionality reduction algorithm, which is used to extract the dimensional features in the multidimensional data set, and at the same time, the visualization of radial coordinates is used to reflect the distribution characteristics of the multidimensional data after dimensionality reduction, which effectively visualizes and analyzes the multidimensional data [11]-[14]. Combining their own knowledge and computer computing ability, they further construct visualization views and conduct exploratory analysis of multidimensional data, so as to understand the potential information behind the data more directly and effectively [15], [16].

This paper first defines the computer terminal, and within this definition, with the help of related technology to obtain the research data in this paper, it is found that the collected data has the characteristics of multi-dimensionality and redundancy, which easily aggravates the difficulty of the user's computer terminal data management. In order to reduce the difficulty of multidimensional data management of computer terminals, two data processing methods are extended on the basis of principal component analysis algorithm, which are data extraction

method based on PCA-ReliefF, and data mapping method based on kernel function-principal component analysis. In order to verify the practical application performance of the above two methods, numerical simulation analysis is carried out under the Matlab 7.1 software together with the research data.

II. Research on multi-dimensional data extraction and mapping of computer terminals

II. A. Computer terminals and their data acquisition

II. A. 1) Computer terminals

Computer terminal mainly refers to the computer system that can independently carry out data processing and provide network service access, which is mainly composed of computer operating system, hardware, software and other parts, and usually can be divided into desktop microcomputer system and portable microcomputer system [17]. Computer terminal is the infrastructure for information data transmission, storage and application and an important node for connecting the network, which has been transformed into an indispensable support platform for a series of links in people's daily life and production. At the same time, due to the special characteristics of the computer terminal, it will face a series of security risks such as network security and data security. Computer terminal security risks have become a network security threat that cannot be ignored, as far as computer terminal data is concerned, as long as it is in the Internet environment, it will be attacked by all parties from the Internet, and once the computer terminal data is leaked and lost, it will cause incalculable losses.

II. A. 2) Data acquisition

In the data collection phase, the computer terminal first performs deep packet inspection of network traffic to obtain various attributes of the packet, including source IP address, destination IP address, protocol type, packet length, and so on. This attribute information constitutes the original security feature collection, which provides the basis for subsequent security feature extraction and analysis. Principal component analysis technique is a network multidimensional data downscaling technique that can detect and record packet information in network traffic. This technique not only extracts the basic attributes of the packets, such as source and destination IP addresses, protocol types, etc., but also further de-analyzes the contents of the packets and extracts deeper information, such as application layer protocols, user behavior, and so on. This deep-level information is of great value for understanding and identifying network security threats. In computer terminals, this technique is usually accomplished by specialized hardware devices or software tools. These devices or tools need to have high-speed data processing capabilities in order to extract packet information in real time without affecting network performance. In addition, because of the variety of protocols, devices, and applications involved in computer terminals, the technology needs to be highly flexible and adaptable in order to handle various types of packets.

II. B. Computer terminal data extraction based on principal component method

II. B. 1) Principal Component Approach

The theory of Principal Component Analysis (PCA) is to transform variables with correlation into a new set of mutually unrelated variables through linear combination, from which a few most representative variables are selected for later analysis and modeling [18]. The main steps are as follows.

(1) Assuming that there are n samples with p variables each, construct the $n\times p$ -order matrix x of the observed samples:

$$x=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \tag{1}$$

where: x_{np} is the p th variable of the n th sample.

(2) Normalize the $n\times p$ -order matrix:

$$X_{ij}=\frac{x_{ij}-\overline{x}_j}{S_j} \quad i=1,2,\cdots,n \quad j=1,2,\cdots,p \tag{2}$$

where: $\overline{x}_j$ is the sample mean by column and S_j is the sample standard deviation. The original sample matrix is normalized to:

$$X=\begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix}=\left[X_1, X_2, \cdots, X_p\right] \tag{3}$$

where X_{np} is the p th variable of the n th sample after standardization. X_p is the p th variable of the n th sample after standardization.

(3) Find the covariance matrix R .

The covariance matrix of the standardized matrix X is R , when the covariance matrix R is of order $p \times p$:

$$R=\begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1p} \\ r_{21} & r_{22} & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & r_{pp} \end{bmatrix} \tag{4}$$

where: r_{pp} is the correlation coefficient.

(4) Calculate the eigenvalues and eigenvectors of the covariance matrix $\mathbf{R}$

Eigenvalue y_j :

$$y_1 \geq y_2 \geq \cdots \geq y_p \tag{5}$$

Eigenvectors:

$$a_1=\begin{bmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{p1} \end{bmatrix}, a_2=\begin{bmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{p2} \end{bmatrix}, \cdots, a_p=\begin{bmatrix} a_{1p} \\ a_{2p} \\ \vdots \\ a_{pp} \end{bmatrix} \tag{6}$$

(5) Principal Component Analysis

The eigenvalues are sorted according to their size, and their eigenvectors are arranged in the same order to form an eigenmatrix. Since the variance of the analyzed principal components decreases step by step, the amount of information they contain also decreases sequentially. Usually, the top k principal features are selected as principal components in the cumulative contribution calculation. Where the contribution ratio C is the proportion of an eigenvalue to all eigenvalues, which is calculated as:

$$C=\frac{y_j}{\sum_{k=1}^{p} y_k} \tag{7}$$

where: y_j is the j th eigenvalue. y_k is the k th eigenvalue.

The higher the contribution of a single attribute, the higher the features contained in its principal component.

(6) Calculate principal components

Generally take the 1st, 2nd, and $m(m \leq p)$ principal components corresponding to the eigenvalues whose cumulative contribution rate exceeds 90%. The m th principal component is:

$$F_m=a_{1m}X_1+a_{2m}X_2+\cdots+a_{pm}X_p \tag{8}$$

II. B. 2) Characterization data classification

Suppose, the computer terminal multidimensional data is divided into m training samples $x_p=\left(x_{p1}, x_{p2}, \cdots, x_{pm}\right)$, and the fuzzification formula is:

$$x_p = \frac{x_p - MIN(x_p^j)}{MAX(x_p^j) - MIN(x_p^j)} \tag{9}$$

Assuming, the computer terminal multidimensional database is divided into two parts, the i th input vector x_i is divided into the interval of K_i network delay feature fuzzy subsets as $A_{i1}, A_{i2}, \cdots, A_{iK}$, and utilizing Eq. (10) to give the computer terminal multidimensional data is divided into the $K_1 K_2 \cdots K_n$ fuzzy subset space as:

$$W = j_1 \times K_1 + \cdots + j_2 \times K_n \tag{10}$$

Based on Eq. (10), the network delay feature data samples are composed into N fuzzy rules, which are computed using the Hopping operator to obtain the fuzzy rule matching degree of the feature data. Namely:

$$\mu_j(x_p) = T(\mu_{j1}(x_{p1}), \cdots, \mu_{jn}(x_{pn})) \tag{11}$$

where $\mu_{j1}(x_{p1})$ represents the affiliation function of A_{ji}.

Based on Eq. (10), the following equation is utilized to obtain the rule affiliation of computer terminal multidimensional data:

$$b_j^k = T(\mu_j(x_p), r_j^k) \tag{12}$$

On the basis of the rule matching degree of computer terminal multidimensional data obtained by Eq. (11), the rule matching degree of computer terminal multidimensional data j is normalized over each classification rule to obtain the classification matching degree as:

$$\eta_j^k = \beta_{Classh}(R_j) \Big/ \sum_{i=1}^{M} \beta_{Classh}(R_j) \tag{13}$$

Since the conclusion of each fuzzy rule will correspond to a classification result, before using the match maximum classification rule to obtain the computer terminal multidimensional data rule weight value is:

$$r_j^l = \beta_{Classh}(R_j) \Big/ \sum_{i=1}^{k} \beta_{Classh}(R_j) \tag{14}$$

For computer terminal multidimensional data classification, when the classification of each computer terminal multidimensional data sample is determined by a single superiority rule post-piece class, its expression is given using the following formula:

$$Y_k = \max\left\{ b_j^k, j = 1, \cdots, N \ and \ k = c_j \right\} \tag{15}$$

When the classification of each computer terminal multidimensional data sample is determined by the posterior class of fuzzy rules for each sample, the total intensity of computer terminal multidimensional data classification is:

$$Y_k = \sum_{j=1, c_j = k}^{N} b_j^k \tag{16}$$

Based on Eq. (15) and Eq. (16), Eq. (17) is utilized to give the classification soundness of computer terminal multidimensional data samples as:

$$Y_k = f\left(b_j^k, j = 1, \cdots, N \right) k = 1, \cdots, M \tag{17}$$

The classification soundness of the feature data samples is classified to the computer terminal multidimensional data rule weight values and its classification result is given as using the following equation:

$$F = \max(Y_k), \cdots, (Y_M) \tag{18}$$

II. B. 3) Data extraction method based on PCA-ReliefF

Based on the obtained multi-dimensional data classification results of computer terminal, the feature data feature vector is obtained by using principal component analysis, and the feature data vector is extracted by using ReliefF algorithm. The specific process is as follows:

Suppose, there is $\Phi(x_i)$ a linear mapping function, there are N computer terminal multidimensional data sample points, and the computer terminal multidimensional data sample points are transformed into the high-dimensional linear feature space P. Namely:

$$\sum_{k=1}^{N}\Phi(x_i)=0 \tag{19}$$

In Eq. (19), the covariance matrix of the characterized data sample points is:

$$C=\frac{1}{N}\sum_{k=1}^{N}\Phi(x_k)\Phi^T(x_k) \tag{20}$$

Suppose, on the high-dimensional feature space P, the computer terminal multidimensional data feature vector is v, whose expression is given using equation (20):

$$v=\sum_{k=1}^{N}\alpha_k\Phi(x_k) \tag{21}$$

This is obtained by combining Eq. (19) with Eq. (20) and Eq. (21):

$$\lambda^{\varphi}\sum_{j=1}^{N}\alpha_j\langle\Phi(x_i),\Phi(x_j)\rangle=\frac{1}{N}\sum_{j=1}^{N}\alpha_j\langle\Phi(x_i)\rangle \tag{22}$$

Simplifying the above equation gives:

$$N\lambda^{\varphi}K=K\alpha \tag{23}$$

In order to extract the main features of computer terminal multidimensional data, obtaining the projection coefficients on the vector $v^{(k)}$ on the high-dimensional feature space P for the pair of feature data samples is:

$$X_k=\langle v^{(k)},\Phi(x)\rangle=\sum_{j=1}^{N}\alpha_j^k\langle\Phi(x_j),\Phi(x)\rangle \tag{24}$$

Based on Eq. (24), the input computer terminal multidimensional data $S=[x_1,x_2,\cdots,x_l]$, and the dimensionality of the projected feature data is k, which is given by utilizing the following equation to give the computer terminal multidimensional data matrix as:

$$K=(k_{ij})_{l\times l} \tag{25}$$

Centering computer terminal multidimensional data matrices:

$$K\leftarrow K-\frac{1}{l}jjK-\frac{1}{l}Kjj+\frac{1}{l^2}(jKj)jj \tag{26}$$

The eigenvectors of the multidimensional data matrix of the computer terminal are found through equation (26):

$$[A,\Lambda]=eig(K) \tag{27}$$

where eig represents the eigenvector function value of the feature data matrix K.

Based on Eq. (28), the feature data are extracted using Relief algorithm, i.e:

$$s_t(i)=\sum_{j=1}^{l}\alpha_i(j)K(x_j,x_t) \tag{28}$$

$$S = \{s_1, s_2, \cdots, s_t\} \tag{29}$$

II. C.Design of Principal Component Based Data Mapping Methods

The goal of the traditional PCA model is to map the data on the original space to the principal element space and the residual space respectively by linear transformation, and to identify and classify the multidimensional data in these two subspaces, where the principal element space represents the information about the degree of change of the process variables in the linear direction, and the residual space represents the linear redundancy that exists in the original data space. However, most of the observed variables in the modern computer terminal computing process are nonlinear relationships between them, and this nonlinear relationship and its process data are difficult to be described by the traditional principal component analysis model, and if linear principal element analysis is used to classify the terminal data of these processes, it is very easy to get the results of the wind of omissions and false alarms. For this kind of nonlinear problem of computer terminal multidimensional data, the most important method in the current research is the kernel principal component analysis algorithm.

II. C. 1) Concepts and forms of nuclear functions

The most important advantage of KPCA over traditional PCA algorithm is that it can deal with the nonlinear problem that exists in the air-cooling process of multidimensional data at the computer terminal. According to the relevant theory of pattern recognition, the linearly indivisible data in the low-dimensional space may become linearly divisible through nonlinear mapping to the high-dimensional feature space. However, if the original data of the process is directly mapped to the high-dimensional space, there is the problem of how to choose the form of the nonlinear mapping function, how to determine the parameters and the dimension of the feature space, and the biggest obstacle is the existence of “dimensional catastrophe” in high-dimensional operations, and the use of the kernel function technique can solve the problem of nonlinearities between data. The technique of kernel function can solve the problem of nonlinearity between data.

There are many kinds of kernel functions, and the functions that satisfy Mercer's theorem can be regarded as kernel functions, and the common kernel functions are the following four kinds:

(1) Linear kernel function $K(x, x_i) = (x, x_i)$.

(2) Gaussian kernel function $K(x, x_i) = \exp\left(-\frac{\|x - x_i\|^2}{2\sigma^2}\right)$.

(3) Polynomial kernel function $K(x, x_i) = (x, x_i + 1)^d$.

(4) Sigmoid kernel function $K(x, x_i) = ((x, x_i) + c)$.

Different forms of kernel functions represent different methods of mapping from low-dimensional space to high-dimensional space, and its selection has a key impact on the performance of principal component analysis algorithms, and the parameters of kernel functions need to be determined empirically, so the selection of kernel functions and and the determination of the kernel parameters bring difficulties in the application of kernel function algorithms.

II. C. 2) Data mapping based on kernel principal components

The main idea of KPCA is to map the data on the original space to the high-dimensional feature space by nonlinear mapping, and then the mapped data are subjected to principal component analysis in the high-dimensional feature space, and it has a strong ability to deal with the nonlinear data of the production process.

Firstly, the sampled dataset $X \in R^{n \times m}$ is normalized according to the computational formula, there exists a nonlinear mapping function Φ which can map the data on the original space to the high-dimensional feature space F, which can make the sample points $x_1, x_2, \cdots, x_n$ on the input space transformed into the feature space of the The sample points $\Phi(x_1), \Phi(x_2), \cdots, \Phi(x_n)$ and satisfy the center condition, viz:

$$\sum_{k=1}^{n} \Phi(x_k) = 0 \tag{30}$$

The covariance matrix of the samples in F of the feature space can be expressed as:

$$C^F = \frac{1}{n} \sum_{k=1}^{n} \Phi(x_k) \Phi(x_k)^T \tag{31}$$

The KPCA algorithm in eigenspace is realized by solving for the eigenvalues and eigenvectors of the covariance matrix C^F, viz:

$$\lambda v = C^F v = \frac{1}{n}\sum_{k=1}^{n}\Phi(x_k)\Phi(x_k)^T v \tag{32}$$

In the calculation process, we do not need to know the explicit form of the function $\Phi(x_k)$, but only need to know the form of the dot product of the original data, so KPCA introduces the kernel function way to solve the calculation of the inner product of the nonlinear transformations and high-dimensional spaces. We define a kernel function:

$$K(x,y) = \left\langle \Phi(x), \Phi(y) \right\rangle \tag{33}$$

Since the feature vector can be represented linearly by the data set, the feature vector v can be represented linearly by $\Phi(x_k)$, i.e:

$$v = \sum_{k=1}^{n}\alpha_k \Phi(x_k) \tag{34}$$

C^F is the matrix of correlation coefficients for $\Phi(x_k)$, and bringing Eqs. (31)(34) into Eq. (32) yields $n\lambda K\alpha = KK\alpha$, which is clearly obtained by the simplification of Eq:

$$\lambda\alpha = \frac{K}{n}\alpha \tag{35}$$

By solving Eq. (35), the requested eigenvalue λ with the eigenvector α, which is also the principal element direction of the high-dimensional space, i.e., the load vector, can be obtained. After calculating the load vector, for any vector x in the input matrix, the k th score vector of the high-dimensional feature space can be obtained as:

$$t_k = \langle v \cdot \Phi(x) \rangle = \sum_{k=1}^{n}\alpha_k \Phi(x_k)\cdot\Phi(x) = \sum_{k=1}^{n}\alpha_k K\langle x_k, x\rangle \tag{36}$$

In addition, the decentralization of the kernel function K is needed when using the KPCA algorithm. The kernel function $\overline{K}$ is the form of kernel function K after decentering, then $\overline{K}$ can be expressed as:

$$\overline{K} = K - KE - EK + EKE \tag{37}$$

III. Data extraction and mapping simulation analysis

III. A. Simulation analysis of computer terminal multidimensional data extraction

III. A. 1) Description of the simulation analysis

ReliefF is a feature weighting algorithm that extends from dealing with two-category problems to solving multi-category problems. Its core idea is to assign weights to features based on the relevance of each feature to the category. The larger the weight, the stronger the categorization ability of the feature. In order to validate the computerized multidimensional data extraction method using a combination of ReliefF and PCA algorithms, the computer terminal multidimensional data extraction method is numerically simulated and analyzed using MATLAB.

III. A. 2) Analysis of extraction effects

This simulation compares the multidimensional data extraction performance of three feature selection methods, wavelet transform, PCA algorithm and ReliefF-PCA proposed in this paper, in terms of classification correctness and time. 200 multidimensional signal data are extracted for each computer terminal of the sample library and 100 multidimensional signal data are extracted for each computer terminal of the sample library to be tested, and all the experimental results are taken as the average of 100 experiments. In the wavelet transform algorithm, secondary and tertiary wavelet transforms are performed respectively, and the extracted multidimensional signal data features are 400 and 200 dimensions.In the PCA algorithm, 252 principal elements are extracted from the 400-dimensional features obtained from the secondary wavelet transforms (with a selected variance contribution of 0.872). Similarly, the algorithm in this paper performs computer terminal multidimensional data extraction with PCA contribution taken as 0.921, ReliefF process selects the 300 features most favorable for classification and finally 116 principal elements are extracted. The multidimensional data extraction effect of each algorithm at different low signal-to-noise ratios

(SNR) is shown in Fig. 1, while the time performance comparison is shown in Fig. 2. It can be seen that the algorithm in this paper has lower dimensionality and higher classification correctness compared to the PCA algorithm at SNR>16. In the case of low signal-to-noise ratio (SNR), this algorithm (PCA-Relief) has a slightly higher classification correctness than the second-level wavelet transform algorithm, and then it is almost the same as it. The third-level wavelet transform can reduce the feature dimension and the computational complexity of machine learning through a higher number of wavelet transforms than the second-level wavelet transform, but the recognition rate decreases. In summary, this paper's algorithm (PCA-Relief) has a greater advantage in maintaining the recognition rate, reducing the feature dimension and the computational complexity of machine learning. In the field of computer terminal multidimensional data extraction, a multidimensional data graph extraction method based on ReliefF and PCA is proposed. The ReliefF algorithm is used to select the most favorable features for classification, and then PCA is used to remove the correlation between the features, which solves the problem of high dimensionality of computer terminal data. The single ReliefF algorithm cannot remove the disadvantage of redundant features, and the single PCA algorithm only considers the direction that guarantees the maximum variance for dimension reduction. The experimental results show that the algorithm can effectively reduce the data dimension and time consumption while ensuring the classification accuracy of multi-dimensional data of computer terminal.

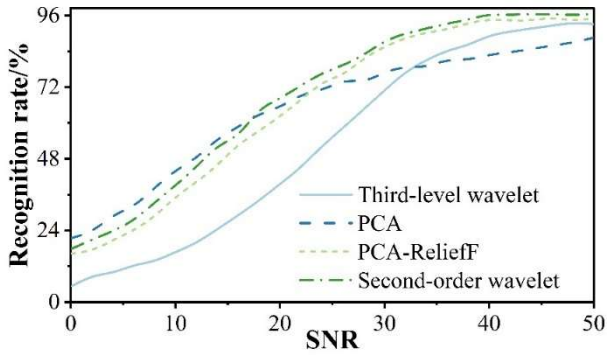

Figure 1: The effect of extracting multi-dimensional data

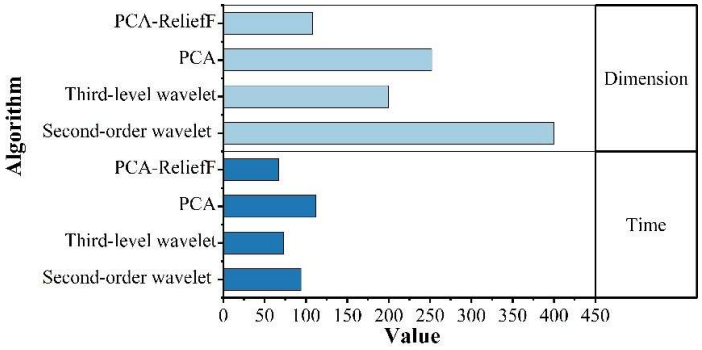

Figure 2: Time performance comparison

III. B. Simulation and analysis of multidimensional data mapping for computer terminals

III. B. 1) Traditional PCA Implementation on Matlab

Traditional PCA has two functions in Matlab to accomplish data mapping. If the covariance matrix of the input data is known, PCA data mapping can be done by pcacov() function. Let the covariance matrix be V, then through [c,l,e]=pcacov(V) can get all the eigenvalues of V in descending order l, the contribution of each principal component e and the principal component coefficients c. Can be carried out through the function barttest() for the Bartlett dimensionality check, and through the pcares() function to carry out the residual analysis of the principal components. The most commonly used function to do PCA analysis is pcacomp(), which requires only the input of the data matrix X. PCA data mapping analysis can be performed by: [c,s,l,t2]=princomp. C represents the coefficient matrix of the principal components, s represents the principal component matrix, l represents the eigenvalues of the covariance matrix of the samples arranged in descending order, and t2 represents the statistic of Hotelling's T2 corresponding to a particular sample, which can be used for computer terminal multidimensional data mapping

detection, to indicate the distance of a particular observation from the center of the observed data, and as an outlier analysis.

III. B. 2) Implementation of Kernel Principal Component Analysis on Matlab

Matlab provides the kpca() function for kernel principal component analysis in the pattern recognition toolkit stprtool. The stprtool toolkit is not a standard toolkit for Matlab 7.1, and it needs to be downloaded and set up under the Matlab path. Using the kpca function is simple but not flexible. The following algorithmic implementation of kernel principal component analysis is performed using Matlab-like. Since the data distribution in many occasions is close to Gaussian distribution, and the covariance matrix of Gaussian distribution has no loss of original information, the following kernel function uses Gaussian radial basis function, which requires an input parameter. The steps are as follows:

(1) Prepare the training data T, test data TT, and set the parameter rbf of Gaussian radial basis function, the cumulative contribution size is thres.

(2) Normalize the training data T, test data TT.

(3) Apply the Gaussian kernel function to calculate the kernel matrix, because the kernel matrix is symmetric, only need to calculate the upper triangular matrix, the other half of the elements with the assignment to complete, can improve the efficiency of the operation.

(4) Apply K_new=K-U*K-K*U+U*K*U centered Gaussian kernel matrix.

(5) Calculate eigenvalues and eigenvectors for K_new: [ev,evalues]=eig(K_new).

(6) Calculate the cumulative contribution rate.

(7) Find the eigenvalues whose cumulative contribution rate is greater than or equal to thres, and form the corresponding eigenvectors into a matrix.

(8) Test the test data TT.

The above algorithm can be seen that the kernel function is adopted, and PCA becomes kernel principal component analysis, but the specific implementation process is very similar to the traditional PCA, but brings the effect that nonlinear mapping can be performed. And with the adoption of the kernel function, the mapping from linear space to higher dimensional space does not need to be specifically concerned with the form of the mapping function.

III. B. 3) Mapping effect analysis for nuclear principal component analysis

The following is the mapping analysis of the computer terminal multidimensional data from Matlab 7.1 with the above kernel principal component analysis algorithm, the 1st principal component and 3rd principal component after kernel principal component analysis are shown in Fig. 3, and Fig. 4 is the principal component pareto chart. A total of 20 cities' computer terminal data statistics were collected, which had a total of 9 indicators. The data values of different indicators vary greatly and need to be normalized. From the results of the run, the kernel principal component analysis is able to map the multidimensional data of the computer terminal, and the effect is comparable to the traditional PCA accuracy, using the tie, toe instruction for timing, the traditional PCA execution time is about 2.582 seconds, and the kernel principal component analysis takes 4.573 seconds, which is because the kernel principal component analysis needs to compute the kernel matrix, and the Gaussian kernel needs to compute the two-paradigm number as well as the exp operation. However, kernel principal component analysis has an increase in computation due to the use of kernels, but it maps effectively to both linear and nonlinear input spaces.

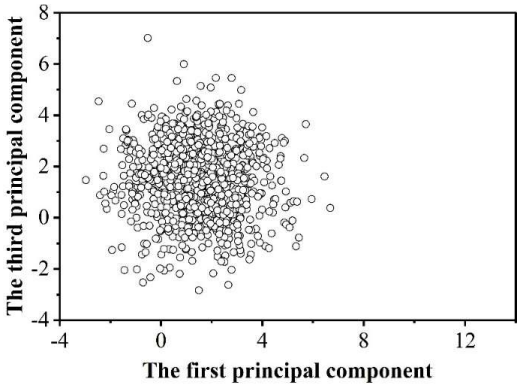

Figure 3: The first principal component and the third principal component

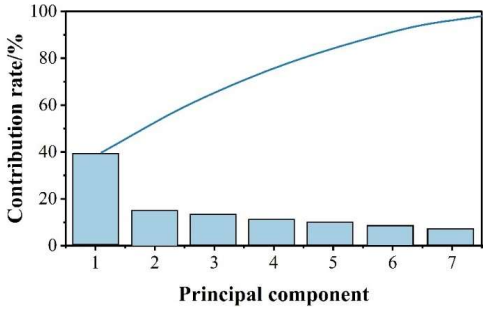

Figure 4: Principal component pareto plot

III. B. 4) Comparative analysis of mapping effects

(1) Contribution rate comparison

The experiment utilizes Matlab 7.1 to write two data mapping algorithms, PCA and KPCA, which are tested using the multidimensional data of computer terminals obtained above. In order to compare the difference between PCA algorithm and KPCA algorithm, a fixed threshold T (i.e., cumulative contribution rate) ≥ 0.95 was set to obtain the contribution rate and cumulative contribution rate of each principal component component after the dimensionality reduction mapping, and the average value was obtained by repeating 5 times, and the results are shown in Table 1. After dimensionality reduction mapping by PCA algorithm, the first 5 principal component components of the data express 94.34% of the information of the whole image data, while after dimensionality reduction by KPCA algorithm, only the first 3 principal component components contain 94.63% of the information of the original image data, which indicates that both PCA and KPCA algorithms are able to carry out dimensionality reduction mapping of the computer terminal multi-dimensional data, and achieve the purpose of data compression and simplification, while KPCA algorithm, as the main component of computer terminal multi-dimensional data, is able to reduce and mapping of the data. The KPCA algorithm, as a kernel function extension of the PCA algorithm, has stronger mapping ability and better dimensionality reduction effect when dealing with computer terminal multidimensional data on the basis of containing more original information as much as possible.

Table 1: Contribution rate comparison

Principal component portion	PCA algorithm		KPCA algorithm	
	Contribution rate	Cumulative contribution rate	Contribution rate	Cumulative contribution rate
PC1	0.6824	0.6824	0.5541	0.5541
PC2	0.2327	0.9151	0.3817	0.9358
PC3	0.0144	0.9295	0.0105	0.9463
PC4	0.0075	0.937	0.0066	0.9529
PC5	0.0064	0.9434	0.0048	0.9577

(2) Running time comparison

The running time comparison is shown in Fig. 5, which compares the time complexity of the two dimensionality reduction algorithms by calculating their time. Both algorithms are run 5 times to take the average value, comparing their running costs, PCA algorithm takes 32.77s, KPCA algorithm takes 42.73s, PCA algorithm has a more obvious advantage. The reason is that the KPCA algorithm does a nonlinear mapping, the eigenvalue decomposition of the covariance matrix in the high-dimensional space, and then use the same way as the PCA algorithm for the mapping, in the high-dimensional space KPCA algorithm applies the kernel function operation, and the computational amount is much more than the PCA algorithm.

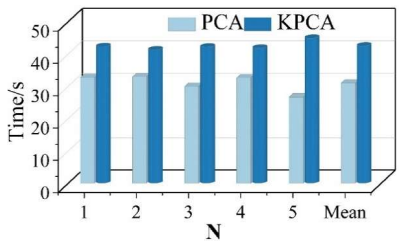

Figure 5: Comparison of running time

(3) Accuracy comparison

In order to further compare the differences between the two mapping algorithms and fully analyze the feature differences between the two, this paper chooses the mapping effect evaluation method based on classification application, with MLP as the classifier.MLP classifier is one of the mainstream classifiers in the current computer field, which is composed of a parallel combination of many identical processing units, and it can carry out a large number of parallel activities of simple units, and it is more suitable for the processing ability of the information and universality. Stronger, more suitable for computer terminal data such as large-scale data classification processing, there are more application examples in computer terminal multidimensional data mapping, it is easier to verify the mapping effect of PCA algorithm and KPCA algorithm. The mapping effect of the two mapping algorithms based on the MLP classification application is shown in Fig. 6, which shows that the classification accuracy of the data of the 1st, 7th and 8th is 100%, and the classification of these 3 types of data has no value due to the fact that the samples of these 3 types of data are too small to do effective prediction. The data number 0 is the background value of the image, which is the main factor of misclassification in the classification. Overall the accuracy of KPCA mapping algorithm is slightly better than PCA mapping algorithm. Overall KPCA algorithm applies the kernel function in the low-dimensional space, which cleverly ascends from low to high dimensions, and realizes the nonlinear mapping in the high-dimensional space, which greatly improves the accuracy of principal component analysis.

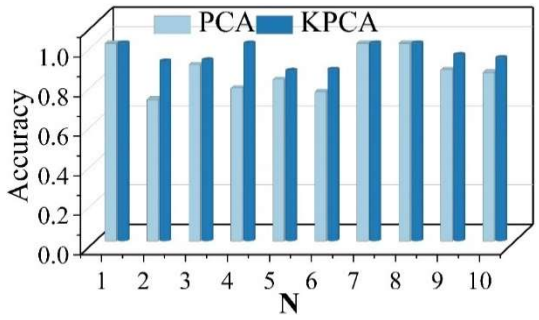

Figure 6: Accuracy rate comparison

IV. Conclusion

In this paper, under the framework of computer terminal, it is found that there are problems such as redundancy and high dimensionality in the acquired computer terminal data. For the problems described above, the principal component analysis algorithm is adopted to process the feature dimensionality reduction and mapping, for which the data extraction algorithm based on PCA-RelifeF and the data mapping algorithm based on KPCA are designed respectively. Matlab is used as the simulation and analysis tool in this study to numerically simulate and analyze the above two algorithms. The traditional PCA algorithm is based on adding the kernel function to realize the nonlinear mapping from low-dimension to high-dimension, and although the algorithm operation time increases by 20s, the accuracy of the kernel is significantly improved (0.823 increases to 0.956), which makes it better to meet the user's data management needs of computer terminals.

Funding

Qinghai Electric Power Company support project (project No:522807240004) of “Research on Deep Threat Detection and Attack Forensic Technology Based on Multi source Heterogeneous Hosts in Power Monitoring System”.

References

[1] Lin, R. H., Chuang, W. W., Chuang, C. L., & Chang, W. S. (2021). Applied big data analysis to build customer product recommendation model. Sustainability, 13(9), 4985.

[2] Liao, S. H., & Yang, C. A. (2021). Big data analytics of social network marketing and personalized recommendations. Social Network Analysis and Mining, 11(1), 21.

[3] Alyass, A., Turcotte, M., & Meyre, D. (2015). From big data analysis to personalized medicine for all: challenges and opportunities. BMC medical genomics, 8, 1-12.

[4] Zhang, Y., Liu, T., Li, K., & Zhang, J. (2017). Improved visual correlation analysis for multidimensional data. Journal of Visual Languages & Computing, 41, 121-132.

[5] Lupton, R. C., & Allwood, J. M. (2017). Hybrid Sankey diagrams: Visual analysis of multidimensional data for understanding resource use. Resources, Conservation and Recycling, 124, 141-151.

[6] Bondarev, A. E., & Galaktionov, V. A. (2019). Generalized computational experiment and visual analysis of multidimensional data. Scientific Visualization, 11(4).

[7] Shahid, M. L. U. R., Molchanov, V., Mir, J., Shaukat, F., & Linsen, L. (2020). Interactive visual analytics tool for multidimensional quantitative and categorical data analysis. Information Visualization, 19(3), 234-246.
[8] Ventocilla, E., & Riveiro, M. (2020). A comparative user study of visualization techniques for cluster analysis of multidimensional data sets. Information visualization, 19(4), 318-338.
[9] Cui, W., Strazdins, G., & Wang, H. (2021). Visual analysis of multidimensional big data: A scalable lightweight bundling method for parallel coordinates. IEEE Transactions on Big Data, 9(1), 106-117.
[10] Nonato, L. G., & Aupetit, M. (2018). Multidimensional projection for visual analytics: Linking techniques with distortions, tasks, and layout enrichment. IEEE Transactions on Visualization and Computer Graphics, 25(8), 2650-2673.
[11] Chen, L. H., & Jiang, C. R. (2017). Multi-dimensional functional principal component analysis. Statistics and Computing, 27, 1181-1192.
[12] Mishra, S. P., Sarkar, U., Taraphder, S., Datta, S., Swain, D., Saikhom, R., ... & Laishram, M. (2017). Multivariate statistical data analysis-principal component analysis (PCA). International Journal of Livestock Research, 7(5), 60-78.
[13] Happ, C., & Greven, S. (2018). Multivariate functional principal component analysis for data observed on different (dimensional) domains. Journal of the American Statistical Association, 113(522), 649-659.
[14] Di Salvo, F., Ruggieri, M., & Plaia, A. (2015). Functional principal component analysis for multivariate multidimensional environmental data. Environmental and ecological statistics, 22, 739-757.
[15] Hellton, K. H., & Thoresen, M. (2017). When and why are principal component scores a good tool for visualizing high-dimensional data?. Scandinavian Journal of Statistics, 44(3), 581-597.
[16] Du, T. Y. (2019). Dimensionality reduction techniques for visualizing morphometric data: Comparing principal component analysis to nonlinear methods. Evolutionary Biology, 46(1), 106-121.
[17] Zahabi Maryam,Pankok Carl & Park Junho. (2020). Human factors in police mobile computer terminals: A systematic review and survey of recent literature, guideline formulation, and future research directions. Applied ergonomics,84,103041.
[18] Nitin Kumar Chauhan,Krishna Singh,Amit Kumar,Ashutosh Mishra,Sachin Kumar Gupta,Shubham Mahajan... & Jungeun Kim. (2025). A hybrid learning network with progressive resizing and PCA for diagnosis of cervical cancer on WSI slides. Scientific Reports,15(1),12801-12801.