A New Visual Analysis Approach to the High Dimensional Data

Meina Song¹,a, Zhihui Pang¹,b and Haihong E¹,c
¹ Beijing University of Posts and Telecommunications, Beijing 100876, China
³ mnsong@bupt.edu.cn, b pangzhihui@bupt.edu.cn, c ehaihong@bupt.edu.cn

Abstract. With the advent of cloud computing and big data era, high-dimensional data sets are widely available in real life. Because of the increase of data dimension and complexity, it is difficult to carry out comprehensive analysis and exploration for high-dimensional data. Therefore, we propose a new visual analysis approach to high-dimensional data. Firstly, we use the principal component analysis (PCA) to reduce the dimension of the high-dimensional data. Then, we use clustering algorithm to classify the reduced dimension data, and each class is rendered independently with different color. Finally, we use the edge binding algorithm to perform a visual clustering. The curves in the same classes are converged and in different classes are separated to alleviate visual confusion. In order to analyze the visualization results better, we also provide the visual interaction technology of "brush". The experimental result shows that our approach can help users to explore the implicit feature patterns quickly and is effective for visual analysis of large-scale high-dimensional data.

1. Introduction

With the development of the network and the rise of the mobile Internet, the data grows explosively. Not only is the amount of data huge, but each piece of data has many dimensions. Moreover, the relationship between data and data is very complicated. Because of high-dimensional data's complex structure, users are difficult to directly find the characteristics and rules of the data. Therefore, how to help users understand the internal structure and rules of high-dimensional data conveniently and quickly has become a key research topic.

Visualization can present the structural information and feature patterns which is hidden in the data in the form of graphics. It is an important means to help users to understand complex phenomena and interpretation of complex data [1]. At present, visualization and analysis for high-dimensional data is a hot spot in the field of visualization. The commonly used visualization method is decomposing the multidimensional data to the low dimensional space. The classical dimension reduction methods include Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), Multidimensional scaling Analysis (MDS) and so on. Although the dimension reduction strategy can maintain some distribution and structural characteristics of high-dimensional data space, when the data’s dimensions are large, the loss of data information is large afe. Parallel coordinate technology is a classical high dimensional data visualization method, which effectively integrates the advantages of reduced dimension and scattered plot matrix. This technique can not only directly represent high dimensional data in two-dimensional space, but also supports the reading of original information and relational mining of each dimension. But it also has some limitations. When the data scale is large, there will be a large number of overlaps between edges and edges. This will lead to visual confusion.
Parallel coordinates’s visual confusion will seriously affect the data analysts observing the characteristics of data.

Based on the above analysis, in this paper, we propose a new visual analysis approach to high-dimensional data based on parallel coordinates. Firstly, we use the principal component analysis (PCA) to reduce the dimension of the high-dimensional data. Then, we use clustering algorithm to classify the reduced dimension data, and each class is rendered independently with different color. Finally, we use the edge binding algorithm to perform a visual clustering. In order to analyze the visualization results better, we also provide the visual interaction technology of "brush". The experimental results can further verify the effectiveness and practicability of the visual analysis approach in this paper.

2. Related Works
In the field of information visualization, the representation, analysis and visualization of high-dimensional data have always been a hot topic. As an important visualization technology in the field of high-dimensional data visualization, parallel coordinates [2] have been widely used in the visualization and analysis of high-dimensional data. In order to meet different needs, many researchers have improved and extended the traditional parallel coordinates.

Clustering algorithms are introduced to improve the visualization of parallel coordinates. There are many clustering methods for high-dimensional data, such as K-means, SOM, hierarchical clustering, etc. Liu et al. [3] proposed an unsupervised self-organizing mapping method to cluster financial data, and then distinguish different classes by opacity and color. Wang et al. [4] proposed a multi-dimensional temporal data visualization scheme based on clustering, and used the parallel coordinate clustering method to solve the problem of line overlap coverage in big data environment. The above methods are single data-based clustering. Due to the large range of data sets, the rendering results after clustering in parallel coordinates may still be very messy. Different from clustering data directly, Zhou et al. [5] has proposed a geometric visual clustering algorithm. The algorithm maximizes the parallelism between edges and adjacent edges while minimizing the curvature of edges. Finally, the effect of edge clustering is achieved. Guo et al. [6] has designed an interactive visual clustering method for local data in a graph. By setting the gravity and repulsion operation points at any position in the view, the user can create a real-time convergence or dispersion of the curves in the adjacent region. Then the hierarchical visual clustering effect is generated. However, the above methods are based on a single visual space clustering, which will lead to the clustering results do not take into account the actual significance of the data itself, affecting the accuracy of the clustering. Dimension rearrangement and filtering techniques are also commonly used to reduce the visual clutter of parallel coordinates. Liang [7] proposes a nonlinear correlation method to calculate the similarity between dimensions and adjust the position of the dimensions so that the dimensions are adjacent to each other when the parallel coordinate visualization dimension is rearranged. Peng et al. [8] designed a metric to compare the degree of confusion in different dimensional sequences, and finally chose the best arrangement scheme. Z Idrus et al. [9] extracted seven filtering designs by abstracting the graph, and proposed a structured process using filtering techniques in parallel coordinates.

3. Visual Analysis approach to High dimensional data

3.1. Dimensionality reduction of Multidimensional attributes based on PCA
In high-dimensional datasets, data dimensions have a certain degree of correlation with each other. Therefore, information overlap and feature redundancy often exist in high-dimensional data. In addition, in the process of high-dimensional data analysis, too many dimensions will increase the complexity of computation and analysis. Therefore, this paper introduces the PCA algorithm to reduce the dimension of high-dimensional datasets.

The steps of the algorithm as follows:
1. From the original dataset, select \( p \) dimensional random vectors and \( n \) samples \( x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T, i = 1, 2 \ldots n, n > p \) to construct sample matrix \( X \).

2. Standardized preprocess the variables of the sample matrix \( X \). Get the standardized data matrix \( Z \).

\[
Z_{i,j} = \frac{x_{i,j} - \bar{x}_i}{s_i} , \quad i = 1, 2 \ldots n, \quad j = 1, 2 \ldots p
\]  

Where

\[
\bar{x}_i = \frac{1}{n} \sum_{j=1}^{n} x_{i,j} \quad s_i^2 = \frac{1}{n-1} \sum_{j=1}^{n} (x_{i,j} - \bar{x}_i)^2
\]

3. Compute Correlation Matrix (Covariance Matrix)

\[
C = \frac{1}{n} Z^T Z
\]

4. Compute Eigenvalue Matrix \( \Lambda = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_m) \) and Eigenvectors of Correlation Matrix \( C \).

5. Arrange the eigenvalues in descending order: \( \lambda_1 > \lambda_2 > \cdots \lambda_m \). And adjust the order of the eigenvector so that the first principal component has the largest variance, the second principal component has the sub-large variance, and the smallest variance corresponds to the \( p \) principal component.

6. Select \( k \) principal components with the largest variance. So that the cumulative variance contribution of \( k \) principal components is more than 85% of the total variance. That is \( \sum_{j=1}^{k} \lambda_j / \sum_{j=1}^{P} \lambda_j \geq 85\% \). This is a generally accepted criterion, but the selection of a few depends on the actual situation.

7. From the selected eigenvector of \( k \) principal components, \( k \) new independent linear combination variables are obtained. According to the above steps, the initial high dimensional data can be reduced. We can obtain new data with less relative dimension and little change in information content.

3.2. The reduced dimension data clustering

Although the data dimension is reduced, the data is still disorganized and difficult to understand. In order to help users quickly identify the features of interest, we use K-means++ clustering algorithm to classify the reduced dimension results and enhance the display. The K-means++ clustering algorithm is used in this paper because it is more efficient and scalable in clustering a large number of high-dimensional data. And it overcomes the problem of selecting the initial clustering center by K-means, which makes the clustering results more accurate. After clustering, we can easily get the center of each cluster, and provide the control points of common path for edge binding algorithm in next section.

For the determination of clustering number \( k \), the Average silhouette method [10] is used as the evaluation index of clustering effect. The Average silhouette method is the evaluation index of the density and dispersion of the class, and its formula is as follows:

\[
s(i) = \frac{b(i) - a(i)}{\max(a(i),b(i))}
\]

Where \( a(i) \) is a measure of similarity within the group, \( b(i) \) is a measure of similarity between groups. \( s(i) \) ranges from \(-1\) to \(1\). The larger the value, the higher the intragroup anastomosis and the farther the distance between groups. In other words, the larger the contour coefficient, the better the clustering effect.

Through the above clustering algorithm, all the data are eventually divided into \( k \) groups. We set the corresponding color for each group when rendering. The selection of cluster colors is carefully designed. The color’s contrast is good, which is helpful for users to distinguish various types, increase the visual differences of various types, and enhance the user's understanding of high-dimensional data sets.

3.3. Visualization of parallel coordinates based on Edge binding
We use parallel coordinate technology to visualize the high-dimensional data preprocessed by reduced-dimension and clustering, as shown in Figure 1(a). Each axis of parallel coordinate represents a principal component that we retain. Because of the large scale of high-dimensional data, the distribution of line segments still be interlaced and overlapped. It is difficult for users to identify data characteristics and trends for each category. In this paper, we propose an edge binding algorithm, which gathers curves in the same class and separates curves among different classes, so as to separate confused data with different characteristics and assist users to do further analysis.

When using parallel coordinates to visualize, this paper uses Bezier curve instead of the original broken line. Before rendering each group, add a virtual axis before each attribute axis of the parallel coordinates. As shown in the Figure 1(b), taking the attribute axis PC1 as an example, its corresponding virtual axis is VPC1. The virtual axis is used to display the coordinates of the cluster centers on this attribute. The coordinate points on the virtual axis are regarded as the control points of the Bezier curve. These edges of the same class must pass through the corresponding control points, so that the edges belonging to the same class can be close to each other. At the same time, by adjusting the range of the virtual axis coordinates, the coordinate points of each cluster center on the virtual axis are arranged uniformly in the coordinate axis height range as far as possible, and the distance between clusters is increased to improve the visual identification. Finally, remove the virtual axis and only retain the attribute axis, we can obtain the visualization results of the parallel coordinates, as shown in Figure 1(c).

![Figure 1. (a) The result after reducing dimension and clustering; (b) The result of adding a virtual axis for edge binding; (c) The result after binding the edges](image)

3.4. Visual interaction
After using the edge binding algorithm for visual clustering of each class, we can clearly see the data trends of each class. However, there is still some overlap between classes in the curves near the attribute axis. In order to reduce the impact of overlap, this paper provides a parallel coordinate visualization interactive technology called brush. When the mouse selects a class on the property axis, it can be highlighted, and the unselected class is grayed out. In this way, we can observe the information of a particular class without the influence of other classes, highlight the subset of data, and help users to better analyze high-dimensional data.

4. Experiment and result analysis
In order to verify the effectiveness of this algorithm, this paper takes the classical high-dimensional data set Boston house price data set as the simulation experimental data. The data set contains 506 records, covering 14 features of different suburban houses in Boston, Massachusetts. In this paper, 13 features are selected for analysis. They are CRIM, ZN, INDUS, NOX, RM, AGE, DIS, RAD, TAX, PTRAIO, B, LSTAT, MEDV. These 13 characteristics are defined as per capita crime rate, the proportion of residential land in excess of 25000 sq.ft, the proportion of non-commercial land, nitric oxide concentration, the number of rooms per residence, the proportion of self-occupied units built before 1940, weighted distance from the five Boston job centres, convenience index from the highway, real estate tax rate for every $10,000, teacher-student ratio, black proportion, proportion of the underprivileged, average house price of self-housing. Parallel coordinate visualization technique is directly used to display the house data, and the visualization results are shown in the Figure 2. Because of the large dimension of the data, the parallel axis is crowded and the lines are messy. The visualization effect is not ideal and the rules of data can not be obtained directly from this figure.
In this paper, we use the principal component analysis to reduce the dimension of the high dimensional data set. After analysis, we find that the cumulative contribution rate of the first five principal components had reached 84.11%. In the end, the 13 dimensions of the source data were reduced to 5 dimensions. They are the location & environment, the quality of houses, the security situation (The value is greater, the security situation is worse), the education level and the proportion of black. The data after dimensionality reduction are displayed in parallel coordinates, as shown in the Figure 3.

After the dimensionality reduction, we use K-means++ to cluster the reduced dimension data. The k value is evaluated by using the average silhouette method mentioned in the previous section. From the Figure 4 we can analyze that the clustering number of the data should be 6.

After the data is processed by K-means++ clustering algorithm, the data set is divided into six clusters. And the cluster center of each cluster can be obtained. We render each cluster independently with different color attributes. Finally, we display the processed data in parallel coordinates, as shown in Figure 5. Although these data are divided into six categories, it is still very difficult to distinguish. Further processing is needed to enable the user to see more clearly the interlinkages of the data. Figure 6 shows the effect of using the edge-binding algorithm in this paper, and we can clearly see the range of the six classes.

We can analyze a separate category by using the visual interaction technique of “brushes”, as shown in figure 7. From figure 7, we can see that this kind of housing is scattered in the dimension of housing quality, which should be paid more attention to when choosing. The above method can help us to make the final decision by deeply analyzing the high dimension.
5. Conclusion
In this paper, we propose a new visual analysis approach to high-dimensional data. This approach reduces the confusion of visualization and improves the visual clustering effect of high dimensional data. First, we use principal component analysis to reduce the dimensionality of high-dimensional data. Then we use clustering algorithm to classify the reduced dimension data. The classification data are visualized by parallel coordinates. Finally, we use parallel coordinates to visualize and use edge bundling to do a visual clustering. So as to form a clear visual clustering result which is convenient to find the trend of data changes. Our method solves the problem of visual confusion existing in traditional parallel coordinates, improves the visualization ability of parallel coordinates to high-dimensional data, and can help users quickly explore the hidden feature patterns in the data.

In the future work, in order to solve the problem of overlapping curves on coordinate axes, we plan to further improve the high-dimensional data visualization technology based on parallel coordinates. At the same time, it is necessary to further improve the interaction, so that users participate in more visual analysis, so that the visual analysis results are more accurate.

Acknowledgements
This work is supported by Project of National Science and Technology Resources Sharing Service Platform Plan (Grant No. YCZYPT[2017]01-8).

References
[1] Keim D A, Kohlhammer J, Ellis G, et al. 2013 Mastering the information age-solving problems with visual analytics (Airela-Ville: Eurographics Association Press)
[2] Inselberg A. 1985. The plane with parallel coordinates. The Visual Computer,1(2) 69-91
[3] Liu F, Tian K, Zhou Z G, et al. 2012. Research on visualization method and its application based on SOM clustering. Application Research of Computers, 21(7) 3182-3193.
[4] Wang Ruisong. 2016. Research on Visualization of Space-time Multidimensional data in big data Environment. Zhejiang University,
[5] Zhou H, Yuan X R, Qu H M, et al. 2008. Visual clustering in parallel coordinates. Computer Graphics Forum, 27(3) 1047- 1054
[6] Guo P H, Xiao H, Wang Z C, et al. 2010. Interactive local clustering operations for high dimensional data in parallel coordinates. IEEE Pacific Visualization Symposium (Los Alamitos: IEEE Computer Society Press) p 97-104
[7] Liang Fu Lu, Mao Lin Huang, Tze-Haw Huang. 2012. A new axes reordering method in parallel coordinates visualization. Machine Learning and Applications, 252-257
[8] Peng W, Ward M O, Rundensteiner E A. 2004. Clutter reduction in multi-dimensional data visualization using dimension reordering. IEEE Symposium on Information Visualization (Los Alamitos: IEEE Computer Society Press) p89-96
[9] Idrus Z, Zainuddin H, Ja’Afar A D M. 2017. Designing flexible filtering in parallel coordinate graph Visual analytics. 9(5S) 23.
[10] Spurr B D. 1987. Density estimation for statistics and data analysis. by B. W. silverman. Journal of the Royal Statistical Society. 150(4) 403-404