High Dimensional Data Visualization Analysis Based on Unsupervised Laplacian Score

Hao Peng[1,2](✉), Jian Zhou[1], and Shenglan Liu[1]

[1] School of Computer Science and Technology,
Dalian University of Technology, Dalian 116024, Liaoning, China
[2] School of Mathematics and Physics,
Qingdao University of Science and Technology, Qingdao 266061, Shandong, China

Abstract. With the rapid development of big data technology and information visualization technology, the concept of data visualization is constantly evolving and developing. As one of the classic high-dimensional data visualization methods, the parallel coordinate axis has excellent plane geometric characteristics. However, as the amount of data increases and the dimension of the data feature increases, the number of polylines on the finite plane of the parallel coordinate graph also increases. The crossing and occlusion of lines lead to serious visual redundancy and clutter. This project uses the feature distribution and feature axis arrangement on the parallel axis as the research entry point, and uses two unsupervised feature selection methods (Laplacian Score and SVD-Entropy) to re-arrange the features on the PCP axis to improve parallelism. Phenomena such as data disorder and clutter on the coordinate axis. Furthermore, we proposed a plane geometry optimization CLS algorithm by combining two unsupervised feature selection algorithms and the PCP axis radius coverage calculation method. The proposed algorithm conforms to people's perception characteristics of information and plane space representation, and can help people more quickly analyze and understand data.

Keywords: High dimensional data visualization · Parallel coordinate axis · Laplacian score · SVD-entropy

1 Introduction

With the development of digital multimedia, computer networks, and information network media, the Internet has become an indispensable way for human life to obtain information [1, 2]. Among them, digital information visualization is the main expression vehicle for Internet information dissemination. High-dimensional data mining is a kind of data mining based on high dimensionality [3–5]. The main difference between it and traditional data mining is its high dimensionality. High-dimensional data mining has become the focus and difficulty of data mining [6]. With the advancement of technology, data collection has become easier and easier, leading to larger and larger databases and higher complexity, such as various types of trade transaction data, Web documents, gene

I. Ahmad et al. (Eds.): STSIoT 2021, LNDECT 122, pp. 113–118, 2023.
https://doi.org/10.1007/978-981-19-3632-6_15

expression data, document word frequency data, The dimensions (attributes) of user rating data, WEB usage data, and multimedia data can usually reach hundreds to thousands of dimensions, or even higher. It is worth noting that the dimensional characteristics of high-dimensional data have many influences on data visualization analysis [7]. First of all, due to the continuous increase of sample dimensions, a large number of traditional statistical methods have lost their effects. This is the "curse of dimensionality" that people often refer to. At the same time, people have noticed that when the number of samples is limited, the rapid increase of feature dimensions will lead to a sparse pattern of data distribution in space, and the correlation between multiple random variables that affect the sample will also increase [8]. Although technological progress has made it possible to obtain a large number of characteristics of observed individuals, including discrete, continuous and even dynamic data [9]. However, people have discovered in research that not all recorded data are related to research interests or provide effective information. The purpose of our data analysis is to extract useful information, so extracting the most effective data from the complex data set has become the focus of research [10]. The characteristic variables of the data are screened, and information useful to the research problem is selected from a large number of characteristic changes, and redundant variables and even noise variables are deleted, so that the processed data set tends to be low-dimensional and concise. Choose feasible methods to effectively reduce the dimension, so that the streamlined data characteristics better reflect the process of data generation, have better interpretability for actual problems, and can greatly reduce costs [11].

The existing parallel axis and RadViz dimensional sorting algorithms are mostly a black box process. The final dimensional order is directly given, and users are rarely involved in the algorithm process. This makes it difficult for users to understand the algorithm process and makes it difficult for users to recommend later. The order of dimensions is effectively adjusted to find a better order of dimensions. For example, Zhang et al. used a hierarchical clustering algorithm with variable parameters in consideration of RadViz characteristics to recommend the initial order of dimensions, and provided a dendrogram showing the results of the algorithm to guide users to interactively adjust, select, and delete dimensions, and perform features. Subset selection. Experiments show that the method in this paper has good interactivity, pays attention to user experience, and reduces the overlap problem of projection points in RadViz. Zhou et al. used parallel coordinates to present geospatial multi-dimensional attribute information, introduced mutual information to measure the correlation between geographic spatial clustering and attribute categories, dynamically determined the order of parallel coordinate axes, and further calculated the binding of data lines between the attribute axes and the map. Determine the location, optimize the layout of the data line, and reduce the disorder of the data line distribution between the map and the parallel coordinate system.

The data that needs to be processed in the fields of computer vision, multimedia analysis, etc. often have a very high dimensionality. The processing of high-dimensional data increases the time and space complexity of the operation, and also leads to the over-fitting phenomenon of the learning model. The view of manifold learning believes that due to the limitation of the internal characteristics of the data, some high-dimensional data will have dimensional redundancy. In fact, these data can be uniquely represented as long as they use a relatively low dimensionality. In addition, not all features are related to the learning task. The above two points show that it is necessary and possible to reduce the dimensionality of these high-dimensional data. Feature selection is a commonly used method of dimensionality reduction. It selects a group of features related to classification from a feature set through a certain algorithm, and uses the selected features for model learning. This method does not change the original representation of the data, and when the selected feature is determined, it only needs to simply extract the feature directly from the original feature set. Feature selection methods can be divided into supervised feature selection methods and unsupervised feature selection methods according to whether there is classification information in the training data. There is a large amount of unlabeled data in the real world, and the labeling of the data requires a high price, so the research on unsupervised feature selection methods has great practical significance. This project mainly uses unsupervised feature selection methods Laplacian Score and Singular Value Decomposition (SVD-Entropy) for experiments.

2 Algorithm Formulation

2.1 Calculation of Laplacian Score

The calculation of Laplacian Score is mainly based on Laplacian Eigenmaps and Locality Preserving Projection. To some extent, the Laplacian score of a feature can be regarded as the Rayleigh quotient of the related feature with respect to the Laplacian graph G. The Laplacian score of each feature is calculated by its local retention ability, which can be represented as follows:

Laplacian Score Algorithm	
Step 1:	Constructing similarity matrix: S: $S_{ij} = \begin{cases} \exp(-\frac{\|x_i - x_j\|^2}{t}) & \text{if } x_i \in KNN(x_j) \text{ or } x_j \in KNN(x_i) \\ 0 & \text{otherwise} \end{cases}$
Step 2:	Let $\tilde{f_r} = f_r - \frac{f_r^T D1}{1^T D1} 1$ with $D = diag(S1)$, $1 = [1, 1, \ldots, 1]^T$, $L = D - S$
Step 3:	Computing each LS of the corresponding features: $L_r = \frac{f_r^T L f_r}{f_r^T D f_r}$

2.2 Formulation of SVD-Entropy

Assuming that there is a matrix $M \times N$, an orthogonal basis $\{v_1, v_2, ..., v_n\}$, the orthogonal basis is mapped to: $\{Av_1, Av_2, ...Av_n\}$ using the mapping matrix A. If you want to make them orthogonal to each other:

$$Av_i \cdot Av_j = (Av_i)^T Av_j = v_i^T A^T Av_j = 0 \tag{1}$$

For singular values, it is similar to the eigenvalues in our eigendecomposition. The singular value matrix is also arranged from largest to smallest, and the singular value is reduced in special blocks. In many cases, the top 10% or even 1% of singularities the sum of values accounts for more than 99% of the sum of all singular values. In other words, we can approximate the matrix with the largest k singular values and the corresponding left and right singular value vectors:

$$A_{m\times n} = U_{m\times m} \sum_{m\times n} V_{m\times n}^T \approx U_{m\times k} \sum_{k\times k} V_{k\times n}^T \tag{2}$$

In summary, the PCP axis obtains feature reconstruction weights as follows:

$$\arg\min_{w_i} RE(w_i) = \left\| x_i - \sum_{j\in N(x_i)} w_{ij}x_j \right\|_2^2 + \xi \left\| \sum_{j\in N(x_i)} w_{ij}d_{ij} \right\|_2^2 \tag{3}$$

where w_{ij} represents the reconstruction weight of reconstruction x_j, and ξ is the regularization parameter.

3 Experimental Results and Analysis

Parallel coordinates are an important technology for information visualization. A significant advantage of parallel coordinates is that it has a good mathematical foundation, and its projective geometric interpretation and duality characteristics make it very suitable for visual data analysis. The parallel coordinate method uses coordinate axes that are parallel to each other. Each coordinate axis represents an attribute of the data. Each high-dimensional data is represented as a polyline connecting its data value points on each coordinate axis (dimension) to form a polyline. Each line segment reflects the value of high-dimensional data on two adjacent coordinate axes. For ease of expression, it is called a data line. Usually between a pair of adjacent coordinate axes, the data lines of all data are overlapped together, and the data correlation in the adjacent dimensions of the data set can be reflected from the overall distribution characteristics of the data lines. In order to overcome the problem that the traditional Cartesian rectangular coordinate system is easy to run out of space and difficult to express data above three dimensions, Parallel Coordinates uses a series of parallel coordinate axes to represent each variable of high-dimensional data, and the value of the variable corresponds to the position on the axis. In order to reflect the trend of change and the relationship between various variables, the points describing different variables are often connected into a broken line. The PCP axis based on LS and SVD can be re-arranged according to the feature score, as shown in the following figure (Fig. 1):

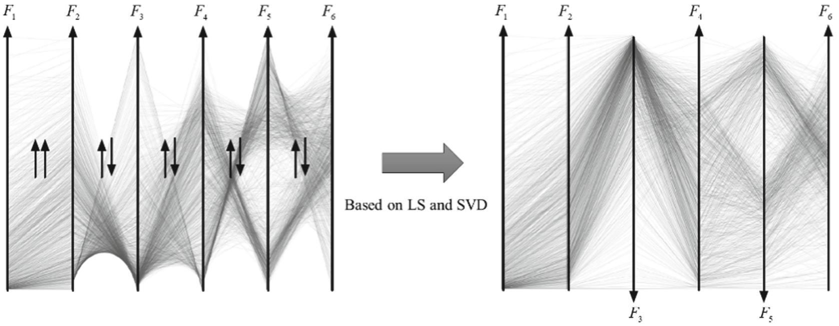

Fig. 1. PCP axis arrangement based on Laplacian score and SVD-entropy

In order to further verify the visualization performance of original PCP with different feature selection methods, we employed four evaluation metrics (Accuracy; Time; Satisfaction; Average R2), which can be seen in Table 1:

Table 1. Visualization comparison between PCP and its variants

Methods	Accuracy (%)	Time (s)	Satisfaction	Average R^2
PCP	84.22	63	4.0	0.327
PCP + LS	88.73	49	5.0	0.376
PCP + SVD	86.45	52	5.0	0.394
PCP + LS + SVD	91.33	37	6.0	0.409

As can be observed in Table 1, PCP + LS + SCD achieved the best performance (Accuracy: 91.33%; Time: 37 s; Satisfaction score: 6.0; Average R: 0.409). Furthermore, we note that the visualization performance with feature selection is more advanced at addressing high dimensional data in PCP, which is nicely demonstrated by the corresponding experimental results (PCP + LS; PCP + SVD; original PCP).

4 Conclusion

In today's data explosion era, the emergence of a large number of high-dimensional unlabeled data makes data processing face great challenges, and so unsupervised feature selection is very necessary.

In summary, this project uses the unsupervised feature selection method to carry out the PCP axis feature rearrangement, which has significant theoretical and practical significance and is worthy of further investigation. Relevant research not only helps to solve the specific problems of high-dimensional data visualization and reveals the basic representation rules, but also can be used as a reference for the further development of data visualization and data mining technology, and promote the cross integration between different disciplines. The planned research content of this project will closely focus on the Laplacian feature selection algorithm (Laplacian Score) and the support vector machine algorithm (SVD-Entropy).

References

1. Chen, X.J., et al.: Local adaptive projection framework for feature selection of labeled and unlabeled data. IEEE Trans. Neural Netw. Learn. Syst. **29**(12), 6362–6373 (2018)
2. Krishnapuram, B., Harternink, A.J., Carin, L., Figueiredo, M.A.T.: A Bayesian approach to joint feature selection and classifier design. IEEE Trans. Pattern Anal. Mach. Intell. **26**(9), 1105–1111 (2004)
3. Liu, H., Yu, L.: Toward integrating feature selection algorithms for classification and clustering. IEEE Trans. Knowl. Data Eng. **17**(4), 491–502 (2005)
4. Wolf, L., Shashua, A.: Feature selection for unsupervised and supervised inference: the emergence of sparsity in a weight based approach. J. Mach. Learn. Res. **6**, 1855–1887 (2005)
5. Liu, S.L., Feng, L., Qiao, H.: Scatter balance: An angle-based supervised dimensionality reduction. IEEE Trans. Neural Netw. Learn. Syst. **26**(2), 277–289 (2015)
6. Yang, D.-H., Li, X., Sun, X., Wan, J.: Detecting impact factor manipulation with data mining techniques. Scientometrics **109**(3), 1989–2005 (2016). https://doi.org/10.1007/s11192-016-2144-6
7. Zhang, J., Luo, Z.M., Li, C.D., Zhou, C.G., Li, S.Z.: Manifold regularized discriminative feature selection for multi-label learning. Pattern Recogn. **95**, 136–150 (2019)
8. Huang, R., Jiang, W.D., Sun, G.L.: Manifold-based constraint Laplacian score for multi-label feature selection. Pattern Recogn. Lett. **112**, 346–352 (2018)
9. Wang, D., Nie, F.P., Huang, H.: Feature selection via global redundancy minimization. IEEE Trans. Knowl. Data Eng. **27**(10), 2743–2755 (2015)
10. He, X.F., Niyogi, P.: Locality preserving projections. Adv. Neural Inform. Process. Syst. **16** (2003)
11. Hoffman, M., Steinley, D., Brusco, M.J.: A note on using the adjusted rand index for link prediction in networks. Soc. Netw. **42**, 72–79 (2015)