A Synthesis Plot of PCP and MDS for the Exploration of High Dimensional Time Series Data

Hao Ma, Yingmei Wei[✉], and Xiaolei Du

National University of Defense Technology, Changsha, China
weiyingmei126@126.com

Abstract. Nowadays, high dimensional time series data draws more and more attention. But it is a great challenge to analyze high dimensional time series data. At present, typical methods for high dimensional time series data visualization, including ThemeRiver and Parallel Coordinates Plots, cannot reveal the distribution of the data state nor the evolution of data with time variation. And they also cannot explore the relationship between attributes of the high dimensional data and data state. In this paper, a synthetic visualization system combining Parallel Coordinates Plots and Multidimensional Scaling (MDS) is proposed for the analysis of multivariate time series data. The state distribution diagram is firstly achieved by mapping high dimensional series data onto the two-dimension space using MDS method. Distance of data points on the state distribution diagram reflects the similarity within time slices while the density indicates the state distribution of the dataset. The original dataset is then mapped on the Parallel Coordinates. Through the interaction of Parallel Coordinates and the state distribution diagram, users are able to detect evolution of time series data and explore the relationship within multiple dimensions under different states of data.

Keywords: Multivariate data · Temporal data · Parallel Coordinates · MDS · Visualization

1 Introduction

Nowadays, world is driven by the continuous collection of data from various domains: Meteorological data can guide the agricultural production, and financial data is helping the economic development. Multivariate time series data as an important part of big data is raising concerns. Multivariate data means each datum does not have only the time attribute, but also has two or more independent or related attributes. Network traffic data is a kind of typical multivariate time series data, which means great significance to the analysis of network traffic data. The network state is affected by the IP address, port, protocol, uplink byte count, downlink byte count and so on. Analysis of these data can not only help people infer the reasons for the abnormal network, helping manage the network; but also can prevent network attacks, improving the network environment.

High dimensional time series data can be understood from two aspects: high dimension and time sequence.

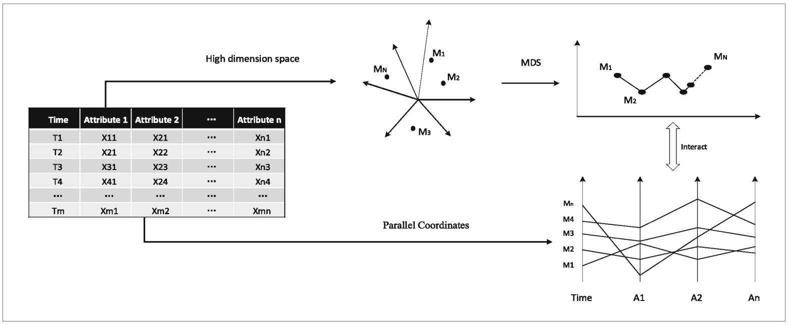
Z. Pan et al. (Eds.): Transactions on Edutainment XIII, LNCS 10092, pp. 38–45, 2017.
DOI: 10.1007/978-3-662-54395-5_4

A common method for high dimensional data visualization is to display the high dimensional data in a low dimensional space by dimension reduction. The common methods of dimension reduction are Principal Component Analysis (PCA) [1] and Multidimensional Scaling (MDS) [2]. Other common methods of high dimensional data visualization include Scatterplot Matrix [3] and Parallel Coordinates Plots [4]. Novotny [5] proposed to use the color bar to represent a specific subset, in order to highlight the performance of the subset of the changes in the property and compare the relationship between different subsets. Gennady and Natalia [6] proposed the attributes of axis are equal in the drawing of the status. At the same time, the brush technology [7] will be used to combine with the parallel coordinate drawing method and other methods of drawing.

The methods of visualization of high dimensional time series data include ThemeRiver [8, 9], Calendar Plot [10] and Spiral Clock Plot [11]. The ThemeRiver uses the horizontal axis to represent time line. Specific properties of the timeline use different colors to represent the theme. Width means the attribute value of the band. Band width varies with time as shown in the ThemeRiver. The ThemeRiver is mainly used on the presentation of temporal data classification.

It is a challenge to consider multiple dimensions and sequential simultaneously. A team of Konstanz University [12] presented Temporal MDS Plots to help analyze the patterns of high dimensional data in time. The idea of this paper is to find the patterns by combining the dimension reduction and time series analysis. Yao proposed a method which forms a three-dimensional coordinate space composed of attribute dimensions to composition range and time dimension by adding the time axis to Parallel Coordinates Plots. However, due to the three-dimensional space occlusion and deformation, the observation effect is not expected.

In this paper, a synthetic visualization system combining Parallel Coordinates Plots and Multidimensional Scaling (MDS) is proposed for the analysis of multivariate time series data. Firstly, the original data is processed and visualized. At the same time, the original data is visualized directly. The data and model which users attend can be found out through interaction (Fig. 1).



Time	Attribute 1	Attribute 2	···	Attribute n
T1	X11	X21	···	Xn1
T2	X21	X22	···	Xn2
T3	X31	X23	···	Xn3
T4	X41	X24	···	Xn4
···	···	···	···	···
Tm	Xm1	Xm2	···	Xmn

Fig. 1. High dimensional data is reduced to two-dimensional and displayed in Parallel Coordinates. The regularity of data is explored through the interaction.

The structure of this paper is organized as follows. First, we give a brief description of our approach in Sect. 2. In Sect. 3, we apply our approach to the real-world data. Conclusions and directions for future work are given in Sect. 4.

2 The Synthesis Plot of PCP and MDS

2.1 Data Description

Multivariate time series data is a sequential sampling set of high dimensional data on a time axis. From the dimension view, the multivariate time series data is the high dimensional data including the time dimension. From data analysis view, the high dimensional time series data is the data set (formula 1), which is composed of a series of discrete data with time attributes. At each time point, a number of data records are formed by sampling a number of specific entities (formula 2).

$$S = \{\mathrm{e}_1, \mathrm{e}_2, \mathrm{e}_3, \ldots \mathrm{e}_\mathrm{m}\} \tag{1}$$

m represents sampling times.e_i represents a data sample set at a single time point, which is made up of n records. Each record corresponds to a specific entity at the current time point:

$$e_i = \{\mathrm{r}_{\mathrm{i}1}, \mathrm{r}_{\mathrm{i}2}, \mathrm{r}_{\mathrm{i}3}, \ldots \mathrm{r}_{\mathrm{in}}\} \tag{2}$$

2.2 Dimension Reduction

As a common dimension reduction method, Multidimensional Scaling (MDS) is widely used in the field of statistical analysis and information visualization. MDS is a kind of multivariate statistical analysis method to analyze the similarity or difference of the objects. We can create a multi-dimensional spatial perception map using the MDS, in which the distance reflects the similarity or difference of the high dimensional data. The greater the distance is, the greater the difference is.

The similarity matrix has a great influence on the final dimension reduction result. Euclidean distance is usually used to calculate the similarity of multivariate data, but the effect of Euclidean distance in some cases is not good. For instance, sequence A: 1,1,1,10,2,3 and sequence B:1,1,1,2,10,3, these two sequences look very similar, but the Euclidean distance is very large. In order to solve this problem, people put forward Time Warping Dynamic (DTW) [13], which is used to calculate the distance between two sequences.

2.3 The Interaction

We can get the state distribution diagram through MDS, but the result of the time attribute after dimension reduction is weak. Especially when the data state is not regular, the visualization will appear a lot of occlusion which is difficult to observe. Although the Parallel Coordinates can retain details of the data, it is difficult to observe the influence of each attribute value on the final data state. Combining the two methods together in an interactive way can contribute to good results.

The MDS plot and Parallel coordinates are based on the same data, so they can easily interact by filtering the data. The process is shown in Fig. 2.

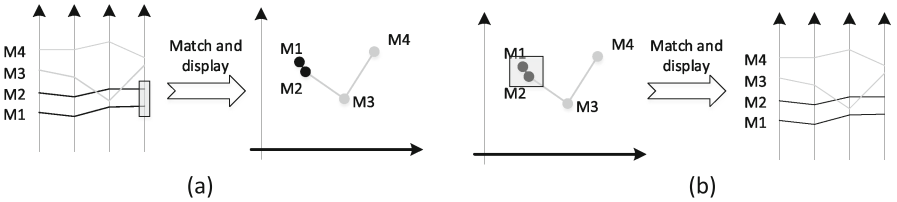


Fig. 2. (a) User selects the data of interest in the Parallel Coordinates. The matched data are highlighted in the state distribution diagram, while the unmatched data is translucent. (b) User selects the data of interest in the state distribution diagram. The matched data are highlighted in the Parallel Coordinates while the unmatched data is translucent.

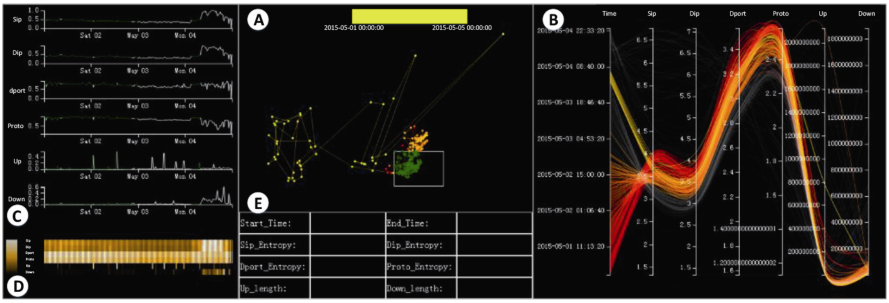
We can explore the regularity through interaction between the Parallel Coordinates and the state distribution diagram. (1) We can explore the evolution of data state over time by the brush operation of time axis in the Parallel Coordinates. (2) We can explore the relationship between the attributes of the original data and the state of data by the brush operation of the attribute axis in the Parallel Coordinates. (3) When the user selects a region of interest, the distribution of the selected state in the time axis can be observed in the Parallel Coordinates. Meanwhile, the user can observe the value of each attribute and explore the relationship of each attribute.

3 Case Study

In this paper, we use the data of China Vis 2015's challenge 2. This is a large network simulation data set which simulates a company that provides Internet networks services.

3.1 The Synthetic Visualization System

The interface of the experimental system is shown in Fig. 3. The system is divided into five parts: (1) part A is to show the results after MDS dimension reduction. The color represents time. The closer the color is to 'red', the earlier the time is. The closer the color is to 'yellow', the later the time is. The distance between the points indicates the similarity of the data state, the closer the distance is, and the more similar the data state is. The farther the distance is, the greater the difference of the data state is. (2) Part B is the effect of the parallel coordinate system of the original data. Each line represents a data. The color of the line represents the number of time. The closer the color is to 'red', the earlier the time is. The closer the color is to 'yellow', the later the time is. (3) Part C is the time sequence display for each attribute. (4) In Part D, the values of each attribute is mapped into different colors, the lighter the color is, and the greater the value is. Since the span of Part C is relatively large, observing the plurality of attributes with time is difficult. Part D span is relatively smaller and easy to observe. (5) The part E shows the selected data information. Users find the point of interest and obtain detailed information on the data by clicking operation.



Start_Time:		End_Time:	
Sip_Entropy:		Dip_Entropy:	
Dport_Entropy:		Proto_Entropy:	
Up_length:		Down_length:	

Fig. 3. The system interface. Part A shows the result of MDS dimension reduction. Part B is displays the original data through Parallel Coordinates. Part C shows the time series for each attribute. In the D part, the value of each attribute is mapped to colors, helping users to observe. Part E shows the selected data's information. (Color figure online)

3.2 Data Analysis and Exploration

State distribution diagram. As shown in Fig. 4, we can get the state distribution diagram through MDS. The color represents time. The closer the color is to 'red', the earlier the time is. The closer the color is to 'yellow', the later the time is. In this way, user can generally see the timeline. After observation, only area A whose dots are dense can be seen in the plot, which means the data in area A is relatively similar. Most points of other parts depart from each other, which shows the considerable difference of their data state.

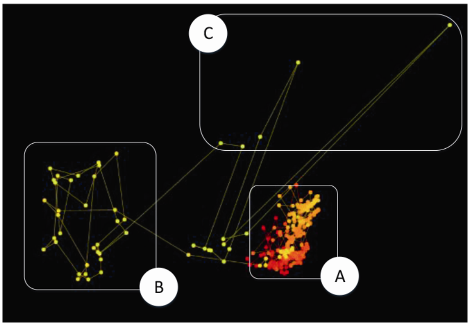


Fig. 4. State distribution diagram. The dots in area A are dense, which means the data in area A is relatively similar. The dots in area B and area C depart from each other, which shows the considerable difference of their data state. (Color figure online)

The state evolution over time. As shown in Fig. 5, through the brush operation of parallel coordinate time axis the evolution of data with time variation is explored. We can see from Fig. 5_3 that the state of data changes suddenly. We can find the time that the locating data changes is 2015-05-04 07:40:00, which is the user concerned about.

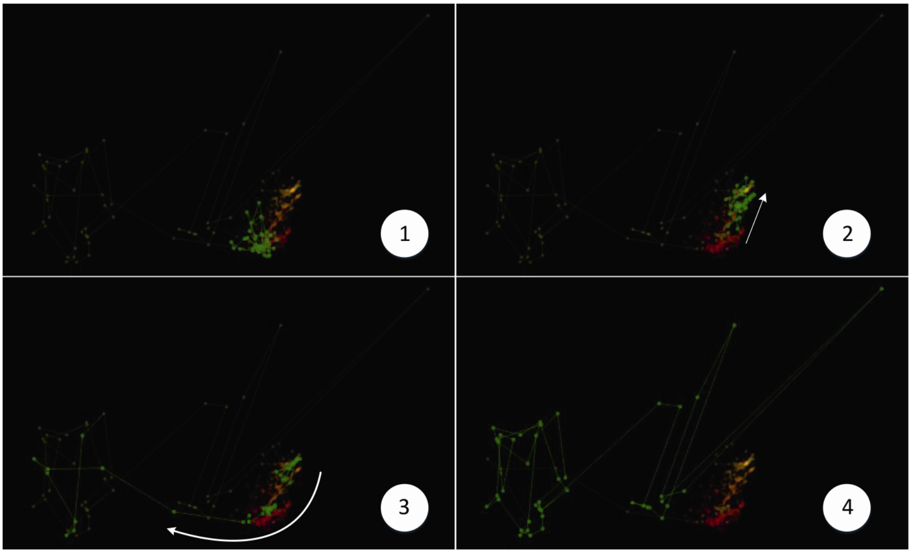


Fig. 5. The state evolution over time. Through the brush operation of parallel coordinate time axis the evolution of data with time variation is explored. We can see from 3 that the state of data changes suddenly.

Explore the regularity of data in frequent state. The frequent state's region is selected in the data view to observe the distribution of the state on the time axis and

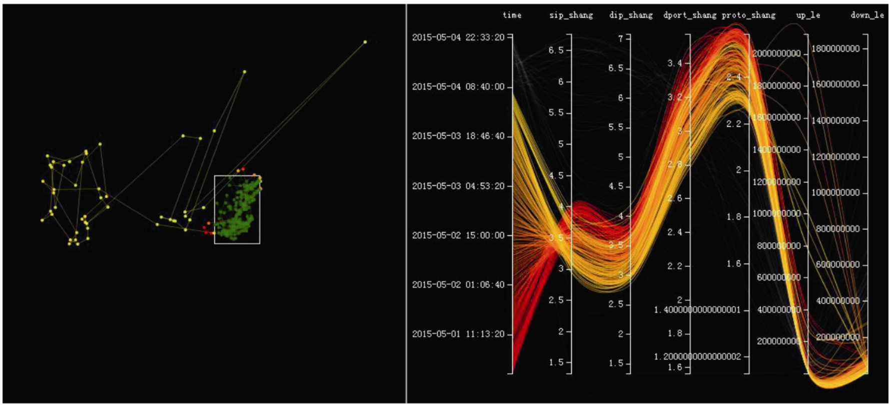


Fig. 6. Explore the regularity of data in frequent state.

explore the value of each attribute of the high dimensional data and the relationship inside in this state. As shown in Fig. 6, finally we can find that every attribute value is stable in a range under the frequent state. For example, the entropy of the source IP is stable at about 3.5. If the frequent state is the normal state, we can have a rough judgment to the later data according to the result obtained.

State distribution of specific data. The uplink bytes and downlink bytes are important indicators of the network security. A lot of network anomalies are companied by traffic anomalies. Through the brush operation of downlink bytes attribute axis, the relationship between downlink bytes and network state is explored. As shown in Fig. 7, when the number of downlink bytes are very large, it is in a state of outliers. It is likely to be abnormal which is consistent with the prior knowledge.

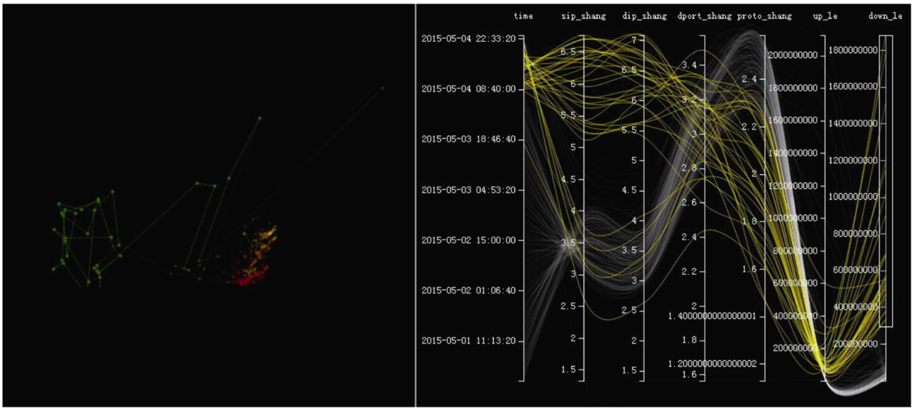


Fig. 7. State distribution of specific data. Through the brush operation of downlink bytes attribute axis, the relationship between downlink bytes and network state is explored.

4 Conclusion

In this paper, a synthetic visualization system combining Parallel Coordinates Plots and Multidimensional Scaling (MDS) is proposed for the analysis of multivariate time series data. The state distribution diagram is firstly achieved by mapping high dimensional series data onto a two-dimension space using MDS method. The distance of data points on the state distribution diagram reflects the similarity within time slices while the density indicates the state distribution of the dataset. The original dataset is then mapped on the Parallel Coordinates. Through the interaction of Parallel Coordinates and the state distribution diagram, users are able to detect the evolution of time series data and explore the relationship within multiple dimensions under different states of data.

We have applied our approach to real-world datasets. We use real data to verify the practicality of the method.

For future work there are several directions. First, the similarity matrix has a great influence on the final dimension reduction result. Finding the best similarity matrix distance formula is a challenge and typically depends on the dataset. Second, some nodes are mutual occlusion in the state distribution diagram. How to solve this problem and reduce Visual Redundancy is a challenge.

Acknowledgments. The authors wish to thank the anonymous reviews for their valuable comments. This work is supported by the National Natural Science Foundation of China (NSFC) under grant No. 61402487.

References

1. Jolliffe, I.: Principal Component Analysis. Wiley, New York (2002)
2. Catmull, E.: A tutorial on compensation tables. ACM SIGGRAPH Comput. Graph. **13**(2), 1–7 (1979). ACM
3. Tatu, A., et al.: Automated analytical methods to support visual exploration of high-dimensional data. IEEE Trans. Vis. Comput. Graph. **17**(5), 584–597 (2011)
4. Inselberg, A.: The plane with parallel coordinates. Vis. Comput. **1**(2), 69–91 (1985)
5. Rübel, O., et al.: PointCloudXplore: visual analysis of 3D gene expression data using physical views and parallel coordinates. In: The Eurographics Association, pp. 203–210 (2006)
6. Andrienko, G., Andrienko, N.: Constructing parallel coordinates plot for problem solving. In: 1st International Symposium on Smart Graphics (2001)
7. Siirtola, H.: Combining parallel coordinates with the reorderable matrix. In: Proceedings of International Conference on Coordinated and Multiple Views in Exploratory Visualization, pp. 63–74. IEEE (2003)
8. Havre, S., et al.: ThemeRiver: visualizing thematic changes in large document collections. IEEE Trans. Vis. Comput. Graph. **8**(1), 9–20 (2002)
9. Imrich, P., et al.: Interactive Poster: 3D ThemeRiver. Cg.tuwien.ac.at
10. Van Wijk, J.J., Van Selow, E.R.: Cluster and calendar based visualization of time series data. In: IEEE Symposium on Information Visualization, p. 4. IEEE Computer Society (1999)
11. Muller, W., Schumann, H.: Visualization methods for time-dependent data - an overview. In: Proceedings of the IEEE Simulation Conference, vol. 1, pp. 737–745 (2003)
12. Jackle, D., et al.: Temporal MDS plots for analysis of multivariate data. IEEE Trans. Vis. Comput. Graph. **22**(1), 141–150 (2016)
13. Bellman, R., Kalaba, R.: On adaptive control processes. IRE Trans. Autom. Control **4**(2), 1–9 (1959)