High-Dimensional Data Visualization Based on User Knowledge

Qiaolian Liu[1], Jianfei Zhao[1], Naiwang Guo[2], Ding Xiao[1], and Chuan Shi[1(✉)]

[1] School of Computer Science, Beijing University of Posts and Telecommunications, Beijing, China
shichuan@bupt.edu.cn

[2] State Grid Shanghai Electric Power Research Institute, Shanghai, China
guonw@sh.sgcc.com.cn

Abstract. Due to the curse of the dimensionality, high-dimensional data visualization has always been a difficult and hot problem in the field of visualization. Some of the existing works mainly use dimensionality reduction methods to generate latent dimensions and visualize the transformed data. However, these latent dimensions often have no good interpretability with user knowledge. Therefore, this paper introduces a high-dimensional data visualization method based on user knowledge. This method can derive dimensions aligned with user knowledge to reorganize data, then it uses scatter-pie plot matrix, an extension of scatter plot matrix, to visualize the reorganized data. This method enables users to explore the relationship between the known and unknown data as well as the relationship between the unknown data and the derived dimensions. The effectiveness of the method is validated by the experiments.

Keywords: High-dimensional data · Visualization · TSVMs

1 Introduction

In science, engineering and biology, high-dimensional data often occur. Larger and larger variables bring us great challenges in data analysis. And people can only perceive 2D and 3D space. Therefore, due to the curse of dimensionality and the limited view space, high-dimensional data analysis has always been a difficult problem in the field of visualization.

There are usually two steps to do when we visualize the high-dimensional data. The first step we usually need to do is to transform the data, and the second is visualization. The most commonly used way when transforming the data is to transform the original data variables in a linear or non-linear way which is usually called Dimensionality Reduction (DR) [1], such as the well known Principal Components Analysis (PCA), Multidimensional Scaling (MDS) and Locally Linear Embedding (LLE). Those methods usually use the statistical properties and create latent dimensions. However, these latent dimensions are usually difficult to interpret with user knowledge. For better understanding and

Y. Tan and Y. Shi (Eds.): DMBD 2016, LNCS 9714, pp. 321–329, 2016.
DOI: 10.1007/978-3-319-40973-3_32

interpreting, a few works take the importance of user knowledge in exploring high-dimensional data into account. For example, an approach introduced by Gleiher [2] can generate projection functions meaningful to users. But it only considered the known data. Since user knowledge is limited, it is difficult for us to know all the observed data. And the widely used visualization techniques, such as the scatterplot matrix and parallel coordinates, cannot reflect user knowledge over different aspects of data.

To solve the limitations mentioned above, here we introduce an approach which integrates user limited knowledge to visualize and explore high-dimensional data. Our approach can derive dimensions that align with user knowledge and reorganize data. By providing a visualization method scatter-pie plot matrix, users can explore the reorganized data and discover new knowledge. We describe the main tasks of the process when visualizing and exploring the high-dimensional data with user limited knowledge. Then we use a dataset to demonstrate how our approach works. To make a summary, our method has the following features: (1) Derive new dimensions that align with user knowledge and reorganize the data, including the known and unknown data; (2) Discover the relationship between the known and unknown data as well as the relationship between the unknown data and the derived dimensions through scatter-pie plot matrix.

2 Methods

2.1 Motivation

We aim to visualize and explore high-dimensional data. However, for the curse of dimensionality and the limited view space, it is difficult for users to visualize and explore it. The traditional approaches usually mapped high-dimensional data into low-dimensional space by creating new latent dimensions using statistical methods. And these approaches usually called Dimensionality Reduction (DR). However, they usually ignore the importance of user knowledge in exploring data. Therefore, we expect to integrate user knowledge to derive dimensions for better understanding and interpreting the data. Although a few works considered about user prior knowledge, users usually cannot know all the observed data and they may be more interested in the unknown data in addition to the known data. Hence, we expect to take user known and unknown data into account when deriving dimensions that align with user knowledge. Then we can use the derived dimensions which align with user knowledge to reorganize the data, and by this way the view space will be saved and users can have a better understanding of the derived dimensions. Because of visualization is a more intuitive way to help user understand the data, it is an essential way for users to explore the data. Since users are more interested in the unknown data, we expect to visualize the data that can reflect the relationship between the known and unknown data as well as the relationship between unknown data and the derived dimensions. To solve the problem mentioned above, we derive two main tasks: *(1) Derive dimensions that align with user knowledge.* Considering about the user limited

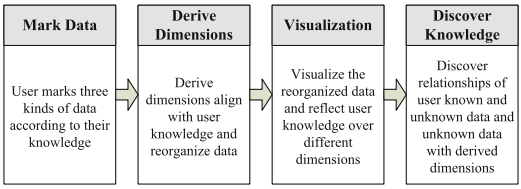


Fig. 1. Illustration of the pipeline when visual exploration of high-dimensional data.

prior knowledge, the unknown data are also taken into account. *(2) Visualize the reorganized data.* Since the data include both the known and unknown data, the view should reflect user knowledge over the data.

2.2 Main Framework

To solve the problem, our approach is designed to support the above tasks. Figure 1 is given as an illustration of the pipeline: (1) mark data, (2) derive dimensions, (3) visualization, (4) discover knowledge. To make it more concrete, here we provide an example to explain the workflow. For instance, the user who has a dataset of city livability data (see Table 1) can mark cities (Step 1) according to their knowledge. Cities belong to Europe, such as Rome, will be marked with Europe. Cities not belong to Europe, such as New York and Beijing, will be marked with non-Europe. However, cities such as Athens and Bogota are not familiar to user and user does not know whether they are belonged to Europe, so they are marked with nothing. Then our method which takes the known and unknown data into account enables user to derive dimension Europe-ness (Step 2) by integrating annotations user marked. Obviously, cities in Europe should be much more Europe-ness than other cities. And the derived dimensions are used to reorganize the data. When visualizing the reorganized data (Step 3), our design can reflect user knowledge over the different aspects of data. Finally, by exploring the reorganized data, user can discover (Step 4) the relationship between the known and unknown data as well as the relationship between the unknown data and the derived dimensions.

Next, we will introduce our method in two parts: firstly, derive dimensions that align with user knowledge to reorganize data. Secondly, visualize the reorganized data and explore the reorganized data to discover new knowledge.

Table 1. Sample data in city livability dataset

City	New York	Beijing	Mexico City	Rome	Paris	Tokyo	HongKong	Athens	Bogota	Cairo
Education indicators	1	3	2	1	1	1	1	2	3	4
Healthcare indicators	1	4	3	1	1	1	2	1	3	3
...	...	...	...	...	...	...	...	...	...	...

2.3 Derive Dimensions Aligned with User Knowledge

In order to make our method easy to understand, here we still use the example mentioned above, but our method can also be used in other dataset.

According to the user marks on the data, we divide the data into three kinds of types. One is user known European cities, one is user known non-European cities, and the last one is user unknown cities. And the value of European cities should be higher than those non-European cities. This reminds us the semi-supervised binary prediction problem. Therefore, we use the convention $\mathbf{y}$ denoted as the labeled vector, and vector $\mathbf{x}$ denoted as data objects. And data points labeled 1 means they are positive elements (e.g. European cities), -1 means they are negative elements (e.g. non-European cities), and 0 means they are unlabeled data (e.g. user unknown cities). Then we use projection function $f(\mathbf{x})$ denoted as the derived dimension and it aligns with user knowledge that European cities should have higher values than non-European cities.

We seek the projection function $f(\mathbf{x})$, and we usually think about the linear function $f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b$. Since hyper-plane $\mathbf{w} \cdot \mathbf{x} + b = 0$ can separate the two classes and the distance $|\mathbf{w} \cdot \mathbf{x} + b|$ can present the correctness and certainty of data belonged to the class. And the positive sign of the label makes data which are far from the hyper-plane have high values that European cities are much more Europe-ness than non-European cities. Therefore, we use projection function $f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b$ denoted as the derived dimension which meets that data in positive set have higher values than data in negative set.

Since Transductive Support Vector Machines (TSVMs) [3] has the following features, here we use TSVMs to solve the problem mentioned above. Firstly, it can generate projection function $f(\mathbf{x})$ aligned with user knowledge that data in positive set have higher values than data in negative set. Secondly, there are unlabeled data, such as the cities user unknown. Thirdly, since users expect to explore data they observed, they do not care about the unseen data. TSVMs takes the unlabeled data as a special test set, and focuses on the seen data and makes a transductive learning.

TSVMs is a transductive learning algorithm which introduced by Joachims. This method takes the unlabeled data as a special test set, and reduces classification error as far as possible. The basic ideas of the algorithm are shown as follows: given a set of sample data, including labeled data $(\boldsymbol{x_1}, y_1), \ldots, (\boldsymbol{x_n}, y_n)$, and unlabeled data $\boldsymbol{x_1^*}, \boldsymbol{x_2^*}, \ldots, \boldsymbol{x_k^*}$. Under the condition of linearly separable case, the learning process is shown in (1). This optimization problem can be solved by minimizing the L2 norm of $\mathbf{w}$ under the constraints and find the hyper-plane $<\mathbf{w}, b>$ separates both training and test data with maximum margin and the label $y_1^*, \ldots, y_k^*$ of the test data.

$$\begin{aligned} &\min_{y_1^*,\ldots,y_n^*,\boldsymbol{w},b} \tfrac{1}{2}\|w\|^2 \\ &s.t.\ \forall_{i=1}^{n} : y_i[\boldsymbol{w} \cdot \boldsymbol{x}_i + b] \geq 1 \ and \ \forall_{j=1}^{k} : y_j^*[\boldsymbol{w} \cdot \boldsymbol{x}_j^* + b] \geq 1 \end{aligned} \tag{1}$$

And under the conditions of non-linearly separable case, the learning process is shown in (2). It introduces slack variables ξ_i to allow errors occur and seeks the

maximum margin and makes the errors minimum. C and C^* can be set to trade off margin size.

$$\begin{array}{c} \min_{y_1^*,\ldots,y_n^*,\boldsymbol{w},b,\xi_1,\ldots,\xi_n,\xi_1^*,\xi_k^*} \frac{1}{2}\|w\|^2 + C\sum_{i=0}^{n}\xi_i + C^*\sum_{j=0}^{k}\xi_j^* \\ s.t.\ \forall_{i=1}^{n} : y_i[\boldsymbol{w}\cdot\boldsymbol{x}_i + b] \geq 1 - \xi_i \ and\ \forall_{i=1}^{n} : \xi_i > 0 \\ \forall_{j=1}^{k} : y_i^*[\boldsymbol{w}\cdot\boldsymbol{x}_j^* + b] \geq 1 - \xi_j^* \ and\ \forall_{j=1}^{k} : \xi_j^* > 0 \end{array} \tag{2}$$

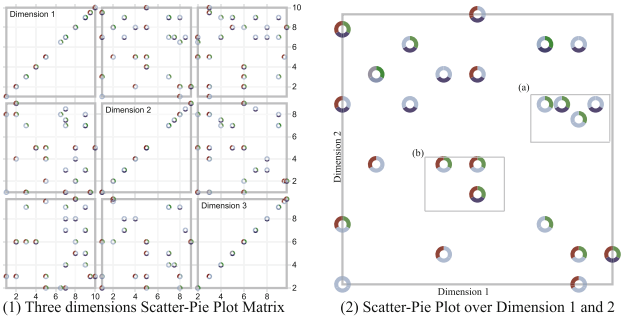
By solving the above optimization problem, we can get the projection function $f(\mathbf{x}) = \mathbf{w}\cdot\mathbf{x} + b$, which corresponding to the derived dimension that align with user knowledge. By this token, several dimensions can be derived to reorganize the data. In this paper, we use SVM^{light} [4] to solve the optimization problem.

2.4 Design of Scatter-Pie Plot Matrix

We expect our view can reflect user knowledge over different aspects of data, and explore the relationship between user known and unknown data as well as the relationship between the unknown data and new dimensions we derived. So scatter-pie plot matrix can meet our needs better. Scatter-pie plot matrix is an extension of scatter plot matrix and the pie chart view. It can present all the combinations of two dimensions.

Fig. 2. Illustration of Scatter-Pie when derive three dimensions. Each Scatter-Pie is divided into three parts to indicate user knowledge over these three dimensions. Special colors (red, green, blue) are filled when user knows this data object over this dimension, and gray is filled when user does not know this data object over this dimension. (Color figure online)

Scatter-Pie. Each scatter-pie in scatter-pie plot matrix represents a data object (see Fig. 2). Various parts of the scatter-pie stand for user knowledge over different dimensions, each part evenly and arranged in clockwise. Colors are used to encode user knowledge over different dimensions. Scatter-pie plot reflects the user knowledge over the different aspects of the data object. Before users derive one dimension, they will first mark their known data and unknown data. So if user knows the data object, special color will be used to show that user know this data object over this dimension. Figure 2 shows the scatter-pie plot of two data objects. The user defines three different dimensions, and uses three different colors to encode their knowledge over different dimensions. Users use this kind of color to fill the dimension when he/she knows the data object over the dimension. Red, green and blue are used in Fig. 2. If the user knows nothing about the data object over this dimension, gray is used to fill this part of the



(1) Three dimensions Scatter-Pie Plot Matrix (2) Scatter-Pie Plot over Dimension 1 and 2

Fig. 3. (1) Three dimensions Scatter-Pie Plot Matrix. (2) One panel in Scatter-Pie Plot Matrix. (Color figure online)

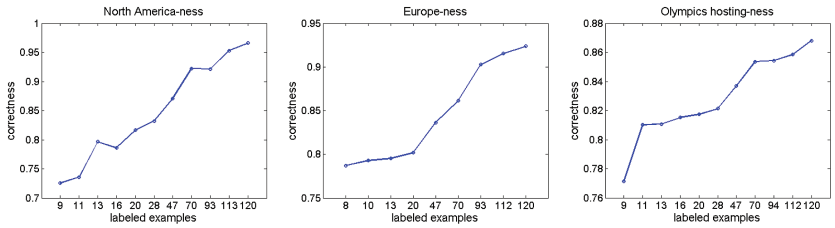


Fig. 4. The impact of labeled examples on correctness when deriving dimension North America-ness, Europe-ness and Olympics hosting-ness.

scatter-pie. This way of view reflects user knowledge over different aspects of data objects.

Scatter-Pie Plot Matrix. Figure 3(1) shows a scatter-pie plot matrix of three dimensions. Each panel in scatter-pie plot matrix is identified by pair-wise dimensions. And each data object is represented by scatter-pie. Figure 3(2) shows one panel in scatter-pie plot matrix. Data points in (a) which are very close to each other may have very similar characteristic over these two dimensions. And data points in (a) are located in the upper right of data points in (b), so they have high values over dimension 1. This way of view provides a lot of help for users to analyze the relationship between user known and unknown data as well as the relationship between the derived dimensions and the data.

3 Case Study

3.1 Dataset

The dataset we use in this paper is city livability data [5] which contains 140 cities from different countries and regions. Each city represents by 45 dimensions. Dimensions are measurements of different aspects of city, such as education indicators, healthcare indicators, crime rating. The sample data are shown in Table 1.

All the data are discrete, and up to 5. As shown in Table 1, the prevalence of violent crime in Mexico City is higher than New York and Beijing. And the healthcare indicators of Beijing are higher than other cities.

3.2 Evaluation

We expect to derive dimensions which align with user knowledge. Hence, we use correctness to ensure the effectiveness of our method. Correctness is defined as shown below: Correctness [2]: the elements in the positive set (denoted as P) should have higher values than elements in the negative set (denoted as N).

$$\forall_{i \in P} \forall_{j \in N} f(i) > f(j) \tag{3}$$

Therefore, cities belonged to positive set (e.g. European cities) are much more Europe-ness than cities belonged to negative set (e.g. non-European cities). Therefore, correctness is calculated as the percentage of times positive cases greater than the negative cases off the total comparison times.

The impact of labeled examples on correctness is shown as Fig. 4. When deriving dimension North America-ness, Europe-ness and Olympics hosting-ness, correctness rate is gradually increasing along with the increasing labeled examples. When the labeled examples increased from about 9 to 120, the correctness rate increased from about 0.7 to 0.9. And in the case of less labeled data (about 9 labeled data), the correctness rate is still above 0.7. It shows that our method can meet the user knowledge to a certain extent when deriving dimensions.

3.3 Visualization and Exploration

On the first step users mark three kinds of data, positive data (e.g. European cities) as 1, negative data (e.g. non-European cities) as -1, and unknown data as 0. And Europe-ness is derived by using our method. The other two dimensions of North America-ness and Olympics hosting-ness are derived by the same token. Then scatter-pie plot matrix is used to visualize the data over these three dimensions. And finally, by exploring the data, users can discover new knowledge through the view.

Exploring relationship between the derived dimensions and user known and unknown data. The scatter-pie plot over dimension Europe-ness and North America-ness is shown as Fig. 5. From Fig. 5(a), we can see that the value of city New York, Seattle, Houston and Washington which locate in continent North America is higher than other cities over North America-ness. And as shown in Fig. 5(b), the value of Berlin and Rome which locate in continent Europe is higher than other cities over Europe-ness. And this aligns with user knowledge. Milan and Paris are cites user unknown over the aspect whether they are European cities (shown as gray in the Scatter-Pie). And they are also very Europe-ness.

Exploring relationship between user known and unknown data. In Fig. 5(b), Milan, Berlin, Rome and Paris are very close to each other, so they are quite similar over the two dimensions. However, Berlin and Rome are known to user

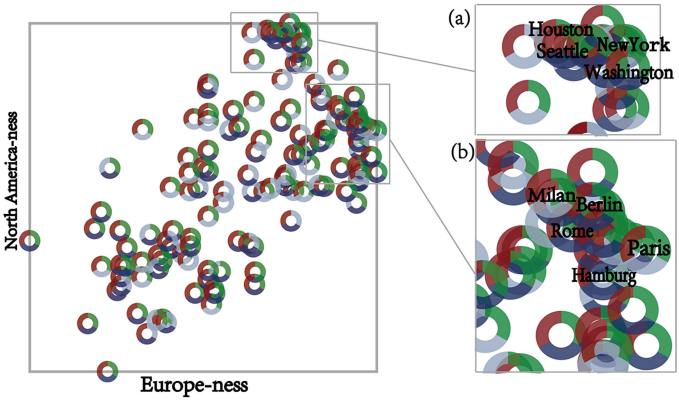


Fig. 5. Scatter-Pie Plot over dimension North America-ness and Europe-ness. Part of Scatter-Pie filled by color green represents user known North American or non-North American cities. Part of Scatter-Pie filled by color blue represents user known European or non-European cities. Part of Scatter-Pie filled by red represents user known Olympics hosting or non-Olympics hosting cities. Part of Scatter-Pie filled by color gray represents user unknown cities over this aspect of the data. (Color figure online)

over these two dimensions (shown as blue and green in the scatter-pie) and they are European cities and quite Europe-ness. Therefore, we think that Milan and Paris are also European cities to a large extent. And the facts show that they are European cities. Hence, our method provides a direction for users to know the two cities.

4 Conclusion

This paper presents an approach to explore high-dimensional data with user knowledge. It can derive dimensions that align with user knowledge to reorganize data and use scatter-pie plot matrix to visualize data. It enables users to discover the relationship between user known and unknown data as well as the relationship between user unknown data and the derived dimensions. Users are more concerned about their own oriented knowledge discovery process and do not have to waste more time to observe the characteristics of data they are not interested in. Because the users knowledge over this data may have certain errors, we hope considering the uncertainty of user knowledge and errors and analyze the effect of the error data on the data distribution in the future work. We temporarily do not provide a friendly interaction and unified operation mode to help users better analyze the data, but we hope to develop one. And friendly interactions such as zooming and filtering will be added to overcome the overlap of scatter-pies in the future work.

Acknowledgments. This work is supported in part by the National High-tech R&D Program (863 Program 2015AA050203), the National Natural Science Foundation of China (No. 61375058), and Co-construction Project of Beijing Municipal Commission of Education.

References

1. Wang, J.: Geometric Structure of High-Dimensional Data and Dimensionality Reduction. Springer, Heidelberg (2012)
2. Gleicher, M.: Explainers: expert explorations with crafted projections. IEEE Trans. Visual Comput. Graphics **19**(12), 2042–2051 (2013)
3. Joachims, T.: Transductive inference for text classification using support vector machines. In: ICML, pp. 200–209 (1999)
4. SVMlight. http://svmlight.joachims.org
5. City livability data. Buzzdata. Best City Contest. http://graphics.cs.wisc.edu/Vis/Explainers/data.html