EuroVis 2013 Workshop on Visual Analytics using Multidimensional Projections (2013)
M. Aupetit and L. van der Maaten (Editors)

Short Papers

Interactive Visualization and Feature Transformation for Multidimensional Data Projection

D. Pérez [1], L. Zhang [2], M. Schaefer[2], T. Schreck[2], D. Keim [2] and I. Díaz[1]

[1]University of Oviedo, Spain
[2]University of Konstanz, Germany

Abstract

Projecting multidimensional data to a lower-dimensional visual display as a scatter-plot-like visualization is a common approach for analyzing multidimensional data. Many dimension reduction techniques exist for performing such a task, but the quality of projections varies in terms of both preserving the original data structure and avoiding cluttered visual displays. In this paper, we propose an interactive feature transformation approach that allows the analyst to monitor and improve the projection quality by transforming feature space and assessing/comparing the quality of different projection results. The method integrates feature selection and transformation as well as a variety of projection quality measures to help analyst generate uncluttered projections that preserve the structural properties of the data. These projections enhance the visual analysis process and provide a better understanding of data.

Categories and Subject Descriptors (according to ACM CCS): I.5.2 [Pattern Recognition]: Design Methodology—Feature evaluation and selection, Pattern analysis

1. Introduction

Projection-based data analysis and visualization is widely used for identifying patterns in multidimensional data. The idea is to project each data item (object) as a point to a two or three dimensional visual display in such a way that similar items are close to each other and dissimilar ones are far apart, result in a scatterplot-like visualization where structures and patterns can be analyzed. The projection is usually achieved by a *Dimension Reduction* (DR) technique that tries to best approximate the distance (similarity) between items in high-dimensional data space to the low dimensional visual display. A large number of DR methods exist [LV07], and one critical part of the technique is the distance measure. Multidimensional data often contains dimensions that are irrelevant to the analysis task, values in these dimensions introduce noise to the distance measure and obscure real distances between objects. Using such inaccurate distance measures may hide the real structure of the data as well as meaningful patterns. To reduce the noise in data, a number of interactive dimension selection and feature transformation techniques have been proposed [JJ09, SZS*13]. These approaches either filter out the noise by selecting relevant dimensions manually or automatically, or reduce the influence of noisy dimensions via feature transformation.

The requirements for evaluating the resulting projections lead to the definition of *quality measures* that help the analyst to understand how well the distances are approximated in the projection. Apart from measures that take into account structural preservation [Sam69, LV08], a set of *visual* quality measures has also been developed [SZS*13, BTK11]. While the techniques and measures provide means of generating meaningful embeddings of multi-dimensional data and assess their quality from different perspectives, existing projection approaches lack the flexibility of integrating interactive visualization and feature transformation mechanism to steer the projection process and improve its quality. Recent advances in the field include interactive approaches [JZF*09, CLKP10] that integrate the human expert in the analysis process and help to understand multidimensional data, as well as an improvement of class separation in projections by means of transforming feature space [SZS*13]. The work reported in this paper advances the above mentioned approaches by combining the strength of both interactive user feedback and feature transformation for generating better quality visual embeddings of multidimensional data.

DOI: 10.2312/PE.VAMP.VAMP2013.021-025

The main contribution of this paper is a novel visual analytics approach that combines interactive visualization, dimension selection, feature transformation, and quality evaluation for improving the quality of multidimensional data projection. The reminder of this paper is organized as follows. In Section 2 we discuss related work, in Section 3 we explain the details of the proposed approach, in Section 4 we demonstrate the effectiveness of the method with real data, and finally, in Section 5 we draw conclusions and discuss future work.

2. Related work

2.1. Feature transformations and interactive analysis

Feature selection and transformations have been developed to improve performance of many applications in several research fields [BL97, GE03]. A recent approach [SZS*13] transforms the feature space by extending specific feature of selected dimensions. The result can be applied to improve group separation and reduce visual cluttering in the final embedding.

DR techniques estimate the underlying structure and reveal relationships in multidimensional data. However, with the increasing size and complexity of data, it becomes more difficult to generate meaningful projections in a fully automatic way. This leads to the development of *interactive multidimensional data projection* techniques that facilitate interactive analysis by integrating the analyst's knowledge about the data as well as the knowledge gained during the learning process. Examples include the iPCA approach [JZF*09] that provides coordinated views for interactive analysis of projections computed by PCA method, the iVisClassifier system [CLKP10] improves data exploration based on a supervised DR technique (LDA). Moreover, the DimStiller framework [IMI*10] analyzes dimension reduction techniques with interactive controls that guide the user during analysis process and Dis-Function [BLBC12] provides an interactive visualization to define a distance function. A comparison of features sets are determined in [BvLBS11], and an interactive exploration can be made for the selection of the suitable data descriptors.

The above mentioned techniques show that a rich body of research exists on multidimensional data visualization. However, integrating human knowledge to the analysis loop to improve the quality of visual embedding remains a challenge.

2.2. Quality Metrics

Despite the large number of DR techniques that have been developed, the question of quality assessment of a given projection has remained mostly unanswered until recent years [BTK11].

The first measures to assess the quality of a projection are the so called *stress* and *strain* measure [Sam69, Kru69]. These measures come from objective functions of nonlinear DR techniques, and assess the quality of structural preservation with the differences of the Euclidean distances between pairwise objects in a low-dimensional embedding approximate and the corresponding distances in high-dimensional data space.

While *strain* and *stress* measures analyze the preservation of global structure of data, the *trustworthiness* and *continuity* measure [VK01] and the *K-ary neighborhoods* measure [LV08] assess the quality of a projection in a broader applicability, taking into consideration also the small neighborhood preservation. In the case of labeled data, the classification error is a typical choice, see for instance [SR03] and other references in [VK07]. The integration of classification error measures in the DR technique leads to better group separation in the final embedding.

Apart from the *structural preservation quality measures* mentioned above, a set of *visual quality measures* has also been developed. Examples include *Histogram Density Measure* that ranks scatter plot visualizations, and the *Class Density Measure* that assess class separation of a given projection [TAE*09]. Moreover, the *overlap measures*, defined in [SZS*13], compute the overlap area between groups and overlap object density in a multidimensional data projection.

3. Method

In this paper, we propose a multidimensional data projection framework that combines the strength of the feature transformation approach [SZS*13], the interactive parameter setting and visualization to help analyst achieve uncluttered projections. The main workflow of the framework is shown in Figure 1. First of all, given a multidimensional dataset, with labels that define the containing classes. An initial projection is generated by a selected DR technique. The interactive

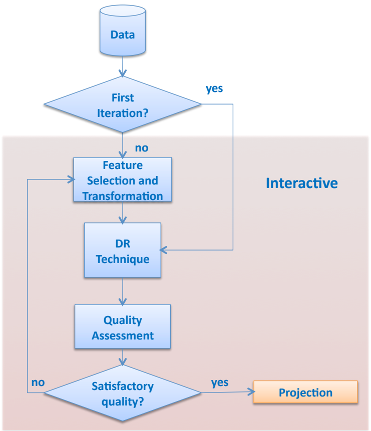
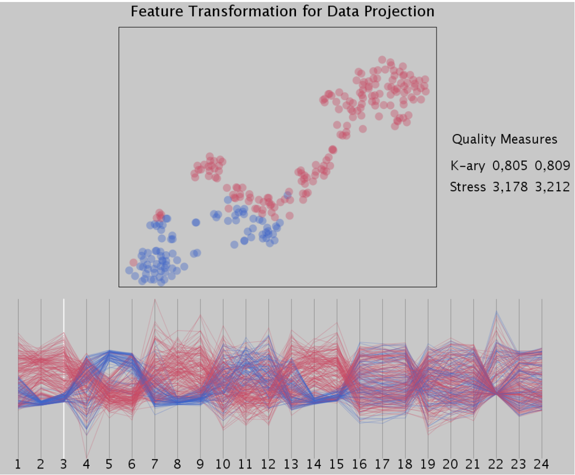


Figure 1: *Workflow of the method*

visualization panel allows the analysts to select dimensions for feature extension based on the data distribution and their knowledge about the data. After that, the system will transform the data by extending the mean values of each class for each variable selected. The DR technique is applied again to the transformed data for generating a new projection. The quality of both projections will be evaluated with quality measures and can be compared to select the one that has better quality. The analysts can iteratively repeat the process until a satisfactory projection is achieved.

3.1. Interactive Visualization for Dimension Selection

Feature selection can be performed with diverse criteria. In an automatic way, it can be used the *range* of data values over a dimension using the labels with categorical information. An interactive approach can be performed by parallel coordinates visualization which shows global data distribution over all dimensions with different color for each class. This view can help the analyst identify dimensions that provide clear distinctions between different classes. For example in Figure 2, from the parallel coordinates visualization it is not difficult to find out that in the 5th dimension, data items that belong to the same class have similar values and data items that belong to different classes are usually different. Such visual patterns often help the analyst to identify "distinctive" dimensions in multidimensional data. The result shows that transforming certain features relates to these distinctive dimensions often helps achieving better quality projection [SZS*13].



Quality Measures
K-ary 0,805 0,809
Stress 3,178 3,212

Figure 2: *Screenshot of the prototype tool*

Due to the scalability of the parallel coordinates visualization, a previous process should be considered to generate features for complex datasets.

3.2. Feature Transformation

The basic idea of the feature space transformation is to extend the selected features by adding the mean values of each class. Considering multidimensional dataset as a matrix $\mathbf{D}$ where rows are data items and columns are features, and class labels $\mathbf{c}$ are given to the class of the i-th row.

$$\mathbf{D} = \left[d_{ij}\right] \in \mathbb{R}^{m \times n} \qquad \mathbf{c} = [c_i] \in \mathbb{N}^m \tag{1}$$

With $i = 1, \ldots, m$ and $j = 1, \ldots, n$, being m the number of feature vectors and n the number of features. If one feature f is selected, the extended data table $\mathbf{D'}$ is defined as follows,

$$\mathbf{D'} = \left[d_{ij} \mid m^f_{c_i}\right] \in \mathbb{R}^{m \times (n+1)} \tag{2}$$

being $m^f_{c_i}$ the mean value of all the items corresponding to the class label c_i in the feature f.

The maximum number of extended features could be the whole set of variables. Although using this selection the result leads to a clear group separation, the similarity preservation between groups objects is damaged. Besides this simple extension strategy, a feature space can be transformed in many different ways. For example, *median* or *mode* could be applied instead of the mean value.

4. Experiments and Results

In this section the proposed approach is shown on multidimensional data with class labels from a real case. The data consists of measures of electrical and environmental variables, collected during a whole year at one university building. The task is the identification of different types of daily consumption patterns in that building. The variables that were used are: *voltage, current, apparent power, power factor, neutral current, temperature, humidity and solar radiation*. The day is divided into three shifts of eight hours each, and characterized with the average value of each shift for each variable, so that each item represents a day. Therefore the data matrix is composed by the days (items with missing values were removed) and 24 features (8 variables x 3 shifts). The used label has two classes depending on whether it is working day or holiday such as weekends.

To validate this approach, a prototype tool has been developed (see Figure 2) which displays both the projection and the parallel coordinates views with color representing labels. The parallel coordinates view helps to decide the best choices over all features. In this case, the automatic feature selection corresponds to the maximum range between mean values for each class of the whole set of attributes. Although this selection recommends using feature five, the extension of the dimension eight obtains a similar map with better quality measures.

The projections of the original and transformed data are computed with the same dimensionality reduction technique. The techniques used were t-SNE method [vdMH08], that

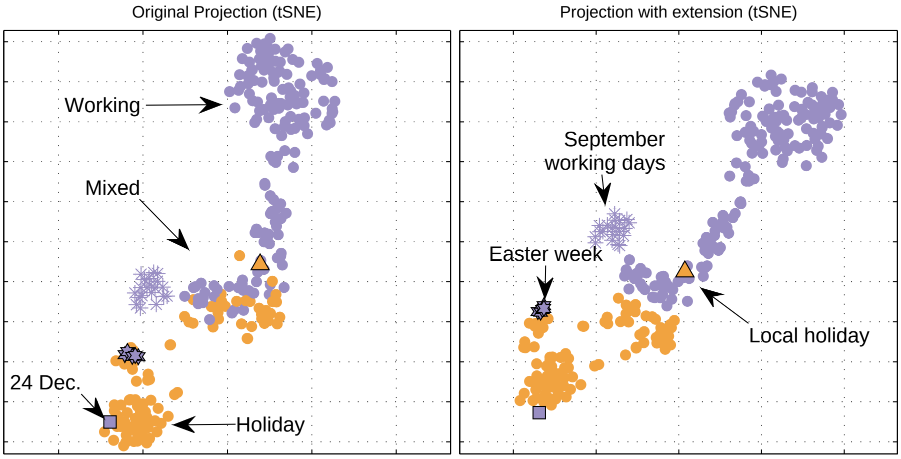


Figure 3: *Original (Left) and extended (Right) projections of daily consumption for one building with t-SNE technique. Color represents class labels (holiday/working day) and shape refers to highlighted items.*

is an effective unsupervised technique for visualizing data, and a supervised technique, *Maximally Collapsing Metric Learning* (MCML) [GR06], in order to use the label information available for computing the embedding. Notice that the transformation is independent of the DR technique chosen. The transformation performed was the extension of the selected dimension with the mean values for each class.

In the projection with the original feature vectors two daily patterns, of high and low consumption, are easily identified, clearly related to working day and holiday, respectively. But there is a third pattern in the middle, with both types of days mixed (see Figure 3, left), which is not easy to identify. The projection with the extension reveals similar daily patterns with a clearer class separation that improves the recognition of the label information in that mixed area (see Figure 3, right). For example, it is easy to distinguish, in the extended projection, a point of a local holiday, that stays close to the working days, revealing similar consumption these days in the building.

Finally the performance of the projections is evaluated by the quality measures previously described. The stress measure is referred to the Sammon's error [Sam69], k-ary neighborhood can be found in [LV08], and the overlap measures are formally defined in [SZS*13]. The values of these measures used are described in Table 1 for this example. These evaluation measures show an enhancement of the projection quality in the extended case.

5. Conclusions

In this paper we propose an interactive visualization framework for improving existing data projections. The method transforms multidimensional data by extending selected features from original data, introducing the human into the analytical loop and utilizing their perception power and domain knowledge. A case with real datasets was conducted to test the effective of the approach. With both supervised and unsupervised DR techniques, through interactive dimension selection and feature transformation, we can achieve projections with improved quality. These projections provide efficiency to pattern recognition, fast identification of class labels and understanding of data. The improvement of the projection is independent of the DR technique that are chosen to perform the projection, having the same scalability limitations that the technique itself.

Table 1: *Assessment measures for the projections*

t-SNE				
Feat. Ext.	**k-ary**	**Stress**	**Overlap area**	**Overlap density**
None	0.80	3.04	0.024	$7 \cdot 10^{-3}$
5	**0.81**	3.02	0.029	$1 \cdot 10^{-3}$
8	**0.81**	**2.95**	$\mathbf{6 \cdot 10^{-5}}$	$\mathbf{9 \cdot 10^{-4}}$
MCML				
Feat. Ext.	**k-ary**	**Stress**	**Overlap area**	**Overlap density**
None	0.6462	0.3953	0.063	$6 \cdot 10^{-4}$
5	0.6836	0.3477	**0**	**0**
8	**0.6838**	**0.3474**	**0**	**0**

As future work we would like to explore more visualization techniques for assisting feature selections, new transformation strategies for noise elimination, and wider range of quality measures for evaluating the projections.

Acknowledgments

This work has been financed by the Spanish Ministry of Science and Education and FEDER funds under grants DPI2009-13398-C02-01/02

References

[BL97] BLUM A., LANGLEY P.: Selection of relevant features and examples in machine learning. *Artificial intelligence 97*, 1 (1997), 245–271. 2

[BLBC12] BROWN E., LIU J., BRODLEY C., CHANG R.: Disfunction: Learning distance functions interactively. In *Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on* (2012), pp. 83–92. 2

[BTK11] BERTINI E., TATU A., KEIM D.: Quality metrics in high-dimensional data visualization: An overview and systematization. *Proceedings of the IEEE Symposium on IEEE Information Visualization (InfoVis) 17* (2011), 2203–2212. 1, 2

[BvLBS11] BREMM S., VON LANDESBERGER T., BERNARD J., SCHRECK T.: Assisted descriptor selection based on visual comparative data analysis. In *Computer Graphics Forum* (2011), vol. 30, Wiley Online Library, pp. 891–900. 2

[CLKP10] CHOO J., LEE H., KIHM J., PARK H.: ivisclassifier: An interactive visual analytics system for classification based on supervised dimension reduction. In *Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium on* (oct. 2010), pp. 27 –34. 1, 2

[GE03] GUYON I., ELISSEEFF A.: An introduction to variable and feature selection. *The Journal of Machine Learning Research 3* (2003), 1157–1182. 2

[GR06] GLOBERSON A., ROWEIS S.: Metric learning by collapsing classes. *Advances in neural information processing systems 18* (2006), 451. 4

[IMI*10] INGRAM S., MUNZNER T., IRVINE V., TORY M., BERGNER S., MÖLLER T.: Dimstiller: Workflows for dimensional analysis and reduction. In *Proc. IEEE Conf. Visual Analytics Science and Technology (VAST)* (2010), vol. 1, Citeseer. 2

[JJ09] JOHANSSON S., JOHANSSON J.: Interactive dimensionality reduction through user-defined combinations of quality metrics. *Visualization and Computer Graphics, IEEE Transactions on 15*, 6 (2009), 993–1000. 1

[JZF*09] JEONG D. H., ZIEMKIEWICZ C., FISHER B., RIBARSKY W., CHANG R.: iPCA: an interactive system for PCA-based visual analytics. *Computer Graphics Forum 28*, 3 (June 2009), 767–774. 1, 2

[Kru69] KRUSKAL J.: Toward a practical method which helps uncover the structure of a set of multivariate observations by finding the linear transformation which optimizes a new 'index of condensation'. In *Statistical Computation*, Milton R., Nelder J., (Eds.). Academic Press, New York, 1969, pp. 427–440. 2

[LV07] LEE J., VERLEYSEN M.: *Nonlinear dimensionality reduction*. Springer, 2007. 1

[LV08] LEE J., VERLEYSEN M.: Quality assessment of nonlinear dimensionality reduction based on k-ary neighborhoods. In *JMLR Workshop and Conference Proceedings (New challenges for feature selection in data mining and knowledge discovery)*, Saeys Y., Liu H., Inza I., Wehenkel L., Van de Peer Y., (Eds.), vol. 4. Sept. 2008, pp. 21–35. 1, 2, 4

[Sam69] SAMMON J. W.: A nonlinear mapping for data structure analysis. *IEEE Trans. Comput. 18*, 5 (May 1969), 401–409. 1, 2, 4

[SR03] SAUL L., ROWEIS S.: Think globally, fit locally: Unsupervised learning of nonlinear manifolds. *Journal of Machine Learning Research 4* (June 2003), 119–155. 2

[SZS*13] SCHAEFER M., ZHANG L., SCHRECK T., TATU A., LEE J. A., VERLEYSEN M., KEIM D. A.: Improving projection-based data analysis by feature space transformations. In *Proceedings of the SPIE Visualization and Data Analysis 2013 (VDA2013)* (2013). 1, 2, 3, 4

[TAE*09] TATU A., ALBUQUERQUE G., EISEMANN M., SCHNEIDEWIND J., THEISEL H., MAGNOR M., KEIM D.: Combining automated analysis and visualization techniques for effective exploration of high-dimensional data. In *Proc. IEEE Symp. Visual Analytics Science and Technology (VAST)* (2009), pp. 59–66. 2

[vdMH08] VAN DER MAATEN L., HINTON G.: Visualizing data using t-SNE. *Journal of Machine Learning Research 9* (2008), 2579–2605. 3

[VK01] VENNA J., KASKI S.: Neighborhood preservation in nonlinear projection methods: An experimental study. In *Proceedings of ICANN 2001*, Dorffner G., Bischof H., Hornik K., (Eds.). Springer, Berlin, 2001, pp. 485–491. 2

[VK07] VENNA J., KASKI S.: Nonlinear dimensionality reduction as information retrieval. In *Proceedings of the 11th International Conference on Artificial Intelligence and Statistics (AISTATS 2007)*, Meila M., Shen X., (Eds.). Omnipress, San Juan, Puerto Rico, Mar. 2007, pp. 568–575. 2