Visualization of Multidimensional Data Using Interactive 2D Scatterplot

Yitong Zhao, Jieqiong Zhao, Dr. Morteza Karimzadeh, Dr. David S. Ebert

PURDUE UNIVERSITY
VACCINE
Visual Analytics for Command, Control and Interoperability Environments
A U.S. Department of Homeland Security
Science and Technology Center of Excellence

VIP Vertically Integrated Projects

INTRODUCTION

- Multidimensional visualization is critical to facilitate users understanding of the relationship among data items.
- The two dimensional scatter plot is a popular technique that renders data items in two dimensions.
- Interactive exploration of data items helps users better understand clusters and outliers in high dimensional data that are transformed into two dimensions.

OBJECTIVES

- Identify clusters and outliers in multidimensional data.
- Support interactive selection of data points.
- Display the selected data entries in detail.

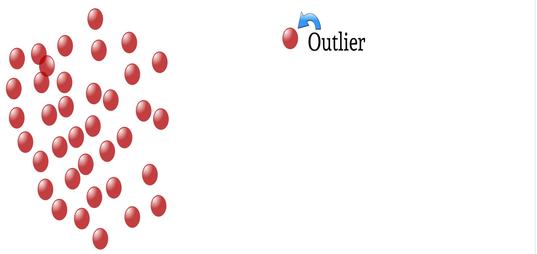

Fig. 1 : A Demonstration for a cluster and an outlier.

TECHNIQUES & METHOD

- To visualize multidimensional data in two dimensional space, PCA (principal components analysis) is adopted as a dimension reduction algorithm.
- The two principal components PCA1 and PCA2 that have the largest variances are selected for the 2D scatterplot (Fig. 2).
- The clusters and outliers are visualized based on distance in the 2D space.
- An interactive Lasso selection method is applied to support the dynamic selection of data entries.
- The selected data entries are highlighted in a table that lists the details of every dimension for all data entries.
- A web application was implemented using the JavaScript library D3.

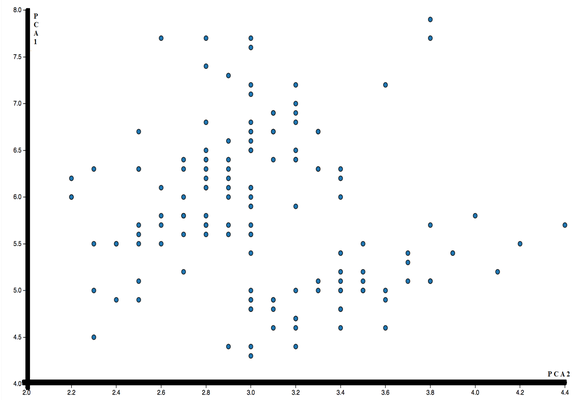

Fig. 2 : 2D Display of Employee Performance, showing the main two PCA components

RESULTS

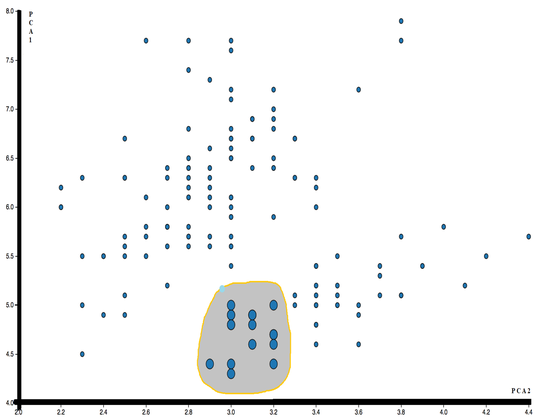

Fig. 4 : Display of 2D Employee Performance After Applying Lasso

PCA1	PCA2	PCA3	PCA4	Number
4.9	3	1.4	3.5	1
4.9	3	1.4	5.2	2
4.7	**3.2**	**1.3**	**3.2**	3
4.6	**3.1**	**1.5**	**1.2**	4
5	3.6	1.4	2.4	5
5.4	3.9	1.7	5.4	6
4.6	3.4	1.4	3.3	7
5	3.4	1.5	2.2	8
4.4	**2.9**	**1.4**	**3.2**	9
4.9	**3.1**	**1.5**	**1.1**	10
5.4	3.7	1.5	3.2	11
4.8	3.4	1.6	4.2	12
4.8	**3**	**1.4**	**1.9**	13
4.3	**3**	**1.1**	**3.1**	14
5.8	4	1.2	2.8	15
5.7	4.4	1.5	3.4	16
5.4	3.9	1.3	2.3	17
5.1	3.5	1.4	1.3	18
5.7	3.8	1.7	3.3	19
5.1	3.8	1.5	2.1	20
5.4	3.4	1.7	3.2	21
5.1	3.7	1.5	1.4	22
4.6	3.6	1	1.2	23
5.1	3.3	1.7	1.5	24
4.8	3.4	1.9	3.2	25
5	**3**	**1.6**	**1.2**	26
5	3.4	1.6	2.4	27
5.2	3.5	1.5	2	28
5.2	3.4	1.4	1.2	29
4.7	**3.2**	**1.6**	**2.6**	30

Fig. 3 : Corresponding Data (including full dimensions) of Employee Performance

SUMMARY

- Clusters and outliers of high-dimensional data can be visually observed in 2D scatterplots.
- Enabling comparison among clusters.
- Lasso selection enables the interactive analysis of relationships among data points.

FUTURE WORK

- Integrate visualizations of multiple features on the scatterplot instead of a simple table.
- Visualize the boundary of clusters.
- Adopt other dimension reduction algorithms such as TSNE.

REFERENCES

[1] Trouble in understanding outliers‘ influence on K-means. (n.d.). Retrieved from https://stats.stackexchange.com/questions/214362/trouble-in-understanding-outliers-influence-on-k-means

[2] L. (2018, August 28). Lasso-js/lasso. Retrieved from https://github.com/lasso-js/lasso

[3] Smith, L. I. (n.d.). A tutorial on Principal Components Analysis. Retrieved from http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf