An Extended Scatterplot Selection Technique for Representing Three Numeric Variables

Mizuki Ishida Takayuki Itoh
Ochanomizu University

ABSTRACT

As a powerful multi-dimensional data visualization method, there have been automatic techniques that selects scatterplots from the ones generated from a multi-dimensional dataset. One of the criteria used in this method supposes that plots are color-coded by a specific categorical variable in the multi-dimensional dataset. We extended this method and show the visualization results by color-coding the plots according to a specific numeric variable in the multidimensional dataset. This extended method represents a specific numeric variable by color, and generates a set of scatterplots in which all pairs of other two numeric variables are assigned to the two axes. Then, the method automatically selects interesting scatterplots that represent three numeric variables.

1 INTRODUCTION

Multidimensional data visualization is one of the most important issues in information visualization. There have been well-known methods for multidimensional data visualization, including icon-based and pixel-based methods as well as geometric methods with explicit coordinate axes, such as ScatterPlot Matrix (SPM) and Parallel Coordinate Plots (PCP). Many geometric methods for visualizing multidimensional data have been published in recent years, which focus on important dimensions by applying dimensionality selection.

Dimension selection corresponds to selecting scatterplots that are worth viewing in the case of scatterplot-based multidimensional data methods. Many existing methods have applied Scagnostics [1], a typical example of numerical evaluation criteria for scatterplots. However, simply selecting scatterplots based on a single index does not necessarily enable comprehensive discovery of the various phenomena latent in the data. The interest of users may lie in correlations between dimensions, and at other times, the interest may lie in the separability of clusters or classes.

Itoh et al. [2] focused on this problem and proposed a method for fast scatterplot selection based on various numerical metrics. The proposed method generates a large number of scatterplots with two axes for two arbitrary variables. Then, the method calculates multiple scores along multiple metrics for each scatterplot, and generates a vector by arranging the scores. A graph is then generated by connecting the scatterplots those vectors are sufficiently similar, and the coloring method is applied to this graph. By selecting a user-specified number of scatterplots from among those that have been assigned the same color, a "diverse set of scatterplots with a certain level of dissimilarity between them" can be automatically selected.

The above technique employs four numerical criteria in their implementation, one of which is based on the assumption that the user selects a specific categorical variable in multidimensional data as a class, and evaluates the scatterplot numerically based on the degree of separation of the class in the scatterplot. In this paper, we extend this numerical evaluation and show the results of color-coding based on a specific real variable in multidimensional data. This extended method generates a scatterplot in which a specific real variable in the multidimensional data is represented by a color, and the other two arbitrary real variables are assigned to the two axes. As are result, the method automatically selects a set of scatterplots that are worth viewing from among many scatterplots that represent the three real-type variables.

2 DIMENSION SELECTION AND SCATTERPLOT EVALUATION FOR MULTIDIMENSIONAL DATA VISUALIZATION

Dimension selection techniques have been widely applied to multidimensional data visualization to effectively represent essential subsets of dimensions. Suematsu et al. [3] and Zheng et al. [4] also converted high-dimensional datasets into low-dimensional subsets and visualized these subsets using multiple PCPs or scatterplots, respectively. These techniques did not provide rich interaction mechanisms to freely select the numbers of dimensions.

Several studies have demonstrated interaction mechanisms to freely visualize interesting low-dimensional subspaces. Lee et al. [5] and Liu et al. [6] applied dimension reduction schemes to interactively select subsets of high-dimensional data. Itoh et al. [7], Watanabe et al. [8], and Nakabayashi et al. [9] presented a series of techniques that easily control the number of dimensions displayed in the PCPs or the number of dimension pairs represented by scatterplots.

Numeric evaluation of the informativeness of scatterplots has been an active research topic. Scagnostics is a remarkable method to quantitatively evaluate the informativeness of scatterplots. Wilkinson et al. [1] proposed nine features of scagnostics based on the appearance of scatterplots. Wang et al. [10] proposed an improved scagnostics by considering the human perception to several metrics, including "Outlying" and "Clumpy." There have been several more studies that focus on specific metrics of scatterplots, including correlation [11,12] and class separation [13,14]. Such evaluation schemes can be applied to appropriate sets of various scatterplots generated from a single multidimensional data. The next section introduces the typical scatterplot selection technique [2] applied in this study.

3 SCATTERPLOT SELECTION TECHNIQUE APPLYING A GRAPH COLORING METHOD

This section describes the details of the automatic scatterplot

selection method [2]. The method calculates multiple scores along multiple metrics for each set of scatterplots generated from multidimensional data, and treats them as vectors. The method then selects a user-specified number of the vectors satisfying the following requirements, and finally displays the corresponding scatterplots.

[Requirement 1:] Vectors that are sufficiently far apart can be selected instead of vectors that are too close to each other, so that a variety of scatterplots can be selected.

[Requirement 2:] By selecting long vectors, a characteristic scatterplot can be selected.

3.1 Metrics for Scatterplot Selection

The scatterplot selection technique implements four types of scores: "correlation," "thinness," "clumping," and "separateness." The extended technique presented in this paper is closely related to the "class separability". The implementation assumes that the value of a specific categorical variable in multidimensional data is the class of each individual, and quantifies class separability based on information entropy. Scatterplots with low information entropy indicate that the classes are separated in the scatterplots, and thus are often interesting.

3.2 Graph Coloring Problem for Scatterplot Selection

Next, a graph coloring problem is applied to select a scatterplot. Given a graph G={S,E} consisting of a set of scatterplots S and a set of edges E connecting the scatterplot pairs. Here, a pair of scatterplots are connected by an edge if the vector of scores has a cosine similarity greater than a threshold value. Then, a unique color is assigned to each of the scatterplots that make up G. The methods has a constraint that two scatterplots adjacent to each other at an edge must be assigned different colors. This constraint ensures that highly similar scatterplots are assigned different colors. As a result, a set of scatterplots that are assigned the same color will be composed of scatterplots that have a certain degree of dissimilarity to each other.

The number of scatterplots specified by the user is automatically selected and displayed. The method selects a predetermined number of scatterplots from among those that are assigned the same color identifier. Scatterplots with the same color identifier have a certain degree of dissimilarity, and therefore, [Requirement 1] is satisfied by selecting scatterplots from these groups. Moreover, [Requirement 2] is satisfied by selecting the scatter plots with the same color identifiers in the order of the maximum score.

In summary, the method automatically selects scatterplots by the following steps:

1. Generate the graph by iterating the generation of edges for pairs of scatterplots whose cosine similarity is above a certain value.
2. Select one of the scatterplots with the largest vector length of scores as the starting point of the search.
3. Search the scatterplots that compose the graph G in width-first order and assign a color identifier to each of them. Select a color identifier that is different from any of the color identifiers assigned to adjacent scatterplots connected by edges.

The user-specified number of scatter plots with the same color identifier are selected in the order of the maximum score.

4 Extended Technique for Representing Three Numeric Values

The scatterplot selection methods introduced in the previous section have been applied to multidimensional data including categorical variables, but there are many cases of multidimensional data consisting only of real-type variables and not including categorical variables. On the other hand, there are many cases where three or more real-type variables are correlated in real-world data. In this paper, we present an example of automatic selection of scatterplots representing three real-type variables.

4.1 Processing Flow

The m-dimensional data A assumed in this report has n samples $A = \{a_1, a_2, \dots, a_n\}$, and each sample is an m-dimensional vector described as $a_i = (a_{i1}, a_{i2}, \dots, a_{im})$. In this case, this method color-codes the points that make up the scatterplot by the j-th dimension real-type variable. Specifically, the interval [min_j, max_j] indicated by the minimum and maximum values of the j-th real-valued values a_{1j}-a_{nj} of each sample is divided into N intervals, each value of a_{1j}-a_{nj} is identified as the number of intervals to which it belongs, and the points are color-coded by the intervals to which they belong. The above process corresponds to converting a real-valued variable of dimension j into a categorical variable with N types of variable values. Figure 1 shows an overview of the process.

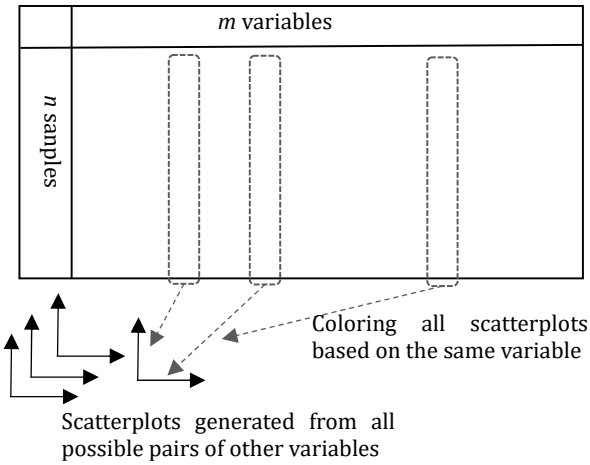

Figure 1: Processing flow of the extended scatterplot selection. The technique specifies colors of the points based on the same variable and generates scatterplots from all possible pairs of other variables. The technique then applies the graph coloring method to select appropriate number of various scatterplots.

4.2 Used Data

A dataset from the optimization process of aircraft wing shape design is applied in this study. The wing shape was designed with 72 explanatory variables in the case, and four objective functions were calculated by hydrodynamic simulation. This process was iterated by a multi-objective genetic algorithm to obtain 776 Pareto solutions [15]. In other words, this optimization process yielded 776 different design results. The designer can select the wing shape by making a decision

among these design results.

The explanatory variables are described as dv00 to dv71 in this section. The following six explanatory variables are known to be particularly important in finding the optimal solution.

dv00, dv01: Span lengths of the inner and outer wings
dv02, dv03: Wake angle
dv04, dv05: Length of the wing root

Other explanatory variables include:

dv06 - dv25: Variables related to wing warp
dv26 - dv32: Variables related to wing twist
dv33 - dv71: Variables related to the wing thickness

The four objective functions are as follows:

CDt: drag coefficient for transonic cruise
CDs: drag coefficient for supersonic cruise
Mb: Bending moment of the wing root at supersonic cruise
Mp: Torsional moment applied to the wing tip

This study visualized this dataset as a group of 776 points that make up a 76-dimensional vector. The following phenomena can be observed from the point clouds that compose the Pareto solution by applying the proposed method to the visualization:

- Strong correlations between two variables
- Correlation across three or more variables
- Outliers or clusters separated from other point clouds

The visualization of the Pareto solution is performed by applying the proposed method to the visualization of point clouds.

4.3 Visualization Result

This section explains our discovery about optimization of wing shape design, using the results of color-coding the scatterplots by the explanatory variable dv05 as an example. We selected various explanatory variables and observed the visualization results applying our implementation on trial, and subjectively selected dv05 as an example introduced in this paper. Figure 2 shows a set of scatterplots selected by this method. This scatterplot has one of the explanatory variables on the horizontal axis and one of the objective variables on the vertical axis. The interval of the explanatory variable dv05 is divided into three parts, and the points are colored yellow, magenta, or cyan. The majority of the scatterplots displayed in Figure 2 in which the three colors are separated have Mb as the vertical axis, and furthermore, the three colors are separated in the vertical direction, suggesting that the explanatory variable dv05 chosen for coloring has a particularly strong correlation with Mb.

The two scatterplots at the lower left end of Figure 2 (Figure 3) are generated by assigning dv00 and Mb, dv02 and Mp on the two axes, respectively. Such a strong correlation among the five variables dv00, dv04, dv05, CDt, and Mb, as well as among the four variables dv02, dv03, CDs, and Mp has been also shown in a previous study [7]. The strong correlation between dv00 and Mb can be observed in the distribution of the point cloud in Figure 3 (left), and also with dv05 based on the color separation of the point cloud. On the other hand, a strong correlation between dv02 and Mp can be observed in Figure 3 (right), but the lack of separation of the colors of the point clouds suggests that there is little correlation with dv05. Furthermore, comparing the two scatterplots in Figure 3, we can observe that the correlation between dv02 and Mp is sharper than the correlation between dv00 and Mb.

Figure 4 shows another two scatterplots included in Figure 2, generated by assigning dv57 and Mb, dv07 and Mp on the two axes, respectively. In Figure 4 (left), the upper right part of the scatterplot is almost blank, and in Figure 4 (right), the upper left part of the scatterplot is almost blank. This indicates that the distribution of the Pareto solution is partially coarse and partially dense. We expect that the optimization process can be more efficient by analyzing the coarseness and density observed only when a specific variable is assigned to the horizontal axis.

On the other hand, we found no outliers or clusters in any of the scatterplots that were clearly separated from other point groups. In other words, the 776 Pareto solutions that comprise the data are distributed uniformly in the 76-dimensional space, with no disconnections among the Pareto solutions.

Totally, the extended visualization brought new knowledge that has not been described in the previous studies [7,15].

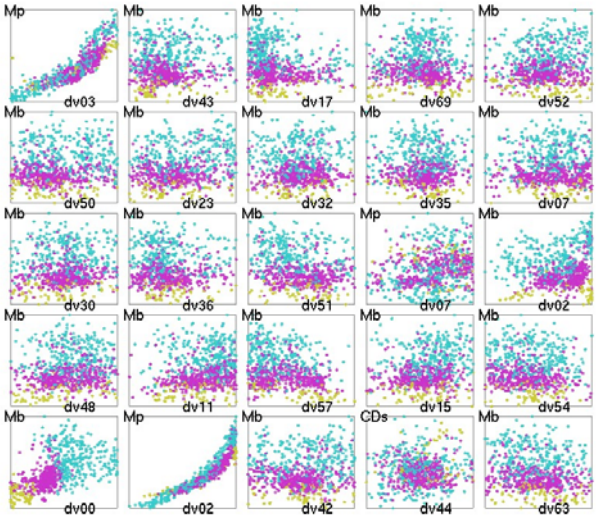

Figure 2: Scatterplot selection result colored based on the variable dv05.

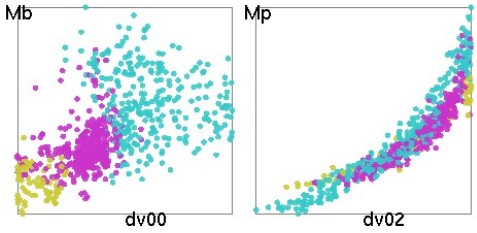

Figure 3: Scatterplots where dv00 and Mb, dv02 and Mp are assigned to the axes respectively.

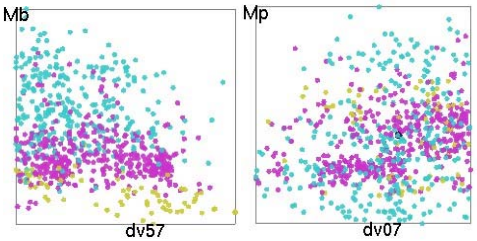

Figure 4: Scatterplots where dv57 and Mb, dv07 and Mp are assigned to the axes respectively.

4.4 Discussion and Limitation

We discussed the limitation and future aspects of the current implementation and visualization results as follows.

Constraints posed by the use of color: In this method, each scatterplot effectively represents three real variables by assigning colors to the intervals of the real variables. It has been experimented that color is less favorable than position (coordinate values) [16] for accurately reading real values while variables are assigned to visual attributes. In other words, it is necessary to use this method without expecting to be able to strictly read the value of a variable to which a color is assigned.

Selection of the variable to be assigned a color: In our current implementation, the variable to be assigned a color is selected manually. In other words, one of our future tasks will be to develop an effective method to automatically select appropriate real variables that can be assigned a color to obtain a visualization worth viewing.

Division of the variables to be assigned a color: We just divide the corresponding variable to the fixed number of equal intervals. We would like to develop an effective method to automatically select an appropriate number of N. Also, we would like to experiment with non-equal intervals such as quantile intervals.

Handling of scores for scatterplot selection: We simply add up the four types of scores introduced in section 3.1 in our current implementation though these scores should be weighted. In particular, the weight of "class separability," which is related to the color of the scatterplot, needs to be carefully adjusted. We would like to modify the method so that, for example, the order is not "the order of the maximum score" as discussed in section 3.2, but "the order of the class separateness score".

5 Conclusions

In this paper, we introduced an extended case study of a scatterplot selection method based on multidimensional data consisting only of real-type variables, classifying each sample by the range of one real-type variable, and representing it in terms of colors on a scatterplot. The visualization results were verified by applying the method to a dataset from the optimization process of aircraft wing shape design. We also discussed the challenges of the method.

In the future, we would like to solve the issues discussed in section 4.4 and verify the effectiveness of this method by applying it to data other than aircraft design.

Acknowledgement

We appreciate Professor Shigeru Obayashi of the Institute of Fluid Science, Tohoku University, for providing the dataset for the optimization process of aircraft design.

References

[1] L. Wilkinson, A. Anand, R. Grossman: “Graph-theoretic scagnostics”, IEEE Symposium on Information Visualization, 157-164, 2005.

[2] T. Itoh, A. Nakabayashi, M. Hagita: “Scatterplot selection applying a graph coloring algorithm”, Visual Information Communication and Interaction Symposium (VINCI2021), 2021.

[3] H. Suematsu, Y. Zheng, T. toh, R. Fujimaki, S. Morinaga, Y. Kawahara: “Arrangement of low-dimenional parallel coordinate plots for high-dimensional data visualization”, 17th International Conference on Information Visualisation, 59-65, 2013.

[4] Y. Zheng, H. Suematsu, T. Itoh, R. Fujimaki, S. Morinaga, Y. Kawahara, “Scatterplot layout for high-dimensional data visualization”, Journal of Visualization, 18(1), 111--119, 2015.

[5] J. H. Lee, K. T. McDonell, A. Zelenyuk, D. Imre, K. Muller, “A structure-based distance metric for high-dimensional space exploration with multidimensional scaling”, IEEE Transaction on Computer Graphics, 20(3), 351-364, 2013.

[6] S. Liu, B. Wang, P.-T. Bremer, V. Pascucci, “Distortion-guided structure-driven interactive exploration of high-dimensional data”, Computer Graphics Forum, 33(3), 101-110, 2014.

[7] T. Itoh, A. Kumar, K. Klein, J. Kim: “High-Dimensional Data Visualization by Interactive Construction of Low-Dimensional Parallel Coordinate Plots”, Journal of Visual Languages and Computing, 43, 1-13, 2017.

[8] A. Watanabe, T. Itoh, M. Kanazaki, K. Chiba, “A scatterplots selection technique for multi-dimensional data visualization combining with parallel coordinate plots”, 21st International Conference on Information Visualisation (IV2017), 78-83, 2017.

[9] A. Nakabayashi, T. Itoh, “A technique for selection and drawing of scatterplots for multi-dimensional data visualization”, 23rd International Conference on Information Visualisation (IV2019), 62-67, 2019.

[10] Y. Wang, Z. Wang, T. Liu, M. Correll, Z. Cheng, O. Deussen, M. Sedlmair, “Improving the robustness of scagnostics”, IEEE Transactions on Visualization and Computer Graphics, 26(1), 759-769, 2020.

[11] L. Harrison, F. Yang, S. Franconeri, R. Chang, “Ranking visualizations of correlation using weber’s law”, IEEE Transactions on Visualization and Computer Graphics, 20(12), 1943-1952, 2014.

[12] L. Shao, A. Mahajan, T. Schreck, D. J. Lehmann, “Interactive regression lens for exploring scatter plots”, Computer Graphics Forum, 36(3), 157-166, 2017.

[13] M. Sedlmair, A. Tatu, T. Munzner, M. Tory, “A taxonomy of visual cluster separation factors”, Computer Graphics Forum, 31(3), 1335-1344, 2012.

[14] M. Aupetit, M. Sedlmair, “Sepme: new visual separation measures”, IEEE Pacific Visualization Symposium, 43-52, 2016.

[15] D. Sasaki, S. Obayashi, K. Nakahashi: “Navier-Stokes Optimization of Supersonic Wings with Four Objectives Using Evolutionary Algorithm”, Journal of Aircraft, 39(4), 621-629, 2002.

[16] R. Mazza, Introduction to Information Visualization, Springer, ISBN:978-1-84800-218-0, 2009.