An Interactive Tool for Feature Analysis of Outliers in Multi-Dimensional Data

Kentaro Asai*
The University of Tokyo

Tsukasa Fukusato†
The University of Tokyo

Takeo Igarashi‡
The University of Tokyo

Abstract

This paper presents an interactive visualization tool for the analysis of outliers in multi-dimensional data. Each data sample consists of multiple variables (features). For a given outlier data sample, our tool helps the user to identify which variable of the data sample makes it an outlier. Our tool consists of a scatter plot view and variable relation graph view. The user first identifies an outlier data sample in a scatter plot. The user selects the outlier data sample, and the system updates the variable relation graph to visualize relationship between the variables of the outlier data sample. The user then examines the variable relation graph and the scatter plot based on the visualization to identify outlying variables.

Index Terms: Outlier—Outlier examination—Visualization—Visualization techniques; Graph—Dimension Graph—

1 Introduction

An outlier is a data sample that contains variables (features) that take unusual values. Given an outlier data sample, it is important to identify a variable of the sample that makes the sample an outlier. It helps the user to identify problems in data collection process (e.g., sensor error). It also allows the user to use the other (not-outlying) variables of the outlier data sample, instead of eliminating the entire data sample.

Identification of an outlying variable is often done by uni-dimensional outlier detection methods, e.g., box plot rule. However, in some cases, it is difficult to identify it by conventional uni-dimensional outlier detection. For example, Figure 2 shows a scatter plot matrix (SPLOM) of four-dimensional dataset with an outlier. In the SPLOM, we can see a blue plot is an outlier data sample. This outlier data sample is generated by adding noise to the value on dimension *e1* of an inlier data sample (following orange arrows). However, we cannot identify that the outlier data sample has an outlying variable value on dimension *e1* with uni-dimensional outlier detection methods. This is because the value on *e1* is not so outlying compared to the other data samples in this single dimension. Our goal is to assist the user in searching for an outlying variable by inspecting the scatter plots, especially there are many variables and it is too time consuming to inspect them all manually.

We provide an interactive system to help users to identify outlying variables of an outlier data sample (see Figure 1). The system consists of scatter plot view and variable relation graph (VRG) view. The system analyzes pairwise relations between variables, and then visualizes the result of analysis in the variable relation graph. The user then examines scatter plots by referring the visualization result.

*e-mail: procken@is.s.u-tokyo.ac.jp
†e-mail: tsukasafukusato@is.s.u-tokyo.ac.jp
‡e-mail: takeo@acm.org

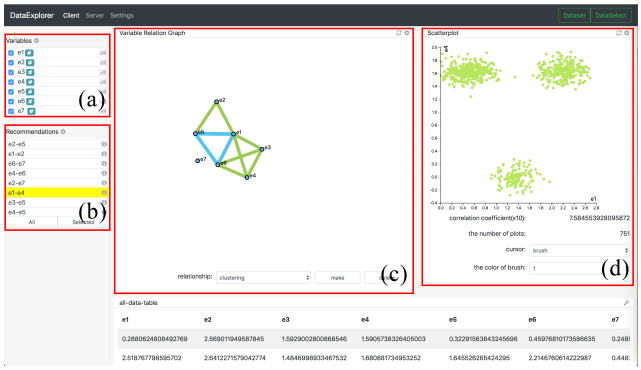

Figure 1: A screenshot of our system. It consists of variable relation graph view (c), scatter plot view (d), variable panel (a), and recommendation panel (b). The variables panel (a) shows variables in the dataset and the recommendation panel (b) shows efficient scatter plots in order from the largest strength.

2 Related Work

For visualizing high-dimensional data, parallel coordinates are often used. However, parallel coordinate is mainly focusing on the relation between most neighboring axis, so this is not appropriate as an outlier visualization [3]. This is because the combination of axis highly effects the visibility of outliers. In addition, several researches into visualizing high-dimensional data by using scatter plot matrix have been proposed. Figure 2 shows an example of scatter plot matrix of four-dimensional data. However, even if we employ scatter plot matrix, it is not easy to visualize the outlying variables of outliers clearly when the dimension is high (variable *e1*). In this paper, we use a dimension graph [5] to visualize relations between dimensions and extend the concept to examine the outlying variables of outliers. In particular, when the dataset has a set of outliers which have independent outlying values on several dimensions, our tool can find them easily and quickly.

3 Method

Our goal is to assist the user in identifying outlying variables manually, not to automate the process. The basic procedure is to manually inspect scatter plot (variable pair) one by one to examine the relation between individual variables. The system makes the process efficient by suggesting which scatter plot to inspect. The suggestion is based on the analysis of data distribution in each scatter plot (variable-pair), and is presented to the user in the variable relation graph.

3.1 Variable Relation Graph

Inspired by Zhang's dimension graph [5], we define a variable relation graph (Figure 3) to show relations between variables of a high-dimensional dataset. The variable relation graph has two types of

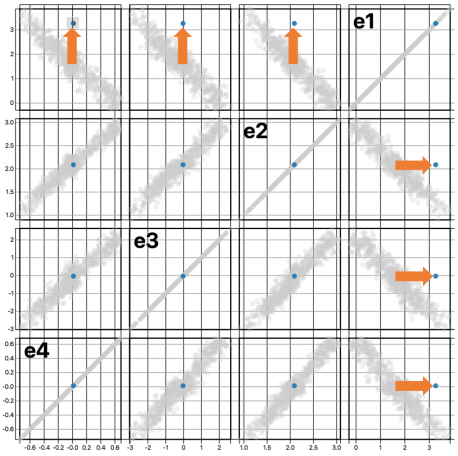

Figure 2: An example of artificial SPLOM dataset. We manually set an outlier plot (blue) by adding noise to one normal datum on dimension *e1* (following orange allows). The user can see that the blue data sample is an outlier by observing the top left scatter plot, but it is difficult to judge which of *e1* or *e2* is the outlying variable.

elements: (1) nodes and (2) edges. A node represents a variable (dimension) and an edge represents the relation between a pair of variables. Each edge corresponds to a scatter plot. In the variable relation graph, the edge color represents the two types of correlation analysis, (blue) linear correlation and (green) cluster relation. The edge width represents a strength of the relation. In our system, the user interactively constructs this graph by seeing scatter plots one by one.

3.2 Outlier Examination

When the user selects an outlier data sample, the system suggests outlying relation between variables by a thick edge and likely outlier variable with an orange node. The blue edges indicate linear correlation in the corresponding scatter plots. The suggestion is based on the distance between the data sample and the regression line. Our system calculates the variance σ of data points in the direction perpendicular to the regression line. In our current implementation, when the distance is more than 2σ, the system judges that the data sample is an outlier in the scatter plot, and makes the corresponding edge automatically thick. The green edges indicate cluster structures in the corresponding scatter plots and the suggestion is based on the result of DBSCAN [2]. If the selected outlier data sample is judged as an outlier data sample by DBSCAN, the green edges become thick. After assigning thickness to all the edges, the system computes the ratio of the thick edges among all the edges connected to a node. When the ratio is equal to or more than a predetermined threshold $\alpha \in [0.0, 1.0]$ (in this paper, we set $\alpha = 0.66$), our system judges that the node represents an outlying variable.

4 User Case Scenario

To illustrate the utility of our graph tool, we show a user case scenario. Figure 3 is an example of the variable relation graph in outlier examination mode from Automobile Data Set [1]. We can examine outliers one by one by clicking an outlier data sample in the scatter plot. In this figure, the graph is displayed for an outlier data sample in the dataset. From this graph, we can understand that three nodes

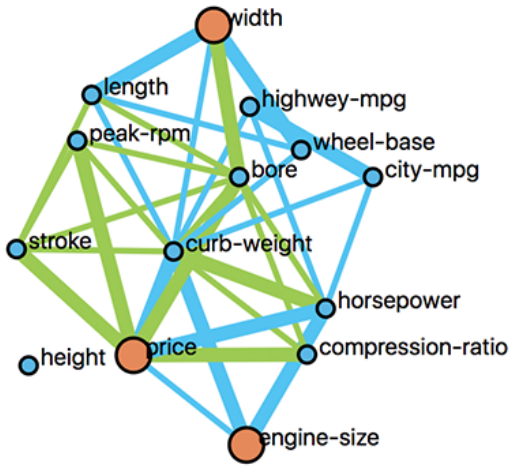

Figure 3: Variable relation graph after focusing on a data sample. Each edge corresponds to a variable pair (a scatter plot). Blue edge means linear correlation and green edge means cluster relation. Thick edge means that the current data sample is an outlier in the corresponding 2D scatter plot. A node with many thick edges become orange, meaning that the variable is likely an outlying variable.

("Width", "Price" and "Engine Size") seem to be outlying. By this process, we can easily see the outlying variable of outliers.

5 Future Works

Currently, our system mainly focuses pair-wise relation between variables. We plan to extend our framework to handle relation among more than two variables in the future. We plan to use dimension reduction techniques such as principal component analysis (PCA) for the analysis. Our current analysis is limited to regression and cluster. We plan to handel other relations (such as Wilkinson's Scagonstics [4]).

Acknowledgments

This work was supported by JST CREST Grant Number JPMJCR17A1, Japan

References

[1] Online: accessed 07-june-2018, automobile data set, https://archive.ics.uci.edu/ml/datasets/automobile.

[2] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, pp. 226–231. AAAI Press, 1996.

[3] L. Wilkinson. Visualizing outliers, https://www.cs.uic.edu/ wilkinson/Publications/outliers.pdf. 2016.

[4] L. Wilkinson and G. Wills. Scagexplorer: Exploring scatterplots by their scagnostics. *IEEE Pacific Visualization Symposium*, pp. 73–80, 2014.

[5] Z. Zhang, K. T. McDonnell, E. Zadok, and K. Mueller. Visual correlation analysis of numerical and categorical data on the correlation map. *IEEE Transactions on Visualization and Computer Graphics*, 21(2):289–303, Feb 2015. doi: 10.1109/TVCG.2014.2350494