A Scatterplot-based Visualization Tool for Regression Analysis

Chie Suzuki[*1] Takayuki Itoh[*2] Keisuke Umezu[*3] Yousuke Motohashi[*4]
Ochanomizu University[*1*2] NEC[*3*4]
{chie, itot}@itolab.is.ocha.ac.jp[*1*2] k-umezu@ak.jp.nec.com[*3] y-motohashi@bk.jp.nec.com[*4]

Abstract

Regression analysis has been widely applied to various academic and industrial fields. Applications of regression analysis include medical problems such as health estimation, environmental problems such as disaster prediction and energy consumption estimation, and business/economic analytics. Here, accuracy and quality of regression analysis strongly depend on relevancy between input explanatory variables and actual objective functions. It often happens that several explanatory variables are well correlated with objective functions while others do not well correlated, and therefore accuracy of regression analysis may improve by removing unnecessary explanatory variables. This paper presents a scatterplot-based regression analysis tool. This tool visualizes the distribution of errors between actual and estimated values of objective functions, and provides user interfaces to explore the relationships between explanatory variables and the errors. This paper introduces examples of the visualization results using the presented tool with actual and estimated revenues at a store.

Keywords--- **Regression analysis, scatterplots**.

1. Introduction

Numeric estimation and prediction is an important problem for various academic and industry fields of natural and social sciences. We may want to estimate revenue while developing new products. We may also want to predict temperature or other weather conditions for upcoming days. Regression analysis is a key methodology to assist such estimation processes. Analysis process gets more and more complex by applying multiple regression analysis consuming multiple explanatory variables, or mixture models applying multiple alterative functions. We often need to select explanatory variables while well affect to estimation, and remove other variables from regression analysis processes. This is a key process to archive successful regression analysis.

Let us describe a real scenario of regression analysis on revenue of products at a store. Revenue of products are often affected various environmental factors, including day of the week, weather and temperature, and neighbor events. It is important to optimize the stock of the products to maximize the profits during promoting them, and therefore revenue estimation is an important process. Retail sales companies usually record transactions with above mentioned factors to analyze the trend of the sales and utilize to the revenue estimation. Here, factors relevant to the revenues are important information as explanatory variables of regression analysis. On the other hands, factors less relevant to the revenues may be noisy input of regression analysis and prevent accurate estimation of revenues. Therefore, it is important to understand how each of explanatory variables affects to the errors between actual and estimated revenues. It is not an easy task since the relationships between explanatory variables and revenues are complex.

This paper presents a software tool for visualizing errors between actual and estimated objective function values for regression analysis. Our implementation features a 3D scatterplot assigning two explanatory variables to the X- and Y-axes, while assigning actual or estimated objective function values to the Z-axis. Colors of dots in the scatterplot are calculated from errors between actual and estimated objective function values. This representation makes users easier to recognize which portions samples which have larger errors are concentrated.

Here, effectiveness of visualization results strongly depends on selection of explanatory variables to be assigned to the X- and Y-axes. Our technique estimates the relevancy between the errors and each of the explanatory variables, and suggests several explanatory variables which are valuable to assign to the axes. This paper presents a result applying Akaike's Information Criterion (AIC) to evaluate the effectiveness of the explanatory variables.

2. Related Work

Regression analysis has been widely applied for modeling and estimation of various academic and industrial problems. Herpen et al. [1] presented a modeling approach using regression analysis to explain the market shares of sustainable products. Chen et al. [2] examined how insurance product characteristics and retailer actions influence the purchase by applying regression analysis.

Even though it is important to intuitively understand how the models work with real problems and datasets, still small number of studies on visualization of regression analysis have been presented. Muhlbacher et al. [3] presented a visualization tool to effectively recommend preferable mathematical models and sets of well correlated explanatory variables to assist interactive construction of regression models. Krause et al. [4] proposed a glyph-based representation for interactive

feature selection for predictive models including regression models. Similar visualization tools for predictive model analysis have been customized for various applications including bioinformatics [5] and social media analysis [6].

Subspace selection from high-dimensional explanatory variable space is one of the key techniques in this study. There have been many studies on dimension selection for high-dimensional data visualization. For example, we presented a scatterplot selection [7] and parallel coordinate plot decomposition [8] techniques which specifies well-related pairs of dimensions from high-dimensional datasets. Many of these techniques focused on correlations or numeric distribution among the dimension. Differently from such existing studies, we applied Akaike's Information Criterion (AIC) to select small number of dimensions, since the goal of this study is finding informative subspaces to analyze the errors of prediction results. Wang et al. [9] presented an AIC-based dimension selection scheme for high-dimensional data visualization; however, this study did not focus on predictive model visualization.

3. Presented Visualization Tool

This section presents the processing flow and user interface of the presented visualization tool. The section also presents a technique to select explanatory variables to be assigned to the axes of the scatterplot applying AIC. Figure 1 shows a snapshot of the presented visualization tool.

3.1 Data Structure

The visualization tool supposes the following structure of datasets are given.

$$\mathrm{X} = \{x_1, x_2, \ldots, x_n\}$$

$$x_i = \{v_{i1}, \ldots, v_{im}, c_{i1}, \ldots, c_{il}, p_i, a_i\}$$

where,

X: a set of samples
n: the number of samples
x_i: the i-th sample
m: the number of explanatory variables
v_{ij}: the j-th explanatory variable of the i-th sample
l: the number of categorical variables
c_{ij}: the j-th categorical variable of the i-th sample
p_i: the estimated objective function value of the i-th sample
a_i: the actual objective function value of the i-th sample.

In case of regression analysis of product sales revenues, we suppose numeric input information describing conditions such as number of stocks or average temperature as explanatory variables. Meanwhile, non-numeric input information such as attributes of products or day of the week can be treated as categorical variables.

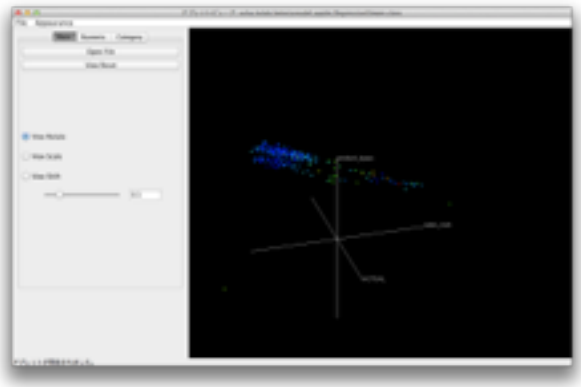


Figure 1 Snapshot of the presented visualization tool

3.2 Visualization by a 3D Scatterplot

Our implementation features a 3D scatterplot to visualize the above mentioned data structure. This visualization tool assigns two of explanatory variables to the X- and Y-axes, and actual or estimated objective function values to the Z-axis. Also, the tool assigns independent colors to the dots of the scatterplot according to the errors between actual and estimated objective function values. Our implementation assigns warmer colors to samples which have larger errors, while assigning cooler colors to samples which have smaller errors.

3.3 User Interface

Our implementation features four tabs with GUI (Graphical User Interface) widgets at the left side of the window. The first tab features GUI widgets for mandatory operations including data file selection and viewing adjustments. The second tab displays the list of explanatory variables and provides radio buttons to select one of them to assign to X-, Y-, and Z-axes of the 3D scatterplot. Figure 2 shows a snapshot of the second tab.

The third and fourth tabs features GUI widgets such as radio buttons and check boxes for categorical variable selection. Figure 3 shows a snapshot of these tabs. The third tabs displays a list of categorical variables and provides radio buttons to select one of them. When a user selects one of the categorical variables, the fourth tab is refreshed and displays the list of values of the selected categorical values. This tab features a set of check boxes so that users can select one or more values. Suppose that "day of the eek" is one of the categorical variables. When a user selects this, the fourth tab displays seven values corresponding to Monday to Sunday.

Suppose several values are selected on the fourth tab while others are not selected. In this case the scatterplot of our implementation assigns colors to dots corresponding to the samples which have the selected values, while displaying other samples as gray dots. This representation makes users easier to observe the relationship between the selected categorical variable and distribution of errors between actual and estimated objective function values.

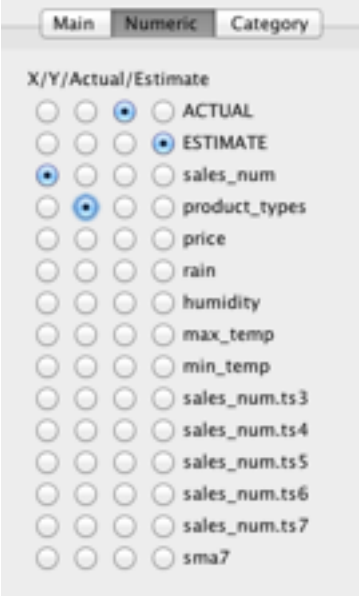


Figure 2 GUI widget for selection of explanatory variables to assign to X-, Y-, and Z-axes of the 3D scatterplot.

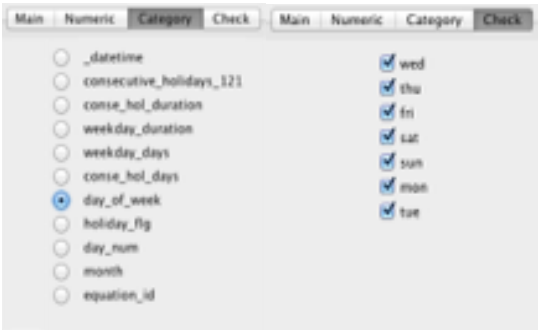


Figure 3 (Left) GUI widgets to select one of the categorical variables. (Right) GUI widgets to select one or more values of the selected categorical variables.

3.4 Evaluation of Explanatory Variables

We can freely select explanatory variables to visualize correlation between the variables and errors of prediction results. In other words, bothering operations to select the variables may be required while using this visualization tool. It should be therefore useful if we can suggest meaningful sets of explanatory variables. Our current study applies Akaike's Information Criterion (AIC) to select the model which minimizes the Kullback-Leibler distance with the actual values. It is defined as:

$$\mathrm{AIC} = -2 \log L + 2k$$

Here, L is the maximum likelihood and k is the number of parameters. Selecting the parameter sets which minimizes the AIC value, we can select the model which optimizes the prediction.

4. Example

We applied a regression analysis result of retail sales revenues to the presented visualization tool. The dataset included 344 samples which have actual and estimated objective function values, 12 explanatory variables, and 8 categorical variables. A heterogeneous mixture model [10][11] is applied to the regression analysis. The explanatory variables are describes as alphabetical characters A to L in this section, because the actual names of the variables are confidential information of the data owner company. Categorical variables included day of the week, month, and equation ID of the mixture model. This section introduces a case study with this dataset.

4.1 Variable Evaluation with AIC

Table 1 AIC values with arbitrary number of explanatory variables.

Explanatory variables	AIC values
A	113.76
A to B	115.76
A to C	117.84
A to D	119.88
A to E	121.92
A to F	123.93
A to G	125.96
A to H	127.96
A to I	129.95
A to J	131.95
A to K	133.95

Figure 4 shows a visualization result where the explanatory variable A is assigned to the X-axis, and B is assigned to the Y-axis. The result shows that errors tend to large if both values of explanatory variables A and B are large. Similarly, errors tend to large if both values of explanatory variables A and C are large, as demonstrated in Figure 5 where A is assigned to the X-axis, and C is assigned to the Y-axis.

Figure 6 shows another visualization result where the explanatory variable B is assigned to the X-axis, and C is assigned to the Y-axis. Dots in warmer and cooler colors are somewhat separated in this result; however, these colors are not concentrated but broadly distributed comparing with Figures 4 and 5. Therefore, the result applying the two variables B and C weakly explains the relationships between explanatory variables and errors comparing with Figures 4 and 5. We therefore concluded the variable A is especially important for regression analysis.

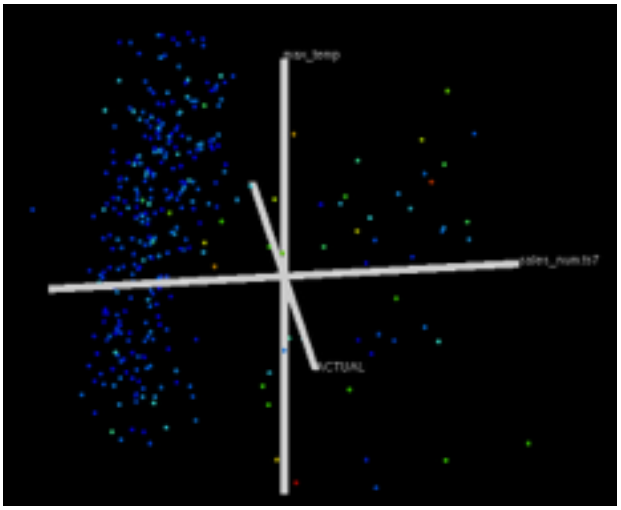


Figure 4 Visualization with variables A and B.

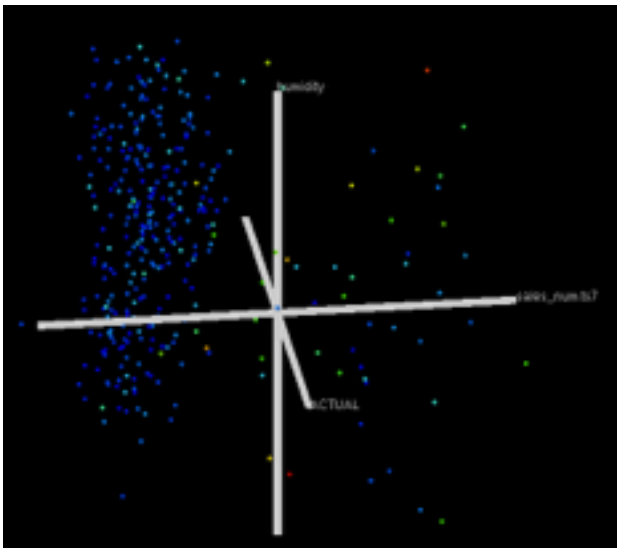


Figure 5 Visualization with variables A and C.

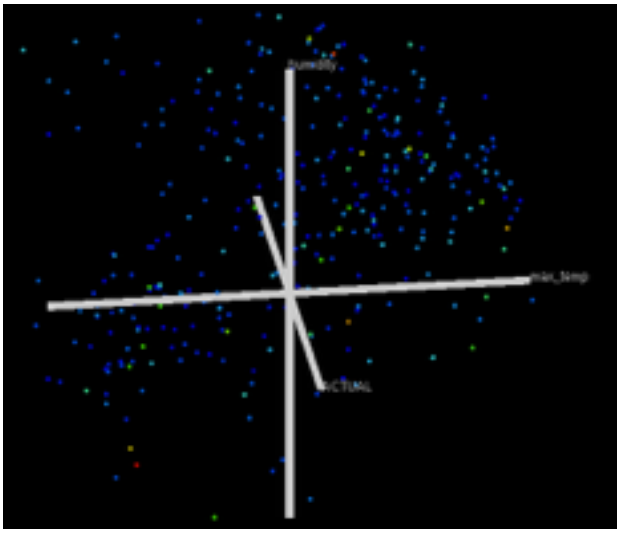
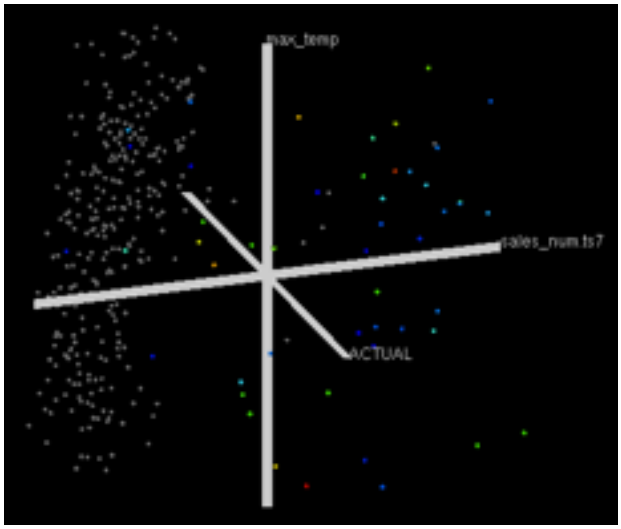


Figure 6 Visualization with variables B and C.

4.2 Categorical Variable Selection

We also visualized the dataset with selecting categorical variables. Figure 7 shows an example of selection of day of the week, while assigning explanatory variables A and B to the two axes of the scatterplot. Figure 7(Upper) shows that many transactions on Monday have larger values of the variable A, because colored dots concentrate at the right side of the scatterplot space. Also, the transactions have larger errors of predictions, because many corresponding dots are drawn in warmer colors. On the contrary, Figure 7(Lower) shows that many transactions during Tuesday to Friday have smaller values of variable A, and smaller errors of prediction. This visualization suggests to modify the prediction model on Monday.



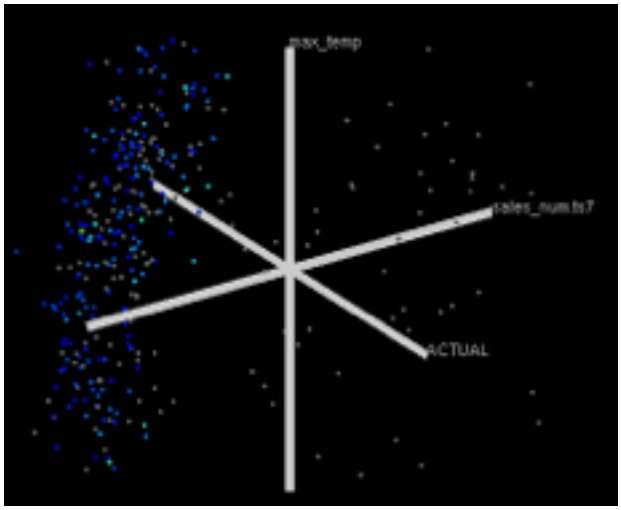


Figure 7 Visualization with selection of day of the week. (Upper) Coloring transactions on Monday. (Lower) Coloring transactions during Tuesday to Friday.

Figure 8 shows another example of selection of month, while assigning explanatory variables A and D to the two

axes. Figure 8(Upper) shows that errors of prediction tends to large when variable A is large and variable D is small. Figure 8(Lower) shows that many transactions in February and September have larger errors because corresponding dots are painted in warmer colors. This visualization recommends to explore causal relationships of errors in February and September.

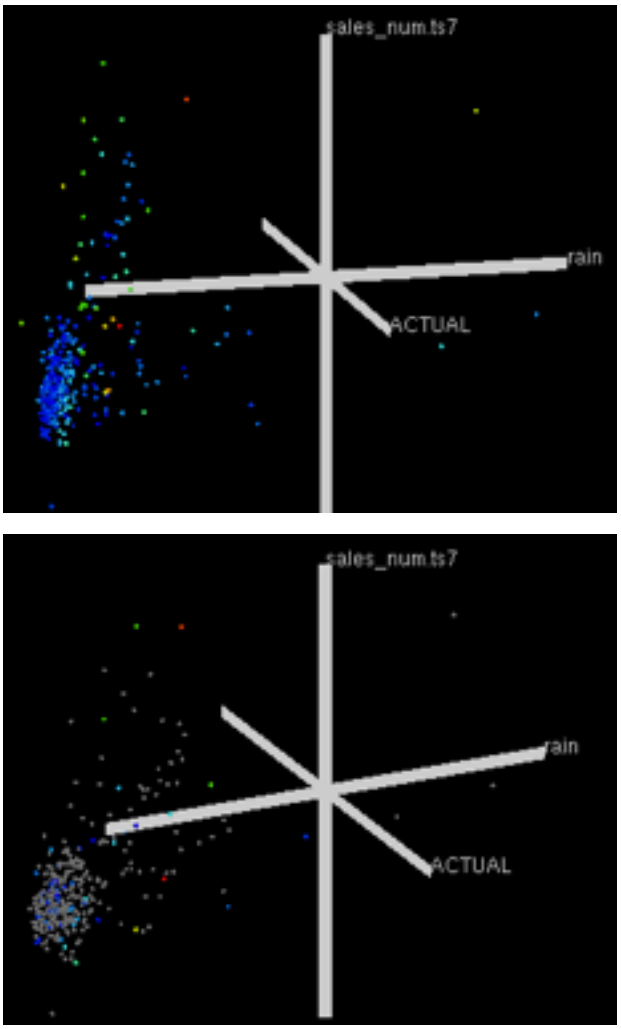


Figure 8 Visualization with selection of month. (Upper) Coloring every transactions. (Lower) Coloring transactions in February and September.

Figure 9 shows the last example that selective visualizes the errors of predictions for each equation of the mixture model.

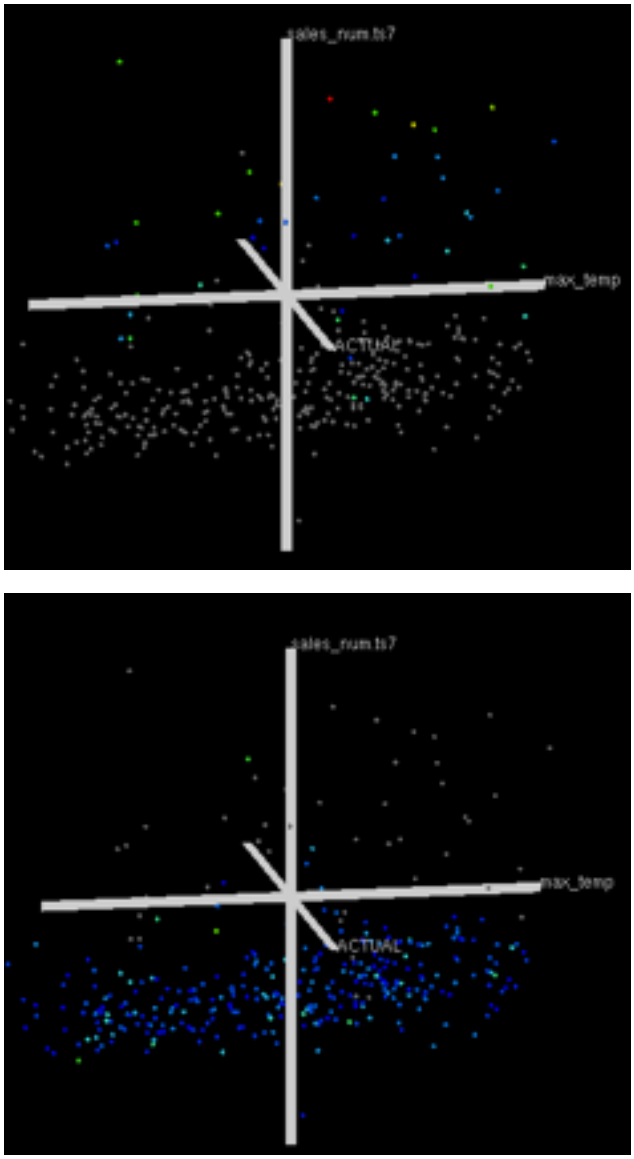


Figure 9 Visualization with selection of equations of the mixture model.

5. Conclusions

This paper presented a scatterplot-based visualization tool for regression error analysis. This tool assigns two of explanatory variables to the X- and Y-axes, and actual or estimated objective function values to the Z-axis of the 3D scatterplot. It also represents the distribution of errors between actual and estimated objective function values as colors of dots. Here, it often happens that several explanatory variables affect to the errors while other variables do not affect. Selection of explanatory variables to be assigned to the X- and Y-axes is therefore important for effective regression error analysis. We applied AIC to evaluate the relevancy between the errors and explanatory variables as a preprocessing. As a result, we could easily discover explanatory variables which are well related to the errors, and effectively visualize the distribution of samples which have large or small errors.

This paper introduced a case study using a retail sales dataset.

Our potential future issues are as follows. Firstly, we would like to improve the evaluation of explanatory variables. We are simply applying AIC in the current implementation; however, AIC has many remaining issues. Especially, AIC does not always have good properties to non-linear problems. We would like to apply improved method and compare how we can effectively select the explanatory variables. Also, we would like to other metrics for selection of explanatory variables. For example, it is often effective to measure the concentration of dots which have large errors in the display space. We are currently implementing this measurement.

It is also important to discover values of categorical variables which are well associated to samples which have large errors. We expect that reasons of errors can be more clearly explained if we develop effective methods to discover the relationships between errors and categorical variables.

References

1. E. van Herpen, E. van Nierop, L. Sloot, The Relationship between In-store Marketing and Observed Sales for Organic Versus Fair Trade Products, Market Letters, 23(1), 293-308, 2012.
2. T. Chen, A. Kalra, B. Sun, Why Do Consumers Buy Extended Service Contracts, Journal of Consumer Research, 36, 611-623, 2009.
3. T. Muhlbacher, H. Piringer, A Partition-Based Framework for Building and Validating Regression Models, IEEE Transactions on Visualization and Computer Graphics, 19(12), 1962-1971, 2013.
4. J, Krause, A. Peter, E. Bertini, INFUSE: Interactive Feature Selection for Predictive Modeling of High Dimensional Data, IEEE Transactions on Visualization and Computer Graphics, 20(12), 1614-1623, 2014.
5. C. Turkay, F. JEanquartier, A. Holzinger, H. Hauser, On Computationally-Enhanced Visual Analysis of Heterogeneous Data and Its Application in Biomedical Informatics, Interactive Knowledge Discovery and Data Mining in Biomedical Informatics, 117-140, 2014.
6. Y. Lu, R. Kruger, D. Thom, F. Wang, S. Kochi, T. Ertl, R. Maciejewski, Integrating Predictive Analytics and Social Media, IEEE Visual Analytics Science and Technology (VAST), 193-202, 2014.
7. 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.
8. H. Suematsu, Y. Zheng, T. Itoh, R. Fujimaki, S. Morinaga, Y. Kawahara, Arrangement of Low-Dimensional Parallel Coordinate Plots for High-Dimensional Data Visualization, International Conference on Information Visualisation (IV2013), 59-65, 2013.
9. Y. Wang, L. Luo, M. T. Freedman, S.-Y. Kung, Probabilistic Principal Component Subspaces: A Hierarchical Finite Mixture Model for Data Visualization, IEEE Transactions on Neural Networks, 11(3), 625-636, 2000.
10. R. Fujimaki, S. Morinaga, Factorized Asymptotic Bayesian Inference for Mixture Modeling, International Conference on Artificial Intelligence and Statistics (AISTATS), 400-408, 2012.
11. R. Eto, R. Fujimaki, S. Morinaga, H. Tamano, Fully-Automatic Bayesian Piecewise Sparse Linear Models, International Conference on Artificial Intelligence and Statistics (AISTATS), 238-246, 2014.