Statistical Graphics for Aggregated Symbolic Data

Yamamoto, Yoshikazu
Tokushima Bunri University, Department of Computer Science and Electronic Engineering
1314-1 Shido
Sanuki, Kagawa 769-2193, Japan
E-mail: yamamoto@fe.bunri-u.ac.jp

Nakano, Junji
The Institute of Statistical Mathematics, Department of Data Science
10-3 Midoricho
Tachikawa, Tokyo 190-8562, Japan
E-mail: nakanoj@ism.ac.jp

Introduction

Symbolic data can have not only real values and categorical values but also complicated values such as intervals, histograms and distributions (Billard and Diday, 2006). They are often generated by "aggregation" operations, which are used to summarize a group of observations by small number of statistics. It is clear that an interval has more information than a mean value and a histogram has more information than the interval. Although the required number of statistics increases a little, they are considerably small compared with the number of the record of the original observations. This is especially useful when the number of the original observations is huge.

Statistical graphics to visualize traditional data are indispensable for grasping the characteristics of the data intuitively. The situation is same for symbolic data. Thus, some visualization techniques of symbolic data have been realized (Diday and Noirhomme-Fraiture, 2008). In this paper, we propose new statistical graphics for symbolic data and some interactive operations to create symbolic data from individual observations by aggregations. We extend a parallel coordinate plot and a scatter plot to visualize symbolic data. We use colors and interactive operations to search adequate groups of observations by aggregation. We use Jasplot (**Ja**va based **s**tatistical **plot**) software (Nakano, Yamamoto and Honda, 2008) to realize our graphics.

Symbolic data and statistical graphics

Statistical data are given by recording the values of variables for individual observations. Each record of traditional data is given by a real value or a categorical value. We usually illustrate them as a table of two dimensions, in which rows express individuals and columns express variables.

A parallel coordinate plot (Inselberg, 1985) is often used to visualize such a table of data. It draws parallel axes to represent variables, and displays each individual by polygonal lines connecting the values of variables located on the corresponding axes. If the number of individuals is small, this static graphic is useful to grasp the characteristics of the data. However, if the number of individuals is large, it becomes difficult to identify each individual, because polygonal lines corresponding to individuals are overdrawn many times and may cover whole area. It is well known that the difficulty is improved by several interactive operations such as selection and highlighting (Unwin, Theus and Hofmann, 2006).

Typical Symbolic data express the distribution information by intervals or histograms as variable values. If we use a table of two dimensions to display them, each component specified by a row and a column contains the distribution information, for example, by two numbers for an interval.

As symbolic data can be considered to be an extension of traditional data, it is natural to

display them by using a parallel coordinate plot with some modifications. Because symbolic data can include more complex values than traditional data, statistical graphics for symbolic data must be more complicated than graphics for traditional data. Interactive operations must be in the same situation.

We also need to consider that many symbolic data are generated by aggregation of traditional statistical data to reduce the amount of information that we have to keep. If we have huge amount of data, for example, more than million individuals, it is natural to aggregate them into reasonable number of groups which has some proper meanings. We often decide groups by using values of several categorical variables. Individuals which take the same values of categorical variables are considered to belong to the same group. The groups are sometimes defined naturally by the domain knowledge and easy to determine. However, when we have little knowledge for the data, groups need to be defined by descriptive data analysis including visualization. In such a situation, it is important that statistical graphics can support this group determining process by interactive operations. For example, it is preferable that we can easily aggregate individuals into some groups as symbolic data and draw graphics for them.

We designed and implemented our statistical graphics by extending traditional graphics such as a parallel coordinate plot and a scatter diagram to symbolic data, and equipped interactive operations for generating several levels of symbolic data.

Extensions of a parallel coordinate plot and a scatter diagram for aggregated symbolic data

We propose to show aggregated symbolic data in the same framework of graphics for expressing each value of an individual observation, such as a scatter diagram and a parallel coordinate plot to compare them with the original data.

A parallel coordinate plot and a scatter diagrams are especially suitable to display real valued variables by one point on the corresponding axis or plain. For real valued variables, aggregated symbolic data are usually expressed by intervals or histograms which are used to illustrate distributions of variable values in a group of individuals. It is natural to show them on the same or similar graphics used for original individual data. Most important information is summarized by single value such as a mean or a median. In our graphics, they are treated as individual observation values of symbolic data. An interval and a histogram have also scale or variability information as the next important information of the distribution. As an interval contains just two numbers, location and scale information are all information contained in it. A histogram includes more information about the distribution than location and scale. We display the distribution of each group of original data by colored rectangular objects on the graphic plane. We express a histogram by using several rectangles arranged sequentially, in which each rectangle expresses a bin of a histogram and the thickness of color of rectangle is proportional to the height of a bin of the histogram, for example, a thick color is used for a large value and a thin color is for small value. We plot these rectangles for expressing symbolic data to avoid overlapping. To display many groups at the same time, rectangles for expressing histograms may be drawn slimly or as a line segment. In a scatter diagram, we draw histograms for horizontal and vertical axes following their directions. We may change the size of rectangles to avoid too much overlapping in a scatter diagram.

We have a mechanism to change the definition of groups interactively by mouse operations (Yamamoto and Nakano, 2010). Groups are usually specified by categorical variables. Individuals which have same values of specified categorical variables are in the same group. We can easily change the group by specifying different categorical variables interactively. By using this function, we can search adequate groups from large amount of data. As usual, highlighting the selected group and linking among graphics are also important. We have implemented these features in our experimental

graphics by using Jasplot library.

Examples of our graphics

We have modified a parallel coordinate plot to realize flexible aggregation operations. The results of aggregation are expressed by several visual objects to illustrate aggregated symbolic data in several ways.

Figure 1 is an example to show the mean value of each group. We use 2004 Cars Data (Unwin, Theus and Hoffman, 2006). In this figure, the original individual data are divided into 18 groups according to the values of two variables: `Type` and `Drive`. `Type` takes 6 values (`Pickup`, `Mini Van`, `Wagon`, `SUV`, `Sports Car` and `Sedan`) and `Drive` takes 3 values (`front`, `rear` and `AWD`). The values of `Drive` are stacked first and the values of `Drive` are also stacked for each value of `Type`. Note that 3 groups contain no individual and has no polygonal line.

In this graphic, we have three areas: ignored variables area, aggregation specification area and data description area, from the left to the right. We can move variables to or from each area freely by the drag-and-drop operation using a mouse.

Variables placed in the ignored variables area are completely ignored from drawing polygonal lines.

The aggregation specification area is a place where categorical variables are used to specify groups. Their values are put here as stacked boxes. Units of aggregation are specified by the Cartesian product of values of these variables.

In the data description area of this figure, mean values of each group are displayed just like a usual parallel coordinate plot. When we perform aggregation based on the group specified in the aggregation specification area, one summary statistics of aggregated data are first displayed. Default aggregation statistics are the means of groups displayed by polygonal lines. Another available statistics are medians.

In addition, more detailed aggregation results are available by boxplots (Figure 3 and Figure 5) or histograms (Figure 2 and Figure 4) on each axis. It is clear that the visibility becomes worse when we draw all the resulted graphics for all groups at the same time. Thus, we can select groups whose resulted graphics are shown. We note that our system has multiple selectors and can specify several groups at the same time.

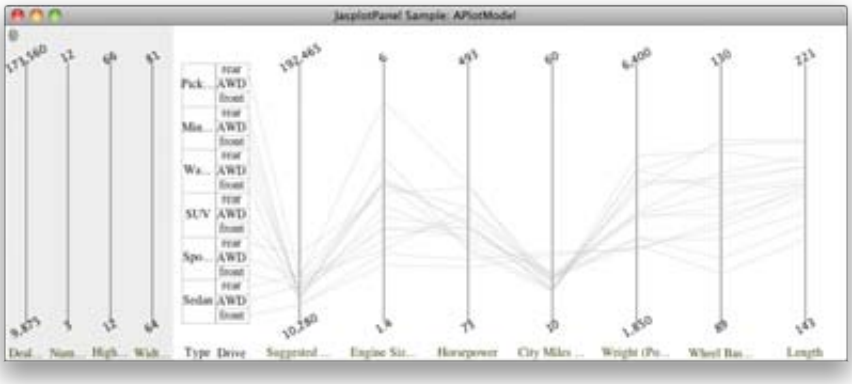

Figure 1: Displaying mean of groups

Figure 2 shows histograms by color rectangles connected by polygonal lines. The polygonal lines express each group and their mean values. The color rectangles display the distribution information of groups. Selected groups by specifying stacked boxes in the aggregation specification area are

highlighted in different colors. In this figure, three groups are specified in red, blue and green. For example, green group is specified by `AWD` value of the variable `Drive` and `Wagon` value of the variable `Type`. We can see the distribution information of groups in detail and find several modes of distributions for each group.

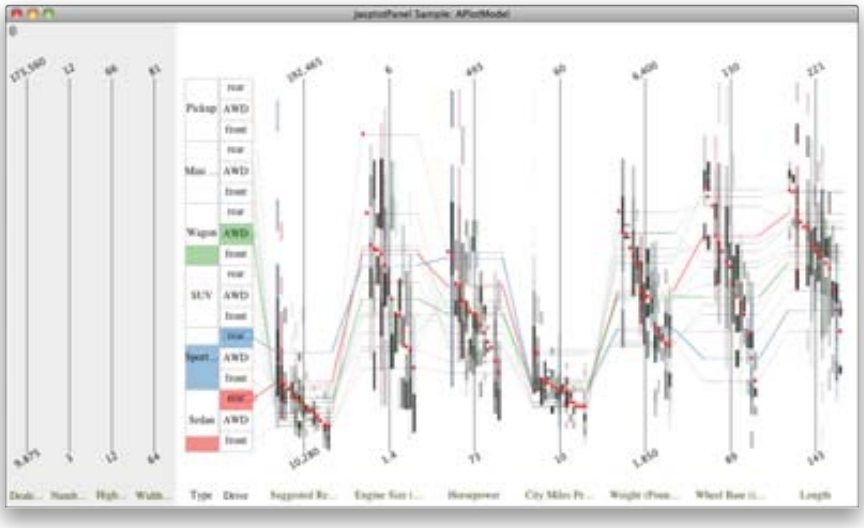

Figure 2: Displaying "histograms" of groups

Figure 3 shows information of groups by boxplots and polygonal lines. We use the same definition of groups and select same groups as Figure 2. In this figure, we can check outlier individuals of groups.

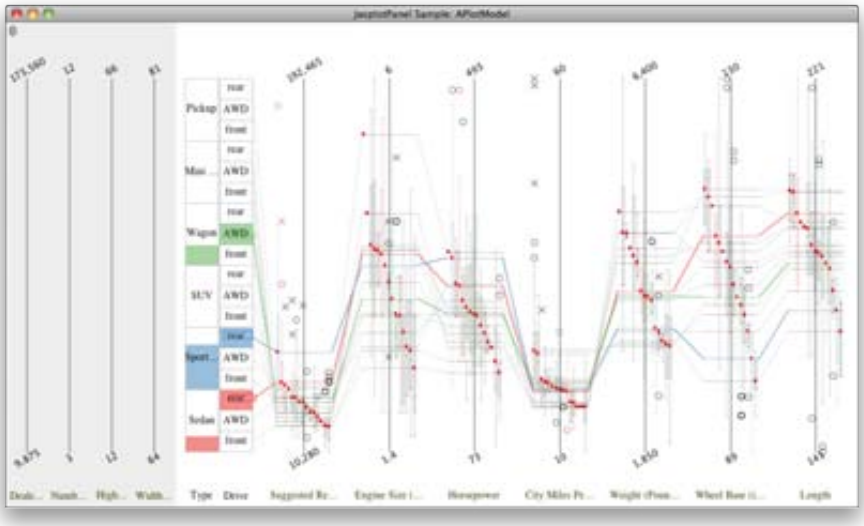

Figure 3: Displaying boxplots of all groups

Figure 4 draws usual histograms on axes connected by polygonal lines. In this figure, distribution information of selected groups is displayed by histograms with the highlighted color. We can see the distribution information in detail for focused groups and can compare them clearly. Unfortunately, if they overlap, colors are mixed and it is not easy to distinguish them.

Figure 5 draws boxplots and the polygonal lines for selected groups. It is easy to see compared with Figure 3.

Figure 6 shows examples of scatter diagrams for aggregated symbolic data together with tra-

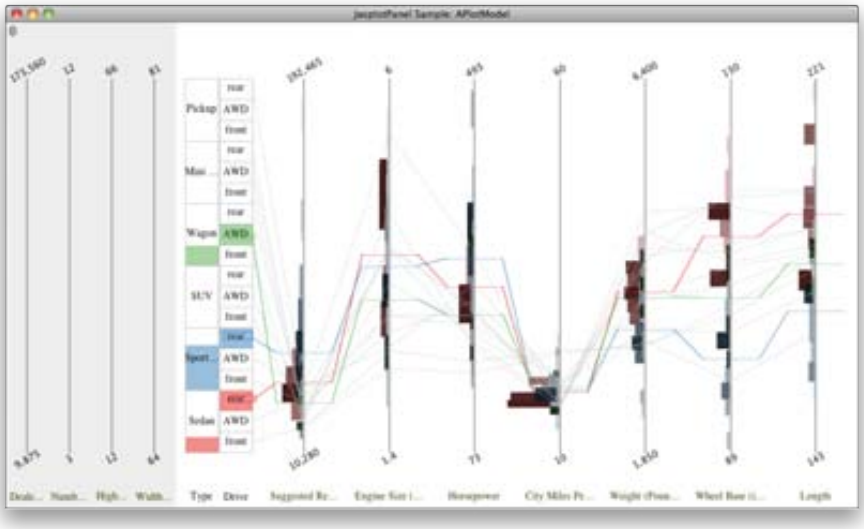

Figure 4: Displaying histograms of selected groups

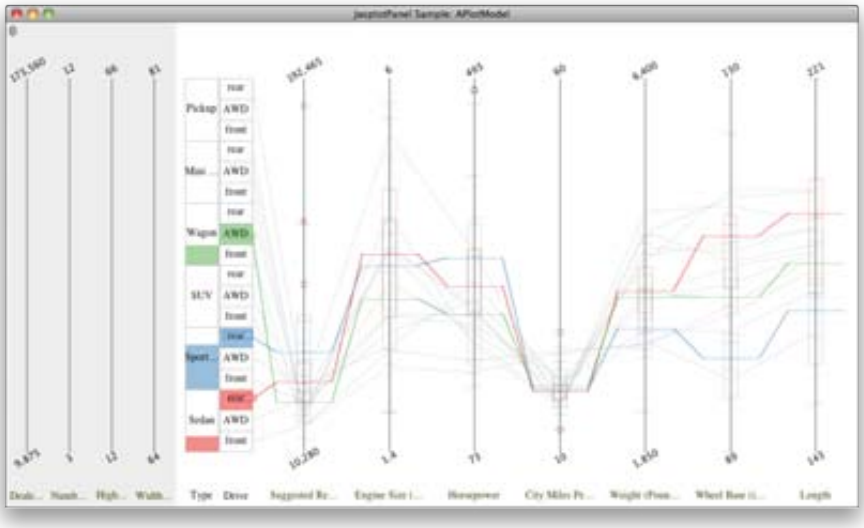

Figure 5: Displaying boxplots of selected groups

ditional scatter diagrams. In this figure, the aggregation specification area is shown separately as a window located in the left side. If we select groups by using stacked boxes in this window, scatter diagrams in the upper right part show the aggregation results. Two groups are selected and two scatter diagrams show the information of these symbolic data. One scatter diagram shows variables `Dealer Cost` and `Engine Size`, and another shows variables `Weight` and `City Miles Per Gallon`. As we have two variables in a scatter diagram, their histograms are expressed by two rectangle belts which cross orthogonally. We can see the detailed distribution information of groups in which the location information is shown by the center of the cross.

Concluding remarks

If the number of individual observations is huge, even a high-speed computer requires considerable time to perform interactive operations on statistical graphics. Aggregation operations can be one remedy for this trouble. Aggregation means to summarize the information in a group of traditional

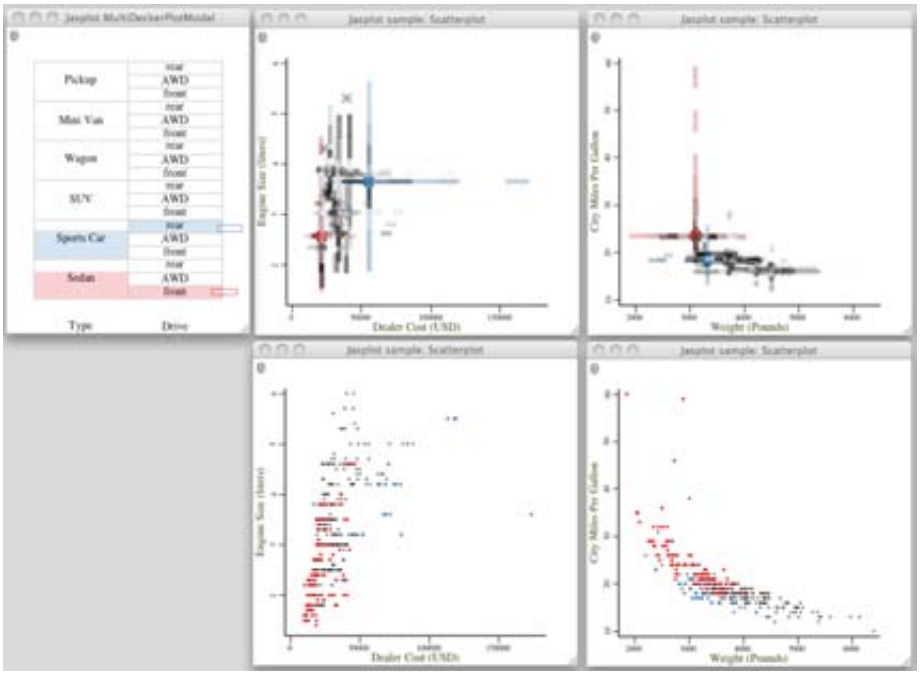

Figure 6: Displaying extended scatter diagrams for symbolic data

statistical data by the less number of values than all the variable values for observations.

Our experimental implementation of an extended parallel coordinate plot and scatter diagram seems to be useful. Obtained groups can be analyzed by using symbolic data analysis techniques.

REFERENCES (RÉFERENCES)

Billard, L. and Diday, E. (2006): *Symbolic Data Analysis: Conceptual statistics and data mining*. John Wiley, New York.

Diday, E. and Noirhomme-Fraiture, M. (2008): *Symbolic Data Analysis and the SODAS Software*. John Wiley, New York.

Inselberg, A. (1985): The plane with parallel coordinates. *The Visual Computer 1*, 69–91.

Nakano, J., Yamamoto, Y. and Honda, K. (2008): Promming statistical data visualization in the Java language. In: Chen, C-H., Hädle, W. and Unwin, A. (Eds.): *Handbook of Data Visualization*. Springer, Berlin, 725–756.

Unwin, A., Theus, M. and Hofmann, H. (2006): *Graphics of Large Datasets: Visualizing a Million*. Springer, Berlin.

Yamamoto, Y. and Nakano, J. (2010): Data Visualization and Aggregation. In: Y. Lechevallier, Y. and Saporta, G. (Eds.): *Proceedings of COMPSTAT'2010: 19th International Conference on Computational Statistics*. Physica-Verlag, Berlin, 1437–1444.