The Structure of DAVIS: Java beans approach

Moon Yul Huh, Kwang Ryoel Song
Department of Statistics
Sungkyunkwan University
Seoul, Korea

Davis is a stand alone JAVA-based application package, and has been implemented to make the most of visualization techniques. The key feature of DAVIS is the real-time interaction between the modules. When we modify a portion of data in a specific module, or when we change the value of parameters of a statistical module, this effect is propagated to all the other modules in real time. Our talk will be in 4 parts: survey of the visual data mining tools, overview of DAVIS, implementation of the data flow among the modules of DAVIS, and prospective future works.

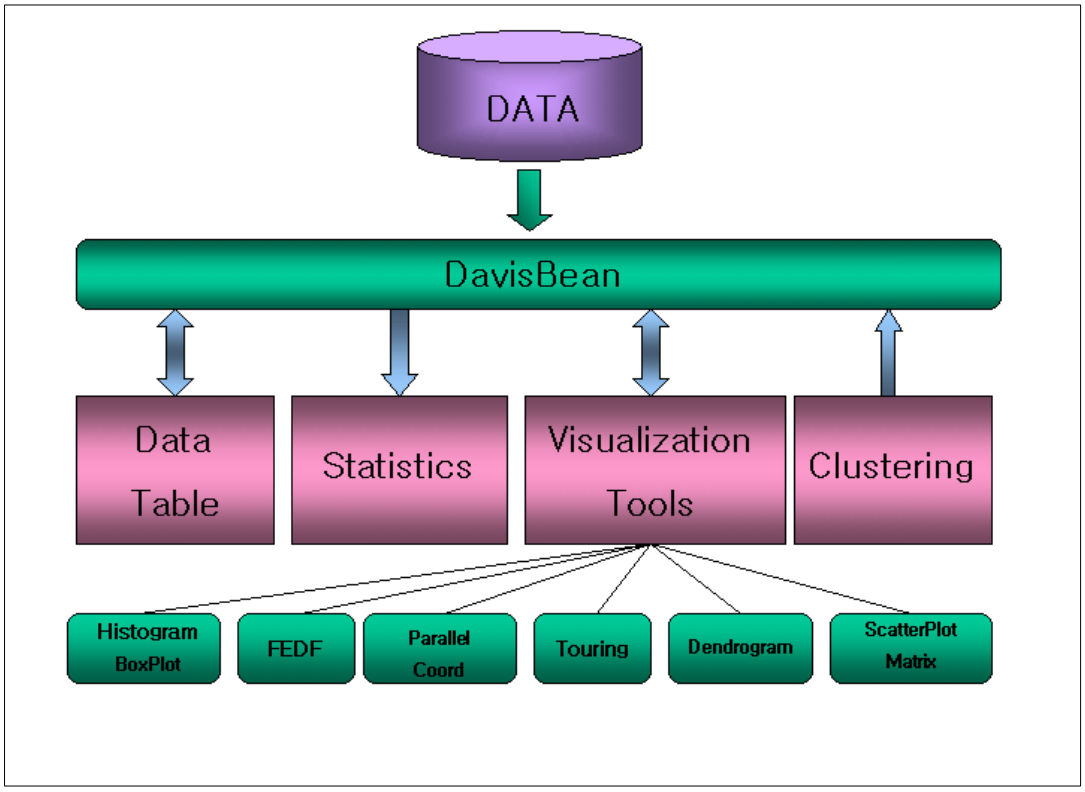
1. Survey of current visual data mining tools

Product	Graphics
CViz	include touring/JAVA
SAS/SPSS	Extensive use of ScatterPlot matrix, GUI with drag & drop
XGobi	Grand Touring, Parallel coordinates, ScatterPlot Matrix
3DV8	Planet view, Nails view, Weights view, Cones view, Landscape view
DAVIS	FEDF, Parallel coordinates, Dendrogram, Grand Touring Tracking Grand Touring

2. Overview of DAVIS

Main window of DAVIS looks as follows.

The architecture of this system is given in the following figure.



The functions of each module are as follows.

DATA: load data from a file or by pasting it from any editor in ASCII code.

Data Table: Variables selection, instances selection, sampling, sorting the data with respect to a specific variable in descending or ascending order by mouse clicking on the right of the variable label.

Statistics: descriptive statistics, correlation, covariance and principal components parts. Principal components part gives scree plot of the components and the scatterplot matrix of the principal components

Clustering: k-means and divisive clustering methods

Visualization tools:

1. Histogram: Each histogram has number of bins button to control the number of bins for the plot. Also, when we divide the data set into several groups either by clustering or by mouse brushing, each bin will be divided into the number of groups having different colors corresponding to each group
2. Boxplot: When we divide the data set into several groups either by clustering or by mouse brushing, the operation is propagated to boxplot by dividing the plot into several boxplots corresponding to each group.
3. Scatterplot matrix: We can draw the plot in 3 different styles: upper, lower and both. There are some packages that attach 1-dimensional marginal plots like histograms or dot plots to the scatterplot matrix. For Davis, we did not consider attaching 1-dimensional marginal plots to the scatterplot matrix because this is available by drawing 1-dimensional plots separately and since these plots are linked to the scatterplot matrix dynamically.
4. Touring: "TGT/GT" button gives the option of drawing TGT or Grand Tour. Also, here is a button for start/stop that controls starting or stoping touring. Currently, the mouse brushing operation is not implemented for this plot.
5. Dendrogram: this has Agglomerating and Distance buttons. Agglomerating button gives 4 options: Nearest, Farthest, Average, and Group Average methods for agglomeration. Distance button gives 3 options: Euclidean, Standard Euclidean, and City-block distance measures. Because of the window size, there is a limitation to represent all elements into dendrogram when the size of data is more than several hundreds. We implemented scroll-bar feature to view those parts of the elements that are outside of the current window.
6. Parallel coordinates: this has Direction and Arrangement buttons. direction button provides the way whether the plot is drawn horizontally or vertically. We can arrange the variable in 3 different methods; data base arrangement component method arrangement, permutation arrangement.

3. Implementation of the data flow

The basic idea of real time information flow among DAVIS is based on the linking method. Linking is available by mouse operation. When this mouse operation is performed and subset of data points are brushed, the index array of the corresponding subset of data points are manipulated and propagated to the beans that are related. The beans are constantly listening to the change of these operations by thread property, and when a bean receives this new index array, relevant operation is performed immediately.

3.1 Dynamic linking features of the visualization tools.

Selecting, deleting, focusing and identifying a subset of data points are available in most of the plots of DAVIS. All the operations are performed using a series simple mouse operations. The result of this operation is linked to the other beans in real time. Currently we have not finished this operation with histogram, boxplot and TGT.

3.2 Dynamic linking features in Data Table

In Data Table, we can select a portion of variables or observations, and this selection will be propagated to the other beans. When a portion of data is selected from the other beans, this selection is realized in the Data Table by highlighting those instances in the Table.

3.3 Dynamic linking features of Statistics and Clustering Modules

The result of dynamic operation in the other modules are immediately propagated to Statistics Module. For Clustering module, the selection of the clustering method or the number of clusters is dynamically linked to other modules.

4. Prospective future works

Dynamic linking among the beans are realized through redrawing the relevant plots when a change in data points is processed in one plot. We implemented dynamic linking operations using JAVA thread property. Java threading consumes a lot of

computing time, especially when we are dealing with drawing plots with lots of data points. Hence, there should be some way to get around this problem. We hope this problem can be partially alleviated by parallel computing.

Another problem happens when we want to restrict the dynamic linking to specific plots. In other words, when we are selecting a portion of data points using FEDF, we want the change to be dynamically linked to only scatterplot matrix, and want other plots not to be affected by the selection. We are currently working on implementing this feature.