SplatVis: a Data Analysis and Visualization Tool for Proteomics Data

Don Johann, Michael McGuigan, Stanimire Tomov

SplatVis is a new data mining and visualization tool for proteomics data. The tool development is result of a NIH-BNL cooperation in the development of a toolkit for visualization and data mining of proteomic datasets for early cancer detection (see also [1, 2]). We developed data mining techniques specific to our problem area but also use data mining tools from the MLC++ library. The visualization tools that we developed use the VTK library, and the GUI is done in Tcl/Tk.

We use a "splatting" technique to interactively visualize proteomics data of very large size (currently 1.5 GByte). The technique splats points into a volume on a regular grid using an elliptical, Gaussian distribution.

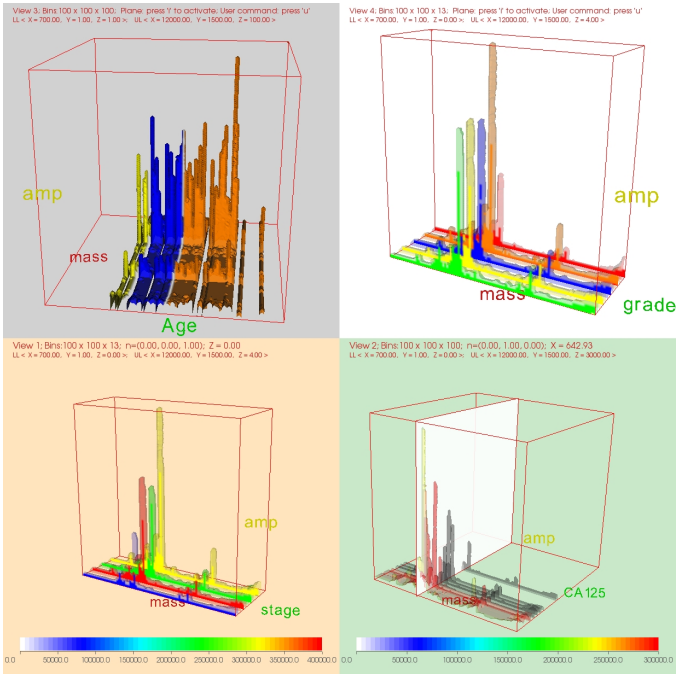

Figure 1: Multiview visualization of proteomic data data.

In our implementation we extended the VTK Gaussian Splatter for the specifics of the proteomics data mining. We developed user interface and visualization of the splatted data including volume visualization and isosurface extraction, cutting planes, colorbar, multi-view, and various user interactions (mouse, key-strokes, and user defined commands). Figure 1 illustrates some of the features mentioned. The user interface is implemented in Tcl/Tk. Figure 2, Left gives the control panel for the visualization on Figure 1. The number of views controlled by the panel is dynamically changed, based on the program input. The input is "schema file" for each proteomic study. The schema file allows for the proper mapping of mass spec data and associated clinical parameters to relevant visual elements in the visualization tool. Figure 2, Right gives a view of the interface developed to create, parse, and pass the schema file to the visualizer.

We tested several of the MLC++ inducers in creating cancer/non-cancer decision trees based

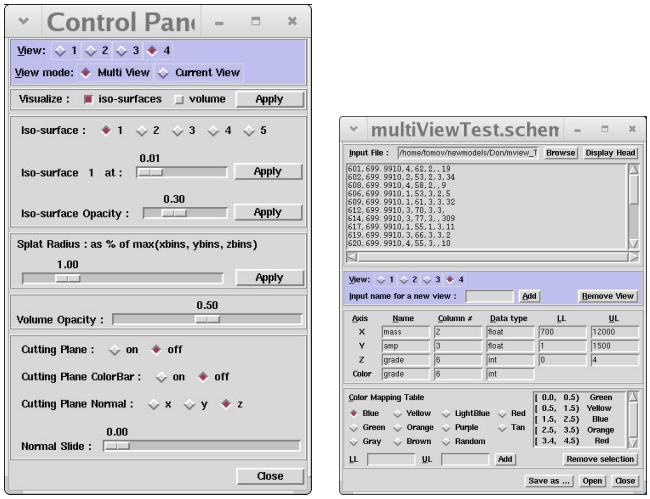

Figure 2: Tcl/TK GUI. Left: the control panel. Right: .schema constructor/reader.

on proteomic data sets. We created Tcl/Tk user interfaces (see Figure 3) to easily use and test the various inducers that MLC++ provides. In particular, our Tcl/Tk scripts allow the user to input data sets that are not in MLC++ format, choose inducer, and build the decision tree using MLC++ routines. The result can be viewed using "dotty", which is a customizable graph editor for the X Window System. The dotty editor is written on top of a preprocessor, called "dot", for drawing directed graphs. Our VTK decision tree visualization uses dot to get the space coordinates of the tree nodes (Figure 3,Right).

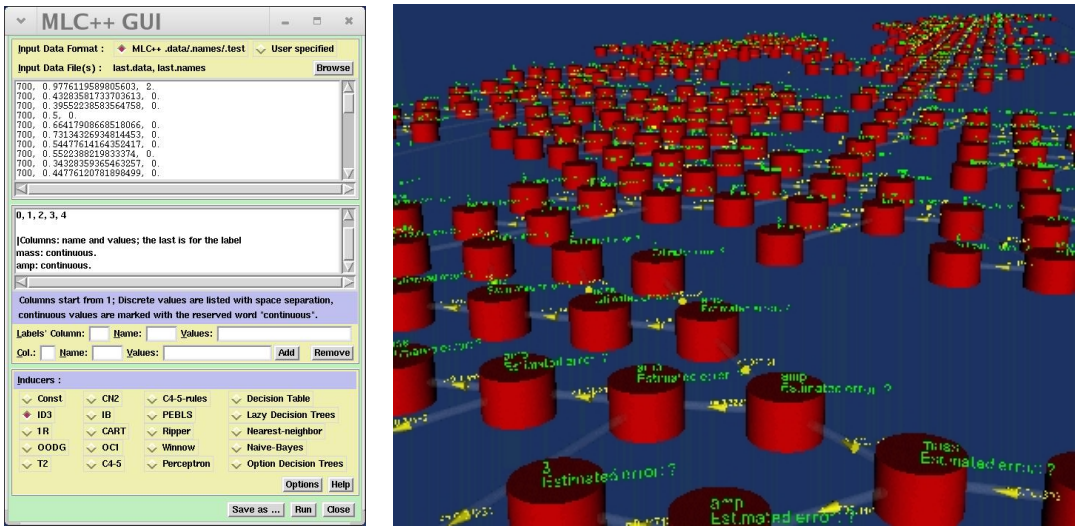

Figure 3: Left: Tcl/TK MLC++ GUI. Right: decision tree visualization.

References

[1] D. Johann, M. McGuigan, S. Tomov, E. Blum, G. Whiteley, E. Petricoin, L. Liotta, *Toward a Systems Biology Software Toolkit*, 17th IEEE Symposium on Computer-Based Medical Systems, June 2004.

[2] D. Johann, M. McGuigan, A. Patel, S. Tomov, S. Ross, T. Conrads, T. Veenstra, D. Fishman, G. Whiteley, E. Petricoin, L. Liotta, *Clinical Proteomics and Biomarker Discovery*, Annals of the New York Academy of Sciences, Vol. 1022, June 2004, pp. 295-306.