Explorations in Visual Data Mining

Research Question

In the research, we study the design issues of constructing a comprehensive system for performing visual data mining. Specifically, we address the following questions:

1. What is a minimal set of easy-to-use, yet powerful, visualizations necessary for conducting a thorough preliminary exploration of the data?

   Existing information visualization tools either (1) provide visualizations that simultaneously present numerous dimensions (many more than three) from a dataset, or (2) include numerous charting options that tend to overlap in functionality while unnecessarily increasing the learning time and complexity of the tool. We are looking for a minimal set of visualizations that fully supports the analyst in completing all required tasks of a preliminary exploration. Hence, of necessity, we also answer the question: “What are those tasks?”

2. What operations are required for manipulation of the datasets before and during analysis and how can those operations be semantically represented and implemented within the visual presentations?

   We consider the basic select, project, and join relational operators. In addition, we consider the drill down/up operators of OLAP, support for user defined data transformations, and data transformations using statistical techniques.

3. What data mining methodologies should be included in the system and how can the models created by these methodologies be visualized to facilitate user comprehension and interpretation?

4. What communications and corresponding protocols are needed to implement the required synchronization of processes, visualizations, and user interactions across multiple computing systems?

   The system, as planned, requires multiple displays for presentation of results and models. It also requires multiple processors to drive the visualizations and to execute the back-end data mining methodologies.

Approach

We answer the above questions by building a prototype. The prototype system consists of three primary components:

1) the control center,
2) the data and model viewers, and
3) the model processors.

Each instantiated component may be executed on either a dedicated or shared processor. A more detailed overview of the design architecture is found at visMining.org.
As much as possible, all functions of the system are implemented via a “direct manipulation” interface.
Also, where possible, to better utilize available display real estate and to facilitate interpretation, data
and model viewers have a 3-d presentation.

In implementing the data viewers, we develop answers to questions one and two above. The control
center design and implementation provides insight into requirements with respect to the data
transformations of question two. The coordination requirements of the control center also steered the
design of the communication protocol between itself, the viewers, and the modelers (question 4).

Question three (methodologies to include) was initially answered by considering the incorporated
methodologies of other data mining systems. In building the prototype, we have taken a two pronged
approach. First, some of the chosen methodologies have been implemented in-house. These include
artificial neural networks for classification and regression, a support vector machine for classification,
and a self-organizing map (SOM) for clustering. Second, we have implemented drivers for the Microsoft
Business Intelligence tools – allowing us to use our system as the front-end presentation and the
Microsoft tools for back-end processing. In the future, we plan to add drivers for other data mining
systems as well.

Contributions

Although the prototype is still under construction, contributions to date include:

1. Design of a three component architecture for visual data mining – the control center, the viewers,
   and the modelers – and the subsequent definition of a communication protocol for synchronization and
   coordination between the components.

2. Design of a 3-d “parallel coordinate plot” (PCP) tool that visually facilitates the assessment of data
distributions, the identification of distinctive sub-populations within a dataset, and the search for
outliers.

3. Design of an interactive “correlation viewer” which, when synchronized with a “scatter plot”, allows
   the user to quickly identify and isolate dataset attributes of interest for a more in-depth analysis.

4. Design of a 3-d SOM clustering tool which in coordination with the PCP tool allows the users to
   quickly evaluate the contents, structure, and quality of a dataset clustering.

Status of Manuscript

The prototype has been under development for almost four years. Most of the time spent has been
devoted to the design and implementation of stand-alone viewers and modelers. Previous papers from
the project, have presented individual contributions of the viewers. In the proposed paper, we
summarize these previous contributions, then plan to describe the architecture, the coordination
protocol, and the overall contributions of the system. These sections have not yet been written.
Contact Information:

Russell K Anderson
russ@visMining.org
806-570-7257