Poster D-7

Exploratory analysis of multi-modal data including breast cancer microarray data

Authors:
Christian Martin (*Technical Faculty, Bielefeld University*)
Harmen grosse Deters (*Technical Faculty, Bielefeld University*)
Tim W. Nattkemper (*Technical Faculty, Bielefeld University*)

Short Abstract: Data analysis in modern biomedical research has to integrate data from different sources, like microarray, clinical and categorical data, so called multi-modal data. The reef SOM, a metaphoric display, is applied and further improved such to allow the simultaneous display of biomedical multi-modal data for an exploratory analysis.

Long Abstract:
In modern biomedical research huge amounts of data are produced by a continuously increasing number of technologies. In the analysis of this data for a set of subjects, i.e. patients, regarding its hidden relationships to clinical outcome the researcher is confronted with the difficulty how to analyze the collected multi-modal data set. While several data mining techniques have been proposed to analyze data sets from single sources (for example laboratory tests, microarrays) there is a growing demand for tools for an integrative analysis of all data. The researcher is not only interested in the analysis of data produced by every single technology, but needs to analyze all available data at once to discover new knowledge which can be employed in clinical practice (for instance in diagnosis or planning / monitoring of treatment). In recent clinical studies concerning cancer research, the available data often consists of clinical, categorical and gene expression data from microarray experiments. The clinical data may contain age, weight or sex of the patient, the size or the grade of the tumor, information about the lymph nodes, or results from a histological analysis. The categorical data may contain a classification of the tumor regarding its malignancy. This might be the patient's survival time after the analysis or the success of chemotherapy. The genomic data may consist of up to 25,000 genes, which can be analyzed simultaneously by modern microarray technology. The major challenge is how this multi-modal data consisting of clinical, categorical and genomic data can be analyzed in an integrative manner. This problem is further increased by the high dimensionality of the data. Usually the number of available experiments (number of samples) is approximately of the same magnitude or even smaller than the number of genes (dimensionality of data space) analyzed, which makes the application of statistical test methods impracticable. Many machine learning methods can handle such difficulties, but most of them need the adjustment of parameters and are based on pairwise similarities, which cannot be defined appropriately for multi-modal data. Considering all these aspects makes a more interactive, exploratory data analysis seem more reasonable. To allow an exploratory study, the multi-modal data, which is usually distributed in several media (tables, flat files) must be integrated into one representation, combining visualizations of all kinds of available data. A simultaneous visual inspection of all modalities enables the detection of patterns and structure in the data when browsing through and zooming into the image. We propose an integrative multi-modal visualization approach based on the Self-Organizing Map (SOM). The basic idea is to render

a multi-modal visualization to display multi-modal data. We design a visualization consisting of dimension reduction and multivariate object display, i.e. data glyphs. The SOM algorithm comprises the aspects dimension reduction, clustering and visualization and is well suited for the analysis and visualization of the microarray data. Thus one data modality, the microarray data is fed into the SOM. Displaying the trained SOM with the U-matrix approach visualizes structural features of the high dimensional microarray data space. We expand the visualized U-matrix by introducing multivariate data glyphs in order to display clinical and categorical data. Using a metaphoric display approach the SOMs U-matrix is rendered as an underwater sea bed with color and texture. This underwater landscape is then completed with glyphs generated from clinical and categorical data. In this work we use a kind of metaphoric glyph, a so called fish glyph. The fish glyphs have two groups of parameters that describe shape or colors. We use this to display clinical data by shape and categorical features with color. The resulting images are both informative and entertaining and can easily be interpreted by the biomedical researcher, since specific knowledge about the SOM algorithm is not required. Its visual inspection might reveal interesting structural pattern in microarray, clinical and categorical data. The reef SOM has the fundamental advantage that it is multi-modal itself, and thus especially well suited for the display of multi-modal data. The geology modus (U-matrix displayed as sea bed) is combined with a fauna modus (fish glyphs) or fauna modi (fish shape representing the clinical data and fish color representing the category). The user can direct his attention to the modus of his choice or to both. Additional modi allow the integration of data of further technologies, e.g. a flora modus might be introduced for displaying features of biomedical images (X-ray, CT, MRI). To illustrate the application and usefulness of the reef SOM for the exploratory analysis of biomedical data, results are presented for the van't Veer breast cancer data set.