Proceedings of the 3rd INFORMS Workshop on Data Mining and Health Informatics (DM-HI 2008)
J. Li, D. Aleman, R. Sikora, eds.

VISUAL ANALYTICS OF MINING HUMAN LUNG CANCER DATA

Qingyu Zhang
Department of Computer & Information Technology
Arkansas State University
Jonesboro, AR 72467
qzhang@astate.edu

Richard S. Segall
Department of Computer & Information Technology
Arkansas State University
Jonesboro, AR 72467

Abstract

This paper provides applications of data mining to health informatics as it illustrates how visual analytics tools can be useful in the analysis of gene-type microarray data for human lung cancer. The analysis of data is performed by using Megaputer PolyAnalyst® 5 and SAS Enterprise Miner™. The human lung cancer datasets are microarray databases consisting of 12,600 transcript sequences in 186 lung tumor samples as available on the web pages of the Broad Institute, which is a research collaboration of MIT, Harvard and its affiliated hospitals. A brief discussion of microarray databases are presented and their applications to bioinformatics research. This paper extends previous research of the authors by providing the actual visual analytics obtained in the mining of human lung cancer data. The visual analytics presented in this paper include link charts, decision trees, gene dimension matrices, regression scatter plots, tables of fit statistics, analysis of variance, SAS process monitor, cluster analysis and proximities, frequency distributions for clusters, self-organized maps and their associated importance rankings of factors. The data mining and visual analytics performed are insightful of new information and especially in determining the significant gene types and the association of specific gene types with human lung cancer.

Keywords: Visual Analytics, SAS Enterprise Miner, Megaputer PolyAnalyst, Human Lung Cancer, Data Mining

Background

According to the American Lung Association [3], lung cancer is currently the leading cause of death from cancer in the United States. According to Edgerton et al. [6], the therapies to improve outcomes in lung cancer lags behind those for the other types of cancer. Hence there is a great prevalent need for the development of novel techniques for the therapeutic remedies for the treatment of lung cancer such as may be revealed by those discussed in this paper using the techniques of data mining. Abdellatif [1] discussed leading the way using microarray as a more

comprehensive approach for discovery of gene expression patterns. Aitman [2] discussed science, medicine, and the future DNA microarrays in medical practice. Edgerton et al. [6] further discussed the applications of data mining for gene networks relevant to poor prognosis in lung cancer via the techniques of Backward Chaining Rule Induction (BCRI).

Visual Analytics

Visual analytics is the exploration and discovery processes that visually represent the information and allow the researcher to directly interact with the information, gain insight, and draw conclusions [7]. It is described as "the science of analytical reasoning facilitated by interactive visual interface" [9]. It is an iterative process involving data gathering, preprocessing, analysis, interaction and decision making. Two words of visual analytics present two aspects of this process: visual means a chart/diagram/image illustration that enable humans to perceive, relate, and conclude in the exploration process; and analytics means the analytical methods from data mining, statistics, and mathematics. It combines the automated analysis methods and interactive visual representation.

This paper provides applications of data mining to health informatics as it illustrates how visual analytics tools can be useful in the analysis of gene-type microarray data for human lung cancer. The analysis of data is performed by using Megaputer PolyAnalyst® 5 and SAS Enterprise Miner™.

Human Lung Cancer Data

The database used in this paper is from the Broad Institute [5]. The Broad Institute is a research collaboration of Massachusetts Institute of Technology (MIT), Harvard and its affiliated hospitals. According to its webpage, the mission of the Broad Institute is "to bring the power of genomics to medicine." The data set selected from the Broad Institute is one of those posted as available with unrestricted access as one of the web links posted on the web page of the Broad Institute Cancer Program Data Sets [5] and is that which is related to the "Classification of Human Lung Carcinomas by mRNA Expression Profiling" research project of the Broad Institute. The selected data base for this research was used by Bhattacharjee et al. [4] using oligonucleotide microarrays to analyze "mRNA expression levels corresponding to 12,600 transcript sequences in 186 lung tumor samples, including 139 adenocarcinomas resected from the lung."

Software

PolyAnalyst® 5 is a product of Megaputer Intelligence, Inc. that contains sixteen advanced knowledge discovery algorithms. It is an enterprise analytical system that integrates data, text, and web mining. PolyAnalyst® can load data from disparate data sources including all popular databases, statistical, and spreadsheet systems. In addition, it can load collections of documents in html, doc, pdf and txt formats, as well as load data from an Internet web source. Megaputer PolyAnalyst® has the standard data mining functionalities such as Categorization, Clustering, Prediction, Link Analysis, Keyword and entity extraction, Pattern discovery, and Anomaly

detection. These different functional nodes can also be directly connected to the web data source node for performing web mining analysis.

SAS Enterprise Miner™ is a product of SAS Institute Inc. and utilizes algorithms for decision trees, regression, neural networks, cluster analysis, and association and sequence analysis [8]. It also includes SAS® Text Miner as an "add-on" with the inclusion of an extra icon in the "Explore" section of the tool bar. SAS® Text Miner performs simple statistical analysis, exploratory analysis of textual data, clustering, and predictive modeling of textual data. SAS® Text Miner uses the "drag-and-drop" principle by dragging the selected icon in the tool set to dropping it into the workspace.

Visual Analytical Results Using Polyanalyst® 5

The input data window of PolyAnalyst® 5 is shown in Figure 1 for the Human Lung Project where each of the columns other than that for the description and accession are gene types. A Link Chart between the NL268 and AD179 gene types is shown in Figure 2. A decision tree analysis has been conducted as shown in Figure 3. The decision tree has 38 leaves, 37 non-terminal nodes, and 8 levels depth with a classification efficiency of 61.74%. A total classification error is 30.51% with the greatest classification error for bin 2 of 53.18% and the lowest classification error for bin 5 of 20.89%. Figure 3 shows the complete window of PolyAnalyst® 5 without this data tag visible for the AD115<2.38 node.

The gene dimension matrix for the human lung data is illustrated in Figure 4. Each column shows the magnitude and frequencies of each gene dimension. Figure 4 shows color coding and pie charts for each column of gene types. As can be seen from pie charts of Figure 4, the gene type NL279 has more diverse sets of magnitudes than that of NL268, AD268 and AD225 as also shown in this figure.

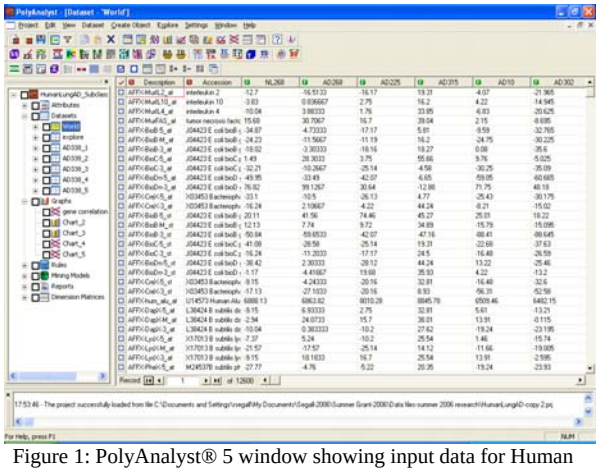

Figure 1: PolyAnalyst® 5 window showing input data for Human Lung Cancer Project

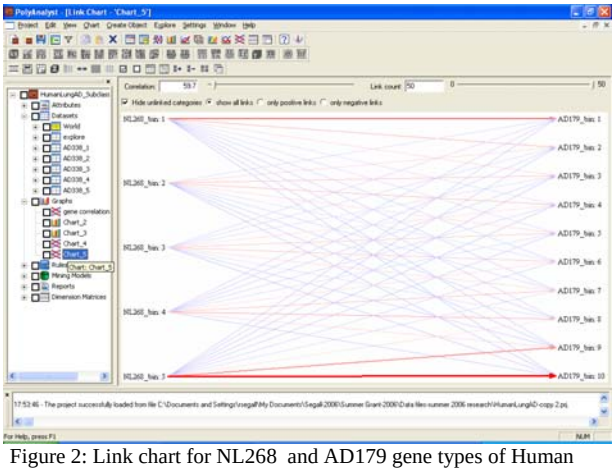

Figure 2: Link chart for NL268 and AD179 gene types of Human Lung dataset using PolyAnalyst® 5

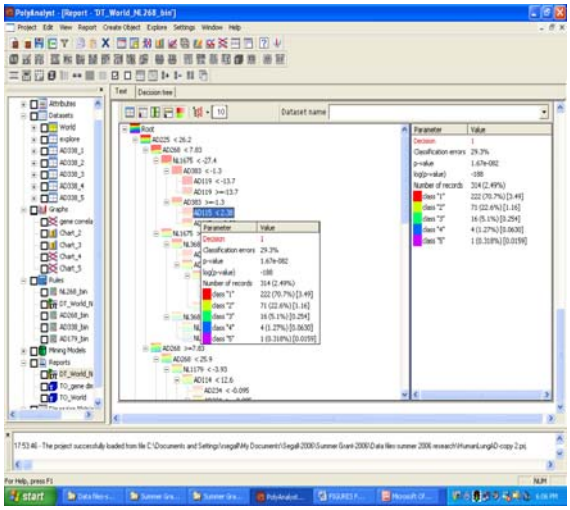

Figure 3: Decision Tree Report for NL268 gene type data bin of Human Lung Project using PolyAnalyst® 5

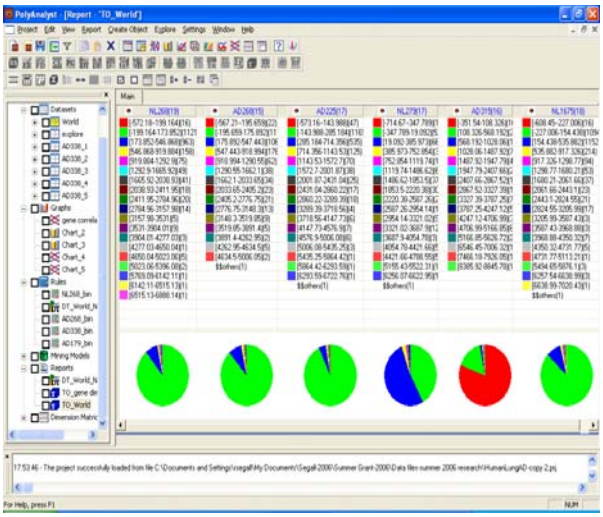

Figure 4: Percentile distributions of individual gene dimensions of Human Lung Project using PolyAnalyst® 5

Visual Analytical Results Using SAS Enterprise Miner^{tm}

The work space of SAS Enterprise Miner^{TM} is shown in Figure 5. Figure 6 shows the results of using data mining tool of decision trees with SAS Enterprise Miner^{TM}. Note that Figure 6 combines four views of results of the data mining using decision trees in a single window. As can be seen, training and validation were completed within 13 leaves of the decision tree using a sample of 5040 for the training and sample of 3780 for the training. Figure 7 shows a partial view of the decision tree diagram obtained by data mining using SAS Enterprise Miner^{TM} as specified for a depth of 6 from the initial node of NL279.

A regression scatter plot of predicted versus actual NL268 gene types for the human lung data is drawn and a linear fit is evident. The fit statistics for regression modeling of the human lung data for training with 106 model degrees of freedom, and 4934 degrees of freedom for error. The data mining has provided an extremely good modeling to the human lung data as indicated by the R-squared values of 97.81%. The Fit Statistics with Target Variable of NL268 effect using neural networks provides a similar result.

A cluster analysis is shown in Figure 8. As can be visualized is that the normalized means for the gene types listed are almost uniform with the exception of that for AD122 and AD043 gene types. Figure 9 shows the cluster proximities which indicates that several of the clusters have the same cluster proximities which is to be expected from the almost uniform normalized means as shown in the previous Figure 8 and also in the next Figure 10 of the frequency distributions of the clusters. Self-Organized Maps (SOM) was run providing results in the form of an interactive map that illustrates the characteristics of the clusters and importance of each variable. The numerical importance of the gene types with the most important gene types listed first ranging from about .53 for gene type AD327 to .09 for gene type AD114.

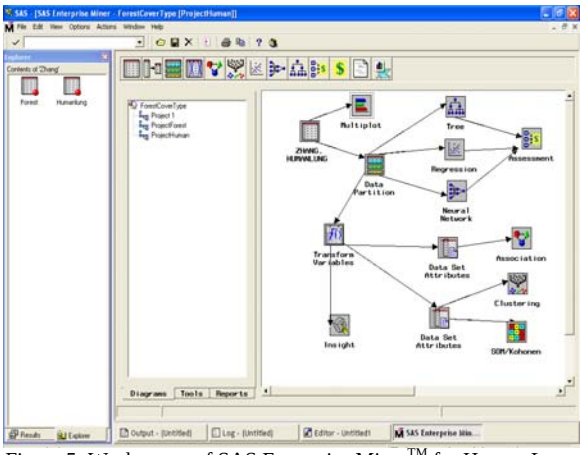

Figure 5: Work space of SAS Enterprise Miner™ for Human Lung Project

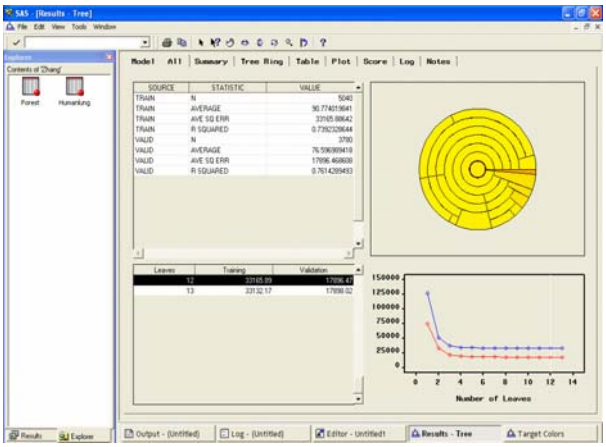

Figure 6: Results of Decision Tree analysis of human lung data using SAS Enterprise Miner™

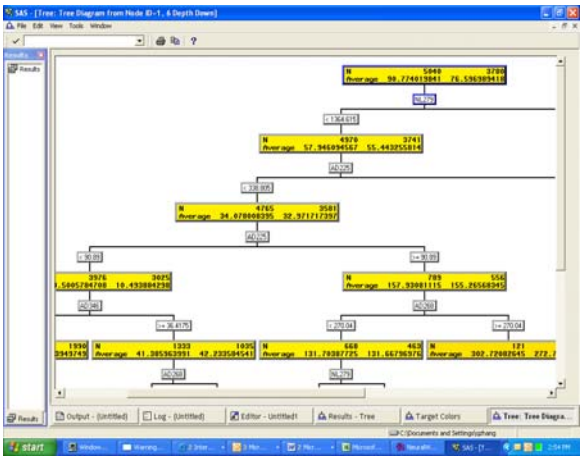

Figure 7: Decision Tree diagram obtained upon applying SAS Enterprise Miner™ for specified depth of 6 from Node ID=1

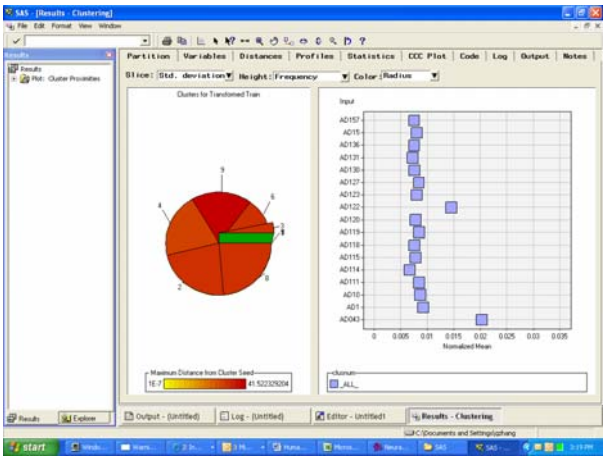

Figure 8: Cluster Analysis for Transformed Training Human Lung Data using SAS Enterprise Miner™

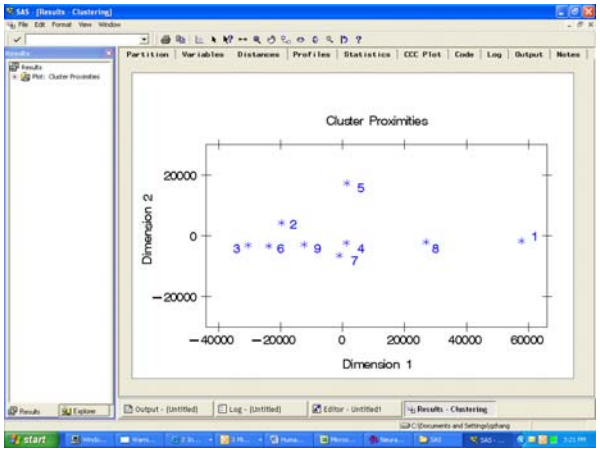

Figure 9: Cluster Proximities for Human Lung data using SAS Enterprise Miner™

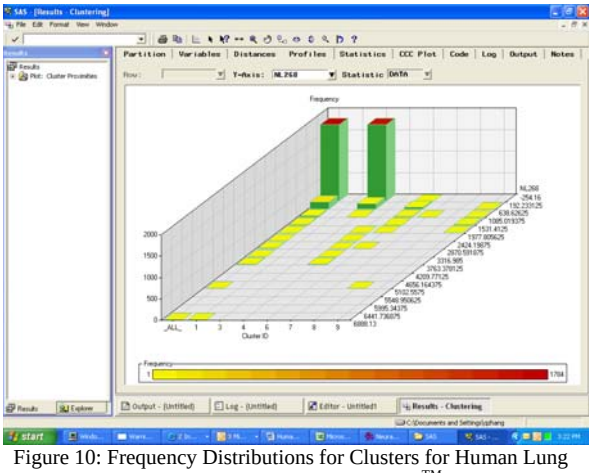

Figure 10: Frequency Distributions for Clusters for Human Lung data using SAS Enterprise Miner™

Conclusions

This paper extends previous research of the authors by providing visual analytical results of using two well-recognized software for data mining of human lung cancer datasets from approximately 12,600 patients for 186 gene types. The visual analytics presented in this paper include link charts, decision trees, gene dimension matrices, regression scatter plots, tables of fit statistics, analysis of variance, SAS process monitor, cluster analysis and proximities, frequency distributions for clusters, self-organized maps and their associated importance rankings of factors. The data mining and visual analytics performed are insightful of new information and especially in determining the significant gene types and the association of specific gene types with human lung cancer.

Acknowledgements

The authors want to acknowledge the support provided by a Summer Faculty Research Grant by the College of Business, and the two software manufactures of SAS Inc. and Megaputer Intelligence Inc. for their support of this research.

References

1. Abdellatif, M. 2000, Leading the Way Using Microarray: A More Comprehensive Approach for Discovery of Gene Expression Patterns. Circulation Research, 86, 919-920.
2. Aitman, T., 2001, Science, Medicine, and the Future DNA Microarrays in Medical Practice. *BMJ*, 323(15), 611-615.
3. American Lung Association, 2005, Facts About Lung Cancer, http://www.lungusa.org/site/pp.asp?c=dvLUK9O0E&b=3
4. Bhattacharjee, A., Richards, W., Staunton, J., et al., 2001, Classification of Human Lung Carcinomas by Mrna Expression Profiling Reveals Distinct Adenocarcinoma Subclasses. *Proc. Natl. Acad. Sci. USA*, 98(24), 13790-13795.
5. Broad Institute Cancer Program Data Sets, 2006, http://www.broad.mit.edu/cancer/datasets.cgi
6. Edgerton, M., Fisher, D., Tang, L., Frey, L., and Chen, Z., 2006, Data mining for gene networks relevant to poor prognosis in lung cancer via backward-chaining rule induction, Libertas Academica, Cancer Informatics, http://www.la-press.com/Cl-2-Edgerton(pr).pdf
7. Keim, D. and Schneidewind, J., 2007, Introduction to the Special Issue on Visual Analytics, SIGKDD Explorations, 9(2), 3-4.
8. Segall, R. S. and Zhang, Q., 2006, Data visualization and data mining of continuous numerical and discrete nominal-valued microarray databases for biotechnology, *Kybernetes: International Journal of Systems and Cybernetics*, v. 35, n. 9/10.
9. Thomas, J. and Cook, K., 2005, Illustrating the Path: Research and Development Agenda for Visual Analytics. IEEE press.