Advanced Data Mining of Leukemia Cells Micro-Arrays

Richard S. SEGALL
Department of Computer & Information Technology, Arkansas State University
State University, AR 72467-0130 USA

and

Ryan M. PIERCE
Student Affairs Technology Services, Arkansas State University
State University, AR 72467-0348 USA

ABSTRACT

This paper provides continuation and extensions of previous research by Segall and Pierce (2009a) that discussed data mining for micro-array databases of Leukemia cells for primarily self-organized maps (SOM). As Segall and Pierce (2009a) and Segall and Pierce (2009b) the results of applying data mining are shown and discussed for the data categories of microarray databases of HL60, Jurkat, NB4 and U937 Leukemia cells that are also described in this article.

First, a background section is provided on the work of others pertaining to the applications of data mining to micro-array databases of Leukemia cells and micro-array databases in general. As noted in predecessor article by Segall and Pierce (2009a), micro-array databases are one of the most popular functional genomics tools in use today.

This research in this paper is intended to use advanced data mining technologies for better interpretations and knowledge discovery as generated by the patterns of gene expressions of HL60, Jurkat, NB4 and U937 Leukemia cells. The advanced data mining performed entailed using other data mining tools such as cubic clustering criterion, variable importance rankings, decision trees, and more detailed examinations of data mining statistics and study of other self-organized maps (SOM) clustering regions of workspace as generated by SAS Enterprise Miner version 4. Conclusions and future directions of the research are also presented.

Keywords: Micro-array databases, self-organized maps, Leukemia, data mining, HL60, U937, NB4, Jurkat

1. INTRODUCTION

According to Wikipedia (2008a), "Leukemia is a cancer of the blood or bone marrow and is characterized by an abnormal proliferation of blood cells, usually white cells called 'leukocyytes'."

According to Piatetsky-Shapario and Tamayo (2003), "Microarrays are a revolutionary new technology with great potential to provide accurate medical diagnosis, help find the right treatment and cure for many diseases and provide a detailed genome-wide molecular portrait of cellular states."

As noted in Segall and Pierce (2009a): With advances in "gene chip" technology, gathering microarray data has become faster and more efficient than ever before. Single chip microarrays yielding estimations of the absolute expression values of a particular gene afford insight into what a cell is actually doing and how it reacting and changing given various stimuli. Gene expression information can be therefore systematically harvested, stored, and analyzed at later dates.

While traditional statistical methods are helpful in preliminary analysis, data mining techniques can not only draw conclusions but also aid in visualizing patterns within the data set itself.

This paper further extends the process of knowledge discovery by applying advanced data mining tools for analyzing and graphically representing the HL60, Jurkat, NB4 and U937 data of cell lines for Leukemia that entails 6416 genes. The advanced data mining tools include cubic clustering criterion, variable importance rankings, decision trees, and more detailed examinations of data mining statistics and study of other self-organized maps (SOM) clustering regions of workspace as generated by

In particular, self-organized maps provides mapping from the input space to the clusters. In SAS® manual "Applying Data Mining Techniques Using Enterprise Miner^{TM}", Walsh (2002) informs us that a SOM attempts to organize clusters that are near each other in the grid space to those seeds that are close in the input space. It differs from *k*-means clustering because it defines an area around each cluster seed in the grid via a neighborhood function. Clusters that are close in proximity on the grid have similar input variables.

Also taken from Walsh (2002), the more widely used cluster analysis, employed for the purposes of this paper is the *k*-means method. This type of clustering, also known as unsupervised classification attains homogeneity amongst disjoint classes with respect to inputs. The standard clustering algorithm, Euclidian (L_2norm), was employed for all clustering within the scope of this paper. Holding true to traditional form, the clusters from this metric tend to be spherical.

2. BACKGROUND

Data

As in the predecessor article by Segall & Pierce (2009a), the data used in this paper was obtained from the Cancer Program website of the Broad Institute (2007). The Broad Institute was founded in 2003 by philanthropists Eli and Edythe Broad, and is a research collaboration involving faculty, staff and students from throughout the MIT (Massachusetts Institute of Technology) and Harvard academic and medical communities and is governed jointly by the two universities.

Below are definitions of terminology HL-60, Jurkat and microarray database that were obtained from Wikepedia (2008a, b, and c) and are necessary for the understanding of both this article and the predecessor article by Segall & Pierce (2009a).

According to Wikipedia (2008b), "The HL-60 (*Human promyelocytic leukemia cells*) cell line is a leukemic cell line that has been used for laboratory research on how certain kinds of blood cells are formed."

According to Wikipedia (2008c), "Jurkat cells are cells that are used to study acute leukemia, and the expression of various receptors susceptible to viral entry, particularly HIV. … Their primary use, however, is to determine the mechanisms of differential susceptibility of cancers to drug and radiation."

Also according to Wikipedia (2008d), "The term microarray database is usually used to describe a repository containing microarray gene expression data."

Data mining of Leukemia and micro-array databases

Previous work by the lead author of this article in the area of data mining of micro-array databases for biotechnology has been presented in Segall (2006, 2005a, 2005b, 2005c) and Segall and Zhang (2007, 2006a, 2006b) and Zhang and Segall (2008, 2007). Numerous others have done other work in data mining in bioinformatics, and for example Bakker (2008) in an opening symposium address in the Netherlands.

Piatetsky-Shapario and Tamayo (2003) discussed the challenges facing investigators of microarray data mining, and also provided an overview of mciroarrays using Affymetrix GeneChip®, and discussion of molecular bilogy and DNA. Some of the challenges Piatetsky-Shapario and Tamayo (2003) cite are gene selection, classification, and clustering and visualization, and low-level analysis.

According to Van der Puten (2005), the problem with Leukemia cell data is that different types of leukemia cells look very similar and hence data mining on micro-array data is the solution for the problem. Data mining can solve problem of that given data for a number of patients, data mining of microarrays of Leukemia cell data can help accurately diagnose the disease, predict outcomes for given treatment, and recommend best treatment.

Conger (2006) cited the significance of using data mining at the genetic level as comparing to "striking genetic gold". Dunphy (2006) performed 'gene expression profiling' in lymphoma and Leukemia. Others performed data mining of leukemia such as Glover et al. (2007) and Labib and Malek (2005) for data mining for cancer management in Egypt case study for childhood acute lymphoblastic Leukemia. Markiewicz and Osowski (2006) performed data mining techniques for feature selection in blood cell recognition, and Marx et al. (2003) performed data mining of the NCI Cancer cell compound GI50.

3. ADVANCED DATA MINING OF LEUKEMIA CELLS

The data mining was performed using SAS Enterprise Miner^{TM} version 4 with modules of predictive modeling that include that for clustering, regression, decision trees.

The data sets used in this paper are that of predecessor paper by Segall and Pierce (2009a). These data sets are individual HL60 data and then also combined HL60, Jurkat, NB4 and U937 (i.e. HL60_U937_NB4_Jurkat) datasets. The data mining is performed to these data sets for interpreting patterns of gene expression from the cancer programs data sets that are available from the Broad Institute (2007).

The data sets for HL60 from the Broad Institute (2007) are described to "contain four time point measurements of t =0 (baseline), t= 0.5 hours, t=4 hours, and t=24 hours each of which corresponds to differentiation time course of HL60 cells. These cells undergo macrophage differentiation upon treatment with the phorbol ester TPA." Also according to the Broad Institute (2007) dataset description, "nearly 100% of HL60 cells become adherent and exit the cell cycle within 24 hours of TPA treatment. To monitor this process at the transcriptional level, cells were harvested at 0, 0.5, 4 and 24 hours after TPA stimulation."

The combined dataset of HL60_U937_NB4_Jurket according to the Broad Institute (2007) data description, "combines expression data from four different cell lines: HL-60 and U937, two myeloid cell lines which undergo macrophage differentiation in response to TPA; NB4, an acute promyelocytic leukemia cell line that undergoes neutrophilic differentiation in response to all-trans retinoic acid (ATRA), and Jurkat, a T-cell line that acquires many hallmarks of T-cell activation in response to TPA.” There are four (4) time periods of measurements each for HL60, U937 and Jurkat of 0, 0.5, 4 and 24 hours) and five (5) time periods of measurements for NB4 of 0, 5.5, 24, 48 and 72. There are a total of 6416 genes and the data was obtained using Affymetrix Hu6000 DNA micro-arrays.

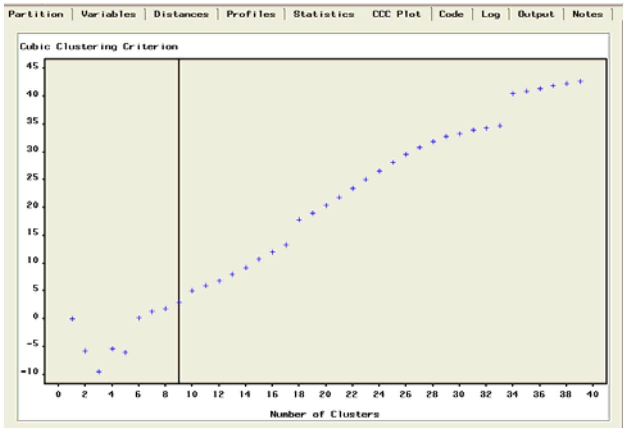


Figure 1: Cubic Clustering Criterion – HL60

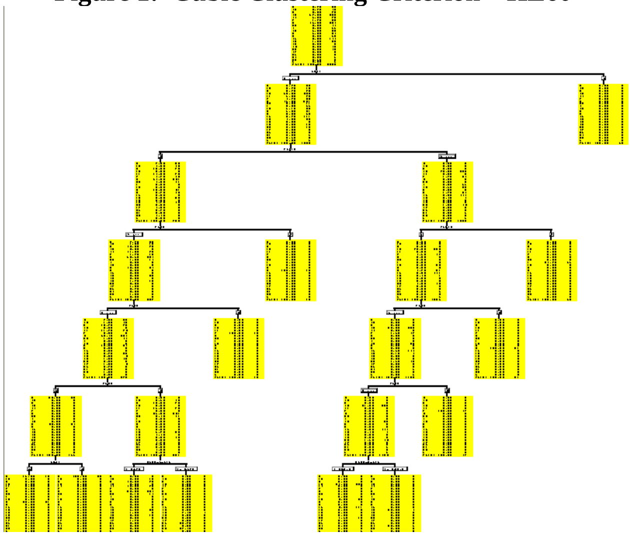


Figure 2: Decision Trees Obtained By Mining HL60 Luekemia Cell Data

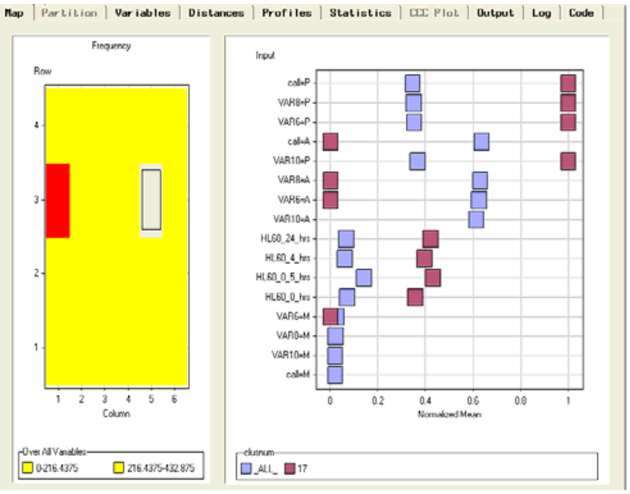
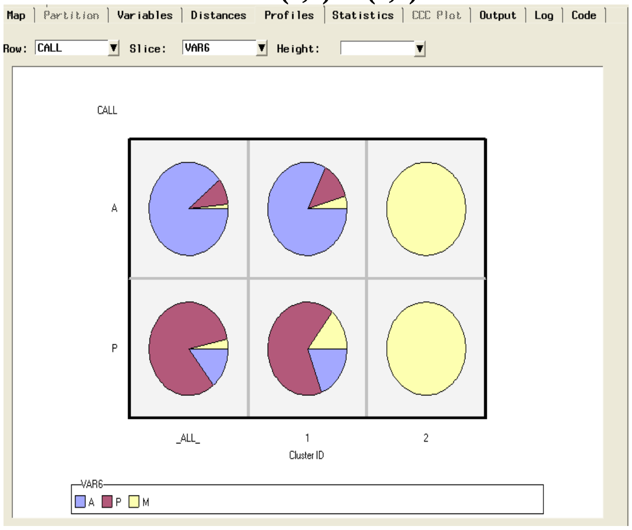


Figure 3: Self-Organized Maps of HL60 Leukemia Data: (3,1) & (3,5)



Row: CALL Slice: VAR6 Height:

Figure 4: Clustering Results Using SOM for HL60 Leukemia Data

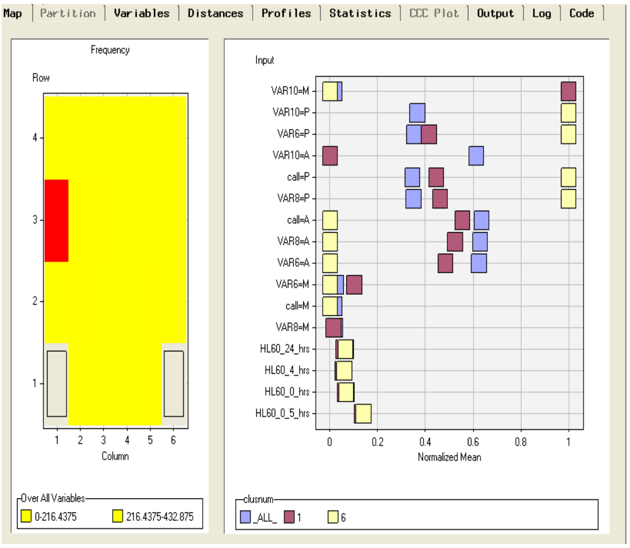
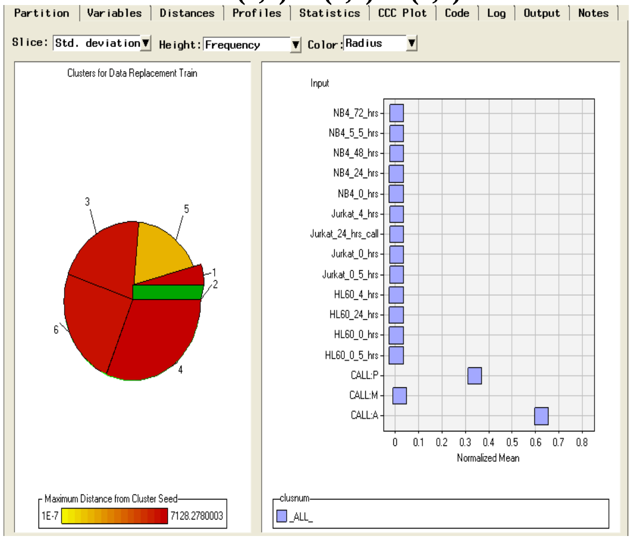


Figure 5: Self-Organized Maps of HL60 Leukemia Data: (3,1) & (1,1) & (1,6)



Slice: Std. deviation Height: Frequency Color: Radius

Figure 6: Pie Chart of Clusters for HL60, Jurkat, NB4, and U937 Leukemia Cells

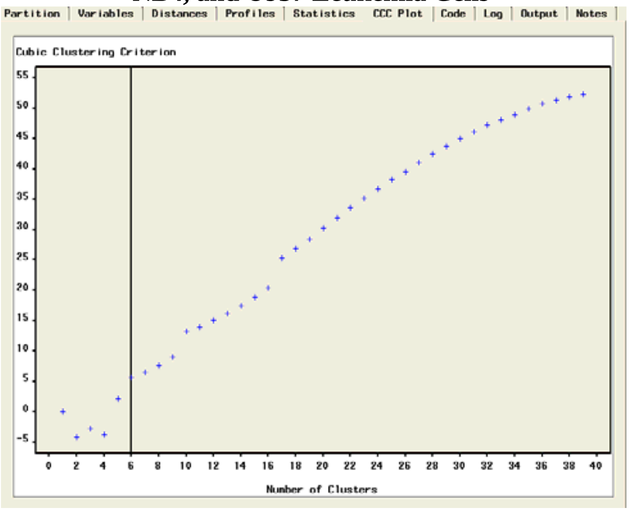


Figure 7: Cubic Clustering Criterion for HL60, Jurkat, NB4, and U937 Leukemia Cells

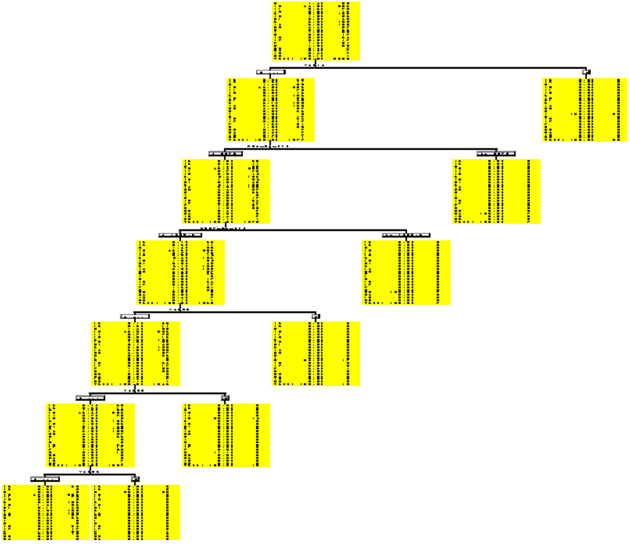


Figure 8: Decision Trees for HL60, Jurkat, NB4, and U937 Leukemia Cell Data

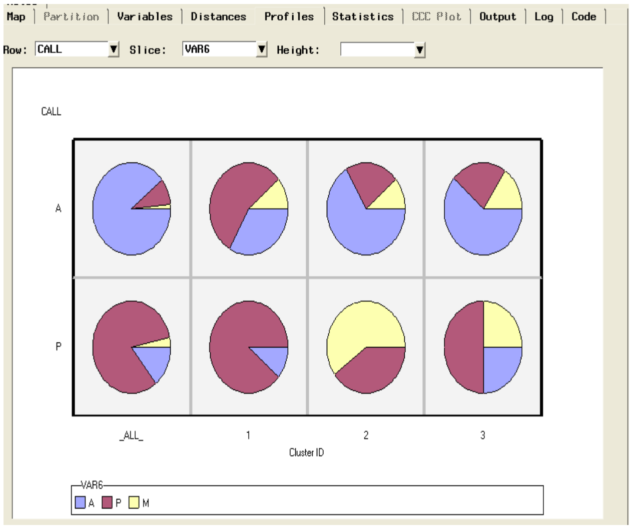

Map | Partition | Variables | Distances | Profiles | Statistics | CCC Plot | Output | Log | Code
Row: CALL
Slice: VAR6
Height:
CALL
A
P
ALL
1
2
3
Cluster ID
VAR6
A
P
M

Figure 9: Clustering Results Using SOM for HL60, Jurkat, NB4, and U937 Leukemia Data

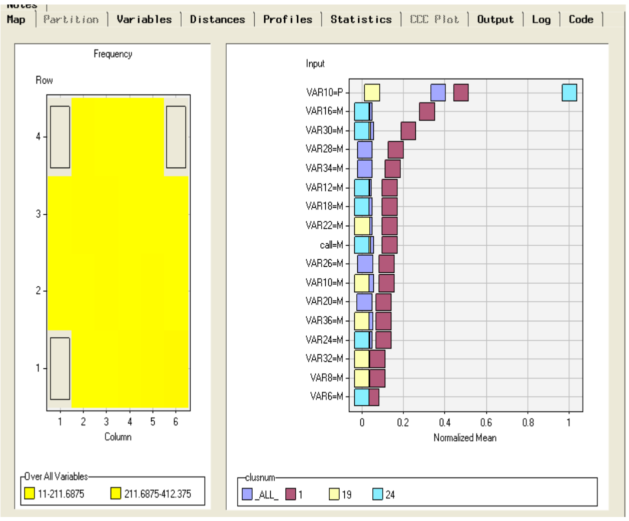


Figure 10: Self-Organized Maps of HL60, Jurkat, NB4, and U937 Leukemia Data: (4,1), (4,6) & (1,1)

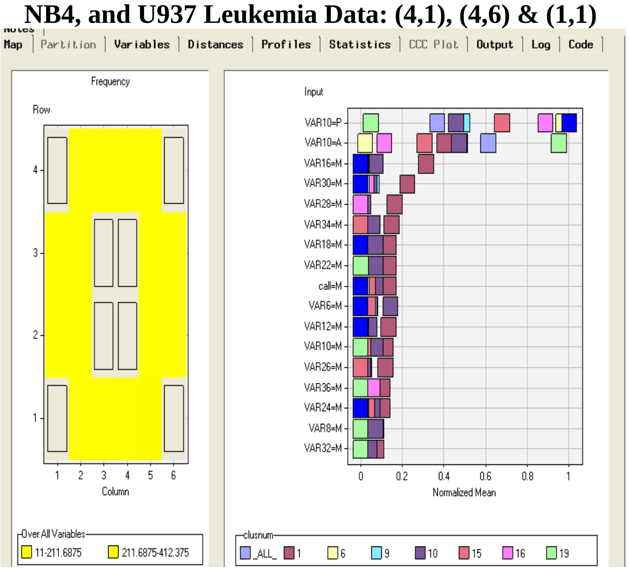


Figure 11: Self-Organized Maps of HL60, Jurkat, NB4, and U937 Leukemia Data: Extremities & Center Clusters

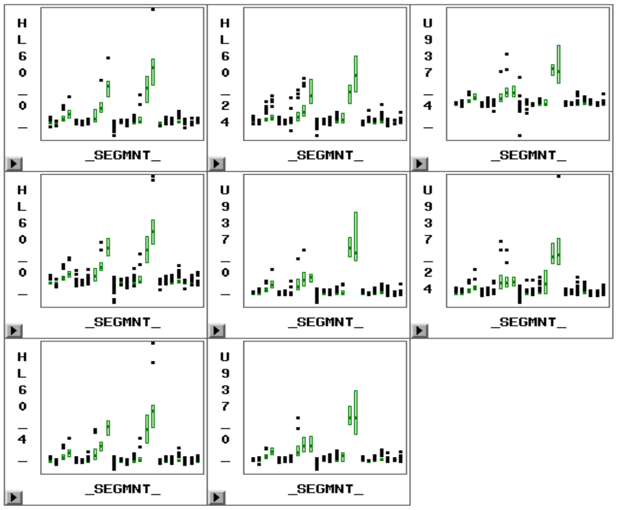


Figure 12: Time Measure Box Whisker Plots for HL60 & U937 Leukemia Cells

Figure 1 shows the cubic clustering criterion for the HL60 data set which is non-linear and has several discontinuities when the number of clusters is less than 10 and also when equal to 18 and 25.

Figure 2 shows the decision trees obtained by mining HL60 Leukemia cell data which has single nodes at each of the upper branches except for the last branch at the bottom level. There is also branching at end of the lower levels. This is in contrast to that of Figure 8 to be discussed later for the combined HL60_U937,_NB4_Jurkat data.

Figure 3 shows self-organized maps of HL60 Leukemia cells for regions (3.1) and (3.5). Figure 3 appears to illustrate four intervals of normalized means with the variables of this HL60 data set.

Figure 4 shows clustering results using self-organized maps (SOM) for HL60 Leukemia cell data that illustrates absence or presence as the dominant region of their respective clusters.

Figure 5 shows self-organized maps for HL60 Leukemia data for clusters (3,1), (1,1) and (1,6) of the data base. In contrast to Figure 3, this Figure 5 has three instead of four intervals of normalized means with the variables of the HL60 data set.

Figure 6 shows a pie chart of clusters for HL60_U937_NB4_Jurkat data with almost uniform normalized mean of zero (0) for all of the variables.

Figure 7 shows the cubic clustering criterion for the combined HL60_U937,_NB4_Jurkat data for Leukemia cells. Figure 7 is similar to that of Figure 1 for the HL60 data set alone in that it is also non-linear, but the combined data set of Figure 7 was more uniform and had fewer discontinuities than that of Figure 1 as to be expected for a larger data set.

Figure 8 shows decision trees for the combined HL60_U937_NB4_Jurkat data for Leukemia cells, and as indicated earlier has double branching made at each of the levels in contrast to that for HL60 data alone as in Figure 2.

Figure 9 shows clustering results using self-organized maps for HL60_U937,_NB4_Jurkat data for Leukemia cells. Figure 9 shows predominance of absence in the top row of absence clusters, and presence in only the first two of the pie charts of the bottom row. The last two pie charts of the bottom row signifying presence indicate a significant amount data with “missing” or other unlabeled characteristics.

Figure 10 shows self-organized maps of HL60_U937,_NB4_Jurkat data for Leukemia cells for clusters (4,1), (4,6) and (1,1). This Figure 10 illustrates both linearity for those data having normalized means of zero, and also non-linearity for those data having non-zero normalized means.

Figure 11 shows self-organized maps of HL60_U937_NB4_Jurkat data for Leukemia cells for those clusters at the extremities and center region of the input data. Specifically this includes clusters at extremities of (1,1), (1,6), (4,1) and (4,6) and center region clusters of (2,3), (2,4), (3,3) and (3,4). This Figure 11 shows more density than that of Figure 10 for three clusters instead of six clusters as in this figure. Figure 11 exemplifies the shape of the normalized means figures of those of Figure 10 by have both linearity for those data having normalized means of zero, and also non-linearity for those data having non-zero normalized means.

Figure 12 shows the Box-Whisker plots for individual HL60 and individual U937 Leukemia cells data. The vertical ranges of the Box-Whisker plots of the HL60 appear to have overall more variability than those of the U937 Leukemia cell data. The vertical range of one bar of the U937 appears to exceed that all of those of the HL60 data.

Other data mining that was performed for which figures are not provided include that for importance profile and data mining statistics of HL60 and using self-organized maps and that similarly for HL60_U937_NB4_Jurkat data.

4. CONCLUSIONS AND FUTURE DIRECTIONS

Applying data mining to the HL60 data by itself lead to visualization plots that had disperse normalizations means such as that shown by Figures 3 and 5, as compared to that for the aggregation of HL60, Jurkat, NB4 andU937 Leukemia data as shown in Figures 10 and 11. Similarly Figure 4 for HL60 data by itself yielded only four clusters after applying the data mining, while Figure 9 for the aggregated data yielded double the number of clusters.

Also the cubic clustering criterion yielded upon applying data mining to the HL60 data alone in Figure 1 had points of discontinuity at larger number of clusters than that in Figure 7 for the aggregation of HL60, Jurkat, NB4 and U937 Leukemia data. The clustering obtaining upon using data mining in Figure 4 for HL60 data by itself yielded only four clusters, while that in Figure 9 for the aggregated data yielded double the number of clusters.

The decision trees shown in Figure 2 obtained by using data mining for the HL60 Leukemia data by itself had 20 nodes and were double-branched at the third level. While the decision trees in Figure 8 for the aggregated data of HL60, Jurkat, NB4 and U937 data had fewer nodes of 12 and simplified double-branching from the second level downward. All of the above statements lead to the conclusion of the value of the aggregated Leukemia cell data and also the precise selection of these Leukemia cell data types of HL60, Jurkat, NB4 and U937.

Future directions of the research include that of continuing to pursue the applications of additional techniques of data mining of the Leukemia cell data, as well as apply data mining to other data sets of the Broad Institute (2007) Cancer Program and compare and contrast new conclusions among the results obtained using data mining to micro-array databases.

5. ACKNOWLEDGEMENTS

The authors would like to acknowledge their gratefulness to SAS Institute Inc. for their partnership created with the College of Business (CoB) at Arkansas State University (ASU), and who provided them with the data mining software of SAS Enterprise Miner version 4 and technical support that was vital in performing the research described within this article.

6. REFERENCES

[1.] Bakker, E.M.(2008), "Data mining in bioinformatics", **DAS3 Opening Symposium**, August 11, Universiteit Leiden, Netherlands http://www.cs.vu.nl/das3/symposium07/das3-embakker.ppt and http://www.docstoc.com/docs/433423/Data-Mining-in-Bioinformatics

[2.] Broad Institute (2007), **Cancer Program Data Sets**, http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi

[3.] Conger, K. (2006), Stanford/Packard scientist's data-mining technique strikes genetic gold," ***Medical News Today***, January 11, http://www.medicalnewstoday.com/articles/36009.php

[4.] Dunphy, C. H., (2006), "Gene expression profiling data in lymphoma and leukemia: Review of the literature and extrapolation of pertinent clinical applications," ***Archives of Pathology & Laboratory Medicine***, April, v. 130, pp. 483-520.

[5.] Glover, C.J., Rabow, A.A, Igsor, Y. G., Shoemaker. R.H., and Couell, D.G. (2007), "Data mining of NCI's anticancer screening database reveals mitochondrial complex I inhibitors cytotoxix to leukemia cell lines, ***Biochemical Pharmacology***, v. 73, n.3, pp. 331-340.

[6.] Labib, N.M. and Malek, M.N. (2005), "Data mining for cancer management in Egypt Case Study: Childhood Acute Lymphoblastic Leukemia", **Proceedings of World Academy of Science and Engineering and Technology**, v. 8, October, pp. 309-314.

[7.] Markiewicz, T. and Osowski, S., "Data mining techniques for feature selection in blood cell recognition:, **Proceedings of European Symposium on Artificial Neural Networks (ESANN'2006)**, April 26-28, 2006, pp. 407-412.

[8.] Marx, K.A., O'Neil, P., Hoffman, P., Ujwal, M.L. (2003), "Data mining the NCI cancer cell compound GI50 values:
Identifying Quinone Subtypes effective against Melanoma and Leukemia cell classes, **Journal of Chemical Information and Modeling**, v. 43, n.5, pp. 1652-1667.

[9.] Piatetsky-Shapario, G. and Tamayo, P (2003), "Microarray Data Mining: Facing the Challenges, **SIGKDD Explorations**, v. 5, n.2, pp. 1-5.

[10.] Segall, R. S. (2006), "Data Mining of Microarray Databases for Biotechnology,"**Encyclopedia of Data Warehousing and Mining**, Edited by John Wang, Montclair State University, USA; Idea Group Inc., 2006, ISBN 1-59140-557-2.

[11.] Segall, R. S. (2005a), "Data Mining of Micro-Array Databases for Biotechnology", R & R in the Bioinformatics World, **Arkansas Biosciences Institute (ABI) Lecture Series**, Arkansas State University, Jonesboro, AR, October 19, 2005.

[12.] Segall, R. S. (2005b), "Data Mining of Microarray Databases for the Analysis of Environmental Factors on Plants Using Cluster Analysis and Predictive Regression", **Proceedings of the Thirty-sixth Annual Conference of the Southwest Decision Sciences Institute**, vol. 36, no. 1, March 3-5, 2005, Dallas, TX.

[13.] Segall, R. S. (2005c), "Data Mining of Microarray Databases for the Analysis of Environmental Factors on Corn and Maize," **Proceedings of the 2005 Conference of Applied Research in Information Technology**, sponsored by Acxiom Laboratory for Applied Research (ALAR), University of Central Arkansas, February 18, 2005.

[14.] Segall, R. S. and Pierce, R. M. (2009a), "Data mining of Leukemia cells using Self-Organized Maps", **Proceedings of 2009 ALAR Conference on Applied Research in Information Technology**, February 13, 2009.

[15.] Segall, R.S. and Pierce, R.M. (2009b), "Advanced data mining of Leukemia cells", Research-In-Progress Abstract, **Proceedings of 2009 ALAR Conference on Applied Research in Information Technology,** February 13, 2009.

[16.] Segall, R. S. and Zhang, Q. (2007), "Data Mining of Microarray Databases for Human Lung Cancer, **Proceedings of the Thirty-eighth Annual Conference**

of the Southwest Decision Sciences Institute, vol. 38, no. 1, March 15-17, 2007, San Diego, CA.

[17.] Segall, R. S. and Zhang, Q. (2006a), "Applications of Neural Network and Genetic Algorithm Data Mining Techniques in Bioinformatics Knowledge Discovery – A Preliminary Study", **Proceedings of the Thirty-seventh Annual Conference of the Southwest Decision Sciences Institute**, vol. 37, no. 1, March 2-4, 2006, Oklahoma City, OK.

[18.] Segall, R. S. and Zhang, Q. (2006b), "Data Visualization and Data Mining of Micro-Array Databases for Biotechnology", ***Proceedings of the 2006 Conference of Applied Research in Information Technology***, sponsored by Acxiom Laboratory for Applied Research (ALAR), University of Central Arkansas, March 3, 2006.

[19.] Stratowa, C., Loffler, G., Lichter, P., Stilgenbauer, S., Haberi, P., Scheweifer, N., Dohner, H., and Wilgenbus, K.K. (2001), CDNA microarray gene expression analysis of B-cell chronic lymphocytic leukemia proposes potential view prognostic markers involved in lymphocyte trafficking, *International Journal of Cancer*, v. 91, n. 4, pp. 474-480.

[20.] Ujwal, M.L, Hoffman, P., and Marx, K.A, (2007), "A machine learning approach to pharmacological profiling of the quinine scaffold in the NCI database: A compound class enriched in those effective against melanoma and leukemia cell lines," *Proceedings of the 7th IEEE International Conference on Bioinformatics and Bioengineering (BIBE 2007)*, October, pp.456-463.

[21.] van der Putten, P. (2005), "Lecture 2: Predictive Data Mining", TUDelft, Universiteit Leiden, Netherlands, http://www.liacs.nl/~putten/edu/dbdm05.

[22.] Walsh, S. (2002), "Applying Data Mining Techniques Using Enterprise MinerTM", SAS Institute Inc., Cary, NC, pp. 7-4 to 7-27.

[23.] Wikipedia (2008a), Leukemia, http://en.wikipedia.org/wiki/Leukemia

[24.] Wikipedia (2008b), HL60, http://en.wikipedia.org/wiki/HL60

[25.] Wikipedia (2008c), Jurkat, http://en.wikipedia.org/wiki/Jurkat_cells

[26.] Wikipedia (2008d), Microarray databases, http://en.wikipedia.org/wiki/Microarray_databases

[27.] Zhang, Q. and Segall, R. S.(2008), "Visual Analytics of Mining Human Lung Cancer Data", *Proceedings of the 3rd INFORMS (Institute for Operations Research and Management Science) Workshop on Data Mining and Health Informatics*, (DH-HI 2008), J.Li, D. Aleman, and R. Sikora, Editors, October 11, 2008, Washington, DC.

[28.] Zhang, Q. and Segall, R. S. (2007), "Data Mining of Forest Cover and Human Lung Micro-Array Databases with Four-Selected Software", *Proceedings of the 2007 Conference of Applied Research in Information Technology*, sponsored by Acxiom Laboratory for Applied Research (ALAR), University of Arkansas-Fayetteville, March 9, 2007.