From Text Data Mining to Gene Data Mining and Back Again

Jeff Solka[1,2]

Avory Bryant[1], Brandon Higgs[2,3],

and Edward J. Wegman[4]

1 – NSWCDD Code B10

2 – SCS GMU

3 - Mitre

4 – Department of Applied and Engineering Statistics GMU

Acknowledgements

- Avory Bryant
  - Text data mining
- Brandon Higgs
  - Gene expression data mining
- Edward Wegman
  - Visualization strategies
- Office of Naval Research
- Defense Advanced Research Projects Agency

Agenda

- Text data mining
- Graph theoretic formulation
- Text data mining preliminary results
- Graph theoretic formulation modified for gene expression data
- Preliminary results on Golub data
  - Text data mining reenters the picture
- Preliminary results on the Alon data
- Conclusion

In a Nutshell?

- What are we trying to do?
  - Develop new and extend existing methods of subspace/biclustering.
- What is our approach predicated on?
  - The synthesis of methodologies from statistics, mathematics and visualization.
- What are the test cases?
  - Roughly 1200 Science News abstracts that have been precategorized into 8 categories.
  - *Roughly 343 Office of Naval Research In-house Laboratory Independent Research documents.*
  - Golub gene expression data.
  - Alon cancer data

What is Biclustering and Subspace Clustering?

- Given a set of n observations in p dimensions (an n by p matrix).
- Biclustering is the simultaneous clustering of observations and dimensions.
- Subspace clustering is the identification of cluster structures that may be manifest only on a subset of dimensions.
  - The cluster structures may reside on manifolds or lower dimensional subspaces in the ambient space.

Getz G, Levine E, and Domany E. “Coupled two-way clustering analysis of gene microarray data.” *Proc Natl Acad Sci USA* 97: 12079–12084, 2000

Text Data Mining Applications

- Literature based discovery
- Formulation of research agendas
  - BAA announcements
  - Conference agendas
- Technology point papers
  - Discipline area
  - Country X
  - Country X vs. Country Y
- Assessment of gene discoveries
  - Literature evidence relationship between gene G and disease Y

The Science News Corpus

- 1117 documents from 1994–2002.
- Obtained from the SN website on December 2002 19,2002 using wget.
- Each article ranges from 1/2 a page to roughly a page in length.
- The corpus html/xml code was subsequently parsed into straight text.
- The corpus was read through and categorized into 8 categories.

The Science News Corpus Breakdown

- Anthropology and Archeology (48).
- Astronomy and Space Sciences (124).
- Behavior (88).
- Earth and Environmental Sciences (164).
- Life Sciences (174).
- Mathematics and Computers (65) .
- Medical Sciences (310) .
- Physical Sciences and Technology (144)

Denoising and Stemming

- These steps are performed prior to subsequent feature extraction steps.

- Various approaches to denoising were used
  - Simplest consists of removal of all words that appear on a stopper or noise word list.
  - the, a, an, …
  - More on this later

- Stemming transforms a given word into its base
  - walking → walk
  - walked → walk

- Denoising is implemented within the current system

Document Features

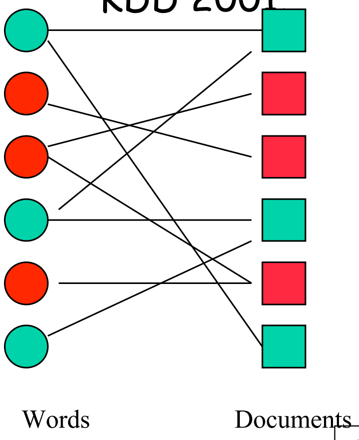
- Bigram Proximity Matrices ala Martinez 2002
  - Angel Martinez, "A Framework for the Representation of Semantics," *Ph.D Dissertation under the direction of Edward Wegman*, October 2002.

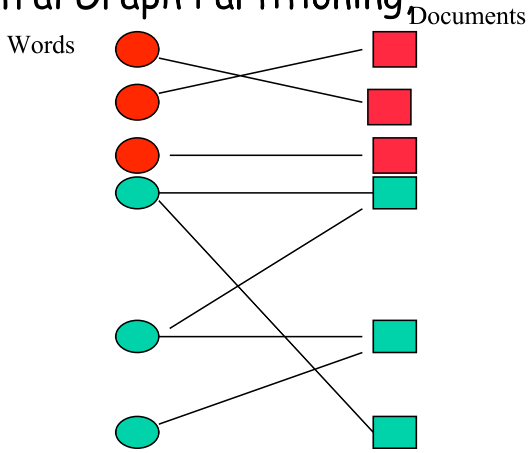
- Mutual Information Features ala Lin 2002
  - Patrick Pantel and Dekang Lin, "Discovery word senses from text," in Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pgs. 613-619, 2002.

- "Normalized" term document matrices ala Dhillon 2001
  - Inderjit S. Dhillon, "Co-clustering documents and words

Bipartite Spectral Based Clustering

- Inderjit S. Dhillon, “Co-clustering documents and words using Bipartite Spectral Graph Partitioning,” KDD 2001





Cut measures the sum of the crossing between vertex set V_1 and vertex set V_2.

$$\mathrm{cut}(\mathcal{V}_1, \mathcal{V}_2) = \sum_{i \in \mathcal{V}_1, j \in \mathcal{V}_2} M_{ij}$$

The Graph Theoretic Formulation

Our Graph	Vertex Set	Edge Set	Edge Weights
$G = (\mathcal{V}, E)$	$\mathcal{V} = \{1, 2, \ldots, \lvert\mathcal{V}\rvert\}$	$\{i, j\}$	E_{ij}

Adjacency Matrix

$$M = \begin{cases} E_{ij}, & \text{if there is an edge } \{i, j\}, \\ 0, & \text{otherwise.} \end{cases}$$

The cut between two subsets of vertices.

$$\mathrm{cut}(\mathcal{V}_1, \mathcal{V}_2) = \sum_{i \in \mathcal{V}_1, j \in \mathcal{V}_2} M_{ij}.$$

The cut between k subsets of vertices.

$$\mathrm{cut}(\mathcal{V}_1, \mathcal{V}_2, \ldots, \mathcal{V}_k) = \sum_{i<j} \mathrm{cut}(\mathcal{V}_i, \mathcal{V}_j)$$

The Document Word Bipartite Model

Our graph consisting of a vertex set consisting of documents and words along with associated edges.

$$G = (\mathcal{D}, \mathcal{W}, E)$$

The word vertices.

$$\mathcal{W} = \{w_1, w_2, \ldots, w_m\}$$

The document vertices.

$$\mathcal{D} = \{d_1, d_2, \ldots, d_n\}$$

One strategy for setting the edge weights.

$$E_{ij} = t_{ij} \times \log\left(\frac{|\mathcal{D}|}{|\mathcal{D}_i|}\right)$$

where t_{ij} is the number of times word w_i occurs in document d_j, $|\mathcal{D}| = n$ is the total number of documents and $|\mathcal{D}_i|$ is the number of documents that contain word w_i.

$$M = \begin{bmatrix} \mathbf{0} & A \\ A^T & \mathbf{0} \end{bmatrix}$$

Adjacency Matrix – A_{ij} = E_{ij}, 0's reflect no word to word or document to document connections

$$\text{cut}(\mathcal{W}_1 \cup \mathcal{D}_1, \mathcal{W}_2 \cup \mathcal{D}_2, \ldots, \mathcal{W}_k \cup \mathcal{D}_k) = \min_{\mathcal{V}_1, \mathcal{V}_2, \ldots, \mathcal{V}_k} \text{cut}(\mathcal{V}_1, \mathcal{V}_2, \ldots, \mathcal{V}_k)$$

Our Clustering Criteria

Corpus Dependent Stop Word Removal

- Stop words are removed.
- Words occurring in less than 0.2% of the documents are removed.
- Words occurring in greater than 15% of the documents are removed.
- N. B.
  - The methodology has been shown successful even if stopper words are not removed.
  - 0.2% and 15% are user "tunable" parameters.

Graph Partitioning

Given a graph $G = (\mathcal{V}, E)$, the classical graph bipartitioning or bisection problem is to find nearly equally-sized vertex subsets $\mathcal{V}_1^*, \mathcal{V}_2^*$ of $\mathcal{V}$ such that

$$\text{cut}(\mathcal{V}_1^*, \mathcal{V}_2^*) \;=\; \min_{\mathcal{V}_1, \mathcal{V}_2} \text{cut}(\mathcal{V}_1, \mathcal{V}_2).$$

The graph partitioning problem is known to be NP-complete.

We will follow Dhillon and use graph spectral methods to obtain an approximate solution based on a suitably formulated objective function.

Assuring An Equitable Partition – An Objective Function

$$W_{ij} = \begin{cases} \text{weight}(i), & i = j, \\ 0, & i \neq j. \end{cases}$$

The weight for a particular vertex.

$$\text{weight}(\mathcal{V}_l) = \sum_{i \in \mathcal{V}_l} \text{weight}(i) = \sum_{i \in \mathcal{V}_l} W_{ii}$$

The weight for a set of vertices.

A figure of merit function that helps assure near equal number of points in each cluster.

$$\mathcal{Q}(\mathcal{V}_1, \mathcal{V}_2) \;=\; \frac{\text{cut}(\mathcal{V}_1, \mathcal{V}_2)}{\text{weight}(\mathcal{V}_1)} + \frac{\text{cut}(\mathcal{V}_1, \mathcal{V}_2)}{\text{weight}(\mathcal{V}_2)}$$

One can think of this as being analogous to the ratio of between group and within group distances in our usual statistical clustering framework.

Choice of Vertex Weights

$$\text{weight}(i) = 1$$

$$\text{Ratio-cut}(\mathcal{V}_1, \mathcal{V}_2) = \frac{\text{cut}(\mathcal{V}_1, \mathcal{V}_2)}{|\mathcal{V}_1|} + \frac{\text{cut}(\mathcal{V}_1, \mathcal{V}_2)}{|\mathcal{V}_2|}$$

$$\text{weight}(i) = \sum_k E_{ik}$$

Normalized cut.

$$\mathcal{N}(\mathcal{V}_1, \mathcal{V}_2) = \frac{\text{cut}(\mathcal{V}_1, \mathcal{V}_2)}{\sum_{i \in \mathcal{V}_1} \sum_k E_{ik}} + \frac{\text{cut}(\mathcal{V}_1, \mathcal{V}_2)}{\sum_{i \in \mathcal{V}_2} \sum_k E_{ik}}$$

Algorithm Bipartition

$$D_1(i,i) = \sum_j A_{ij} \quad \text{(sum of edge-weights incident on word } i\text{)},$$

$$D_2(j,j) = \sum_i A_{ij} \quad \text{(sum of edge-weights incident on document } j\text{)}.$$

$$z_2 = \begin{bmatrix} \boldsymbol{D_1}^{-1/2} \boldsymbol{u}_2 \\ \boldsymbol{D_2}^{-1/2} \boldsymbol{v}_2 \end{bmatrix} \quad (4.13)$$

Algorithm Bipartition

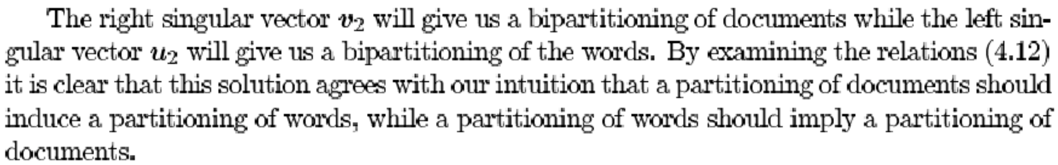
1. Given $\boldsymbol{A}$, form $\boldsymbol{A}_n = \boldsymbol{D_1}^{-1/2} \boldsymbol{A} \boldsymbol{D_2}^{-1/2}$.
2. Compute the second singular vectors of $\boldsymbol{A}_n$, $\boldsymbol{u}_2$ and $\boldsymbol{v}_2$ and form the vector $\boldsymbol{z}_2$ as in (4.13).
3. Run the k-means algorithm on the 1-dimensional data $\boldsymbol{z}_2$ to obtain the desired bipartitioning.

The singular vectors $\boldsymbol{u}_2$ and $\boldsymbol{v}_2$ of $\boldsymbol{A}_n$ give a real approximation to the discrete optimization problem of minimizing the normalized cut.

The Left and Right Singular Vectors

$$A_n v_2 = \sigma_2 u_2, \qquad {A_n}^T u_2 = \sigma_2 v_2, \qquad (4.12)$$

$$\sigma_2 = 1 - \lambda_2$$



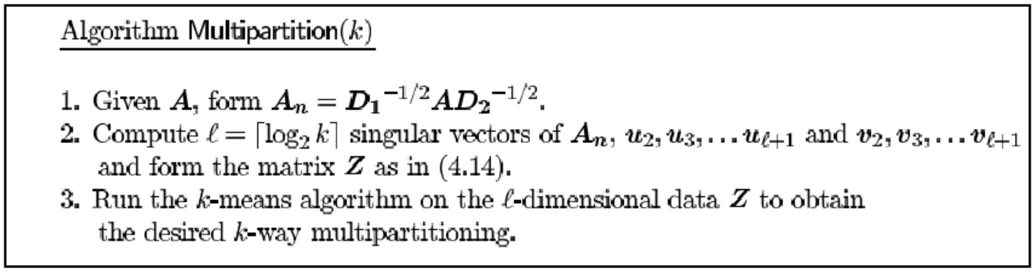
The right singular vector v_2 will give us a bipartitioning of documents while the left singular vector u_2 will give us a bipartitioning of the words. By examining the relations (4.12) it is clear that this solution agrees with our intuition that a partitioning of documents should induce a partitioning of words, while a partitioning of words should imply a partitioning of documents.

The curious fact is that the obtained transformation allows one to map the documents and words into the same one-dimensional space.

Algorithm Multipartition(k)

$$Z = \begin{bmatrix} {D_1}^{-1/2} U \\ {D_2}^{-1/2} V \end{bmatrix} \quad (4.14)$$

$$U = [u_2, u_3, \ldots, u_{\ell+1}], \text{ and } V = [v_2, v_3, \ldots, v_{\ell+1}], \quad \ell = \lceil \log_2 k \rceil$$



```
Algorithm Multipartition(k)

1. Given $A$, form $A_n = {D_1}^{-1/2} A {D_2}^{-1/2}$.
2. Compute $\ell = \lceil \log_2 k \rceil$ singular vectors of $A_n$, $u_2, u_3, \ldots u_{\ell+1}$ and $v_2, v_3, \ldots v_{\ell+1}$
   and form the matrix $Z$ as in (4.14).
3. Run the $k$-means algorithm on the $\ell$-dimensional data $Z$ to obtain
   the desired $k$-way multipartitioning.
```

How Do We Know That the Dhillon 2001 Strategy is Worthwhile - I

- Confusion Matrix Performance Measures
  - Inderjit S. Dhillon, “Co-clustering documents and words using Bipartite Spectral Graph Partitioning,” KDD 2001.
  - Inderjit S. Dhillon, “ Co-clustering documents and words using Bipartite Spectral Graph Partitioning,” Ut CS Technical Report # TR 2001-05.
  - These were obtained using “mixtures” of MEDLINE (medical database), CISI (Institute of Scientific Information database), and CRANFIELD (document searching database) document sets along with YAHOO_K5 (Reuter News Articles from Yahoo where words are stemmed and heavily pruned) and YAHOO_K1 (Reuters News Articles from Yahoo: words are stemmed and only stop words are pruned)

How Do We Know That the Dhillon 2001 Strategy is Worthwhile - II

- Confusion matrix performance on the
  - Science News
  - ONR ILIR Data

- Theoretical results that insure us that the spectral based approach is a good approximation to solving the NP-compete problem.

Iterated Bipartite Bipartition Methodology

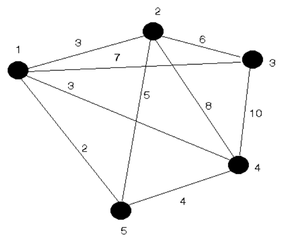
- Alternative to the multipartition approach.
- Iteratively use the bipartite bipartition methodology to obtain a multipartition of the data.
- Which cluster to split next is currently based on a simple mean distance of all observations to the centroid measure.
  - Certainly could be the subject of a more advanced statistical methodology.
- A visualization framework for exploration of the clusters (documents and words) and their associated concepts is provided.

Inherent Dimensionality of the Projected Data

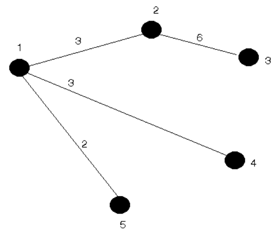
- Multipartition
  - Moderately low dimensional space $\log_2(k)$
- Recursive Bipartition
  - Set of one-dimensional spaces
- Use minimal spanning trees to facilitate layout and exploration of the documents associated with each cluster.

The Minimal Spanning Tree (MST): A Strategy for Effective Exploration of the Interpoint Distance Matrix and Cluster Computation

- Definition (Minimal Spanning Tree (MST)) – The collection of edges that join all of the points in a set together, with the minimum possible sum of edge values. The edge values that will be used here is the distance measures stored in our interpoint distance matrix.



A complete graph.



Associated MST.

Implementation Details

- JAVA
  - Originally implemented as an application
  - Currently being implemented as an applet for transition to the ONR Science and Technology website.

- JAVA Matrix Libraries Used
  - JAMA
  - JMP

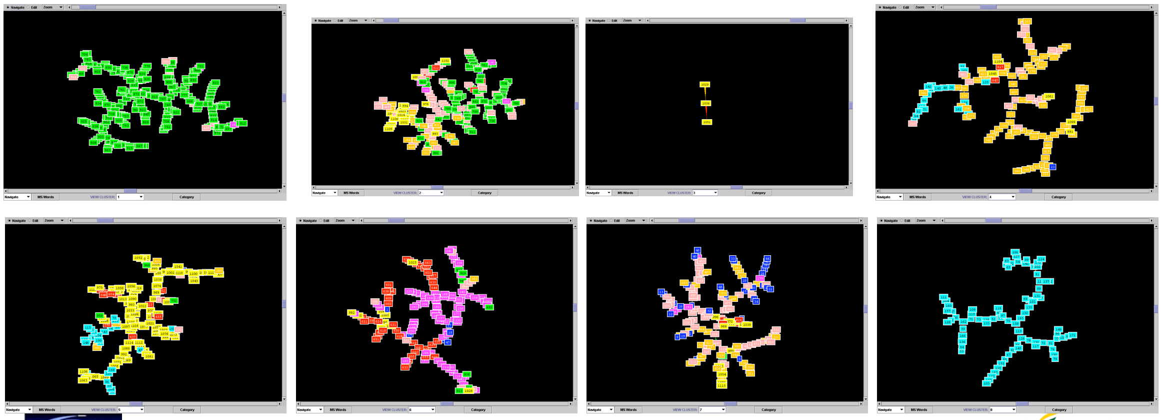
- TouchGraph

TouchGraph

- TouchGraph is a general public license JAVA-based library for the visualization of graphs. (www.touchgraph.com)
- Graph layout in TouchGraph:
  - When a graph is first loaded, nodes start out at the center with slightly random positions, and then spread out because of node-node repulsions.
- Graph manipulation tools provided by TouchGraph.
  - Zooming.
  - Rotation.
  - Hyperbolic manipulation.
  - Graph dragging.

Science News 8 Multi-partitioning

ANTHROPOLOGY & ARCHEOLOGY
BEHAVIOR
LIFE SCIENCES
MEDICAL SCIENCES
ASTRONOMY & SPACE SCIENCES
EARTH & ENVIRONMENTAL SCIENCES
MATHEMATICS & COMPUTERS
PHYSICAL SCIENCE & TECHNOLOGY



Science News 8 Multi-Partitioning Confusion Matrix

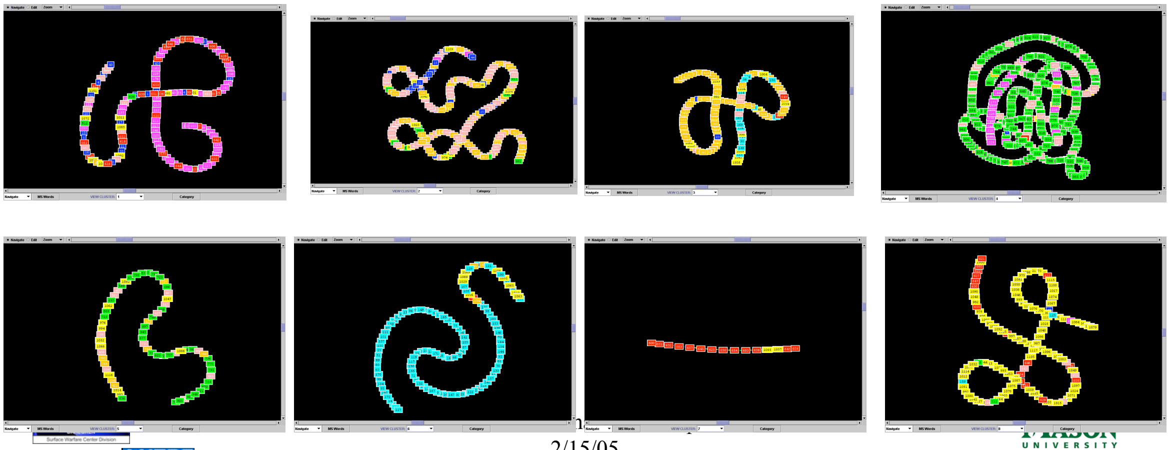
	Class1	Class2	Class3	Class4	Class5	Class6	Class7	Class8
Cluster1	0	0	3	0	9	0	208	0
Cluster2	2	0	5	32	77	4	91	20
Cluster3	0	0	0	0	0	0	0	3
Cluster4	1	19	0	89	18	2	0	5
Cluster5	1	25	0	6	5	12	3	95
Cluster6	7	0	75	1	8	43	7	9
Cluster7	37	0	5	36	57	4	1	12
Cluster8	0	80	0	0	0	0	0	0

Class 1 is anthropology and archaeology, class 2 astronomy and space sciences, class 3 is behavior, class 4 is earth and environmental sciences, class 5 is life sciences, class 6 is mathematics and computers, class 7 is medical sciences, and class 8 is physical sciences and technology.

Science News 8 Recursive Bi-partitioning

ANTHROPOLOGY & ARCHEOLOGY
BEHAVIOR
LIFE SCIENCES
MEDICAL SCIENCES

ASTRONOMY & SPACE SCIENCES
EARTH & ENVIRONMENTAL SCIENCES
MATHEMATICS & COMPUTERS
PHYSICAL SCIENCE & TECHNOLOGY

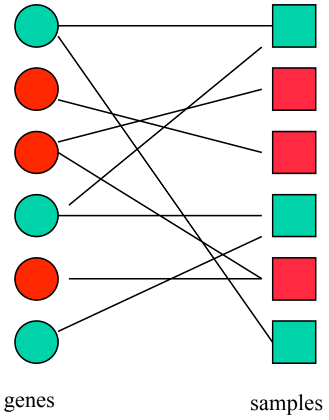


Science News 8 Recursive-Bipartitioning Confusion Matrix

	Class1	Class2	Class3	Class4	Class5	Class6	Class7	Class8
Cluster1	19	0	55	1	10	29	2	9
Cluster2	25	0	3	69	79	0	10	9
Cluster3	2	20	0	75	10	2	0	13
Cluster4	1	0	29	8	57	3	263	4
Cluster5	0	0	0	8	13	0	33	11
Cluster6	0	102	0	3	0	1	0	9
Cluster7	0	0	0	0	0	13	0	2
Cluster8	1	2	1	0	5	17	2	87

Class 1 is anthropology and archaeology, class 2 astronomy and space sciences, class 3 is behavior, class 4 is earth and environmental sciences, class 5 is life sciences, class 6 is mathematics and computers, class 7 is medical sciences, and class 8 is physical sciences and technology.

Vertex Formulation of a Gene Expression Data Set



Bipartition Algorithm for Gene Expression Data

- Term weighting scheme for edge weight
  - $E_{ij} = t_{ij} * \log(|D| / |D_j|)$ where

    t_{ij} is expression in cell t_{ij} of matrix

    D is the number of samples

    D_{ij} is the number of samples for gene i that have expression > noise

    noise was chosen at avg. diff=50 after testing increments of 25, 50, 100, & 200
- $A_{ij} = E_{ij}$
- Compute diagonal matrices D_1 and D_2
  - $D_1 = \sum A_{ij}$ for sum of gene edge-weights
  - $D_2 = \sum A_{ij}$ for sum of sample edge-weights
- Compute normalized matrix, A_n
  - $A_n = D_1^{-1/2} A D_2^{-1/2}$
- Calculate second left and right singular vectors of A_n
  - u_2 and v_2 are obtained from SVD of A_n
- Vector z_2 is formed
  - $z_2 = [D_1^{-1/2} u_2 \; D_2^{-1/2} v_2]$
- Calculate *k*-means clustering of vector z_2

Implementation Details

- Developed software was implemented using Bioconductor and R
- http://www.bioconductor.org/
- http://lib.stat.cmu.edu/R/CRAN/

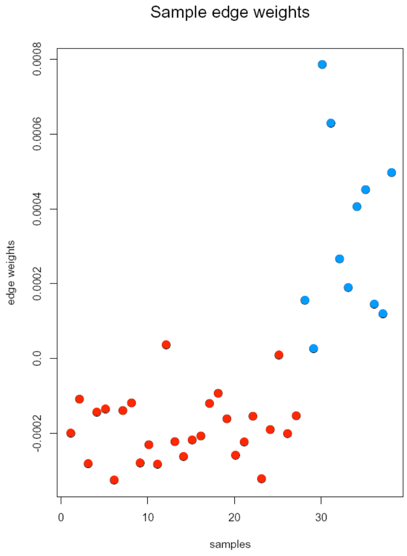
Golub Data

- Golub et al., “Molecular Classification of Cancer: Class Discovery and Class Prediction by ...,” *Science* 1999 286: 531-537
- 7129 gene expression values measured on 72 leukemia patients
- ALL
  - T and B cell variant
- AML

Results from the Golub Training Data Set Using all 7,129 genes

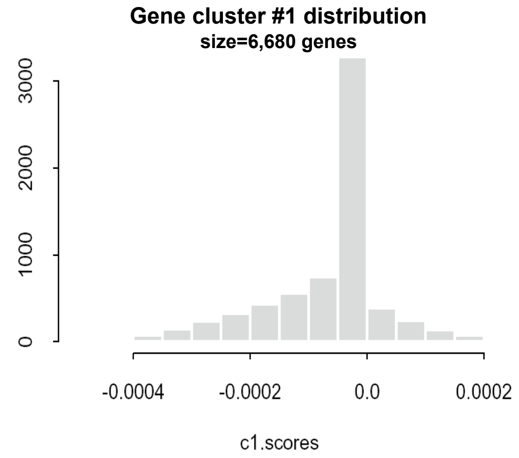
Sample confusion matrix

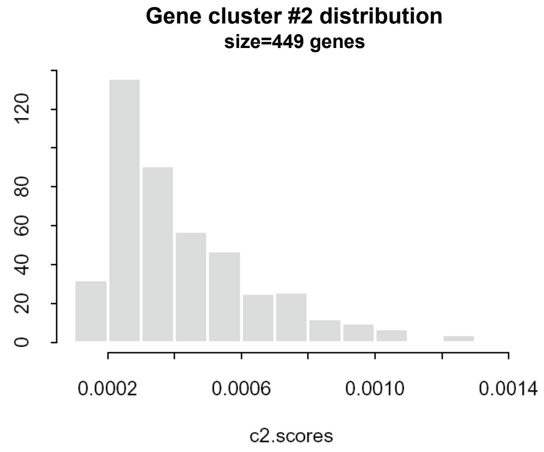
	ALL	AML
S_0	27	4
S_1	0	7



Distribution of gene internal edge weights

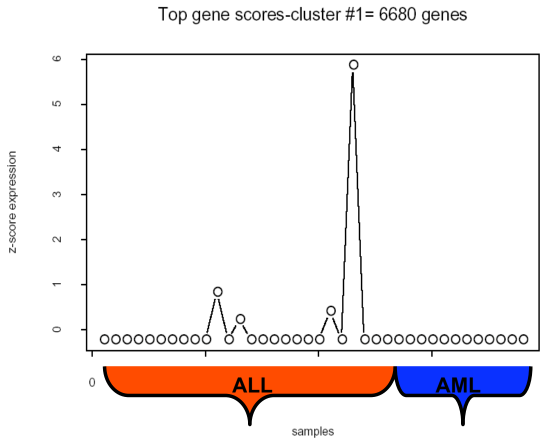
internal edge weight = $(D_1^{-1/2}u_2)$

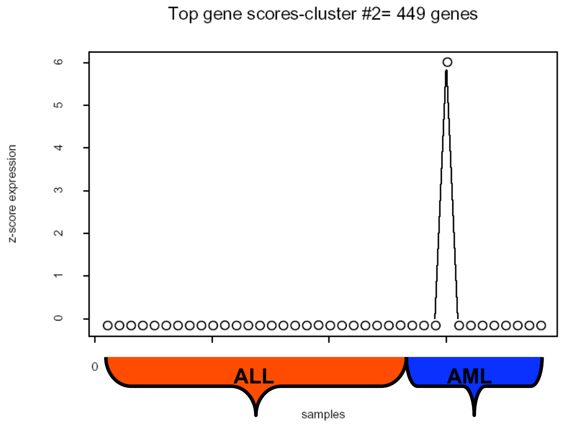




Gene profiles from genes with top ranking internal edge weights

internal edge weight = $(D_1^{-1/2}u_2)$





Issues

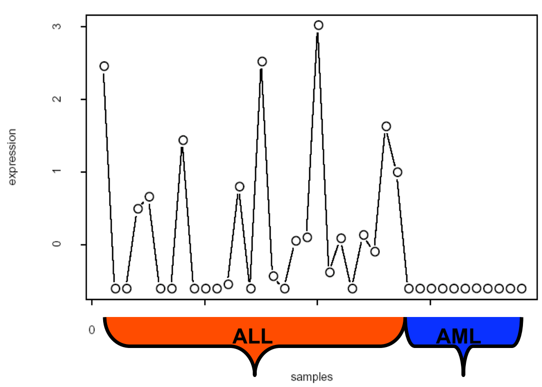
- When using all 7,129 genes, the highest ranking gene scores for each cluster are sensitive to extreme expression values
  - As depicted by the peaks in the previous plots
  - These two genes represent the most negative internal edge weighted gene from cluster #1 and the most positive internal edge weighted gene from cluster #2

- These misleading genes can be handled by a few possible approaches:
  - Feature selection prior to bipartitioning to find genes with somewhat consistent variance in each cluster
    - **Example results on next slide**
  - Preliminary filtering to remove genes with expression peaks in few samples
  - Some down-weighting scheme (e.g regression) applied to the final gene scores to penalize those genes with few sample peaks

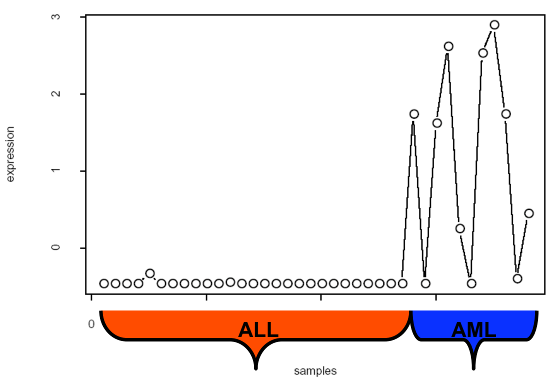
Gene profiles from genes with top ranking internal edge weights

internal edge weight = $(D_1^{-1/2}u_2)$

Top gene scores-cluster #1= 306 genes



Top gene scores-cluster #2= 29 genes



Top 335 genes that discriminate the 2 classes were first selected prior to the bipartition algorithm

Confusion matrix for samples is not shown here, but accuracy was perfect

Analysis Strategy - I

- Use all samples (72) and repeat gene selection with t-test and bipartitioning (raw MAS 4 expression data)
  - Try alternative edge weighting scheme
  - See how well samples partition
- Look for biological relevance in my top 30 scoring genes from each class and paper cluster LG1 (60 genes). Also look at intersection between my genes and paper genes
  - Results show only 20 genes intersect out of total of my 571 genes
  - Gene selections differ, so might not expect strong intersection
  - Text mining method using Bioconductor packages
- Use the n genes from the ALL cluster to attempt to divide the ALL samples into B-cell and T-cell classes
  - 264 ALL genes

Analysis - II

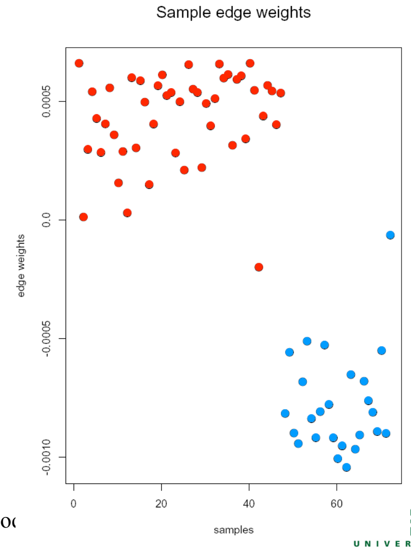
- Use paper's gene filtering method and normalization to repeat clustering (1753 genes)
  - Poor results partition only 1 gene in class 2 (no samples)
    - Removal of gene and repeat of method only partitions 4 genes (no samples)
    - Since normalization is essentially mean centering and scaling to sd=0.11, it is sensitive to genes with large contribution to variance in the second singular vector (from SVD)
    - The magnitude of one aberrant expression value for a gene is increased with the normalization scheme. This stands out as a max in the edge-weighted matrix and subsequently in the second singular vector
    - Edge weighting scheme becomes more important since the D_i term (number of documents that contain word i) will be the same value for each gene (scaled the same).
    - Attempted alternative edge weighting scheme and received similar results
- Use paper's 1753 genes and raw MAS 4 expression data with bipartitioning method and multipartitioning method
  - Run this with my modified edge weight scheme (as done in previous work)

Golub Training & Test Data Sets Using 571 t-test Genes ($p<0.001$)

*Sample confusion matrix

	ALL	AML
S_0	45	1
S_1	2	24

*Both edge weighting schemes gave same classifications



Evaluation of the Biological Relevance of Genes – I (Text Processing Reenters the Picture)

- Too many genes to look up individually, so require a more heuristic search method to determine the biological relevance of the genes as they apply to leukemia
- Pick top scoring genes from AML and ALL classes from our 571 gene set
  - Most indicative of separation between AML and ALL samples
  - Choosing 30 from each class give the same number as in the LG1 cluster from the Getz, Levine, and Domany 2000 paper (60)
- Using packages and metafiles in Bioconductor a script was written that queries PubMed abstracts and returns the PubMed ID of the instances where the query gene is cited in the abstract
  - Required R libraries
    - Annotate
    - XML
    - hu6800

Evaluation of the Biological Relevance of Genes – II (Text Processing Reenters the Picture)

- Use information from the class labels (AML/ALL) as additional query terms to determine co-occurrences of both the gene of interest and the associated class label term
  - Class terms: "lymphoblastic", "leukemia", "myeloblastic", "acute lymphoblastic leukemia", "acute myeloblastic leukemia"
- Write out to an incidence matrix where the cell value is indicative of the number of abstracts that the term appears with the gene, divided by the total number of abstracts that the gene appears in (percentage)
  - This percentage protects against a gene that may be in say 50 abstracts, but only co-occurs/is associated with a search term 5 times (incidence value would be is 0.10)
  - The opposite is a gene that is only in say 3 abstracts, but co-occurs with a search term in all 3 abstracts (incidence value would be 1)
  - Matrix dimensions are gene-by-search term

Gene-by-Search Term Incidence Matrices

Paper genes from cluster LG1

Genes	lymphoblastic	leukemia	myeloblastic	acute lymphoblastic leukemia	acute myeloblastic leukemia
AF006084_at	0.0%	0.0%	0.0%	0.0%	0.0%
D16217_at	0.0%	0.0%	0.0%	0.0%	0.0%
D29963_at	0.0%	10.5%	0.0%	0.0%	0.0%
D49400_at	0.0%	0.0%	0.0%	0.0%	0.0%
D78361_at	0.0%	0.0%	0.0%	0.0%	0.0%
HG2059-HT2114_at	0.0%	0.0%	0.0%	0.0%	0.0%
HG2279-HT2375_at	0.0%	0.0%	0.0%	0.0%	0.0%
HG2788-HT2896_at	0.0%	7.7%	0.0%	0.0%	0.0%
J02923_at	0.0%	0.0%	0.0%	0.0%	0.0%
J04162_at	0.0%	0.0%	0.0%	0.0%	0.0%
J04173_at	0.0%	0.0%	0.0%	0.0%	0.0%
J05272_at	0.0%	0.0%	0.0%	0.0%	0.0%
K01396_at	0.0%	0.0%	0.0%	0.0%	0.0%
L12168_at	0.0%	0.0%	0.0%	0.0%	0.0%
L19437_at	0.0%	14.3%	0.0%	0.0%	0.0%
L42373_at	0.0%	0.0%	0.0%	0.0%	0.0%
M11147_at	0.0%	8.3%	0.0%	0.0%	0.0%
M19722_at	0.0%	0.0%	0.0%	0.0%	0.0%
M23197_at	0.0%	28.6%	0.0%	0.0%	0.0%
M23254_at	0.0%	0.0%	0.0%	0.0%	0.0%
M32315_at	0.0%	0.0%	0.0%	0.0%	0.0%
M55067_at	0.0%	0.0%	0.0%	0.0%	0.0%
M62783_at	0.0%	0.0%	0.0%	0.0%	0.0%
M63138_at	0.0%	0.0%	0.0%	0.0%	0.0%
M63167_at	0.0%	1.6%	0.0%	0.0%	0.0%
M91029_cds2_at	0.0%	0.0%	0.0%	0.0%	0.0%
U02570_at	0.0%	0.0%	0.0%	0.0%	0.0%
U06681_at	0.0%	6.7%	0.0%	0.0%	0.0%
U10868_at	0.0%	0.0%	0.0%	0.0%	0.0%
U40282_at	0.0%	0.0%	0.0%	0.0%	0.0%
U44772_at	0.0%	0.0%	0.0%	0.0%	0.0%
U46692_rna1_at	0.0%	0.0%	0.0%	0.0%	0.0%
U50136_rna1_at	0.0%	20.0%	0.0%	0.0%	0.0%
U50523_at	0.0%	0.0%	0.0%	0.0%	0.0%
U51336_at	0.0%	0.0%	0.0%	0.0%	0.0%
U59878_at	0.0%	33.3%	0.0%	0.0%	0.0%
U89336_cds1_at	0.0%	0.0%	0.0%	0.0%	0.0%
X16663_at	0.0%	0.0%	0.0%	0.0%	0.0%
X56494_at	0.0%	0.0%	0.0%	0.0%	0.0%
X62320_at	0.0%	6.7%	0.0%	0.0%	0.0%
X67698_at	0.0%	0.0%	0.0%	0.0%	0.0%
X69550_at	0.0%	0.0%	0.0%	0.0%	0.0%
X76538_at	0.0%	0.0%	0.0%	0.0%	0.0%
X98411_at	0.0%	0.0%	0.0%	0.0%	0.0%
Z32765_at	0.0%	3.0%	0.0%	0.0%	0.0%
Z36531_at	0.0%	0.0%	0.0%	0.0%	0.0%
U34587_at	0.0%	0.0%	0.0%	0.0%	0.0%
U53204_at	0.0%	0.0%	0.0%	0.0%	0.0%
D43682_s_at	0.0%	0.0%	0.0%	0.0%	0.0%
X55448_cds1_s_at	0.0%	0.0%	0.0%	0.0%	0.0%
HG2090-HT2152_s_at	0.0%	0.0%	0.0%	0.0%	0.0%
HG4264-HT4534_s_at	0.0%	0.0%	0.0%	0.0%	0.0%
U01691_s_at	0.0%	0.0%	0.0%	0.0%	0.0%
M32304_s_at	0.0%	0.0%	0.0%	0.0%	0.0%
L35249_s_at	0.0%	0.0%	0.0%	0.0%	0.0%
U45448_s_at	0.0%	0.0%	0.0%	0.0%	0.0%
J03801_f_at	0.0%	6.7%	0.0%	0.0%	0.0%
M19045_f_at	0.0%	6.7%	0.0%	0.0%	0.0%
X14008_rna1_f_at	0.0%	6.7%	0.0%	0.0%	0.0%
X16546_at	0.0%	0.0%	0.0%	0.0%	0.0%

Our genes (top scoring 30 from each cluster)

Top scoring ALL genes

Genes	lymphoblastic	leukemia	myeloblastic	acute lymphoblastic leukemia	acute myeloblastic leukemia
M29474_at	0.0%	0.0%	0.0%	0.0%	0.0%
X82240_rna1_at	11.1%	66.7%	0.0%	11.1%	0.0%
K01911_at	0.0%	0.0%	0.0%	0.0%	0.0%
Z49194_at	0.0%	14.3%	0.0%	0.0%	0.0%
M89957_at	0.0%	14.3%	0.0%	0.0%	0.0%
L33930_s_at	7.7%	7.7%	0.0%	7.7%	0.0%
U33822_at	0.0%	0.0%	0.0%	0.0%	0.0%
D88270_at	0.0%	0.0%	0.0%	0.0%	0.0%
S76617_at	0.0%	0.0%	0.0%	0.0%	0.0%
U36922_at	0.0%	5.0%	0.0%	0.0%	0.0%
X99920_at	0.0%	0.0%	0.0%	0.0%	0.0%
M11722_at	0.0%	20.0%	0.0%	0.0%	0.0%
U07139_at	0.0%	0.0%	0.0%	0.0%	0.0%
X93512_at	0.0%	0.0%	0.0%	0.0%	0.0%
X55740_at	0.0%	16.7%	0.0%	0.0%	0.0%
L14848_s_at	0.0%	4.5%	0.0%	0.0%	0.0%
M54992_at	0.0%	20.0%	0.0%	0.0%	0.0%
U03911_at	0.0%	5.7%	0.0%	0.0%	0.0%
L06175_at	0.0%	0.0%	0.0%	0.0%	0.0%
U46006_s_at	0.0%	0.0%	0.0%	0.0%	0.0%
U05259_rna1_at	0.0%	0.0%	0.0%	0.0%	0.0%
M38690_at	13.0%	17.4%	0.0%	13.0%	0.0%
M28170_at	0.0%	0.0%	0.0%	0.0%	0.0%
AF009426_at	0.0%	0.0%	0.0%	0.0%	0.0%
U89922_s_at	0.0%	0.0%	0.0%	0.0%	0.0%
D14678_at	0.0%	0.0%	0.0%	0.0%	0.0%
L08895_at	0.0%	0.0%	0.0%	0.0%	0.0%
U40572_at	0.0%	0.0%	0.0%	0.0%	0.0%
X17025_at	0.0%	0.0%	0.0%	0.0%	0.0%
M63488_at	0.0%	0.0%	0.0%	0.0%	0.0%

Top scoring AML genes

Genes	lymphoblastic	leukemia	myeloblastic	acute lymphoblastic leukemia	acute myeloblastic leukemia
M84526_at	0.0%	0.0%	0.0%	0.0%	0.0%
M27891_at	0.0%	4.5%	0.0%	0.0%	0.0%
D88422_at	0.0%	0.0%	0.0%	0.0%	0.0%
U05572_s_at	0.0%	0.0%	0.0%	0.0%	0.0%
M28130_rna1_s_at	0.0%	2.3%	0.6%	0.0%	0.6%
M96326_rna1_at	0.0%	0.0%	0.0%	0.0%	0.0%
M19507_at	0.0%	29.7%	0.0%	0.0%	0.0%
Y00787_s_at	0.0%	2.3%	0.6%	0.0%	0.6%
X52056_at	0.0%	29.4%	0.0%	0.0%	0.0%
M83652_s_at	0.0%	0.0%	0.0%	0.0%	0.0%
U46499_at	0.0%	0.0%	0.0%	0.0%	0.0%
M92357_at	0.0%	25.0%	0.0%	0.0%	0.0%
M57710_at	3.4%	3.4%	0.0%	0.0%	0.0%
X95735_at	0.0%	25.0%	0.0%	0.0%	0.0%
X90858_at	0.0%	0.0%	0.0%	0.0%	0.0%
J03909_at	0.0%	0.0%	0.0%	0.0%	0.0%
HG3494-HT3688_at	0.0%	2.3%	0.0%	0.0%	0.0%
D10495_at	0.0%	3.1%	0.0%	0.0%	0.0%
X05908_at	16.7%	16.7%	0.0%	8.3%	0.0%
X07743_at	0.0%	0.0%	0.0%	0.0%	0.0%
M59820_at	0.0%	35.3%	0.0%	0.0%	0.0%
X61587_at	0.0%	0.0%	0.0%	0.0%	0.0%
X64072_s_at	0.0%	0.0%	0.0%	0.0%	0.0%
L09209_s_at	0.0%	0.0%	0.0%	0.0%	0.0%
M98399_s_at	0.0%	3.0%	0.0%	0.0%	0.0%
M23197_at	0.0%	28.6%	0.0%	0.0%	0.0%
U70063_at	0.0%	0.0%	0.0%	0.0%	0.0%
M80254_at	0.0%	0.0%	0.0%	0.0%	0.0%
X16546_at	0.0%	0.0%	0.0%	0.0%	0.0%
L19872_at	0.0%	3.1%	0.0%	0.0%	0.0%

□ Not in any abstract

■ In at least 1 abstract

LG1 was obtained from Getz, Levine, and Domany 2000.

Golub Training & Test Data Sets

From the bipartitioning, 264 genes were grouped in the ALL cluster and 307 genes were grouped in the AML cluster

Using only the 264 genes and the 47 ALL samples, try to partition the B-cell and T-cell subclasses (results below)

B-cell class contains 189 genes

T-cell class contains 75 genes

Sample confusion matrix

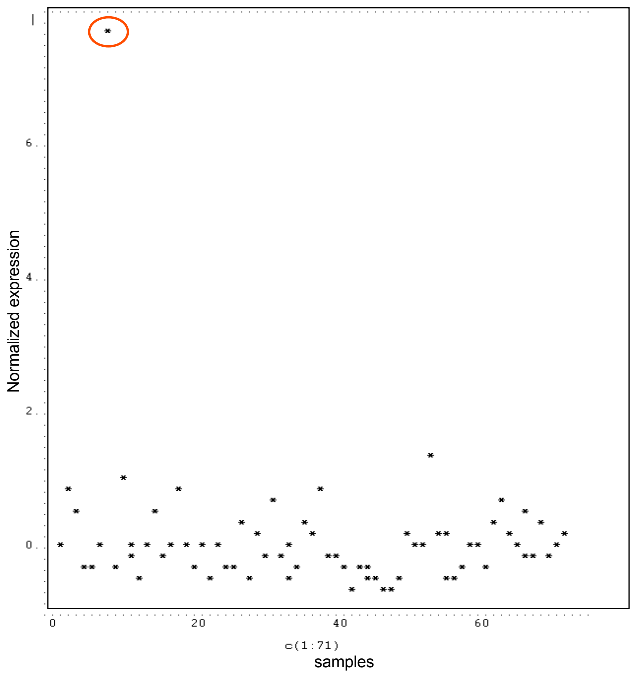
	B-cell	T-cell
S_0	35	0
S_1	3	9

Getz, Levine, and Domany 2000 Normalization Issue

This particular gene has been normalized by the paper's method. This method essentially mean centers each gene with sd=0.11.

The problem in applying the bipartition algorithm to genes that have been scaled this way is that it requires a SVD on the edge-weight matrix, such that this type of gene will stand out in the second singular vector from the SVD.

When one goes to then *k*-means cluster this 1-D vector, this gene will be assigned to its own cluster, since it's value far exceeds any other



Golub Training & Test Data Sets

Using Paper's 1753 Genes and Our Original Edge Weight Scheme

Bipartition method
sample confusion matrix

	ALL	AML
S_0	47	8
S_1	0	17

Multipartition method
sample confusion matrix

	B-cell	T-cell	AML
S_0	22	3	0
S_1	2	6	0
S_2	14	0	25

Golub Training & Test Data Sets

Using Paper's 1753 Genes and Alternative Edge Weight Scheme

Bipartition method
sample confusion matrix

	ALL	AML
S_0	28	1
S_1	19	24

Multipartition method
sample confusion matrix

	B-cell	T-cell	AML
S_0	27	5	7
S_1	11	4	0
S_2	0	0	18

Preliminary Interpretations - I

- The Getz, Levine, and Domany 2000 paper's normalization scheme seems difficult to implement into bipartition/multipartition algorithm
  - Essentially mean centers data, so likelihood of a word occurring in a document is equal across all words (genes)
  - Magnified outlier issue discussed on the previous slide

- The paper's filtered 1753 genes don't provide the optimal sample partitioning in 2 or 3 classes
  - Using raw MAS 4 data or paper normalized data

Preliminary Interpretations - II

- Best attempt to resolve 3 classes comes from:
  - Feature selection on 2 classes (ALL/AML)
  - Use either edge weighting schemes (similar results)
    - Our original edge weight scheme: set noise < 50
    - Our modified edge weight scheme

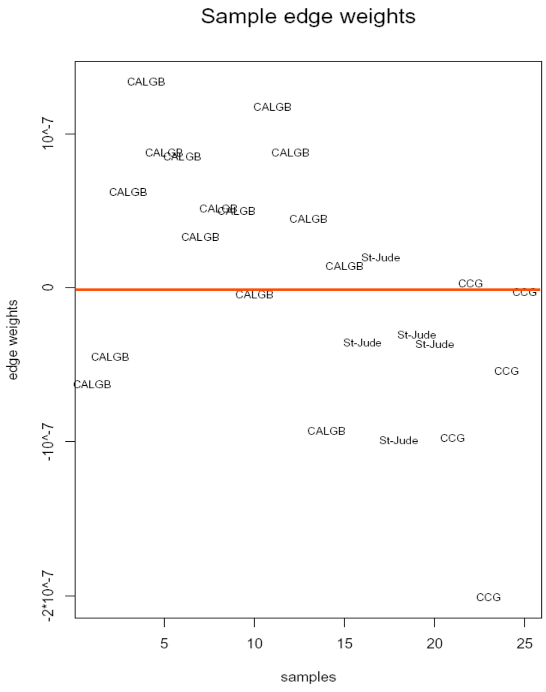
- Biological relevance of genes that partition AML and ALL is greater in these 60 genes than the Getz, Levine, and Domany 2000 top gene discriminators (LG1 genes)
    - Many more hits in our incidence matrix vs. the paper's cluster LG1

- Bipartition method implemented on raw MAS 4 data
    - Implemented once on ALL/AML samples
    - Implemented a second time on ALL samples, using ALL cluster genes

Additional Analysis

- Attempted to bipartition the AML samples using the 264 AML genes to partition the treatment effect
  - Similar to GLD paper's cluster LS2/LG4
- Examined the biological relevance of gene clusters from AML bipartition, as compared to GLD paper
  - GLD claims many ribosomal proteins and cell growth-related genes in cluster LG4
- Built a binary tree using the bipartition algorithm at each branch
- Used the Alon colon cancer data set to bipartition the normal and tumor samples

Bipartitioning on AML Samples to Reveal Treatment



Bipartition on the 264 AML genes using only the 25 AML samples and Dr. Solka's edge-weight scheme performs as follows:

11/15 treated patients (CALGB) partition into group #1 (GLD paper has 14/15)

1 St-Jude patient partitions into group #1

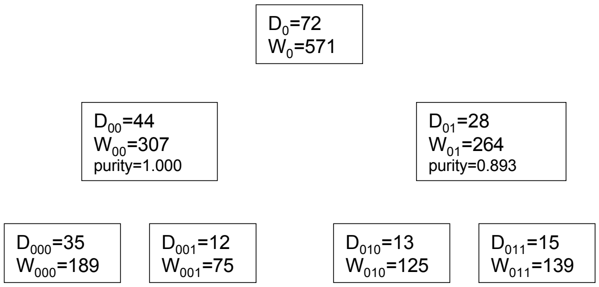
1 CCG patient partitions into group #1

Concerned about confounding factor of hospital vs. treatment since all treated patients are stratified on same location

Genes from W_0 and W_1 have many related to DNA replication/repair and cellular growth/proliferation.

Similar to GLD 16 gene cluster in cellular

Iterative Descent Tree on the Golub Data



$D_0=72$
$W_0=571$

$D_{00}=44$
$W_{00}=307$
purity=1.000

$D_{01}=28$
$W_{01}=264$
purity=0.893

$D_{000}=35$
$W_{000}=189$

$D_{001}=12$
$W_{001}=75$

$D_{010}=13$
$W_{010}=125$

$D_{011}=15$
$W_{011}=139$

D_{00}=45/47 ALL
D_{01}=24/25 AML

D_{010}=11/15 CALGB
D_{000}=35/38 B-cell
D_{001}=9/9 T-cell

Alon colon cancer data

using all 2,000 genes and 97 t-test genes (p<0.001)

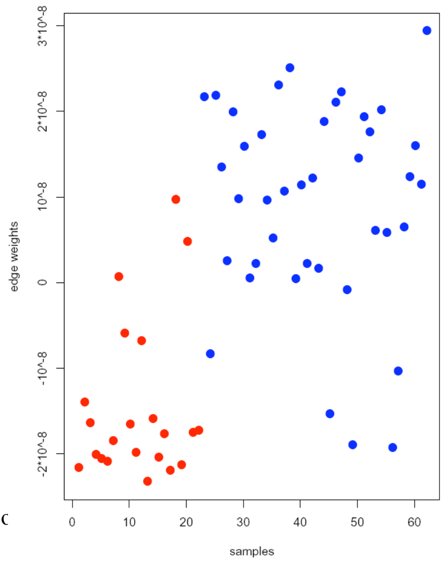
2,000 genes sample confusion matrix

	Norm	Tum
S_0	15	13
S_1	7	27

97 genes sample confusion matrix

	Norm	Tum
S_0	17	4
S_1	5	36

Sample edge weights



Future

- Development of visualization frameworks that allow for simultaneous display of words and documents (genes and samples).
- Tree-based displays for the recursive bipartitioning tree.
- Higher dimensional visualization in the case of the multipartition algorithm.
- Additional applications of the iterative methodology to gene expression data.

Conclusions

- Demonstrated extensions and new applications of the Dhillon 2001 spectral based clustering methodology.
- Tested the method on example text mining dataset
  - Science News dataset
- Tested the method on two gene expression datasets.
  - Golub leukemia
    - Use text-based analysis to evaluate the “significance” of the discovered genes
    - Compared results to those obtained in Getz, levine, and Domany 2000
  - Alon cancer

References - I

- Alon U, Barkai N, Notterman DA, Gish, K, Ybarra, S. Mack, D and Levine, AJ. ,"Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," *Proc. Natl. Acad. Sci.* USA. 96 (1999) 6745-6750.
- Inderjit S. Dhillon, "Co-clustering documents and words using Bipartite Spectral Graph Partitioning," *KDD* 2001.
- Getz G, Levine E, and Domany E. "Coupled two-way clustering analysis of gene microarray data." *Proc Natl Acad Sci USA* 97: 12079–12084, 2000.
- Golub et al., "Molecular Classification of Cancer: Class Discovery and Class Prediction by ...," *Science* 1999 286: 531-537.
- J. L. Solka and Brandon Higgs, " From Text Data Mining to Gene Data Mining and Back Again," *invited presentation at Joint Annual Meeting of the Interface and the Classification Society of North America: Theme: Clustering and Classification*, Washington University School of Medicine St. Louis, Missouri, June 8, 2005 - June 12, 2005.

References - II

- J. L. Solka, A. C. Bryant, and Edward J. Wegman, "Text Data Mining With Minimal Spanning Trees," *in Handbook of Statistics 24 on Data Mining and Visualization*, C. R. Rao, Edward J. Wegman, and J. L. Solka, Eds, Elsevier North Holland, 2005.
- J. L. Solka, A. C. Bryant, and E. J. Wegman, "Identifying Cross Corpora Document Associations Via Minimal Spanning Trees," *Proceedings Interface 2004: Computational Biology and Bioinformatics 36th Symposium on the Interface*, May 26-29, 2004.

GMU Bioinformatics Colloquium
2/15/05