MICROARRAY DATA ANALYSIS OF SURVIVAL TIMES OF PATIENTS WITH LUNG ADENOCARCINOMAS USING ADC AND K-MEDIANS CLUSTERING

Wenting Zhou, Weichen Wu, Nathan Palmer, Emily Mower, Noah Daniels, Lenore Cowen, and Anselm Blumer
Computer Science, Tufts University, Medford, MA 02155 USA

Abstract: We experiment with two types of clustering, K-medians and a dimension-reduction technique known as Approximate Distance Clustering [Cowen and Priebe 1997], for classifying lung adenocarcinomas into high-risk and low-risk groups according to gene expression values from microarray data. The microarrays were Affymetrix oligonucleotide arrays used in studies at Michigan and Harvard, with 12,600 and 7129 probesets respectively. We show that we can obtain accurate classification based on a reduced set of genes obtained by Nearest Shrunken Mean [Tibshirani *et al*. 2002] or a combination of a variance-based approach with hierarchical clustering. The quality of the clustering is measured by using the p-values from log-rank tests, and the results are confirmed using cross-validation and by using the reduced set of genes obtained from one dataset to cluster the other.

Key words: Microarray; ADC clustering; K-medians; adenocarcinoma; survival time.

1. INTRODUCTION

This paper investigates clustering and dimension-reduction techniques on two of the four CAMDA 2003 datasets of gene expression values and survival times of patients with lung adenocarcinomas. We chose the Michigan [Beer *et al*. 2002] and Harvard [Bhattacharjee *et al*. 2001] data due to the reasonably large sample sizes (n = 86 and 84) and lack of missing values. We use ADC maps [Cowen and Priebe 1997] to project the data into one or two dimensions so we can use very simple clustering techniques, then follow this with Nearest Shrunken Mean [Tibshirani *et al*. 2002] to reduce the number of genes used to predict the clusters. We contrast this with more classical techniques of variance ratios and hierarchical clustering.

2. METHODS

2.1 Approximate Distance Clustering (ADC]

Approximate Distance Clustering is a method that reduces the dimensionality of data by calculating the distances from data points to subsets of the data points called "witness sets" [Cowen and Priebe 1997). One witness set is chosen for each desired output dimension.

It is defined as follows:

- Let **X** be a collection of data in $\mathbf{R}^m$ In this case, each data point corresponds to a gene chip, so m is 12,600 or 7,129 initially.
- Define D_1, D_2,…, D_d to be subsets of **X** of sizes k_1, k_2,…, k_d These are the witness sets.
- The associated ADC map, $f_{(D1, D2,..., Dd)} : \mathbf{R}^m \rightarrow \mathbf{R}^d$ maps **X** to (y_1, y_2,…, y_d), where $y_i = \min\{ \| x_j - x \| : x_j \varepsilon D_i\}$

In other words, data point x maps to a point in m-dimensional space with i^{th} coordinate equal to the distance from x to the nearest point in the i^{th} witness set. A good witness set is a small set of points that produces a mapping that preserves inter-cluster distances. In this paper, we look at the simplest cases of ADC projection on the microarray data: the case where the number of dimensions we project to is 1 or 2, and the size of all witness set is 1. Note that ADC does not in itself produce a clustering; the resulting points in 1 or 2 dimensions must still be classified or clustered using some method that works for low-dimensional data. In one dimension we just pick a cutoff value and assign all points below the cutoff to one cluster and all points above to the other. In two dimensions, we add the coordinates together before comparing to the cutoff. We use the following criterion to choose a good clustering from the set of allowable clusterings:

Compute the Kaplan-Meier survival curves and the p-value from the log-rank test, then use the following w-criterion:

$$w = 4000 * a + 5500 * b + 450 * (1-c) + 50 * d$$

where

- *a* is 1 if the size of smaller group is less than *n*/8, and 0 otherwise
- *b* is the p-value
- *c* is the difference between the final survival rates of the low-risk and high-risk groups
- *d* is the high-risk group's final survival rate

2.2 Nearest Shrunken Mean (NSM) Gene Reduction

After choosing the high-risk and low risk clusters using ADC clustering according to the w-criterion, we use Nearest Shrunken Mean (NSM) [Tibshirani *et al*. 2002] to eliminate genes (or probesets) that have all their cluster means close to their overall mean.

Let:

- x_{ij} be the expression of gene *i* for tissue sample *j*
- m_{ik} be the mean expression of gene *i* in class (cluster) *k*
- x_i be the mean of gene i
- n be the sample size
- K be the number of clusters
- n_k be the size of cluster k
- $s_i = (1/(n-K)) \Sigma_k \Sigma_{j \in Ck} (x_{ij}-m_{ik})^2$
- s_0 be the median of the s_i
- $M_k = \text{sqrt}(1/n_k+1/n)$
- $d_{ik} = (m_{ik} - x_i) / (m_k*(s_i+s_0))$, so
- $m_{ik} = x_i + d_{ik} * m_k*(s_i+s_0)$

In this expression, d_{ik} can be reduced by Δ in absolute value or replaced by zero if its absolute value is smaller than Δ. If it is replaced by zero, the cluster mean becomes the overall mean; if this happens for all clusters, the gene can be eliminated.

2.3 K-medians Clustering

K-medians clustering is a variation of K-means clustering where the cluster centers must be chosen from among the data points. It is an unsupervised method, so the quality of the clustering is measured just using the distances between the data points without looking at their classifications. It selects K points to be cluster centers and calculates the quality of the clustering as the sum of the distances of data points to their nearest cluster center. In this paper, we use K=2 so it is feasible to calculate the quality of all $n(n+1)/2$ clusterings and choose the optimal one.

2.4 Minimal Variance Ratio (MVR) Gene Reduction

The variance ratio is the sum of the within-cluster variances divided by the total variance of expression values for that gene. Using the notation from the NSM section above, let

- $\sigma^2_{ik} = (1/ n_k) \Sigma_{j \in Ck} (x_{ij}-m_{ik})^2$ be the within-cluster variance for gene *i* in cluster *k*

- $\sigma^2_i = (1/n)\Sigma_j(x_{ij}-x_i)^2$ be the total variance for gene *i*, then
- $(\Sigma_k \sigma^2_{ik})/\sigma^2_i$ is the variance ratio for gene *i*

Genes with large variance ratios are thought to contribute less to the cluster definitions and are eliminated.

2.5 Dimension Reduction With ADC and NSM

One set of experiments involved using one or two dimensional ADC clustering with a witness set of size one, followed by NSM to obtain a set of genes of the desired size. The *w* measure above was used to select the witness and the cutoff point between the two clusters. In the case of two dimensional ADC clustering we summed the values of the distances along the two axes to determine whether a point was below the cutoff. We also experimented with Survival-Time Cutoff Clustering (STCC), sorting the patients according to survival time and splitting them 50-50 or 60-40 into high risk – low risk clusters to replicate the results of [Beer *et al*. 2002].

2.6 Dimension Reduction With MVR, K-Medians, and Hierarchical Clustering

A second set of experiments involved starting with high-risk and low-risk clusters of equal size according to survival times (50% STCC), then using MVR to select a subset of genes to approximate this clustering. Some genes in this subset may have similar expression profiles, so a form of hierarchical clustering was used to obtain a desired number of clusters of these genes and one gene was selected from each cluster. This doubly reduced gene set was then used (after normalizing each gene profile to have vector length one) to obtain a K-medians clustering with K=2 and the p-value from the log-rank test was calculated.

3. EXPERIMENTAL RESULTS

We experimented with these methods on adenocarcinoma examples (patients) from the Michigan [Beer *et al*. 2002] and Harvard [Bhattacharjee *et al*. 2001] data that had survival times (both censored and uncensored). The Michigan data had expression values for 7,129 probesets for each of 86 examples, while the Harvard data had expression values for 12,600 probesets for each of 84 examples.

3.1 ADC on Harvard and Michigan data

Tables 1 through 4 give the results of using the w-criterion to select the best ADC witnesses and cutoffs, then reducing the set of probesets to the specified size with NSM. In all cases the witness sets had size one. The p-values were obtained from leave-one-out crossvalidation on the reduced set of probesets. Specifically, ADC clusters were formed based on the reduced set of probesets, leaving out one patient, with the best ADC clustering being selected according to the w-criterion. The excluded patient was then classified as high-risk or low-risk according to which cluster mean was closer. The values for STCC were obtained by following the same procedure but substituting clusters formed of the 50% or 60% highest risk patients for the ADC clusters.

Table 1. p-values for 1 and 2 dimensional ADC and STCC on Michigan data (n = 86)

Genes	1D ADC	2D ADC	50% STCC	60% STCC
7129	0.0028	0.0500	0.0086	0.0126
1000	0.0275	0.0009	0.0111	0.0158
500	0.0495	0.0048	0.0046	0.0089
200	0.0019	0.0033	0.0075	0.0056
100	0.0058	0.0194	0.0023	0.0048
50	0.0019	0.1442	0.0064	0.0048
40	0.0009	0.0268	0.0011	0.0048
30	0.0009	0.0356	0.0029	0.0067
20	0.0021	0.0189	0.0029	0.0090
10	0.0061	0.0618	0.0059	0.0049
5	0.0086	0.3559	0.0151	0.0024

Table 2. Low risk/high risk group sizes for 1 and 2 dimensional ADC and STCC on Michigan data (n = 86)

Genes	1D ADC	2D ADC	50% STCC	60% STCC
7129	55/31	54/32	46/40	46/40
1000	59/27	60/26	45/41	43/43
500	52/34	57/29	47/39	45/41
200	58/28	58/28	47/39	48/38
100	57/29	55/31	49/37	46/40
50	58/28	42/44	50/36	47/39
40	58/28	44/42	50/36	47/39
30	58/28	43/43	51/35	46/40
20	57/29	42/44	51/35	46/40
10	56/30	37/49	50/36	47/39

Genes	1D ADC	2D ADC	50% STCC	60% STCC
5	58/28	41/45	49/37	49/47

Table 3. p-values for 1 and 2 dimensional ADC and STCC on Harvard data (n = 84)

Genes	1D ADC	2D ADC	50% STCC	60% STCC
12600	0.0646	0.0046	0.1946	0.0741
1000	0.0124	0.0013	0.0381	0.0038
500	0.0023	0.0116	0.0021	0.0027
200	0.0121	0.0037	0.0007	0.0004
100	0.0201	0.0027	0.0213	0.0004
50	0.0332	0.0090	0.0120	0.0047
40	0.0332	0.0019	0.0100	0.0033
30	0.0898	0.0010	0.0065	0.0098
20	0.0448	0.0039	0.0083	0.0015
10	0.0424	0.0011	0.0034	0.0001
5	0.0321	0.0032	0.0053	0.0196

Table 4. Low risk/high risk group sizes for 1 and 2 dimensional ADC and STCC on Harvard data (n = 84)

Genes	1D ADC	2D ADC	50% STCC	60% STCC
12600	25/59	24/60	39/45	41/43
1000	20/64	15/69	44/40	38/46
500	21/63	22/26	42/42	36/48
200	21/63	21/63	40/44	32/52
100	24/60	26/58	42/42	30/54
50	21/63	21/63	40/44	35/49
40	21/63	27/57	40/44	35/49
30	28/56	26/58	39/45	35/49
20	27/55	26/58	38/46	34/50
10	22/62	20/64	37/47	33/51
5	20/64	25/59	36/48	28/56

Since these datasets contained multiple probesets corresponding to the same genes, we then selected the top 50 probesets corresponding to distinct genes. Tables 5 and 6 give the probeset names, gene symbols, and mean expression values in the low-risk and high-risk group for each probeset selected. It is interesting to note that in the Michigan dataset most of these 50 (all except IGKC, IGL@, IGHG3, NPC2, HLA-A, CD74, HLA-B, MGP, NBL1, GRN, and the two with NULL symbol) have lower mean expression values in the low-risk group, while in the Harvard dataset all except GAPD, CLDN9, MIF, and PSMB3 have higher mean expression values in the low-risk group.

Crossvalidation of the classification based on these expression values gave p-values of 0.0074 on the Michigan dataset and 0.0331 on the Harvard dataset. Figures 1 and 2 give the Kaplan-Meier curves corresponding to these p-values.

Table 5. Top 50 distinct genes from Michigan data. Underlined genes are also found in Table 6, bold genes are among the top 100 in [Beer *et al*. 2002].

Probeset	Symbol	Low-Risk	High-Risk
M63438_s_at	IGKC	29936.2	14461.4
M34516_at	NULL	23771.3	7285.7
X57809_s_at	IGL@	23693.4	6952.74
M87789_s_at	IGHG3	41259.8	8671.2
L19437_at	TALDO1	1352.48	2566.89
X01677_f_at	**GAPD**	8820.27	12018.6
L10678_at	PFN2	775.93	1462.43
X67698_at	NPC2	8877.69	6543.1
M21388_r_at	NULL	3370.06	2362.68
X00274_at	HLA-A	14115.9	11346.3
M13560_s_at	CD74	8951.48	6846.82
M17886_at	RPLP1	13417.8	19409.6
D49387_at	LTB4DH	372.44	1068.32
M37583_at	**H2AFZ**	1557.07	2302.42
X67951_at	PRDX1	4228.8	5964.1
X02152_at	LDHA	6607.16	8852.83
D13630_at	**KIAA0005**	1129.9	1655.69
D14874_at	**ADM**	368.88	624.67
X15940_at	RPL31	7048.57	8760.57
J03934_s_at	NQO1	481.3	1309.43
X91247_at	TXNRD1	1369.52	2603.73
X69654_at	**RPS26**	5012.86	6148.86
M22382_at	HSPD1	2687.07	3960.79
X77584_at	TXN	3019.61	4447.59
M26730_s_at	UQCRB	1783.05	2319.47
D49824_s_at	**HLA-B**	24959.3	18358.9
X15183_at	HSPCA	4756.56	6527.33
U09813_at	ATP5G3	2284.24	3336
X56468_at	YWHAQ	1832.02	2488.57
X13238_at	COX6C	1824.35	2530.02
D14657_at	KIAA0101	311.29	536.96
M22760_at	COX5A	1112.69	1458.31
D00762_at	PSMA3	1243.9	1629.8
J04823_rna1_at	COX8	4599.03	5722.32

Probeset	Symbol	Low-Risk	High-Risk
X53331_at	MGP	7151.91	4174.75
M24485_s_at	GSTP1	5788.36	8422.77
L08666_at	**VDAC2**	1480.34	2011.79
X65614_at	**S100P**	2495	6197.89
L37043_at	CSNK1E	858.41	1145.46
J04444_at	CYC1	1042.34	1524.23
M19961_at	COX5B	1631.52	2097.81
L19686_rna1_at	MIF	7390.13	8807.46
D28124_at	NBL1	4359.21	2358.11
X62320_at	GRN	3043.87	2825.88
Z14244_at	COX7B	461.46	705.04
Z49099_at	SMS	1017.55	1426.29
V00572_at	PGK1	3705.16	5137.71
U84573_at	PLOD2	555.49	710.12
U31814_at	HDAC2	421.74	611.64
HG4074-HT4344_at	FEN1	248.57	394.65

Table 6. Top 50 distinct genes from Harvard data. Underlined genes are also found in Table 5, bold genes are among the top 100 in [Beer *et al*. 2002].

Probeset	Symbol	Low-Risk	High-Risk
36627_at	SPARCL1	513.74	298.01
41723_s_at	**HLA-B**	1845.4	1001.59
38833_at	HLA-A	1936	1066.85
216_at	PTGDS	895.54	494.11
32905_s_at	TPSB2	454.99	193.21
39220_at	SCGB1A1	687.17	135.19
31525_s_at	HBA2	697.61	380.52
35905_s_at	**GAPD**	4541.9	5160.53
38691_s_at	SFTPC	4873	1276.4
32052_at	HBB	1032.3	580.48
32542_at	FHL1	121.61	52.95
1288_s_at	EEF1A1	5176.5	4636.86
35016_at	CD74	2641.5	1740.34
36097_at	ETR101	504.53	341.97
34363_at	SEPP1	322.25	182.02
1005_at	DUSP1	675.19	421.52
36634_at	BTG2	574.35	393.16
649_s_at	CXCR4	310.64	231.84
37394_at	C7	125.46	37.66
37021_at	CTSH	1988.2	1009.38

Probeset	Symbol	Low-Risk	High-Risk
33383_f_at	SFTPB	2232.6	1179.78
39864_at	CIRBP	353.61	276.04
35521_at	CLDN9	-91.05	14.98
31870_at	CD37	197.08	114.42
37168_at	LAMP3	304.36	84.97
41382_at	DMBT1	462.34	199
40607_at	DPYSL2	296.13	195.57
36495_at	FBP1	443.57	265.59
36669_at	FOSB	357.53	170.15
895_at	MIF	1270.2	1758.25
36680_at	AMY2B	242.38	56.31
534_s_at	FOLR1	782.5	449.16
36452_at	SYNPO	604.05	490.65
35183_at	ABCA3	376.66	152.54
428_s_at	B2M	3152.4	2805.04
39066_at	MFAP4	108.89	35.79
1915_s_at	FOS	1010	752.43
35926_s_at	LILRB1	1212.5	834.49
32321_at	HLA-E	481.98	365.18
34793_s_at	PLS3	321.7	217.19
35842_at	IL6ST	281.29	206.09
32786_at	JUNB	458.56	329
35730_at	ADH1B	43.05	15
31775_at	SFTPD	743.05	260.13
1117_at	CDA	312.02	209.11
1309_at	PSMB3	223.86	285.94
39345_at	NPC2	2083.5	1352.04
32597_at	RBL2	160.27	121.24
35868_at	AGER	139.54	54.72
33295_at	FY	124.02	79.66

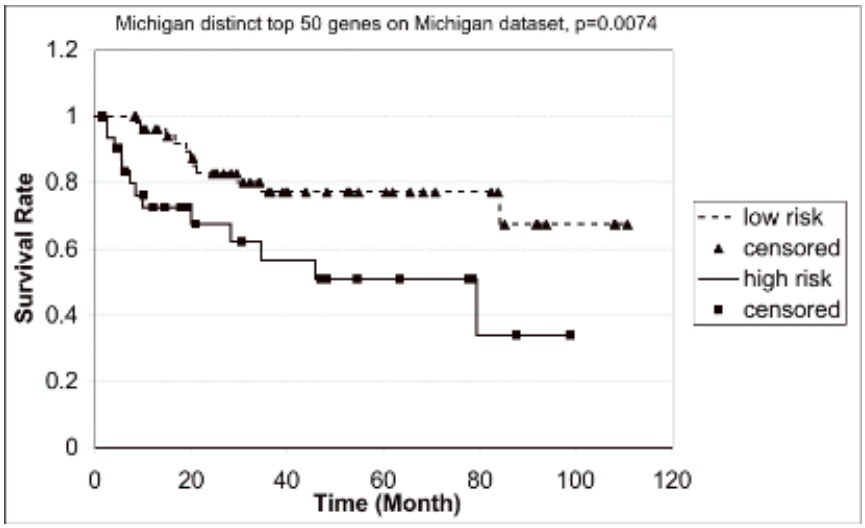

Figure 1. Kaplan-Meier curves for crossvalidation of probesets corresponding to 50 distinct genes selected from Michigan dataset, validated on Michigan dataset.

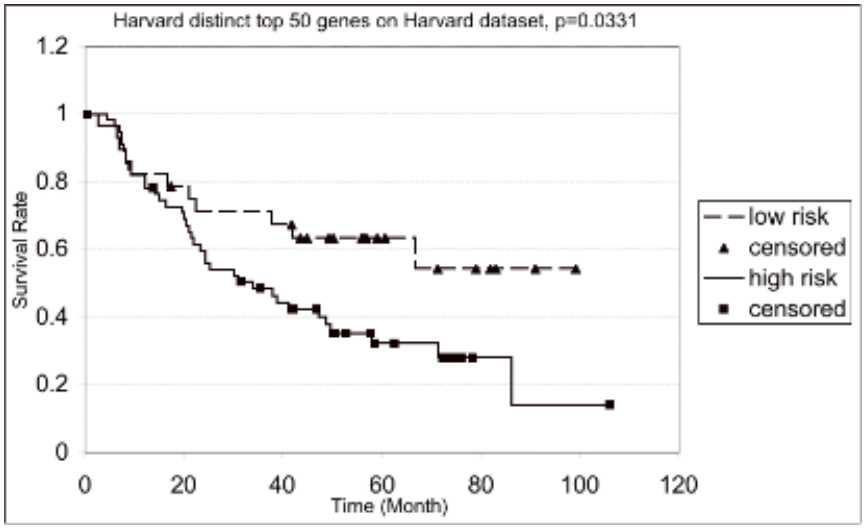

Figure 2. Kaplan-Meier curves for crossvalidation of probesets corresponding to 50 distinct genes selected from Harvard dataset, validated on Harvard dataset.

3.2 Validating ADC between Harvard and Michigan data

We also validated the groups of 50 probesets described above across datasets. Since the Michigan and Harvard studies used different gene chips, we used the probeset link table from affymetrix.com (filename PN600444HumanFLComp.zip) to find corresponding probesets in the two datasets. Starting from the top 50 probesets in the Michigan data we found the 57 matching probesets in the Harvard dataset, since the link table is not one-to-one. We then averaged probesets with the same gene symbol (including three with NULL symbol), leaving 48 distinct genes (plus NULL). We used those 49 as in the internal leave-one-out crossvalidation to classify each example as low-risk or high-risk. Testing the top Michigan probesets on the Harvard data in this way gave a p-value of 0.0254. We then reversed this procedure, starting with the top 50 Harvard probesets. This gave 42 distinct genes in the Michigan dataset (plus NULL). Using those 43 for crossvalidation on the Michigan data gave a p-value of 0.0307. Figures 3 and 4 give the Kaplan-Meier curves corresponding to these p-values.

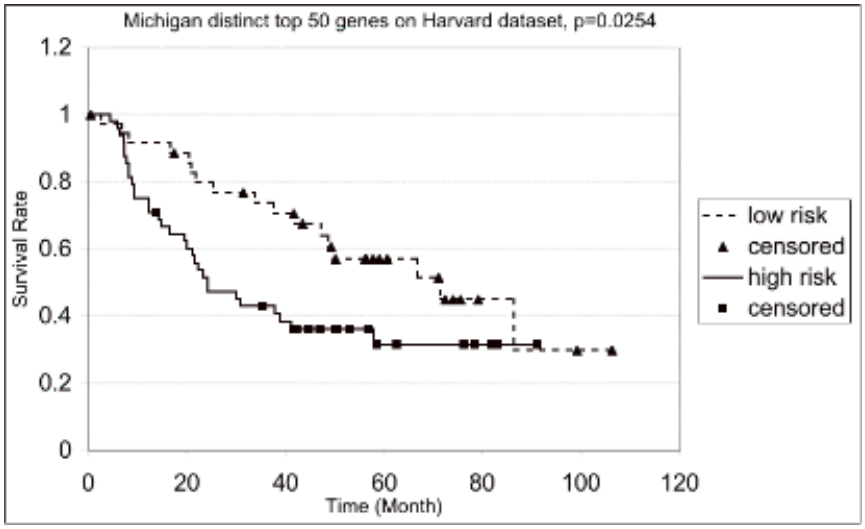

Figure 3. Kaplan-Meier curves for crossvalidation of probesets corresponding to 50 distinct genes selected from Michigan dataset, validated on Harvard dataset.

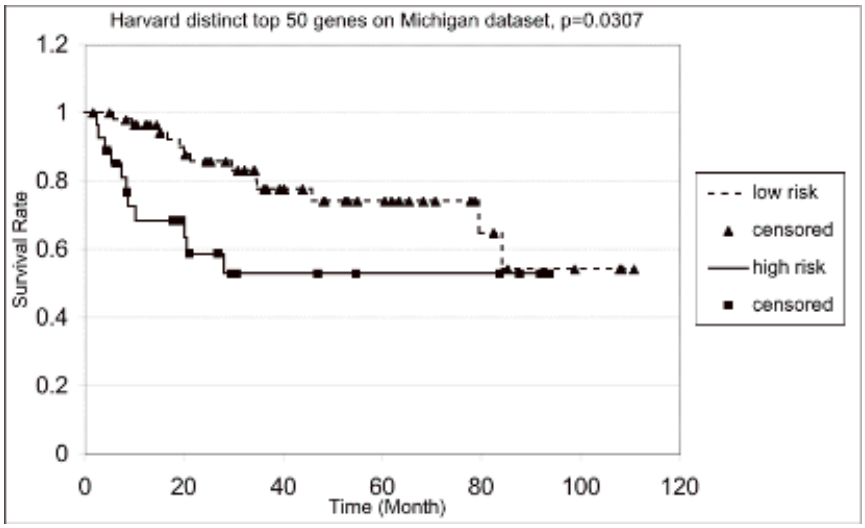

Figure 4. Kaplan-Meier curves for crossvalidation of probesets corresponding to 50 distinct genes selected from Harvard dataset, validated on Michigan dataset.

3.3 MVR and K-medians

We used Minimal Variance Ratio to select 200 probesets from the Michigan and Harvard data based on an initial 50-50 clustering according to survival times (50% STCC), then used hierarchical clustering to group these probesets into 40 clusters. We selected one probeset from each cluster and performed a K-medians clustering of the patients into a high-risk and low-risk group using these 40 probesets after normalizing their expression profiles so that the clusters wouldn't be unduly influenced by probesets with high mean expression values. On the Michigan data this gave a p-value of 0.00002 with cluster sizes of 36 and 50, while on the Harvard data the p-value was 0.0417 with cluster sizes of 47 and 37. Kaplan-Meier curves for these are given in Figures 5 and 6.

We used leave-one-out crossvalidation to verify this whole procedure. After clustering, the remaining patient was classified as high-risk or low-risk according to which cluster had the smaller average distance to that patient. For the Michigan data, this gave a p-value of 0.0219 and for the Harvard data the p-value was 0.0696.

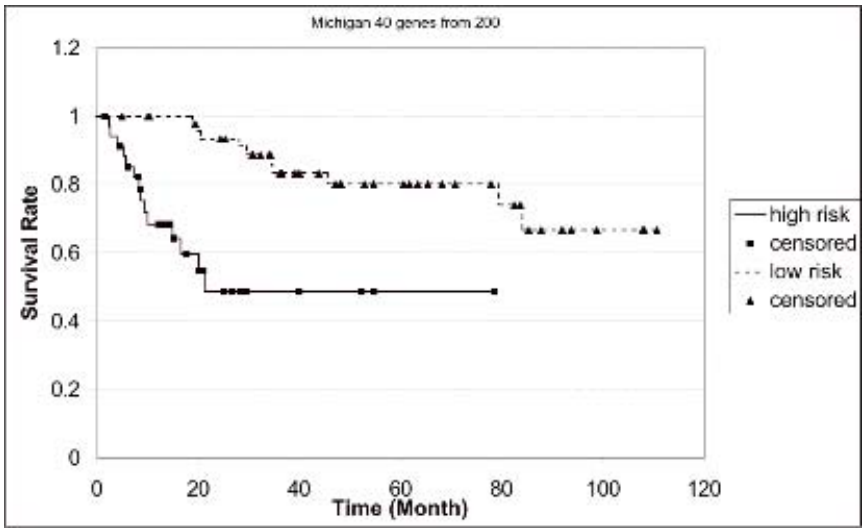

Figure 5. Kaplan-Meier curve for classifying Michigan data according to 40 probesets selected using MVR, K-medians, and hierarchical clustering of probesets.

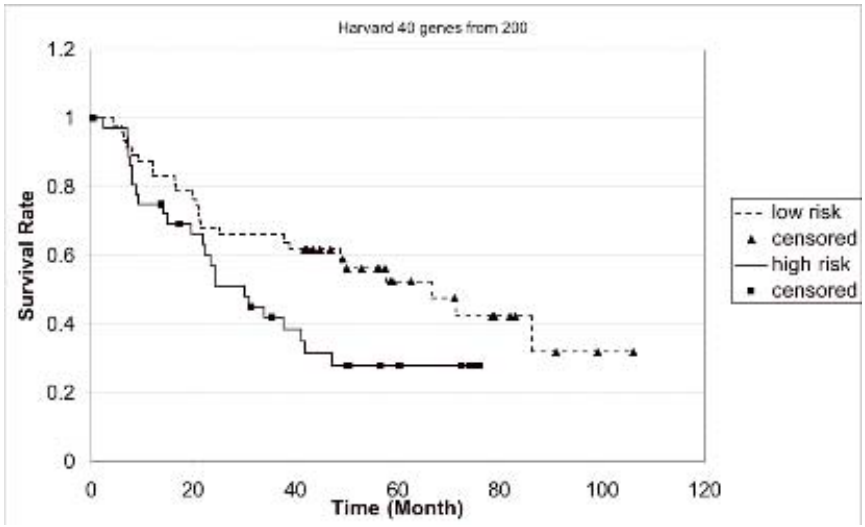

Figure 6. Kaplan-Meier curve for classifying Harvard data according to 40 probesets selected using MVR, K-medians, and hierarchical clustering of probesets.

4. CONCLUSIONS

On the Michigan data one-dimensional ADC clustering obtained results very comparable in terms of the p-values of the Kaplan-Meier curves to those obtained by Beer using Cox model regression, and we were able to reduce the set of genes further than they reported [Beer *et al*. 2002]. Beer reported a p-value of 0.0006 for leave-one-out crossvalidation based on a set of 50 genes, whereas in Table 1 we show p-values of 0.0009 for sets of 30 or 40 genes. On the Harvard data we obtained good results using 2-dimensional ADC, as reported in Table 3. We also obtained reasonable crossvalidation between the Harvard and Michigan data.

Our reduced sets of genes differed significantly from those reported by Beer. This is perhaps not surprising since our MVR and K-median experiments found that hierarchical clustering of the genes could often significantly reduce the number of genes without much of a decrease in the quality of the clustering as measured by the p-value. This probably indicates that the data contained many genes with closely related biological function. The following genes that have been associated to cancer appear on one or both of our top-50 lists, but were not among the top 50 reported by Beer:

- SPARCL1 (also known as MAST9 or hevin) - down regulation of SPARCL1 also occurs in prostate and colon carcinomas, suggesting that SPARCL1 inactivation is a common event not only in NSCLCs but also in other tumors of epithelial origin. (http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?cmd=Retrieve&db=PubMed&list_uids=11179481&dopt=Abstract)
- CD74 - well-known for expression in cancers (http://biz.yahoo.com/prnews/031120/nyth078_1.html)
- PRDX1 - linked to tumor prevention (http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?cmd=Retrieve&db=PubMed&list_uids=12891360&dopt=Abstract)
- PFN2 - seen as increasing in gastric cancer tissues (http://cancerres.aacrjournals.org/cgi/content/full/62/1/233)
- SFTPC - responsible for morphology of the lung; a mutation causes chronic lung disease (http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?cmd=Retrieve&db=PubMed&list_uids=14525980&dopt=Abstract)
- HLA-DRA (HLA-A) - lack of expression causes cancers (http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?cmd=Retrieve&db=PubMed&list_uids=12756506&dopt=Abstract)

Not much is known about the function of the following genes: PTGDS, H2AFZ, KIAA0005 (also called BZW1), EEF1A1, TXNRD1, RPS26. The

fact that appeared on our lists indicates that they may be worth further investigation.

Source code for our programs (in C++) and further results are available from http://camda.cs.tufts.edu

5. REFERENCES

Beer, D. G., *et al*., Gene-expression profiles predict survival of patients with lung adenocarcinoma, 2002, *Nature Medicine* **8**(8):816-824.

Bhattacharjee, A., *et al.*, 2001, Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses, *PNAS* **98**(24):13790-13795.

Cowen, L. J. and Priebe, C.E., 1997, Randomized non-linear projections uncover high-dimensional structure. *Adv. Appl. Math.*, **19**:319-331.

Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G., 2002, Diagnosis of multiple cancer types by shrunken centroids of gene expression., *PNAS* **99**(10):6567-657.