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, Anselm Blumer

Computer Science
Tufts University
Medford, MA 02155 USA
001-617-627-3217
{wzhou, wewu, npalmer, emower, ndaniels, cowen, ablumer}@cs.tufts.edu

ABSTRACT
We experiment with two types of clustering, K-medians and a dimension-reduction technique known as Approximate Distance Clustering, for classifying lung adenocarcinomas into high-risk and low-risk groups according to gene expression values from microarray data. We base this classification on a reduced set of genes obtained by Nearest Shrunken Mean [4] or a combination of a vraince-based approach with hierarchical clustering.

Categories and Subject Descriptors
J.3 [**Life and Medical Sciences**]: Biology and Genetics.

General Terms
Algorithms, Experimentation

Keywords
Microarray, ADC clustering, K-medians, adenocarcinoma, survival time

1 INTRODUCTION
This abstract investigates clustering and dimension-reduction techniques on two of the four CAMDA datasets of gene expression values and survival times of patients with lung adenocarcinomas. We chose the Michigan [1] and Harvard [2] data due to the reasonably large sample sizes (n = 86 and 84) and lack of missing values. We use ADC maps [3] to project the data into one or two dimensions so we can use very simple clustering techniques, then follow this with Nearest Shrunken Mean [4] 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
K-medians Clustering

This standard classical unsupervised clustering method selects K points to be cluster centers and calculates the quality of the clustering as the total distance of data points to their cluster centers. In this paper, we use K=2 so it is possible to find the optimal clustering in a reasonable time.

Approximate Distance Clustering (ADC) [3]

Approximate Distance Clustering is a method that reduces the dimensionality of data calculating the distances from data points to subsets of the data points called witness sets. It is defined as follows:

Let $\mathbf{X}$ be a collection of data in $\mathbf{R}^n$

Define $D_1, D_2, \ldots, D_m$ to be subsets of $\mathbf{X}$ of sizes $k_1, k_2, \ldots, k_m$

The associated ADC map, $f_{(D1,\ D2,\ \ldots,\ Dm)} : \mathbf{R}^n \rightarrow \mathbf{R}^d$ maps X to $(y_1, y_2, \ldots, y_m)$, where $y_i = \min\{ \| x_j - x \| : xj \text{ e } D_i\}$

A good witness set is a small set of points that produces a mapping that preserves inter-cluster distances. In this abstract, 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 the 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 this paper, we use the following criterion:

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 * c + 50 * d
```

where

a = 0 if the size of smaller group >= *n* /8,
= 1 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

Minimal Variance Ratio (MVR) Gene Reduction

Our goal is to find good clusters based on a reduced set of genes, selected using Minimal Variance Ratio or Nearest Shrunken Mean. Either of these methods may be followed by a hierarchical clustering of genes to eliminate those with similar expression patterns. The variance ratio is the sum of the within-cluster variances divided by the total variance of expression values for that gene. Genes with large variance ratios are thought to contribute less to the cluster definitions and are eliminated.

Nearest Shrunken Mean (NSM) Gene Reduction [4]

NSM eliminates genes with cluster mean close to the overall mean. Let:

x_{ij} be the expression of gene i of sample j

m_{ik} be the mean expression of gene i in class 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

$s_i = (1/(n-K)) \sum_k \sum_{j \in C_k} (x_{ij}-m_{ik})^2$

s_0 be the median of the s_i

$M_k = \mathrm{sqrt}(1/n_k+1/n)$

$d_{ik} = (m_{ik} - x_i) / (m_k*(s_i+s_0))$, 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.

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 averaged 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 [1].

A second set of experiments involved starting with high-risk and low-risk clusters of equal size according to survival times, 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 the parts of the Michigan [1] and Harvard [2] data that gave survival times (both censored and uncensored). The sample sizes were 84 for the Harvard data and 86 for the Michigan data.

ADC on Harvard and Michigan data

Tables 1 and 2 give the results of using the W criterion to select the best ADC witnesses and cutoffs, then reducing the set of genes with NSM. In both cases the witness sets had size one. The p-values were obtained from leave-one-out crossvalidation on the reduced set of genes. Specifically, ADC clusters were formed based on the reduced set of genes, 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. Kaplan-Meier curves for the 40-gene cases are given on the last page.

Table 1. Comparison of 1 and 2 dimensional ADC with STCC on Michigan data (n = 86)

Number of genes	p-value				Low-risk/high-risk group size			
	1D ADC	2D ADC	50% STCC	60% STCC	1D ADC	2D ADC	50% STCC	60% STCC
7129	0.0028	0.0500	0.0086	0.0126	*55/31*	*54/32*	*46/40*	*46/40*
1000	0.0275	0.0009	0.0111	0.0158	*59/27*	*60/26*	*45/41*	*43/43*
500	0.0495	0.0048	0.0046	0.0089	*52/34*	*57/29*	*47/39*	*45/41*
200	0.0019	0.0033	0.0075	0.0056	*58/28*	*58/28*	*47/39*	*48/38*
100	0.0058	0.0194	0.0023	0.0048	*57/29*	*55/31*	*49/37*	*46/40*
50	0.0019	0.1442	0.0064	0.0048	*58/28*	*42/44*	*50/36*	*47/39*
40	0.0009	0.0268	0.0011	0.0048	*58/28*	*44/42*	*50/36*	*47/39*
30	0.0009	0.0356	0.0029	0.0067	*58/28*	*43/43*	*51/35*	*46/40*
20	0.0021	0.0189	0.0029	0.0090	*57/29*	*42/44*	*51/35*	*46/40*
10	0.0061	0.0618	0.0059	0.0049	*56/30*	*37/49*	*50/36*	*47/39*
5	0.0086	0.3559	0.0151	0.0024	*58/28*	*41/45*	*49/37*	*49/47*

Table 2. Comparison of 1 and 2 dimensional ADC with STCC on Harvard data (n = 84)

Number of genes	p-value				Low-risk/high-risk group size			
	1D ADC	2D ADC	50% STCC	60% STCC	1D ADC	2D ADC	50% STCC	60% STCC
12600	0.0646	0.0046	0.1946	0.0741	*25/59*	*24/60*	*39/45*	*41/43*
1000	0.0124	0.0013	0.0381	0.0038	*20/64*	*15/69*	*44/40*	*38/46*
500	0.0023	0.0116	0.0021	0.0027	*21/63*	*22/26*	*42/42*	*36/48*
200	0.0121	0.0037	0.0007	0.0004	*21/63*	*21/63*	*40/44*	*32/52*
100	0.0201	0.0027	0.0213	0.0004	*24/60*	*26/58*	*42/42*	*30/54*
50	0.0332	0.0090	0.0120	0.0047	*21/63*	*21/63*	*40/44*	*35/49*
40	0.0332	0.0019	0.01	0.0033	*21/63*	*27/57*	*40/44*	*35/49*
30	0.0898	0.0010	0.0065	0.0098	*28/56*	*26/58*	*39/45*	*35/49*
20	0.0448	0.0039	0.0083	0.0015	*27/55*	*26/58*	*38/46*	*34/50*
10	0.0424	0.0011	0.0034	0.0001	*22/62*	*20/64*	*37/47*	*33/51*
5	0.0321	0.0032	0.0053	0.0196	*20/64*	*25/59*	*36/48*	*28/56*

Table 3 gives the top 40 genes found by one-dimensional ADC from the Michigan data in rank order.

Validating ADC between Harvard and Michigan data

We validated the 100 genes we obtained from Michigan's data by finding the genes in the Harvard data that matched these most closely and using those to run leave-one-out crossvalidation on the Harvard data. For the 1-dimensional ADC, we found 88 matching genes in the Harvard data and obtained a p-value of 0.0076 with cluster sizes of 25 and 59. For the 2-dimensional ADC, we found 83 genes and obtained a p-value of 0.4189 with cluster sizes of 24 and 60.

We then reversed this procedure with the 100 genes we obtained from the Harvard data. For the 1-dimensional ADC, we found 70 matching genes in the Michigan data and obtained a p-value of 0.0495 with cluster sizes of 64 and 22. For the 2-dimensional ADC, we found 65 genes and obtained a p-value of 0.3560 with cluster sizes of 44 and 40.

MVR and K-medians

We used Minimal Variance Ratio to select 200 genes from the Michigan and Harvard data based on an initial 50-50 clustering according to survival times, then used hierarchical clustering to group these genes into 40 clusters. We selected one gene from each cluster and performed a K-medians clustering of the patients into a high-risk and low-risk group using these 30 genes after normalizing their expression profiles so that the clusters wouldn't be unduly influenced by genes 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 on the last page.

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.

4 CONCLUSIONS

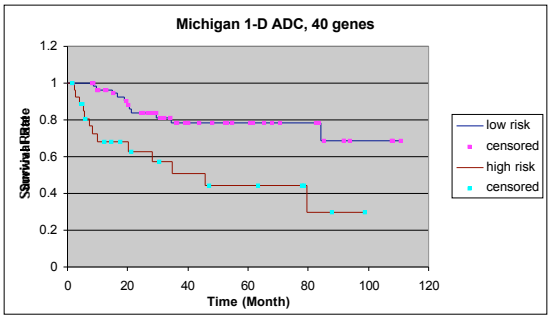
On the Michigan data ADC clustering obtained results very comparable in terms of the p-values of the Kaplan-Meier curves to those obtained by Beer *et al.* [1] using Cox model regression, and we were able to reduce the set of genes further than they reported. On the Harvard data we obtained results slightly worse in terms of the p-values than those reported by Bhattacharjee *et al.* [2], but we managed to significantly reduce the set of genes. We also obtained reasonable crossvalidation between the Harvard and Michigan data.

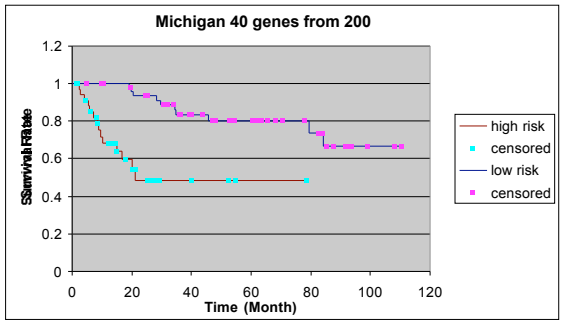
Our reduced sets of genes differed significantly from those reported by Beer *et al.* [1]. This is perhaps not surprising since our MVR and K-median experiments found that hierarchical clustering of the genes could often significantly reduce the number of genes without much of a decrease in the quality of the clustering as measured by the p-value. This probably indicates that the data contained many genes with closely related biological function.

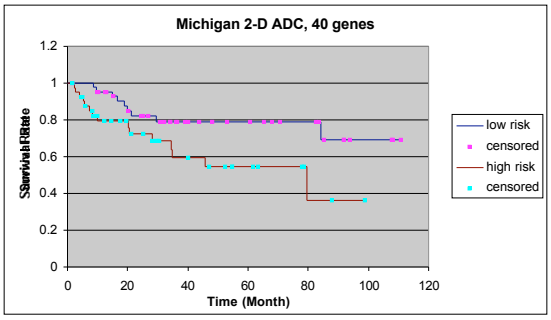
Source code for our programs (in C++) and further results are available from http://camda.cs.tufts.edu

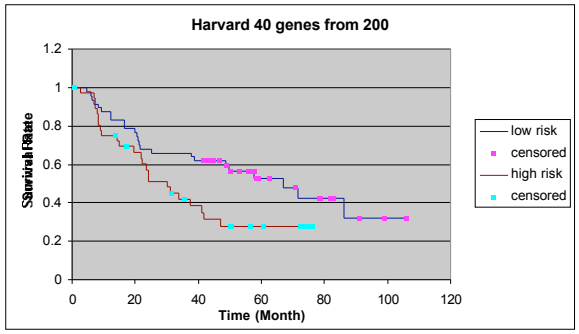
Table 3: Top 40 genes in Michigan data in rank order from 1-D ADC. (All probe set names end in _at)

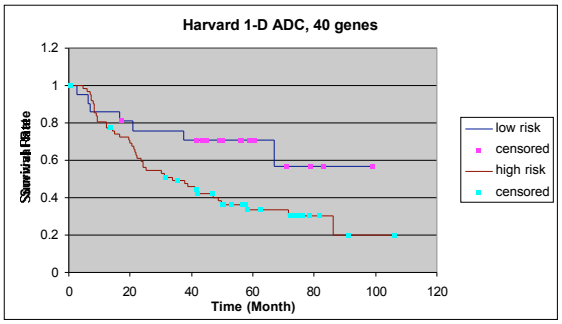
	Probe Set	Symbol	Name		Probe Set	Symbol	Name
1	M63438_s	IGKC	immunoglobulin kappa constant	21	X15940	RPL31	ribosomal protein L31
2	M34516	NULL		22	J03934_s	DIA4	diaphorase (NADH/NADPH (cytochrome b-5 reductase)
3	X57809_s	NULL		23	X91247	TXNRD1	thioredoxin reductase 1
4	M87789_s	IGHG3	immunoglobulin heavy constant gamma 3 (G3m marker)	24	X69654	RPS26	ribosomal protein S26
5	L19437	TALDO1	transaldolase 1	25	M22382	HSPD1	heat shock 60kD protein (chaperonin)
6	X01677_f	GAPD	glyceraldehyde-3-phosphate dehydrogenase	26	X77584	TXN	thioredoxin
7	L10678	PFN2	profilin 2	27	M26730_s	UQCRB	ubiquinol-cytochrome reductase binding protein
8	AFFX-HUMGAPDH/M33197_M	NULL		28	AFFX-HUMGAPDH/M33197_5	NULL	
9	M34516_r	NULL		29	D49824_s	HLA-B	major histocompatibili complex, class I, B
10	X67698	HE1	epididymal secretory protein (19.5kD)	30	X62744	HLA-DMA	major histocompatibili complex, class II, DM alpha
11	M21388_r	NULL		31	X15183	HSPCA	heat shock 90kD protein alpha
12	X00274	HLA-DRA	major histocompatibility complex, class II, DR alpha	32	U09813	ATP5G3	ATP synthase, H+ transportin mitochondrial F0 comple subunit c (subunit 9) isofor 3
13	M13560_s	CD74	CD74 antigen (invariant polypeptide of major histocompatibility complex, class II antigen-associated)	33	X56468	YWHAQ	tyrosine monooxygenase/tryptophan monooxygenase activatio protein, theta polypeptide
14	M17886	RPLP1	ribosomal protein, large, P1	34	X13238	COX6C	cytochrome c oxidase subun VIc
15	D49387	NULL		35	D14657	KIAA0101	KIAA0101 gene product
16	M37583	H2AFZ	H2A histone family, member Z	36	M22760	COX5A	cytochrome c oxidase subun Va
17	X67951	PRDX1	peroxiredoxin 1	37	D00762	PSMA3	proteasome (prosom macropain) subunit, alpl type, 3
18	X02152	LDHA	lactate dehydrogenase A	38	J04823_rna1	COX8	cytochrome c oxidase subun VIII
19	D13630	KIAA0005	KIAA0005 gene product	39	X53331	MGP	matrix Gla protein
20	D14874	ADM	adrenomedullin	40	M24485_s	GSTP1	glutathione S-transferase pi

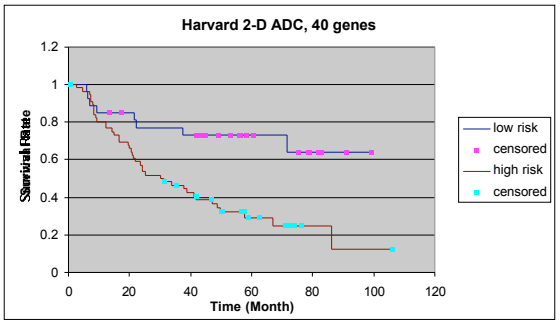












5 REFERENCES

[1] Beer, D. G., *et al*. Gene-expression profiles predict survival of patients with lung adenocarcinoma. Nature Medicine, v. 8, no. 8 (August 2002), 816-824.

[2] Bhattacharjee, A., *et al.* Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. Proceedings of the National Academy of Sciences, v. 98, no. 24 (November 2001), 13790-13795.

[3] Cowen, L. J. and Priebe, C.E. Randomized non-linear projections uncover high-dimensional structure. Advances in Applied Mathematics, v. 19 (1997), 319-331.

[4] Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proceedings of the National Academy of Sciences, v. 99, no. 10 (May 14, 2002), 6567-6572.