A New Profiling Tool for Gene Expression Data

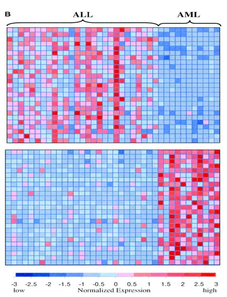
Satish Ch. Panigrahi, Md. Shafiul Alam, Asish Mukhopadhyay
School of Computer Science, University of Windsor, 401 Sunset Avenue, Windsor, ON, N9B 3P4, CANADA

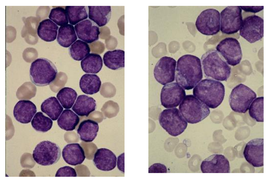
ABSTRACT

The availability of large volumes of gene expression data from microarray analysis (cDNA and oligonucleotide) has opened a new door to the diagnoses and treatments of various diseases based on gene expression profiling. In this poster, we discuss a new profiling tool based on linear programming. Given gene expression data from two subclasses of the same disease (e.g.leukemia), we are able to determine efficiently if the samples are linearly separable with respect to triplets of genes. This was left as an open problem in an earlier study that considered only pairs of genes as linear separators. Our tool comes in two versions - offline and incremental. Tests show that the incremental version is markedly more efficient than the offline one. This poster also introduces a gene selection strategy that exploits the class distinction property of a gene by separability test by pairs and triplets. We applied our gene selection strategy to 4 publicly available gene-expression data sets. Our experiments show that gene spaces generated by our method exhibit better classification accuracy than the gene spaces generated by t-values.

GOALS

1. To build a good classifier on gene expression dataset that can be applied to the clinical setting for proper identification and successful diagnosis.
2. This leads to identification of biomarker genes that differentiate between cancer cell from normal cell.

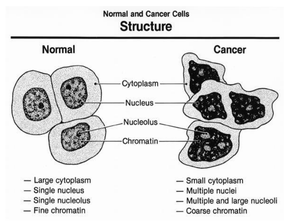




ALL

AML

A n × m matrix where
n = total number of genes (12582)
m = total samples (24 + 28)

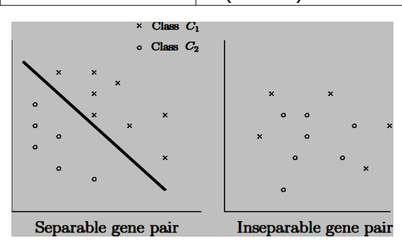


DATASET IN USE

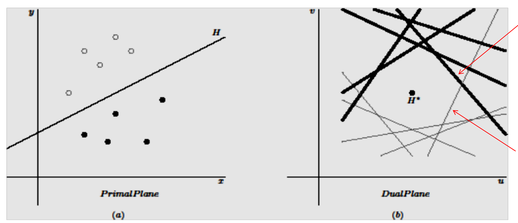
	Dataset	No of Gene	Total Sample
01	Lung Cancer	12533	181(31+150)
02	Leukemia Data	12582	52(24+28)
03	SRBCT	2308	43(23+20)
04	Colon Data	2000	62(40+22)
05	Breast Cancer	21682	77(44+33)

GEOMETRIC PERSPECTIVE

- Mapping of points in 2D Plane
- Each point represents one observation
- We have C(n, 2) gene pairs for n genes



METHOD



A separating line H in primal plane is a feasible solution H* in dual plane

Feasible solution of following linear program

minimize/maximize u_2

$u_2 > p^i_1 u_1 - p^i_2$, $i = 1 \ldots m_1$

$u_2 < p^i_1 u_1 - p^i_2$, $i = 1 \ldots m_2$

Offline Approach

- $\log_2 m$ iterations where m is the sample size.
- Prunes away a quarter of the constraints in each iteration.
- Linear in sample size

Incremental Approach

- Inseparability of gene pairs can be detected early without considering all samples.
- Thus leads to better running time.

2D Separability Test with Run time

	Dataset	Percentage of LSP	Runtime of Offline in msec	Runtime of Incremental in msec	Improvement of incremental over offline
1	Lung Cancer	0.72%	4617	1537	66.71%
2	Leukemia Data	11.915%	1138	418	63.27%
3	SRBCT	8.86%	987	356	63.93%
4	Colon Data	0%	1328	275	79.29%
5	Breast Cancer	0.925%	1606	440	72.6%

3D Separability Test with Run time

	Dataset	Percentage of PLST	Runtime of Offline in msec	Runtime of Incremental in msec	Improvement of Incremental over offline
1	Lung Cancer	0.946%	1114467	166662	85.04%
2	Leukemia Data	3.72%	115791	49263	57.45%
3	SRBCT	4.11%	92825	52080	43.89%
4	Colon Data	0%	170143	39955	76.516%
5	Breast Cancer	0.137%	274141	83913	69.39%

Gene Selection

Two steps:

a. Coarse Filtration: Choose 100 gene based upon t value defined by Nguyen and Rocke.

b. Fine Filtration: $\Delta = \{g_1, g_2, \ldots, g_n\}$, For a gene $g_i \in \Delta$

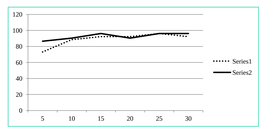
$S_i = \{g_j \mid (g_i, g_j)$ is an LS pair, $g_j \in \Delta, i \neq j\}$, $\mathbf{P_i = | S_i |}$

$Q_i = \{(g_j, g_k) \mid (g_i, g_j, g_k)$ is an LS triplet, $g_j, g_k \in \Delta, i \neq j \neq k\}$, $\mathbf{T_i = | Q_i |}$

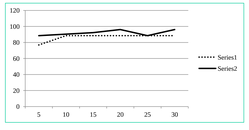
- Generate feature space of size 5, 10, 15, 20, 25, 30

RESULTS

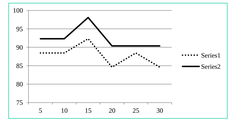
Feature Space Vs Accuracy



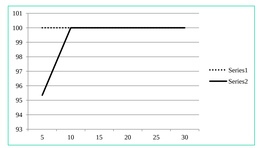
Leukemia SVM



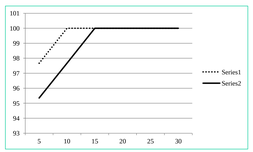
Leukemia kNN, k =21



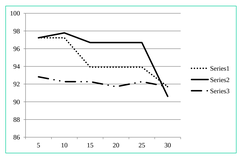
Leukemia BayesNet



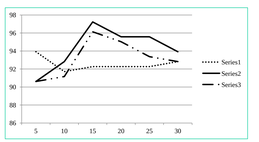
SRBCT SVM & kNN, k= 18



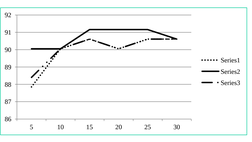
SRBCT BayesNet



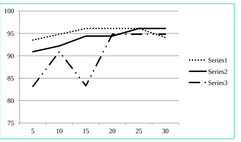
Lung Cancer SVM



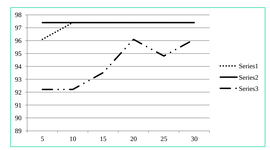
Lung Cancer kNN, k = 27



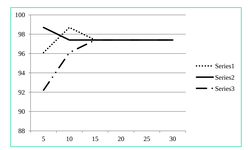
Lung Cancer BayesNet



Breast Cancer SVM

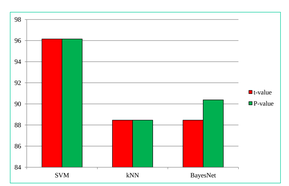


Breast Cancer kNN

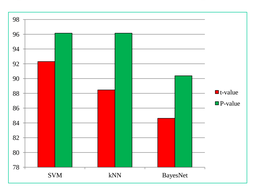


Breast Cancer BayesNet

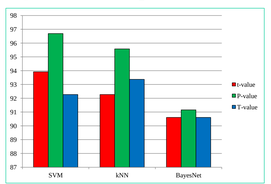
Classification Accuracy of Feature Space (25 & 30)



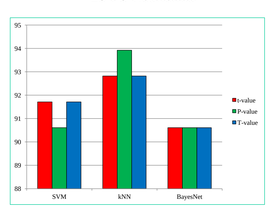
25 FS Leukemia



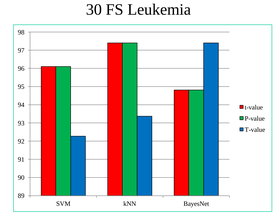
30 FS Leukemia



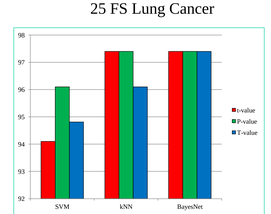
25 FS Lung Cancer



30 FS Lung Cancer



25 FS Breast Cancer



30 FS Breast Cancer