Gene Selection for Cancer Classification using Support Vector Machines

Aixia Guo

Department of Mathematics, UH

May 08, 2014

Machine Learning, 46, 389–422, 2002

Gene Selection for Cancer Classification using Support Vector Machines

ISABELLE GUYON isabelle@barnhilltechnologies.com
JASON WESTON
STEPHEN BARNHILL
Barnhill Bioinformatics, Savannah, Georgia, USA

VLADIMIR VAPNIK vlad@research.att.com
AT&T Labs, Red Bank, New Jersey, USA

Editor: Nello Cristianini

Content

- Motivation
- Methodology
- Problem Description and Prior Work
- Support Vector Machines (SVMs) and Recursive Feature Elimination (RFE)
- Experimental Results
- Conclusion

Motivation

- Cancer vs Normal tissues
- Micro-array technology measures expression level of 10,000~30,000 genes simultaneously in a single experiment
- Micro-array devices generate bewildering amounts of raw data
- Methods are needed to sort out whether cancer tissues have distinctive signatures of gene expression over normal tissues or other types of cancer tissues

Motivation

- Cancer vs Normal tissues
- Micro-array technology measures expression level of 10,000~30,000 genes simultaneously in a single experiment
- Micro-array devices generate bewildering amounts of raw data
- Methods are needed to sort out whether cancer tissues have distinctive signatures of gene expression over normal tissues or other types of cancer tissues

Methodology

Address the problem by a new method of gene selection utilizing Support Vector Machine(SVM) methods based on Recursive Feature Elimination (RFE)

(1)To select a small subset of genes from broad gene expression data

(2) To build a classifier by using available training examples from cancer and normal patients

Terminology

Gene = feature = attribute = column

Pattern: a vector of *n* components (features)

Patient	Gene 1	Gene 2	…	Gene n	Class
1	100.30	200.52	…	1000.11	+
2	20.56	500.31	…	600.75	-
…	…	…	…	…	…
m	150.24	1000.20	…	300.33	-

Example of gene expressions

Problem Formulation of Classification problems

identify the two classes with the symbols (+) and (−). A training set of a number of patterns $\{\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_k, \ldots \mathbf{x}_\ell\}$ with known class labels $\{y_1, y_2, \ldots y_k, \ldots y_\ell\}$, $y_k \in \{-1, +1\}$, is given. The training patterns are used to build a decision function (or discriminant function) $D(\mathbf{x})$, that is a scalar function of an input pattern $\mathbf{x}$. New patterns are classified according to the sign of the decision function:

$$D(\mathbf{x}) > 0 \Rightarrow \mathbf{x} \in \text{class } (+)$$
$$D(\mathbf{x}) < 0 \Rightarrow \mathbf{x} \in \text{class } (-)$$
$$D(\mathbf{x}) = 0, \text{ decision boundary.}$$

Decision functions that are simple weighted sums of the training patterns plus a bias are called linear discriminant functions (see e.g. Duda, 1973). In our notations:

$$D(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + \mathbf{b}, \tag{1}$$

where $\mathbf{w}$ is the weight vector and b is a bias value.

Prior Works of Space Dimensionality Reduction

- A common method to reduce feature space dimension

  Project on the first few principle directions of the data(see, e.g. Duda, 73)

  New features obtained are linear combinations of the original features

- Disadvantages

  None of the original input features can be discarded

- New pruning techniques are needed

  Eliminate some of the original input features and retain a minimum subset of features that yield best classification performance

Feature-ranking Technique

- Feature ranking with correlation coefficients

  Select the genes that individually classify best the training data

  Eliminate genes that are useless for discrimination

- Evaluating how well an individual feature contributes to the separation (e.g. cancer vs normal) can produce a simple feature (gene) ranking.

Various correlation coefficients are used as ranking criteria. The coefficient used in Golub (1999) is defined as:

$$w_i = (\mu_i(+) - \mu_i(-))/(\sigma_i(+) + \sigma_i(-))$$

where μ_i and σ_i are the mean and standard deviation of the gene expression values of gene i for all the patients of class (+) or class (−), i = 1, . . . n. Large positive wi values indicate strong correlation with class (+) whereas large negative wi values indicate strong correlation with class (−).

Disadvantages

Cannot yield compact gene sets because genes are redundant

Complementary genes that individually do not separate well the data are missed

Recursive Feature Elimination

1) Train the classifier
2) Compute the ranking criterion
   for all feature
3) Remove the fe
   with *smalles*
4) Repeat

For computational reasons,it may be more efficient to remove several features at a time, at the expense of possible classification performance degradation.

RFE has no effect on correlation methods since the ranking criterion is computed with information about a single feature.

Feature Ranking with Support Vector machines (SVM)

- Idea from using the weights of a classifier to produce a feature ranking
- In this paper, the classifier used is linear SVMs (Boser, 1992; Vapnik, 1998)
- Presently SVM is one of the best-known classification techniques with computational advantages over their contenders (Cristianini, 1999).

Feature Ranking with SVMs

Algorithm SVM-train:

Inputs: Training examples $\{\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_k, \ldots \mathbf{x}_\ell\}$ and class labels $\{y_1, y_2, \ldots y_k, \ldots y_\ell\}$.

$$\begin{cases} \text{Minimize over } \alpha_k: \\ J = (1/2) \sum_{hk} y_h\, y_k\, \alpha_h\, \alpha_k\, (\mathbf{x}_h \cdot \mathbf{x}_k + \lambda \delta_{hk}) - \sum_k \alpha_k \\ \text{subject to:} \\ 0 \le \alpha_k \le C \quad \text{and} \quad \sum_k \alpha_k\, y_k = 0 \end{cases} \tag{5}$$

Outputs: Parameters α_k.

The resulting decision function of an input vector $\mathbf{x}$ is:

$$D(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b$$

with

$$\mathbf{w} = \sum_k \alpha_k\, y_k\, \mathbf{x}_k \quad \text{and} \quad b = \langle y_k - \mathbf{w} \cdot \mathbf{x}_k \rangle$$

The weight vector $\mathbf{w}$ is a linear combination of training patterns. Most weights α_k are zero. The training patterns with non-zero weights are support vectors. Those with weight satisfying the strict inequality $0 < \alpha_k < C$ are marginal support vectors. The bias value b is an average over marginal support vectors.

(1) $\delta_{hk} = 1$ if h = k and 0 otherwise
(2) λ and C are soft margin parameters. The soft margin parameters ensure convergence even when the problem is non-linearly separable or poorly conditioned
(3) use a small value of λ (of the order of 10^{-14}) to ensure numerical stability

SVM Recursive Feature Elimination (SVM RFE)

SVM RFE is an application of RFE using the weight magnitude as ranking criterion. Below is an outline of the algorithm in the linear case, using *SVM-train* in Eq. (5).

Algorithm SVM-RFE:

Inputs:

Training examples

$$\mathbf{X}_0 = [\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_k, \ldots \mathbf{x}_\ell]^T$$

Class labels

$$\mathbf{y} = [y_1, y_2, \ldots y_k, \ldots y_\ell]^T$$

Initialize:

Subset of surviving features

$$\mathbf{s} = [1, 2, \ldots n]$$

Feature ranked list

$$r = [\,]$$

Repeat until **s** = [ ]

Restrict training examples to good feature indices

$$X = X_0(:, \mathbf{s})$$

Train the classifier

$$\alpha = \textit{SVM-train}(X, \mathbf{y})$$

Compute the weight vector of dimension length(**s**)

$$\mathbf{w} = \sum_k \alpha_k \, y_k \, \mathbf{x}_k$$

Compute the ranking criteria

$$c_i = (w_i)^2, \quad \text{for all } i$$

Find the feature with smallest ranking criterion

$$f = \textit{argmin}(\mathbf{c})$$

Update feature ranked list

$$\mathbf{r} = [\mathbf{s}(f), \mathbf{r}]$$

Eliminate the feature with smallest ranking criterion

$$\mathbf{s} = \mathbf{s}(1{:}f-1, f+1{:}\text{length}(\mathbf{s}))$$

Output:

Feature ranked list **r**.

Experimental Results

1) A small subset of selected features have the best classification results
2) The features selected matter more than the classifier used
3) SVM-RFE selects relevant genes

Gene Expression Dataset and the Classification Problem

Leukemia data is available on-line. The problem is to distinguish between two variants of leukemia (ALL and AML).

The data is split into two subsets: A training set, used to select genes and adjust the weights of the classifiers, and an independent test set used to estimate the performance of the system obtained.

Their training set consists of 38 samples (27 ALL and 11 AML) from bone marrow specimens. Their test set has 34 samples (20 ALL and 14 AML), prepared under different experimental conditions and including 24 bone marrow and 10 blood sample specimens. All samples have 7129 features, corresponding to some normalized gene expression value extracted from the micro-array image.

A small subset of selected features have the best classification results

# of genes	Train accuracy	Test accuracy
All (7129)	0.95	0.85
4096	0.82	0.71
2048	0.97	0.85
1024	1.00	0.94
512	0.97	0.88
256	1.00	0.94
128	1.00	0.97
64	1.00	0.94
32	1.00	0.97
16	**1.00**	**1.00**
8	**1.00**	**1.00**
4	0.97	0.91
2	0.97	0.88
1	0.92	0.79

Features selected matter more than the classifier used

Table 1. SVM classifier trained on SVM genes obtained with the RFE method (Leukemia data).

	Training set (38 samples)				Test set (34 samples)			
Number of genes	V_{suc}	V_{acc}	V_{ext}	V_{med}	T_{suc}	T_{acc}	T_{ext}	T_{med}
All (7129)	0.95	0.87	0.01	0.42	0.85	0.68	−0.05	0.42
4096	0.82	0.05	−0.67	0.30	0.71	0.09	−0.77	0.34
2048	0.97	0.97	0.00	0.51	0.85	0.53	−0.21	0.41
1024	1.00	1.00	0.41	0.66	0.94	0.94	−0.02	0.47
512	0.97	0.97	0.20	0.79	0.88	0.79	0.01	0.51
256	1.00	1.00	0.59	0.79	0.94	0.91	0.07	0.62
128	1.00	1.00	0.56	0.80	0.97	0.88	−0.03	0.46
64	1.00	1.00	0.45	0.76	0.94	0.94	0.11	0.51
32	1.00	1.00	0.45	0.65	0.97	0.94	0.00	0.39
16	**1.00**	**1.00**	**0.25**	**0.66**	**1.00**	**1.00**	**0.03**	**0.38**
8	**1.00**	**1.00**	**0.21**	**0.66**	**1.00**	**1.00**	**0.05**	**0.49**
4	0.97	0.97	0.01	0.49	0.91	0.82	−0.08	
2	0.97	0.95	−0.02	0.42	0.88	0.47	−0.23	
1	0.92	0.84	−0.19	0.45	0.79	0.18	−0.27	

Fewer genes selected by SVM-RFE have better classification results comparing to the genes selected by correlation

Table 2. SVM classifier trained on baseline genes (Leukemia data).

	Training set (38 samples)				Test set (34 samples)			
Number of genes	V_{suc}	V_{acc}	V_{ext}	V_{med}	T_{suc}	T_{acc}	T_{ext}	T_{med}
All (7129)	0.95	0.87	0.01	0.42	0.85	0.68	−0.05	0.42
4096	0.92	0.18	−0.43	0.29	0.74	0.18	−0.68	0.36
2048	0.95	0.95	−0.09	0.32	0.85	0.38	−0.25	0.33
1024	1.00	1.00	0.09	0.34	0.94	0.62	−0.13	0.34
512	1.00	1.00	0.08	0.39	0.94	0.76	−0.06	0.37
256	1.00	1.00	0.08	0.40	0.91	0.79	−0.04	0.42
128	1.00	1.00	0.09	0.39	0.94	0.82	−0.04	0.49
64	**0.97**	**0.97**	**0.01**	**0.44**	**0.97**	**0.82**	**−0.09**	**0.44**
32	1.00	1.00	0.07	0.46	0.91	0.88	−0.07	0.42
16	1.00	1.00	0.16	0.52	0.94	0.91	−0.07	0.39
8	1.00	1.00	0.17	0.52	0.91	0.85	−0.10	0.51
4	1.00	1.00	0.21	0.48	0.88	0.68	−0.03	0.28
2	0.97	0.97	0.00	0.36	0.79	0.47	−0.22	0.27
1	0.92	0.84	−0.19	0.45	0.79	0.18	−0.27	0.23

The success rate (at zero rejection, the acceptance rate (at zero error), the extremal margin and the median margin are reported for the leave-one-out method on the 38 sample training set (V results) and the 34 sample test set (T results). We outline in boldface the classifiers performing best on test data reported in Table 5. For comparison, we also show the results on all genes (no selection).

Features selected matter more than the classifier used

Table 3. Baseline classifier trained on SVM genes obtained with the RFE method (Leukemia data).

	Training set (38 samples)				Test set (34 samples)			
Number of genes	V_{suc}	V_{acc}	V_{ext}	V_{med}	T_{suc}	T_{acc}	T_{ext}	T_{med}
All (7129)	0.89	0.47	−0.25	0.28	0.85	0.35	−0.24	0.34
4096	0.97	0.97	0.01	0.41	0.88	0.59	−0.12	0.40
2048	1.00	1.00	0.29	0.56	0.88	0.76	−0.07	0.45
1024	1.00	1.00	0.44	0.67	0.94	0.82	0.01	0.47
512	1.00	1 00	0.39	0.81	0.91	0.88	0.07	0.55
256	1.00	1.00	0.55	0.76	0.94	0.94	0.09	0.62
128	1.00	1.00	0.56	0.81	0.94	0.82	0.02	0.45
64	**1.00**	**1.00**	**0.47**	**0.74**	**1.00**	**1.00**	**0.14**	**0.49**
32	1.00	1.00	0.44	0.66	0.94	0.79	0.01	0.40
16	1.00	1.00	0.27	0.63	(			
8	1.00	1.00	0.25	0.62	(			
4	0.95	0.89	0.04	0.45	(			
2	0.97	0.95	0.03	0.39	(			
1	0.92	0.76	−0.17	0.43	(			

Baseline classifier (not SVM) has better classification results with the SVM-RFE features comparing to using the baseline genes (correlation selected genes)

Table 4. Baseline classifier trained on baseline genes (Leukemia data).

	Training set (38 samples)				Test set (34 samples)			
Number of genes	V_{suc}	V_{acc}	V_{ext}	V_{med}	T_{suc}	T_{acc}	T_{ext}	T_{med}
All (7129)	0.89	0.47	−0.25	0.28	0.85	0.35	−0.24	0.34
4096	0.95	0.76	−0.12	0.33	0.85	0.44	−0.20	0.37
2048	0.97	0.97	0.02	0.36	0.85	0.53	−0.13	0.37
1024	1.00	1.00	0.11	0.36	0.94	0.65	−0.11	0.37
512	1.00	1 00	0.11	0.39	0.94	0.79	−0.05	0.40
256	1.00	1.00	0.11	0.40	0.91	0.76	−0.02	0.43
128	1.00	1.00	0.12	0.39	0.94	0.82	−0.02	0.50
64	**1.00**	**1.00**	**0.07**	**0.43**	**0.97**	**0.82**	**−0.08**	**0.45**
32	1.00	1.00	0.11	0.44	0.94	0.85	−0.07	0.42
16	1.00	1.00	0.18	0.50	0.94	0.85	−0.07	0.40
8	1.00	1.00	0.15	0.50	0.91	0.82	−0.10	0.51
4	1.00	1.00	0.18	0.45	0.88	0.62	−0.03	0.28
2	0.95	0.92	0.02	0.33	0.82	0.59	−0.22	0.27
1	0.92	0.76	−0.17	0.43	0.79	0.18	−0.27	0.23

SVM-RFE selects relevant genes

Table 9. SVM RFE top ranked genes (Leukemia data).

Rk	Expression	GAN	Description	Possible function/relation to Leukemia
4	AML > ALL	U59632	Cell division control related protein (hCDCrel-1) mRNA	hCDCrel-1 is a partner gene of MLL in some leukemias (Osaka, 1999).
3	AML > ALL	U82759	GB DEF = Homeodomain protein HoxA9 mRNA	Hoxa9 collaborates with other genes to produce highly aggressive acute leukemic disease (Thorsteinsdottir, 1999).
2	ALL > AML	HG1612	MacMarcks	Tumor necrosis factor-alpha rapidly stimulate Marcks gene transcription in human promyelocytic leukemia cells (Harlan, 1991).
1	AML > ALL	X95735	Zyxin	Encodes a LIM domain protein localized at focal contacts in adherent erythroleukemia cells (Macalma, 1996).

The entire data set of 72 samples was used to select genes with SVM RFE. Genes are ranked in order of increasing importance. The first ranked gene is the last gene left after all other genes have been eliminated. Expression: ALL > AML indicates that the gene expression level is higher in most ALL samples; AML > ALL indicates that the gene expression level is higher in most AML samples; GAN: Gene Accession Number. All the genes in this list have some plausible relevance to the AML vs. ALL separation.

Conclusion

- The genes selected by SVM-RFE yield better classification performance (rather than the classifiers)
- The selected genes are closely related to the diseases
- In contrast with the baseline method, their method eliminates gene redundancy automatically and yields better and more compact gene subsets

References

[1] Boser, B., Guyon, I., & Vapnik, V. (1992). An training algorithm for optimal margin classifiers. In Proceedings of the Fifth Annual Workshop on Computational Learning Theory (pp. 144–152). Pittsburgh: ACM.

[2] Cristianini, N. & Shawe-Taylor, J. (1999).An introduction to support vector machines. Cambridge, MA: Cambridge University Press.

[3] Duda, R. O. & Hart, P. E. (1973). Pattern classification and scene analysis. New York: Wiley.

[4]Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L.,Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., & Lander, E. S. (1999). Molecular classification of cancer:Class discovery and class prediction by gene expression monitoring. Science, 286, 531–537. The data is available on-line at http://www.genome.wi.mit. edu/MPR/data set ALL AML.html.

[5] Vapnik, V. N. (1998). Statistical learning theory. Wiley Interscience.

Thank you