International Conference on Science &
Technology: Application in Industry
& Education (2008)

# VARIABLE SELECTION APPROACH BASED ON RFE-SVM FOR CANCER CLASSIFICATION

[1]Khin May Win, [2]Nan Sai Moon Kham
[1]Department of Information Sciences, University of Computer Studies,Yangon,Myanmar
e-mail: [1]winn.km05@gmail.com, [2]moonkhamucsy@gmail.com

**Abstract.** *The data analysis can be many different points of views from many researchers. New method is evaluated for variable subset relevance with a view to variable selection. The new criteria are derived from support vector approach based on classification problems. This search can be efficiently performed by minimizing the generalization error. Selecting a small subset of features variables not only improves the efficiency of the classification algorithms but also improve the cancer classification accuracy. The process of building classifier is divided into two components (i ) selection of variables features (i .e genes) (ii ) selection of classification method. Our study indicates that the classification problem is more difficult than the binary one for the gene expression data sets. This new method is related to structural risk minimization and thus leads to good generalization. The proposed method is compared to some standard feature selection method with real data sets. This method is computationally efficient with better classification performance.*

***Keywords* —**support vector machines, linear kernels, variable selection, feature ranking, over fitting.

## 1. Introduction

Genetic information of cells is stored in DNA and all cells in an organism which have different gene expression patterns. The development of DNA microarray technology has been produced large amount of gene data. This technology has been applied to the field of accurate prediction and diagnosis of cancer disease. Especially accurate classification of cancer is very important issue for treatment of cancer. To precisely classify cancer we have to select genes related to cancer .We attempt to choose the cancer related genes by using feature selection and combined the classifiers to improve the performance of classification.

The gene expression data usually consist of huge number of genes and the necessity of tools analyzing them to get useful information. The selection of relevant variables may also be useful to gain some insight about the concept to be learned. Other advantages of feature selection include cost reduction of data gathering and storage (in medical applications) and computational speedup. In this work, we investigate the efficiency of criteria derived from support vector machines (SVMs) for variable selection in application to classification problems. This can be seen as an extension of the SVM-RFE algorithm. In genomics expression, the data set is usually plagued with large number of variables versus the small number of records or vectors (the problem is known as the ‘curse of dimensionality’). Genes are clustered first, and usual methods used are K-means clustering and hierarchical clustering [12], Singular Value Decomposition or Principal Component Analysis, supervised clustering and fuzzy clustering methods [4,6].In the dual space the decision function is expressed as a linear combination of basis functions parameterized by the supporting patterns. The supporting patterns correspond to the class are chosen automatically by the maximum margin training procedures [10]. In case of polynomial classifiers, the Perceptron representation involves an untraceable number of parameters [5]. This problem is overcome in the dual space representation, where the classification rule is weighted sum of a kernel function for each support vectors patterns.

The incremental informative content of more variables is not always significant. Among existing methods, S2N performs good combination of gene selection methods and classifiers for microarray data .However, a significant improvement is achieved from choosing the appropriate parameter to small value. It supports the performance of the SVM classifier [15]. The selection of relevant variables may be insight to enhance the generalization performance of the learning process. High order polynomial classifiers with very large training sets can therefore be handled efficiently with the feature selection method. A new feature selection method criterion function was proposed based on the Feature weighting value. For hope, the learning process may avoid redundant, noisy or unreliable information of features**.**

## 2. Related Work

From the data mining viewpoint, gene selection problem is essentially a feature selection or dimensionality reduction problem. After reviewing, the soft margin SVM classifier can perform ranking criteria derived from SVM and an associated algorithm for feature selection. Finally, relationships with other SVM-based feature selection methods are given.

### 2.1 Support Vector Machines

One strategy is first to train linear Support Vector Machines (SVM) on a subset of training data to create initial classifiers. Each vector $X = \{x_i\}_{i=1}^{m}$ labeled by $\{y_i\}_{i=1}^{m}$ in the gene expression matrix may be thought of as a point in an m-dimensional expression space. In theory, a simple way to build a binary classifier is to construct a hyperplane which can separate class members. The decision function becomes:

$$f(x) = (w, \Phi(x)) + b \tag{1}$$

Unfortunately, most real-world problems involve non separable data for which there does not exist a hyperplane that successfully separates the positive from the negative examples. One solution to the inseparability problem is to map the data into a higher-dimensional space and define a separating hyperplane there. This higher dimensional space is called the feature space, as opposed to the input space occupied by the training examples.

For linearly non-separable cases, one can introduce slack variables $\xi$ and accordingly, the discriminated function is defined by

$$y_i(w.x_i + b) \geq 1 - \xi \geq 0 \tag{2}$$

measure the deviation of a data point from optimal hyper plane. SVM are designed by minimizing

$$\Phi(w, \xi) = \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{n} \xi_i$$

$$subject\ to: y_i(w.x_i + b) \geq 1 - \xi_i, \xi_i \geq 0 \tag{3}$$

Minimize over $: (w, b, \xi_i, \ldots, \xi_m): \quad \| w \|_p^p + C \sum_{i=1}^{m} \xi_i$

$$\text{subject to}: \forall_{i=1}^{m}: y_i(< w.\Phi(x_i) > +b) \geq 1 - \xi_i, \xi > 0 \tag{4}$$

Where <.> is the inner product of mapping function between two vectors .C is a user –specified constant for controlling the penalty to the violation terms denoted by each (slack variables).C + and C- control the penalty to the violation of positive and negative examples of each features respectively. The w and b constitute of the classifier,

$$y = \text{sign}(< w.\Phi(x) > +b) \tag{5}$$

Furthermore, artificially separating the data in this way exposes the learning system to the risk of finding trivial solutions that over fit the data. SVMs elegantly sidestep both difficulties (4). They avoid over fitting by choosing the maximum margin separating hyperplane from among the many that can separate the positive from negative examples in the feature space.

### 2.2. Correlation based Feature Ranking Algorithms for Gene Selection

Gene selection can be viewed as a feature selection or dimensionality reduction problem. Currently, there are mainly two kinds of algorithms for gene selection: correlation-based algorithms and backword elimination algorithms. Correlation-based feature ranking algorithms work in a forward selection. Then, some top ranked genes are selected to form the most informative gene subset [12], [13].
Some commonly used ranking matrices are:

Signal-to-Noise (S2N):

$$w_i = | \mu_i(+) - \mu_i(-) | / \sigma_i(+) + \sigma_i(-) \tag{6}$$

Fisher Criterion (FC):

$$w_i = (\mu_i(+) - \mu_i(-))^2 | / \sigma_i(+)^2 + \sigma_i(-)^2 \tag{7}$$

T-Statistics (TS):

$$w_i = |\mu_i(+) - \mu_i(-)| / \sqrt{\sigma_i(+)^2 / n(+) + \sigma_i(-)^2 / n(-)} \qquad (8)$$

At above equations, $\mu_i(+)$ and $\mu_i(-)$ are the mean values of the I th gene expression data over positive and negative samples in the training data set, respectively. $\sigma_i(+)$ and $\sigma_i(-)$ are the corresponding standard deviations. $n(+)$ and $n(-)$ denote the numbers of positive negative training samples, respectively. A larger $w_i$ means that the ith gene is more informative for cancer classification.

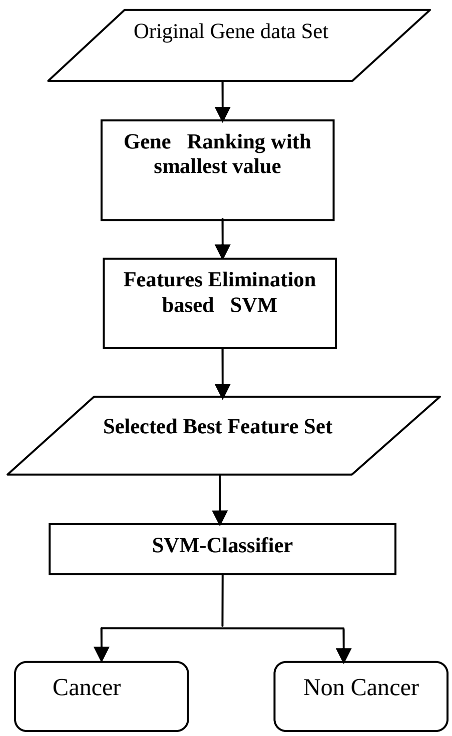


**Fig .1.** System Flow for features elimination based SVMs

## 3. Recursive Feature Elimination

The RFE approach operates in an iterative manner to eliminate features weighted by weak weights specified in a 2-norm SVM model. Nested subsets of features are selected in a sequential backward elimination manner, which starts with all the features and remove one feature each time. In this way, in the end, all the feature variables are ranked. At each step, the coefficients of the weight vector $w$ of a linear SVM are used as the feature ranking criterion. The recursive elimination procedure used as follows:

**(1)** Start: ranked feature $R$ = [ ];
selected subset $S = [1,\ldots, d]$;
**(2)** Repeat until all features are ranked:
(a) Train a linear SVM with all the training data and variables in $S$;
(b) Compute the weight vector
(c) Compute the ranking scores in $S$: $c_i = (w_i)^2$;
(d) Find the feature with the smallest ranking score: $e = \arg \min i \; c_i$;
(e) Update $R$: $R = R[e,R]$;
(f) Update $S$: $S = S - [e]$;
**(3)** Output: Ranked feature list $R$

The algorithm can be generalized to remove more than one feature per step.

In SVM-FRE, the following SVM formulation is used

$$\min_{w,b,\xi_i} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}\xi_i^2 \tag{9}$$

$$\text{subject to} \quad y_i(w.z_i + b) \geq 1 - \xi_i, \xi_i \geq 0 \tag{10}$$

The formulation of SVM is usually solve by the following dual problem with mapping function $< \phi(x_i).\phi(x_j) >$, $z_i = \phi(x_i)$ .i.e. kernel function.

Maximize over $(\alpha_1,....,\alpha_m)$:

$$J = -\frac{1}{2}\sum_{i,j=1}^{m} \alpha_i\alpha_j y_i y_j < \phi(x_i).\phi(x_j) > + \sum_{i=1}^{m} \alpha_i \tag{11}$$

$$\text{subject to}: \sum_{i=1}^{m} \alpha_i y_i = 0 \text{ and } \forall_{i=1}^{m}: 0 \leq \alpha_i \leq C \tag{12}$$

The weighting vector w is given by $\sum_{i=1}^{m} \alpha_i y_i \phi(x_i)$. Calculating w might be prohibitively expensive when nonlinear kernels are used. Using $w_i^2$ as ranking score corresponds to removing the feature whose removal change the objective function test. The approximation of the change in objective function caused by removing the i^th feature by expanding the objective function in Taylor series to second order.

$$\Delta(i) = \frac{\partial}{\partial w_i}\Delta w_i + \frac{\partial^2}{\partial w_i^2}(\Delta w_i)^2 \tag{13}$$

At the optimum of *J*, the first order term can be neglected and with $= (1/2)\|w\|^2$ ,the equation becomes

$$\Delta(i) = (\Delta w_i)^2 \tag{14}$$

For tasks of selecting features in the input space, it is often unnecessary to calculate the true vector of w, which corresponds to the features of mapped data. We might only need the set of weights relevant to the features of input vectors. In RFE, the weight of a feature is approximately measured by the change of the objective value *J* in SVM model by leaving this feature out.

## 4. Experimental Results

We randomly split the original dataset into a training set and a test set and keep percentages of the positive and negative samples same in the training and test sets. We summarize some basic information about the datasets, including the number of features, the sizes of the training and test sets. However, the total numbers of available samples in our mass spectrometry datasets are small. In such a case, the test error may be biased due to an “unfortunate” partition of training and test sets. Thus, instead of reporting such an test error from one division of training and test sets, merging the training set and test set and then partition the total samples again into a training set and a test set randomly by stratified sampling for 100 times; for each division, first train a linear SVM classifier on the training set (hyperparameter *C* is to be selected by 10-fold cross-validation on the training set) and then test it on the corresponding test set; from this 100 trials we can compute the averages of performance measures.

**(1) Leukemia cancer data set**

Leukemia dataset consists of 72 samples: 25 samples of acute myeloid leukemia (AML) and 47 samples of acute lymphoblastic leukemia (ALL). The source of the gene expression measurements was taken form 63 bone marrow samples and 9 peripheral blood samples. Gene expression levels in these 72 samples were measured using high density oligonucleotide microarrays of 72 samples were used as training data and the remaining were used as test data in this paper. Each sample contains expression level of 7129 genes.

**(2) Colon cancer dataset**

Colon dataset consists of 62 samples of colon epithelial cells taken from colon-cancer patients. Each sample contains over 2000 gene expression levels. 40 of 62 samples are colon cancer samples and the remaining are normal samples. Each sample was taken from tumors and normal healthy parts of the colons of the same patients and measured using high density oligonucleotide array out of 62 samples were used as training data and the remaining were used as test data .

**(3) Lymphoma cancer dataset**

Lymphoma data sets cell diffuse large cell lymphoma (B-DLCL) is a heterogeneous group of tumors,4026 genes containing. Gene expression profiling has revealed two distinct tumor subtypes of B-DLCL: germinal center B cell-like DLCL and activated B cell-like DLCL. This data sets contains 77 tissue samples, 58 are diffuse large B-cell lymphomas (DLBLC) and remaining 19 samples are follicular lymphomas (FL).

Table 1.Accuracy Comparison with other algorithms on Leukemia Dataset

| Models | 8 genes | Best (<=8) | Mean (<=20) | Std (<=20) |
|---|---|---|---|---|
| S2N correlation | 0.8264 | 0.8356 | 0.8451 | 0.0254 |
| FC correlation | 0.8126 | 0.8237 | 0.8634 | 0.0213 |
| Default SVM-RFE | 0.8041 | 0.9012 | 0.0509 | 0.0489 |
| Extended SVM-RFE | 0.9816 | 0.9006 | 0.0600 | 0.0601 |

Table 2.Accuracy Comparison with other algorithms on Colon Dataset

| Models | 8 genes | Best (<=8) | Mean (<=20) | Std (<=20) |
|---|---|---|---|---|
| S2N correlation | 0.8648 | 0.8649 | 0.8611 | 0.0254 |
| FC correlation | 0.8226 | 0.8664 | 0.8611 | 0.0223 |
| Default SVM-RFE | 0.88871 | 0.9034 | 0.0559 | 0.0599 |
| Extended SVM-RFE | 0.9616 | 0.9677 | 0.0622 | 0.0632 |

Table 3.Accuracy Comparison with other algorithms on Lymphoma Dataset

| Models | 8 genes | Best (<=8) | Mean (<=20) | Std (<=20) |
|---|---|---|---|---|
| S2N correlation | 0.8268 | 0.8568 | 0.8604 | 0.0125 |
| FC correlation | 0.8266 | 0.8568 | 0.8604 | 0.0125 |
| Default SVM-RFE | 0.9056 | 0.9355 | 0.0549 | 0.0759 |
| Extended SVM-RFE | 0.9622 | 0.9624 | 0.0905 | 0.0600 |

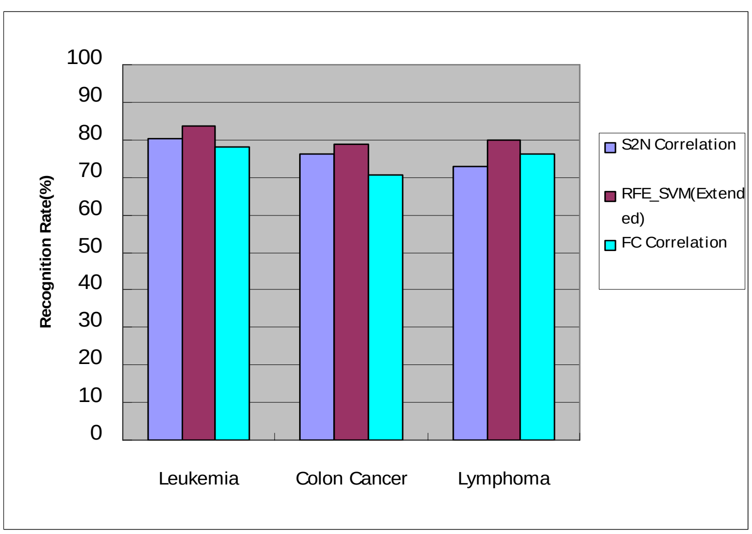


Figure .2. Performance Comparison with some correlation based methods

## 5. Conclusion

In this study, a new feature subset selection algorithm for classification task using SVMs was developed. The proposed method was assumed that very few features are needed to classify the given samples and smallest subset may provide more insight into the data. Looking at the performance of SVMs without SVM and with SVM in tables. The classification performance of extended SVM-RFE are much better than of other SVMs with all feature subset as input variables. In terms of dimensionality reduction, the best accuracy we get by starting from the original set and reduces the irrelevant features in each individual gene subset. The high prediction accuracy also strengths the promising application prospects of mass spectrometry patterns in the further cancer classification.


### Acknowledgements

This research was supported by SDRC (Software Development and Research Center) in University of Computer Studies, Yangon. SDRC is a Research Center designated by Myanmar Science and Engineering Foundation and Ministry of Science & Technology.


### References


[1] C.Emmanouilidis, A.Hunter,and J.MacIntyre, ”A multiobjective evolutionary setting for feature selection and a commonality-based crossover operator”, in Proceedings of the 2000 Congress on Evolutionary Computation(CEC00).

[2] Cancer Program Data Set [http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi]
Chang,Chih-Chung and Lin,Chih-Jen.” LIBSVM: a library for support Vector Machines”, http://www.csie.ntu.edu.tw/~cjlin/libsvm, 2003

[3] Dudoit S,Fridlyand J,Speed TP: Comparison of discrimination method for the classification of tumors using gene expression data.J Amer Statist Assoc 2002,97:77-87.

[4] Eisen, M.B and Brown ,P.O.(1999); DNA arrays for analysis of gene expression. Methods Enzymbol, 303: 179-205
Furey,T.S.,Cristianini,N.,Duffy,N.,Bednarski,D.W., Schummer,M. and Haussler,D.(2000):Support Vector Machines Classification and validation of cancer tissue samples using microarray expression data.Bioinformatics,16(10):906-914.

[5] Golub TR,Sloonim DK,Tamayo P,Huard C,Gassenbeek M,Mesirov JP,Coller Fgff,Loh ML,Downing JR,Caligiuri MA,Bloomfield CD,Lander ES:Molecular classification of cancer:class discovery and class prediction by gene expression monitoring.Science 1999,286:531-537.

[6] Harrington,C.A.,Rosenow,C., and Retief,J.(2000); Monitoring gene expression using DNA microarrays.Curr. Opin.Microbiol.,3: 285-291.

[7] I.Guyon, J.Weston, S.Barnhill,and V.Vapnik. Gene Selection for cancer classification using support vector machines.Machine Learning,2000.

[8] K.M. Win and Kham N.S.M, “Minimizing Essential Set Based Feature selection for Cancer Classification”, ICCA2008, Yangon, Myanmar, Feb 14-15, 2008

[9] Li,L.,Weinberg ,C.R.,Darden,T.A.and Pedersen,L.G.(2001):Gene Selection for sample classification base don gene expression data: study of sensitivity to choice of parameters of GA/KNN method.Bioinformatics,17(12):1131-1142.

[10] M.Momma and K.P.Bannett.A Pattern search method for model selection of support vector regression.In Proceeding of the SIAM International Conference on Data Mining.Philadelphia,Pennsylvania,2002.SIAM

[11] N.Cristaianini and J.Shawe-Taylor.*Introduction to Support Vector Machines* .Cambridge University Press,2000

[12] T.Furey, N.Cristianini,N.Cristianini,N.Duffy,D.Bednarski,M.Schummer,and D.Haussler,” Support Vector Machine Classification and Validation of Cancer Tissue Samples Using Microarray Expression Data,” *Bioinformatics*,vol.16,pp.906-914,2000

[13] P.Pavlidis at al.,”Gene Functional Analysis from Heterogeneous Data,” Proc. Conf. *Research in Computational Molecular Biology (RECOMB)*,pp.249-255,2001.

[14] V. Vapnik and O.Chapelle. Bounds on error expectation for support vector machines. Neural Computation, 12(9)2000.

[15] Wessels Weston,J. and C.Watkins.”Support Vector Machines for Multi-Class Pattern Recognition”, Proceedings of he Seventh European Symposium On Artificial Neural Networks,1999.