Minimum Reference Set Based Feature Selection for Small Sample Classifications

Xue-wen Chen XWCHEN@KU.EDU
Jong Cheol Jeong JCJEONG@KU.EDU
Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS 66049 USA

Abstract

We address feature selection problems for classification of small samples and high dimensionality. A practical example is microarray-based cancer classification problems, where sample size is typically less than 100 and number of features is several thousands or higher. One of the commonly used methods in addressing this problem is recursive feature elimination (RFE) method, which utilizes the generalization capability embedded in support vector machines and is thus suitable for small samples problems. We propose a novel method using minimum reference set (MRS) generated by the nearest neighbor rule. MRS is the set of minimum number of samples that correctly classify all the training samples. It is related to structural risk minimization principle and thus leads to good generalization. The proposed MRS based method is compared to RFE method with several real datasets, and experimental results show that the MRS method produces better classification performance.

1. Introduction

High dimensional data analysis is an extremely crucial task in various applications, such as multi/hyperspectral data-based target detection and classification (Schweizer and Moura, 2000), and microarray data-based cancer classification (Xiong and Chen, 2006). On one hand, the high dimensionality provides rich information about the data and offers the potential to distinguish between different classes. On the other hand, in most practical cases, the number of labeled training data is very small compared to the number of features available. For example, in microarray data-based cancer classification problems, typical number of samples for each class is less than 100 and the dimensionality is several thousands or tens of thousands. Learning from small samples with high dimensionality poses a significant challenge to machine learning society, for example, computational complexity (the computational demands for searching in high dimensional spaces grow exponentially with data dimension) and overfitting (models obtained from high dimensional data fit the training data very well, but perform poorly on previously unseen data).

Developing classification methods to overcome the over-fitting problems has already attracted significant interest from machine learning community (Bradley and Mangasarian 1998; Fung et al., 2002; Vapnik and Chapelle 2000; Guyon et al., 2002; Reunanen, 2003; Weston et al., 2003). As one of the most commonly-used learning methods, support vector machine (SVM) has shown excellent performance in handling large feature space and overfitting problems (Chapelle et al., 2002; Guyon et al., 2002; Vapnik 1998; Haykin 1999; Weston et al., 2000). A SVM yields its decision function derived from the structural risk minimization (SRM) principle. Unlike the empirical risk minimization, which minimizes the errors on training data and consequently leads to overfitting, the SRM principle suggests that we should minimize an upper bound on the expected risk by controlling both the number of training errors and the capacity of the set of candidate functions measured by the so-called Vapnik-Chervonenkis dimension (Vapnik, 1998).

Another approach for counteracting the overfitting problems and for reducing the computational complexity for the analysis of small samples with high dimensionality is feature selection. Feature selection is the process of searching for a subset of relevant features from a larger set of original ones in terms of some pre-defined criteria, such as classification performance or class separability. In fact, feature selection methods play a significant role for solving small sample classification problems where the number of features is much larger than the number of training samples. It has been shown that feature selection can also improve the performance of SVMs for small sample classification problems. Enlightened to the fact that SVM generalizes well, Guyon et al. (2002) recently developed a feature selection method, called recursive feature elimination (RFE), for small sample classification problems. The RFE method is originally applied to microarray-based cancer

Appearing in *Proceedings of the 24^{dth}International Conference on Machine Learning*, Corvallis, OR, 2007. Copyright 2007 by the author(s)/owner(s).

classification where the number of training samples is less than 100 and the number of features is several thousands, and has become an effective approach in small-sample feature selection problems. Based on the idea of Optimal Brain Damage theory (Le Cun et al., 1990), RFE seeks to improve generalization performance by removing the least important features whose deletion will have the least effect on training errors. The importance of a feature is evaluated in terms of a criterion derived from SVMs.

While it has shown great promise in small-sample feature selection problems, the RFE method tends to remove redundant and weak features and retains independent features. As pointed out by Guyon and Elisseeff (2003): (1) presumably redundant features may provide better class separation, and (2) two weak features that are useless by themselves can provide a significant performance improvement when used together. Thus, simply removing redundant or weak features may degrade classification performance. This is particularly true when few features are retained as we observed in our experiments. Another potential issue is that the maximal margin decision boundary derived from SVMs exists in nonlinear feature space, not necessary in observation space (Karacal and Krim 2002).

It is generally accepted that the generalization performance is closely related to the trade-off between the number of training samples used and the model capacity (Bottou and Vapnik, 1992). As pointed out by Vapnik (Bishop, 1998), “the function that describes data well and belongs to a set of functions with low capacity will generalize well regardless of the dimensionality of the input space.” As a local algorithm, one nearest neighbor (1-NN) classifier has a very low capacity. Karacal and Krim (2002) recently showed that the complexity of a 1-NN classifier is directly related to the reference set derived from the training set. A reference set is a subset of training set that can correctly classify all training samples through the 1-NN rule. Thus, a better generalization can be achieved by replacing the training set with a small reference set. In this paper, we propose a minimum reference set (MRS) based feature selection method. The MRS method evaluates feature sets in terms of the size of MRS in observation space. We argue that for two feature subsets that classify all the training samples correctly through 1-NN rule, the one with smaller MRS is expected to generalize well. We compare MRS and RFE methods on various practical datasets and show that the MRS method significantly improves generalization accuracy.

The remainder of this paper is organized as follows. Section 2 first introduces the RFE method. We then describe the proposed MRS method. Section 3 presents the experimental results of six datasets with simple samples. Finally, Section 4 presents our conclusions.

2. Method

2.1 Recursive Feature Elimination

For a linearly separable problem, SVMs find a discriminant function, $g(\mathbf{x}_i) = \mathbf{w} \cdot \mathbf{x}_i + b$, where b is a bias term, $\boldsymbol{x_i} \in \Re^{\boldsymbol{n}}$ are samples, and y_i are corresponding class labels $y_i = \{\pm 1\}, i = 1, ..., m$. The discriminant function satisfies following constraint:

$$\begin{aligned} g(\mathbf{x}_i) > 0, \ \ if \ \ y_i = 1 \\ g(\mathbf{x}_i) < 0, \ \ if \ \ y_i = -1 \end{aligned} \tag{1}$$

For linearly non-separable cases, one can introduce slack variables ξ_i and accordingly, the discriminant function is defined by:

$$y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0 \tag{2}$$

ξ_i measure the deviation of a data point from optimal hyperplane (Vapnik, 1998). SVMs are designed by minimizing

$$\begin{aligned} &\Phi(\mathbf{w}, \xi) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i \\ &Subject\ to: y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0 \end{aligned} \tag{3}$$

The optimization problem is solved in a dual problem:

$$\begin{aligned} &W(\boldsymbol{\alpha}) = \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{m} y_i y_j \alpha_i \alpha_j (x_i \cdot x_j) \\ &Subject\ to: 1) \quad 0 \leq \alpha_i \leq C, \ \ i = 1, ..., m \\ &\qquad\qquad 2) \quad \sum_{i=1}^{m}\alpha_i y_i = 0 \end{aligned} \tag{4}$$

where α_i are the Lagrange coefficients.

The linear SVMs can be readily extended to nonlinear SVMs where more sophisticated decision boundaries are needed. This is done by applying the kernel trick, i.e., simply replacing every dot product $(\boldsymbol{x_i} \cdot \boldsymbol{x})$ in linear SVMs by a nonlinear kernel function $\boldsymbol{K}(\boldsymbol{x_i} \cdot \boldsymbol{x})$, which satisfies Mercer’s Theorem (Vapnik, 1998).

The RFE method is based on the concept in Optimal Brain Damage (Le Cun et al., 1990) and SVMs. It seeks to improve generalization capability and speed of learning by recursively removing features with the smallest weight values w_i calculated from SVM training. At each step, the vector $\boldsymbol{w}$ is calculated by training a SVM using the remaining features. RFE simply removes a weak feature measured by its weight value w_i. It does not consider the effect of removing a feature on the performance of SVMs. However, a weak feature may still be an important feature when used with other features together. Thus, simply removing redundant or weak features may degrade classification performance.

2.2 Minimum Reference Set

In this section, we describe the proposed MRS feature selection method, which uses reference set sizes to evaluate the importance of a set of features.

A minimum reference set is the smallest subset of training set that can correctly classify all training samples through the 1-NN rule. Since the complexity of a 1-NN classifier is directly related to the number of training samples involved, the size of a MRS is closely tied to the structural risk minimization (SRM), and thus the generalization ability.

The SRM principle for learning from samples of small size is to find the decision function that minimizes the guaranteed risk on test data (Vapnik, 1998). This is achieved by controlling model capacity. Let $\Im$ be a set of indicator functions defined on the training set $(\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_m, y_m)$, and let $R(f)$ denote the risk for an indicator function $f \in \Im$. The guaranteed risk can be derived through the bounds on the actual risk (Vapnik, 1998).

Theorem 1 (Luntz and Brailovsky). *The leave-one-out estimator is almost unbiased.*

Theorem 2. *Let E[R(f)] be the expectation of the probability of error taken over both training and test data for an optimal indicator function f constructed on training samples of size m and 1-NN. Let N_m denote the size of the MRS formed on the basis of training samples of size m. Then the following inequality holds true,*

$$E[R(f)] \le \frac{E(N_m)}{m} \tag{5}$$

To prove this theorem, we follow the similar steps to Vapnik (1998). Apparently, the removal of a sample $\boldsymbol{x}_i \notin$ MRS from the training set will not change the MRS. Thus, in the leave-one-out method, samples $\boldsymbol{x}_i \notin$ MRS will be correctly classified. Therefore, the number of errors by the leave-one-out method does not exceed the size of the MRS, that is, the largest error rate for training data using the leave-one-out method is N_m /m. According to Theorem 1, Eq. (5) is true.

□

From Theorem 2, we conclude that the generalization ability of the indicator function constructed on the basis of the MRS depends on the size of the MRS. Minimizing the size of the MRS on the basis of empirical data leads to minimizing the structure risk $R(f)$. For two feature sets with the same size, we can create two minimum reference sets for zero training errors. The feature set with a smaller MRS is expected to have better generalization ability, as fewer training samples are used for constructing the classifier. Thus, the proposed MRS method seeks for the feature subset that needs smallest MRS for classification.

We first describe the procedures to find a MRS. Starting with an empty set, we update a reference set by adding the closest samples between classes until all training samples are correctly classified through 1-NN classifier. Apparently, this algorithm always converges. In the worst case, all training samples are included into the reference set (Karacal and Krim 2002). For calculating distances between samples on different classes, the Euclidean distance, $d(\mathbf{x}_i, \mathbf{x}_j)$, is used.

MRS_ID: MRS Identifier Algorithm

```
I = set of selected samples = $\varnothing$.
$err(I)$ : classification error using 1-NN and training samples in I.
$\mathbf{d}$ : ranked distances calculated from samples of between classes.
$d_k$ : $k^{th}$ element in d
Step1: calculate all pairwise distance $d(\mathbf{x}_i, \mathbf{x}_j)$ for samples from two classes, i.e, $y_i$ = 1 and $y_j$ = -1
Step2: sort the distance from the smallest to the largest and store the ranked distance in d. Set k = 1.
Step3:repeat
        Find i and j which is related to $d(\mathbf{x}_i, \mathbf{x}_j) = d_k$
        if $\{i,\ j\} \not\subset \mathbf{I}$
            update $I \leftarrow I \cup \{i,\ j\}$
      end if
        k = k + 1
until $err(I) = 0$
return (I)
```

The final set *I* is the MRS.

Next, assume that the number of features to be selected is *k*. the MRS method randomly chooses a set of *k* features and swaps one feature at a time between the selected feature set ***SF*** and the remaining feature pool ***RF***. For each feature combination, MRS Identify algorithm is executed to obtain a MRS. If the size of the MRS for the new feature set ***SF*** (after swapping) is smaller than that before swapping, the swapping is accepted; otherwise, the feature set remains the same. We repeat this process for all the features in ***SF***. The feature set with the smallest number of a reference set is considered as the best feature set.

MRS Feature Selection Algorithm

k = the number of selected features

n = original number of features

$N(F)$ = the size of reference set with feature set F

S = the size of MRS

F = final feature set

SF = set of selected features

RF = set of remaining features

```
Step1: Randomly select $k$ features,
    $SF = \{ f_1, f_2, f_3, ..., f_k \}$
    $RF = \{ f_{k+1}, f_{k+2}, ..., f_n \}$
Step2: Search possible $k$ features with smallest MRS.
    perform MRS_ID for samples with feature set $SF$,
    $F = SF;$   $S = N(SF)$.
    for $i$ = 1 to $k$
        for $j = k$+1 to n
            swap $f_i$ in $SF$ and $f_j$ in $RF$
            perform MRS_ID for samples with feature set $SF$
            $S_1 = N(SF)$.
            If $S_1 < S$,
                accept the swap ($S = S_1$, $F = SF$).
            end if
        end j
    end i
    return ($F$)
```

The best feature set with smallest MRS is saved in F.

Computationally, the MRS feature selection method executes MRS_ID $k \times (n - k)$ times. Each time, one feature in ***SF*** will be replaced by a different feature. The new feature set is then evaluated as a whole, instead of evaluating one feature at a time as in the RFE method. For better results, the search process can be repeated several times with random restart (step 1). Alternatively, one can run the algorithm just once by using a deterministic starting feature subset created by another feature selection algorithm (e.g., RFE). The latter case is employed in our study.

Two major differences between MRS method and RFE method are: (1) MRS evaluates the importance of a group of features, while RFE evaluates the importance of individual features, one at a time; and (2) MRS evaluates feature sets using reference set sizes which are directly tied to the structural risk minimization principle and thus good generalization, while RFE evaluates individual features in terms of their weights calculated from SVM training. Next, we apply the MSR method to six datasets, each with a small number of training samples.

3. Experimental Results

3.1 Datasets Description

Six datasets, all with small number of training samples, are used to compare RFE and MRS. The first dataset (sonar) is downloaded from UCI machine learning repository (http://www.ics.uci.edu/~mlearn) and the other five sets are microarray datasets, as summarized in Table1. For all the microarray data sets, since the largest number of samples for each class is less than 60, we use bootstrapping method for evaluating the proposed method. Specifically, for each dataset, we randomly generate (sampling with replacement) 70% training samples and 30% test samples. This is done 15 times. Thus, for each dataset, we now have 15 sub-groups of a training set and a test set, and test results are averaged over the 15 randomly generated sub-groups of test sets.

3.2 Results

To evaluate the MRS feature selection method, selected features are compared with those selected by RFE method. Since MRS and RFE features are selected through 1-NN and SVMs, respectively, we compare classification performance with both 1-NN and SVMs as classifiers. We use linear SVMs in all cases.

Figures 1 to 6 show the classification accuracy versus the number of selected features. Lines with cross markers represent results for MRS features (solid lines with cross markers for a SVM classifier (MRS-SVM) and dashed lines with cross markers for an 1-NN classifier (MRS-NN)). Lines without cross markers are for RFE features (solid lines for a SVM classifier (RFE-SVM) and dashed lines for an 1-NN classifier (RFE-NN)).

For sonar data (Figure1), MRS features clearly outperform RFE features, either with the 1-NN or the SVM classifier. For ALL/AML data (Fig. 2) and COLON data (Fig. 5), when the number of features is larger than 30, both methods are comparable. When the number of feature is less than 30, MRS features produce better classification accuracy. For CNS data (Fig. 3), with more than 15 features, results for MRS and RFE methods are comparable. With less than 15 features to use, RFE features with a SVM classifier yields highest accuracy. Finally, for BREAST data (Fig. 4) and LYMPH data (Fig.

6), MRS features clearly produce better classification accuracy than RFE methods. It is interesting to note that for BREAST data and LYMPH data, regardless of the classifiers to use, MSR methods perform better than RFE methods in most cases (Figs. 4 and 6). In conclusion, MRS methods outperform RFE methods most of the time, especially when the number of features is small. Note that in practice, small number of features is preferred to overcome overfitting problems for small sample classification problems. Thus, the MRS method is of practical use and interest.

To visualize the features selected by MRS and RFE methods, we plot both training and test data of ALL/AML with the best two features. We randomly select a training data set and a test set generated by bootstrapping and run MRS and RFE feature selection methods to select two best features. Figures 7 and 8 show the training and test data with the top two features selected by the MRS method, respectively. Figures 9 and 10 show the training and test data with the top two features selected by the RFE method, respectively. Apparently, two classes in MRS features are better separated than in RFE features.

Figure 11 shows the average percentage of training samples in minimum reference set (the ratio of training samples in reference set to the number of original training samples) versus number of features with COLON data. As expected, for different number of features to use, number of samples in MRS differs.

4. Conclusion

Classification problems with small sample sizes and very high dimensionality have drawn increasing attention in machine learning community. An essential step in small sample classification is feature selection. In this paper, we propose and apply a minimum reference set based method to feature selection and compare it to the commonly used RFE method. In a RFE method, a feature is removed if it is weak at a particular step. The weakness is evaluated in terms of its weight value in constructing a SVM decision hyperplane. A weak feature, however, might be important when combined with other features. Our proposed method assesses features as a group based on the minimum reference set derived from a 1-NN classifier. MRS methods implement structural risk minimization principle and guarantee to generalize well. We compare the proposed MRS method to RFE method on six datasets, each with small training samples. The MRS method makes significantly improvement over the RFE method, especially for small number of features, which are of practical use. Our future work will address the problem of extension of the MRS method to multi-class feature selection problems. Unlike RFE method which is based on SVMs, MRS method is based on 1-NN method. Consequently, we expect that the MRS method can be readily applied to feature selection for multi-class classification problems.

Table 1 Data description

SONAR: This data set consists of 208 instances and 60 attributes (Gorman and Sejnowski 1988). The task is to classify sonar signals that bounce off a metal cylinder or a roughly cylindrical rock. The data are divided equally into two sets: 104 instances are used for training and rests are used as test.

ALL/AML: ALL-AML Leukemia Data (Golub et al., 1999). This data set contains 72 samples of human acute leukemia. 47 samples belong to acute lymphoblastic leukemia(ALL), and the other acute myeloid leukemia(AML). Each sample presents the expression levels of 7129 genes.

CNS: Embryonal Tumors of Central Nervous System Data (Pomeroy et al., 2002). This data set contains 60 patient samples, 21 are survivors of a treatment, and 39 are failures. There are 7129 genes in the data set.

BREAST: Breast Cancer Data (West et al., 2001). This data contains 7129 genes in 49 breast tumor samples. There are two classes: 25 samples are estrogen receptor positive (ER+), whereas the remaining 24 samples are estrogen receptor negative (ER-).

COLON : Colon Tumor Data (Alon et al., 1999). This data set contains 62 samples collected from colon-cancer patients. Among them, 40 samples are from tumors and 22 biopsies are from healthy parts of the colons of the same patients. 2000 genes are selected to measure their expression levels.

LYMPH: Lymphoma Data (Shipp et al., 2002). This data set contains 77 tissue samples, 58 are diffuse large B-cell lymphomas (DLBCL) and remaining 19 samples are follicular lymphomas (FL). Each sample is represented by the expression levels of 7129 genes.

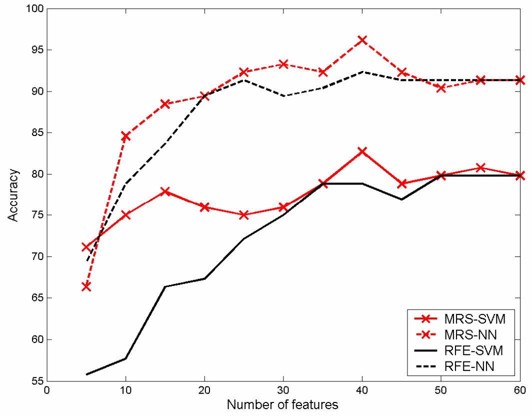

Figure 1 Test accuracy for sonar data

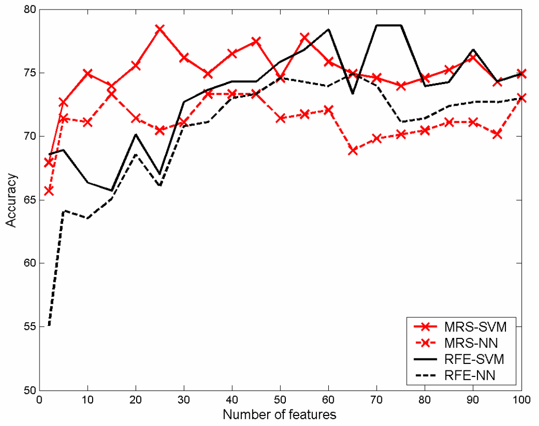

Figure 2 Test accuracy for ALL/AML data

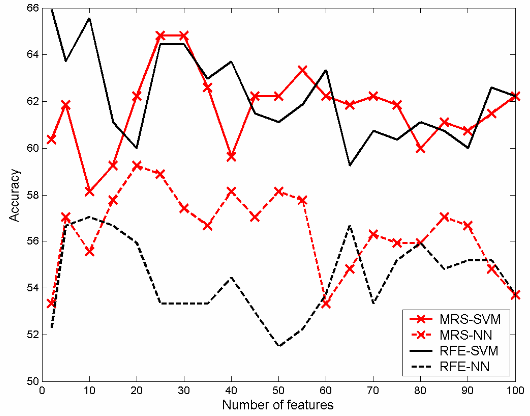

Figure 3 Test accuracy for CNS data

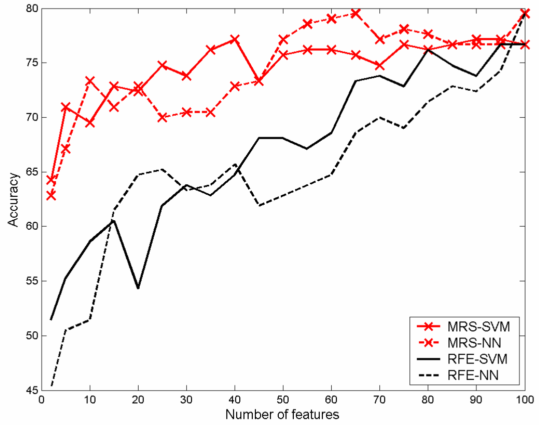

Figure 4 Test accuracy for BREAST data

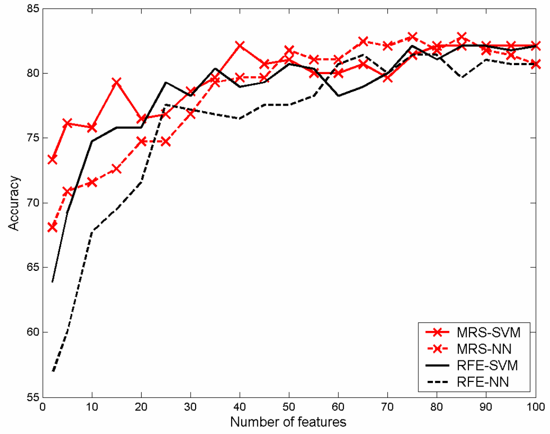

Figure 5 Test accuracy for COLON data

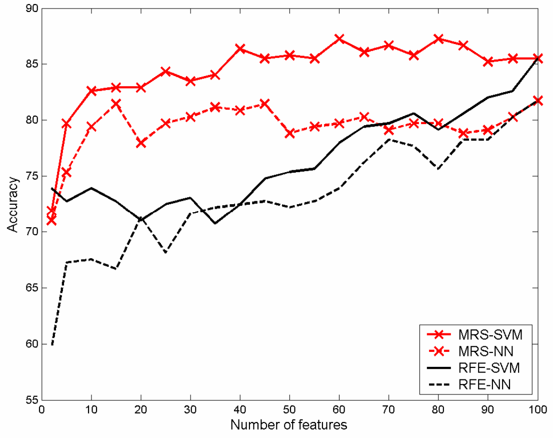

Figure 6 Test accuracy for LYMPH data

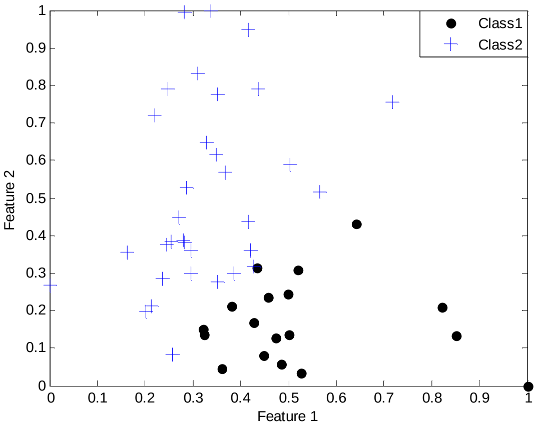

Figure 7 Train data distribution of ALL/AML with the best two features selected by MRS.

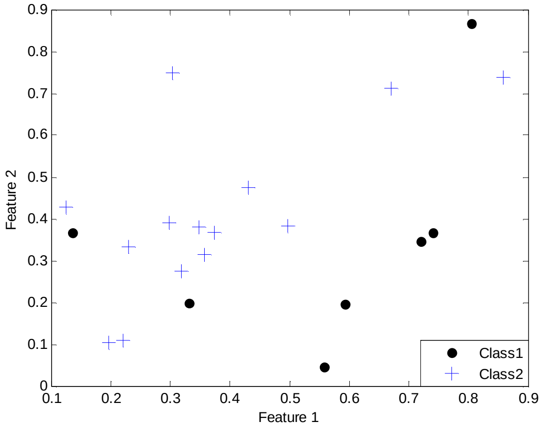

Figure 8 Test data distribution of ALL/AML with the best two features selected by MRS.

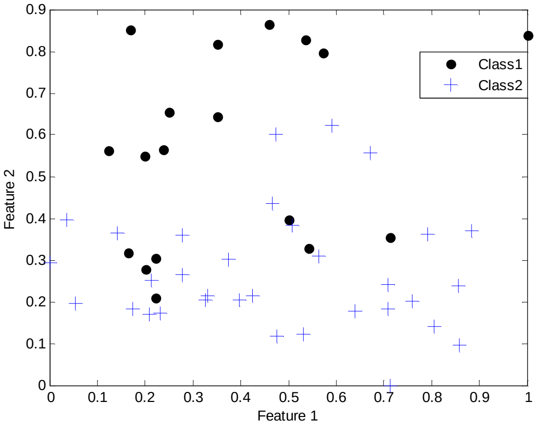

Figure 9 Train data distribution of ALL/AML with the best two features selected by RFE.

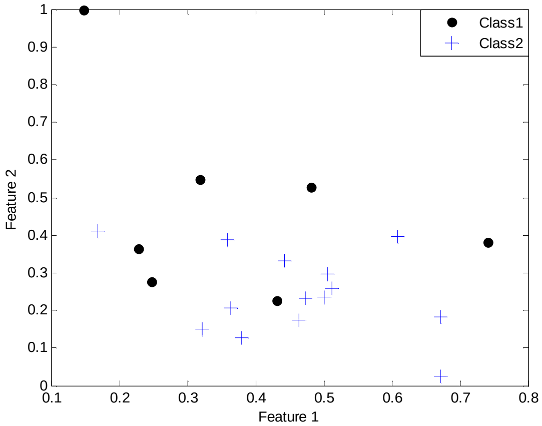

Figure 10 Test data distribution of ALL/AML with the best two features selected by RFE.

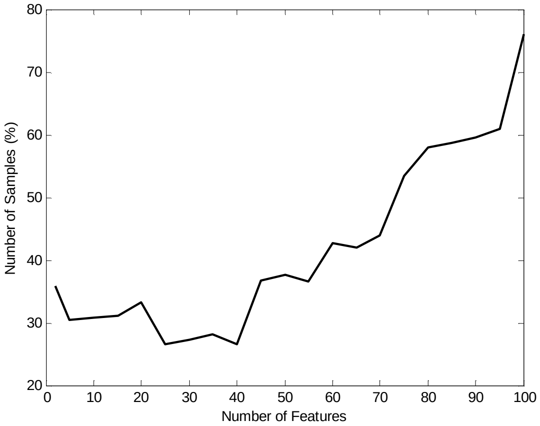

Figure 11 Average percentage of samples in MRS for COLON data (averaged over 15 bootstrapped training data sets).

Acknowledgments

We would like to thank the reviewers for their valuable comments. This material is based upon work supported in part by the U. S. Army Research Office under contract number W911NF-06-1-0351 and by NSF award IIS-0644366.

References

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., and Levine, A.J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissue probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci*. USA, 96, 6745-6750, June.

Bishop, C. M. (Ed.) (1998). Neural Networks and Machine Learning, NATO ASI Series, Series F: *Computer and Systems Sciences*, 168, Berlin: Springer-Verlag.

Bottou, L. and Vapnik, V. (1992). Local Learning Algorithms, *Neural Computing*, 4, 888-890.

Bradley, P. S. and Mangasarian, O. L. (1998). Feature selection via concave minimization and support vector machines. *Proc. 13th ICML*, 82-90, San Francisco, CA.

Chapelle, O. Vapnik, V. Bousquet, O. and Mukherjee, S. (2002). Choosing kernel parameters for support vector machines. *Machine Learning,* 46(1-3), 131-159.

Le Cun, Y., Denker, J., and Solla, S. (1990). Optimal Brain Damage. *Advances in Neural Information Processing Systems* 2, 598-605.

Fung, G., Mangasarian, O. L., and Smola, A. J. (2002). Minimal kernel classifiers*. Journal of Machine Learning* Research 3, 303-321.

Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gassenbeek, M. Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D., and Lander, E.S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286, 531-537.

Gorman, R. P., and Sejnowski, T. J. (1988). Analysis of Hidden Unitsin a Layered Network Trained to Classify Sonar Targets, *Neural Networks*, 1, 75-89.

Guyon, I., Weston J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning,* 46(1-3), 389-422.

Guyon, I., and Elisseeff, A. (2003). An introduction to variable and feature selection. *JMRL special Issue on variable and Feature Selection* 3, 1157-1182.

Haykin, S. (1999). *Neural Networks a comprehensive foundation (2nd edition).* Prentice-Hall, NJ, 1999

Karacal, B., and Krim, H. (2002). Fast Minimization of structural risk by nearest neighbor rule. *IEEE transactions on neural networks*, 14(1), 127-137.

Luntz, A. and Brailovsky, V. (1969). On estimation of characters obtained in statistical procedure of recognition. *Technicheskaya Kibernetica*, 3.

Pomeroy, S.L., Tamayo, P. Gaasenbeek, M., Sturla, L.M., Angelo, M., McLaughlin, M.E., Kim, J.Y.H., Goumnerova, L.C., Black, P.M., Lau, C., Allen, J.C., Zagzag, D., Olson, J.M., Curran, T. Wetmore, C., Biegel, J.A., Poggio, T., Mukherjee, S., Rifkin, R., Califano, A., Stolovitzky, G., Louis, D.N., Mesirov, J.P., Lander, E.S., and Golub, T.R. (2002). Prediction of central nervous system embryonal tumor outcome based on gene expression. *Letters to Nature, Nature*, 415, 436-442.

Reunanen, J. (2003). Overfitting in making comparisons between variable selection methods. *JMLR special Issue on variable and Feature Selection* 3, 1371-1382.

Schweizer, S. and Moura, J. (2000). Hyperspectral imagery: clutter adaptation in anomaly detection. *IEEE Trans. on Information Theory*, vol. 46(5), 1855-1871.

Shipp, M.A., Ross, K.N., Tamayo, P., Weng, A.P., Kutok, J.L., Aguiar, R.C.T., Gaasenbeek, M., Angelo, M., Reich, M., Pinkus, G.S., Ray, T.S., Koval, M.A., Last, K.W., Norton, A., Lister, T.A., Mesirov, J., Neuberg, D.S., Lander, E.S., Aster, J.C., Golub, T.R. (2002). Diffuse Large B-Cell Lymphoma Outcome Prediction by Gene Expression Profiling and Supervised Machine Learning. *Nature Medicine*, vol.8, 68-74.

Vapnik, V. (1998). *Statistical Learning Theory*. John Wiley and Sons, New York.

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

West, B., Blanchette, C., Dressman, H. Huang, E. and et.al. (2001). Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc. Natl. Acad. Sci.* USA, 98, 11462-11467.

Weston, J., Mukherjee, S., Chapelle, O. Pontil, M. Poggio, T. and Vapnik, V. (2000). *Feature selection for support vector machines.* In Advances in Neural Information Processing Systems.

Weston, J. Elisseeff, A. Scholkopf, B. and Tipping, M.(2003) Use of the zero-norm with linear models and kernel methods. *JMLR special Issue on variable and Feature Selection 3,* 1439-1461.

Xiong, H. and Chen, X. (2006). Kernel-Based Distance Metric Learning for Microarray Data Classification. *BMC Bioinformatics*, 7:299.