

NETTAB 2004

Workshop on Models and Metaphors from Biology to Bioinformatics Tools

Random subspace ensembles for the biomolecular diagnosis of tumors

Alberto Bertoni, Raffaella Folgieri, Giorgio Valentini

{bertoni,folgieri,valentini}@dsi.unimi.it

Dipartimento di Scienze dell'Informazione
Università degli Studi di Milano

Outline

- The problem of the bio-molecular diagnosis of tumors using gene expression data
- Current approaches to bio-molecular diagnosis
- The Random Subspace (RS) ensemble method
- Reasons to apply RS ensembles to the bio-molecular diagnosis of tumors
- Experimental results
- Open problems


Bio-molecular diagnosis of malignancies: motivations

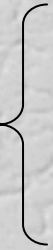
- Traditional clinical diagnostic approaches may sometimes fail in detecting tumors (Alizadeh et al. 2001)
- Several results showed that bio-molecular analysis of malignancies may help to better characterize malignancies (e.g. gene expression profiling)
- Information for supporting both diagnosis and prognosis of malignancies at bio-molecular level may be obtained from high-throughput biotechnologies (e.g. DNA microarray)

Bio-molecular diagnosis of malignancies: current approaches

- Huge amount of data available from biotechnologies: analysis and extraction of significant biological knowledge is critical
- Current approaches: statistical methods and machine learning methods (*Golub et al., 1999; Furey et al., 2000; Ramaswamy et al., 2001; Khan et al., 2001; Dudoit et al. 2002; Lee & Lee, 2003; Weston et al., 2003*).

Main problems with gene expression data for bio-molecular diagnosis

- High dimensionality
 - Low cardinality
- 
- Curse of dimensionality

- Data are usually noisy: 
 - Gene expression measurements
 - Labeling errors

Current approaches against the curse of dimensionality

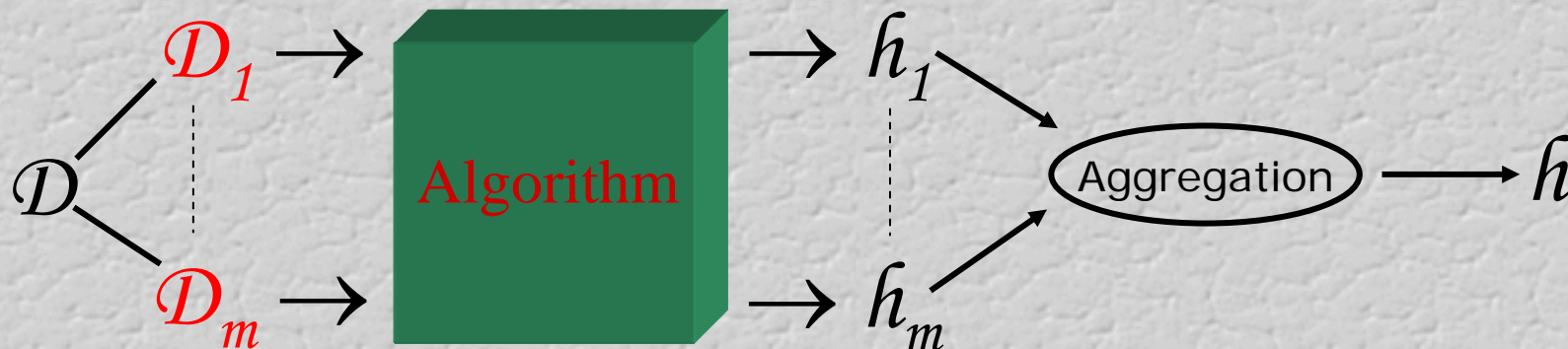
- *Selection of significant subsets of components (genes)*
e.g.: filter methods, forward selection, backward selection, recursive feature elimination, entropy and mutual information based feature selection methods (see *Guyon & Elisseeff, 2003* for a recent review).
- *Extraction of significant subsets of features*
e.g.: Principal Component Analysis or Independent Component Analysis

Anyway, both approaches have problems ...

An alternative approach based on ensemble methods

Random subspace (RS) ensembles:

- RS (Ho, 1998) reduce the high dimensionality of the data by randomly selecting subsets of genes.
- Aggregation of different base learners trained on different subsets of features may reduce variance and improve diversity



The RS algorithm

Input: a d -dimensional labelled gene expression data set

1. Select a random projection from the d -dimensional input space to a k -dimensional subspace
2. Project the data from the d -dimensional space into the selected k -dimensional subspace
3. Train a classifier on the obtained k -dimensional gene expression data, using a suitable learning algorithm
4. Repeat steps 1-3 m times
5. Aggregate the trained classifiers by majority (or weighted) voting

Reasons to apply RS ensembles to the bio-molecular diagnosis of tumors

- Gene expression data are usually very high dimensional, and RS ensembles reduce the dimensionality and are effective with high dimensional data (*Skurichina and Duin, 2002*)
- Co-regulated genes show correlated gene expression levels (see e.g. *Gasch and Eisen, 2002*), and RS ensembles are effective with correlated sets of features (*Bingham and Mannila, 2001*)
- Random projections may improve the diversity between base learners
- Overall accuracy of the ensemble may be enhanced through aggregation techniques (at least w.r.t. the variance component of the error)

Experimental environment

We considered 2 *bio-medical problems* both based on gene-expression profiles of a relatively small group of patients:

1. *Colon adenocarcinoma diagnosis* (Alon *et al.*, 1999): 62 samples, 40 colon tumors and 22 normal colon samples, 2000 genes.
2. *Medulloblastoma clinical outcome prediction* (Pomeroy *et al.*, 2002): 60 samples, 39 survivors and 21 treatment failures, 7129 genes.

Methods:

- RS ensembles with linear SVMs as base learners
- Single linear SVMs

Software: C++ NEUROObjects library (Valentini, 2002)

Hardware: Avogadro cluster of Xeon double processor workstations

Results

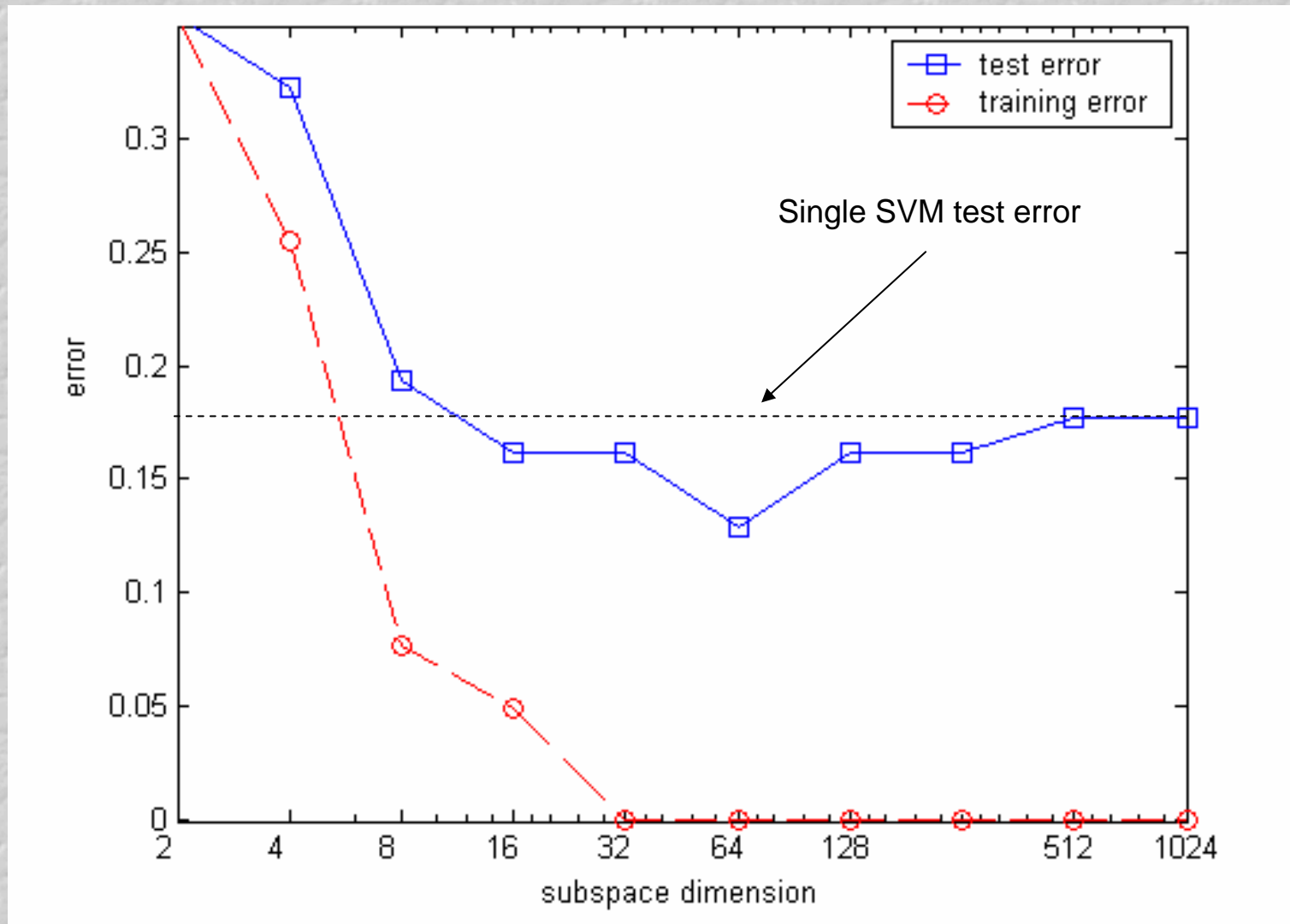
Colon tumor prediction (5 fold cross validation)

	Test Err.	St. dev.	<u>Train.Err.</u>	St. dev.	Sens.	<u>Spec.</u>	<u>Prec.</u>
RS ensemble	0.1290	0.0950	0.0000	0.0000	0.9000	0.8182	0.9000
Single SVM	0.1774	0.1087	0.0000	0.0000	0.8500	0.7727	0.8718
Single base SVM	0.1776	0.1019	0.0000	0.0000	---	---	---

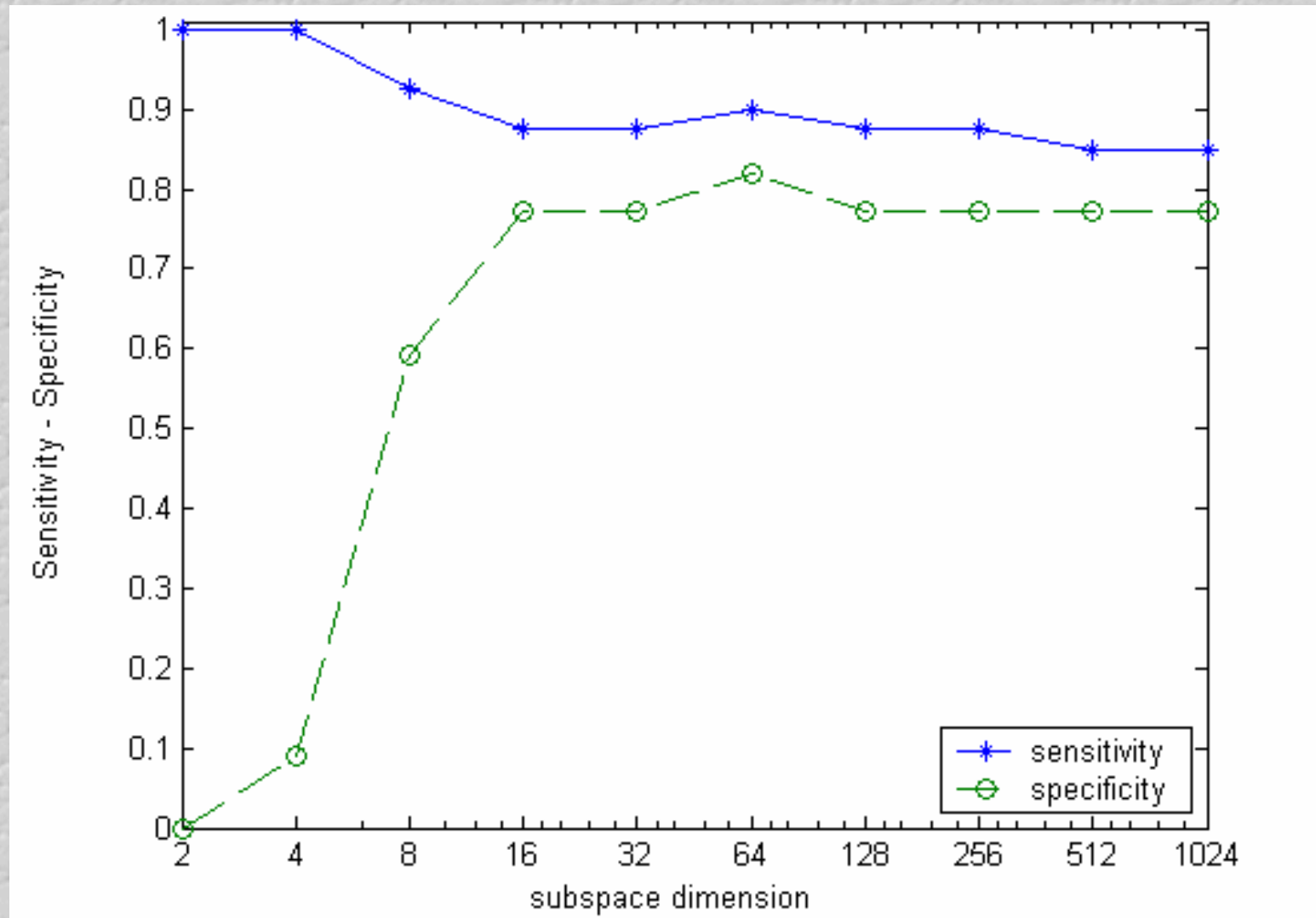
Medulloblastoma clinical outcome prediction (5 fold cross validation)

	Test Err.	St. dev.	Train.Err.	St. dev.	Sens.	Spec.	Prec.
RS ensemble	0.2333	0.1087	0.0000	0.0000	0.5714	0.8718	0.7059
Single SVM	0.2833	0.0950	0.0083	0.0114	0.5238	0.8205	0.6111
Single base SVM	0.2916	0.1008	0.0092	0.0103	---	---	---

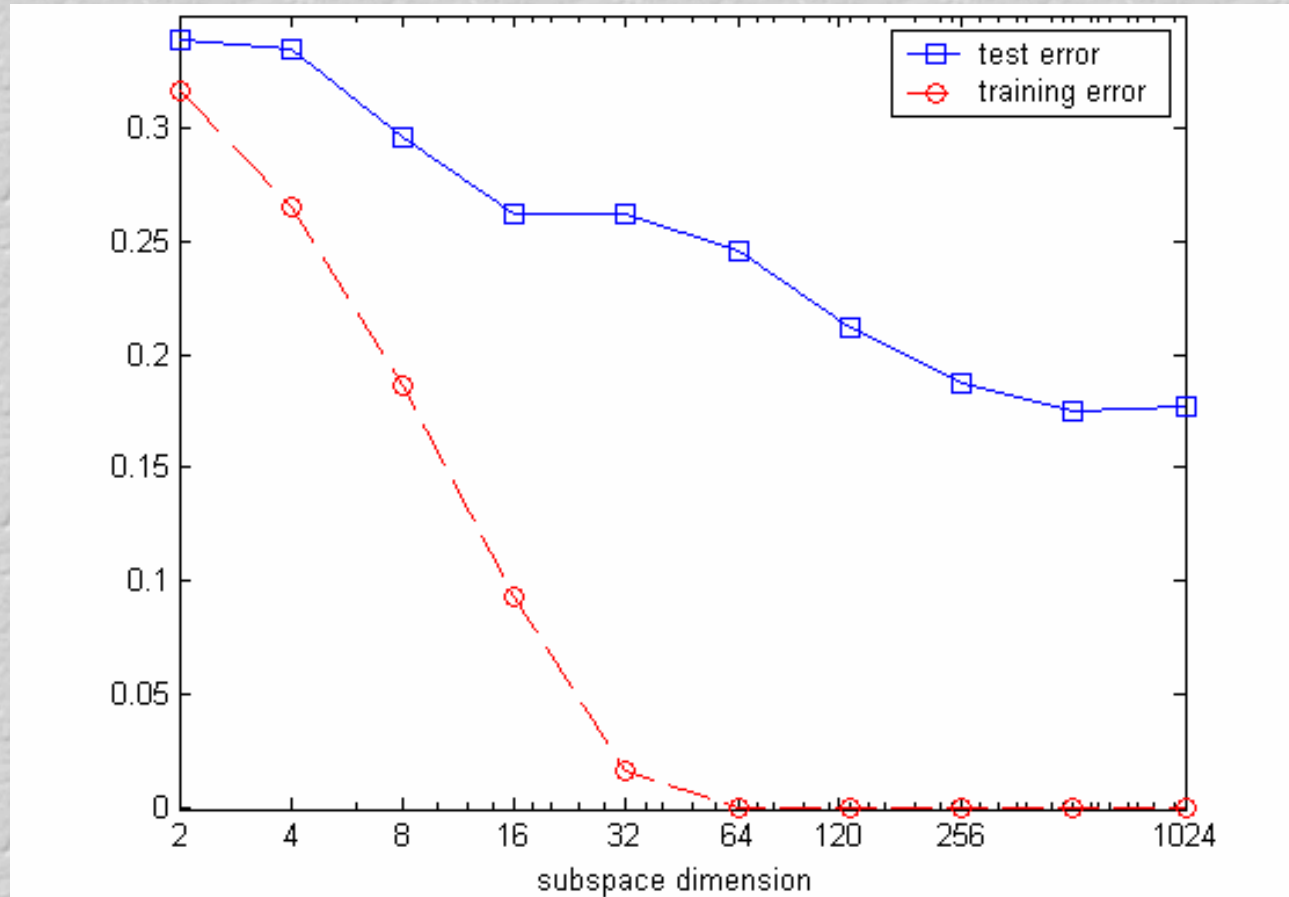
Colon tumor prediction: error as a function of the subspace dimension



Colon tumor prediction : sensitivity/specificity as a function of the subspace dimension



Average base learner error



The better accuracy of the RS ensemble does not simply depend on the better accuracy of their component base learners

Results summary

- Statistical significant difference in favour of RS ensembles vs. single SVMs
- RS ensembles are better than single SVMs for a large choice of subspace dimensions
- No learning if a too small subspace dimension is selected (because of the low accuracy of the corresponding base learners)
- The results cannot be explained only through the accuracy of the base learners

What about the reasons of the effectiveness of the random subspace approach?

Effectiveness of the RS method and other open problems

1. Can we explain the effectiveness of RS through the diversity of the base learners ?
2. Can we get a bias-variance interpretation ?
3. Can we get quantitative relationships between dimensionality reduction, redundant features and correlation of gene expression levels?
4. What about the “optimal” subspace dimension?
5. *Are feature selection and random subspace ensemble approaches alternative, or it may be useful to combine them?*

Combining feature selection and random subspace ensemble methods

*Random Subspace on Selected Features (RS-SF
algorithm)*

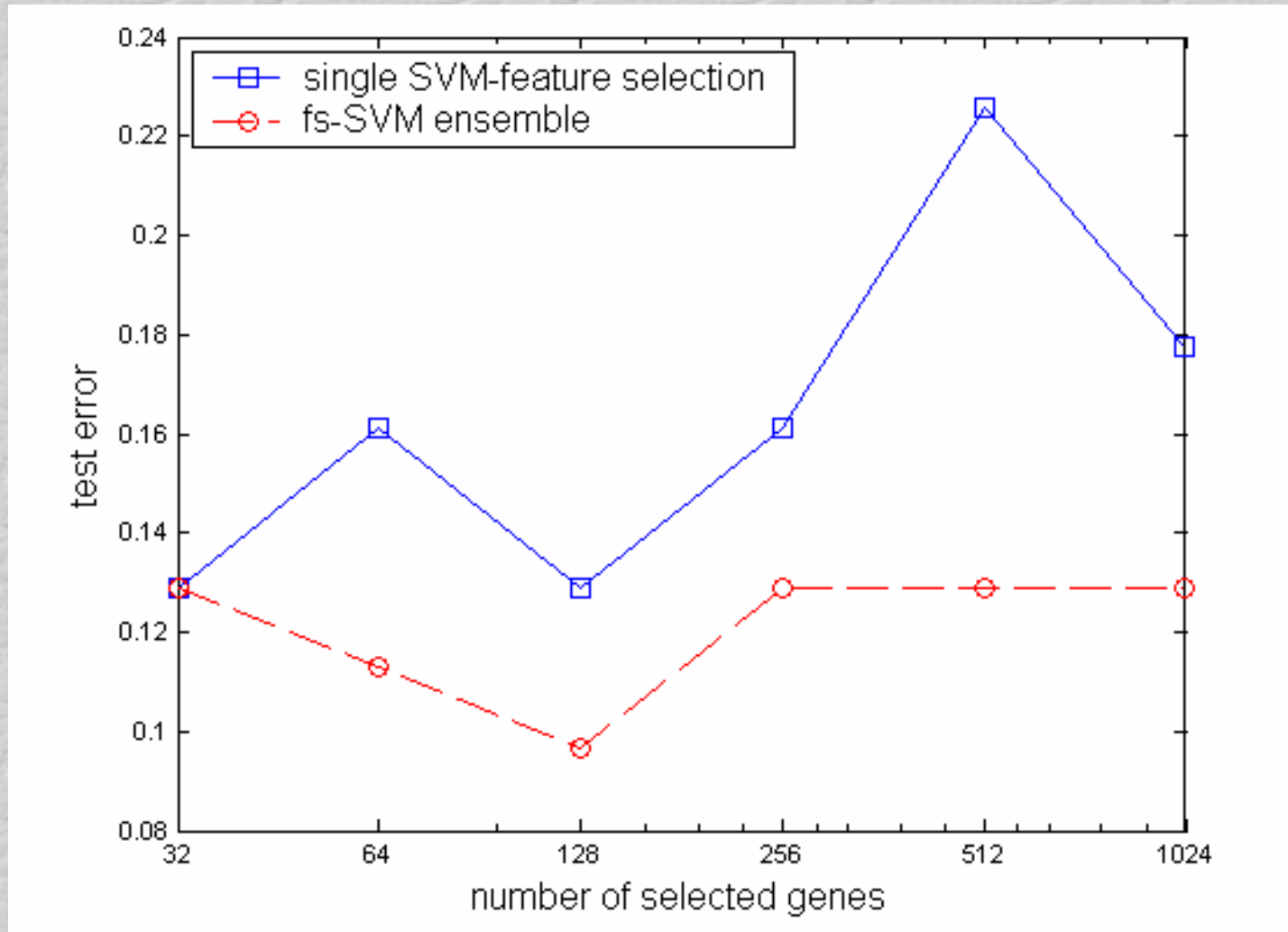
A two-steps algorithm:

1. Select a subset of features (genes) according to a suitable feature selection method
2. Apply the random subspace ensemble method to the subset of selected features

Preliminary results on combining feature selection with random subspace ensembles - 1

	Test	St.dev	Train	St.dev	Sens.	Spec.	Prec.
RS-SF ensemble	0.0968	0.0697	0.0727	0.0183	0.9250	0.8636	0.9250
RS ensemble	0.1290	0.0950	0.0000	0.0000	0.9000	0.8182	0.9000
Single FS-SVM	0.1129	0.0950	0.0768	0.0231	0.9250	0.8182	0.9024
Single SVM	0.1774	0.1087	0.0000	0.0000	0.8500	0.7727	0.8718

Preliminary results on combining feature selection with random subspace ensembles - 2



Conclusions

- RS ensembles can improve the accuracy of bio-molecular diagnosis characterized by very high dimensional data
- Several problems about the reasons of the effectiveness of the proposed approach remain open
- A new promising approach consists in combining feature (gene) selection and RS ensembles