CAPI 2006 Convegno Calcolo ad Alte Prestazioni "Biocomputing"

Bio-molecular diagnosis through Random Subspace Ensembles of Learning Machines

Alberto Bertoni, Raffaella Folgieri, Giorgio Valentini

DSI – Dipartimento di Scienze dell'Informazione

Università degli Studi di Milano

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

http://homes.dsi.unimi.it/~valenti

Outline

- Bio-molecular diagnosis of tumors using machine learning methods
- Current approaches to automatic bio-molecular diagnosis
- Random Subspace (RS) ensemble: experimental results on a case study
- Combining feature selection and RS ensemble
- On-going work: RP-ensembles

Bio-molecular diagnosis of malignancies: motivations

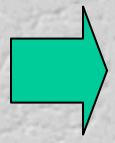
- Traditional clinical diagnostic approaches may sometimes fail in detecting tumors (Alizadeh et al. 2001)
- Several results showed that bio-molecular analysis of malignancies may help to better characterize malignancies (e.g. gene expression profiling)
- Information for supporting both diagnosis and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

Bio-molecular diagnosis of malignancies: current approaches

- Huge amount of data available from bio-technologies: 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; *Dudoit* et al. 2002; *Lee* & *Lee*, 2003; *Weston* et al., 2003, *Dettling* et al., 2003, *Dettling* 2004, *Zhou* et al, 2005, *Zhang* et al., 2006).

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

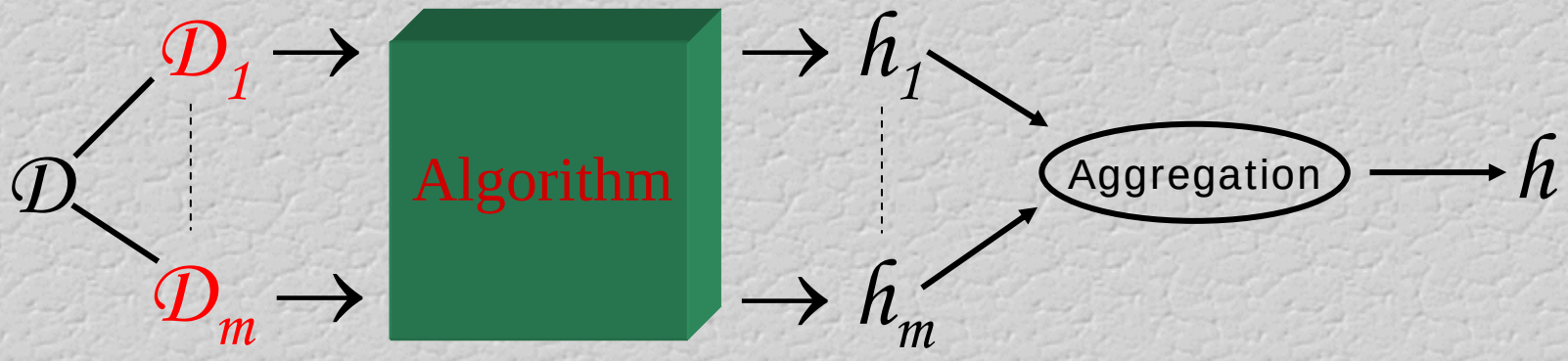
- *Selection of significant subsets of components* (genes)
  e.g.: filter methods, forward selection, backward selection, recursive feature elimination, entropy and mutual information based feature selection methods (see *Guyon & Ellisseef*, 2003 for a recent review).
- *Extraction of significant subsets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

- a learning algorithm *L*
- subspace dimension n<d
- number of the base learners *I*

Output:

- Final hypothesis h_{ran}:X→C computed by the ensemble

begin
  for i = 1 to *I*
  begin
    D_i = Subspace_projection(D,n)
    H_i = *L*(D_i)
  end
  $h_{ran}(x)=\text{argmax}_{t\in C}\text{card}(\{i|h_i(x)=t\})$
end

Reasons for applying 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 (*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)

Colon adenocarcinoma diagnosis

Data (*Alon et al*., 1999):

- 62 samples
- 40 colon tumors
- 22 normal colon samples
- 2000 genes

Methods:

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

Software: C++ NEURObjects library

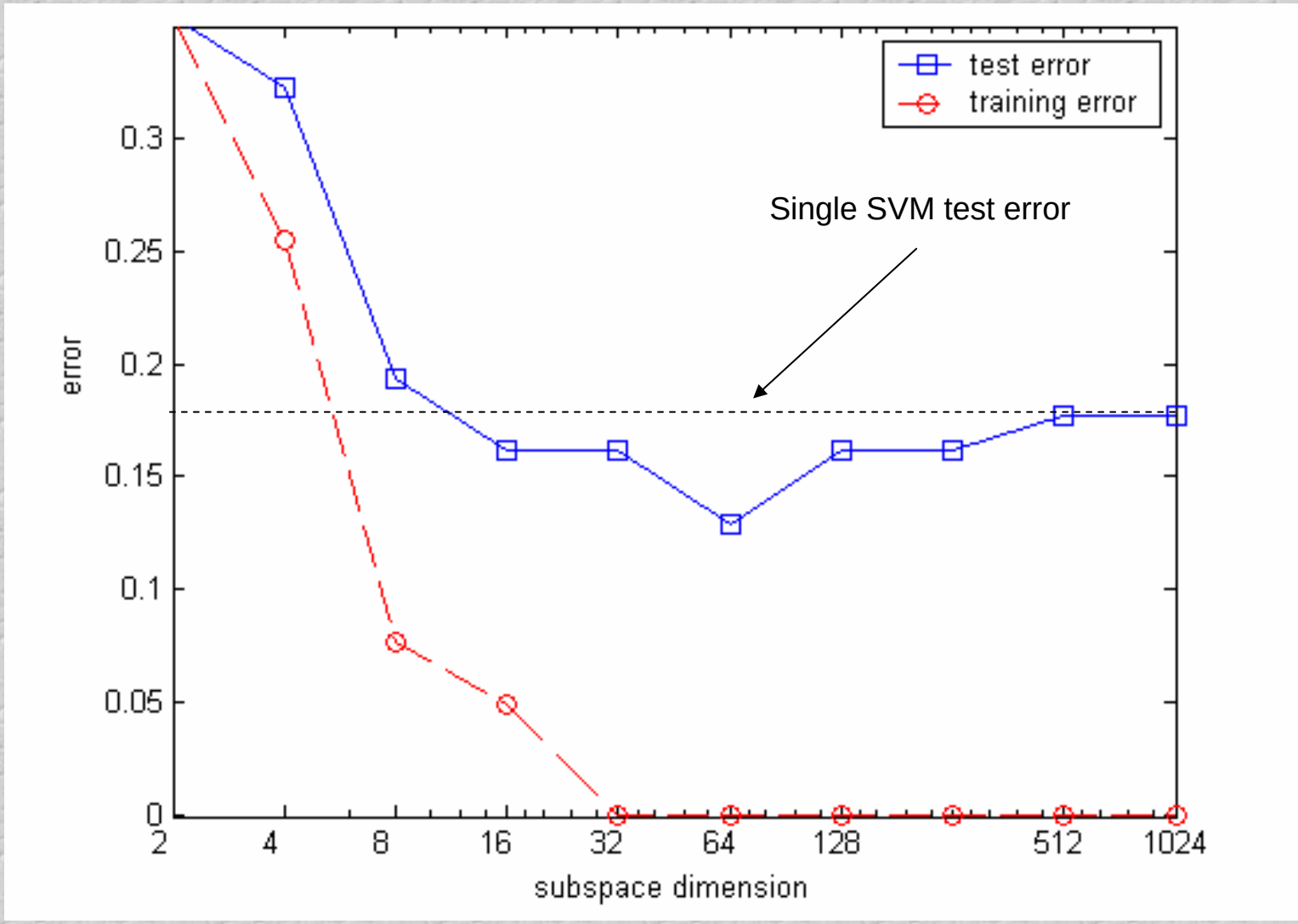
Hardware: Avogadro cluster of Xeon double processor workstations (*Arlandini*, 2005)

Results

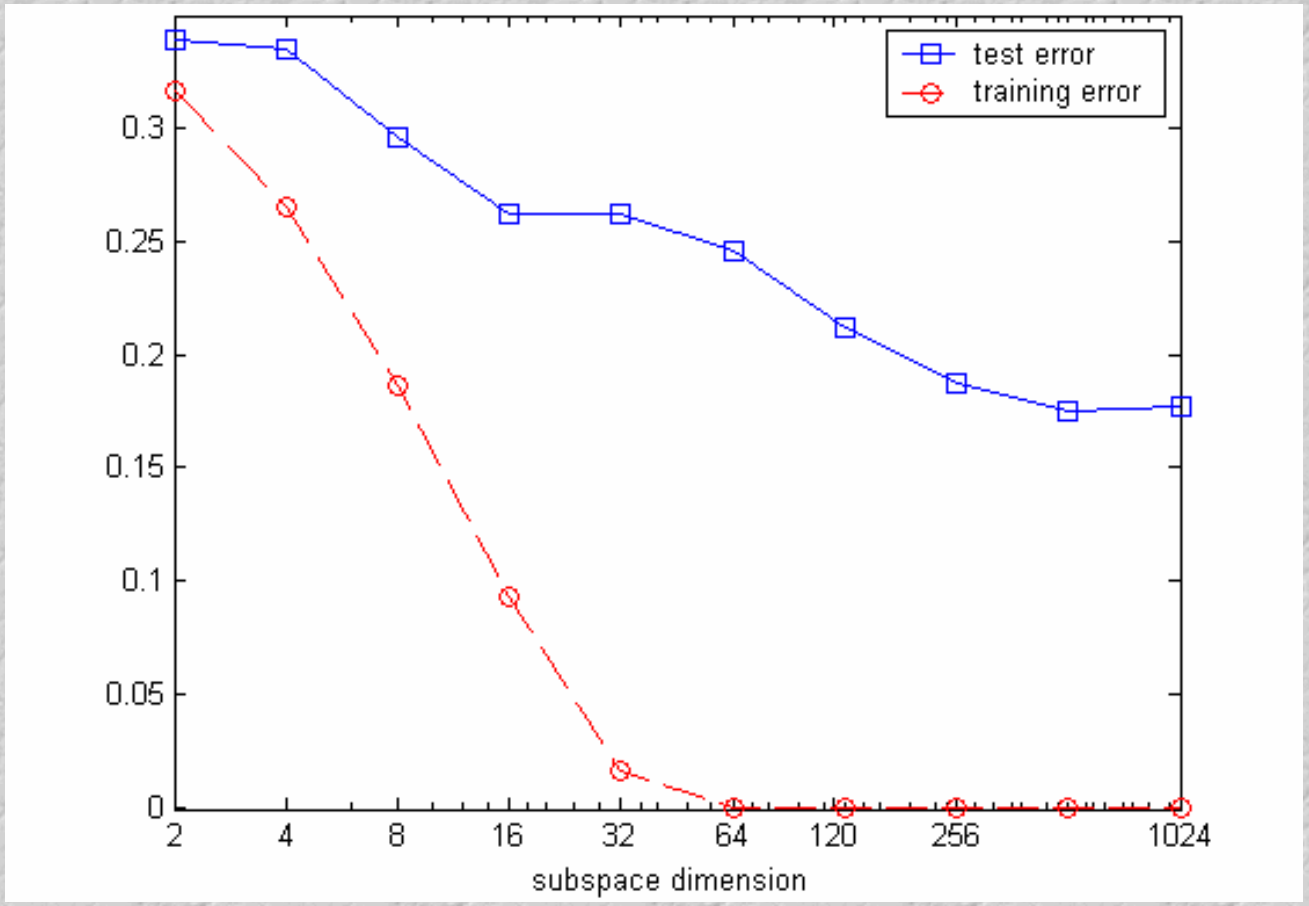
Colon tumor prediction (5 fold cross validation)

	Test Err.	St. dev.	Train.Err.	St. dev.	Sens.	Spec.	Prec.
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	---	---	---

Colon tumor prediction: error as a function of the susbspace 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

- Open problems with RS methods

1. Can we explain the effectiveness of RS through the diversity of the base learners ?
2. Can we get a bias-variance interpretation ?
3. What about the “optimal” subspace dimension?

4. ***Are feature selection and random subspace ensemble approaches alternative, or it may be useful to combine them?***

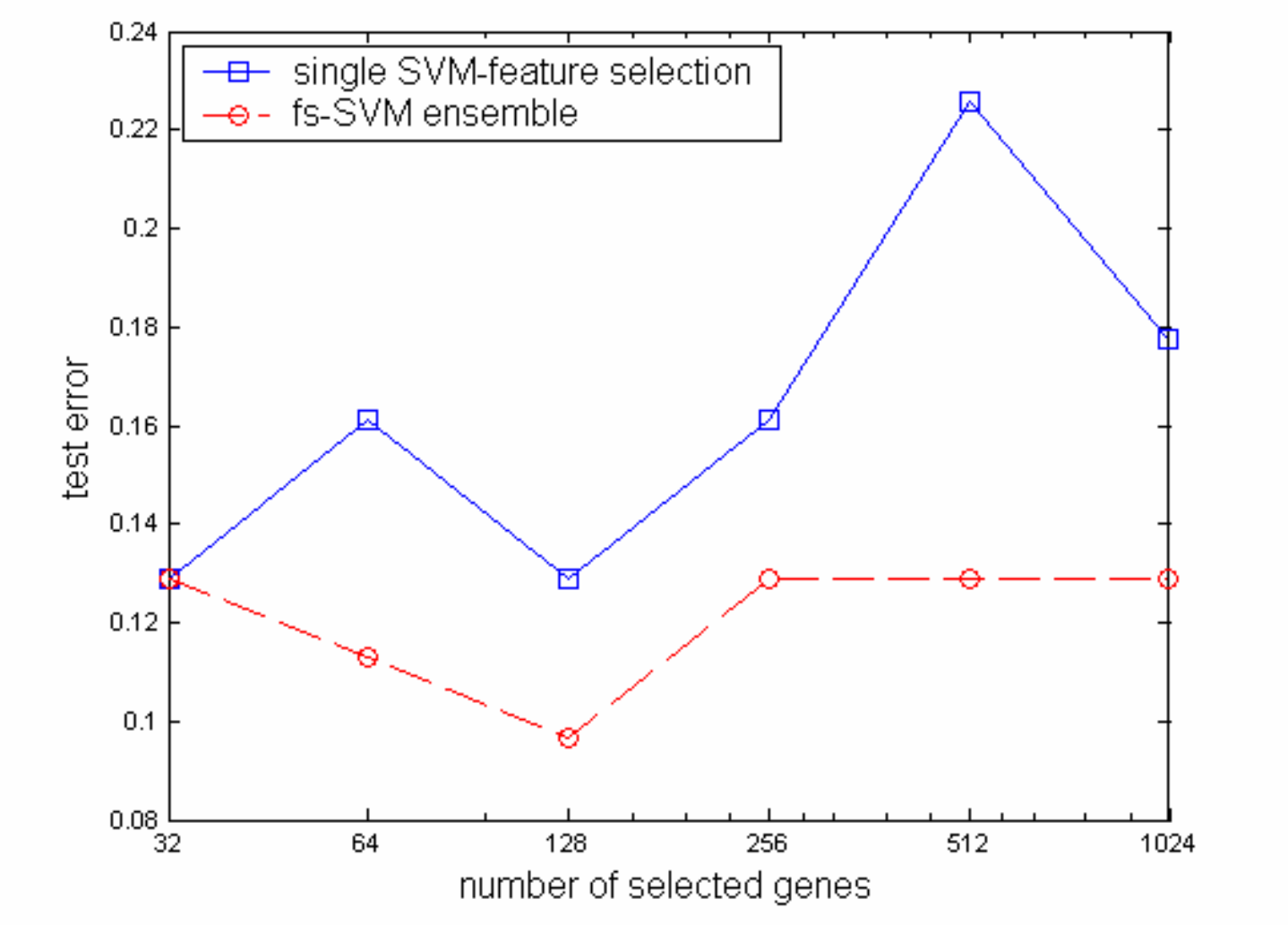
Combining feature selection and random subspace ensemble methods

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

Results on combining feature selection with random subspace ensembles



Colon data set (Alon, 1999) 5-fold cross validation

Comparison with other methods

Methods	**Estimated error**
LogitBoost (*Dettling and Buhlmann, 2003*)	0.1914
Bagging (*Valentini et al., 2004*)	0.1286
BagBoost (*Dettling, 2004*)	0.1610
Random Forest (*Breiman, 2001*)	0.1486
Random Subspace	0.0968
SVM	0.1129
PAM (*Tibshirani et al. 2002*)	0.1190
DLDA (*Dudoit et al. 2002*)	0.1286
kNN	0.1638

Colon data set: generalization error estimated through cross-validation or multiple-hold out techniques

An on-going development: Supervised Randomly Projected Ensembles (RP-ensembles):

- Recent work on *unsupervised analysis* of complex bio-molecular data (*Bertoni and Valentini, 2006*) showed that random projections obeying the *Johnson-Lindenstrauss lemma* can be used for:
  - Discovering structures in bio-molecular data
  - Validating clustering results
  - Improving clustering results
- Random projections to lower dimensional subspaces can be applied to *supervised analysis* (e.g. bio-molecular diagnosis) ?

Conclusions

- RS ensembles can *improve the accuracy* of bio-molecular diagnosis characterized by very high dimensional data
- They could be also easily applied to *heterogeneous* bio-molecular and clinical data.
- A new promising approach consists in *combining* state of the art feature (gene) selection methods and RS ensembles
- RS ensembles are computationally intensive but can be easily parallelized using *clusters of workstations* (e.g. in a MPI framework).