Random subspace ensembles for the bio-molecular diagnosis of tumors

Alberto Bertoni, Raffaella Folgieri, and Giorgio Valentini

DSI, Dipartimento di Scienze dell' Informazione,
Università degli Studi di Milano,
Via Comelico 39, 20135 Milano, Italia.
{bertoni,valentini}@dsi.unimi.it, folgieri@dico.unimi.it.

Abstract. The bio-molecular diagnosis of malignancies, based on DNA microarray biotechnologies, is a difficult learning task, because of the high dimensionality and low cardinality of the data. Many supervised learning techniques, among them support vector machines (SVMs), have been experimented, using also feature selection methods to reduce the dimensionality of the data. In this paper we investigate an alternative approach based on random subspace ensemble methods. The high dimensionality of the data is reduced by randomly sampling subsets of features (gene expression levels), and accuracy is improved by aggregating the resulting base classifiers. Our experiments, in the area of the diagnosis of malignancies at bio-molecular level, show the effectiveness of the proposed approach.

1 Introduction

Traditional diagnosis of malignancies based on histopathological and clinical parameters can fail in detecting tumours that differentiate from normal tissues at bio-molecular level [1]. Similarly, traditional approaches to prognosis determination (e.g. staging, or marker based prognosis) can be sometimes ineffective when the clinical outcome depends on multiple bio-molecular factors [2].

Information for supporting both diagnosis and prognosis of malignancies at bio-molecular level can be obtained by means of high throughput bio-technologies based on large scale hybridization techniques (e.g. DNA microarray) [3].

Because of the large amount of gene expression data produced from DNA microarray experiments, the problem of analysing and extracting significant knowledge from the data becomes critical.

Several supervised learning methods have been applied to the analysis of cDNA microarrays and high density oligonucleotide chips [4, 5]. In particular, Support Vector Machines (SVMs) have been applied to the analysis of DNA microarray gene expression data in order to classify functional groups of genes, normal and malignant tissues and multiple tumor types [5, 6].

One of the main drawbacks that characterize gene expression data is represented by their very high dimensionality and low cardinality. Indeed is well known that in these cases the *curse of dimensionality* problem arises [7]. Hence

several works pointed out the importance of feature selection methods to reduce the dimensionality of the input space [8].

From a computational point of view, feature selection is a difficult problem. In fact it is a NP-hard problem [9]. In this paper we propose to experiment a different approach, based on Random Subspace ensembles of learning machines [10]. The prediction accuracy of the "weak" base learners arising from the random selection of the genes is then enhanced through aggregation by majority voting of the trained base classifiers.

The characteristics of the methods are summarized in the next section, discussing also the reasons why this approach should be effective with gene expression data. Sect. 3 shows the results of the application of random subspace ensembles of SVMs to a diagnostic problem (*colon adenocarcinoma*), and a prognostic problem (*medulloblastoma* clinical outcome prediction) using high dimensional gene expression data. After discussing the experimental results, conclusions and on-going developments of the present work ends the paper.

2 Random subspace ensembles for gene expression data analysis

Ensemble methods combine different classifiers in order to build a more accurate classifier. Several ensemble methods, such as bagging and boosting, have been proposed, and it has been shown that they enhance accuracy of learning machines [11]. In particular, these methods and others, such as Error Correcting Output Coding ensembles, have been applied to the analysis of DNA microarray data [4, 12].

DNA microarray data are usually characterized by a small *number* of vectors of *high dimension*: high dimensionality and low cardinality of data arise the so called *curse of dimensionality* problem. A possible approach to reduce the dimension consists in considering feature selection methods [9], [10].

In this paper we experiment an alternative approach based on Random Subspace ensemble [23]. This method avoids some computational difficulty of feature selection (feature selection is an NP-hard problem) [9] and it can be implemented in parallel in a natural way. Moreover with different random projections of the data we can obtain more base learners, so that their diversity can be improved, while the overall accuracy of the ensemble can be enhanced through aggregation techniques. A high-level pseudo-code of the random subspace ensemble method is the following:

Random Subspace Algorithm

`Input`:

- A data set $\mathcal{D} = \{(\mathbf{x}_j, t_j) | 1 \leq j \leq m\}$, $\mathbf{x}_j \in \mathcal{X} \subset \mathbb{R}^d$, $t_j \in \mathcal{C} = \{1, \ldots, k\}$
- a *weak* learning algorithm $\mathcal{L}$
- subspace dimension $n < d$
- number of the base learners I

`Output`:

- Final hypothesis $h_{ran} : \mathcal{X} \to \mathcal{C}$ computed by the ensemble.

`begin`

for $i = 1$ to I

begin

$D_i =$ `Subspace_projection`($\mathcal{D}$, n)

$h_i = \mathcal{L}(D_i)$

end

$h_{ran}(\mathbf{x}) = \arg\max_{t \in \mathcal{C}} card(\{i | h_i(\mathbf{x}) = t\})$

`end`.

$\mathcal{D}$ represents the original d-dimensional training set. The randomized procedure `Subspace_projection` randomly selects, according to the uniform distribution, a n-subset $A = \{\alpha_1, \ldots, \alpha_n\}$ from $\{1, 2, \ldots, d\}$, so defining a projection $P_A : \mathbb{R}^d \to \mathbb{R}^n$, where $P_A(x_1, \ldots, x_d) = (x_{\alpha_1}, \ldots, x_{\alpha_n})$; then it returns as output the new data set $D_i = \{(P_A(\mathbf{x}_j), t_j) | 1 \leq j \leq m\}$.

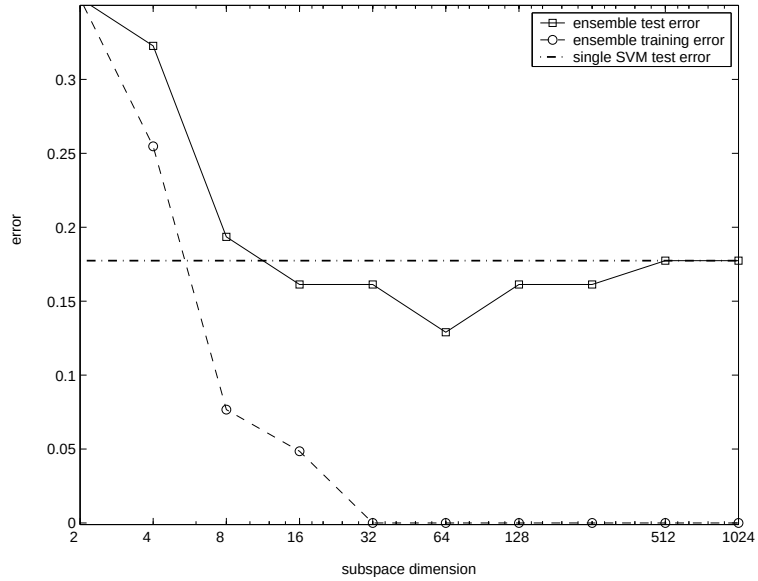
The new data set D_i obtained through the procedure `Subspace_projection` is given as input to a learning algorithm $\mathcal{L}$ which outputs a classifier h_i; this process is repeated I times, where I is the number of the base learners. In the testing phase all the trained classifiers are finally aggregated through majority voting, that is the output of the ensemble is the label of the class more frequently chosen by the component classifiers ($card(S)$ denotes the cardinality of the set S). Note that, with abuse of notation, with $h_i(\mathbf{x})$ we ambiguously denote the extension of h_i to the entire $\mathbb{R}^d$ space.

This approach has been successfully applied to different real problems [13, 14], but it seems well-suited for the diagnosis of polygenic and tumoral diseases using very high dimensional gene expression data.

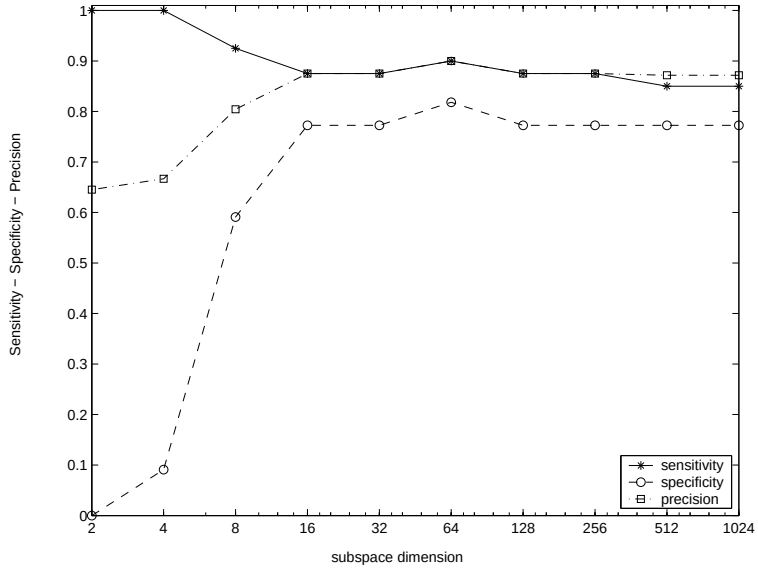
3 Experimental environment

We have experimented the previous algorithm on 2 bio-medical problems: 1) Colon adenocarcinoma bio-molecular diagnosis [15] 2) Central nervous system embryonal tumour clinical outcome prediction [2]. The two prediction problems are both based on gene expression profiles of a relatively small group of patients.

We specialized the learning algorithm $\mathcal{L}$ using linear Support Vector Machines (SVMs). In fact, SVMs are usually considered very good predictors for the bio-molecular prediction of tumours. Moreover random subspace ensembles seems to give good results with linear base learners characterized by a decreasing learning curve (error) with respect to the cardinality n [16], and linear SVMs show these characteristics. Furthermore, we fixed 200 as the number I



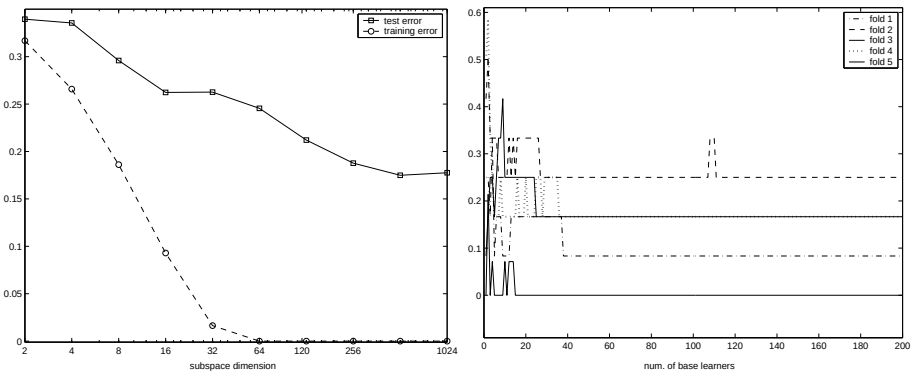
(a)



(b)

Fig. 1. SVM random subspace ensembles results on the colon data set (5-fold cross validation). (a) Test and training error with respect to the dimension of the subspace. (b) Sensitivity, specificity and precision with respect to the dimension of the subspace.

of base learners and chose as dimension of subspace every number $n = 2^k$ with



(a) (b)

Fig. 2. Colon data set: (a) Average training and test error of the base learners (component predictors) with respect to the subspace dimension (b) Test error of the 64 dimensional SVM random subspace ensemble with respect to the number of the base learners on the 5 folds.

$1 \leq k < \lceil \log_2 d \rceil$. More precisely, we drew 200 random subspaces from the available $\binom{d}{n}$ ones, and we used them to project the original $d-$dimensional input data into the obtained 200 $n-$dimensional subspaces; the resulting samples have been used to train the 200 base SVMs that belong to the ensemble.

Concerning the implementation, we developed new C++ classes and applications for random subspace ensembles extending the *NEURObjects* [1] library [17]. The experiments have been executed by means of the C.I.L.E.A. Avogadro cluster of Xeon double processor workstations [18].

The main goal that we have pursued in the experiments is the performance comparison of single SVMs trained with all the available genes vs. random subspace ensembles. In order to evaluate and to understand the ensemble behaviour, we analyzed also the accuracy of the base learners, that is the performances of the single base SVMs trained with random subsets of features (genes).

We computed for both single SVMs and random subspace ensembles the test error and the training error, by 5-fold cross validation. Moreover we considered sensitivity, specificity and precision values. Only for the ensembles we also evaluated the error as a function of the number of the base learners on each fold.

3.1 Colon tumor prediction

The Colon adenocarcinoma data set is composed of 2000 genes and 62 samples: 40 colon tumor samples and 22 normal colon tissue samples [15].

[1] The extended new version of the *NEURObjects* library is freely downloadable for research or teaching purposes from `http://homes.dsi.unimi.it/~valenti/sw/NEURObjects/`.

Experimental setup We used the same preprocessing technique illustrated in [15]. Concerning model selection, the values of the regularization parameter C of the SVMs have been selected in the range between 0.01 and 1000. Moreover the dimension k of the subspaces is each power of 2 in the range between 2 and 2^{10}, while the number of base learners used is 200.

Results Single SVMs trained using the entire set of gene expression data achieved an error of 17.74 ± 10.87 % according to a 5-fold cross validation evaluation of the generalization error. As outlined in other works [8], on this task the linear SVMs are insensitive to the regularization C parameter that controls the trade-off between the accuracy on the training set and the complexity of the learning machine: for all the C values we obtained the same results.

On the contrary Random subspace ensembles on this task are quite sensitive to the regularization parameter: for instance 16-dimensional random subspace ensembles achieve better results with quite large C values. The minimum of the test error is obtained using 64-dimensional subspaces, but also with 16 to 1024-dimensional subspaces equal or better results with respect to single SVMs trained on the entire feature space can be achieved (Fig 1 a).

Interestingly enough, sensitivity is very high if very low dimensional subspaces are applied, but at the expenses of the specificity (Fig 1 b). Indeed using 2 or 4-dimensional subspaces the base SVMs learn nothing, predicting that all samples are malignant, without any distinction between normal and cancerous tissues. The ensembles start to learn when 8 random genes are selected, and if we apply at least 16 gene-subspaces we achieve a reasonable specificity at the expense of a low decrement of the sensitivity (Fig 1 b).

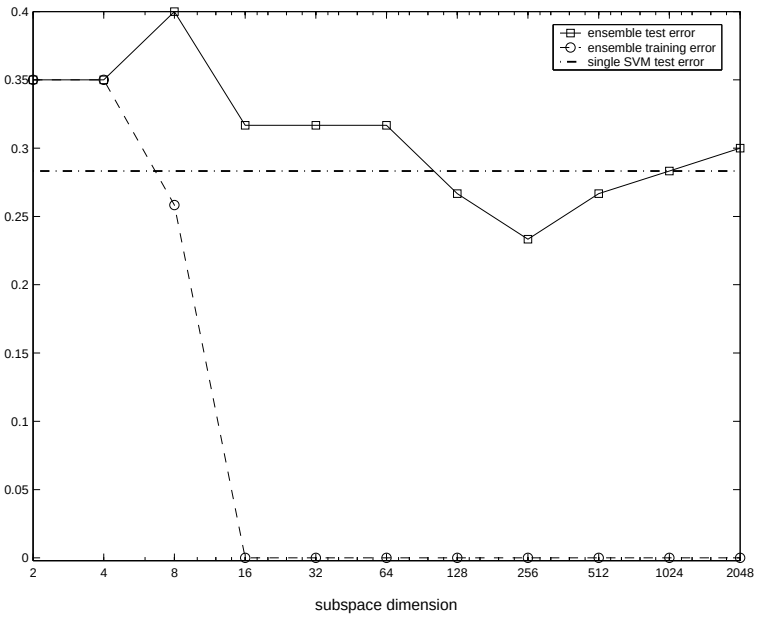
Fig. 2 (a) shows that both the base learner training and test error decrease monotonically with the subspace dimension. Hence the best performance with 64-dimensional random subspace ensembles cannot be the effect of a better accuracy of the base learners trained with 64 random genes.

We trained 200 SVMs for each ensemble, but Fig. 2 (b) shows that with about 50 learners we can achieve the same results. Indeed the test error on the 5 folds decreases up to 50 base learners, and for larger ensembles the test error stabilizes and no variations are registered. Note that the spikes in the five curves are due to the low cardinality of the data set.

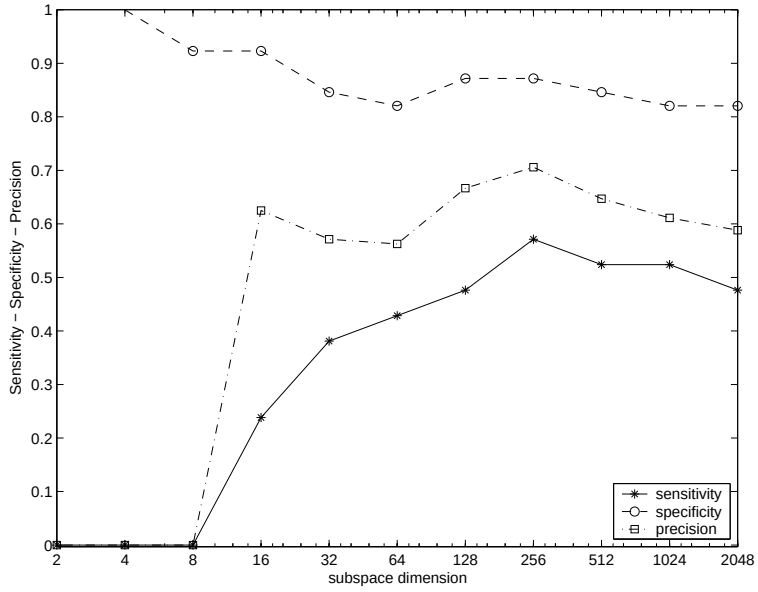
3.2 Medulloblastoma clinical outcome prediction

The data set considered doesn't treat a diagnosis problem, but a prediction of medulloblastoma clinical outcome by gene expression profiling. The *Medulloblastoma* data set is composed of 60 samples, with 39 medulloblastoma survivors and 21 treatment failures. We used about 4000 genes, obtained by the same preprocessing techniques adopted in the original work [2].

Experimental setup The gene expression data were subjected to a threshold and a variation filter that excludes genes showing minimal variation across



(a)

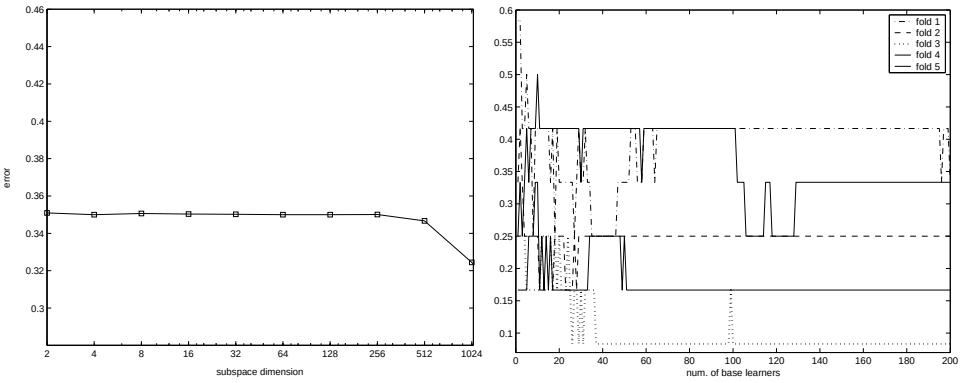


(b)

Fig. 3. SVM random subspace ensembles results on the medulloblastoma data set (5-fold cross validation). (a) Test and training error with respect to the dimension of the subspace (b) Sensitivity, specificity and precision with respect to the dimension of the subspace.

the samples being analysed. By preprocessing we obtained 4433 genes from the original 7129.

Regarding the model selection, we selected the C values in the range between 10^{-9} and 10^{3}. The dimension k of the subspaces is each power of 2 in the range between 2 and 2^{11}, and the number of the base learners used is 200.



(a) (b)

Fig. 4. Medulloblastoma data set: (a) Average test error of the base learners (component predictors) with respect to the subspace dimension (b) Test error of the 256 dimensional SVM random subspace ensemble with respect to the number of the base learners on the 5 folds.

Results With this data set single SVMs trained using the entire set of gene expression data are quite sensitive to the C parameter. In particular we need to choose very low values of the regularization parameter in order to achieve an error of 28.33 ± 9.50 % according to a 5-fold cross validation evaluation of the generalization error.

Similarly to the colon data set, also with the medulloblastoma clinical outcome prediction problem random subspace ensembles outperform single SVMs trained on the entire set of the gene expression data. The minimum of the test error is registered with 256-dimensional subspaces, but in this case we need from 128 to 512-dimensional random subsets of genes to achieve better results than single SVMs (Fig 3 a).

In all cases we obtained low sensitivity (slightly better for subspaces between 128 and 512 dimensions), and large specificity for a large range of randomly selected genes. The ensembles trained with relatively small subspaces (less than 16 random genes) show large specificity and zero sensitivity, as low dimensional ensembles fail to detect failures (that is those who succumbed to their disease) while successfully detect all the survivors (Fig 3 b). On medulloblastoma clinical outcome prediction Pomeroy et al [2] (see supplementary documentation)

successfully applied SVMs and feature selection methods, but our ensemble approach achieves better accuracy and sensitivity [2].

As with the colon data set, also in this case the better results obtained with 256 dimensional subspaces cannot be explained with a better accuracy of the base learners trained with 256 random genes. Indeed Fig. 4 (a) shows that base learner training and test error decrease with the subspace dimension.

Differently from the previous classification problem, in medulloblastoma experiments we need more base learners to stabilize the error (Fig. 4 (b)).

3.3 Discussion

Table 1. Comparison of the results between single and random subspace ensembles of SVMs.

	Test Err.	**St.dev**	**Train Err.**	**St.dev**	**Sens.**	**Spec.**	**Prec.**
Colon tumour data set							
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 data set							
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	——	——	——

Tab. 1 compares the results of the best single SVM and the best random subspace ensemble on the *Colon* and *Medulloblastoma* data sets. Average errors of the best base learners of the ensemble are also reported.

The most significant result is that Random subspace ensembles outperform single SVMs on both the considered classification tasks. The null hypothesis that the random subspace ensemble has the same error rate as single SVMs is rejected at 0.05 significance level according to the 5-fold cross validated paired t-test [19] for both the *Colon* and *Medulloblastoma* data sets.

Moreover we achieve better results with random subspace ensembles for a quite large choice of the subspace dimension (Fig. 1 and 3). Only if too small subspaces are used, we cannot obtain good results, because the base learners are not able to learn when the data are too uninformative (Fig. 2 (a) and 4 (a)).

The best average accuracy of the base learners, comparable with the accuracy of the single SVM trained with the entire set of features (genes), is achieved with 1024 dimensions with the colon data set and 2048 dimensions with the medulloblastoma data set. In both cases there is no statistical significant difference

[2] Anyway, note that we used 5-fold cross validation to estimate the generalization error, while Pomeroy et al. applied leave-one-out techniques.

between the average accuracy of the base learners and the accuracy of the SVMs trained with all the available gene expression data (Tab. 1). As outlined in other works [1, 12], this fact highlights that the information carried out by many genes is highly correlated, and no discrimination gain is achieved when we double the number of genes both for the colon and medulloblastoma data sets. On the other hand these results can also be explained by the fact that many genes are not correlated with the discrimination of the functional classes.

Anyway the significant performance differences between random subspace and single SVMs cannot be only explained through the accuracy of the base learners, as the best ensemble performance are obtained with 64 (Colon data set: fig. 1 a) and 256-dimensional subspaces (Medulloblastoma data set: Fig. 3 a), whilst the best base learner accuracy is achieved respectively with 1024 and 2048-dimensional subspaces. Hence we need a deeper understanding of the ensemble behaviour to explain the better results of random subspaces.

To this purpose we propose a list of possible research directions to explain the results of the random subspace ensembles with high-dimensional gene expression data:

1. *Diversity of the base learners.* The trade-off problem between accuracy and diversity of the base learners and their relationships with the accuracy of the corresponding ensembles is a critical topic in ensemble methods research [20]. Our results suggest that the ensembles could be more accurate when the component base learners are diverse, that is when their predictions are not equal; on the other hand we need also accuracy, because very inaccurate base learners cannot make the ensemble accurate. Anyway, to confirm this hypothesis. we need to quantitatively evaluate the diversity between the component base learners using, for instance, diversity measures proposed in [20].

2. *Bias-variance interpretation.* It is well-known that ensemble methods based on aggregation and majority voting reduce the variance component of the error (e.g. bagging [21]). Random subspace ensembles are based on aggregation and majority voting, so we could expect that the ensemble decrements the variance with respect to the single learner. Anyway it is unclear if the bias component is reduced by this approach. To this purpose it could be interesting to interpret the generalization capabilities of random subspace ensembles using bias–variance analysis, as proposed in [22].

3. *Dimensionality reduction.* Random subspace ensembles reduce the *curse of dimensionality* effect that plagues hyperdimensional spaces [7]. Hence, gene expression data, characterized by very high dimensionality, can benefit from the random subspace approach. Indeed random subspace ensembles increment the effective cardinality, that is the ratio between cardinality and dimensionality [16].

4. *Redundant sets of features.* Random subspace methods can obtain a decrement of the generalization error when the classification problems involves a redundant set of features, and if we can dispose of "weak" learners (that is classifiers that perform better than "random guessing") [10, 23]. This is exactly the case of gene expression data classification for diagnosis or clinical outcome prediction purposes. Indeed it is well-known that gene expression levels of subsets

of genes are correlated, because those subsets are co-regulated through the same sets of transcriptional factors [24].

4 Conclusions and developments

The results of the application of the random subspace ensembles of linear SVMs have revealed the effectiveness of the approach with colon and medulloblastoma data sets. Better results than single SVMs trained with the global available feature (gene) space are achieved for a quite large range of subspace dimensions.

Anyway, especially for the Medulloblastoma clinical outcome prediction task we obtained a relatively high estimated generalization error and a too low sensitivity, even if comparable with the modified K-nearest-neighbour plus gene selection approach proposed in [2]. For these reasons we are experimenting a new approach that combines the random subspace ensemble approach with feature selection ones in order to improve the generalization capabilities of the ensemble.

In the discussion (Sect. 3.3) we showed some research directions to investigate the reasons why the proposed approach works with gene expression data. Nevertheless, we need a deeper understanding of the success of random subspace ensembles with high dimensional data. To this purpose an interesting theoretical open problem consists in explicitly relating the random subspace method with Kleinberg's Stochastic Discrimination theory [25].

Acknowledgments

We thank the C.I.L.E.A. for providing *Avogadro* [18], the Linux computer cluster used in our experiments. We would like also to thank the anonymous reviewers for their comments and suggestions.

References

[1] Alizadeh, A. et al.: Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature **403** (2000) 503–511
[2] Pomeroy, S. et al.: Gene Expression-Based Classification and Outcome Prediction of Central Nervous System Embryonal Tumors. Nature **415** (2002) 436–442
[3] Alizadeh, A. et al.: Towards a novel classification of human malignancies based on gene expression. J. Pathol. **195** (2001) 41–52
[4] Dudoit, S., Fridlyand, J., Speed, T.: Comparison of discrimination methods for the classification of tumors using gene expression data. JASA **97** (2002) 77–87
[5] Brown, M., et al.: Knowledge-base analysis of microarray gene expression data by using Support Vector Machines. PNAS **97** (2000) 262–267
[6] Furey, T., Cristianini, N., Duffy, N., Bednarski, D., Schummer, M., Haussler, D.: Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics **16** (2000) 906–914
[7] Bellman, R.: Adaptive Control Processes: a Guided Tour. Princeton University Press, New Jersey (1961)

[8] Guyon, I., Weston, J., Barnhill, S., Vapnik, V.: Gene Selection for Cancer Classification using Support Vector Machines. Machine Learning **46** (2002) 389–422
[9] Amaldi, E., Kann, V.: On the approximation of minimizing non zero variables or unsatisfied relations in linear systems. Theoretical Computer Science **209** (1998) 237–260
[10] Ho, T.: The random subspace method for constructing decision forests. IEEE Transactions on Pattern Analysis and Machine Intelligence **20** (1998) 832–844
[11] Dietterich, T.: Ensemble methods in machine learning. In MCS 2000: Multiple Classifier Systems. First International Workshop, Cagliari, Italy. Vol. 1857 of LNCS, Springer-Verlag (2000) 1–15
[12] Valentini, G.: Gene expression data analysis of human lymphoma using support vector machines and output coding ensembles. Artificial Intelligence in Medicine **26** (2002) 283–306
[13] Munro, R., Ler, D., Patrick, J.: Meta-learning orthographic and contextual models for language independent named entity recognition. In: CoNLL-2003, Proc. of the Seventh Conference on Natural Language Learning. (2003)
[14] Hall, L., Bowyer, K., Banfield, R., Bhadoria, D., Kegelmeyer, W., Eschrich, S.: Comparing pure parallel ensemble creation techniques against bagging. In: Third IEEE International Conference on Data Mining, Melbourne, Florida (2003)
[15] Alon, U.e.a.: Broad patterns of gene expressions revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. PNAS **96** (1999) 6745–6750
[16] Skurichina, M., Duin, R.: Bagging, boosting and the randon subspace method for linear classifiers. Pattern Analysis and Applications (2002)
[17] Valentini, G., Masulli, F.: NEURObjects: an object-oriented library for neural network development. Neurocomputing **48** (2002) 623–646
[18] Arlandini, C.: Avogadro: il CILEA oltre il muro del teraflops. Bollettino del CILEA (2004)
[19] Dietterich, T.: Approximate statistical test for comparing supervised classification learning algorithms. Neural Computation (1998) 1895–1924
[20] Kuncheva, L., Whitaker, C.: Measures of diversity in classifier ensembles. Machine Learning **51** (2003) 181–207
[21] Breiman, L.: Bagging predictors. Machine Learning **24** (1996) 123–140
[22] Valentini, G., Dietterich, T.: Bias–variance analysis of Support Vector Machines for the development of SVM-based ensemble methods. Journal of Machine Learning Research **5** (2004) 725–775
[23] Bingham, E., Mannila, H.: Random projection in dimensionality reduction: Applications to image and text data. In: Proc. of KDD 01, San Francisco, CA, USA, ACM (2001)
[24] Gasch, P., Eisen, M.: Exploring the conditional regulation of yeast gene expression through fuzzy k-means clustering. Genome Biology **3** (2002)
[25] Kleinberg, E.: On the Algorithmic Implementation of Stochastic Discrimination. IEEE Transactions on Pattern Analysis and Machine Intelligence **22** (2000) 473–490