Dataset Complexity Can Help to Generate Accurate Ensembles of K-Nearest Neighbors

Oleg Okun, Giorgio Valentini

Abstract— **Gene expression based cancer classification using classifier ensembles is the main focus of this work. A new ensemble method is proposed that combines predictions of a small number of k-nearest neighbor (k-NN) classifiers with majority vote. Diversity of predictions is guaranteed by assigning a separate feature subset, randomly sampled from the original set of features, to each classifier. Accuracy of k-NNs is ensured by the statistically confirmed dependence between dataset complexity, determining how difficult is a dataset for classification, and classification error. Experiments carried out on three gene expression datasets containing different types of cancer show that our ensemble method is superior to 1) a single best classifier in the ensemble, 2) the nearest shrunken centroids method originally proposed for gene expression data, and 3) the traditional ensemble construction scheme that does not take into account dataset complexity.**

I. Introduction

Gene expression is a two-stage process including the transcription of deoxyribonucleic acid (DNA) into messenger ribonucleic acid (mRNA) which is then translated into protein by the ribosomes. Recent advances in microarray technology facilitate measurement of gene expression levels for thousands of genes at once. It was shown in numerous works [1], [2] that expression levels provide valuable information for discrimination between normal and cancer specimens. However, the classification task is not easy since there are typically thousands of expression levels versus few dozens of cases. In addition, expression levels are noisy due to the complex procedures and technologies involved in the measurements of gene expression levels, thus causing ambiguity in classification.

As a classifier, a k-nearest neighbor (k-NN) was chosen in this work because it performed well for cancer classification, compared to more sophisticated classifiers [3]. Besides, it is a simple method that has a single parameter (the number of nearest neighbors) to be pre-defined, given that the distance metric is Euclidean.

Ensembles of classifiers can lower the uncertainty related to predictions provided by "single" learning machines [4]. However, not every ensemble can outperform its most accurate member. Faced with this open question, we propose a novel approach based on an explicit estimation of data complexity to construct an ensemble of classifiers. Its novelty comes from extensive simulation followed by statistical analysis using the copula method [5], [6], [7], which, to our best knowledge, has not yet been widely found in cancer classification studies. The copula method is useful in exploring association (dependence or concordance) relations in multivariate data. According to our hypothesis, there is positive dependence between dataset complexity and classification error, with low (high) complexity associated with small (large) error.

After confirming that this concordance relation indeed exists when complexity is measured by the Wilcoxon rank sum statistic [8] and classification performance is estimated by the bolstered resubstitution error [9], we build our ensemble generating scheme on 1) the selection of a few least complex subsets of features (genes), where each subset constitutes a small group of genes randomly sampled from the original gene set, 2) the association of each subset with one k-NN, and 3) majority vote combining predictions of individual k-NNs. Random feature selection results in diversity of predictions of individual k-NNs and it avoids the bias that is difficult to escape in the filter/wrapper models of feature selection. On the other hand, selecting the least complex feature subsets implies accurate classifiers working with these subsets. Therefore two requirements (diversity and accuracy) for the ensembles to be superior to their single best members are satisfied in our approach.

This work was inspired by [10], [11], where dataset complexity was estimated for different datasets in order to compare how it correlates with classification error of individual classifiers and classifier ensembles. However, no statistical analysis was done and association between complexity and accuracy largely remained a conjecture. Unlike [10], [11], in this paper, we assert this association relation and apply it in order to generate accurate k-NN ensembles, which constitutes the novelty of our approach.

II. Gene expression datasets

Three gene expression datasets were employed in our study.

A. Colon dataset

This microarray (oligonucleotide) dataset [12], introduced in [1], contains expressions of 2000 genes for 62 cases (22 normal and 40 colon tumor cases). Preprocessing includes the logarithmic transformation to base 10, followed by normalization to zero mean and unit variance as usually done with this dataset.

B. Brain dataset

This microarray (oligonucleotide) dataset [13] introduced in [2] is different from the others in this study because

Oleg Okun is with the Department of Electrical and Information Engineering, University of Oulu, Finland (email: oleg@ee.oulu.fi). Giorgio Valentini is with DSI, Department of Computer Science, University of Milan, Italy (email: valentini@dsi.unimi.it).

it contains two classes of brain tumor instead of cancer and normal classes. The dataset (also known as Dataset B) contains 34 medulloblastoma cases, 9 of which are desmoplastic and 25 are classic. Preprocessing consists of thresholding of gene expressions with a floor of 20 and ceiling of 16000; filtering with exclusion of genes with $\max / \min \leq 3$ or $max - min < 100$, where $\max$ and $\min$ refer to the maximum and minimum expressions of a certain gene across the 34 cases, respectively; base 10 logarithmic transformation; normalization across genes to zero mean and unit variance. As a result, 5893 out of 7129 original genes are retained.

C. SAGE dataset

This dataset [14] was produced by a technology alternative to microarrays and called SAGE (Serial Analysis of Gene Expression) [15]. It contains 31 normal and 59 cancer (10 types of cancer) cases with 27679 expressed genes. No preprocessing was done and all cancer types were assigned to a single class.

D. Dataset summary

Table I provides a summary for all datasets.

TABLE I

SUMMARY OF GENE EXPRESSION DATASETS

Dataset no.	Cancer type(s)	# expression levels	# cases
1	Colon	2000	62
2	Brain	5893	34
3	Multiple	27679	90

III. DATASET COMPLEXITY

To gain insight into a supervised classification problem such as gene expression based cancer classification, we adopt dataset complexity measures. Our work was inspired by [16] but in contrast to it, we considered only one measure of data complexity. This allows us to simplify relations between dataset complexity and classification accuracy, since in [16] it was difficult to draw decisive conclusions from combinations of several complexity measures.

The notion of complexity described below assumes two-class problems and it is classifier-independent. It has been borrowed from gene selection methods introduced in [17] and modified. Its goal is to provide a score reflecting how well two classes of data are separated. Given a set of features, the data of each class are projected onto the diagonal linear discriminant axis by using only these features (for details, see [17]). Projection coordinates then serve as input for the Wilcoxon rank sum test for equal medians [8] (the null hypothesis of this test is two medians are equal at the 5% level). Given a sample divided into two groups according to class membership, all the observations are ranked as if they were from a single sample and the rank sum statistic W is computed as the sum of the ranks in the smaller group. The value of the rank sum statistic is employed as a score characterizing separability power of a given set of features. The higher this score, the larger the overlap in projections of two classes, i.e. the worse separation between classes. To compare W coming from different datasets, each W can be normalized by the sum of all ranks, i.e. if N is the sample size, then the sum of all ranks will be $\sum_{i=1}^{N} i$.

IV. BOLSTERED RESUBSTITUTION ERROR

This is a low-variance and low-bias classification error estimation proposed in [9]. Unlike the cross-validation techniques reserving a part of the original data for testing, it permits to use the whole dataset. Since sample size of gene expression datasets is very small compared to the data dimensionality, using all available data is an important positive factor. However, one should be aware of the effect of overfitting in this case. Braga-Neto and Dougherty [9] avoided this pitfall by randomly generating a number of artificial points (cases) in the neighborhood of each training point. These artificial cases then act as a test set and classification error on this set is called bolstered. In this paper, we utilize the bolstered variant of the conventional resubstitution error known as *bolstered resubstitution error*.

Briefly, bolstered resubstitution error is estimated as follows [9]. Let A_0 and A_1 be two decision regions corresponding to the classification generated by a given algorithm, N be the number of training points, and M_{MC} be the number of random samples per training point, drawn from the D-variate normal distribution ($M_{MC} = 10$ as advised in [9]). The bolstered resubstitution error is then defined as

$$\epsilon_{bresub} \approx \frac{1}{NM_{MC}} \sum_{i=1}^{N} \left(\sum_{j=1}^{M_{MC}} I_{x_{ij} \in A_1} I_{y_i=0} + \sum_{j=1}^{M_{MC}} I_{x_{ij} \in A_0} I_{y_i=1} \right),$$

where $\{x_{ij}\}_{j=1,\ldots,M_{MC}}$ are samples drawn from $1/((2\pi)^{D/2}\sigma_i^D)e^{-\|x\|^2/(2\sigma_i^2)}$ by using the Marsaglia polar normal random number generator, and I is the indicator function.

In a D-dimensional case, samples are drawn from a hypersphere centered at a particular training point. Hence, the radius of this hypersphere, determined by σ_i, is of importance since its selection amounts to choosing the degree of bolstering. Typically, σ_i should vary from point to point in order to be robust with respect to the data. In [9] $\sigma_i = \hat{d}(y_i)/cp$ for $i = 1, \ldots, N$, where $\hat{d}(y_i)$ is the mean minimum distance between points belonging to class of y_i (y_i can be either 0 or 1)[1], and cp is the constant called the correction factor defined as the inverse of the chi-square cdf (cumulative distribution function) with parameters 0.5 and D, because interpoint distances in the Gaussian case are distributed as a chi-square random variable with D degrees of freedom. Thus, cp is the function of the data dimensionality. The parameter 0.5

[1] $\hat{d}(y_i)$ is determined by first computing the minimum distance from each point x_i to all other points x_j ($j \neq i$) of the same class as that of x_i and then by averaging thus obtained minimum distances.

is chosen so that points inside a hypersphere will be evenly sampled.

V. Dependence between dataset complexity and classification error

Our main idea to build ensembles of k-NNs is based on the hypothesis that *the dataset complexity and bolstered resubstitution error are related*. In other words, knowing the former can predict the latter, i.e. running a classifier is unnecessary if the sufficient evidence of such a relation exists.

To verify our hypothesis, 10000 feature subsets were randomly sampled for each dataset (subset size ranged from 1 to 50) and both complexity and bolstered resubstitution error for 3-NN were computed. The result of such simulation is shown in Figs. 1-3 together with marginal histograms for each variable where the dependence between complexity and error is clearly detectable. In addition, one characteristic important for successful ensemble generation is present: diversity among predictions since one complexity value corresponds to several different error values.

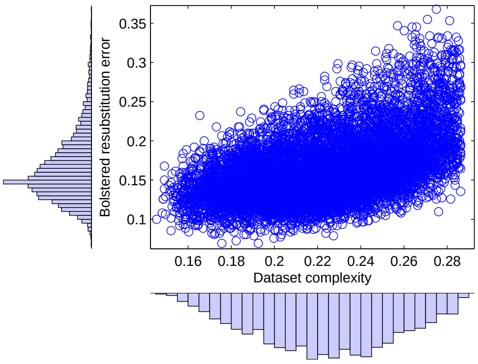

Fig. 1. (Colon) Bivariate distribution of normalised complexity and bolstered resubstitution error and univariate marginal histograms

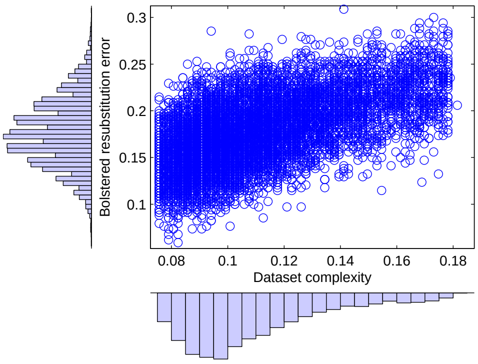

Fig. 2. (Brain) Bivariate distribution of normalised complexity and bolstered resubstitution error and univariate marginal histograms

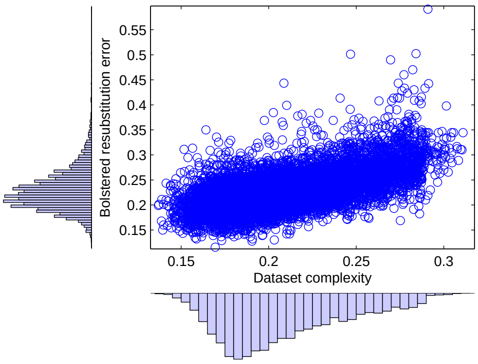

Fig. 3. (SAGE) Bivariate distribution of normalised complexity and bolstered resubstitution error and univariate marginal histograms

To quantify this dependence, the rank correlation coefficients Spearman's ρ and Kendall's τ were computed (see Table II) and the test on positive correlation at the significance level 0.05 was done which confirmed the existence of such correlation (all p-values were equal to zero). The rank correlations measure the degree to which large (small) values of one random variable correspond to large (small) values of another variable (concordance relations among variables). They are useful descriptors in our case since high (low) complexity implies that the data are difficult (easy) to accurately classify, which, in turn, means high (low) classification error. Unlike the linear correlation coefficient, ρ and τ are preserved under any monotonic (strictly increasing) transformation of the underlying random variables.

TABLE II
Spearman's ρ and Kendall's τ

Dataset no.	τ	ρ
1	0.3446	0.4964
2	0.3991	0.5581
3	0.4173	0.5864

VI. Copula method

To deeply explore dependence relations, we employed the copula method [5], [6], [7]. Copulas are functions that describe dependencies among variables and allow to model correlated multivariate data by combining univariate distributions.

A copula is a multivariate probability distribution, where each random variable has a uniform marginal distribution on the interval [0,1]. The dependence between random variables is completely separated from the marginal distributions in the sense that random variables can follow any marginal distributions, and still have the same rank correlation. This is one of the main appeals of copulas: they allow separation of dependence and marginal distribution. Though there are multivariate copulas, we will only talk about bivariate ones since our dependence relation includes two variables.

Sklar's theorem, which is the foundation theorem for copulas, states that for a given joint multivariate distribution function $H(x,y) = P(X \leq x, Y \leq y)$ and the relevant marginal distributions $F(x) = P(X \leq x)$ and $G(y) = P(Y \leq y)$, there exists a copula function C relating them, i.e. $H(x,y) = C(F(x), G(y))$. If F and G are continuous, C is unique. Otherwise, C is uniquely determined on Ran$X\times$Ran$Y$, where 'Ran' stands for the range. If F and G are continuous, the following formula is used to construct copulas from the joint distribution functions: $C(u,v) = H(F^{-1}(u), G^{-1}(v))$ [7], where F^{-1} means a quasi-inverse of F, G^{-1} means a quasi-inverse of G, and U and V are uniform random variables distributed between 0 and 1. That is, the typical copula-based analysis of multivariate (or bivariate) data starts with transformation from the (X,Y) domain to the (U,V) domain, and all manipulations with the data are then done in the latter. Such a transformation to the copula scale (unit square $\mathbf{I}^2$) can be achieved through a kernel estimator of the cumulative distribution function (cdf). After that the copula function $C(u,v)$ is generated according to the appropriate definition for a certain copula family.

In general, the choice of a particular copula may be based on the observed data. Among numerous copula families, we preferred the Frank copula belonging to the Archimedean family based on the visual look of plots in Figs. 1-3 and for dependence in the tail. We are particularly concerned with lower tail dependence when low complexity is associated with small classification error as this forms the basis for ensemble construction in our approach. The Frank copula is a one-parameter (θ is a parameter, $\theta \in]-\infty, +\infty[\backslash 0)$ copula defined for uniform variables U and V (both are defined over the unit interval) as

$$C_\theta(u,v) = -\frac{1}{\theta}\ln\left(1 + \frac{(e^{-\theta u}-1)(e^{-\theta v}-1)}{e^{-\theta}-1}\right),$$

with θ determining the degree of dependence between the marginals (we set θ to Pearson's correlation coefficient[2] between U and V so that as θ increases, positive dependence also increases).

Using copulas, dependence structure between random variables can be characterized by quadrant dependence, tail monotonicity and stochastic monotonicity [7], described below. Unlike correlation coefficients that measure the overall strength of the association between variables, these characteristics provide valuable information about how the strength varies across the distribution.

A. Quadrant dependence

Random variables X and Y are positively quadrant dependent (PQD) in terms of C if $\forall(u,v)$ in $\mathbf{I}^2$, $C(u,v) \geq uv$ [7]. We found that complexity and bolstered resubstitution error are PQD for all datasets.

It is interesting to ask when one continuous bivariate distribution H_1 is more PQD (more concordant) than another H_2. The answer is readily provided by comparing ρ or τ [6]: if $\rho(H_1) \leq \rho(H_2)$ or $\tau(H_1) \leq \tau(H_2)$, then H_2 is more PQD (more concordant) than H_1. From Table II it can be seen that SAGE is more PQD than other datasets, i.e. concordance relations between complexity and bolstered resubstitution error are stronger for this data than those for other datasets.

B. Tail monotonicity

As we mentioned above, we are interested in tail dependence when low (high) complexity associates small (large) classification error. Tail monotonicity reflects this type of association and it is stronger condition for dependence than PQD.

In terms of a copula and its first-order partial derivatives tail monotonicity is defined as follows [7]:

- Y is left tail decreasing in X iff for any v in $\mathbf{I}$, $\partial C(u,v)/\partial u \leq C(u,v)/u$ for almost all u.
- X is left tail decreasing in Y iff for any u in $\mathbf{I}$, $\partial C(u,v)/\partial v \leq C(u,v)/v$ for almost all v.
- Y is right tail increasing in X iff for any v in $\mathbf{I}$, $\partial C(u,v)/\partial u \leq (v - C(u,v))/(1-u)$ for almost all u.
- X is right tail increasing in Y iff for any u in $\mathbf{I}$, $\partial C(u,v)/\partial v \leq (u - C(u,v))/(1-v)$ for almost all v.

For the Frank copula, the first-order partial derivatives are

$$\frac{\partial C_\theta(u,v)}{\partial u} = \frac{e^{-\theta u}(e^{-\theta v}-1)}{e^{-\theta}-1+(e^{-\theta u}-1)(e^{-\theta v}-1)},$$

$$\frac{\partial C_\theta(u,v)}{\partial v} = \frac{e^{-\theta v}(e^{-\theta u}-1)}{e^{-\theta}-1+(e^{-\theta u}-1)(e^{-\theta v}-1)}.$$

Tail monotonicity is also guaranteed if $\rho \geq \tau \geq 0$ is met [7].

We verified that for all datasets bolstered resubstitution error is left tail decreasing in complexity, complexity is left tail decreasing in bolstered resubstitution error, bolstered resubstitution error is right tail increasing in complexity, and complexity is right tail increasing in bolstered resubstitution error.

[2] In [18] it was shown that Spearman's ρ and Kendall's τ can be expressed solely in terms of the copula function as follows:

$$\rho = 12\int\int C(u,v)dudv - 3,$$

$$\tau = 4\int\int C(u,v)dC(u,v) - 1,$$

where integration is over $\mathbf{I}^2$.

The integrals in these formulas can be interpreted as the expected value of the function $C(u,v)$ of uniform [0,1] random variables U and V whose joint distribution function is C, i.e.

$$\rho = 12E(UV) - 3, \quad \tau = 4E(C(u,v)) - 1.$$

As a consequence, ρ for a pair of continuous random variable X and Y is identical to Pearson's linear correlation coefficient for random variables $U = F(X)$ and $V = G(Y)$ [7].

C. Stochastic monotonicity

Stochastic monotonicity is stronger than tail monotonicity. According to [7],

- Y is stochastically increasing in X iff for any v in $\mathbf{I}$, $C(u,v)$ is a concave function of u.
- X is stochastically increasing in Y iff for any u in $\mathbf{I}$, $C(u,v)$ is a concave function of v.

A concave function implies that the second-order derivatives must be less than or equal to zero. For the Frank copula, these derivatives are

$$\frac{\partial^2 C_\theta(u,v)}{\partial u^2} = \frac{\theta e^{-\theta u}(e^{-\theta v}-1)(e^{-\theta v}-e^{-\theta})}{\left[e^{-\theta}-1+(e^{-\theta u}-1)(e^{-\theta v}-1)\right]^2},$$

$$\frac{\partial^2 C_\theta(u,v)}{\partial v^2} = \frac{\theta e^{-\theta v}(e^{-\theta u}-1)(e^{-\theta u}-e^{-\theta})}{\left[e^{-\theta}-1+(e^{-\theta u}-1)(e^{-\theta v}-1)\right]^2}.$$

Since $\theta > 0$ in our case (positive dependence), it is easy to verify that $\frac{\partial^2 C_\theta(u,v)}{\partial u^2} \leq 0$ and $\frac{\partial^2 C_\theta(u,v)}{\partial v^2} \leq 0$, which, in turn, implies that $C_\theta(u,v)$ is concave. Thus, for all datasets in our study, bolstered resubstitution error is stochastically increasing in complexity and complexity is stochastically increasing in bolstered resubstitution error.

VII. Ensembles of classifiers

An ensemble of classifiers consists of several classifiers (members) that make predictions independently of each other. After that, these predictions are combined together to produce the final prediction. Though ensemble members can belong to different types of algorithms, because of our interest in k-NN classifiers we choose only this algorithm. Moreover, the value of k is fixed to 3 for all ensemble members[3]. As a combination technique, the conventional majority vote was selected in order to demonstrate that ensembles built with our approach show good performance even when employing simple non-trainable combiners.

It is well known that an ensemble is able to outperform its best performing member if ensemble members make mistakes on different cases so that their predictions are uncorrelated and diverse as much as possible. On the other hand, an ensemble must include a sufficient number of accurate classifiers since if there are only few good votes, they can be easily drowned out among many bad votes. As a result, an ensemble can predict wrongly most of the time.

So far many definitions of diversity were proposed [19], [4], but unfortunately the precise definition is still largely illusive. Because of this fact, we decided not to follow any *explicit* definition of diversity, but to introduce diversity implicitly instead. Since we fixed the base classifier and its parameter, one of the solutions is to let each ensemble member to work with its own feature subset.

Feature subset selection can be done in two ways: either applying a certain feature selection algorithm or a group of such algorithms, or randomly sampling features from the original feature set. As concluded in [20], differences in classification performance among feature selection algorithms are less significant than performance differences among the error estimators used to implement these algorithms. In other words, the way of how error is computed has a larger influence on classification accuracy than the choice of a feature selection algorithm. Since bolstered resubstitution error is a low-bias, low-variance estimate of classification error, which is what is needed for high dimensional gene expression data, we opt for random feature selection. Figs. 1-3 show that random feature selection leads to diversity since one complexity value corresponds to several values of error. Given that it is difficult to carry out biological analysis of many genes, we restricted the number of genes to be sampled to 50, i.e. each ensemble member works with 1 to 50 randomly selected (sampled with replacement) genes.

Based on the above-mentioned, two approaches to form ensembles consisting of L classifiers are explored:

1. Randomly select L feature subsets, one subset per classifier, as described above. Classify the data with each classifier and combine votes.
2. Randomly select $M > L$ (e.g., $M = 100$) feature subsets and compute the dataset complexity for each of them. Rank subsets according to their complexity and select L least complex subsets while ignoring the others. Classify the data with each classifier and combine votes.

We will call the first approach conventional to distinguish it from ours, which is the second approach. The typical (and perhaps the earliest) example of the former is [21]. As one can see, the main difference between two approaches lies in the way of choosing feature subsets: in the conventional approach, subsets are chosen regardless of their classification power. As a result, one may equally expect both very good and very bad base learner predictions. In contrast, in our approach, subsets are chosen based on the measure *directly* related to classification performance. As lower complexity is associated with smaller bolstered resubstitution error as shown in Section V, selection of the subsets of smaller complexity implies more accurate classifiers included into an ensemble. Since each ensemble member works with only a small subset of all features, such feature space decomposition is akin to dividing a complex problem into simpler subproblems. Thus, with our approach, both diversity and accuracy requirements for ensembles are satisfied. Hence, we can expect better *average* classification performance with our approach compared to the conventional approach.

VIII. Simulation results

In ensemble applications to gene expression based cancer classification, a small and accurate ensemble is of importance, since too many ensemble members would complicate biological understanding of relations among genes. Bearing this in mind, we set the number of 3-NNs (L) in the ensemble to be equal 3, 5, and 7.

Table III represents the dataset complexity as estimated by the normalized rank sum statistic W (see Section III)

[3] In our opinion, $k = 1$ tends to lead to optimistic estimation of bolstered resubstitution error.

for different values of L when ensembles were built with our approach. For each dataset and L, two values are given: average minimum and average maximum complexity (averaging over 100 trials) of the selected feature subsets included into an ensemble. It can be observed that complexity for each dataset is rather stable as L changes. Brain appears to be far less complex than the other datasets. For the conventional ensemble approach 'avr.max' often went to a very big value, meaning poor class separation according to the Wilcoxon rank sum test.

TABLE III
AVERAGE MINIMUM AND MAXIMUM NORMALIZED W FOR FEATURE SUBSETS SELECTED WITH OUR ENSEMBLE GENERATING APPROACH FOR VARIOUS VALUES OF L

Dataset no.		$L=3$	$L=5$	$L=7$
1	avr.min	0.1587	0.1584	0.1581
1	avr.max	0.1676	0.1723	0.1769
2	avr.min	0.0761	0.0760	0.0761
2	avr.max	0.0781	0.0803	0.0818
3	avr.min	0.1501	0.1494	0.1493
3	avr.max	0.1584	0.1619	0.1655

We also compared both ensemble schemes as well as each ensemble scheme against a single best classifier (SBC) in the ensemble. Let e_{SBC} and e_{ENS} be bolstered resubstitution error achieved with a SBC and an ensemble, respectively. The following statistics widely used in machine learning and data mining were computed over 100 ensemble generations:

- win-tie-loss count, where 'win'/'tie'/'loss' means the number of times when an ensemble was superior /equal/inferior in terms of bolstered resubstitution error to a SBC (in other words, the number of times when $e_{ENS} < e_{SBC}$, $e_{ENS} = e_{SBC}$, $e_{ENS} > e_{SBC}$, respectively).
- 'min. win' and 'max. win'(minimum and maximum differences $e_{SBC} - e_{ENS}$ when an ensemble outperforms its SBC,
- 'min. loss' and 'max. loss' (minimum and maximum differences $e_{ENS} - e_{SBC}$ when a SBC outperforms an ensemble.

Tables IV-VI contain values of these statistics. If there were no losses, this fact is marked as 'no'. 'C' and 'O' stand for the conventional and our approaches to ensemble construction while numbers from 1 to 3 in the first column of each table correspond to 'Dataset no.' as in Table I.

TABLE IV
COMPARISON OF A SBC AND TWO ENSEMBLES WHEN $L = 3$

		win-tie-loss	min.win	max.win	min.loss	max.loss
1	C	84/3/13	0.0032	0.0629	$< 10^{-4}$	0.0548
1	O	96/2/2	0.0016	0.0532	0.0065	0.0145
2	C	79/2/19	0.0029	0.0824	0.0029	0.0324
2	O	98/0/2	0.0029	0.0794	0.0059	0.0059
3	C	81/1/18	0.0011	0.0700	0.0011	0.0578
3	O	99/0/1	0.0022	0.0856	0.0056	0.0056

TABLE V
COMPARISON OF A SBC AND TWO ENSEMBLES WHEN $L = 5$

		win-tie-loss	min.win	max.win	min.loss	max.loss
1	C	93/1/6	0.0016	0.0694	0.0032	0.0274
1	O	100/0/0	0.0016	0.0597	no	no
2	C	76/2/22	0.0029	0.0853	0.0029	0.0735
2	O	99/1/0	0.0059	0.1000	no	no
3	C	84/0/16	0.0033	0.1022	0.0022	0.0800
3	O	99/0/1	0.0122	0.1089	0.0022	0.0022

TABLE VI
COMPARISON OF A SBC AND TWO ENSEMBLES WHEN $L = 7$

		win-tie-loss	min.win	max.win	min.loss	max.loss
1	C	99/0/1	0.0032	0.0726	0.0113	0.0113
1	O	99/1/0	0.0016	0.0694	no	no
2	C	88/2/10	0.0029	0.0971	0.0088	0.0412
2	O	99/0/1	0.0088	0.1088	0.0029	0.0029
3	C	80/0/20	0.0022	0.1256	0.0044	0.0700
3	O	100/0/0	0.0178	0.1033	no	no

As one can see, both ensemble schemes were superior to a SBC on all datasets for the most part. When analyzing the performance of two ensemble schemes, it was observed that *on average*, our approach yields better results in the sense that its win (loss) count is typically higher (lower) and the absolute losses to a SBC are lower, too. In contrast, the conventional ensemble generating approach sometimes shows spectacular results (e.g., the high max.win count), but it also suffers many defeats from a SBC. That is, its results are less predictable since there is no control over complexity of the selected feature subsets and hence, if such 'complex' subsets are selected, a SBC can render ensemble efforts to further lower error fruitless. With the explicit selection of the least complex subsets, our approach is able to succeed where the comparative approach failed.

Table VII summarizes the average bolstered resubstitution error (over 100 runs) and its standard deviation achieved with two ensemble schemes. From Table VII it is clearly noticable that both the average error and its standard deviation are much smaller for our ensemble scheme.

TABLE VII
AVERAGE BOLSTERED RESUBSTITUTION ERROR AND ITS STANDARD DEVIATION FOR TWO ENSEMBLE SCHEMES

		$L=3$	$L=5$	$L=7$
1	C	0.110±0.025	0.091±0.019	0.080±0.013
1	O	0.092±0.014	0.077±0.012	0.071±0.014
2	C	0.129±0.035	0.117±0.032	0.101±0.031
2	O	0.081±0.022	0.062±0.019	0.054±0.017
3	C	0.177±0.034	0.160±0.040	0.152±0.040
3	O	0.130±0.023	0.113±0.022	0.098±0.019

Finally, both ensemble schemes were compared with the nearest shrunken centroids (NSC) algorithm [22], which was proposed for cancer classification based on gene expression levels. Another reason for its selection was that it belongs to the class of the nearest neighbor algorithms that are utilized in our ensembles.

Briefly speaking, NSC shrinks class centroids toward the overall centroid after normalizing by the within-class standard deviation for each gene. This normalization gives higher weight to those genes whose expression is stable within cases (examples) belonging to the same class. More precisely, given a set of n genes, for each gene i, a t statistic d_{ik}, comparing the centroid $\bar{x}_{ik}$ of class k to the overall centroid $\bar{x}_i$, is computed:

$$d_{ik} = \frac{\bar{x}_{ik} - \bar{x}_i}{\sqrt{1/n_k + 1/n}\,(s_i + s_0)}$$

where s_i is the pooled within-class standard deviation for gene i and the positive constant s_0 assures that genes with low expression levels do not produce large d_{ik} by chance. NSC shrinks each d_{ik} by the amount of shrinkage, Δ, according to $d'_{ik} = sign(d_{ik})(\mid d_{ik} \mid - \Delta)_+$, where $+$ means positive part ($a_+ = a$ if $a > 0$ and zero otherwise), in order to compute new shrunken centroids $\bar{x}'_{ik}$:

$$\bar{x}'_{ik} = \bar{x}_i + \sqrt{1/n_k + 1/n}\,(s_i + s_0)\,d'_{ik}$$

The classification rule for an example x chooses the class k^* whose shrunken centroid is the nearest one:

$$k^* = \arg\min_k \sum_{i=1}^{n} \frac{(x_i - \bar{x}'_{ik})^2}{(s_i + s_0)^2} - 2\log \pi_k$$

where π_k is an estimate of the prior for class k.

Note that NSC implicitly performs gene selection, because after shrinking some d'_{ik} may become zero: only genes having at least one nonzero d'_{ik} are retained in order to classify test data. The larger Δ, the larger the amount of shrinkage, and hence, the fewer genes are selected. It can be noticed from the formula for d'_{ik} above that there is a working range of Δ values (outside of this range, changes in Δ do not affect d'_{ik}).

Table VIII contains the best NSC results collected over 100 runs. These results might look good compared to our results in Table VII but one should take into account that NSC results in Table VIII are obtained for the optimal values of Δ. Figures 4-6 show the complete picture. In each of these figures, the vertical axis denotes the working range of Δ values and the horizontal axis denotes the bolstered resubstitution error values. It can be seen that the NSC results are very sensitive to Δ. Hence, they can be highly sub-optimal to our results if Δ is not carefully set up. Besides, for Colon dataset and $L = 5, 7$, our results are better than the NSC ones for the optimal Δ. Best NSC results for Brain and SAGE datasets are achieved with no gene selection. Based on them, two conclusions can be drawn: 1) our ensemble scheme provides both dimensionality reduction[4] and 2) high accuracy, especially for 'complex' datasets. On the other hand, NSC can be competitive on 'easy' datasets but its results strongly depend on the optimal Δ. It turned out that reaching both gene selection and high accuracy can be difficult for this algorithm.

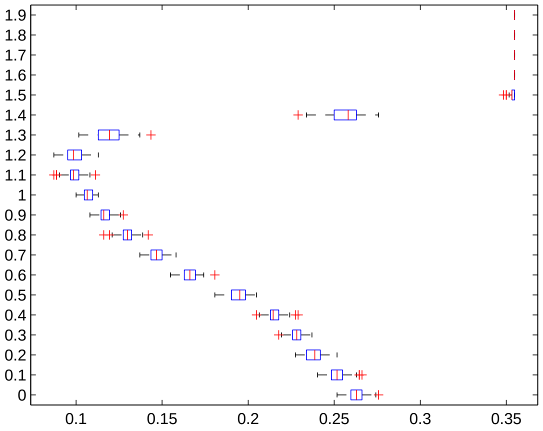

Fig. 4. (Colon) Bolstered resubstitution error of NSC versus Δ

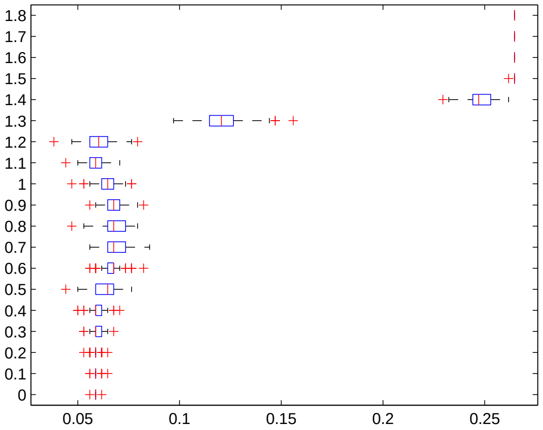

Fig. 5. (Brain) Bolstered resubstitution error of NSC versus Δ

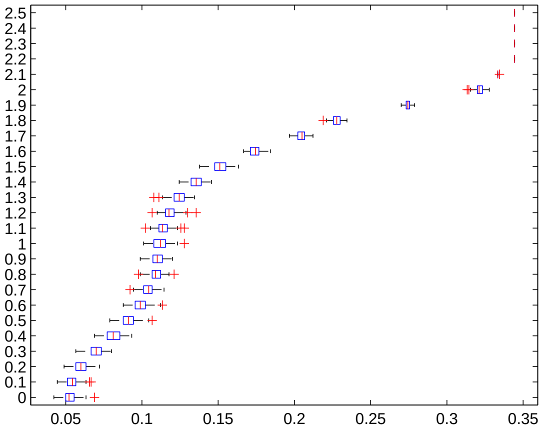

Fig. 6. (SAGE) Bolstered resubstitution error of NSC versus Δ

[4] In the worst case when different sampled subsets of features do not overlap, our ensemble scheme provides the compression ratio $c = \frac{D}{50L}$, where D and 50 are the number of original features and the maximal subset size, respectively. For instance, given $L = 7$, $c \approx 5.7$ for Colon, $c \approx 16.8$ for Brain, and $c \approx 76.5$ for SAGE. For smaller L, compression is even higher.

TABLE VIII

AVERAGE BOLSTERED RESUBSTITUTION ERROR AND ITS STANDARD DEVIATION FOR NSC. BEST RESULTS ARE GIVEN (THIRD COLUMN) TOGETHER WITH THE OPTIMAL Δ (SECOND COLUMN)

1	1.1	0.099±0.004
2	0.0	0.059±0.000
3	0.0	0.052±0.005

IX. CONCLUSION

We proposed a new ensemble method, based on random selection of features and direct estimation of data complexity to guide the construction of committees of k-NN classifiers. As demonstrated on three gene expression datasets, our approach leads to lower bolstered resubstitution error compared to 1) the conventional ensemble approach, purely based on random selection of features, 2) the single best classifier in the ensemble, and 3) the nearest shrunken centroids algorithm proposed for gene expression based cancer classification.

Our approach springs from dependence between dataset complexity and bolstered resubstitution error established through the copula method. Using the Frank copula fit to the data, we found that there is positive dependence between complexity and error, where low (high) complexity corresponds to small (large) error. Hence, the dataset complexity serves as a reliable indicator of the expected classification performance without the need to carry out actual classification. As a result, selection of least complex subsets of features implies more accurate ensemble members and therefore it ensures better ensemble performance.

Our method is advantageous to apply at very low L (e.g., from 3 to 7), which is important when further analysis of relations among selected genes is desired.

Future work includes exploring new measures of dataset complexity (see e.g. [23], [24] and comparing them with our measure in order to find out if they can complement each other. Using several complexity measures will allow us to consider more sophisticated relations between complexity and accuracy which might escape our notice in the bivariate copula model analyzed in this paper. We also plan to compare our ensemble scheme with other feature selection/ensemble methods (see e.g. [25], [26]), using also a larger set of biomolecular data, in order to have a better understanding of pros and cons of our approach.

REFERENCES

[1] U. Alon, N. Barkai, D.A. Notterman, K. Gish, S. Ybarra, D. Mack, and A.J. Levine, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," *Proceedings of the National Academy of Sciences*, vol. 96, pp. 6745–6750, 1999.

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

[3] S. Dudoit and J. Fridlyand, "Classification in Microarray Experiments," in *Statistical Analysis of Gene Expression Microarray Data* (Chapter 3), Edited by T. Speed, Boca Raton, FL: Chapman & Hall\CRC Press, 2003.

[4] L. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*, Hoboken, NJ: John Wiley & Sons, 2004.

[5] A. Sklar, Fonctions de répartition à n dimensions et leurs marges, *Publications of the Institute of Statistics, University of Paris*, pp. 229–231, 1959.

[6] H. Joe, *Multivariate Models and Dependence Concepts*, Boca Raton, FL: Chapman & Hall\CRC Press, 1997.

[7] R.B. Nelsen, *An Inroduction to Copulas*, New York, NY: Springer Science+Business Media, 2006.

[8] J.H. Zar, *Biostatistical Analysis*, Upper Saddle River, NJ: Prentice Hall, 1999.

[9] U. Braga-Neto and E.R. Dougherty, "Bolstered error estimation," *Pattern Recognition*, vol. 37, pp. 1267–1281, 2004.

[10] O. Okun and H. Priisalu, Random forest for gene expression based cancer classification: overlooked issues, *Proc. 3rd Iberian Conference on Pattern Recognition and Image Analysis*, Girona, Spain, June 2007, pp. 483–490.

[11] O. Okun and H. Priisalu, Dataset complexity and gene expression based cancer classification, *Proc. 4th International Meeting on Computational Intelligence Methods in Bioinformatics and Biostatistics*, Portofino, Italy, July 2007, pp. 484–490.

[12] http://microarray.princeton.edu/oncology/affydata/index.html

[13] http://www.broad.mit.edu/mpr/CNS/

[14] http://lisp.vse.cz/challenge/ecmlpkdd2004

[15] V.E. Velculescu, L. Zhang, B. Vogelstein, and K.W. Kinzler, "Serial analysis of gene expression," *Science* , vol. 270, pp. 484–487, 1995.

[16] T.K. Ho and M. Basu, "Complexity measures of supervised classification problems," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, pp. 289–300, 2002.

[17] T.H. Bø and I. Jonassen, "New feature subset selection procedures for classification of expression profiles," *Genome Biology*, vol. 3, pp. 0017.1–0017.11, 2002.

[18] B. Schweizer and E.F. Wolff, "On nonparametric measures of dependence for random variables," *The Annals of Statistics*, vol. 9, pp. 879–885, 1981.

[19] L. Kuncheva and C.J. Whitaker, "Measures of diversity in classifier ensembles," *Machine Learning*, vol. 51, pp. 181–207, 2003.

[20] C. Sima, S. Attoor, U. Braga-Neto, J. Lowey, E. Suh, and E.R. Dougherty, Error estimation confounds feature selection in expression-based classification, *Proc. IEEE International Workshop on Genomic Signal Processing and Statistics*, Newport, RI, 2005.

[21] S. Bay, "Nearest neighbor classification from multiple feature sets," *Intelligent Data Analysis*, vol. 3, pp. 191–209, 1999.

[22] R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu, "Diagnosis of multiple cancer types by shrunken centroids of gene expression," *Proceedings of the National Academy of Sciences*, vol. 99, pp. 6567–6572, 2002.

[23] S. Singh, "Multiresolution estimates of classification complexity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, pp. 1534–1539, 2003.

[24] M. Dong and R. Kothari, "Feature subset selection using a new definition of classifiability," *Pattern Recognition Letters*, vol. 24, pp. 1215–1225, 2003.

[25] N. Kasabov, "Global, local and personalised modeling and pattern discovery in bioinformatics: an integrated approach," *Pattern Recognition Letters*, vol. 28, pp. 673–685, 2006.

[26] Z. Wang, V. Palade, and Y. Xu, Neuro-fuzzy ensemble approach for microarray cancer gene expression data analysis, *Proc. 2nd IEEE International Symposium on Evolving Fuzzy Systems*, Ambelside, UK, September 2006, pp. 241–246.