Lowering Uncertainty of Cancer Classification

Oleg Okun
University of Oulu, Finland

Helen Priisalu
Teradata, Finland

Abstract

A new ensemble scheme is proposed for classifying high dimensional data, which exploits dependence between data complexity, determining how difficult to classify a given dataset, and classification error. As a classification task, gene expression based cancer classification is studied, with a k-nearest neighbor as a base classifier. Experiments carried out on five datasets show the importance of taking into account dataset complexity when constructing ensembles of nearest neighbors.

Keywords: Machine Learning, Classification, Ensembles of Classifiers, Bioinformatics, Gene Expression.

1 Introduction

When classifying high dimensional data, many algorithms tend to demonstrate a similar performance, e.g. in terms of error rate, thus leading to uncertainty which algorithm to prefer over others. This uncertainty can be efficiently exploited by first running several algorithms in parallel on a dataset and then combining their predictions by voting, which is termed as an ensemble of classifiers in the literature [7].

In this paper, we study a particular type of high dimensional data: gene expression levels which are used for discrimination between normal and cancer specimens or between different types of cancer. 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 because it performed well for cancer classification, compared to more sophisticated classifiers [5]. 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. k-NNs are known to be insensitive to small perturbations of training data. Hence, k-NNs seem to be difficult to combine into a highly accurate ensemble because of the lack of diversity in predictions of individual k-NNs. One solution is to associate each k-NN with its own feature subset. The most straightforward and probably fastest approach is to randomly sample subsets from the original features [2]. It is appropriate if one is only interested in the performance figures, regardless of the features used to achieve them[1]. Another approach is to apply feature selection. However, gene expression datasets often do not have a separate test set to check generalization of a classifier, which can easily introduce the selection

[1] For cancer classification, genes relevant to cancer or its suppression are highly sought. Hence, selected features are required to meet two goals: to provide good discrimination between classes and to be meaningful for further biological analysis.

L. Magdalena, M. Ojeda-Aciego, J.L. Verdegay (eds): Proceedings of IPMU'08, pp. 543–550
Torremolinos (Málaga), June 22–27, 2008

bias when relying on the wrapper (classifier based) models for feature selection. Besides, many feature selection algorithms are time-consuming, given many thousands of features.

Thus, with random feature sampling, the uncertainty about whether we picked right features for cancer classification is high. To lower this uncertainty while keeping the computational cost of feature selection low, we propose a novel approach based on the estimation of data complexity to construct a classifier ensemble. Using the copula method [8, 13] for exploring dependence or concordance relations in multivariate data, we found that it is possible to roughly predict classification performance on the basis of dataset complexity. According to our findings, low (high) complexity is associated with small (large) classification error. Hence, selecting feature subsets of low complexity implies accurate predictions of individual k-NNs, which, in turn, leads to an accurate k-NN ensemble.

2 Gene expression datasets

Five gene expression datasets whose characteristics are summarized in Table 1 are employed in this work. Numbers in brackets in the last column are the number of normal and cancer cases, respectively. SAGE 1 and SAGE 2 datasets contain multiple cancer types that were treated as a single type. Colon and Prostate datasets include one type of cancer as indicated in the name of each dataset. Brain dataset (also known as Dataset B) contains 34 medulloblastoma cases, 9 of which are desmoplastic and 25 are classic. Preprocessing if necessary was based on the procedure from the original article (see the second column in Table 1).

Table 1: Gene expression datasets.

Dataset	Source	# genes	# cases
SAGE 1	[6]	822	74 (24,50)
Colon	[1]	2000	62 (22,40)
Brain	[9]	5893	34 (9,25)
SAGE 2	[6]	27679	90 (31,59)
Prostate	[12]	12600	102 (52,50)

3 Dataset complexity

It is known that the performance of classifiers is strongly data-dependent. To gain insight into a supervised classification problem[2], one can adopt dataset complexity characteristics. The goal of such characteristics is to provide a score reflecting how well classes of the 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 [4]). Projection coordinates then serve as input for the Wilcoxon rank sum test for equal medians [15] (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 (R_1) in the smaller group[3]. The value of the rank sum statistic, i.e. R_1, 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 (the closer two medians to each other), i.e. the worse separation between classes. To compare scores coming from different datasets, each score 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 $\sum_{i=1}^{N} i = R_1 + R_2$. Then the normalized score is $\frac{R_1}{R_1+R_2}$ and it lies between 0 and 1.

4 Bolstered resubstitution error

This is a low-variance and low-bias classification error estimation proposed in [3]. 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 [3] avoided this pitfall by randomly generating a number of artificial points

[2] Two-class problems are assumed.

[3] Let R_2 be the sum of the ranks in the larger group.

(cases) in the neighborhood of each training point[4]. 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. For further details, see [3].

5 Dependence relation

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 roughly predict the latter[5].

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. A typical result of such simulation is shown in Figure 1 for SAGE 2 dataset[6] together with marginal histograms for each variable. The dependence between complexity and error is clearly detectable in Figure 1, i.e. error increases (decreases) as complexity increases (decreases). To form good k-NN ensembles, it is all important for us to know how accurate ensemble members are. Rough estimation of the expected accuracy can be thus gained from the complexity.

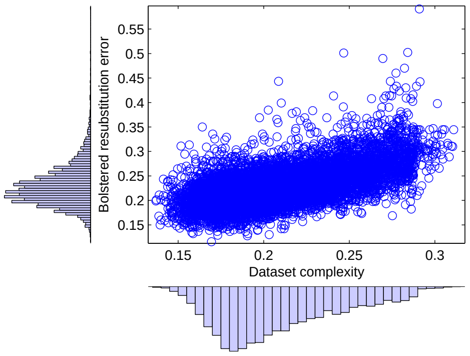

Figure 1: (SAGE 2) Bivariate distribution of normalized complexity and bolstered resubstitution error and univariate marginal histograms.

To quantify this dependence, the rank correlation coefficients Spearman's ρ and Kendall's τ were computed (see Table 2) 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[7] 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 2: Spearman's ρ and Kendall's τ.

Dataset	τ	ρ
SAGE 1	0.3100	0.4468
Colon	0.3446	0.4964
Brain	0.3991	0.5581
SAGE 2	0.4173	0.5864
Prostate	0.4288	0.6006

To deeply explore dependence relations, we employed the copula method [8, 13]. 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

[4] Braga-Neto and Dougherty recommended 10 artificial cases per each training case.

[5] We do not seek the regression-like dependence, where each complexity value would associate one error value. It is more important for us to know how changes in complexity affect changes in error.

[6] Plots for other datasets look similar and they are omitted due to space limitation.

[7] However, definitions of these relations by ρ and τ are different; hence, the difference in the absolute values as observed in Table 2.

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 [13] 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, the following formula is used to construct copulas from the joint distribution functions: $C(u,v) = H(F^{-1}(u), G^{-1}(v))$ [8], where 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 the 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. After that the copula function $C(u,v)$ is generated according to the appropriate definition for a certain copula family.

In [10] it was shown that Spearman's ρ and Kendall's τ can be expressed solely in terms of the copula function as follows:

$$\rho = 12 \int\int C(u,v)\,du\,dv - 3,$$
$$\tau = 4 \int\int C(u,v)\,dC(u,v) - 1,$$

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 $V = G(Y)$ [8].

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 and for dependence in the tail. Besides, this copula type permits negative as well as positive dependence. 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 between U and V so that as θ increases, positive dependence also increases).

Correlation coefficients measure the overall strength of the association, but give no information about how that varies across the distribution, e.g. in the tail. Hence, additional characteristics of dependence structure are necessary. We checked and found that for all the datasets in our study the following associations take place: quadrant dependence, tail monotonicity, and stochastic monotonicity [8]. They strengthen our hypothesis about concordance of dataset complexity and bolstered resubstitution error.

6 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[8]. 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.

The main goal for any ensemble is to perform better than its most accurate member[9]. 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, and as a result, an ensemble can predict wrongly most of the time.

So far many definitions of diversity were proposed [7], 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 [11], for small samples like those in this study, 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. Figure 1 shows that random feature selection leads to diversity of prediction estimates 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 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 [2]. 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 5, 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

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

[9] Otherwise, the extra computational cost is not justified.

ensembles are satisfied. Hence, we can expect better *average* classification performance with our approach compared to the conventional approach.

7 Experiments

To generate ensembles, we set the number of 3-NNs (L) in the ensemble to be equal to 3, 5, 7, 9 and 11.

Tables 3-7 summarize the average bolstered resubstitution error (over 100 runs) and its standard deviation achieved with two ensemble schemes. 'C' and 'O' stand for the conventional and our approaches to ensemble construction. It is clearly noticeable that both the average error and its standard deviation are much smaller for our approach.

Table 3: Average bolstered resubstitution error and its standard deviation for two ensemble schemes ($L = 3$).

SAGE 1	C	0.141±0.025
	O	0.119±0.016
Colon	C	0.110±0.025
	O	0.092±0.014
Brain	C	0.129±0.035
	O	0.081±0.022
SAGE 2	C	0.177±0.034
	O	0.130±0.023
Prostate	C	0.141±0.034
	O	0.101±0.022

Table 4: Average bolstered resubstitution error and its standard deviation for two ensemble schemes ($L = 5$).

SAGE 1	C	0.125±0.024
	O	0.105±0.017
Colon	C	0.091±0.019
	O	0.077±0.012
Brain	C	0.117±0.032
	O	0.062±0.019
SAGE 2	C	0.160±0.040
	O	0.113±0.022
Prostate	C	0.111±0.026
	O	0.078±0.016

Table 5: Average bolstered resubstitution error and its standard deviation for two ensemble schemes ($L = 7$).

SAGE 1	C	0.119±0.021
	O	0.098±0.014
Colon	C	0.080±0.013
	O	0.071±0.014
Brain	C	0.101±0.031
	O	0.054±0.017
SAGE 2	C	0.152±0.040
	O	0.098±0.019
Prostate	C	0.096±0.020
	O	0.071±0.012

Table 6: Average bolstered resubstitution error and its standard deviation for two ensemble schemes ($L = 9$).

SAGE 1	C	0.110±0.018
	O	0.094±0.014
Colon	C	0.074±0.015
	O	0.066±0.010
Brain	C	0.092±0.027
	O	0.047±0.016
SAGE 2	C	0.143±0.039
	O	0.093±0.017
Prostate	C	0.084±0.016
	O	0.066±0.011

For comparison, we included experiments with one filter-based feature selection algorithm searching for the optimal set of genes using a Markov blanket [14]. This algorithm called RBF (redundancy-based filter), especially intended for gene expression data analysis, aims at elimination of redundant genes. It is based on the fact that a gene can be safely eliminated if there is a Markov blanket for it. Because finding a Markov blanket is computationally demanding, the solution in [14] is to approximate it so that the Markov blanket always consists of one gene. All original genes are first ranked based on the estimate of how strongly a certain gene is correlated to the class[10]. Then each gene is checked if it has any approximate Markov blanket in the current set. Table 8 lists the average bolstered

[10] We used entropy-based symmetrical uncertainty for defining correlation.

Table 7: Average bolstered resubstitution error and its standard deviation for two ensemble schemes ($L = 11$).

SAGE 1	C	0.111±0.018
	O	0.088±0.015
Colon	C	0.068±0.013
	O	0.064±0.010
Brain	C	0.088±0.028
	O	0.045±0.015
SAGE 2	C	0.151±0.043
	O	0.089±0.016
Prostate	C	0.076±0.014
	O	0.063±0.010

resubstitution error and its standard deviation computed over 100 runs when RBF was applied to each dataset prior to 3-NN classification. The third column contains the number of genes retained after filtering.

Table 8: Average bolstered resubstitution error and its standard deviation when RBF was applied before 3-NN classification (RBF+3-NN).

Dataset	**RBF**	**#genes**
SAGE 1	0.199±0.011	12
Colon	0.107±0.010	3
Brain	0.055±0.010	6
SAGE 2	0.145±0.005	152
Prostate	0.117±0.008	2

It can be observed that our ensemble scheme almost always outperforms RBF+3-NN, except for Brain data[11], which was easy to classify according to dataset complexity. In contrast, the conventional scheme was inferior to RBF+3-NN on many more occasions, which again confirms the superiority of our approach to ensemble construction.

Finally, we computed the win-tie-loss count[12] frequently employed in machine learning and data mining in order to determine if either ensemble scheme is inferior to a single best 3-NN in the ensemble. The complete results are reported elsewhere and here we only provide their brief summary compiled after 100 ensemble generations. The main result is that both ensemble schemes are indeed superior to a single best 3-NN in the ensemble on all datasets for the most part. It was also 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 single best 3-NN are lower, too. In contrast, the conventional ensemble generating approach sometimes shows spectacular results, but it also suffers many defeats from a single best 3-NN. That is, its results are more uncertain (less predictable) since there is no control over complexity of the selected feature subsets and hence, if such ‘complex’ subsets are selected, a single best 3-NN 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.

8 Conclusion

We proposed a new ensemble generating scheme using a 3-NN as the base classifier and tested this scheme on gene expression based cancer classification. Our approach leads to lower bolstered resubstitution error compared to the conventional ensemble approach, purely based on random selection of features, and to the single best classifier in the ensemble. In addition, our scheme outperforms a 3-NN preceded by the RBF algorithm [14], especially proposed to deal with redundancy among genes.

Our approach springs from dependence between dataset complexity and bolstered resubstitution error established through the copula method. We found that there is positive dependence between complexity and error, where low (high) complexity corresponds to small (large) error. Exploiting this fact, it is possible to achieve more predictable and therefore less uncertain results for cancer clas-

[11] $L = 3, 5$.

[12] Given two algorithms A and B, this characteristic counts the number of times when algorithm A is more accurate than algorithm B (win count), the number of times when both algorithms demonstrate the same error rate (tie count), and the number of times when algorithm A is less accurate than algorithm B (loss count).

sification.

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. Bay, Nearest neighbor classification from multiple feature sets, *Intelligent Data Analysis*, vol. 3, 191–209, 1999.

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

[4] 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.

[5] 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.

[6] O. Gandrillon, Guide to the gene expression data, in: P. Berka and B. Crémilleux, eds., *Proceedings of the ECML/PKDD Discovery Challenge Workshop* (Pisa, Italy, 2004) 116-120.

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

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

[9] 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.

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

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

[12] D. Singh, P.G. Febbo, K. Ross, D.G. Jackson, J. Manola, C. Ladd, P. Tamayo, A.A. Renshaw, A.V. D'Amico, J.P. Richie, E.S. Lander, M. Loda, P.W. Kantoff, T.R. Golub, and W.R. Sellers, Gene expression correlates of clinical prostate cancer behavior, *Cancer Cell*, vol. 1, pp. 203–209, 2002.

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

[14] L. Yu, Feature selection for genomic data analysis, in *Computational Methods of Feature Selection* (Chapter 17), Edited by H. Liu and H. Motoda, Boca Raton, FL: Chapman & Hall\CRC, 2008.

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