An evaluation of using ensembles of classifiers for predictions based on genomic and proteomic data

Ringnér, Markus; Johansson, Peter

2006

An evaluation of using ensembles of classifiers for predictions based on genomic and proteomic data

Peter Johansson[1] and Markus Ringnér*[1]

[1]Computational Biology and Biological Physics Group, Dept. of Theoretical Physics, Lund University, Sölvegatan 14A, SE-223 62 Lund, Sweden

Email: Peter Johansson - peter@thep.lu.se; Markus Ringnér*- markus@thep.lu.se;

*Corresponding author

Abstract

Background: Classification of expression profiles to predict disease characteristics of for example cancer is a common application in high-throughput gene and protein expression research. Cross-validation is often used to optimize design of classifiers, with the aim to construct an optimal single classifier. In this work, we explore if classification performance can be improved by aggregating classifiers into ensembles that use committee votes for classification.

Results: We investigated if combining classifiers into ensembles improved classification performance compared to single classifiers. A couple of commonly used classifiers, nearest centroid classifier and support vector machine, were evaluated using four publicly available data sets. We found ensemble methods generally performed better than corresponding single classifiers.

Background

Using microarrays and high-throughput mass spectronomy, gene and protein expression profiles of samples from patients have been measured for many diseases. A common application is to develop approaches for diagnostic predictions based on expression profiles [1–5]. To build a diagnostic predictor for different diagnostic classes, one has to find the characteristic features that either define each class or discriminate between classes, and build a predictor that based on these characteristics is able to predict the class of unknown samples.

The construction of a predictor can be divided into different parts. A common division is into classifier selection, feature selection, classifier training and independent validation. Classifier selection includes choosing between different types of classifiers such as support vector machines (SVM) or diagonal linear discriminant classifiers, but also choosing values for the parameters of the classifier. Feature selection is used to select inputs for the classifier, for example, selecting a subset of genes to use in classification based on gene expression profiles. The purpose of feature selection can vary, including selecting the smallest possible set of features that results in a required prediction performance, or selecting the set of features that results in the optimal prediction performance. Gene and protein expression data sets typically contain many more features than samples. The features can, for example, be genes probed by microarrays or m/z values from discretized mass spectra. In this situation large independent test data sets are rare and often cross-validation is used to validate classifiers and evaluate their predictive performance.

In v-fold cross-validation, samples are randomly split into v groups of which one is set aside as a test set and the remaining groups are a training set used to train a classifier. The procedure is repeated with each of the v groups as a test set. These test sets would provide an honest estimate of the predictive performance, in the case where there are no choices in classifier construction. However, suppose parameters of the predictor are tuned, or features are selected, to achieve the best prediction results for the test set, then the test set is no longer independent of the construction of the predictor. Such dishonest use of the test set will lead to overly optimistic estimates of the predictive performance [6].

To circumvent this dishonest use of the test set, the training samples from the cross-validation can be used in a second internal procedure of cross-validation to optimize the predictive performance of the classifier. The external cross-validation is used solely to evaluate the test procedure. Procedures in which an interior cross-validation loop is used to construct predictors and an exterior cross-validation loop for evaluating the test performance have applied to classification of gene expression profiles [7,8].

When internal cross-validation is used to optimize choices for predictor construction, many classifiers are constructed for each test set. There are many ways to proceed in the construction of a predictor for a test set. For example, one can train a single classifier using the entire training set and the optimal choices from the internal cross-validation [8], or one can use the classifiers optimized in the internal cross-validation as an ensemble that predicts the class of samples in the test set by using a committee vote. Ensembles of different types of classifiers, including artificial neural networks and decision trees have been used for classification based on gene expression profiles [2, 9–11]

Many comparisons of classifiers for gene expression data have been performed [8, 12]. While the results of these comparisons have been somewhat data set dependent, simple classifiers combined with filter methods for feature selection have generally been found to perform very well. There are many methods to aggregate classifiers into ensembles. Common approaches to aggregate classifiers include bagging and boosting. In bagging, ensemble members are trained on individual training sets drawn at random with replacement from the original training data, and classifiers are aggregated with equal weights into an ensemble vote [13]. In boosting, the resampling of training data for a classifier is adaptively modified to include the most misclassified samples more frequently, and the aggregation of classifiers is done by weighted voting [14]. Ensemble methods generally perform very well for classification problems where the number of features is much smaller than the number of samples [15]. For this case, it has been proven that having an ensemble of disagreeing committee members each trained on a subset of the samples should result in improved predictive performance compared to one classifier trained on all samples [16]. Hence, the benefit of ensemble classifiers stems from aggregating widely varying classifiers.

For prediction based on gene and protein expression data sets, the situation is different. If the number of samples is much smaller than the number of features, the improved performance expected by having an ensemble of disagreeing classifiers may be ruined by each classifier being too poor as a result of being trained on too few samples. Instead, one classifier trained using all training samples may provide better results. In this work, we evaluate if combining classifiers into ensembles, using an unweighted vote for predictions, results in improved performance for gene and protein expression data sets. We used a filter method for feature selection and two different classifiers, SVM [17] and nearest centroid classifiers (NCC) [3], both shown to work well combined with filter methods for high-dimensional data [8, 18–20]. We compared the performance of six different methods to construct classifiers, including both individual classifiers and classifiers aggregated into ensembles, using four publicly available data sets, three gene expression data sets and one proteomic data set.

Methods

Classifiers

We used NCC and SVM as classifiers, both individually and aggregated into ensembles.

For NCC, the centroid for each class was the vector of means for each feature. Unknown samples were evaluated by calculating the distance between its feature profile and and each class centroid using $1 -$ Pearson correlation as distance. Unknowns were assigned the class to which they were nearest. We did not shrink centroids as this does not seem to be important for classification of microarray data [20]. In ensembles the average distance to each centroid across classifiers was used for class assignments.

For SVM, we used the maximal margin classifier, that is SVM with no soft margin (C parameter set to infinity) and linear kernel. In ensembles the average distance from the decision hyperplane across classifiers was used for class assignments.

Classifier evaluation

External 3-fold cross-validation of all data was used to evaluate each classifier. The cross-validation was iterated 100 times so that each sample was a test sample 100 times and there was a total of 300 test sets.

For each test set the predictive performance was evaluated using balanced accuracy (BACC) and area under the receiver operating characteristic (AUC). BACC is the average of the sensitivity and specificity: the average of the number of correctly classified samples in each class. AUC corresponds to the probability that in a randomly chosen pair of samples, one from each class, the predictions for each sample is closest to the correct class. AUC complements BACC in the sense that BACC requires a decision regarding the class prediction for each sample, whereas AUC indicates the largest possible classification accuracy obtainable if an optimal decision based on the predictions could be found. Both measures are 50% for random predictors. The averages of BACC and AUC across the 300 test sets are presented.

To compare different methods to construct classifiers, we also ranked each construction method for each test set such that the best performing method got rank one. Methods were evaluated based on the average rank for the 300 test sets. We ranked NCC and SVM classifiers separately to high-light differences in classifier construction.

Feature selection

We used a filter based on a ranking criterion to select features. This feature selection consists of two parts. First the features are ranked based on their ability to individually discriminate between classes. It is our and others experience [8] that the most widely used ranking criteria perform very similarly. Therefore the choice of criterion is not crucial and we have used the signal-to-noise ratio (SNR) [1] to rank features. Second the number of top-ranked features to use is selected based on classification performance.

We used sets of features, where each set contained 1.5 times more top-ranked features than the previous set. The first set contained only the top-ranked feature and the final set contained all features. To select which set of features to use, we employed 3-fold cross-validation internal for the training samples and computed the predictive performance for each feature set. The number of features resulting in the best average BACC for ten complete cross-validation rounds (a total of 30 validation sets) was selected.

Often forward or backward filter selection procedures are used, in which one starts using one feature and increase the number of features, or starts using all features and decrease the number of features, respectively, until the performance deteriorates. We evaluate all feature sets employed. Hence, we use neither a forward nor a backward method.

For some gene expression data sets, it has been observed that using different subsets of samples results in large differences in which features are selected [21]. To get a potentially more robust ranking of features, we utilized the subsets of training samples from the internal cross-validation. In this consensus feature selection, features were ranked according to their median rank for the internal training samples.

Classifier construction

The only parameter values and other choices to optimize for the SVM and NCC classifiers we use are the number of features to employ. For each split into a training and test set from the external cross-validation, the optimal number of top-ranked features, n_g, to use was found using internal cross-validation of the training set as described in the previous section “Feature selection”. We optimized n_g separately for SVM and NCC. The internal 3-fold cross-validation of training data iterated 10 times resulted in 30 classifiers in ensembles.

The following six methods to construct a classifier were used.

Single classifier. Construct a single classifier using all features and all training data.

Ensemble of classifiers. Use internal cross-validation of training samples to construct an ensemble of classifiers in which each classifier uses its own internal training data for training but no feature selection (all features are used).

Single classifier with feature selection. Construct one classifier using all training data and the top $n_{\rm g}$ genes for this training data.

Ensemble of classifiers with individual feature selection. Use internal cross-validation of training samples to construct an ensemble of classifiers in which each classifier uses its own internal training data for training and the top $n_{\rm g}$ genes ranked based also on its internal training data.

Single classifier with consensus feature selection. Construct one classifier using all training data and the top $n_{\rm g}$ genes from a consensus gene list based on 3-fold internal cross-validation of all training data.

Ensemble of classifiers with consensus feature selection. Use internal cross-validation to construct an ensemble of classifiers in which each classifier uses its own internal training data, but the same genes (the top $n_{\rm g}$ genes from a consensus gene list based on the internal cross-validation of all training data.)

Data sets

We used four different publicly available data sets to evaluate different methods to construct classifiers. Three of the data sets were from gene expression profiling studies and one was from a mass spectrometry based proteomic study.

Leukemia. This data sets contains gene expression profiles of 72 samples from leukemia of two variants: 25 samples of acute myeloid leukemia (AML) and 47 samples of acute lymphoblastic leukemia (ALL) [1]. We used the quality filtering described for this data set by Dudoit *et al.* [12] to reduce the total of 7,129 features to 3,571 features used in our analysis.

Central nervous system (CNS) embryonal tumors. This data set contains gene expression profiles of samples from embryonal tumors of the central nervous system [4]. We used the subset of 60 samples for which outcome information after embryonic treatment of the CNS was available. Of the 60 samples, 21 represent survivors and 39 represent deaths. We used the quality filter described in the supplementary material of ref. [4] to reduce the total of 7,129 features to 4,459 features used in our analysis.

Breast cancer. This data set consists of gene expression profiles of samples from breast tumors [3]. We used the subset of 97 samples from sporadic tumors consisting of 51 samples from patients with a good outcome and 46 from patients with a poor outcome. We required each feature to have at least six samples with a maximal p value, from the Rosetta error model [22], of 0.01. This quality filter reduced the total number of features (24,481) to 8,472 features used in our analysis.

Liver cancer. This data set consists of SELDI-TOF mass spectrometric profiles of peptides and proteins in a total of 411 sera samples from 199 hepatocellular carcinoma patients and 212 healthy individuals [23]. Each mass spectra in the data set consisted of $\approx 340,000$ m/z values with corresponding ion intensities. We used spectra pre-processed according to the low-level analysis described in ref. [23]. This pre-processing reduced the number of features to 368.

Results and Discussion

Leukemia data

The results of predictions for the six different ways to construct classifiers are presented in Table 1. For both SVM and NCC, the best ranked method found was an ensemble classifier with no feature selection. These two methods obtained similar average BACCs for the test sets: 97.2% and 97.3%, respectively. For NCC without feature selection, the BACC was larger for the ensemble classifier than for the single classifier for 14 of the 300 test sets, whereas the single classifier never obtained a larger BACC than the ensemble classifier. For SVM without feature selection, the corresponding numbers were 27 and 0, respectively. Hence, while the differences for these two construction methods were small and they often tied, we note that the single classifiers never performed better than the corresponding ensemble classifiers. Similarly, we note that all three NCC and all three SVM ensemble methods were ranked better than their respective corresponding single classifier.

To explore, why filter selection did not improve predictions, we investigated the number of features selected for each test set (Fig. 1). We made three observations. First, selecting all features was the most common choice. Second, a large variation in the number of selected features across test sets was observed for both methods. Finally, SVM tended to select more features than NCC. The second observation means that different subsets of samples not only

Table 1: Comparison of methods to construct classifiers for the leukemia data.

			Validation				Test				
Predictor	Filter	Ensemble	BACC(%)		AUC(%)		BACC(%)		AUC(%)		Rank[a]
			Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean
NCC	None	No	-	-	-	-	97.1	3.0	99.4	1.1	2.94
	None	Yes	97.3	1.7	99.3	0.7	97.2	2.9	99.4	1.1	2.81
	Individual	No	-	-	-	-	95.8	3.4	99.5	1.0	3.86
	Individual	Yes	97.4	1.8	99.5	0.6	95.9	3.5	99.5	1.0	3.75
	Consensus	No	-	-	-	-	95.8	3.4	99.5	1.0	3.83
	Consensus	Yes	97.8	1.7	99.6	0.6	95.9	3.4	99.5	1.0	3.81
SVM	None	No	-	-	-	-	97.0	3.1	99.5	0.9	3.47
	None	Yes	97.1	2.2	99.5	0.5	97.3	2.9	99.5	0.9	3.22
	Individual	No	-	-	-	-	96.6	3.5	99.4	1.0	3.70
	Individual	Yes	97.0	3.0	99.3	3.0	96.5	6.8	99.1	5.7	3.44
	Consensus	No	-	-	-	-	96.6	3.5	99.4	1.0	3.68
	Consensus	Yes	97.4	2.1	99.6	0.4	96.9	3.3	99.5	0.9	3.47

[a] NCC and SVM were ranked separately.

results in different and equally performing rankings of features as found by Ein-Dor *et al.* [21], but also results in different numbers of features selected when optimizing supervised classifiers. This observation suggests that it is difficult to optimize the number of features to use based on internal cross-validation of training data, as it is not likely to perform as good on an independent test set. In agreement, we observed systematically better and competitive results for the validation data sets as compared to the test data sets: optimizing the number of selected features resulted in over-fitting (Table 1).

Comparing with other predictions of this data set, we note that Wessels *et al.* found that using the dimensional reduction method partial least squares (PLS) performed better than feature selection using forward filtering based on SNR [8]. Our performance using all features is similar to the performance obtained using PLS. Our results indicate that to obtain a highly competitive performance for this data set the choice of classifier is not crucial if all features are used. It has also been observed for other gene expression data sets that SVM classifiers perform best when all features are used [24, 25].

CNS embryonal tumor data

The results for the CNS embryonal tumor data set are presented in Table 2. For NCC, the best ranked classifier was an ensemble with individual feature selection, for which a BACC of 60.6% was obtained. This classifier performed better for 136 and worse for 81 test sets when compared with its corresponding single classifier. For SVM, the best ranked classifier was a single classifier with no feature selection, which performed better than the NCC classifiers and a BACC of 63.0% was obtained. This classifier was similarly ranked as its corresponding ensemble classifier, and performed better for 102 and worse for 99 test sets.

For the leukemia data set performances close to a 100% were obtained, making it difficult to compare predictive performances for the test sets with the potentially overly optimistic estimates from the validation sets. For the CNS embryonic tumors the predictive performances were much worse, making comparisons between test and validation results more illustrative. We made three observations both for NCC and SVM.

First, with no feature selection the validation result was worse than the test result. Here, there is no feature selection and no optimization of classifiers and the validation result is an honest estimate of the predictive performance. However, in the internal cross-validation each sample is classified by an ensemble of the 10 classifiers for which it was not used in training, whereas the test samples from the external cross-validation are classified by an ensemble of all 30 classifiers from the internal cross-validation. Apparently, the larger ensembles perform better for this data set.

Second, with individual feature selection the validation results are overly optimistic estimates of the predictive performance. Here, the only dishonest as-

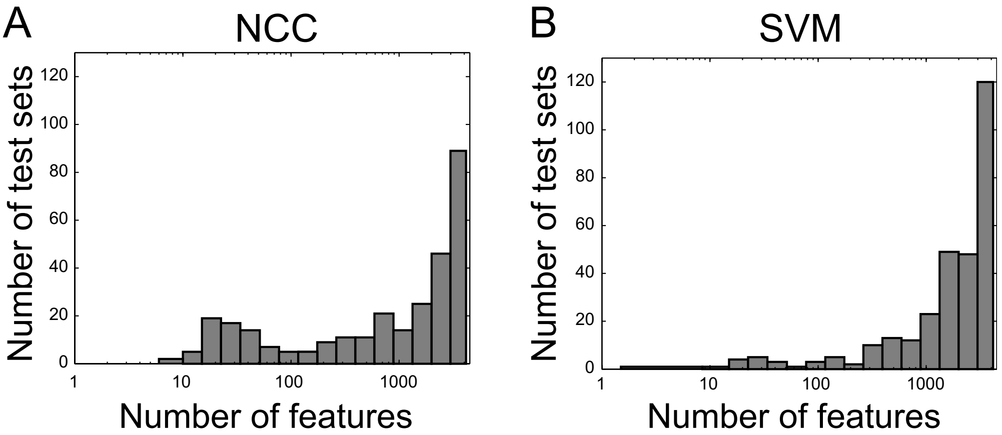

Figure 1: The optimal number of features selected for each test set for the leukemia data. There was a total of 300 test sets and 3,571 features. A) NCC. Median number of selected features was 1598 and B) SVM. median number of selected features was 2397.

pect of the validation performance is that the number of features selected has been optimized to give the best performance. Hence, even though features are ranked individually for each classifier based only on its training samples, an overly optimistic estimate was obtained.

Third, with consensus feature selection the validation results are even more optimistic than for individual feature selection. Here, there is a dishonest use of the class of the validation samples in the internal cross-validation because all internal samples have been used to rank features. Using validation samples to rank features may not only result in overly optimistic results but may also inflate performance for classes which can not be classified, leading to incorrect conclusions [6].

In the original analysis of this data set [4], Pomeroy *et al.* used k-nearest neighbor classifiers and evaluated the predictive performance using leave-one-out cross-validation. Both the number of neighbors, k, and the selected number of features were optimized in the cross-validation. This use of the validation samples in classifier optimization resulted in an overall classification accuracy of 78%, not likely to be obtainable when using an independent test set.

As for the leukemia data, we note that for SVM no feature selection performed best. The BACC of this classifier (63.0%) was also higher than for all classifiers evaluated in ref. [8], where the best BACC obtained was 61.3%. In ref. [8], SVM obtained the best result when combined with recursive feature elimination. This combination obtained a BACC of 60.1% with on average 1235 features selected. SVM combined with forward filtering selected fewer features, on average 120, and performed worse: 57.6% BACC. SVM combined with our filtering method selected roughly as many features (on average 1,655) as recursive feature elimination and performed similarly. Together, these findings show a sensitivity to minor details in the combination of classifiers and feature selection methods and that forward filtering may find local maxima in performance.

Breast cancer data

The results for the breast cancer data set are presented in Table 3. For NCC, the best ranked classifier was an ensemble with consensus feature selection, for which a 66.4% BACC was obtained. All four NCC classifiers with feature selection achieved similar results. For SVM, the best ranked classifier was an ensemble classifier with individual feature selection, which performed slightly worse than the NCC classifiers and a BACC of 66.0% was obtained. This SVM classifier performed better for 174 and worse for 94 test sets compared to its corresponding single classifier. To our knowledge, there is no comparable study for this data set, but our results are in agreement with previous studies of variants of this data set [8, 26].

Table 2: Comparison of methods to construct classifiers for the CNS embryonal tumor data.

			Validation				Test				
Predictor	Filter	Ensemble	BACC(%)		AUC(%)		BACC(%)		AUC(%)		Rank[a]
			Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean
NCC	None	No	-	-	-	-	58.6	9.4	64.2	10.4	3.81
	None	Yes	58.4	6.3	59.2	8.5	59.5	9.8	64.2	10.4	3.53
	Individual	No	-	-	-	-	59.5	10.1	63.5	10.8	3.44
	Individual	Yes	64.3	6.8	66.6	8.8	60.6	10.3	65.7	11.8	2.96
	Consensus	No	-	-	-	-	58.9	10.2	63.2	10.8	3.67
	Consensus	Yes	72.6	7.6	78.4	8.9	59.0	10.0	63.2	10.8	3.57
SVM	None	No	-	-	-	-	63.0	10.3	68.4	10.8	3.08
	None	Yes	61.4	8.9	68.0	9.3	62.3	9.1	69.0	10.8	3.14
	Individual	No	-	-	-	-	59.8	11.0	63.6	11.9	3.97
	Individual	Yes	64.9	8.6	70.0	9.0	62.2	9.4	66.9	11.9	3.21
	Consensus	No	-	-	-	-	60.1	10.1	63.3	11.7	3.94
	Consensus	Yes	73.4	7.9	82.0	8.3	60.7	9.8	64.8	11.5	3.66

[a] NCC and SVM were ranked separately.

Table 3: Comparison of methods to construct classifiers for the breast cancer data.

			Validation				Test				
Predictor	Filter	Ensemble	BACC(%)		AUC(%)		BACC(%)		AUC(%)		Rank[a]
			Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean
NCC	None	No	-	-	-	-	65.0	7.2	74.9	7.2	3.96
	None	Yes	65.2	4.1	72.8	4.7	65.0	7.2	74.9	7.2	3.95
	Individual	No	-	-	-	-	66.1	7.0	74.5	7.0	3.28
	Individual	Yes	68.7	4.4	75.8	4.7	66.2	7.4	76.0	7.1	3.39
	Consensus	No	-	-	-	-	66.3	7.0	74.5	7.0	3.22
	Consensus	Yes	79.1	4.4	86.8	4.1	66.4	7.0	74.5	6.9	3.19
SVM	None	No	-	-	-	-	65.1	6.9	70.5	7.4	3.61
	None	Yes	64.0	5.0	70.5	7.4	65.3	6.8	71.2	7.5	3.61
	Individual	No	-	-	-	-	64.0	14.2	67.9	9.4	3.81
	Individual	Yes	67.7	6.6	72.1	8.2	66.0	8.7	70.6	9.9	3.07
	Consensus	No	-	-	-	-	64.4	8.4	68.3	9.5	3.70
	Consensus	Yes	77.8	7.7.	86.0	8.2	65.7	7.8	70.2	7.9	3.20

[a] NCC and SVM were ranked separately.

Liver cancer data

The results for the liver cancer data set are presented in Table 4. For NCC, the best ranked classifier was an ensemble with individual feature selection, for which a 77.0% BACC was obtained. Even though the best ranked classifier performed better for 61 and worse for 39 test sets compared to its corresponding single classifier, the performance for each test set was typically very similar, and all four NCC classifiers with feature selection achieved almost identical BACC. For SVM, the best ranked classifier was also an ensemble with individual feature selection, for which a 91.3% BACC was obtained. This classifier performed better for 238 and worse for 48 test sets compared to its corresponding single classifier. Moreover, it was the best classifier for most of the 300 test sets, as seen from its average rank being close to one. It also outperformed all NCC classifiers.

Ressom *et al.* obtained a BACC of ≈91.5% for this data set when using SVM combined with particle swarm optimization for feature selection [23]. We obtained a comparable BACC using filtering based on SNR for feature selection, indicating that the choice of feature selection method is not crucial.

Table 4: Comparison of methods to construct classifiers for the liver cancer data.

			Validation				Test				
Predictor	Filter	Ensemble	BACC(%)		AUC(%)		BACC(%)		AUC(%)		Rank[a]
			Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean
NCC	None	No	-	-	-	-	76.4	3.3	84.9	3.0	3.67
	None	Yes	76.4	1.7	84.6	1.5	76.4	3.3	84.9	2.8	3.63
	Individual	No	-	-	-	-	77.0	2.9	84.8	2.8	3.47
	Individual	Yes	78.0	1.4	84.7	1.5	77.0	2.8	84.8	2.8	3.31
	Consensus	No	-	-	-	-	77.0	2.9	84.8	2.8	3.48
	Consensus	Yes	78.0	1.5	84.8	1.4	77.0	2.9	84.8	2.8	3.44
SVM	None	No	-	-	-	-	75.3	6.8	83.4	7.2	5.45
	None	Yes	85.3	1.8	92.5	1.2	86.7	3.0	93.7	2.1	3.40
	Individual	No	-	-	-	-	89.6	2.5	95.8	1.5	2.23
	Individual	Yes	91.1	1.4	96.7	0.8	91.3	2.3	96.7	1.3	1.24
	Consensus	No	-	-	-	-	76.1	6.0	84.4	5.8	5.45
	Consensus	Yes	86.0	1.0	93.0	1.3	87.0	2.9	94.0	1.9	3.25

[a] NCC and SVM were ranked separately.

Conclusions

We have investigated if aggregating classifiers into ensembles improves classification performance for gene and protein expression data sets, for which the number of features typically is much larger than the number of samples. The general conclusions may be summarized as follows:

- Ensemble methods performed best, even though differences in terms of predictive accuracies often were relatively small. For NCC, an ensemble method performed best for all four data sets. For SVM, an ensemble method performed best for three data sets.
- Even minimal dishonest use of test samples, such as optimizing only the number of features to use based on predictive performance of test samples, may result in overly optimistic estimates of predictive performance.
- If the goal is to obtain good predictive performance regardless if very many features are used, SVM with no feature selection often performs very well.
- Forward filtering may find classifiers that perform well using small feature sets, however, better performance is often obtained using larger feature sets.

The performance of classifiers can potentially be improved in many ways. For example, various approaches to weight the classifiers in the ensembles can be explored. We have used ensembles of size 30, and our results indicate that smaller ensembles perform worse. There is a trade-off between ensemble size and ensemble construction time. Therefore, it may be worthwhile to investigate the dependence of performance on ensemble size.

Acknowledgments

This work was in part funded by the Swedish Cancer Society.

References

1. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES: **Molecular classification of cancer: class discovery and class prediction by gene expression monitoring**. *Science* 1999, **286**(5439):531–537.
2. Khan J, Wei JS, Ringnér M, Saal LH, Ladanyi M, Westermann F, Berthold F, Schwab M, Antonescu CR, Peterson C, Meltzer PS: **Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks**. *Nat Med* 2001, **7**(6):673–679.
3. van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AAM, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH: **Gene expression profiling predicts clinical outcome of breast cancer**. *Nature* 2002, **415**(6871):530–536.
4. Pomeroy SL, Tamayo P, Gaasenbeek M, Sturla LM, Angelo M, McLaughlin ME, Kim JYH, Goumnerova LC, Black PM, Lau C, Allen JC, Zagzag D, Olson JM, Curran T, Wetmore C, Biegel JA, Poggio T, Mukherjee S,

Rifkin R, Califano A, Stolovitzky G, Louis DN, Mesirov JP, Lander ES, Golub TR: **Prediction of central nervous system embryonal tumour outcome based on gene expression**. *Nature* 2002, **415**(6870):436–442.

5. Petricoin EFr, Ornstein DK, Paweletz CP, Ardekani A, Hackett PS, Hitt BA, Velassco A, Trucco C, Wiegand L, Wood K, Simone CB, Levine PJ, Linehan WM, Emmert-Buck MR, Steinberg SM, Kohn EC, Liotta LA: **Serum proteomic patterns for detection of prostate cancer**. *J Natl Cancer Inst* 2002, **94**(20):1576–1578.

6. Simon R, Radmacher MD, Dobbin K, McShane LM: **Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification**. *J Natl Cancer Inst* 2003, **95**:14–18.

7. Gruvberger-Saal SK, Edén P, Ringnér M, Baldetorp B, Chebil G, Borg Å, Fernö M, Peterson C, Meltzer PS: **Predicting continuous values of prognostic markers in breast cancer from microarray gene expression profiles**. *Mol Cancer Ther* 2004, **3**(2):161–168.

8. Wessels LFA, Reinders MJT, Hart AAM, Veenman CJ, Dai H, He YD, van't Veer LJ: **A protocol for building and evaluating predictors of disease state based on microarray data**. *Bioinformatics* 2005, **21**(19):3755–3762.

9. Gruvberger S, Ringnér M, Chen Y, Panavally S, Saal LH, Ferno M, Peterson C, Meltzer PS: **Estrogen receptor status in breast cancer is associated with remarkably distinct gene expression patterns**. *Cancer Res* 2001, **61**(16):5979–5984.

10. Tan AC, Gilbert D: **Ensemble machine learning on gene expression data for cancer classification**. *Appl Bioinformatics* 2003, **2**(3 Suppl):75–83.

11. Dettling M: **BagBoosting for tumor classification with gene expression data**. *Bioinformatics* 2004, **20**(18):3583–3593.

12. Dudoit S, Fridlyand J, Speed TP: **Comparison of discrimination methods for the classification of tumors using gene expression data**. *JASA* 2002, **97**:77–87.

13. Breiman L: **Bagging predictors**. *Mach Learn* 1996, **24**:123–140.

14. Freund Y, Shapire RE: **A descision-theoretic generalization of online learning and an application to boosting**. *J Comput Syst Sci* 1997, **55**:119–139.

15. Opitz D, Maclin R: **Popular ensemble methods: an empirical study**. *Journal of Artificial Intelligence Research* 1999, **11**:169–198.

16. Krogh A, Vedelsby J: **Neural network ensembles, cross validation, and active learning**. In *Advances in Neural Information Processing Systems*, *Volume 2*. Edited by Tesauro G, Touretzky D, Leen T, San Mateo, CA: Morgan Kaufman 1995:650–659.

17. Vapnik V: *The nature of statistical learning theory*. Springer Verlag 1995.

18. Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, Haussler D: **Support vector machine classification and validation of cancer tissue samples using microarray expression data**. *Bioinformatics* 2000, **16**(10):906–914.

19. Tibshirani R, Hastie T, Narasimhan B, Chu G: **Diagnosis of multiple cancer types by shrunken centroids of gene expression**. *Proc Natl Acad Sci U S A* 2002, **99**(10):6567–6572.

20. Dabney AR: **Classification of microarrays to nearest centroids**. *Bioinformatics* 2005, **21**(22):4148–4154.

21. Ein-Dor L, Kela I, Getz G, Givol D, Domany E: **Outcome signature genes in breast cancer: is there a unique set?** *Bioinformatics* 2005, **21**(2):171–178.

22. Roberts CJ, Nelson B, Marton MJ, Stoughton R, Meyer MR, Bennett HA, He YD, Dai H, Walker WL, Hughes TR, Tyers M, Boone C, Friend SH: **Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles**. *Science* 2000, **287**(5454):873–880.

23. Ressom HW, Varghese RS, Abdel-Hamid M, Eissa SAL, Saha D, Goldman L, Petricoin EF, Conrads TP, Veenstra TD, Loffredo CA, Goldman R: **Analysis of mass spectral serum profiles for biomarker selection**. *Bioinformatics* 2005, **21**(21):4039–4045.

24. Ramaswamy S, Tamayo P, Rifkin R, Mukherjee S, Yeang CH, Angelo M, Ladd C, Reich M, Latulippe E, Mesirov JP, Poggio T, Gerald W, Loda M, Lander ES, Golub TR: **Multiclass cancer diagnosis using tumor gene expression signatures**. *Proc Natl Acad Sci U S A* 2001, **98**(26):15149–15154.

25. Pavey S, Johansson P, Packer L, Taylor J, Stark M, Pollock PM, Walker GJ, Boyle GM, Harper U, Cozzi SJ, Hansen K, Yudt L, Schmidt C, Hersey P, Ellem KAO, O'Rourke MGE, Parsons PG, Meltzer P, Ringnér M, Hayward NK: **Microarray expression profiling in melanoma reveals a BRAF mutation signature**. *Oncogene* 2004, **23**(23):4060–4067.

26. Michiels S, Koscielny S, Hill C: **Prediction of cancer outcome with microarrays: a multiple random validation strategy**. *Lancet* 2005, **365**(9458):488–492.