Overproduce-and-Select: The Grim Reality

Ulf Johansson
School of Business and IT
University of Borås
Sweden
Email: ulf.johansson@hb.se

Tuve Löfström
School of Business and IT
University of Borås
Sweden
Email: tuve.lofstrom@hb.se

Henrik Boström
Department of Computer and System Sciences
Stockholm University
Sweden
Email: henrik.bostrom@dsv.su.se

***Abstract*—Overproduce-and-select (OPAS) is a frequently used paradigm for building ensembles. In static OPAS, a large number of base classifiers are trained, before a subset of the available models is selected to be combined into the final ensemble. In general, the selected classifiers are supposed to be accurate and diverse for the OPAS strategy to result in highly accurate ensembles, but exactly how this is enforced in the selection process is not obvious. Most often, either individual models or ensembles are evaluated, using some performance metric, on available and labeled data. Naturally, the underlying assumption is that an observed advantage for the models (or the resulting ensemble) will carry over to test data. In the experimental study, a typical static OPAS scenario, using a pool of artificial neural networks and a number of very natural and frequently used performance measures, is evaluated on 22 publicly available data sets. The discouraging result is that although a fairly large proportion of the ensembles obtained higher test set accuracies, compared to using the entire pool as the ensemble, none of the selection criteria could be used to identify these highly accurate ensembles. Despite only investigating a specific scenario, we argue that the settings used are typical for static OPAS, thus making the results general enough to question the entire paradigm.**

***Index Terms*—Overproduce-and-select, ensembles, neural networks.**

I. Introduction

In predictive classification a target function f, mapping each instance $\mathbf{x}$, to one of the predefined class labels y is learnt. The class label y is a discrete attribute, restricted to values in a predefined set $\{c_1, \ldots, c_n\}$. When using machine learning techniques for predictive classification, the algorithm uses a set of training instances, each consisting of an input vector $\mathbf{x}_i$ and a corresponding target class label y_i to learn the function $y = f(\mathbf{x}; \theta)$. During training, the parameters θ, which are specific for each technique, are optimized, based on a score function. When sufficiently trained, the predictive model is able to accurately predict a value $\hat{y}$, when presented with a novel (test) instance $\mathbf{x}_j$.

Within machine learning, it is well established that combining several individual classifiers into *ensembles* results in improved predictive performance, compared to single models, see e.g., [1] and [2]. An ensemble aggregates multiple classifiers (called *base classifiers*) into a composite model, making the ensemble prediction a function of all the included base classifiers.

When the base classifiers $\mathbf{H} = \{h_1, \ldots, h_m\}$ are trained on the training set, the output (prediction) of a base classifier h_j on instance $\mathbf{x}_i$ is $h_j(\mathbf{x}_i)$. Although this prediction is, in the general case, a vector of size n consisting of probability estimates for each class. Often, however, each base classifier simply votes for one specific class, i.e., returns the class label associated with the maximum belief. If all classifier weights are equal, the procedure is referred to as *majority voting* when the base classifiers output class labels, and *averaging* when they output probability distributions.

The most intuitive explanation for why ensembles work is that the aggregation of several models, using averaging or majority voting, will eliminate uncorrelated base classifier errors; see e.g., [3]. Consequently, ensemble accuracy will be higher than mean base classifier accuracy, as long as the base classifiers commit their errors on different instances. Ideally, if base classifiers mistakes are independent, the ensemble accuracy can be pushed to an arbitrarily high level, just by adding more base classifiers. Informally, the key term *ensemble diversity* therefore measures and describes how base classifier mistakes are distributed over the instances.

In [4], Brown et al. introduced a taxonomy of methods for creating diversity. The first obvious distinction made is between *explicit* methods, where some diversity metric is directly optimized, and *implicit* methods, where the method is likely to produce diversity without actually targeting it. A number of implicit methods produce diversity by supplying each classifier with a slightly different training set. Standard *bagging* [5], produces diversity by using resampling to create different training sets for each base classifier. More specifically, each training set (called a *bootstrap*), has the same size as the original training set, but since the instances are randomly selected with replacement, a training set will contain multiple copies of some instances while lacking others completely. On average, approximately 63% of the orginal instances are present in each bootstrap.

In contrast to, for instance, *random forests* [6], most dedicated ensemble techniques utilizing artificial neural networks (ANNs) are quite complicated, often explicitly optimizing some diversity metric. This is despite the fact that solid empirical, as well as theoretical, validation of the explicit algorithms, are absent from the field [7]. In fact, the current situation can be summarized like this: Diversity is obviously beneficial for ensembles, but it is very questionable if any suggested diversity measure is actually useful as part of an optimization function. Consequently, implicit methods have

978-1-4673-5853-8/13/$31.00

recently regained interest from the research community.

One very general ensemble paradigm, often advocated in different variations, is *overproduce-and-select* (OPAS), see e.g., [8]. In OPAS, a large number of base classifiers (often referred to as a *pool*) is first trained, before a subset of the available models are somehow selected to be combined into the final ensemble. Most often, included classifiers are supposed to be accurate and diverse for the OPAS strategy to work as intended; i.e., producing ensembles that are more accurate than just combining all base classifiers in the pool. For the training, any suitable technique can be used, as long as it produces at least fairly accurate and diverse base classifiers. Standard bagging, as described above, is often the method of choice.

In *static* (or *global*) selection strategies, the models that will form the ensemble are chosen, once and for all, before the ensemble is applied to all test instances. In *dynamic* selection strategies, different ensembles are formed for each test instance, most often based on locality, see e.g., [9].

In addition, even if the classifiers are selected globally, the votes from the base classifiers may be weighted differently, for each test instance, again typically based on locality. This study focuses solely on static approaches where standard majority voting without any weighting is used.

The most interesting part of the OPAS paradigm is, however, how the actual selection is performed. Or, put in another way, if the ambition is to use accurate but diverse models, how do we find these models? The most obvious method is to base the selection on some metric which can be measured on available (training) data. Naturally, the underlying assumption is that an observed advantage for the models (or the resulting ensemble) will carry over to the test set. Unfortunately, this supposedly straightforward procedure involves a number of subtleties. First of all, we need some data to do the evaluation on. The probably most common way is to set aside some data, a *hold out* or *validation* set, explicitly for this purpose. This, however, will result in that not all the available labeled data can be used for the training of the models, potentially leading to weaker base classifiers. Another option is to use results on the training data, but since this data has already been used for the training, these results are no longer unbiased. Finally, when using bagging, there is another option which is to use the *out-of-bag* (OOB) data. When the training set for a model is drawn by sampling with replacement, about $1/3$ of the instances are left out of the sample, and these instances are said to be out-of-bag for that model. When using OOB-estimates for an ensemble, only the models for which a specific instance is out-of-bag are allowed to vote when the ensemble is evaluated on that instance. Since only approximately $1/3$ of the models have each instance out-of-bag, the voting ensembles are much smaller than the ensemble that is actually being evaluated, i.e., ensemble OOB accuracy normally underestimates ensemble test set accuracy [10].

Once the selection criterion has been chosen, there are a number of more or less sophisticated approaches to find the optimal ensemble, based on that criterion. Most approaches use some kind of search procedure. One option is to search globally and directly for a subset of the available models, using for instance a genetic algorithm. Another option is to start out with either all models in the ensemble or an empty ensemble, before iteratively searching for a model to delete or add, respectively. No matter what search procedure that is used, OPAS can only be considered successful if the selected ensemble is better than most other candidate ensembles when evaluated on the test data. Specifically, it must of course be better than just using all the models in the pool as the ensemble.

In this study, we will not evaluate a specific OPAS algorithm, but instead investigate its most basic assumption; i.e., that it is indeed possible to select a superior ensemble based on performance that can be measured using available data.

II. Background

Diversity measures are often divided into pairwise and non-pairwise measures. Pairwise measures calculate the average of a particular distance metric between all possible pairings of classifiers in the ensemble, while non-pairwise measures typically use some variation of entropy, or calculate the correlation between each ensemble member and the averaged output. In this study, we will use two pairwise measures and one non-pairwise.

It must be noted that all diversity measures are in fact calculated on what is sometimes called an *oracle output matrix*, i.e., the correct target values are assumed to be known. Let the (oracle) output of each classifier D_i be represented as an N-dimensional binary vector y_i, where $y_{j,i} = 1$ if D_i correctly recognizes instance z_j and 0 otherwise. Let N^{ab} mean the number of instances for which $y_{j,i} = a$ and $y_{j,k} = b$. As an example, N^{11} is the number of instances correctly classified by both classifiers. N is the total number of instances.

The probably most intuitive diversity measure is the *disagreement* measure, which is the ratio between the number of instances on which one classifier is correct and the other incorrect to the total number of instances:

$$Dis_{i,k} = \frac{N^{10} + N^{01}}{N} \tag{1}$$

To find the diversity of a specific ensemble consisting of L classifiers, the averaged Dis over all pairs of classifiers is calculated:

$$Dis = \frac{2}{L(L-1)} \sum_{i=1}^{L-1} \sum_{k=i+1}^{L} Dis_{i,k} \tag{2}$$

Naturally a higher disagreement value implies a larger diversity. The *double-fault* measure was proposed in [11] and is the proportion of instances misclassified by both classifiers:

$$DF_{i,k} = \frac{N^{00}}{N} \tag{3}$$

To calculate the double fault diversity for an ensemble, the double fault values are averaged over all pairs of classifiers,

identically to disagreement. For double fault, a lower value indicates higher diversity. Comparing double fault and disagreement, the main difference is that when using disagreement, N^{00} and N^{11} are treated equally, while the double fault measure does not "punish" pairs of classifiers where both are correct. In a setting where the purpose is to obtain accurate ensembles, this may appear to be a strong argument for double fault, but it should be noted that, as a consequence, double fault will be positively correlated with the base classifier accuracy.

The non-pairwise *difficulty* measure was introduced by Hansen and Salomon in [12]. Let X be a random variable taking values in $\{0/L, 1/L, \ldots, 1\}$. X is defined as the proportion of classifiers that correctly classify an instance x drawn randomly from the data set. To estimate X, all L classifiers are run on the data set. Ensemble difficulty is then defined as the variance of X. For difficulty, lower values mean higher diversity, with the explanation that for a diverse classifier ensemble, every instance can at least be classified correctly by a portion of all the base classifiers, which is likely to result in a lower variance. The opposite would mean that all base classifiers are correct on some instances and wrong on some other instances, which of course would lead to higher variance.

Despite the fact that both intuition and a strong theoretical foundation advocate the benefit of diverse base classifiers, none of the suggested diversity measures is proven superior to the others. As a matter of fact, when Kuncheva and Whitaker studied ten statistics measuring diversity using oracle outputs, i.e., correct or incorrect vote for the class label, all diversity measures evaluated showed low or very low correlation with test set accuracy, see [13]. In [14], Saitta supports Kuncheva's negative view, as presented in a series of papers, but she also goes one step further and shows not only that no useful measure exists today, but also that it is unlikely that one will ever exist.

A. Related work

Several studies simply try to maximize the diversity, typically on either training data or a hold out set, when selecting the models for the ensemble. Giacinto and Roli [15] use a search method intended to maximize the diversity measured using double fault. Classifiers are added until the desired number of ensemble members is reached. Similarly, Margineantu and Dietterich [16] use another diversity measure (not evaluated in this study), called kappa, to select ensemble members. Here too, the ensemble is built by iteratively selecting pairs with highest diversity until the desired ensemble size is reached.

In another study, Giacinto and Roli [17] again utilize the double fault diversity measure, but now the double fault matrix is regarded as a distance matrix, and used for clustering. The idea is that models in the same cluster will make identical mistakes, while members of different clusters should make different mistakes. The algorithm starts out with each model as a cluster of its own, and then, in each step, the most accurate single model is first chosen from each cluster, before the two least diverse clusters are merged. This is repeated until all classifiers are in the same cluster, so this algorithm will produce one ensemble of every possible size. Finally, one of these ensembles is picked based on accuracy measured on a validation set.

There are also a number of approaches where accuracy and diversity are optimized simultaneously. Zenobi and Cunningham [18] used a diversity measure and the error rate to guide a hill-climbing search method. Tremblay et al. [19] employed a multi-objective genetic algorithm guided by objective functions composed of the error rate with four different diversity measures. Dos Santos et al. [20] optimize a combination of ensemble error rate and diversity measures too, but they also include the ensemble size as a criterion in their analysis of the Pareto front.

Elghazel et al. [21] applies the overproduce-and-select paradigm to Random Forests in their algorithm called *Fitselect*. Fitselect uses a greedy search to select the optimal subset and directly balances accuracy and diversity on a validation set using a tradeoff variable. To combat the risk for overfitting, they also wrap cross-validation around the ensemble selection to maximize the amount of validation data considering, in turn, each fold as a validation fold.

There are also a number of previous studies evaluating the use of specific selection criteria, typically different diversity measures, in the OPAS paradigm. Specifically, Ruta and Gabrys [22], investigated the practical applicability of diversity measures in the context of combining classifiers by majority voting. Here, a number of search algorithms were used together with a number of selection criteria, including majority voting error and various diversity measures. The overall result was that a direct combiner error based search outperformed the use of the diversity measures as selection criteria. These findings are of course in accordance with the more general studies establishing the low correlation between diversity measures and ensemble performance. Two examples are the previously mentioned Kuncheva and Whitaker study [13], where mostly artificial classifiers were used, and a previous study of ours, where these results were confirmed using real-world classifiers [23].

III. Method

As described in the introduction, the overall purpose is to evaluate the basic OPAS paradigm, where a subset of the models in a pool is selected based on some specific but global property. In this study, a number of ANNs are first trained using standard bagging, and then the actual ensembles are selected using some performance metric, which of course must be possible to measure on the available data. In all experiments, exactly 100 ANNs are used as the classifier pool, and the performance metrics are *ensemble accuracy*, *mean base classifier accuracy* and the three diversity measures *disagreement*, *difficulty* and *double fault*.

In the first experiment, the ensemble selection is based on performance on a hold-out set (a validation set). As described in the background, this is the most common basic procedure.

In the second experiment, two different options that actually allow the use of all available training data for the ANN training are evaluated; i.e., measuring the performance on the training data or on out-of-bag data. In both these experiments, all the evaluated smaller ensembles have an identical size of 51 ANNs. We believe that this is a fairly typical OPAS scenario, picking substantially larger ensembles would mean smaller differences, both between the selected ensembles and compared to using all available models. If selecting significantly smaller ensembles, on the other hand, the expected result is that (almost) all of the selected ensembles would be outperformed by using all available models. Nevertheless, to investigate this, we include a third experiment, where the selected ensembles consist of just 7 ANNs.

Before the experimentation, each data set was preprocessed in the following way: first all missing numerical values were replaced with the mean value of that specific attribute, while missing categorical values were replaced with the mode values. Secondly, all categorical attributes were converted into binary numerical attributes. All ANNs in the study are fully-connected MLPs with one hidden layer, utilizing a localist coding, i.e., the number of output units is equal to the number of classes, and the training used is the resilient backpropagation (rprop) learning algorithm. The base classifiers are always combined using majority vote. Every ANN in the pool is trained exactly 150 epochs, and since standard bagging is used, each ANN is trained independently from all others in the pool.

It is well-known that there is no general rule-of-thumb that will always find an optimal, or even acceptable, number of hidden units in an MLP, based on the data set characteristics. In practice, some kind of internal cross-validation is instead often used to determine the number of hidden units. Nevertheless, there are many rules-of-thumb proposed, and most of them suggest that the number of hidden units should be somewhere between the number of input units and the number of output units, thus resulting in pyramid shaped networks. In this study, the number of hidden units h, is calculated using (4):

$$h = \left\lfloor \frac{\#attributes + \#classes}{2} \right\rfloor \tag{4}$$

During experimentation, the data sets were randomly split in 75% training and 25% testing. In Experiment 1, the size of the extra hold-out set was set to 15% (of the entire data set), and it was of course taken from the training data. In each run, 100 smaller ensembles were formed by randomly picking ANNs from the pool. All runs were repeated ten times for each data set, so reported results are averaged over the ten runs. The 22 data sets used are all publicly available from either the UCI repository [24] or the PROMISE Software Engineering Repository [25].

A. Evaluation

The actual evaluation is divided into two parts for each experiment. In the first part, we start by investigating global properties of the pool and the smaller ensembles. First of all, it is very interesting to see if there are any smaller ensembles that actually outperform the use of all available models. If this is not the case, the entire prospect of using overproduce-and-select, at least in this specific setting, is of course futile. It is also important to ensure that there are detectable differences in the performance, if there are only marginal differences (either on training/validation/OOB data or on the test data) it becomes more or less impossible for the selection process to succeed. Finally, another high-level analysis, frequently used in similar studies, is to investigate how different performance measures (on available data) correlate with ensemble test accuracy.

In the second part, we compare ensembles that rank high on a certain performance metric against ensembles that rank low on the same metric. More specifically, for each performance metric and data set (e.g., ensemble accuracy on validation data), the smaller ensembles were first sorted, i.e., ranked according to that measure. In the analysis, we then make two different comparisons; first we compare the best quartile (the ensembles ranked in the top quartile) against the worst, and then the single best ensemble to using all available models. These comparisons are, of course, carried out on the test data. Naturally, the overall purpose is to investigate whether ranking high on any specific performance measure obtained on available data actually transfers to also ranking high on ensemble test accuracy. The first comparison will tell us if there is a substantial (or even detectable) difference in ensemble test accuracy when comparing ensembles that rank high, i.e., could have been selected using the OPAS paradigm, to ensembles with mediocre ranking. The second and most important test will show if it is generally better to select one specific ensemble (the best possible according to some selection criterion) compared to just use all the models in the pool.

IV. Results

Table I below shows the overall properties of the classifier pool, measured on test data. The columns named EA-a, BA-a and Dis-a tabulate ensemble accuracy for the *all* ensemble, i.e., if the entire pool was used as the ensemble, mean base classifier accuracy and overall disagreement. Columns +, 0 and - indicate the proportions of the smaller ensembles in Experiment 1 (i.e., size 51) that outperform (+), have identical accuracy as (0) and is outperformed by (-) the all-ensemble. The last column, Ran, shows the range for test accuracy.

If we first of all look at the effect of using a separate validation set, it is as expected obvious that the base classifiers are less accurate but also more diverse when trained on less data. In this setting, the increased diversity is not enough to make up for the lowered base classifier accuracies, resulting in significantly lower ensemble accuracies for the all-ensemble. From this result, it seems very unlikely that using a separate validation set will result in better ensembles overall, even if the selection procedure will show to benefit from estimates on fresh data.

Looking at how the small ensembles fare against the all-ensemble, it is very interesting to see that on almost every

data set, there is indeed a fairly large proportion of the smaller ensembles that actually outperform the all-ensemble. Looking at mean values over all data sets, the picture is that approximately one third of the smaller ensembles have a higher accuracy than *all*, and that the spread in ensemble accuracy is larger than 0.05. This is, of course, a very encouraging result for the OPAS paradigm. We now know that there are many smaller ensembles that are more accurate than *all*, we just need to find a suitable method to identify them!

TABLE I
PROPERTIES OF THE POOL - TEST DATA

	With validation set							Without validation set						
	EA-a	BA-a	Dis-a	+	0	-	Ran	EA-a	BA-a	Dis-a	+	0	-	Ran
bcanc	.704	.661	.247	.41	.31	.28	.070	.713	.675	.242	.32	.30	.37	.085
cmc	.481	.451	.347	.56	.11	.33	.044	.543	.502	.302	.42	.09	.49	.044
colic	.843	.812	.143	.36	.42	.22	.022	.823	.787	.138	.20	.47	.33	.065
cred-a	.896	.860	.114	.25	.39	.36	.023	.853	.833	.106	.38	.30	.32	.035
cred-g	.758	.701	.259	.33	.14	.52	.040	.766	.712	.242	.34	.13	.53	.032
diabet	.779	.738	.221	.37	.22	.42	.047	.764	.749	.165	.50	.21	.29	.047
glass	.709	.614	.301	.44	.32	.24	.113	.681	.622	.239	.36	.32	.32	.094
haber	.711	.667	.240	.21	.34	.45	.040	.714	.692	.201	.41	.28	.32	.053
heart-c	.836	.782	.181	.17	.39	.44	.066	.829	.789	.172	.37	.34	.30	.093
heart-h	.760	.732	.198	.25	.39	.36	.069	.823	.785	.173	.28	.30	.42	.082
heart-s	.818	.769	.192	.31	.34	.35	.045	.830	.779	.189	.30	.35	.35	.075
iono	.926	.901	.083	.15	.68	.17	.023	.884	.861	.116	.19	.44	.37	.081
jEd42	.687	.665	.252	.25	.38	.37	.044	.715	.691	.173	.26	.29	.45	.074
jEd43	.639	.599	.285	.28	.29	.43	.065	.618	.598	.245	.37	.22	.40	.054
kc2	.765	.758	.141	.44	.39	.17	.031	.842	.835	.074	.36	.40	.24	.031
kc3	.887	.861	.110	.31	.44	.26	.018	.908	.888	.078	.24	.42	.34	.044
liver	.737	.660	.302	.34	.29	.37	.116	.737	.681	.241	.37	.30	.33	.070
lymph	.846	.777	.182	.12	.65	.22	.054	.835	.798	.149	.16	.53	.30	.081
prom	.885	.748	.309	.19	.65	.16	.077	.935	.794	.269	.14	.55	.30	.077
sonar	.871	.810	.191	.26	.52	.22	.058	.862	.798	.191	.18	.35	.47	.058
spectf	.849	.764	.228	.34	.36	.30	.023	.856	.805	.180	.34	.21	.45	.081
vehic	.814	.754	.218	.35	.18	.47	.033	.802	.768	.171	.28	.17	.56	.043
Mean	**.782**	**.731**	**.216**	**.30**	**.37**	**.32**	**.051**	**.788**	**.747**	**.184**	**.31**	**.32**	**.38**	**.063**

As mentioned above, the OPAS paradigm requires that not all ensembles have identical performance on the data used as basis for the selection. Naturally, (almost) all ensembles compared during the selection will have different values for mean base classifier accuracy as well as the diversity measures. For ensemble accuracy, however, this is not the case. Table II below tabulates the number of different accuracy values among the 50 ensembles, when averaged over the ten repeated runs. Clearly, this is a strong warning against using OPAS based on results for the entire ensemble - even the smaller ensembles are robust enough to obtain very similar results on training, validation and OOB, as well as, on test data. In addition, the fact that the 50 ensembles obtain, on average, as few as 5 different values on test accuracy indicates that the selection must be very hard.

TABLE II
PROPERTIES OF THE POOL - ENSEMBLE SPREAD

	EA-Train	EA-OOB	EA-VAL	EA-TEST
bcanc	6.8	11.5	4.0	5.6
cmc	18.1	21.4	11.9	12.7
colic	6.1	8.8	3.7	3.5
cred-a	6.7	9.7	5.6	4.6
cred-g	13.4	17.3	8.3	9.6
diabet	10.0	13.5	6.8	7.2
glass	6.7	9.7	3.6	5.0
haber	7.5	10.6	4.7	5.5
heart-c	5.1	8.8	3.5	4.6
heart-h	4.3	10.2	2.5	4.8
heart-s	5.4	9.1	3.9	4.5
ionos	2.7	9.4	4.0	3.6
jEd42	7.9	9.1	3.3	4.9
jEd43	10.1	12.4	6.0	6.0
kc2	7.0	7.9	2.9	3.9
kc3	7.2	8.0	3.0	3.8
liver	9.4	12.0	4.8	5.0
lymph	2.6	6.0	2.4	3.1
prom	2.5	5.9	2.2	3.2
sonar	2.3	8.6	4.3	4.2
spectf	6.4	10.1	3.5	6.7
vehic	11.4	15.6	6.3	8.3
Mean	**7.3**	**10.7**	**4.6**	**5.5**

From the results discussed above, we would expect low correlations between the different measures and test set accuracy, when taken over the entire data sets. Table III below confirms this. As a matter of fact, all correlations are, when averaged over all data sets, very low.

TABLE III
CORRELATIONS WITH TEST SET ACCURACY

	TRAIN					OOB					VAL				
	EA	BA	Dif	Dis	DF	EA	BA	Dif	Dis	DF	EA	BA	Dif	Dis	DF
bcanc	-.07	-.07	-.07	.00	-.08	-.01	-.02	-.01	-.01	-.05	.03	-.03	-.07	-.03	-.06
cmc	.02	.14	-.07	-.09	.04	.05	.11	-.05	-.06	.02	-.13	-.05	-.05	-.03	-.08
colic	-.04	-.05	-.01	.05	-.02	-.05	-.06	-.06	.00	-.06	.00	.04	.05	.01	.07
cred-a	-.14	-.12	-.06	.03	-.06	.01	-.12	-.01	.05	-.12	-.02	.01	-.01	-.02	.00
cred-g	.00	.01	.02	.00	.02	.02	-.04	.01	.05	-.01	.00	-.04	.09	.11	.02
diabet	.04	.00	.06	.07	.05	-.05	-.04	.02	.06	.03	.05	.02	-.03	-.04	-.01
glass	.01	-.01	-.06	-.05	-.04	-.04	-.04	.00	.04	-.04	-.01	-.06	-.03	-.02	-.05
haber	-.04	.02	-.02	-.04	-.01	-.10	-.02	-.10	-.03	-.05	.01	-.01	-.05	-.04	-.03
heart-c	-.10	-.10	-.07	.07	-.09	.02	-.07	-.03	.05	-.03	-.06	.07	-.01	-.07	.02
heart-h	-.04	-.08	-.05	.05	-.05	.06	-.05	-.02	.06	-.03	.00	-.06	-.03	.06	-.04
heart-s	-.06	.04	-.09	-.12	-.08	.00	.10	.01	-.13	-.01	-.03	.00	.05	.04	.04
ionos	.04	.14	.11	-.13	.12	.06	.14	.08	-.07	.10	.04	-.04	-.02	-.02	-.03
jEd42	-.01	-.07	-.10	-.05	-.12	-.02	.06	-.06	-.11	-.03	-.06	-.08	-.06	-.01	-.08
jEd43	.02	.01	.00	.00	.01	.01	-.01	.02	.01	-.01	.02	.13	-.05	-.06	.06
kc2	-.03	.00	.04	.06	.04	-.12	-.05	-.01	.05	-.03	.03	-.05	.00	.06	.00
kc3	-.07	-.09	-.10	.01	-.10	.05	.00	-.03	-.05	-.03	.04	.09	.09	.00	.10
liver	.03	.10	-.04	-.07	.01	-.03	.03	-.10	-.08	-.02	.01	.02	.12	.10	.09
lymph	.09	.01	.02	.00	.02	-.07	-.02	-.05	-.01	-.02	.02	.04	.10	.05	.08
prom	-.04	.05	-.05	-.05	.03	.09	.05	.01	-.07	-.07	-.05	-.15	-.11	.12	-.19
sonar	.04	.06	.08	-.02	.08	-.03	.03	.01	-.02	.06	.01	.02	.02	.01	.02
spectf	.32	-.02	.27	.13	.19	.00	-.07	.13	.22	.13	-.05	-.07	-.10	-.01	-.09
vehic	.03	.09	.15	.06	.15	.02	.04	.05	.03	.06	.04	.05	.06	.00	.06
Mean	**.00**	**.00**	**.00**	**.00**	**.00**	**.00**	**.00**	**-.01**	**.00**	**-.01**	**-.01**	**-.01**	**.00**	**.01**	**.00**

Correlation is of course a rather blunt measure, especially when taken over all the ensembles. The important question is instead if the best ensembles, according to the selection criterion, also obtain superior test accuracy. For this analysis,

we compare the best quartile (the 12 ensembles with the best performance on the selection criterion) to the worst quartile; i.e., the 12 ensembles with the worst performance on the selection criterion. Table IV below shows the number of the ten runs where the ensembles from the best quartile had a significantly higher test accuracy (based on a standard paired t-test with $\alpha = 0.05$), than the ensembles from the worst quartile. Unfortunately, these results too are very discouraging. There is apparently very little to gain from selecting an ensemble based on any of the evaluated metrics.

TABLE IV
NUMBER OF THE TEN RUNS WHERE SELECTING ENSEMBLES FROM THE BEST QUARTILE WAS SIGNIFICANTLY BETTER THAN SELECTING FROM THE WORST

	TRAIN					OOB					VAL				
	EA	BA	Dif	Dis	DF	EA	BA	Dif	Dis	DF	EA	BA	Dif	Dis	DF
bcanc	1	0	0	0	0	0	0	0	0	1	2	0	0	1	0
cmc	1	2	0	0	1	0	0	0	0	0	0	0	0	1	0
colic	0	0	0	0	2	0	0	0	0	0	0	0	2	0	0
cred-a	0	1	1	1	1	3	1	2	2	0	0	1	0	0	2
cred-g	1	0	0	1	0	1	1	0	2	1	1	2	2	1	1
diabet	0	1	1	1	2	0	0	1	0	0	0	0	0	2	0
glass	1	1	0	0	1	0	0	2	0	0	0	0	1	1	0
haber	2	0	0	0	0	0	0	1	2	0	1	0	0	0	0
heart-c	0	0	1	1	0	0	0	0	1	2	0	1	2	1	1
heart-h	0	1	0	1	1	2	2	1	1	1	1	1	0	0	0
heart-s	0	1	0	0	1	1	3	1	0	2	0	2	0	2	2
ionos	0	1	0	0	1	1	2	2	0	4	0	0	0	0	0
jEd42	0	1	0	0	0	2	2	1	0	1	0	0	0	1	0
jEd43	0	0	0	0	1	0	0	0	1	0	1	3	0	0	2
kc2	0	1	1	2	3	0	0	2	3	2	0	0	1	0	1
kc3	0	0	0	0	0	1	1	0	0	0	1	2	3	2	2
liver	1	1	1	1	1	0	3	0	0	0	1	1	1	1	1
lymph	2	0	0	0	0	0	1	0	0	1	0	1	2	3	3
prom	0	0	0	0	0	1	0	0	0	0	0	0	0	3	0
sonar	1	2	3	1	3	1	2	1	0	3	0	0	0	0	1
spectf	8	1	6	2	4	1	0	2	4	2	0	0	0	0	0
vehic	0	3	3	1	3	0	0	2	1	2	1	1	1	2	2
Mean	**0.8**	**0.8**	**0.8**	**0.5**	**1.1**	**0.6**	**0.8**	**0.8**	**0.8**	**1.0**	**0.4**	**0.7**	**0.7**	**1.0**	**0.8**

In Table V below, the comparison between selecting from the best quartile and selecting from the worst quartile is presented in a slightly different way. Here, test accuracies for the different selection criteria (averaged over models, runs and data sets) are presented, together with an outright pairwise comparison. W/T/L shows the number of data set wins, ties and losses, respectively, for selecting from the best quartile against selecting from the worst. When selecting based on training or OOB data, the average test set accuracy is actually identical for the best and worst quartile. In addition, when looking at wins and losses, there is no apparent advantage in using models from the best quartile either. Actually, it is just as often better to select from the worst quartile. Unfortunately, this is also true when a separate validation set is used, the only differences being the accuracy levels, where it is again obvious that the smaller training data resulted in less accurate ensembles overall. Looking, finally, at the results on test data, which are included for reference only (since it obviously can not be used as selection criteria), it is interesting to see that on test data there are significant differences. Specifically, if we could somehow find the ensembles with the highest mean base classifier accuracy or diversity, measured as double fault or difficulty, on the test data, that would result in very accurate ensembles. Maximizing disagreement, even on the test data will, however, not lead to more accurate ensembles, which is a pity, since it can be calculated (at least for two class problems) without the true targets.

TABLE V
SELECTING FROM THE BEST QUARTILE VS. SELECTING FROM THE WORST

	Best quartile					Worst quartile				
	EA	BA	Dif	Dis	DF	EA	BA	Dif	Dis	DF
TRAIN										
Mean	.786	.786	.786	.785	.786	.786	.786	.786	.786	.786
W/T/L	7/5/10	12/2/8	8/2/12	7/5/10	10/2/10					
OOB										
Mean	.786	.786	.786	.786	.786	.786	.786	.786	.786	.786
W/T/L	6/6/10	10/3/9	7/6/9	10/3/9	7/5/10					
VAL										
Mean	.782	.781	.781	.782	.781	.782	.782	.782	.782	.782
W/T/L	10/5/7	10/3/9	8/3/11	8/5/9	6/6/10					
TEST										
Mean	.803	.792	.791	.786	.793	.769	.780	.781	.786	.779
W/T/L	22/0/0	22/0/0	21/1/0	7/7/8	22/0/0					

Yet another interesting comparison, between selecting the single best (sub-) ensemble and using the all ensemble, is presented in Table VI below. From these results, it appears to be at least slightly better to omit the selection phase, and instead just use all available models. As seen in Table I above, the mean test accuracy over all data sets for the all ensemble is 0.782 when using a validation set, and 0.788 when not.

TABLE VI
COMPARING THE BEST ENSEMBLE AGAINST USING THE ENTIRE POOL

	EA	BA	Dif	Dis	DF
TRAIN					
Mean	.787	.788	.785	.785	.787
W/T/L	8/3/11	9/0/13	5/3/14	5/2/15	9/1/12
OOB					
Mean	.786	.786	.785	.785	.787
W/T/L	7/3/12	8/2/12	5/3/14	6/1/15	9/1/12
VAL					
Mean	.781	.782	.781	.782	.782
W/T/L	8/2/12	8/5/9	10/2/10	11/1/10	10/2/10
TEST					
Mean	.814	.798	.795	.785	.800
W/T/L	22/0/0	22/0/0	19/0/3	5/4/13	22/0/0

In the final analysis of the results from Experiment 1, we turn the problem around. Instead of first sorting all ensembles based on a specific selection criterion, and then see how they fared on test data, we now sort the ensembles on test data accuracy, and investigate if there are any differences between good or poor ensembles that *could have been observed* on available data. As seen in Table VII below, the results are again disappointing. It is actually impossible to spot any differences in the evaluated performance measures on training, validation or OOB data. Only when we look at results on test data,

is higher mean base classifier accuracy or higher diversity beneficial for ensemble accuracy.

TABLE VII
RESULTS FOR ENSEMBLES SORTED ON TEST ACCURACY

	EA	BA	Dif	Dis	DF
TRAIN					
Single Worst	.903	.846	.046	.155	.076
Worst Quartile	.903	.847	.046	.155	.076
Best Quartile	.903	.847	.046	.156	.076
Single Best	.904	.847	.046	.156	.076
OOB					
Single Worst	.781	.746	.097	.185	.161
Worst Quartile	.781	.747	.097	.184	.161
Best Quartile	.781	.747	.098	.185	.162
Single Best	.781	.746	.097	.185	.162
VAL					
Single Worst	.785	.740	.083	.053	.040
Worst Quartile	.784	.740	.082	.053	.040
Best Quartile	.784	.740	.082	.053	.040
Single Best	.784	.739	.082	.053	.040
TEST					
Single Worst	.755	.743	.094	.185	.164
Worst Quartile	.769	.745	.093	.184	.163
Best Quartile	.803	.750	.091	.185	.158
Single Best	.814	.751	.090	.185	.156

So based on all those results from Experiment 1, we must conclude that, at least in this setting and with the evaluated performance metrics as selection criteria, the OPAS strategy has failed miserably.

Turning to Experiment 2, where the ensemble size is 7 instead of 51, Table VIII below shows the properties of this pool in comparison to the properties of the pool of larger ensembles, which were shown above in Table I. It can be seen that a slightly smaller proportion of the ensembles outperform the all-ensemble, and that the range for test accuracy is somewhat larger. Interestingly enough, it is, however, still more than 25% of the ensembles of size 7 that have higher test accuracies than the all-ensemble of size 100.

TABLE VIII
COMPARING PROPERTIES OF ENSEMBLES OF SIZE 7 AND 51

	With validation set				Without validation set			
	+	0	-	Ran	+	0	-	Ran
Mean for size 7	.26	.17	.57	.110	.27	.19	.54	.099
Mean for size 51	.30	.37	.32	.051	.31	.32	.38	.063

Looking at Table IX below, we see that there is also an increased spread in the results obtained by the ensembles on all data sets. As a matter of fact, the number of different values is almost doubled for all the measures, including test set accuracy. Maybe OPAS will be more successful when selecting these much smaller ensembles?

TABLE IX
COMPARING SPREAD OF ENSEMBLES OF SIZE 7 AND 51

	EA-Train	EA-OOB	EA-VAL	EA-test
Mean for size 7	13.2	19.1	8.4	9.5
Mean for size 51	7.3	10.7	4.6	5.5

From the results in Table X below, there is at least a tendency that it is better to pick one of the ensembles from the best quartile than the worst, based on ensemble accuracy or mean base classifier accuracy. In particular, when using OOB values, a standard sign test with $\alpha = 0.05$, shows that it is even significantly better to use ensembles from the best quartile. One obvious reason is that for such small ensembles, it becomes very important to avoid the base classifiers with the worst individual accuracies, which is no surprise.

TABLE X
COMPARING BEST AND WORST QUARTILES (ENSEMBLE SIZE = 7)

	Best quartile					Worst quartile				
	EA	BA	Dif	Dis	DF	EA	BA	Dif	Dis	DF
TRAIN										
Mean	.778	.779	.778	.776	.778	.777	.776	.778	.780	.777
W/T/L	8/2/12	12/2/8	8/0/14	3/3/16	11/1/10					
OOB										
Mean	.779	.779	.778	.776	.778	.776	.776	.778	.779	.777
W/T/L	15/2/5	16/3/3	8/4/10	4/2/16	12/0/10					
VAL										
Mean	.768	.769	.767	.768	.767	.769	.769	.770	.769	.769
W/T/L	12/2/8	13/0/9	10/1/11	7/2/13	12/0/10					
Test										
Mean	.805	.794	.789	.775	.795	.749	.761	.766	.779	.761
W/T/L	22/0/0	22/0/0	21/0/1	4/1/17	22/0/0					

Regrettably, when comparing the selected ensembles against the all-ensemble in Table XI below, the picture is again very clear. Here, it was actually always significantly better (a standard single-sided sign test requires 16 wins of 22 for significance with $\alpha = 0.05$) to use the entire pool than selecting a specific ensemble. So, in Experiment 2 too, the OPAS strategy failed.

TABLE XI
COMPARING THE BEST ENSEMBLE AGAINST USING THE ENTIRE POOL (ENSEMBLE SIZE = 7)

	EA	BA	Dif	Dis	DF
TRAIN					
Mean	.780	.781	.780	.774	.782
W/T/D	3/1/18	6/0/16	2/0/20	2/0/20	4/0/18
OOB					
Mean	.780	.779	.780	.774	.780
W/T/D	1/3/18	3/2/17	6/0/16	1/2/19	3/0/19
VAL					
Mean	.765	.766	.763	.768	.763
W/T/D	2/0/20	2/1/19	2/1/19	2/2/18	3/0/19
Test					
Mean	.824	.807	.799	.771	.805
W/T/D	22/0/0	22/0/0	20/1/1	2/0/20	22/0/0

V. CONCLUSION

In this paper, we have evaluated the static overproduce-and-select paradigm. From the experiments, investigating a specific but typical OPAS setup, the main result is that there is absolutely nothing to gain by selecting an ensemble based on the very natural metrics evaluated here. When using larger ensembles (i.e., picking 51 ANNs from the 100 net pool), each

ensemble is quite accurate, actually comparable to using the entire pool as the ensemble, but there is no way to detect the best ensembles. When using smaller ensembles (7 ANNs), it was on the other hand, possible to distinguish poor ensembles from better ones by comparing either ensemble accuracy or mean base classifier accuracy on OOB. Unfortunately, the results also showed that in this scenario, even the most promising ensembles could not compete with the strategy of using all available models as the ensemble.

The main explanation identified in this study is the robustness inherent in ensembles. At least for ANN models and these fairly small data sets, the often marginal differences in performance on training, validation or OOB data, will simply not carry over to test data.

The overall conclusion is thus that the results from both experiments clearly show that the grim reality is that although there exists a fairly large proportion of smaller ensembles with better test set accuracies than the all-ensemble, none of the very natural, and frequently used, selection criteria evaluated here could be used to identify these highly accurate ensembles.

VI. Discussion

One very important question is how general the results obtained in this study are. After all, we evaluate only some selection criteria, and under very specific circumstances, i.e., fixed size ensembles selected from a pool of exactly 100 models, all MLPs trained with bagging and identical settings etc. Still, we argue that the settings used are illustrative for OPAS; ANNs is the most typical model, and a pool size of 100 is a reasonable choice. Using a smaller pool will normally make no sense, while the pool would have to be considerably larger for these results to lose their validity. It should be noted that we could possibly get 'better' results, e.g., higher correlations and larger differences between best and worst picks, by including weaker models, for instance ANNs without a hidden layer, in the pool, but those ensembles would most certainly be less accurate than the ensembles used here.

During the experimentation, we also looked at using combinations of one diversity measure, together with mean base classifier accuracy, as the selection criterion, as suggested e.g. in [26], unfortunately with very similar results. A more important issue is probably the choice of data sets to include in the study. Using much larger data sets (10000+ instances), all partitions of the data sets will of course be substantially larger, which may lead to different results, but this remains to be verified. Nevertheless, one implication of the conclusion is that it seems better to invest time in finding ways of maximizing the performance of the all-ensemble, e.g., by maximizing diversity and base classifier accuracy, than trying to find methods for making an optimal selection from the pool.

Acknowledgement

This work was supported by the Swedish Foundation for Strategic Research through the project High-Performance Data Mining for Drug Effect Detection (ref. no. IIS11-0053) at Stockholm University, Sweden.

References

[1] T. G. Dietterich, "Ensemble methods in machine learning," in *Multiple Classifier Systems*, ser. Lecture Notes in Computer Science, J. Kittler and F. Roli, Eds., vol. 1857. Springer, 2000, pp. 1–15.

[2] D. W. Opitz and R. Maclin, "Popular ensemble methods: An empirical study," *Journal of Artificial Intelligence Research*, vol. 11, pp. 169–198, 1999.

[3] T. G. Dietterich, "Machine-learning research: Four current directions," *The AI Magazine*, vol. 18, no. 4, pp. 97–136, 1998.

[4] G. Brown, J. Wyatt, R. Harris, and X. Yao, "Diversity creation methods: a survey and categorisation," *Journal of Information Fusion*, vol. 6, no. 1, pp. 5–20, 2005.

[5] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[6] ——, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[7] E. Tang, P. Suganthan, and X. Yao, "An Analysis of Diversity Measures," *Machine Learning*, vol. vol 65, pp. 247–271, 2006.

[8] L. I. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*. Wiley-Interscience, 2004.

[9] E. M. Dos Santos, R. Sabourin, and P. Maupin, "A dynamic overproduce-and-choose strategy for the selection of classifier ensembles," *Pattern Recognition*, vol. 41, no. 10, pp. 2993–3009, 2008.

[10] H. Boström, "Calibrating random forests," in *International Conference on Machine Learning and Applications, 2008.*, 2008, pp. 121–126.

[11] G. Giacinto and F. Roli, "Design of effective neural network ensembles for image classification purposes," *Image and Vision Computing*, vol. 19, no. 9-10, pp. 699–707, 2001.

[12] L. K. Hansen and P. Salamon, "Neural network ensembles," *IEEE Transactions on Pattern Analysis and Machine*, vol. 12, no. 10, pp. 993–1001, 1990.

[13] L. I. Kuncheva and C. Whitaker, "Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy," *Machine Learning*, vol. 51, no. 2, pp. 181–207, 2003.

[14] L. Saitta, "Hypothesis Diversity in Ensemble Classification," in *Foundations of Intelligent Systems*. Springer, 2006, pp. 662–670.

[15] G. Giancinto and F. Roli, "Design of effective neural network ensembles for image classification purposes," *Image vision and Computing Journal*, vol. 19, pp. 699–707, 2001.

[16] D. D. Margineantu and T. G. Dietterich, "Pruning adaptive boosting," in *Proceedings of the Fourteenth International Conference on Machine Learning (ICML 1997), Nashville, Tennessee, USA, July 8-12, 1997*, D. H. Fisher, Ed. Morgan Kaufmann, 1997, pp. 211–218.

[17] G. Giacinto and F. Roli, "An approach to the automatic design of multiple classifier systems," *Pattern recognition letters*, vol. 22, pp. 25–33, 2001.

[18] G. Zenobi and P. Cunningham, "Using diversity in preparing ensembles of classifiers based on different feature seature subsets to minimize generalization error," in *Lecture Notes in Computer Science*. Springer Verlag, 2001, pp. 576–587.

[19] G. Tremblay, R. Sabourin, and P. Maupin, "Optimizing nearest neighbour in random subspaces using a multi-objective genetic algorithm," *Pattern Recognition, International Conference on*, vol. 1, p. 208, 2004.

[20] E. M. Dos Santos, R. Sabourin, and P. Maupin, "Pareto analysis for the selection of classifier ensembles," in *Genetic and evolutionary computation, GECCO*. ACM, 2008, pp. 681–688.

[21] H. Elghazel, A. Aussem, and F. Perraud, "Trading-off diversity and accuracy for optimal ensemble tree selection in random forests," in *Ensembles in Machine Learning Applications*, ser. Studies in Computational Intelligence. Springer, 2011, vol. 373, pp. 169–179.

[22] D. Ruta and B. Gabrys, "Classifier selection for majority voting." *Information Fusion*, vol. 6, no. 1, pp. 63–81, 2005.

[23] U. Johansson, T. Löfström, and L. Niklasson, "The importance of diversity in neural network ensembles-an empirical investigation," in *International Joint Conference on Neural Networks*, 2007, pp. 661–666.

[24] A. Frank and A. Asuncion, "UCI machine learning repository," 2010. [Online]. Available: http://archive.ics.uci.edu/ml

[25] J. Sayyad Shirabad and T. Menzies, "The PROMISE Repository of Software Engineering Databases." School of Information Technology and Engineering, University of Ottawa, Canada, 2005. [Online]. Available: http://promise.site.uottawa.ca/SERepository

[26] T. Löfström, U. Johansson, and H. Boström, "Ensemble member selection using multi-objective optimization," in *CIDM*, 2009, pp. 245–251.