Statistical Ensemble Method (SEM): A New Meta-machine Learning Approach Based on Statistical Techniques

Andrés Yáñez Escolano, Pedro Galindo Riaño, Joaquin Pizarro Junquera, and Elisa Guerrero Vázquez

Universidad de Cádiz, Departamento de Lenguajes y Sistemas Informáticos,
Grupo de *"Sistemas Inteligentes de Computación"*,
C.A.S.E.M. 11510 – Puerto Real (Cádiz), Spain
{andres.yaniez, pedro.galindo, joaquin.pizarro, elisa.guerrero}@uca.es

Abstract. The goal of combining the outputs of multiple models is to form an improved meta-model with higher generalization capability than the best single model used in isolation. Most popular ensemble methods do specify neither the number of component models nor their complexity. However, these parameters strongly influence the generalization capability of the meta-model. In this paper we propose an ensemble method which generates a meta-model with optimal values for these parameters. The proposed method suggests using resampling techniques to generate multiple estimations of the generalization error and multiple comparison procedures to select the models that will be combined to form the meta-model. Experimental results show the performance of the model on regression and classification tasks using artificial and real databases.

1 Introduction

The aim of machine learning is to make a good model based on a set of examples. The goal is not to learn an exact representation of the training data itself, but rather to build a statistical model of the process which generates the data [1]. Classic methods for model building choose a model from a set of competing alternatives, assigning a single measure of generalization error to each candidate. The model which minimizes this value is selected and the rest is discarded.

However, when several models show similar generalization errors, we should not conclude that the model having the best performance on the validation set will achieve the best performance on new test data, given that we are working with a noisy, finite learning dataset. Therefore, any chosen hypothesis will be only an estimate of the real target and, like any estimate, will be affected by a bias and a variance term. Furthermore, there is another disadvantage with such approach: all the effort involved in generating the remaining near-optimal models is wasted. These drawbacks can be overcome by combining these models.

Model combination approach leads to significant improvements of new prediction with a little additional computational effort. It is possible to identify two main

J. Cabestany, A. Prieto, and D.F. Sandoval (Eds.): IWANN 2005, LNCS 3512, pp. 192 – 199, 2005.

approaches to combining models: multiexpert and multistage methods. Multiexpert methods work in parallel, while multistage methods use a serial approach where the next model is trained/consulted only for examples rejected by the previous models. Two main groups of meta-machine learning methods which work in parallel exist: mixture of experts [7] and ensemble methods. While mixture of experts approach divides the input space with a gating network and allocates the subspaces to different experts (models), the output of an ensemble is generated by the weighted outputs of each model. The performance of an ensemble can be better than the performance of the best single model used in isolation when the models are accurate enough and fairly independent in the errors they make [6], [9]. The meta-machine learning method proposed in this paper is based on the last approach.

2 The Statistical Ensemble Method

In general, an ensemble is built in two steps: a) Generation/selection of a number of component models. b) Combination of their predictions.

The most prevailing approaches for generating component models are based on varying the topology, the algorithm, the set of initial parameters to be used in the iterative learning process (i.e. random weights for MLPs) or the data itself (subsampling the training examples, manipulating the input features or the output targets and injecting ramdomness) [3],[16]. However, most popular ensemble methods specify neither the number of component models nor their complexity, and, obviously, these parameters strongly influence the generalization capability of the ensemble. In this paper we propose a methodology to generate a meta-model with optimal values for these parameters.

The steps of the proposed methodology may be outlined as follow:

1. Obtain multiple generalization error measures for each model using resampling techniques. The use of a set of estimations instead of a single measure of generalization error for model selection was proposed in [14],[5].
2. Determine the set of models whose errors are not significantly different from the model with minimum estimated generalization error using statistical tests [18] for comparing groups of paired samples (multiple comparison procedures).
3. Combine these near-optimal models using ensemble methods.

2.1 Multiple Error Measures Using Resampling Techniques

Resampling methods for estimating the generalization error generate multiple test-and-train datasets, and estimate the generalization error as the average of the validation errors.

The main approaches to resampling are the following:

- Random hold-out: many randomly train-and-test sets are generated. The examples are selected without replacement.
- K-fold cross-validation: examples are randomly divided into k mutually exclusive partitions of approximately equal size. Each model is trained and tested *k* times; each time tested on a fold and trained on the dataset minus the fold.

- Leave-one-out: it is a special case of k-fold cross validation, where k equals the sample size.
- Bootstrapping: instead of repeatedly analyzing subsets of the data, you repeatedly analyze subsamples of the data. Each subsample is a random sample with replacement from the full dataset and constitutes a training set. Examples not found in the training set form the validation set.

In the above methods, the estimate of generalization error is taken as the average of the estimated accuracies (validation errors) from the different train/test sets. In the proposed methodology, this estimate will be used to determine the reference model, but all validation errors obtained from each train/test pair will be kept in order to be able to apply statistical tests to compare groups of related samples, instead of comparing a single estimate of the generalization error. The whole process may be described in more detail as follows:

1. Take the whole data set and create *m* resampled data sets (*m* train/test pairs) using any of the approaches described above.
2. For each resampled train/test set (*m* pairs), and for each model (*k* models), obtain a validation error. This allows us to obtain an array of *m* x *k* validation errors.
3. Determine the class (*Si*) with minimum estimated generalization error, that is the class with minimum validation error mean.

2.2 Multiple Model Selection Using Statistical Tests

The second step in the methodology consists on the selection of a set of models to be combined. The best conditions for combining occur when the learned models are accurate enough, but fairly independent in the errors they make. The first condition will be guaranteed by determining those models not significantly different from the model with minimum estimated generalization error using statistical tests for comparing k groups of related samples. The second condition is much more difficult to ensure, and is approximated considering different architectures, learning paradigms, model complexities, etc.

The proposed methodology determines a subset of models having *similar* error measures that the model with minimum estimated generalization error as follows:

1. Apply a medium power test (i.e. Nemenyi) to obtain the models which are not significantly different from the model with minimum estimated generalization error.
2. Apply an omnibus test for related samples (repeated measures ANOVA test, if the assumptions are met or Friedman test in different case).
   2.1. If the global null hypothesis is true (that is, all model classes of this set are not significantly different), finish the process.
   2.2. If the global null hypothesis is false, apply more powerful multiple comparison procedures (t or Wilcoxon paired tests with Bonferroni method for p-values adjustment) and obtain a subset with the model classes which are not significantly different from the model class with minimum estimated generalization error.

Some remarks about the method should be done. When omnibus tests are significant, it indicates that at least two of the model classes are significantly different, but we don't know which could be. At this point, multiple comparison procedures, which are usually less powerful, are applied.

Nemenyi test is a medium power multiple comparison procedure. It may even accept model classes that should be rejected. It is a good procedure to generate an initial but not definitive set of *non-significant* model classes.

Finally, the results may improve with a large number of resampled sets: resampling methods estimate better the generalization error and parametric tests [4], which are more powerful, may be applied on step 2. We suggest $m \geq 30$.

2.3 Model Combination Using Ensemble Methods

Model combination starts with the determination of a model for each near-optimal class as determined in the previous step of the methodology. For each class, we should select the member $f_i(x,w^*)$ whose parameter vector w^* minimizes the empirical risk for the whole dataset.

Once a set of component models has been generated, they must be combined. This combination consists of a weighted combination of models. For combining the outputs of component models, the most prevailing approaches are majority weighted voting for classification tasks and weighted averaging for regression tasks [6],[13].

3 Experimental Results

In this section we shall describe the experiments carried out with our methodology, the obtained results, and a comparative study with other strategies. A number of simulations have been conducted to evaluate the efficiency of SEM method using Radial Basis Function networks (RBF). In our experiments, we have used several databases from the UCI repository [2], StatLib repository [12], Donoho-Johnstone benchmarks [15] and the ELENA Project [8] in order to test the performance of the method on regression and classification tasks using artificial and real databases.

We have repeatedly extracted (100 times) from each database a small number of examples (sample size column in tables 2 to 5) for model estimation, while the remaining ones were used to get a precise estimation of the expected generalization error for each trained model. For the block function, gaussian noise has been added to the outputs and the generalization error is estimated with 10000 previously unseen examples.

In order to compare the performance of different networks, we define the observed efficiency of model m_i as the ratio of the lowest estimated generalization error to that of model m_i. Thus, observed efficiencies range from 0 to 1. An observed efficiency equal to 1.0 would correspond to a model always having the lowest generalization error.

We have a considered an initial set of RBF models with complexities ranging from 1 to n (n = 20 or 30, depending on the database) where n is defined as the number of kernels. The width of the basis functions has been set to

$$\sigma = \left\| \max(x_i - x_j) \right\| / \sqrt{2n}$$

All statistical tests have been applied using a level of significance $\alpha = 5\%$.

For SEM, we suggest three methods which require some restrictions on the weights (the weights must be greater than zero and sum to one) and fix the weights at the end of training. First, an unweighted average is computed (Basic Ensemble Method-BEM)[13]. Second, the weights are inversely proportional to estimated generalization error [11]. Finally, the weights are proportional to the number of times that each model has been selected as the model with minimum validation error [17]. A comparative of these methods is shown in [17]. In this paper, the simplest method (BEM) is applied.

Table 2 shows observed efficiency values for three regression tasks and for different sample sizes: Block function (25 y 50 data) , Abalone data set (50 and 100 data) and California housing (250 and 500 data). Statistical measures (mean, median and standard deviation) of the observed efficiency for the different methodologies are shown, as well as methodologies are ranked from the highest mean of the observed efficiency to the lowest mean (from 1 to 4 respectively). We have considered four different model building strategies: a) an ensemble using all the models, b) SEM using only Nemenyi test, c) SEM using Bonferoni test, and d) the model with the lowest estimated error.

Table 3 shows the number of component networks per meta-model on average for the different regression tasks considered.

Table 2. Observed efficiency for three regression tasks

Database	Sample size	Statistical measures	All models	SEM using Nemenyi	SEM using Bonferroni	Model selection
Block function	25	Mean	0.4019	0.8657	0.8308	0.7793
		Median	0.2156	1.0000	0.9044	0.7670
		Stand. dev.	0.4053	0.2554	0.2214	0.1436
		Rank	3	1	2	4
	50	Mean	0.7855	0.8955	0.8915	0.7260
		Median	0.9629	0.9728	0.9555	0.7104
		Stand. dev.	0.3110	0.2032	0.1869	0.1606
		Rank	3	1	2	4
Abalone	50	Mean	0.9220	0.9789	0.9584	0.8638
		Median	0.9290	0.9968	0.9746	0.8661
		Stand. dev.	0.0702	0.0318	0.0486	0.1036
		Rank	3	1	2	4
	100	Mean	0.9797	0.9826	0.9727	0.9305
		Median	0.9882	0.9882	0.9828	0.9503
		Stand. dev.	0.0234	0.0192	0.0315	0.0703
		Rank	2	1	3	4
California housing	250	Mean	0.9815	0.9893	0.9804	0.9063
		Median	0.9881	0.9978	0.9958	0.9111
		Stand. dev.	0.0198	0.0167	0.0280	0.0582
		Rank	2	1	3	4
	500	Mean	0.9881	0.9970	0.9876	0.9152
		Median	0.9887	1.0000	0.9939	0.9176
		Stand. dev.	0.0086	0.0070	0.0213	0.0430
		Rank	2	1	3	4

Table 3. Average of the number of component networks per meta-model

Database	Sample size	All models	SEM using Nemenyi	SEM using Bonferroni	Model Selection
Block function	25	30	7,37	4,08	1
	50	30	13,87	8,67	1
Abalone	50	20	11,27	6,47	1
	100	20	12,51	6,72	1
California housing	250	20	15,12	8,07	1
	500	20	17,35	11,26	1
	250	30	29,05	23,69	1

Table 4 shows results for 3 different classification tasks: Clouds data set with 50 and 250 experimental data, Gauss 2D with 50 and 100 experimental data and Phoneme data set with 100 and 250. Table 5 shows the number of component networks per meta-model on average.

Table 4. Observed efficiencies for three binary classification tasks

Database	Sample size	Statistical measures	All models	SEM using Nemenyi	SEM using Bonferroni	Model selection
Clouds	50	Mean	0.9193	0.9587	0.9125	0.7475
		Median	1.0000	0.9778	0.9330	0.7416
		Stand. dev.	0.1384	0.0506	0.0871	0.1008
		Rank	2	1	3	4
	250	Mean	0.9156	0.9827	0.9955	0.9479
		Median	0.9182	0.9857	1.0000	0.9481
		Stand. dev.	0.0340	0.0168	0.0103	0.0393
		Rank	3	2	1	4
Gauss 2D	50	Mean	0.7390	0.9624	0.9782	0.9395
		Median	0.7786	0.9772	1.0000	0.9618
		Stand. dev.	0.1446	0.0535	0.0464	0.0674
		Rank	4	2	1	3
	100	Mean	0.9155	0.9842	0.9913	0.9681
		Median	0.9265	0.9919	1.0000	0.9771
		Stand. dev.	0.0514	0.0210	0.0164	0.0332
		Rank	4	2	1	3
Phoneme	100	Mean	0.9918	0.9935	0.9813	0.9129
		Median	1.0000	1.0000	0.9955	0.9110
		Stand. dev.	0.0149	0.0116	0.0290	0.0460
		Rank	1	2	3	4
	250	Mean	0.9994	0.9986	0.9903	0.9014
		Median	1.0000	1.0000	0.9977	0.9083
		Stand. dev.	0.0017	0.0031	0.0154	0.0467
		Rank	1	2	3	4

Table 5. Average of the number of component networks per meta-model

Database	Sample size	All models	SEM using Nemenyi	SEM using Bonferroni	Model Selection
Clouds	50	20	18,34	12,02	1
	250	20	18,91	13,34	1
Gauss 2D	50	20	14,03	6,22	1
	100	20	14,83	6,85	1
Phoneme	100	30	25,96	18,72	1
	250	30	29,05	23,69	1

Experimental results from the simulations (tables 2 and 4) suggest that generalization capability of SEM is higher (or similar in the worst case) than the model with minimum expected generalization error, and better than the ensemble obtained combining all component networks. Similar results are obtained applying only Nemenyi test or applying Bonferroni correction after it, but the ensembles generated after Bonferroni correction are less complex (tables 3 and 5). The SEM model selects the optimal cardinality for the ensemble and the appropriate complexity for their component networks.

4 Conclusions

It is known that combining networks improve the generalization ability. The number of component networks and their complexity are free parameters and usually must be fixed before the training process begins, but there is no standard procedure to fix these parameters. In this paper we have proposed a new ensemble method based on statistical techniques (SEM) which fixes these parameters in order to obtain a low generalization error with a small set of optimal component networks. Experimental results have shown that SEM improves the performance when compared to the strategy which selects the model with the lowest estimated generalization error and the strategy which combines all the networks.

Finally, other simulation results obtained applying our method [17] show that:

1. With other families of models (eg. linear models, polynomials, MLP networks,…), SEM always reduced the generalization error.
2. More powerful multiple comparison procedures based on Bonferroni correction [10] are not necessary, because a set of models with similar cardinality is selected.
3. Similar results are obtained applying random hold-out technique, but leave-one-out or k-fold cross-validation techniques make the results worse, because they select set of models with high cardinality.

References

1. Bishop, C. M.: Neural network for pattern recognition. Clarendon Press-Oxford (1995)
2. Blake, C.L. y Merz, C.J. UCI Repository of machine learning databases http://www.ics.uci.edu/~mlearn/MLRepository.html. Irvine, CA: University of California, Department of Information and Computer Science (1998).
3. Dietterich, T. G. Machine Learning Research: Four Current Directions. Artificial Intelligence Magazine, 18(4), pp. 97-136 (1997).
4. Don Lehmkuhl, L.: Nonparametric statistics: methods for analyzing data not meeting assumptions required for the application of parametric tests. Journal of prosthetics and orthotics Vol. 8, num. 3, pp.105-113 (1996)
5. Guerrero, E., Yáñez, A., Galindo, P. and Pizarro, J. Repeated measures multiple comparison procedures applied to model selection in neural network. Proceeding of the 6th. Int. Conf. on Artificial Neural Network (IWANN), vol. 2, pp. 88-95 (2001).
6. Hansen, L. K., Salamon, P. Neural network ensembles. IEEE Transactions on Pattern Analysis and Machine Intelligence, 12(10), pp. 993-1001 (1990).
7. Jacobs, R. A., Jordan, M. I., Nowlan, S. J. e Hinton, G. E. Adaptative mixtures of local experts. Neural Computation, 3 (1), pp. 79-87 (1991).
8. Jutten, C. et al. ESPIRIT basic research project number 689 ELENA. ftp.dice.ucl.ac.be/ pub/ neural-net/ELENA/databases
9. Krogh, A., Vedelsby, J. Neural networks ensembles, cross validation and active learning. In Tesauro, G., Touretzky, D. and Leen, T. (Eds.). Advances in Neural Information Processing Systems, vol. 7, pp. 231-238. The MIT Press (1995).
10. Lasarev, M. R.: Methods for p-value adjustment, Oregon Health & Science University, http://medir.ohsu.edu/~geneview/education/dec19_h.pdf (2001).
11. Optiz, D. W. y Shavlik, J. W. (1996). Generating accurate and diverse members of a neural-network ensemble. Advances in Neural Information Processing Systems, 8, págs. 535-541. Ed. D. S. Touretzky, M. C. Mozer y M. E. Hasselmo. The MIT Press.
12. Pace, R. K. y Barry, R. Sparse Spatial Autoregressions. Statistics and Probability Letters, 33, pp. 291-297. http://lib.stat.cmu.edu/ (1997).
13. Perrone, M.P., Cooper, L.N. When networks disagree: ensemble method for neural networks, in: R.J.Mammone (Ed.), Artificial Neural Networks for Speech and Vision, Chapman & Hall, New York, pp.126-142 (1993).
14. Pizarro, J., Guerrero, E. and Galindo, P. Multiple comparison procedures applied to model selection. Neurocomputing 48, pp. 152-159 (2001).
15. Sarle, W. Donoho-Johnstone benchmarks: neural nets results. ftp://ftp.sas.com/pub/neural/ dojo/dojo.html (1999).
16. Scharkey, A. J. C. On Combining Artificial Neural Nets. Connection Science, 8, 3/4, pp. 299-314 (1996).
17. Yáñez, A. Regresión mediante la combinación de modelos seleccionados mediante técnicas de remuestreo y procedimientos de comparación múltiple. Thesis. University of Cádiz.
18. Zar, J. H.: Biostatistical analysis, Prentice Hall (1996)