Ensemble Learning of Coupled Parameterized Kernel Models

B. Hamers, J.A.K. Suykens,V. Leemans, B. De Moor

Katholieke Universiteit Leuven
Department of Electrical Engineering ESAT-SISTA
Kasteelpark Arenberg 10, B-3001 Leuven, Belgium
Tel: 32/16/32 11 45 , Fax: 32/16/32 19 70
E-mail: {bart.hamers,johan.suykens}@esat.kuleuven.ac.be

Abstract - ***In this paper we propose a new method for learning a combination of estimators. Classically, committee networks are constructed after training the networks independently from each other. Here we present a learning strategy where the training is done in a coupled way. We illustrate that combining parameterized kernel methods with output coupling and use of a synchronization set of data points leads to an improved generalization. Examples are given on artificial and real life data sets.***

Keywords. ***Ensemble learning, committee networks, kernel methods, synchronization, collective intelligence.***

I. Introduction

It is well-known in the literature that combining learning algorithms is a fruitful way for creating improved estimators. Many different strategies exist including sequential strategies like Boosting [5], Bagging [2] and other combinations of trained models like Committee Networks [1]. In committee networks the training of the individual submodels is done independently from each other.

In this paper we present a new technique where we couple the learning processes of the individual submodels. The idea behind this method is that the different estimators of the ensemble are each trained on disjoint training sets, but have to agree on a predefined set of input values of chosen data points. A method of coupled local minimizers and coupled training processes has been proposed in [8], where the coupling is done by means of synchronization constraints. In this paper the outputs of individual models are synchronized on a chosen set of data points. The technique is applied to the training of kernel models which have a similar form as least squares support vector machines [7] or regularization networks [3], but consider the model and training in a parameterized way. The method is illustrated on a number of examples.

This paper is organized as follows. Section 2 gives a short introduction on ensemble learning and committee networks. Section 3 discusses parameterized kernel models and uncoupled ensembles. In Section 4 explains ensemble learning by coupled minimizers using a synchronization set. Section 5 gives some results on both artificial as real-life data sets.

II. Ensemble Learning

Consider an additive noise model $\mathbf{y} = f(\mathbf{x}) + \mathbf{e}$ where $E[\mathbf{e}] = \mathbf{0}$, $E\left[\mathbf{e}\mathbf{e}^T\right] = I\sigma^2$ and $\sigma^2 < \infty$. Based on a training set $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^N \subset \mathbb{R}^d \times \mathbb{R}$, we will estimate a function $\hat{f}(\mathbf{x}; D)$ which approximates $f(\mathbf{x})$.

In ensemble learning one combines different models into one model. So, instead of using one estimator, one combines a population of estimators $\hat{f}^{(j)}$ with $j = 1, ..., q$ to estimate a function f. Different strategies exist about how to combine these submodels $\hat{f}^{(j)}$. In this work we will concentrate on *weighted averaging methods* (committee networks) or also called *Generalized Ensemble Methods (GEM)* as introduced in [1]

$$\hat{f}(\mathbf{x}, D) = \sum_{j=1}^{q} \beta^{(j)} \hat{f}^{(j)}\left(\mathbf{x}; D^{(j)}\right). \tag{1}$$

The optimal weighting factors minimizing the error are computed based on the error covariance matrix S, using a finite-sample approximation [1],

$$\beta^{(j)} = \frac{\sum_{i=1}^{q} \left(S^{-1}\right)_{ji}}{\sum_{k=1}^{q} \sum_{i=1}^{q} \left(S^{-1}\right)_{ki}}. \tag{2}$$

In relation to this paper we consider training the different submodels on separate subsets of the original data set. In this way, we avoid that the whole data set D has to

be learned by one single model. We divide D into q subsets, each of size $g = \lfloor N/q \rfloor$. Each of the submodels $\hat{f}^{(j)}$ is trained on one of the mutually disjoint subsets $D^{(j)} = \left\{ \left(\mathbf{x}_i^{(j)}, y_i^{(j)} \right) \right\}_{i=1}^{g} \subset D$, for $j = 1, ..., q$.

III. Parameterized Kernel Ensemble Methods

In order to construct our population of submodels, *parameterized kernel models* will be considered in this paper. These models are of the same form as the models arising in theories like Gaussian Processes [4], Regularization Networks [3] and Least Squares Support Vector Machines [7], however, no primal-dual interpretations whatsoever are made.

Consider ensembles of estimators $\hat{f}^{(j)}\left(\mathbf{x}; D^{(j)}\right)$ where the individual models have the form

$$\hat{f}^{(j)}\left(\mathbf{x}; D^{(j)}\right) = \sum_{p=1}^{g} \alpha_p^{(j)} k(\mathbf{x}_p^{(j)}, \mathbf{x}). \tag{3}$$

The function $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R} : (\mathbf{x}, \mathbf{x}') \mapsto k(\mathbf{x}, \mathbf{x}')$ is a Mercer kernel and $\alpha^{(j)} = [\alpha_1^{(j)} ... \alpha_g^{(j)}]^T$ are the parameters of the model $\hat{f}^{(j)}$ that we have to estimate. The optimal parameters for each individual submodel can be found based on the data set $D^{(j)}$ by solving the linear system

$$K^{(j)} \alpha^{(j)} = \mathbf{y}^{(j)}, \tag{4}$$

where outcome variables of the training points of subset $D^{(j)}$ are contained in the vector $\mathbf{y}^{(j)}$. Because we use a Mercer kernel, $K^{(j)}$ is the positive semidefinite Gram-matrix where $K_{lm}^{(j)} = k(\mathbf{x}_l^{(j)}, \mathbf{x}_m^{(j)})$, $l, m = 1...g$ with $\left(\mathbf{x}_l^{(j)}, y_l^{(j)}\right), \left(\mathbf{x}_m^{(j)}, y_m^{(j)}\right) \in D^{(j)}$. The convex cost function $U^{(j)}\left(\alpha^{(j)}\right)$ corresponding to the training of the submodels follows, from the conjugate gradient algorithm objective for solving the corresponding linear system and is given by

$$\min_{\alpha^{(j)}} \alpha^{(j)T} \left(K^{(j)} \alpha^{(j)} - \mathbf{y}^{(j)} \right) + \frac{1}{\gamma} \left\| \alpha^{(j)} \right\|_2^2. \tag{5}$$

Notice that we added an extra term in the cost function which acts as regularization and can be controlled by the hyperparameter γ. From a numerical point of view this term acts as a *jitter factor* which ensures that the Hessian matrix of this optimization problem is positive definite.

The training of the ensemble corresponds to minimizing the sum of the individual cost functions $\min_{\alpha^{(j)}} \sum_{j=1}^{q} U^{(j)}(\alpha^{(j)})$. The solution of this optimization is

$$\begin{bmatrix} H^{(1)} & 0 & \cdots & 0 \\ 0 & H^{(2)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & H^{(q)} \end{bmatrix} \begin{bmatrix} \alpha^{(1)} \\ \alpha^{(2)} \\ \vdots \\ \alpha^{(q)} \end{bmatrix} = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(q)} \end{bmatrix}, \tag{6}$$

where $H^{(j)} = K^{(j)} + \frac{2}{\gamma} I_g$. Ensembles of submodels found via this training will be called here *Uncoupled Ensembles* as opposed to coupled versions discussed in the next section. The overall committee model, based on (1), becomes

$$\hat{f}\left(\mathbf{x}, D\right) = \sum_{j=1}^{q} \beta^{(j)} \sum_{p=1}^{g} \alpha_p^{(j)} k(\mathbf{x}_p^{(j)}, \mathbf{x}). \tag{7}$$

IV. Ensemble Learning by Coupled Minimizers using a Synchronization Set

We now consider an *overparameterization* of our individual submodels. Each submodel j is constructed based on a Gram matrix on the respective training set $D^{(j)}$ and on the *adjoining subsets* $D^{(j-1)}$ and $D^{(j+1)}$. The submodels have the form

$$\hat{f}^{(j)}\left(\mathbf{x}; \tilde{\alpha}^{(j)}\right) = \tilde{\alpha}^{(j)T} \tilde{\mathbf{k}}^{(j)}(\mathbf{x}), \tag{8}$$

where $\tilde{\mathbf{k}}^{(j)}(\mathbf{x}) = [k(\mathbf{x}_1^{(j-1)}, \mathbf{x}) \dots k(\mathbf{x}_g^{(j-1)}, \mathbf{x})\ k(\mathbf{x}_1^{(j)}, \mathbf{x}) \dots k(\mathbf{x}_g^{(j)}, \mathbf{x})\ k(\mathbf{x}_1^{(j+1)}, \mathbf{x}) \dots k(\mathbf{x}_g^{(j+1)}, \mathbf{x})]^T$ and $\tilde{\alpha}^{(j)} \in \mathbb{R}^{3g \times 1}$. This is done in a circular way, such that $\mathbf{x}_i^{(0)} = \mathbf{x}_i^{(q)}$ and $\mathbf{x}_i^{(q+1)} = \mathbf{x}_i^{(1)}$.

Finding the optimal parameters for this model will be done based only on its training set $D^{(j)}$

$$\tilde{K}^{(j)} \tilde{\alpha}^{(j)} = \left[K^{(j,j-1)}\ K^{(j,j)}\ K^{(j,j+1)} \right] \tilde{\alpha}^{(j)} = \mathbf{y}^{(j)}, \tag{9}$$

where $K^{(i,j)} \in \mathbb{R}^{g \times g}$ is a kernel matrix build on the subsets $D^{(i)}$ and $D^{(j)}$ defined by $K_{kl}^{(i,j)} = k(\mathbf{x}_k^{(i)}, \mathbf{x}_l^{(j)})$, $k, l = 1, ..., g$ and $\left(\mathbf{x}_k^{(i)}, y_k^{(i)}\right) \in D^{(i)}$ and $\left(\mathbf{x}_l^{(j)}, y_l^{(j)}\right) \in D^{(j)}$. Solving this underdetermined linear systems does not have a unique solution. The solution with the smallest 2-norm is give by computing the solution of this underdetermined linear system via the Moore-Penrose pseudo-inverse. This can also be found by minimizing the following optimization function $U^{(j)}\left(\tilde{\alpha}\right)$

$$\min_{\alpha^{(j)}} \left\| \tilde{K}^{(j)} \tilde{\alpha}^{(j)} - y^{(j)} \right\|_2^2 + \frac{1}{\gamma} \left\| \tilde{\alpha}^{(j)} \right\|_2^2. \tag{10}$$

Similar to (5), this cost function is also a combination of an error minimization and a regularization term.

The idea of Coupled Local Minimizers is that different optimization processes are combined with an extra coupling between them. This was already used to solve non-convex optimization problems by multi-start procedures so that they are forced to converge to the same optimum [8], thereby realizing a form of global optimization. In this work we use a related idea to train the different learning algorithms. Each of the q submodels is trained by minimizing a cost function. Since training sets $D^{(j)}$ are all coming from the same distribution, we expect the submodels to give similar results on the same input $\mathbf{x}$. This can be used as an extra coupling between the submodels. We will use this constraint as an extra *output coupling* between the models on a synchronization set $D_{syn} = \{\mathbf{x}_i\}_{i=1}^{Nsyn}$. Notice that we don't need the output values of the synchronization points. This means that we can choose the points randomly. In the same way as given in Section III, we now have the ability of creating an ensemble of these models with a circular coupling via:

$$\min_{\tilde{\alpha}^{(j)}, e_i^{(j)}} \sum_{j=1}^{q} \left\| \tilde{K}^{(j)} \tilde{\alpha}^{(j)} - y^{(j)} \right\|_2^2 + \frac{1}{\gamma} \left\| \tilde{\alpha}^{(j)} \right\|_2^2 + \frac{\nu}{2} \sum_{j=1}^{q} \sum_{i=1}^{N_{syn}} \left(e_i^{(j)} \right)^2,$$
$$\text{s.t. } e_i^{(j)} = \tilde{\alpha}^{(j)T} \tilde{\mathbf{k}}^{(j)} (\mathbf{x}_i) - \tilde{\alpha}^{(j+1)T} \tilde{\mathbf{k}}^{(j+1)} (\mathbf{x}_i), \quad i = 1, ..., N_{syn}, j = 1, ..., q. \quad (11)$$

The coupling strength can be adjusted by the coupling parameter ν. Notice that due to the strict convexity of this optimization problem, it has a unique solution. This solution can be found via the Karush-Kuhn-Tucker optimality conditions. After some matrix algebra this can be rewritten into the following linear systems:

$$\begin{bmatrix} H^{(1)} & -G^{(1,2)} & 0 & \cdots & 0 & -G^{(1,q)} \\ -G^{(2,1)} & H^{(2)} & -G^{(2,3)} & & & 0 \\ 0 & -G^{(3,2)} & H^{(3)} & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & 0 \\ 0 & & & \ddots & \ddots & -G^{(q-1,q)} \\ -G^{(q,1)} & 0 & \cdots & 0 & -G^{(q,q-1)} & H^{(q)} \end{bmatrix} \begin{bmatrix} \tilde{\alpha}^{(1)} \\ \tilde{\alpha}^{(2)} \\ \vdots \\ \\ \tilde{\alpha}^{(q)} \end{bmatrix} = \begin{bmatrix} 2\tilde{K}^{(1)T} \mathbf{y}^{(1)} \\ 2\tilde{K}^{(2)T} \mathbf{y}^{(2)} \\ \vdots \\ \\ 2\tilde{K}^{(q)T} \mathbf{y}^{(q)} \end{bmatrix}, \quad (12)$$

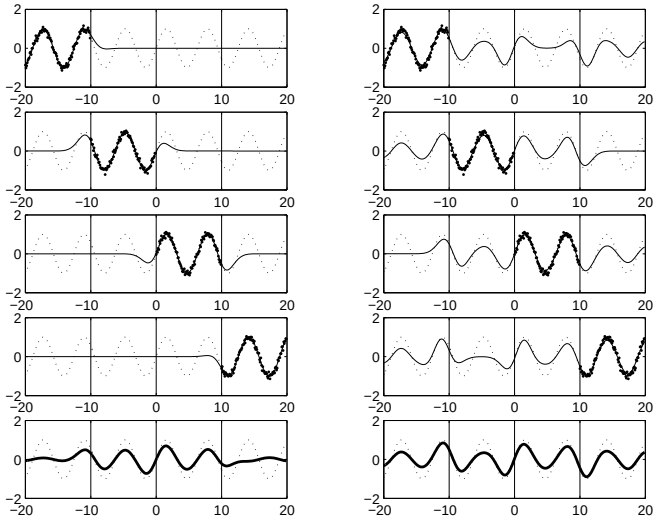


Fig. 1. In this figure we see the 4 submodels of the coupled (right) and uncoupled (left) ensembles for a sine function (true function: dotted) with their corresponding training sets. The bottom figures illustrate the combined (GEM) models of both ensembles: coupled (right), uncoupled (left).

where $H^{(j)} = 2(\tilde{K}^{(j)^T} \tilde{K}^{(j)} + \frac{1}{\gamma} I_{3g} + G^{(j,j)})$ and $G^{(q,w)} = \nu \sum_{i=1}^{N_{syn}} \tilde{\mathbf{k}}^{(q)} (\mathbf{x}_i) \tilde{\mathbf{k}}^{(w)} (\mathbf{x}_i)^T$. This equation has the same form as in (6) with extra coupling terms in the off-diagonal blocks. The sparseness in this large positive definite systems, with a memory complexity of $\mathcal{O}(g(q + 2(q-1) + 2))$, can be exploited for large data sets.

V. Examples

A. Sine function

In this first example we give a comparison of a function estimation for a sine function between models constructed via an uncoupled ensemble (as defined in III) versus a coupled overparameterized ensemble (as defined in IV). We did the training of the submodels on 4 disjoint subsets locally sampled in the input space. Although this local sampling of the input space is never done in practice, we will use it here for demonstration purposes. The synchronization points are uniformly spread over the x-axes. One can clearly see in Figure 1 that the individual models of the coupled overparameterized system (right) have good performances in the regions where they are not trained. Compared to the individual models of the uncoupled ensemble (left), they clearly give a bad performance in region where there were no training points.

In the bottom part of Figure 1, we constructed a weighted combination (GEM) based on these ensembles.

We see that the performance of the coupled overparameterized models is better than the uncoupled ensemble.

This example illustrates that coupling increases generalization performance of the submodels. This is caused by the information exchange between the different models. Averaging individual models with a better generalization will most probably lead to a better performance of the ensemble model.

B. Boston housing data set

The Boston housing data set is a multivariate regression data set of 506 cases in 14 attributes. The task is to predict the median value of the price of a home (MEDV). In these tests we do a regression on the data sets with the models constructed via a simple ensemble and via a coupled overparameterized ensemble on randomly sampled subsets. We trained on 400 training points and use the resting 106 points as test set. As synchronization set we use 10 % of the total training data, randomly chosen, without the y-values. In both cases we used the Gaussian radial base function as kernel. The used hyperparameters found by cross-validation for the individual models are $(\sigma; \gamma) = (3.37; 49, 80)$. The coupling parameter is $\nu = 1$.

To test the difference in performance between the methods, we computed the mse on a test set for 100 randomizations of the Boston data set the MEDV estimations and compared their distributions (see Figure 2). The final results are compared based on a Wilcoxon ranksum test which tells us that if there is a statistically significantly difference between the averages of the two groups. The test on the MEDV Boston data showed a statistically significant difference with a p-value of 0.0131. From these results we may conclude that the coupling leads to statistically significant better performance with 95% probability in the case of a GEM-based coupling with a random choice of synchronization set.

VI. Conclusions

In traditional committee networks the individual submodels are trained independently from each other. We have shown how improvements can be obtained by considering a form of coupled learning between the submodels. This enables the individual models of the ensemble to have better generalization performances which on itself will lead to better ensemble methods. The proposed models were tested on different data sets and showed an statistically significant improvements. These results are promising for training of large data sets in kernel based learning.

Acknowledgements. B. Hamers is a Research Assistant with the I.W.T. (Flemish Institute for Scientific and Technological Research in Industry). J.A.K. Suykens is a Postdoctoral Researcher with the F.W.O. (Fund for Scientific Research-Flanders). This work was supported by grants and funding from the Research Council of the K.U.Leuven (PhD/postdoc grants, IDO , GOA Mefisto-666), the Foundation for Scientific Research Flanders (FWO)(PhD and postdoc grants and projects (G.0115.01, G.0197.02, G.0407.02, G.0080.01), research communities (ICCoS, ANMMM)), the Flemish regional government (Bilateral Coll., IWT-project Soft4s, Eureka projects Synopsis, Impact, FLiTE, STWW-project Genprom, GBOU-project McKnow, IWT PhD grants), the Belgian Federal Government (DWTC, IUAP IV-02, IUAP V-22, Sustainable Development MD/01/024), the European Commission (TMR Alapedes, Ernsi) and industry supported direct contract research (Electrabel, ELIA, ISMC, Data4s).

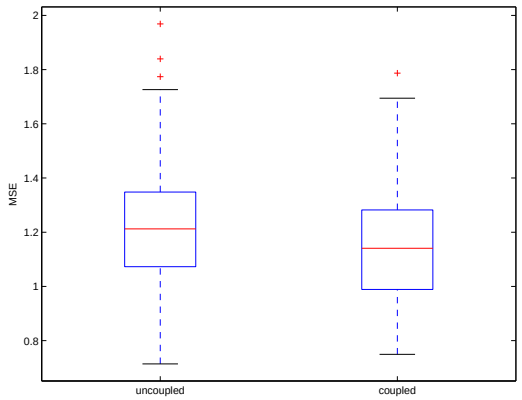


Fig. 2. MSE on test sets of the Boston Housing dataset for uncoupled (left) and coupled (right) ensemble models after 100 randomizations, showing improvements for the ensemble model with coupled learning.

References

[1] C.M. Bishop. *Neural Networks for Pattern Recognition*, Oxford University Press, 1995.

[2] L. Breiman. Bagging Predictors. *Machine Learning*, 24(2): 123–140, 1996.

[3] T. Evgeniou and M. Pontil and T. Poggio Regularization Networks and Support Vector Machines. *Advances in Computational Mathematics*, 13(1): 1–50, 2000.

[4] D.J.C. MacKay. Introduction to Gaussian processes. *Neural networks and machine learning*, 133–165. Computer and Systems Sciences, (Ed. C.M. Bishop), Springer NATO-ASI Series F, Vol.168, 1998.

[5] R.E. Schapire. A brief introduction to boosting. In *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, 1401–1406, 1999.

[6] B. Schölkopf and A. Smola. *Learning with Kernels*, MIT Press, Cambridge, MA, 2002.

[7] J.A.K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, and J. Vandewalle. *Least Squares Support Vector Machines*. World Scientific, Singapore, 2002.

[8] J.A.K. Suykens, J. Vandewalle, and B. De Moor. Intelligence and cooperative search by coupled local minimizers. *International Journal of Bifurcation and Chaos*, 11(8): 2133–2144, 2001.