Citation for the published book chapter:

Cristofer Englund and Antanas Verikas

”A SOM based model combination strategy”

In: Advances in Neural Networks – ISNN 2005. Berlin/Heidelberg: Springer, 2005, pp. 461-466

URL: http://dx.doi.org/10.1007/11427391_73

A SOM based model combination strategy

Cristofer Englund[1] and Antanas Verikas[1,2]

[1] Intelligent Systems Laboratory, Halmstad University, Box 823,
S-301 18 Halmstad, Sweden
cristofer.englund@ide.hh.se

[2] Department of Applied Electronics, Kaunas University of Technology, Studentu 50,
LT-3031, Kaunas, Lithuania
antanas.verikas@ide.hh.se

Abstract. A SOM based model combination strategy, allowing to create adaptive—data dependent—committees, is proposed. Both, models included into a committee and aggregation weights are specific for each input data point analyzed. The possibility to detect outliers is one more characteristic feature of the strategy.

1 Introduction

A variety of schemes have been proposed for combining multiple models into a committee. The approaches used most often include averaging [1], weighted averaging [1, 2], the fuzzy integral [3, 4], probabilistic aggregation [5], and aggregation by a neural network [6]. Aggregation parameters assigned to different models as well as models included into a committee can be the same in the entire data space or can be different—*data dependent*—in various regions of the space [1, 2]. The use of data-dependent schemes, usually provides a higher estimation accuracy [2, 7, 8].

In this work, we further study data-dependent committees of models. The paper is concerned with a set of neural models trained on different data sets. We call these models *specialized*. The training sets of the specialized models may overlap to varying, sometimes considerable, extent. The specialized models implement approximately the same function, however only approximately. The unknown underlying functions may slightly differ between the different specialized models. However, the functions may also be almost identical for some of the models. In addition to the set of specialized models, a *general* model, trained on the union of the training sets of the specialized models, is also available. On average, when operating in the regions of their expertise, the specialized models provide a higher estimation accuracy than the general one. However, the risk of extrapolation is much higher in the case of specialized models than when using the general model. Since training data sets of the specialized models overlap to some extent, a data point being in the extrapolation region for one specialized model may be in the interpolation region for another model. Moreover, since the underlying functions that are to be implemented by some of the specialized models may be almost identical, we can expect boosting the estimation accuracy by aggregating appropriate models into a committee. It all goes to show that an adaptive—possessing data dependent structure—committee is required. Depending on a data point being analyzed, appropriate specialized models should be detected and aggregated into a committee. If for a particular data point extrapolation is encountered for all the specialized models—*outlying* data point for the specialized models—the committee should be made of only the general model. We utilize a SOM [9] for attaining such adaptive behaviour of the committee. Amongst the variety of tasks a SOM has been applied to, outlier detection and model combination are also on the list. In the context of model combination, a SOM has been used as a tool for subdividing a task into subtasks [10]. In this work, we employ a SOM for obtaining committees of an adaptive, data dependent, structure.

2 The Approach

We consider a non-linear regression problem and use a one hidden layer perceptron as a specialized model. Let $\mathbf{T}_i = \{(\mathbf{x}_i^1, \mathbf{y}_i^1), (\mathbf{x}_i^2, \mathbf{y}_i^2), ..., (\mathbf{x}_i^{N_i}, \mathbf{y}_i^{N_i})\}$, $i = 1, ..., K$ be the learning data set used to train the ith specialized model network, where $\mathbf{x} \in \Re^n$ is an input vector, $\mathbf{y} \in \Re^m$ is the desired output vector, and N_i is the number of data points used to train the ith network. The learning set of the general model—also a one hidden layer perceptron—is given by the union $\mathbf{T} = \bigcup_{i=1}^{K} \mathbf{T}_i$ of the sets $\mathbf{T}_i$. Let $\mathbf{z} \in \Re^{n+m}$ be a centered concatenated vector consisting of $\mathbf{x}$ augmented with $\mathbf{y}$. Training of the prediction committee then proceeds according to the following steps.

1. Train the specialized networks using the training data sets $\mathbf{T}_i$.
2. Train the general model using the data set $\mathbf{T}$.
3. Calculate eigenvalues λ_i ($\lambda_1 > \lambda_2 > ... > \lambda_{n+m}$) and the associated eigenvectors $\mathbf{u}_i$ of the covariance matrix $\mathbf{C} = \frac{1}{N} \sum_{j=1}^{N} \mathbf{z}_j \mathbf{z}_j^T$, where $N = \sum_{i=1}^{K} N_i$.
4. Project the $N \times (n+m)$ matrix $\mathbf{Z}$ of the concatenated vectors $\mathbf{z}$ onto the first M eigenvectors $\mathbf{u}_k$, $\mathbf{A} = \mathbf{ZU}$.
5. Train a 2–D SOM using the $N \times M$ matrix $\mathbf{A}$ of the principal components by using the following adaptation rule:
$$\mathbf{w}_j(t+1) = \mathbf{w}_j(t) + \alpha(t) h(j^*, j; t)[\mathbf{a}(t) - \mathbf{w}_j(t)] \tag{1}$$
where $\mathbf{w}_j(t)$ is the weight vector of the jth unit at the time step t, $\alpha(t)$ is the decaying learning rate, $h(j^*, j; t)$ is the decaying Gaussian neighbourhood, and j^* stands for the index of the winning unit.
6. Map each data set $\mathbf{A}_i$ associated with $\mathbf{T}_i$ on the trained SOM and calculate the hit histograms.
7. Low-pass filter the histograms by convolving them with the following filter $h(n)$ as suggested in [11]:
$$h[n] = (M - |n|)/M \tag{2}$$

where $2M+1$ is the filter size. The convolution is made in two steps, first in the horizontal and then in the vertical direction. The filtered signal $y[n]$ is given by

$$y[n] = x[n] * h[n] = \sum_{m=-M}^{M} x[n-m]h[m] \tag{3}$$

8. Calculate the discrete probability distributions from the filtered histograms:

$$P_{ij} = P(\mathbf{a} \in \mathcal{S}_j) = \frac{\mathtt{card}\{k|\mathbf{a}_k \in \mathcal{S}_j\}}{N_i}, \quad i = 1, ..., K \tag{4}$$

   where $\mathtt{card}\{\bullet\}$ stands for the cardinality of the set and $\mathcal{S}_j$ is the *Voronoi region* of the jth SOM unit.
9. For each specialized model $i = 1, ..., K$ determine the expertise region given by the lowest acceptable P_{ij}.

In the operation mode, processing proceeds as follows.

1. Present $\mathbf{x}$ to the specialized models and calculate outputs $\widehat{\mathbf{y}}_i$, $i = 1, ..., K$.
2. Form K centered $\mathbf{z}_i$ vectors by concatenating the $\mathbf{x}$ and $\widehat{\mathbf{y}}_i$.
3. Project the vectors onto the first M eigenvectors $\mathbf{u}_k$.
4. For each vector of the principle components $\mathbf{a}_i$, $i = 1, ..., K$ find the best matching unit ij^* on the SOM.
5. Aggregate outputs of those specialized models $i = 1, ..., K$, for which $P_{ij^*} \geq \beta_i^1$, where β_i^1 is a threshold. If

$$P_{ij^*} < \beta_i^1 \ \ \forall i \quad \mathtt{and} \quad P_{ij^*} \geq \beta_i^2 \ \ \exists i \tag{5}$$

   where the threshold $\beta_i^2 < \beta_i^1$, use the general model to make the prediction. Otherwise use the general model and make a warning about the prediction.

We use two aggregation schemes, namely averaging and the weighted averaging. In the weighted averaging scheme, the committee output $\widehat{\mathbf{y}}$ is given by

$$\widehat{\mathbf{y}} = \frac{\sum_i v_i \widehat{\mathbf{y}}_i}{\sum_i v_i} \tag{6}$$

where the sum runs over the selected specialized models and the aggregation weight $v_i = P_{ij^*}$.

3 Experimental Investigations

The motivation for this work comes from the printing industry. In the offset lithographic printing, four inks—cyan (C), magenta (M), yellow (Y), and black (K)—are used to create multicoloured pictures. The print is represented by C, M, Y, and K dots of varying sizes on thin metal plates. These plates are mounted on press cylinders. Since both the empty and areas to be printed are on the same

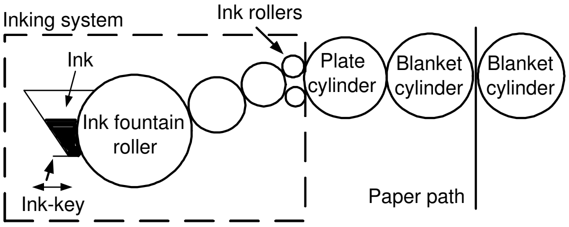

Fig. 1. A schematic illustration of the ink-path.

plane, they are distinguished from each other by ones being water receptive and the others being ink receptive. During printing, a thin layer of water is applied to the plate followed by an application of the corresponding ink. The inked picture is transferred from a plate onto the blanket cylinder, and then onto the paper. Fig. 1 presents a schematic illustration of the ink-path.

Ink feed control along the page is accomplished in narrow—about 4 cm wide—the so call ink zones. Thus, up to several tens of ink zones can be found along a print cylinder. The amount of ink deposited on the paper in each ink zone is determined by the opening of the corresponding ink-key—see Fig. 1. The ink feed control instrumentation is supposed to be identical in all the ink zones. However, some discrepancy is always observed. The aim of the work is to predict the initial settings of the instrumentation for each of the four inks in different ink zones depending on the print job to be run. The accurate prediction is very valuable, since the waste of paper and ink is minimized. Due to possible discrepancies between the instrumentation of different ink zones, we build a separate—specialized—neural model for each ink zone. A general model, exploiting data from all the ink zones is also built.

In this work, we consider prediction of the settings for only cyan, magenta, and yellow inks. The setting for black ink is predicted separately in the same way. Thus, there are three outputs in all the model networks. Each model has seventeen inputs characterizing the ink demand in the actual and the adjacent ink zones for C, M, and Y inks, the temperature of inks, the printing speed, the revolution speed of the C, M, and Y ink fountain rollers, and the $L^*a^*b^*$ values [12] characterizing colour in the test area. Thus, the concatenated vector $\mathbf{z}_i$ contains 20 components. The structure of all the model networks has been found by cross-validation. To test the approach, models for 12 ink zones have been built. About 400 data points were available from each ink zone. Half of the data have been used for training, 25% for validation, and 25% for testing.

There are five parameters to set for the user, namely, the number of principal components used, the size of the filtering mask, the SOM size, and the thresholds β_i^1 and β_i^2. The SOM training was conducted in the way suggested in [9]. The number of principal components used was such that 95% of the variance in the data set was accounted for. The SOM size is not a critical parameter. After some experiments, a SOM of 12 × 12 units and the filtering mask of 3 × 3 size were adopted. The value of $\beta_i^2 = 0$ has been used, meaning that a prediction result was always delivered. The value of β_i^1 was such that for 90% of the training data the specialized models were utilized.

Fig. 2 presents the distribution of the training data coming from all the specialized models on the 2–D SOM before and after the low-pass filtering of the distribution. As it can be seen from Fig. 2, clustering on the SOM surface becomes more clear after the filtering.

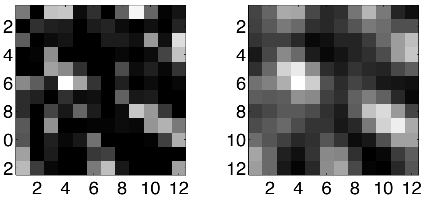

Fig. 2. The distribution of the training data on the 2–D SOM before (left) and after (right) the low-pass filtering.

Fig. 3 illustrates low-pass filtered distributions of training data coming from four different specialized models. The images placed on the left-hand side of Fig. 3 are quite similar. Thus, we can expect that functions implemented by these models are also rather similar. By contrast, the right-hand side of the figure exemplifies two quite different data distributions.

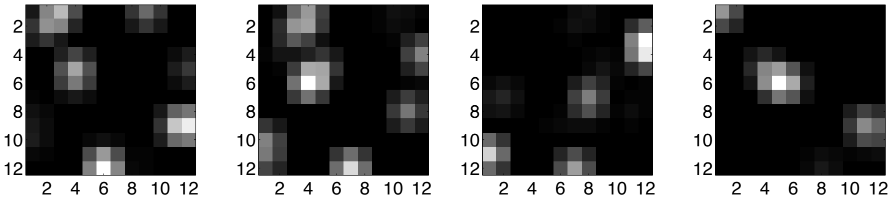

Fig. 3. The low-pass filtered distributions of the training data of four specialized models on the 2–D SOM.

Table 1 presents the average prediction error $\overline{E}$, the standard deviation of the error σ, and the maximum prediction error E^{max} for 209 data samples from the test data set. The data points chosen are "difficult" for the specialized models, since they are situated on the borders of their expertise. As it can be seen, an evident improvement is obtained from the use of the committees. The weighted committees is more accurate than the averaging one.

Table 1. Performance of the Specialized, General, and Committee models estimated on 209 unforseen test set data samples.

Model	$\overline{E}_c$ (σ_c)	$\overline{E}_m$ (σ_m)	$\overline{E}_y$ (σ_y)	E_c^{max}	E_m^{max}	E_y^{max}
Specialized	2.05 (1.52)	8.23 (4.43)	3.23 (0.75)	5.89	17.99	4.33
General	1.96 (2.13)	3.90 (2.87)	3.28 (1.19)	5.21	9.07	5.24
Committee (averaging)	0.79 (0.61)	3.63 (3.61)	1.35 (0.70)	1.72	9.49	2.12
Committee (weighted)	0.75 (0.63)	2.73 (2.46)	1.23 (0.66)	1.72	7.75	2.12

4 Conclusions

We presented an approach to building adaptive—data dependent—committees for regression analysis. The developed strategy of choosing relevant, input data point specific, committee members and using data dependent aggregation weights proved to be very useful in the modelling of the offset printing process. Based on the approach proposed, the possibility to detect outliers in the input-output space is easily implemented, if required.

References

1. Taniguchi, M., Tresp, V.: Averaging regularized estimators. Neural Computation **9** (1997) 1163–1178
2. Verikas, A., Lipnickas, A., Malmqvist, K., Bacauskiene, M., Gelzinis, A.: Soft combination of neural classifiers: A comparative study. Pattern Recognition Letters **20** (1999) 429–444
3. Gader, P.D., Mohamed, M.A., Keller, J.M.: Fusion of handwritten word classifiers. Pattern Recognition Letters **17** (1996) 577–584
4. Verikas, A., Lipnickas, A.: Fusing neural networks through space partitioning and fuzzy integration. Neural Processing Letters **16** (2002) 53–65
5. Kittler, J., Hatef, M., Duin, R.P.W., Matas, J.: On combining classifiers. IEEE Trans Pattern Analysis and Machine Intelligence **20** (1998) 226–239
6. Kim, S.P., Sanchez, J.C., Erdogmus, D., Rao, Y.N., Wessberg, J., Principe, J.C., Nicolelis, M.: Divide-and-conquer approach for brain machine interfaces: nonlinear mixture of competitive linear models. Neural Networks **16** (2003) 865–871
7. Woods, K., Kegelmeyer, W.P., Bowyer, K.: Combination of multiple classifiers using local accuracy estimates. IEEE Trans Pattern Analysis Machine Intelligence **19** (1997) 405–410
8. Verikas, A., Lipnickas, A., Malmqvist, K.: Selecting neural networks for a committee decision. International Journal of Neural Systems **12** (2002) 351–361
9. Kohonen, T.: Self-Organizing Maps. 3 edn. Springer-Verlag, Berlin (2001)
10. Griffith, N., Partridge, D.: Self-organizing decomposition of functions. In Kittler, J., Roli, F., eds.: Lecture Notes in Computer Science. Volume 1857. Springer-Verlag Heidelberg, Berlin (2000) 250–259
11. Koskela, M., Laaksonen, J., Oja, E.: Implementing relevance feedback as convolutions of local neighborhoods on self-organizing maps. In Dorronsoro, J.R., ed.: Lecture Notes in Computer Science. Volume 2415. Springer-Verlag Heidelberg (2002) 981–986
12. Hunt, R.W.G.: Measuring Colour. Fountain Press (1998)