ENTOOL - A TOOLBOX FOR ENSEMBLE MODELING

J.D. Wichard and M.J. Ogorzałek

AGH - University of Mining and Metallurgy
al. Mickiewicza 30
30-059 Kraków, Poland

C. Merkwirth

Max-Planck-Institut für Informatik
Stuhlsatzenhausweg 85
66123 Saarbrücken, Germany

What is ENTOOL ?

ENTOOL is a toolbox for ensemble modeling. It was designed to detect the deterministic structure in short and noisy time series or multi-channel measurements. It is written in MATLAB and partly in **C**++. The object-oriented implementation provides a transparent model building strategy and allows the user the addition of his own model classes.

Building Ensemble Models

ENTOOL extends the ensemble learning approach [1] for neural networks by averaging the output of several different models $f_k(\mathbf{x})$ in order to improve generalization in the regression problem $y = f(\mathbf{x}) + noise$. We take the weighted average

$$\bar{f}(\mathbf{x}) = \sum_{k=1}^{K} w_k f_k(\mathbf{x}), \qquad \text{with} \quad \sum_k w_k = 1$$

and decompose the error $e(\mathbf{x}) = (y - \bar{f}(\mathbf{x}))^2$ in two parts

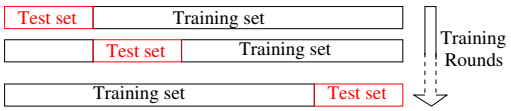
$$\bar{\epsilon}(\mathbf{x}) = \sum_{k=1}^{K} w_k (y - f_k(\mathbf{x}))^2 \qquad \text{(ensemble error)}$$

$$\bar{a}(\mathbf{x}) = \sum_{k=1}^{K} w_k (f_k(\mathbf{x}) - \bar{f}(\mathbf{x}))^2 \quad \text{(ensemble ambiguity)},$$

and write it in the form

$$e(\mathbf{x}) = \bar{\epsilon}(\mathbf{x}) - \bar{a}(\mathbf{x}) \; . \qquad (1)$$

We achieve an ensemble with high generalization ability, if we average well trained models (low ensemble error) which disagree concerning the data (high ensemble ambiguity). The training procedure is based on an extended cross validation scheme [2]. In every training round the data is divided in a training set and a test set. We train several different models using only the data in the training set and than we build the ensemble by averaging the models with the lowest out-of-sample-error regarding the test set. The following model types are implemented so far:



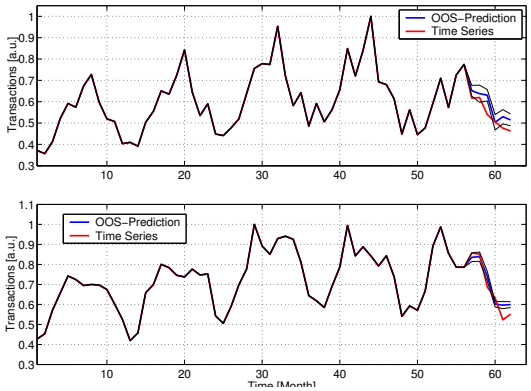
- Radial basis functions (RBF)
- Linear and polynomial regression
- K-nearest-neighbour models with adaptive metric
- Multilayer perceptron (MLP)
- Nearest trajectory models
- Multivariate adaptive regression splines

Financial Data

We applied our ensemble modeling approach to the prediction of financial time series. The time series $\{x(n)\}_{n=1,\ldots,N}$ consists of values related to monthly credit card transactions over several years. We build time lagged vectors $\vec{x}_n = (x(n), x(n-1), \ldots, x(n-d))$ in d-dimensions and train an ensemble model $f(\vec{x})$ to predict the future states

$$x(n+1) = f(\vec{x}_n) \; .$$

The prediction is iterated in the way, that the predicted values are used to build new time lagged vectors for the next step. The figure above shows the results for 30 independent



training runs. The blue line is the mean of the 30 ensemble predictions the black lines mark the standard deviation.

How to obtain ENTOOL ?

The ENTOOL toolbox and documentation is available at:
`http://chopin.zet.agh.edu.pl/~wichtel/`

References

[1] Peronne, Cooper, When networks disagree: Ensemble methods for neural networks, Neural Networks for Speech and Image Processing, Chapman Hall (1993)

[2] Krogh, Vedelsby, Neural Network Ensembles, Cross Validation and Active Learning, Advances in Neural Information Processing Systems 7, MIT Press (1995)

Acknowledgments

The work was done within the Research Training Network *COSYC of SENS* No. HPRN-CT-2000-00158 within in 5th EU Framework Program of the European Community.