

Crosstrained Ensemble of Neural Networks for Robust Time Series Prediction

Dymitr Ruta*

ABSTRACT

The time series data is increasingly being described by multiple continuous or discrete features allowing for complex non-linear regression modelling. In such systems, rather than reasoning from historical values of the time series the main effort is directed towards finding the complex relationship between features and the time series output such that the future series outputs can be calculated anytime irrespective of the previous values. The common problem with such time disparate approach is tackling a significant noise component. This work proposes a powerful time series prediction model which combines the strength of multiple neural networks (NN) regressors and the intelligent output smoothing technique developed to eliminate noise. The key strength of the model is its excellent adaptability and generalisation ability achieved through a moving-window-based cross-training of multiple NN models with injected noise. To model has been evaluated within NISIS Competition 2006 concerning prediction of continuous time-series of a catalytical reaction process described by 14 continuous chemical measures and showed outstanding predictive performance in comparison to other competitive models.

1 The Problem Formulation

The objective of the competition was to create an adaptive mathematical model describing the relationship between 14 input variables and an output variable describing catalytic oxidisation process in the multi-tube reactor. The input variables represent various measures continuously collected during the chemical process like: flow of air/gasses [kg/hr], temperature and various concentration measures, whereas the output variable represents the catalytic activity of the process. All the variables vary over time effectively forming a multidimensional time series. There was no restriction concerning methods and algorithms but an adaptive component of the predictive model was an obligatory requirement.

In the first phase of the competition 8 months of data (242 * 24 hours), both input \mathbf{x} and output y was given along with next months input data to create the model and make predictions for the catalytic activity over the next month. Then on submission of the predictions the true output values are given for this month along with the input data for the next month and the process repeats over the period of 4 months as shown in Figure . For simplicity the model performance is assessed for the predictions of only 15 output values per month representing the activity measured at the end of every second day and will be represented by the average error rate calculated as follows:

$$ERR = \sum_{j=1}^4 \frac{100}{N} \sum_{i=1}^Q \frac{|y_{true} - y_{pred}|}{|y_{true}|} f(n) \quad (1)$$

where Q stands for the number of measures per month.

2 Data preprocessing and feature selection

The available data constitute a timestamp column, 14 continuous variables and the continuous output variable taking values within the range (-1,1). One of the features (QI X ORG) took invariably the same value hence was excluded leaving the input data of 13 variables. As the data had temporal nature, at this stage the decision was needed to be made on whether the historical values of the same feature shall be used as separate features or should be discarded altogether. To decide upon this problem a simple experiment was designed to test the correlation between the historical attributes at different lags and the output variable as well as output autocorrelation. Moreover a simple multiple linear regression model was applied to test the prediction performance of up to 4 months into the future. The results of the correlation analysis are presented in Figure 2(a) that shows the evolution of the correlation coefficients between all the variables and the output variable

*British Telecommunications Group (BT), Chief Technology Office, Research & Venturing, Adastral Park, Orion MLB 1, PP12, Ipswich IP53RE, UK, dymitr.ruta@bt.com

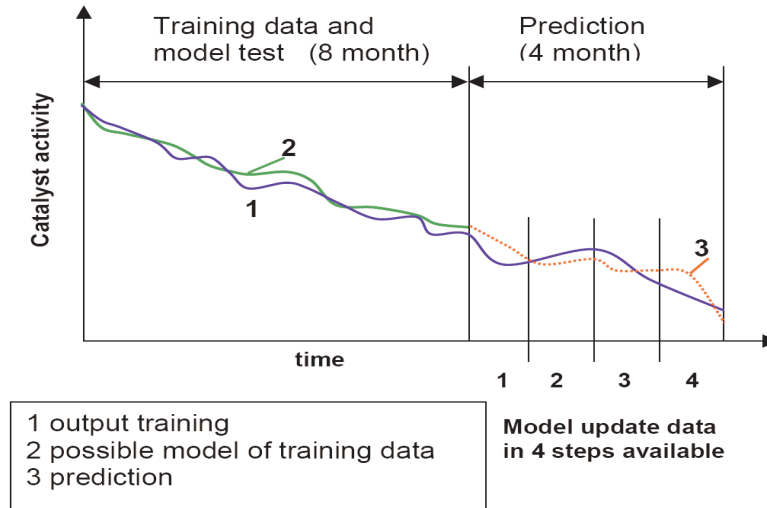
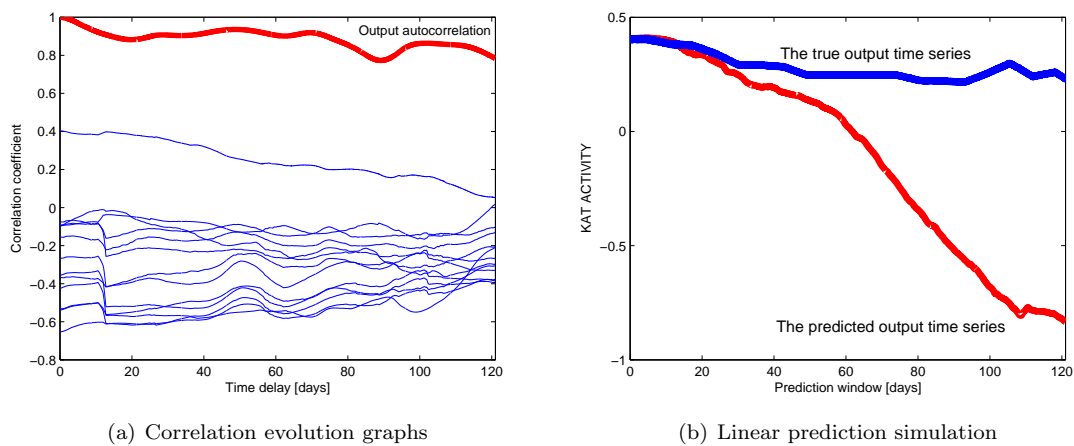


Figure 1: Visualisation of the NISIS Competition 2006 task.

over increasing time lags. It clearly shows that on its own no variable is well correlated with the output and this correlation falls for increasing time lags. Surprisingly however the autocorrelation of the output variable (marked by thick line) sustains high value even for larger time lags. If however the past values of the outputs were to be used as features for current output predictions, the error rate of such system would very quickly accumulate to a large value as shown in the simulation on Figure 2(b). This phenomenon can be explained by the accumulated temporal uncertainty inflicted by the use of features which themselves are predicted in the previous time steps.

Based on these initial tests and taking into account that the output predictions are to be made over 30 days into the future a decision was made to use only current feature values for predictions and hence eliminate the direct impact of past outputs values on its future predictions. The predictive system will therefore represent a multiple-inputs-one-output model as shown in Figure 3.



(a) Correlation evolution graphs

(b) Linear prediction simulation

Figure 2: Evolution of the correlation coefficients between input and out variables. (2(a)) Plots of correlation coefficients over increasing time lags. The self autocorrelation evolution curve is shown by thick red line. (2(b)) Output prediction based on linear regression of past inputs and outputs for increasing time lags.

3 Cross-trained Ensemble of Neural Networks

Due to continuous nature of the output variable the choice of predictive model narrows down to multiple-input one-output temporal regression problem. Neural Networks are considered to be a universal non-linear regression model with a neat complexity control mechanism and high predictive diversity that can be further encouraged by varying network initialisation, cross-training, injecting noise to the data etc. [1]. Given many diverse and well

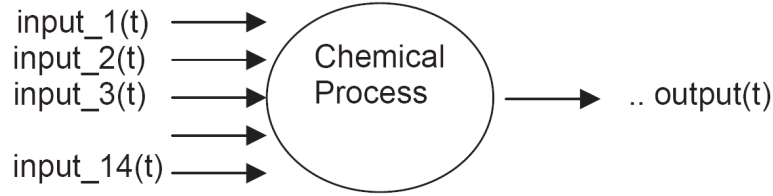


Figure 3: The chosen prediction paradigm.

performing predictors it is possible to construct an ensemble of regressors that would jointly outperform any individual regression model. There is many ways the individual NN models can be combined in the ensemble: the simplest is just by averaging the individual network outputs, other method often used is a linear combination of NN outputs [2], [3]. Although complexity of such model dramatically increases, given the performance critical nature of the predictive task and relatively small data set of about 6000 14-dimensional data points such model becomes a viable and analytically strong proposition.

Technically, individual neural network represented a Feedforward Multilayer Perceptron with iRPROP [1] training and 3 hidden layers of neurons structured as follows: [16 64 32]. The selected number M of such NN models is trained and evaluated on many different partitions of the training data following the k -fold cross-validation scheme. On average M/k models is trained on data from a single k -fold crossvalidation partition and are assigned the error rate obtained on the testing part of this partition. Overall the training process involves M individual NN learning processes an scales linearly with the size of ensemble.

After the training process a two-stage model selection is applied to construct the final ensemble. First the fixed fraction of the best models from each partition split are selected according to their regression error. Then they are pulled together and again the selection guided by the individual error rate proceeds to give the final ensemble with the desired number of NN models.

4 Intelligent Output Series Smoothing

Given the constructed ensemble it is applied to both the training set on which it was build and the validation set that has not been used during the ensemble training process. The validation set should follow the training set in the timeline. The predicted outputs are then compared with the true outputs for the whole dataset to build an optimised smoothing of the predicted time series. The smoothing model applied to the data comes in two stages. First a procedure called $\text{signal_filter}(x, k, r)$ is used to remove high-level noise component. This procedure compares the predicted signal with the bi-directional k -step moving average of this signal and replaces the original signal with the aggregated signal where the difference between the 2 signals is greater than r times standard deviation of the original signal. The resulting signal is further smoothed using the same bi-directional n -step moving average yet in generally using different step parameter of the aggregation.

K -Step bi-directional smoothing is a simple procedure applicable to time series $y(t)$ where $t = 1.., N$, which returns smoothed series $y'(t)$ such that:

$$\begin{cases} y'(t) = \frac{\sum_{i=t-k+1}^{t+k-1} y(i)}{2k-1} & \text{for } t = k, \dots, N - k + 1 \\ y'(t) = \frac{\sum_{i=1}^{t+k-1} y(i)}{t+k-1} & \text{for } t = 1, \dots, k - 1 \\ y'(t) = \frac{\sum_{i=t-k+1}^N y(i)}{N-t} & \text{for } t = N - k + 2, \dots, N \end{cases} \quad (2)$$

The three parameters of the smoothing procedure: k, r and n are optimised with respect to the regression error rate obtained for the validation set via a naive looping through all possible combination within the grid of 20 values per parameter giving the total of 8000 evaluations. The trained ensemble of NN models along with optimised smoothing parameters constitutes the fully trained model ready to make predictions.

5 Adaptive time series prediction

Subsequent monthly time series predictions have been set up using adaptive moving window scheme. Given that the current prediction series starts at time t_{cur} the model was trained on the fixed training size w ranging from $t_{cur} - w$ to $t_{cur} - 1$. The window width w has been optimised via a naive evaluation of the whole range of grided window widths from 1 to 8 months. Evaluation of the optimal parameter w has been carried out on the 8 first months of complete data and kept fixed throughout the predictive stages of the competition.

6 Experiments

The experimental part of this work forms the submission of the model predictions to the NISIS Competition 2006. As discussed in Section 1 the prediction of the catalyst activity time series was organised in 4 monthly slots following 8 months of complete data. The optimisation of the moving window width for adaptive learning showed that 8 months is the optimal window width, which suggests that the available data is still not in excess and at this stage it appears that the more data used for training the better performance. Using this 8-months window the presented model was subsequently rebuilt on the 8 months of preceding data using first 6 months for proper training and the remaining two months for validation and fine-tuning of the smoothing parameters. Interestingly, at each iteration of the model testing process significantly different different smoothing parameter were found. Initially the optimal set of smoothing parameters was (11, 0.9, 56), yet for the last prediction phase the optimal parameters turned out to be 7, 0.8, 231). It proves that whereas the global analytical engine of the model inherited from a mixture of neural networks remains data greedy, the smoothing parameters become very sensitive to the local variability of the data and hence accommodate the adaptive component of the model tuning. Visual comparison of the winning presented model predictions along with other 2 model predictions and the true output time series is presented in Figure 4, along with the complete numerical performance comparison of all the models participating in the competition. Note that the presented model outperformed immensely other competitors leaving the second best model with the twice bigger error rate.

Rank	Mth 1	Mth 2	Mth 3	Mth 4	Sum
1	20.13	21.01	12.87	19.14	73.06
2	43.41	18.38	52.31	24.14	138.26
3	63.15	17.83	33.89	28.91	143.79
4	62.56	14.48	53.75	37.05	167.86
5	81.46	29.80	33.78	23.06	168.12
6	59.01	69.91	37.21	27.54	193.75
7	24.13	87.15	54.16	28.74	194.20
8	75.97	25.48	80.20	27.77	209.43
9	77.78	84.80	32.24	22.21	217.05
10	47.87	126.71	32.59	39.45	246.64
11	52.56	135.91	35.46	57.95	281.89

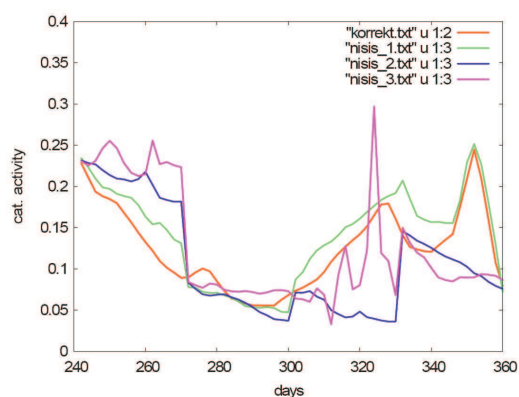


Figure 4: Performance comparison of all models submitted to NISIS 2006 Competition. (left) Error rates for all competitive models. (right) 3 best model outputs compared with the true output series.

7 Conclusions

This work promotes a new type of time series prediction when a rich evidence of multidimensional features related to the time series is available. It combines highly robust ensemble of neural network regressors with the intelligent smoothing of the highly noisy output signal. The individual models are cross-trained on different partitions of the training data with injected noise component to boost model diversity and eventually strengthen generalisation abilities of the ensemble. The key element of the model is the output smoothing components that converts highly noisy output from NN ensemble into a smooth trend adaptively fitted to the true validation output by means of optimised moving-average-based filtering and aggregation. The model performance turned out to be significantly better than any other competitive proposition leaving the second-best model with twice bigger error rate.

References

- [1] A.J.C Sharkey: Combining artificial neural nets: ensemble and modular. Springer-Verlag, London, UK, 1999
- [2] A.J.C Sharkey, N.E. Sharkey: Combining diverse neural nets. The Knowledge Engineering Review 12(3):231-247, 1997
- [3] S. Hansen, P. Salamon: Neural network ensembles. IEEE Transactions on Pattern Analysis and Machine Intelligence 12(10):993-1001, 1990