An Information-Based Approach towards Neuro-Evolution

Behzad Behzadan
University College London
Dept. of Computer Science
London, UK, WC1E 6BT
b.behzadan@cs.ucl.ac.uk

Robert Elliott Smith
University College London
Dept. of Computer Science
London, UK, WC1E 6BT
Rob.Smith@cs.ucl.ac.uk

ABSTRACT

A novel self-governing system, which is theoretically founded on information theory, is introduced with the ability of determining the optimal quantity and connectivity of the hidden-layer of a three layer feed-forward neural network. The system - called MINES - simultaneously links parameter learning (performed by back-propagation) to structural learning (performed by genetic algorithm) with the aid of mutual information between the error-space and the hidden-layer.

Categories and Subject Descriptors

1.2.6 [**Artificial Intelligence**]: Learning—*connectionism and neural nets*

Keywords

Neural Network, Neuro-Evolution, Information Theory, Forecasting

1. INTRODUCTION

The "universal approximation theorem" [2] (also called the "Kolmogrov theorem" [1]) states that, a *neural network* (NN) with one hidden layer is sufficient for any mapping problem. However, the theorem does not propose a way to realise how a single hidden layer is optimal [2].

To obtain an optimal hidden layer, a NN is required to be trained by a simultaneous consideration of the number of hidden nodes and the synaptic connections (i.e. structural learning) and the associated synaptic weights (i.e. parameter learning). Indeed, structural and parameter learnings are not two independent techniques, and there is an intertwined relation between them.

2. MINES METHODOLOGY

Mutual Information Neuro-Evolutionary System (MINES) simultaneously performs structural and parameter learning to fully determine the state of the incorporated hidden nodes. For this purpose, MINES alternately uses *genetic algorithm* (GA) and *back propagation* (BP) alongside the other, and indirectly links them by the fitness provided by the GA individuals. The GA in MINES is responsible for changing the structure, while BP reduces the cost function.

GECCO'12 Companion, July 7–11, 2012, Philadelphia, PA, USA.
ACM 978-1-4503-1178-6/12/07.

GA individuals are clustered. Individuals in each cluster have same binary variables. Each cluster is associated to a unique hidden node (even though the cluster itself may contain one or several individuals). Individuals in each cluster control the connectivity pattern of the associated hidden node. Note that, the synaptic weights of hidden nodes are not evolved by GA and they are only modified by BP.

To calculate the fitness of a GA individual, first the cluster in which the individual belongs to is found. Second, the hidden node associated to the cluster is determined. Third, the *mutual information* (MI) between the output of the hidden node and the respective *residual error* of the NN is calculated. Given a hidden node, the residual error is the remaining error of the NN after the exclusion of the hidden node from the network. The calculated MI is the fitness of the underlying GA individual and is also the fitness of all other individuals which belong to the same cluster.

Note that, the MI depends on the total state of the representing hidden node; i.e its receptive connectivity patterns controlled by GA, and the associated synaptic weights tuned by BP. This is how GA and BP are indirectly connected through MI, following by alternatively running BP for some iteration steps and subsequently applying a GA generation: (1) BP is paused, run after several iterations; (2) MIs are measured, providing the fitness values; (3) GA determines the appropriate local structure; (4) BP is resumed from usually a new point in the weight space, adjusted previously by GA; (5) GA is re-applied to the current population; (6) the whole training procedure stops when the system converges.

3. POLYNOMIAL DISCOVERY

The regression ability of MINES is tested against the task of finding the right polynomial represented by:

$$y = 2 + 3x_1x_2 + 4x_3x_4x_5 \quad (1)$$

This fitting to a polynomial is used by Saito and Nakano [5] who earlier [4] showed how to formulate a nonlinear polynomial using a FFNN. The challenge of the underlying problem is not solely to perform a regression task fitting some input data to some target values with a FFNN. The challenge is indeed to discover the exact polynomial formula regressing the input-output data. For this reason, the FFNN must be evolved into a partially connected network with the right synaptic weights representing Equation 1.

Figures 1 and 2 respectively show the *root mean squared error* (RMSE) on the test set and the number of evolved hidden nodes after the convergence of MINES for selected number of BP iterations. It is observed that the overall

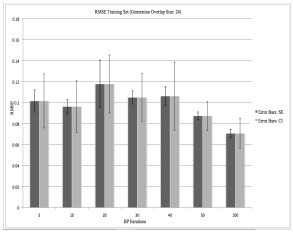

Figure 1: Test-set RMSE across various BP iterations for polynomial discovery.

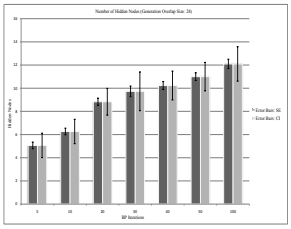

Figure 2: Evolved hidden nodes across various BP iterations for polynomial discovery.

generalisation performance of the system is very good for all number of BP iterations, but the BP iteration number of 5 and 10 result in fewer number of hidden nodes.

An optimal result obtained by MINES is presented here:

$$\begin{aligned} y &= 1.7550 \\ &+ 3.2261 {x_1}^{0.9656} {x_2}^{0.9656} {x_3}^{0.0132} {x_4}^{0.0129} {x_5}^{0.0129} \\ &+ 3.9447 {x_1}^{-0.0075} {x_2}^{-0.0074} {x_3}^{1.0057} {x_4}^{1.0056} {x_5}^{1.0056} \end{aligned} \quad (2)$$

It is seen that Equation 2 is very close to the objective polynomial introduced in Equation 1.

4. BUSINESS FORECASTING

An approach to forecast crude oil prices (similarly used by Yu et al. [7]) is presented as the real world application of MINES. Monthly *West Texas Intermediate* (WTI) crude oil prices are chosen as the input data for training set (January 1986 till December 2004) and test set (January 2005 to December 2009)[1]. MINES is trained over the training set to make one-step-ahead (one month ahead) prediction over the test set.

Prior to being fed into the system, the data is first denoised using *exponential moving average* (EMA). The denoised data is subsequently decomposed into 5 subcomponents, based on a novel technique called *Empirical Mode Decomposition* (EMD) [3]. EMD decomposes a signal into some oscillatory functions, named *intrinsic mode components* (IMC). The original data will be the sum of the IMCs.

Statistic	Distinct Connections	Hidden Nodes	NMSE	Dstat
mean	8	13	0.24	0.66
stdev	4	5	0.01	0.01

Table 1: 10 random experimental results on test set to forecast monthly WTI crude oil prices.

Table 1 shows the average and the standard deviation of the number of distinct connectivity patterns and the number of hidden nodes, obtained after the convergence of MINES for 10 random experiments on the training set. To evaluate the prediction performance, the *normalized mean squared error* (NMSE) and the *directional statistic* (Dstat) [6] over the test set has been provided. Dstat measures the percentage of the times that the forecasting direction is correct (in this case 66%). A graphical view of the forecasting performance has been demonstrated in Figure 3.

[1]Data is freely accessible from http://www.eia.doe.gov.

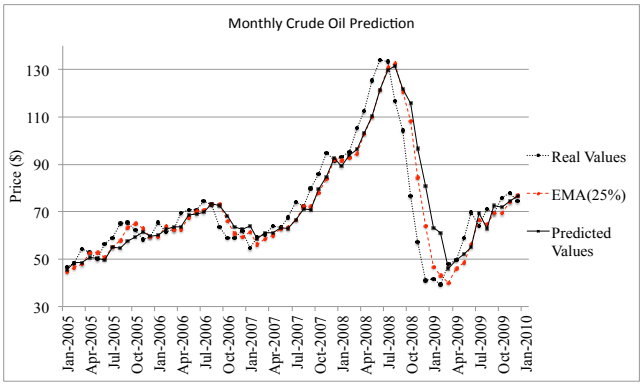

Figure 3: Monthly WTI crude oil forecasting graph.

5. CONCLUSION

MINES is built around a central hypothesis that by aligning the MI of the output of a hidden node and the corresponding *residual error* of the system, the proper receptive field connectivity pattern of the hidden node, in relation to the other incorporating elements, can be determined so that the evolving hidden layer would form an optimal, or close to optimal, FFNN.

6. REFERENCES

[1] A. N. Gorban and D. C. Wunsch II. The general approximation theorem. In *Neural Networks Proceedings, 1998. IEEE World Congress on Computational Intelligence. The 1998 IEEE International Joint Conference on*, volume 2, pages 1271–1274, Anchorage, AK , USA, 1998.

[2] S. Haykin. *Neural Networks and Learning Machines*. Prentice-Hal, third edition, 2009.

[3] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N.-C. Yen, C. C. Tung, and H. H. Liu. The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceedings: Mathematical, Physical and Engineering Sciences*, 454(1971):903–995, 1998.

[4] K. Saito and R. Nakano. Law discovery using neural networks. In *IJCAI'97: Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence*, pages 1078–1083, San Francisco, CA, USA, 1997. Morgan Kaufmann Publishers Inc.

[5] K. Saito and R. Nakano. Structuring neural networks through bidirectional clustering of weights. In *Discovery Science*, pages 206–219, 2002.

[6] J. Yao and C. L. Tan. A case study on using neural networks to perform technical forecasting of forex. *Neurocomputing*, 34(1-4):79–98, 2000.

[7] L. Yu, S. Wang, and K. K. Lai. Forecasting crude oil price with an emd-based neural network ensemble learning paradigm. *Energy Economics*, 30(5):2623–2635, September 2008.