An evolutionary method for term selection in the Group Method of Data Handling

Mutasem Hiassat, Neil Mort

Automatic Control & Systems Engineering, University of Sheffield

1 Introduction

The Group Method of Data Handling (GMDH) and Genetic Programming (GP) are two popular non-linear methods of mathematical modelling. In this paper, both methods are explained and a new implementation of the GMDH algorithm, which incorporates GP, is proposed. The new implementation is then applied to financial time series data and its results are compared with the results obtained by the standard GMDH algorithm and the more recent multi-objective GMDH algorithm (MOGMDH).
Results show that the new proposed algorithm appears to perform better than the other two algorithms.

2 The Group Method of Data Handling

The GMDH is a heuristic self-organising modelling method which was introduced by Ivakhnenko (1968) as a rival to the method of stochastic approximations. The principle adopted in constructing the model for the data is very similar to the way evolutionary mechanisms produce an organism and its off-springs. The method is particularly useful in solving the problem of modelling multi-input to single-output data.
Firstly, the data is divided into training and checking sets. This partitioning is conducted heuristically either by selecting points for each set alternately or based on their variance from the mean value. The points with high variance are used in the checking set to ensure that the selected models can extrapolate outside the data in the training set. Secondly, the data in the input matrix are taken in pairs and a quadratic polynomial between each pair, x_i and x_j, with the corresponding output, y, is formed:

$$y = a_o + a_1x_i + a_2x_j + a_3x_ix_j + a_4x_i^2 + a_5x_j^2 \quad (1)$$

The coefficients of the polynomial are found by least square fitting as given in Press *et al* (1992) using the data in the training set. The output of the polynomials is then evaluated and tested for suitability using the data points in the checking set. An external criterion, usually the mean squared error (mse), which is also known as the regularity criterion, is then used to select the polynomials that are allowed to proceed to a next layer where the outputs of the selected polynomials become the new input values. Finally, the whole procedure is repeated until the condition for terminating the GMDH run has been reached. This occurs when the lowest mse is no longer smaller than that of the previous layer. The model of the data can be computed by tracing back the path of the polynomials that corresponded to the lowest mse in each layer.
Since its introduction scientists were quick to use, explore and improve the GMDH algorithm. One of the areas that has been investigated and developed is the selection criterion. Parks *et al* (1975) used the “unbiased criterion” in predicting a model for the British Economy. They

deduced that it was a better selector than the regularity criterion. Robinson (1998) introduced a multi-objective GMDH algorithm (MOGMDH) in which the regularity criterion was used as well as three other selectors in the selection process. This resulted in a significant improvement in the performance of the GMDH algorithm. Hiassat *et al* (2003) introduced the GP-GMDH algorithm, which uses genetic programming to find the best function that maps the input to the output in each layer of the GMDH algorithm, and showed that it performs better than the conventional GMDH algorithm in time series prediction using financial and weather data.

3 Genetic Programming (GP)

Genetic programming was formally introduced by Koza (1992) as an extension to the popular genetic algorithm paradigm (Holland (1975)). Initially, a population of genetic programs is randomly generated by selecting functions, variables, which are the input data points, and constants from pre-defined sets. Their performance is then evaluated and compared with the actual solution of the problem. In the case considered in this paper, the predicted values are compared with the actual currency exchange rates. Based on their fitness, genetic operators, such as mutation and cross over, are applied to the individuals resulting in the creation of a new generation of genetic programs. The process is repeated until the required, or predefined, number of generations has been reached. The fittest individual, which is a genetic program, is considered to be the system model.
An example of a genetic program in a parse-tree structure is given in figure 1.

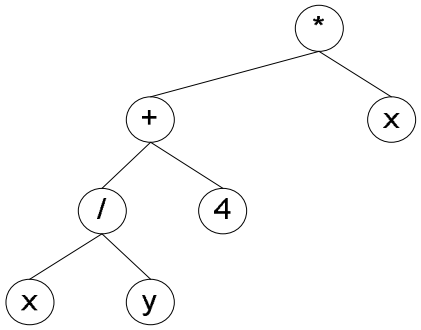

Figure 1: A parse-tree structure GP

The tree is interpreted in a depth-first, left-to-right postfix manner. The result is the expression given below.

$$x\left(\frac{x}{y}+4\right)$$

4 The proposed algorithm

It is evident from the previous two sections that both modelling methods have many common features, but, unlike the GMDH, GP does not follow a pre-determined path for input data generation. The same input data elements can be included or excluded at any stage in the evolutionary process by virtue of the stochastic nature of the selection process. A GP algorithm can thus be seen as implicitly having the capacity to learn and adapt in the search space and thus allow previously bad elements to be included if they become beneficial in the

latter stages of the search process. The standard GMDH algorithm is more deterministic and would thus discard any underperforming elements as soon as they are realised.
Using GP in the selection process of the GMDH algorithm, the model building process is free to explore a more complex universe of data permutations. This selection procedure has three main advantages over the standard selection method. Firstly, it allows unfit individuals from early layers to be incorporated at an advanced layer where they generate fitter solutions. Secondly, it also allows those unfit individuals to survive the selection process if their combinations with one or more of the other individuals produce new fit individuals, and thirdly, it allows more implicit non-linearity by allowing multi-layer variable interaction.
The new GMDH algorithm that is proposed in this paper is constructed in exactly the same manner as the standard GMDH algorithm except for the selection process. In order to select the individuals that are allowed to pass to the next layer, all the outputs of the GMDH algorithm at the current layer are entered as inputs in the GP algorithm where they are allowed to evolve, mutate, crossover and combine with other individuals in order to prove their fitness. The selected fit individuals are then entered in the GMDH algorithm as inputs at the next layer. The whole procedure is repeated until the criterion for terminating the GMDH run has been reached.

5 Results

The conventional GMDH, the MOGMDH and the new proposed GMDH are used in the prediction of two daily currency exchange rates: the US Dollar to the Japanese Yen, denoted USD2JPY, and the US Dollar to the British Pound, denoted USD2GBP. The daily rates are normalised by subtracting the mean and dividing by the standard deviation. The normalised values from 1st March, 2003 to 26th February, 2004 are used in the training and checking process of the algorithms. Then the resulting models are used to predict the exchange rate values from 27th February, 2004 to 17 March, 2004. Each of the GMDH runs is carried out for three generations based on an assumption that the current exchange rate value is dependant upon the values of no more than the previous four days.
Table 1 shows the results obtained by the three algorithms based on two performance measures: the mean percentage error and the root mean squared error.

		Mean Percentage Error	**Root Mean Squared Error**
Conventional	**USD2JPY**	0.9593	1.2876
	USD2GBP	0.9076	0.0063
MOGMDH	**USD2JPY**	0.6391	0.8725
	USD2GBP	0.8617	0.0064
New GMDH	**USD2JPY**	0.4811	0.7248
	USD2GBP	0.4555	0.0034

Table 1: Results of the 3 GMDH algorithms

It is clear that the results produced by the new algorithm are superior to those produced by the other two algorithms.
Figure 2 shows the actual exchange rate values of USD2JPY and the ones predicted by the new proposed GMDH algorithm.

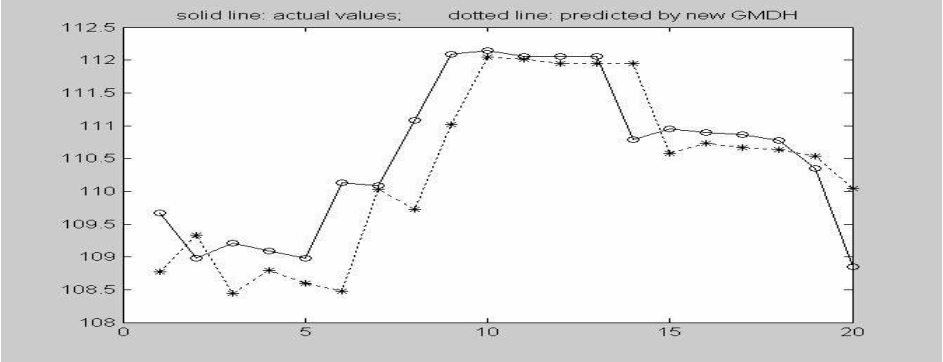

Figure 2: Actual and predicted by new GMDH values of USD2JPY exchange rate.

Conclusion

Results have shown that the performance of the conventional GMDH algorithm is significantly improved when the term selection method proposed in this paper is implemented. However, a comparative study of the terms selected by the conventional GMDH algorithm with those selected by the proposed GMDH algorithm has not been carried out in this paper. This study could prove beneficial in understanding the influence of the selection criterion on the performance of the algorithm and, hence, provide scope for further improvement in the versatile group method of data handling.

References

Hiassat M, Abbod M and Mort N. Using Genetic Programming to Improve the GMDH in Time Series Prediction, *Statistical Data Mining and Knowledge Discovery*, edited by Hamparsum Bozdogan. Chapman & Hall CRC, 2003, pp257-268.

Holland, J H (1975). "Adaption in Natural and Artificial Systems", The University of Michigan Press, Ann Arbor.

Ivakhnenko A G, (1968). The Group Method of Data Handling-A rival of the Method of Stochastic Approximation. *Soviet Automatic Control, vol 13 c/c of avtomatika,* 1, 3, 43-55.

Koza J R (1992), *Genetic Programming: on the programming of computers by means of natural selection*. MIT Press, Cambridge. (1992).

Parks P, Ivakhnenko A G, Boichuk L M and Svetalsky B K, (1975). A self-organizing model of British economy for control with optimal prediction using the Balance-of-Variables criterion. *Int. J. of Computer and Information Sciences*, vol 4, No. 4.

Press, W H, Teukolsky, S A, Vetterling, W T and Flannery, B P (1992). "*Numerical recipes in C: The art of scientific computing*". Cambridge University Press.

Robinson, C (1998). "*Multi-objective optimisation of polynomial models for time series prediction using genetic algorithms and neural networks*", PhD Thesis in the Department of Automatic Control & Systems Engineering, University of Sheffield, UK.