Heteroscedacity Elimination and Genetic Selection in GMDH

Marcel Jirina\textsuperscript{1}, Marcel Jirina, jr.\textsuperscript{2}

\textsuperscript{1} Institute of Computer Science AS CR, Czech Republic
\textsuperscript{2} Faculty of Biomedical Engineering, Czech Technical University in Prague, Czech Republic
marcel@cs.cas.cz, jirina@fbmi.cvut.cz

Abstract. The paper introduces a cloning procedure for generating clones of a given neuron that may be better than the original “parent” neuron with the parameters stated by linear regression. We show that cloning together with the use of genetic selection procedure leads to a new type of GMDH algorithm. We found that cloning is a simple and effective method for obtaining a less biased solution and faster convergence than that obtained by standard linear regression.

Keywords

MIA GMDH algorithm, linear regression, homoscedasticity, biased parameters, cloning procedure, genetic selection

I Introduction

In the well-known MIA GMDH algorithm each neuron of the GMDH network has a quadratic transfer function of two input variables that has six parameters. The process of adaptation of the GMDH network is based on standard linear regression. However, it can be found that the mathematical condition of homoscedasticity for linear regression to get unbiased results is not fulfilled. Thus all neurons have slightly biased parameters and do not give an optimal solution.

2 A new type of GMDH algorithm

Here we introduce a cloning procedure for generating clones of a given neuron that may be better than the original “parent” neuron with the parameters stated by linear regression. We show that cloning together with the use of genetic selection procedure leads to a new type of GMDH algorithm. We found that cloning is a simple and effective method for obtaining a less biased solution and faster convergence than that obtained by standard linear regression. At the same time, the use of genetic selection procedure allows all neurons already generated to remain potential parents for a new neuron. Thus the problem of deleting excessive neurons during learning disappears.

Our results demonstrate that the influence of heteroscedasticity can be easily eliminated by a simple cloning procedure and faster convergence and so better behavior of GMDH algorithm can be obtained. We suppose that our finding of heteroscedasticity in the GMDH method and its solution by cloning may lead to finding other more effective procedures based e.g. on robust approaches.

Fig. 1 Histograms of residuals for both classes (Class 0 – malignant, Class 1 - benign) and for both classes together for the breast cancer data classification problem [1].
3 Linear regression and cloning

Usually – and we do it as well - the parameters of the new neuron are set up by linear regression, i.e. with a least mean squared error method. This method uses the Gauss-Markov assumptions that due to the nonlinearity of the problem as well as the GMDH network, the assumption of homoscedasticity is not met especially for the classification problem. For illustration see in Fig. 1 the histograms of residuals, i.e. histograms of errors for one neuron and for both classes separately and for all data. There it can be seen that first, the expected value apparently is not zero, and second, residuals are heteroscedastic.

The solution obtained by linear regression can be used as the first approximation. After that we use cloning, i.e. we generate several neurons with the same inputs and with parameters a, ..., f only slightly modified, i.e. mutated with respect to their original values, and we select the best of them.

4 GMDH with genetic selection

In genetic algorithms in the selection step there is a common approach that the probability of being a parent is proportional to the value of the fitness function. Just this approach is used here. The fitness is simply a reciprocal of the mean absolute error on the validation set. The initial state form n inputs only, there are no neurons. In this state two different inputs are selected randomly with equal probability as parents of a new neuron. If there are k neurons already, a new parent is selected from inputs or from already existing neurons. The probability that an already existing neuron will be selected is inversely proportional to the value of fitness function. The fitness function is equal to the reciprocal error on the verification set. After the new neuron is formed and evaluated it can immediately become a parent for another neuron. Thus the network has no explicit layers. Each new neuron can be connected to any input or already existing neuron.

![Fig. 2 Classification errors for four methods on some data sets from the UCI MLR. Note that for Shuttle small data the errors are ten times enlarged in this graph. In legend GMC GMDH 2 means GMC GMDH method with fitness equal to the reciprocal of square of the mean absolute error.](image_url)

5 Performance analysis

The classification ability of the genetically modified GMDH algorithm with cloning (GMC GMDH) was tested using real-life tasks from the UCI Machine Learning Repository [1]. Main characteristics are summarized in Fig.2.

References