Parameter Design for Group Method Data Handling (GMDH) using Taguchi in Software Effort Estimation

Siti Hajar Arbain¹, N H Mustaffa¹, N A Ali¹ and D N A Jawawi¹

¹School of Computing, Universiti Teknologi Malaysia (UTM), 81310 Johor Bahru, Johor, Malaysia.

Email: shajar75@live.utm.my

Abstract. Recently, the use of data-driven models is becoming increasingly impactful but has proven to offer best prediction with less knowledge of the geological, hydrological, and physical process behaviour and criteria. A Group Data Handling Model (GMDH) is one of the sub-model common neural network data driven. It was first developed for complex systems with a modelling and recognition algorithm. GMDH is known as a self-organizing heuristic modelling approach. For solving modelling issues involving multiple inputs to single output data, it is very successful. While the GMDH model has been implemented in many modelling fields, some modifications in terms of parameter design have been given little attention. In other respects, Dr. Genichi Taguchi suggested that the Taguchi method for improving the process or product design with the help of significant parameter levels that influence the delivery of the product. In this paper, we evaluated the behaviour of GMDH model based on numbers of neuron per layer, hidden layer, alpha, and train ratio parameters using Taguchi method. Cocomo and Kemerer datasets are used to test our hypothesized scenarios. The result shows that number of neurons, layer and train ratio are the important parameters that affects the performance of the GMDH model.

1. Introduction

Early estimation of project effort is essentials for a delivered successful project within budget and satisfied requirements. It is also known as feature of software engineering monetary how to oversee restricted assets in a way undertaking could make its focused on the objective goal of the schedule, budget, and scope. The algorithmic process is utilized numerical model to do software estimation whereas non algorithmic process was focused on expert judgement within previous software development experiences [1]. Incorporating algorithmic approach with novel machine learning models becomes popular research nowadays. Most of the researcher has done their research using many models and combining with algorithmic process but the accuracy of model still been questionable until now [2].

Group method of data handling (GMDH) algorithm which was first developed by Ivakhnenko for modelling and identification of complex systems [3]. The GMDH model is known as a self-organizing heuristic modelling approach. Due to some limitations in Artificial Neural Network (ANN), GMDH begins to attract the attention of researchers. GMDH can generally be classified as a supervised feed-forward multi-layered network model whereby the original input variables are used to generate the initial layer of the network, and the output of each layer acts as an input for the subsequent layer.

Although GMDH model have been applied in many fields of modelling, it has been given a little attention in setting parameter techniques. Over the years, various modifications made on GMDH are being discussed. It is necessary for a GMDH practitioner to initialize the maximum number of neurons
(per layer) and the maximum number of layers beforehand, despite not having to predetermine the number of neurons per layer and the number of hidden layers for the GMDH model. Ivakhnenko (1986) claimed that if the selected neuron in each layer should be set as large as possible, so that the optimum model would never be lost. The same issues about number of layers. If the numbers is too high, the output will generate a complex polynomial equations [4]. The issue regarding the determination of parameters for GMDH is rarely discussed in literature and are usually predetermined by researchers only through possibly trials and error methods. For example, Ghazanfari and Taylor [5] used 100 as the maximum number of neurons and 5 as the maximum number of layers while Ghazanfari applied 30 and 50 for their maximum neuron numbers while they applied 4 and 5 as the maximum number of layers [6]. Therefore, this study is concentrating on application of improvise GMDH parameter setting model using Taguchi, for it to improve the accuracy of software effort estimation.

2. Methodology

It is more than a decade since the quality management theories of Genichi Taguchi were applied in various industry. His approach to parameter design to minimize product and process variance has created a great deal of interest among both quality practitioners and statisticians [7-9]. This paper presents the incorporating Taguchi Design techniques to early parameter setting in machine learning single GMDH. It is upon this benchmark that the methods proposed will be evaluated using two dataset of software effort estimation: Cocomo and Kemerer.

The Taguchi design approach uses fractional factorial research designs to minimize the number of experiments, called Orthogonal Array (OAs). Orthogonal Array (OA) is a statistical method of defining parameters that converts test areas into factors and levels. A suitable orthogonal array design has been selected that suits for estimation and optimization. Selecting an effective OA would depend on the number of control factors and their scales. Using OA architecture, multiple process variables can be calculated which influence the output characteristic simultaneously, thus reducing the number of test runs.

While there are many standard orthogonal arrays available, each of the arrays is meant for a specific number of independent design variables and levels. In our methodology, we want to conduct an experiment to understand the influence of 4 different independent variables which is number of neurons, layers, alphas, and train ratios; with each variable having 3 set values (level values), based on previous review which has the best among the best setting designs as shown in Table 1.

For these design setting, L9 orthogonal array is appropriate to use for understanding the effect of 4 independent factors each having 3 factor level values. Table 2 shows a Taguchi design of L9 orthogonal array applied in parameter setting of GMDH model. The L9 orthogonal array means there are 9 experiments to be conducted and each experiment is based on the combination of level values as shown in the table. For example, the third experiment is conducted by keeping the independent design No of neurons at level 30, number of layers at level 5, alpha at level 0.6, and train ratio at level 0.8.

From these four factors, we need to analyze which factors are very crucial and by using some of statistical analysis, no longer try and error assumption should be used to identify which factors are dominant compare to others.

Table 1. Taguchi Setting

| Variables | Low  | Medium | High |
|-----------|------|--------|------|
| Neuron    | 30   | 50     | 100  |
| Layer     | 2    | 4      | 5    |
| Alpha     | 0.1  | 0.3    | 0.6  |
| Train ratio | 0.8 | 0.7    | 0.8  |
Table 2. Layout of Taguchi design L_9 orthogonal array in GMDH model.

| Experiment No | Neurons | Layer | Alpha | Train ratio | Performance Parameter Value |
|---------------|---------|-------|-------|-------------|-----------------------------|
| 1             | 30      | 2     | 0.1   | 0.6         | p1                          |
| 2             | 30      | 4     | 0.3   | 0.7         | p2                          |
| 3             | 30      | 5     | 0.6   | 0.8         | p3                          |
| 4             | 50      | 2     | 0.3   | 0.8         | p4                          |
| 5             | 50      | 4     | 0.6   | 0.6         | p5                          |
| 6             | 50      | 5     | 0.1   | 0.7         | p6                          |
| 7             | 100     | 2     | 0.6   | 0.7         | p7                          |
| 8             | 100     | 4     | 0.1   | 0.8         | p8                          |
| 9             | 100     | 5     | 0.3   | 0.6         | p9                          |

3. Experimental Results
The polynomial Transfer Function (TF) GMDH is the best among other training function for Cocomo and Kemerer dataset and will be apply in this study. The effect of the best setting factors based on the average of the S/N ratio and mean for Cocomo and Kemerer dataset are shown in Figures 1 and 2 respectively. The S/N ratio is a tools of quality index of signal to noise (S/N) in Taguchi. The main effect plot for S/N ratio indicates that the layer has the highest effect on the S/N ratio whereas alpha factor has a smallest effect.

![Figure 1. means and SN ratio for Cocomo-TaguchiGMDH.](image1)

![Figure 2. means and SN ratio for Kemerer-TaguchiGMDH](image2)

The effect of each factor can be emphasized further using Analysis of Variance (ANOVA) technique. Variance analysis (ANOVA) is performed to determine the significant process factor that considered
the percentage contribution of variance of factors to be measured affecting product quality by determining the quantities such as the degree of freedom (DF), sum of squares (SS), and value significant (P) for each factor.

It can be seen from the figure, that the S/N ratios are highest at low value setting for number of neurons and alpha factors for both dataset whilst the different pattern occur for layer factor where minimum variation is at low setting for Cocomo and the minimum variation at high settings for Kemerer dataset. The sum of square deviation of a particular variable indicates whether the performance parameter is sensitive to the change in level setting. The higher the value of sum of square of an independent variable, the more it has influence on the performance parameter. Also, if the p-value is less than 0.05, we reject the null hypothesis that there's no difference between the means and conclude that a significant difference does exist. The ANOVA analysis for Cocomo and Kemerer dataset in Table 3 and Table 4 indicate that that the number of neurons, layer and train ratio are significant at 95% confidence level that could affect the performance for both Cocomo and Kemerer whereas the number of alpha might not significant to the performance for GMDH model. The layer is most significant factor followed by number of neuron and train ratio. It shows the different parameter setting have impact on the performance of GMDH model. Thus, the optimal combination of parameter settings for GMDH model is required in improving the accuracy in the analysis.

### Table 3. ANOVA Taguchi Cocomo – GMDH

| Source    | DF | SS  | MS   | F    | P     |
|-----------|----|-----|------|------|-------|
| Neuron    | 2  | 42.1| 21.0 | 1.53 | 0.0290|
| Layer     | 2  | 48.1| 24.0 | 1.88 | 0.0232|
| Alpha     | 2  | 6.1 | 3.1  | 0.16 | 0.0859|
| Train ratio | 2  | 28.3| 14.2 | 0.88 | 0.0381|

### Table 4. ANOVA Taguchi Kemerer – GMDH

| Source    | DF | SS  | MS   | F    | P     |
|-----------|----|-----|------|------|-------|
| Neuron    | 2  | 18.1| 9.09 | 1.26 | 0.0349|
| Layer     | 2  | 19.0| 9.54 | 1.35 | 0.0328|
| Alpha     | 2  | 9.11| 4.56 | 0.52 | 0.0618|
| Train ratio | 2  | 15.11| 7.55 | 0.98 | 0.0429|

### 4. Conclusion and Future Work

The study provides the overview of the implementation of Taguchi setting in Group Method Data Handling (GMDH) techniques in improving the effort estimation accuracy. The best combination setting of Taguchi experiment on GMDH model is the combination of neuron100, layer 5, alpha 0.3 and train ratio 0.6 for Cocomo whilst the best combination for Kemerer with combination of parameter neuron 30, layer 4,pha 0.3 and train ratio 0.7. Hence it shows that the parameter setting effect the performance of GMDH method. Therefore, it is recommended that the future research works put a focus on the enhancement of GMDH in optimising the parameter that influence the performance of the model. Future research should also evaluate the performance on different types of transfer function in GMDH-based ensemble method.

### References

[1] Iqbal S Z, Idrees M , Sana A and Khan N 2017 *Appl. Math.* 2 (3) pp 33–39
[2] Molokken K and Jorgensen M 2003 *International Symposium on Empirical Software Engineering* pp 223–230
[3] Ivakhnenko A G and Ivakhnenko G A 1995 *Pattern Recognit.* 5 pp 527–535
[4] Gupta U and Kumar M 2013 *IOSR J. Compt. Eng.* 14 pp 58–62
[5] Taušer J and Buryan P 2011 *Prague Econ. Pap.* **20** pp 232-249
[6] Ghazanfari B and Taylor M E 2017 *ArXiv* pp 1-10
[7] Dean E B and Unal R 1991 *Annual Conference of the International Society of Parametric Analysts* pp 1–10
[8] Shahavi M H, Hosseini M, Jahanshahi M, Meyer R L and Darzi G N 2016 *Desalination Water Treat.* **57(39)** pp 18379–18390
[9] Lee K 2015 *Adv. Mater. Sci. Eng.* **2015** pp 304691

**Acknowledgement**
This project is by the Universiti Teknologi Malaysia under UTM Encouragement Research (UTMER) Grant for Vot No. 18J50.