SELF-ORGANIZING DATA MINING FOR WEATHER FORECASTING

Godfrey C. Onwubolu[1], Petr Buryan[2], Sitaram Garimella[3], Visagaperuman Ramachandran[4], Viti Buadromo[5]and Ajith Abraham[6]
[1,3,4,5]*School of Engineering and Physics, University of the South Pacific, Private Bag, Suva, Fiji onwubolu_g@usp.ac.fj, garimella_s@usp.ac.fj, ramachandran@usp.ac.fj*
[2]*Gerstner Laboratory, Department of Cybernetics, Czech Technical University, Technicka 2, 166 27 Prague, Czech Republic buryan@labe.felk.cvut.cz*
[6]*Center of Excellence for Quantifiable Quality of Service (Q2S), Norwegian University of Science and Technology, Trondheim, Norway ajith.abraham@ieee.org*

ABSTRACT

The rate at which organizations are acquiring data is exploding and managing such data so as to infer useful knowledge that can be put to use is increasingly becoming important. Data Mining (DM) is one such technology that is employed in inferring useful knowledge that can be put to use from a vast amount of data. This paper presents the data mining activity that was employed in weather data prediction or forecasting. The self-organizing data mining approach employed is the enhanced Group Method of Data Handling (e-GMDH). The weather data used for the DM research include daily temperature, daily pressure and monthly rainfall. Experimental results indicate that the proposed approach is useful for data mining technique for forecasting weather data.

1. INTRODUCTION

Knowledge discovery in databases (KDD) process is well documented in the literature [1—7]. A wide variety of data-mining model representation methods exist, but here, we only focus on a subset of popular techniques, which include *decision trees* and *rules* [8], [9], *linear models*, *nonlinear models* e.g., neural networks (see [7], [10], [11] for more detailed discussions), *example-based methods* (e.g., nearest-neighbor and case-based reasoning methods) [12], *probabilistic graphical dependency models* e.g., Bayesian networks [13]—[15], and *relational attribute models* [16]. Model representation determines both the flexibility of the model in representing the data and the interpretability of the model in human terms. Typically, the more complex models may fit the data better but may also be more difficult to understand and to fit reliably. While researchers tend to advocate complex models, practitioners involved in successful applications often use simpler models due to their robustness and interpretability [2—5].

In this paper, we present the DM process applied to weather data acquired at the School of Engineering and Physics, University of the South Pacific, Fiji to demonstrate the usefulness of this emerging technology in practical real-life applications. The weather data include daily temperature and pressure observed using automated instruments and a chaotic rainfall data set observed for the city of Suva.

2. SELF-ORGANIZING DATA MINING

Experience gained from expert systems, statistics, Neural Networks or other modeling methods has shown that there is a need to try to limit the involvement of modelers (users) in the overall knowledge extraction process to the inclusion of existing a priori knowledge, exclusively, while making the process more automated and more objective. Additionally, most users' interest is in results in their field and they may not have time for learning advanced mathematical, cybernetic and statistical techniques and/or for using dialog

driven modeling tools. Self-organizing modeling is based on these demands and is a powerful way to generate models from ill-defined problems.

A powerful method for model self-organization is the Group Method of Data Handling (GMDH) invented by Ivakhnenko [17],[18]. GMDH combines the best of both statistics and Neural Networks features while considering a very important additional principle of *induction*. This cybernetic principle enables GMDH to perform not only in advanced model parameter estimation but, more important, to perform an automatic model structure synthesis and model validation, too. GMDH creates adaptively models from data in form of networks of optimized transfer functions (active neurons) in a repetitive generation of populations (layers or generations) of alternative models of growing complexity and corresponding model validation and fitness selection until an optimal complex model which is not too simple and not too complex (over-fitted) has been created. Neither, the number of neurons and the number of layers in the network, nor the actual behavior of each created neuron (transfer function of active neuron) are predefined. All these are adjusted during the process of self-organization by the process itself. As a result, an explicit analytical model representing relevant relationships between input and output variables is available immediately after modeling. This model contains the extracted knowledge applicable for interpretation, prediction, classification or diagnosis problems. For detailed discussion of GMDH for self-organizing data mining applications, see [19]. Other self-organizing network variants derived from GMDH include polynomial neural networks [20]. In a wider sense, the spectrum of self-organizing modeling contains regression-based methods, rule-based methods, symbolic modeling and nonparametric model selection methods.

a. regression-based methods

Commonly, statistically-based principles are used to select parametric models. Besides sophisticated methods of mathematical statistics there has been much publicity about the ability of Artificial Neural Networks to learn and to generalize. However, Sarle [21] has shown that models commonly obtained by Neural Networks are overfitted multivariate multiple nonlinear (specifically linear) regression functions.

b. rule -based models in the form of binary or fuzzy logic

Rule induction from data uses genetic algorithms where the representation of models is in the familiar disjunctive normal form. A self-organizing fuzzy modeling may come to be more important for ill-defined problems using GMDH algorithm.

c. symbolic modeling

Self-organizing structured modeling uses a symbolic generation of an appropriate model structure (algebraic formula or complex process models) and optimization or identification of a related set of parameters by means of genetic algorithms. This approach assumes that the elementary components are predefined (model base) and suitably genetically coded.

d. nonparametric models

Known nonparametric model selection methods include: Analog Complexing (AC) which selects nonparametric prediction models from a given data set representing one or more patterns of a trajectory of past behavior which are analogous to a chosen reference pattern and Objective Cluster Analysis (OCA).

Table 1 shows some data mining functions and more appropriate self-organizing modeling algorithms for addressing these functions.

Table 1. Algorithms for self-organizing modeling

Data Mining functions	**Algorithm**
classification	GMDH, FRI, AC
clustering	AC
modeling	GMDH, FRI
time series forecasting	AC, GMDH, FRI
sequential patterns	AC

3. THE GROUP METHOD OF DATA HANDLING (GMDH)

The basics steps involved in the original Group Method of Data Handling (GMDH) modeling approach [17] are as follows:

Preamble: collect regression-type data of *n*-observations and divide the data into training and testing sets: $x_{ij};\ y_i \qquad i=1,2,...,n;\ j=1,2,...,m$

Step 1: Construct ${}^{m}C_2$ new variables $Z_1, Z_2, Z_3, ..., Z_{\binom{m}{2}}$, in the *training dataset* for all independent variables (columns of X), two at a time $\left(x_{i,k-1}, x_{i,k};\ i \in [1,m], k \in \left[2, \binom{m}{2}\right]\right)$ and construct the regression polynomial:

$$Z_1 = A + Bx_1 + Cx_2 + Dx_1^2 + Ex_2^2 + Fx_1x_2 \quad \text{at points } (x_{11}, x_{12}) \tag{1}$$

$$Z_k = A + Bx_{k-1} + Cx_k + Dx_{k-1}^2 + Ex_k^2 + Fx_{k-1}x_k \quad \text{at points } (x_{i,k-1}, x_{i,k}) \tag{2}$$

Step 2: For each of these regression surfaces, evaluate the polynomial at all *n* data points (i.e. using *A*, *B*, *C*, *D*, *E*, and *F* obtained from $x_{i,k-1}, x_{i,k};\ y_i$ for training)

Step 3: Eliminate the least effective variables: replace the columns of X (old variables) by those columns of Z (new variables) that best estimate the dependent variable *y* in the testing dataset such that

$$d_k^2 = \sum_{i=n_t+1}^{n} (y_i - z_{i,k})^2, \qquad k \in \left[1,2,..., \binom{m}{2}\right] \tag{3}$$

Order Z according to the least square error $d_k \mid \|d_j\| < R$ where R is some prescribed number chosen *a priori*. Replace columns of X with the best Z's $(Z_{<R})$; in other words, $X_{<R} \leftarrow Z_{<R}$

Step 4: Test for convergence. Let $DMIN = d_k$. If $DMIN_k = DMIN_{k-1}$ go to Step 1, else stop the process.

Since the introduction of GMDH, there have been variants devised from different perspectives to realize more competitive networks. The one employed in this paper is an enhanced version, e-GMDH [22].

4. DATA MINING EXPERIMENTATION

Rainfall prediction is very important to countries thriving on agro-based economy. In general, climate and rainfall are highly non-linear phenomena in nature giving rise to what is known as "butterfly effect". The parameters that are required to predict the rainfall are enormously complex and subtle so that uncertainty in a prediction using all these parameters is enormous even for a short period. Soft computing is an innovative approach to construct computationally intelligent systems that are supposed to possess humanlike expertise within a specific domain, adapt themselves and learn to do better in changing environments, and explain how they make decisions. Unlike conventional artificial intelligence techniques the guiding principle of soft computing is to exploit tolerance for imprecision, uncertainty, robustness, partial truth to achieve tractability, and better rapport with reality [11]. In this paper, we analyzed 13 years of rainfall data in Suva, the capital of Fiji. We attempted to train 3 prediction models using soft computing techniques with half the period of rainfall data. For performance evaluation, network predicted outputs were compared with the actual rainfall data.

4.1 Data Gathering

The weather data used for the data mining application described in this paper was acquired at the School of Engineering & Physics, University of the South Pacific, Fiji. The weather data include daily temperature and pressure observed from 2000—2007 using automated instruments and a chaotic rainfall data set observed for the city of Suva. The weather instruments used for gathering data used in this paper include HMP45D Humidity and Temperature Probe®, and T133P-XXHS Tipping Bucket Rain-gauge® with 0.5 mm plastic bucket calibration and 5 m cable. Campbell Scientific CR23X® data logger was used to capture the weather data from the local weather station to a dedicated PC located in the Physics laboratory. The transmitted weather data was then copied to Excel spreadsheets and archived on daily basis as well as monthly basis to

ease data identification. The day-to-day management of the instruments is undertaken by a Senior Technician.

4.2 Data Cleansing

In order to utilize the acquired weather data, some of the authors of this paper engaged some students as Research Assistants to organize the data which where copied to CDs into a daily logical datasets and also to convert the Julian dates into recognizable yearly dates. One of the authors was responsible for cross-checking all the acquired data and eliminating all possible errors such as those recorded when the weather recording instruments would have possibly failed (signified by some suspicious number) for a time interval and where blanks were found on the Excel data sheet. There was the need to eliminate such errors and/or bogus data to ensure data-integrity.

4.3 Feature Extraction

The data logger used for the data acquisition system acquires daily rainfall, temperature, humidity, wind speed, wind direction, and radiation. It was therefore necessary to extract only the interesting attributes of the data for our experimentation purpose.

4.4 Pattern Extraction and Discovery

The self-organizing GMDH technique was used for the purpose of extraction and discovery of knowledge of the data acquired; this is the core of data mining. There were 1922 rows of data for both temperature and pressure. A time lag = 5 was used for experimentation in all cases. The forecasting evaluation criteria used for all the experiments is the normalized mean squared error (Variation Accuracy Criterion or Ivachnenko's δ^2):

$$\delta^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} = \frac{1}{\sigma^2}\frac{1}{N}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad (4)$$

where y_i and $\hat{y}_i$ are the actual and predicted values, σ is the estimated variance of the data and $\bar{y}_i$ the mean. The ability to forecast movement direction or turning points can be measured by a statistic developed by Yao and Tan [23]. Directional change statistics (D_{stat}) can be expressed as

$$D_{stat} = \frac{1}{N}\sum_{t=1}^{n} a_t \times 100\%, \qquad (5)$$

where $a_t = 1$ if $a_t = (y_{t+1} - y_t)(\hat{y}_{t+1} - \hat{y}) \geq 0$, and $a_t = 0$ otherwise.

Case 1: University of South Pacific monthly temperature

For the 1922 rows of daily temperature over the period of 2000—2007 the average, minimum, maximum, and standard deviation are respectively 25.98^oC, 20.51^oC, 30.19^oC, and 1.71^oC. Using the time-lag approach, five columns of input data were generated with one column as output; the number of rows therefore reduced to 1918. The external criterion that was used for the GMDH approach for this particular experimentation is the VAC, Variation Accuracy Criterion (Ivachnenko delta-squared). The coefficient of determination (r-squared value), $r^2 = 0.8910$ and the directional statistics value, $ds = 44.2067$. The graphical-representation of e-GMDH network connections after pruning is shown in Figure 1. Figure 2 shows how the performance index on training data decreases for different layers as well as how the performance index on testing data decreases for different layers until layer 5. Figure 3 shows the GMDH prediction and absolute

difference error for the daily temperature data mining problem. The absolute difference error, is found be within the range of ± 1.5. Here, there is an excellent match between the measured and predicted values, showing that the proposed e-GMDH model can be used as a feasible solution for exchange rate forecasting.

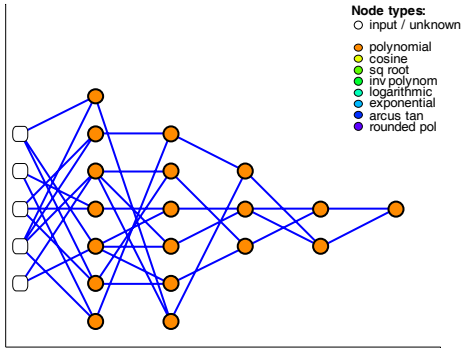

Figure 1. Graphical-representation of e-GMDH network connections after pruning

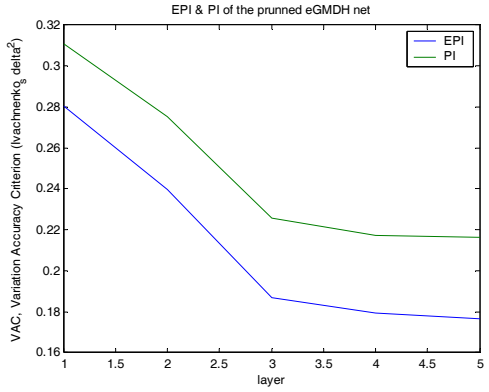

Figure 2. Performance indices on training (PI) and testing (EPI) datasets for different layers

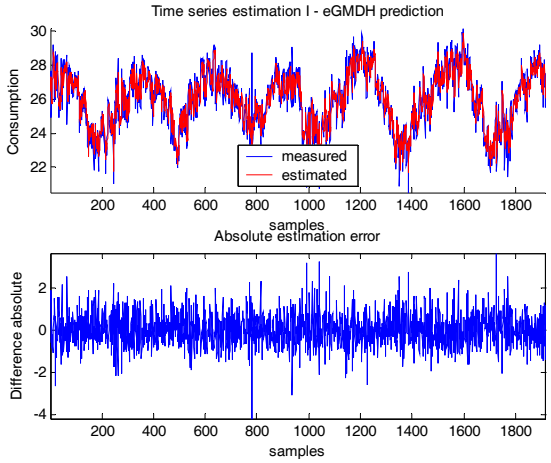

Figure 3. The GMDH actual and predicted values and absolute difference error for the temperature problem

Case 2: University of South Pacific monthly pressure

For the 1922 rows of daily pressure over the period of 2000—2007 the average, minimum, maximum, and standard deviation are respectively 2.95 bar, 1.601 bar, 4.052 bar, and 0.49 bar. Using the time-lag approach, five columns of input data were generated with one column as output; the number of rows therefore reduced

to 1918. The external criterion that was used for the GMDH approach for this particular experimentation is the VAC, Variation Accuracy Criterion (Ivachnenko delta-squared). The coefficient of determination (r-squared value), r^2 = 0.9437 and the directional statistics value, ds = 54.9061. Figure 4 shows the GMDH prediction and absolute difference error for the daily temperature data mining problem. The absolute difference error, is found be within the range of ±0.3 .

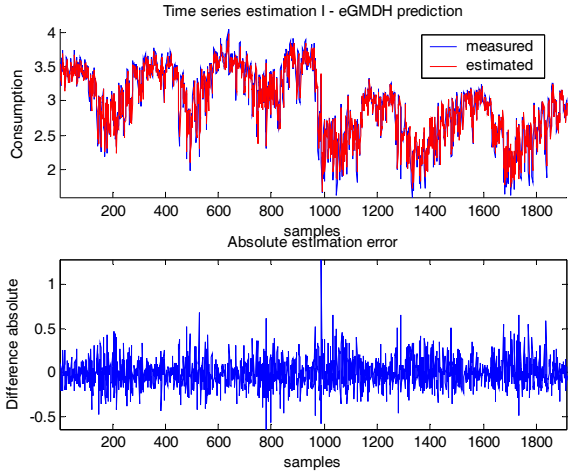

Figure 4. The GMDH actual and predicted values and percentage difference error for the pressure problem

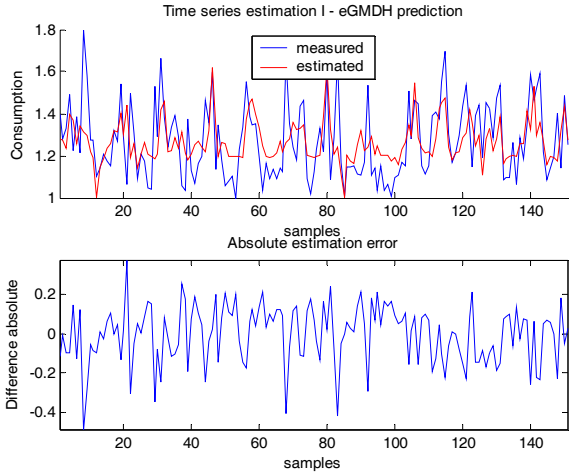

Figure 5. The GMDH actual and predicted values and percentage difference error for the rainfall problem

Case 3: Suva normalized monthly rainfall

Additional study included that of a chaotic dataset representing the monthly rainfall of Suva for the period of 1990—2002 [24], giving 156 rows of data. The average, minimum, maximum, and standard deviation are respectively 236.9 mm, 27.3 mm, 645.6 mm, and 135.5 mm. Using the time-lag approach, five columns of input data were generated with one column as output; the number of rows therefore reduced to 152. The external criterion that was used for the GMDH approach for this particular experimentation is the VAC, Variation Accuracy Criterion (Ivachnenko delta-squared). The coefficient of determination (r-squared value), r^2 = 0.3864 and the directional statistics value, ds = 51.9737. Due to the chaotic nature of this problem, the initial dataset was normalized according the following rule $y = 1 + 0.8\left(\frac{x - x_{\min}}{x_{\max} - x_{\min}}\right)$ where x is the actual value, $x_{\max}$ is the maximum value of x, $x_{\min}$ is the minimum value of x, and y is the normalized value

corresponding to *x*. Figure 5 shows the GMDH prediction and absolute difference error for the daily temperature data mining problem. The absolute difference error, is found be within the range of ± 0.2 . Here, there is not a good match between the measured and predicted values.
As already discussed, in general, climate and rainfall are highly non-linear phenomena in nature giving rise to what is known as "butterfly effect". It is not therefore surprising that the data collected for our mean rainfall experimentation has non-linear feature similar to previous research results reported [25-27].

4.5 Visualization of the Data

It is generally agreed that visualization is a key aspect of a good data mining platform. Therefore, inclusion of Figures 1—5 enhances our e-GMDH-based data mining approach. Figure 1 is particular useful because it depicts how the e-GMDH network cascades as well helps the user to track the related nodes.

4.6 Evaluation of Results

The e-GMDH generalizes quite well for Cases 1 and 2 but not quite for Case 3. The data for Case 3 seems chaotic because the statistical data of minimum, maximum, and standard deviation which are respectively 27.3 mm, 645.6 mm, and 135.5 mm, are quite unusual. The standard deviation is too high, and the difference between the maximum and minimum readings seems much. As part of validating the weather data acquired for this research, the Fiji Meteorological Service was visited to ensure that our instruments have been set up to World Meteorological Organization requirements/standards.
In order to assess the efficacy of our e-GMDH, it was necessary to compare its performance for all three Cases studied. Its performance is compared with other variants of GMDH, such as polynomial neural network (PNN) and the enhanced version, e-PNN as shown in Table 2. The e-GMDH does best for the temperature problem but does not fair well in the pressure and rainfall problem when compared to PNN and its variant, e-PNN. The reason in these particular cases is unclear as e-GMDH generally outperforms PNN and its variant in most modeling and prediction problems.

Table 2. Forecast performance evaluation for the three Cases

	Daily Temperature		Daily Pressure		Monthly Rainfall	
	PI	EPI	PI	EPI	PI	EPI
PNN	0.2770	2.0290	0.0318	0.0258	0.0270	0.0275
e-PNN	0.5895	0.6352	0.0286	0.0241	0.0810	0.0275
e-GMDH	0.2162	0.1767	0.1128	0.0982	0.6800	0.8360

5. CONCLUSIONS

This paper presents the data mining activity that was employed to mining weather data. The self-organizing data mining approach employed is the enhanced Group Method of Data Handling (e-GMDH). The weather data used for the DM research include daily temperature, daily pressure and monthly rainfall. Experimental results indicate that the proposed approach is useful for data mining technique for forecasting weather data. The results of e-GMDH were compared to those of PNN and its variant, e-PNN. The reason in these particular cases is unclear as e-GMDH generally outperforms PNN and its variant in most modeling and prediction problems. This paper has shown that end-users of data mining should endeavor to follow the methodology for DM since suspicious data points or outliers in a vast amount of data could give unrealistic results which may affect knowledge inference. Inclusion of the graphical network is a good visual instrument that would assist end-users to track how the e-GMDH advances in the search space. Empirical results also show that there are various advantages and disadvantages for the different techniques considered. There is little reason to expect that one can find a uniformly best learning algorithm for optimization of the performance for different weather datasets. This is in accordance with the *no free lunch theorem*, which explains that for any algorithm, any elevated performance over one class of problems is exactly paid for in performance over another class [28]. The dataset for the average monthly rainfall used in this research is available for researchers to experiment with, using various self-organizing data mining techniques. Further work will include a graphical user interface (GUI) which is partly in place but needs to be updated to include the current functionalities. Plans are in place for a partnership arrangement between the Fiji Meteorological Service and University of the South Pacific with regards to the free exchange of weather data for academic purposes and weather and climate monitoring purposes.

REFERENCES

[1] Codd, E. F. 1993. Providing OLAP (On-Line Analytical Processing) to User-Analysts: An IT Mandate. E. F. Codd and Associates.

[2] Fayyad, U. M.; Djorgovski, S. G.; and Weir, N. 1996. From Digitized Images to On-Line Catalogs: Data Mining a Sky Survey. *AI Magazine* 17(2): 51–66.

[3] Fayyad, U. M.; Haussler, D.; and Stolorz, Z. 1996. KDD for Science Data Analysis: Issues and Examples. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), 50–56. Menlo Park, Calif.: American

[4] Association for Artificial Intelligence.

[5] Fayyad, U. M.; Piatetsky-Shapiro, G.; and Smyth, P. 1996. From Data Mining to Knowledge Discovery: An Overview. In *Advances in Knowledge Discovery and Data Mining,* eds. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, 1–30. Menlo Park, Calif.: AAAI Press.

[6] Fayyad, U. M.; Piatetsky-Shapiro, G.; Smyth, P.; and Uthurusamy, R. 1996. *Advances in Knowledge Discovery and Data Mining.* Menlo Park, Calif.: AAAI Press.

[7] Brachman, R., and Anand, T. 1996. The Process of Knowledge Discovery in Databases: Human-Centered Approach. In *Advances in Knowledge Discovery and Data Mining,* 37–58, eds. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. Menlo Park, Calif.: AAAI Press.

[8] Elder, J., and Pregibon, D. 1996. A Statistical Perspective on KDD. In *Advances in Knowledge Discovery and Data Mining,* eds. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, 83–116. Menlo Park, Calif.: AAAI Press.

[9] Quinlan, J. 1992. *C4.5: Programs for Machine Learning.* San Francisco, Calif.: Morgan Kaufmann.

[10] Breiman, L.; Friedman, J. H.; Olshen, R. A.; and Stone, C. J. 1984. *Classification and Regression Trees.*Belmont, Calif.: Wadsworth.

[11] Friedman, J. H. 1989. Multivariate Adaptive Regression Splines. *Annals of Statistics*

[12] Dasarathy, B. V. 1991. Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques. Washington, D.C.: IEEE Computer Society.

[13] Heckerman, D. 1996. Bayesian Networks for Knowledge Discovery. In *Advances in Knowledge Discovery and Data Mining,* eds. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, 273–306. Menlo Park, Calif.: AAAI Press.

[14] Whittaker, J. 1990. *Graphical Models in Applied Multivariate Statistics.* New York: Wiley.

[15] Pearl, J. 1988. *Probabilistic Reasoning in Intelligent Systems.* San Francisco, Calif.: Morgan Kaufmann.

[16] Dzeroski, S. 1996. Inductive Logic Programming for Knowledge Discovery in Databases. In *Advances in Knowledge Discovery and Data Mining,* eds. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, 59–82. Menlo Park, Calif.: AAAI Press. 19:1–141.

[17] A. G. Ivakhnenko, Polynomial Theory of Complex Systems, IEEE Transactions on Systems, Man, and Cybernetics, 1971, pp. 364-378.

[18] H. R. Madala, A. G. Ivakhnenko, Inductive Learning Algorithms for Complex Systems Modeling, CRC Press, Boca Raton, 1994.

[19] Mueller and Lemke, 1999, *Self-organizing Data Mining: An Integrated Approach to Extract Knowledge From Data*, Dresden, Berlin.

[20] S. K. Oh and W. Pedrycz, The design of self-organizing polynomial neural networks, Inf. Sci. 141 (2002) 237-258.

[21] Sarle, W.S. (1995), "Stopped training and other remedies for overfitting," Proceedings of the 27th Symposium on the Interface of Computing Science and Statistics, 352-360, ftp://ftp.sas.com/pub/neural/inter95.ps.Z

[22] Buryan, P. and Onwubolu, G. C., 2007, Design of Enhanced MIA-GMDH Learning Networks, (review in process).

[23] J.T. Yao, C.L. Tan. A case study on using neural networks to perform technical forecasting of forex , Neurocomputing, 34:79-98, 2000.

[24] Kumar, V. V., 2003, Effects of Atmospheric Parameters on Ku-Band Satellite Link, MSc Thesis, School of Pure & Applied Sciences, The University of the South Pacific, Fiji.

[25] Abraham A, Philip N S and Joseph K B, Will we have a Wet Summer? Soft Computing Models for Long-term Rainfall Forecasting, In Proceedings of 15th European Simulation Conference ESM 2001, Prague, 2001.

[26] Philip N S and Joseph K B, On the Predictability of Rainfall in Kerala: An Application of ABF Neural Network, In Proceedings of Workshop on Intelligent Systems Design and Applications (ISDA 2001), In Conjunction with International Conference on Computational Sciences, ICCS 2001, San Francisco, May 2001

[27] Macready W.G. and Wolpert D.H.: The No Free Lunch theorems, IEEE Trans. On Evolutionary Computing, 1(1), (1997) 67-82.