Prediction of liquid- liquid equilibrium for a ternary system using GMDH

H.Ghannadzadeh [1*] Akbar Haghi and Katia Ghannadzadeh [2]

1- Department of Chemical Engineering, Guilan University, Rasht, Iran
2- Department of Chemical Engineering Bacelona University, Barcelona, Spain

***Abstract*:**
Liquid- liquid equilibrium (LLE) data are important for designing and modeling of process equipments. In this paper, the Group Method of Data Handling technique has been applied for estimation of LLE data of the ternary system of water + alcohol +solvent at 298.15 K. Using this technique, a new model has been proposed that suitable to use in place of conventional methods to predict LLE. The experimental results and predicted data by the Group Method of Data Handling, also mean deviations of the proposed, UNIQUAC and NRTL models have been compared.

***Keywords*:** Liquid–Liquid Equilibrium, Quaternary System, Group Method of Data Handling, Artificial Neural Network

1. Introduction

The importance of the availability of precise liquid–liquid equilibrium (LLE) data in rational design of many chemical processes and separation operations, have been the subject of much research in recent years. A large amount of investigation has been carried out on the LLE measurements, in order to understand and provide further information about the phase behavior of such systems. Usually, the equilibrium data presented are correlated using thermodynamic methods. The thermodynamic models have been successfully applied for the correlation of several LLE systems but these conventional methods for LLE data prediction of complex systems are tedious. Recently, to avoid these limitations, new prediction methods were developed by using artificial neural network (ANN). ANNs are non linear and highly flexible models that have been successfully used in many fields to model complex non–linear relationships. Hence they offer potential to overcome the limitations of the traditional thermodynamic models and polynomial correlation methods for the complicated systems, especially in estimating the LLE and vapour- liquid equibilirum (VLE) [2- 7]. ANNs may be viewed as the universal approximators but the main disadvantage of them is that the detected dependencies are hidden within the neural network structure [13]. Conversely, Group Method of Data Handling (GMDH) [8] is aimed to identify the functional structure of a model hidden in the empirical data. The main idea of GMDH is the use of feed- forward networks based on short- term polynomial transfer function whose coefficients are obtained using regression technique combined with the emulation of the self-organizing activity for the neural network (NN) structural learning [9]. GMDH was developed for complex systems modelling, prediction, identification and approximation of multivariate processes, diagnostics, pattern recognition and clusterization of data sample. It was proved, that for inaccurate, noisy or small data can be found best optimal simplified model, accuracy of which is higher and structure is simpler than structure of usual full physical model. In this work, to avoid the limitations of ANNs, an LLE prediction method was developed by using GMDH algorithm. The aim of this proposed method is to predict LLE data of a quaternary system [1], Corn Oil + Oleic acid + Ethanol + Water, using GMDH algorithm. Using existing data in [1], the proposed network was trained and the trained network used to predicting of LLE data in oil phase and alcohol phase. Then, the predicted data of the proposed model compared with the experimental data. Also mean deviations obtained by NRTL, UNIQUAC and proposed model have been compared. The phase diagrams for the studied quaternary system including both the experimental and predicted tie lines are presented.

2. Group Method of Data Handling (GMDH)

The Group Method of Data Handling is a combinatorial multi- layer algorithm in which a network of layers and nodes is generated using a number of inputs from the data stream being evaluated. The Group Method of Data Handling was first proposed by Alexy G. Ivakhnenko [8]. The GMDH network topology has been traditionally determined using a layer by layer pruning process based on a pre-selected criterion of what constitutes the best nodes at each level. The goal is to obtain a mathematical model of the object under study. The GMDH creates adaptively models from data in form of networks of optimized transfer functions in a repetitive generation of layers 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 has been created. Neither, the number of neurons and the number of layers in the network, nor the actual behavior of each created 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 [10].

2.1 GMDH Algorithm

The traditional GMDH method [8- 9] is based on an underlying assumption that the data can be modeled by using an approximation of the Volterra Series or Kolmorgorov- Gabor polynomial [11] as shown in equation (1).

$$y = a_0 + \sum_{i=1}^{m} a_i x_i + \sum_{i=1}^{m}\sum_{j=1}^{m} a_{ij} x_i x_j + \sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{m} a_{ijk} x_i x_j x_k \dots \quad (1)$$

Where x_i, x_j, x_k are the inputs, y the output and a_0, a_i, a_{ij}, a_{ijk} are the coefficients of the polynomial functional node.

A GMDH network can be represented as a set of neurons in which different pairs of them in each layer are connected through a quadratic polynomial and thus produce new neurons in the next layer [12]. In the classical GMDH algorithm, all combinations of the inputs are generated and sent into the first layer of the network. The outputs from this layer are then classified and selected for input into the next layer with all combinations of the selected outputs being sent into layer 2. This process is continued as long as each subsequent layer(n+1) produces a better result than layer(n). When layer(n+1) is found to not be as good as layer(n), the process is halted. The formal definition of the problem is to find a function $\hat{f}$ so that can be approximately used instead of actual one, f, in order to predict output $\hat{y}$ for a given input vector $X = (x_1, x_2, x_3, \ldots, x_n)$ as close as possible to its actual output y. Therefore, given M observation of multi-input–single-output data pairs (*training data set*) so that

$$y_i = f(x_{i1}, x_{i2}, x_{i3}, \ldots, x_{in}) \qquad i=1,\ 2,\ \ldots,\ M. \qquad (2)$$

It is possible to train a GMDH-type network to predict the output values $\hat{y}$ using *training* data, i.e.

$$\hat{y}_i = \hat{f}(x_{i1}, x_{i2}, x_{i3}, \ldots, x_{in}) \qquad i=1,\ 2,\ \ldots,\ M. \qquad (3)$$

This equation is tested for fit by determining the mean square error of the predicted $\hat{y}$ and actual y values as shown in equation (4) using the set of *testing* data. This value should be minimized.

$$\sum_{i=1}^{M} (\hat{y}_i - y_i)^2 \rightarrow \min. \qquad (4)$$

General connection between inputs and output variables can be expressed by equation (1). For most application the quadratic form of only two variables is used in the form

$$\hat{y} = G(x_i, x_j) = a_0 + a_1 x_{i_n} + a_2 x_{j_n} + a_3 x_{i_n} x_{j_n} + a_4 x_{i_n}^2 + a_5 x_{j_n}^2 \qquad (5)$$

to predict the output y. A typical feed-forward GMDH-type network is shown in figure 2. The coefficients a_i in equation (5) are calculated using regression techniques [8–9] so that the difference between actual output, y, and the calculated one, $\hat{y}$, for each pair of x_i, x_j as input variables is minimized. Indeed, it can be seen that a tree of polynomials is constructed using the quadratic form given in equation (5) whose coefficients are obtained in a least-squares sense. In this way, the coefficients of each quadratic function G_i are obtained to optimally fit the output in the whole set of input–output data pair, i.e.

$$r^2 = \frac{\sum_{i=1}^{M} (y_i - G_i(\ ))^2}{\sum_{i=1}^{M} y_i^2} \rightarrow \min. \qquad (6)$$

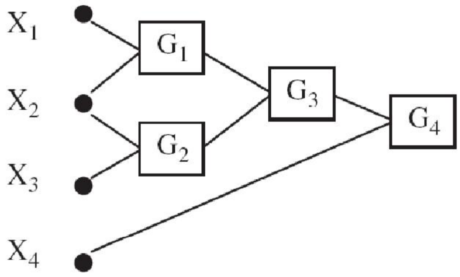

Fig. 1. Feed-forward GMDH-type network.

In the basic form of the GMDH algorithm, all the possibilities of two independent variables out of total n input variables are taken in order to construct the regression polynomial in the form of equation (5) that best fits the dependent observations (y_i, $i = 1, 2, \ldots, M$) in a least- squares sense. Consequently,

$$\binom{n}{2} = \frac{n(n-1)}{2}$$

neurons will be built up in the second layer of the feed- forward network from the observations $\{(y_i, x_{ip}, x_{iq}), (i = 1, 2, \ldots, M)\}$ for different $p, q \in \{1, 2, \ldots, M\}$ [9]. In other words, it is now possible to construct M data triples $\{(y_i, x_{ip}, x_{iq}), (i = 1, 2, \ldots, M)\}$ from observation using such $p, q \in \{1, 2, \ldots, M\}$ in the form

$$\begin{bmatrix} x_{1p} & x_{1q} & \vdots & y_1 \\ x_{2p} & x_{2q} & \vdots & y_2 \\ \cdots & \cdots & \cdots & \cdots \\ x_{Mp} & x_{Mq} & \vdots & y_M \end{bmatrix}$$

Using the quadratic sub- expression in the form of equation (5) for each row of M data triples, the following matrix equation can be readily obtained as

$$Aa = Y \qquad (7)$$

where $\boldsymbol{a}$ is the vector of unknown coefficients of the quadratic polynomial in equation (5):

$$a = \{a_0, a_1, a_2, a_3, a_4, a_5\} \qquad (8)$$

and

$$Y = \{y_1, y_2, y_3, \ldots, y_M\}^T \qquad (9)$$

is the vector of output's value from observation. It can be readily seen that

$$A = \begin{bmatrix} 1 & x_{1p} & x_{1q} & x_{1p}x_{1q} & x_{1p}^2 & x_{1q}^2 \\ 1 & x_{2p} & x_{2q} & x_{2p}x_{2q} & x_{2p}^2 & x_{2q}^2 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & x_{Mp} & x_{Mq} & x_{Mp}x_{Mq} & x_{Mp}^2 & x_{Mq}^2 \end{bmatrix} \qquad (10)$$

The least- squares technique from multiple- regression analysis leads to the solution of the normal equations in the form of

$$a = \left(A^T A\right)^{-1} A^T Y \qquad (11)$$

which determines the vector of the best coefficients of the quadratic equation (5) for the whole set of M data triples.

2.2. Prediction of LLE using the GMDH- type network

The proposed model is a feed- forward GMDH-type network and has constructed using experimental data set from ref. [1]. This data set is constituted of 25 points in four different concentrations of water in solvent. In Table 1, the overall experimental compositions of the mixtures and in Table 2, experimental mass fractions of the components in alcohol and oil phase are shown. The data set is divided in two parts, 80% used as *training* and 20% used as *testing* data. Each point in *training* and *test* data is constituted of 13 values. The four mass fractions in overall compositions and water concentration in solvent are normalized and used as inputs of GMDH-type network (X_1, ..., X_5) and other eight values are used as desired outputs of network, 4 mass fractions in alcohol phase (Y_1, ..., Y_4) and 4 mass fractions in oil phase (Z_1, ...,Z_4). After applying the data set to the network, GMDH-type network eight polynomial equations are obtained that can be used to predicting of mass fractions in alcohol and oil phase, Table 3. For example, the prediction equations of mass fraction of the acid in alcohol and oil phase are:

$$Y_2 = 1.157533X_3 - 7.533211X_1X_2X_3$$

$$Z_2 = 1.962677X_3^2 + 1.500100X_2X_3 + 15.43566X_2X_3X_5$$

where X_1 is the water concentration in solvent and X_2, X_3 and X_5 are the normalized mass fraction of oleic acid, ethanol and water in overall composition, respectively. The network topology of this part of GMDH model is shown in Figure 2.

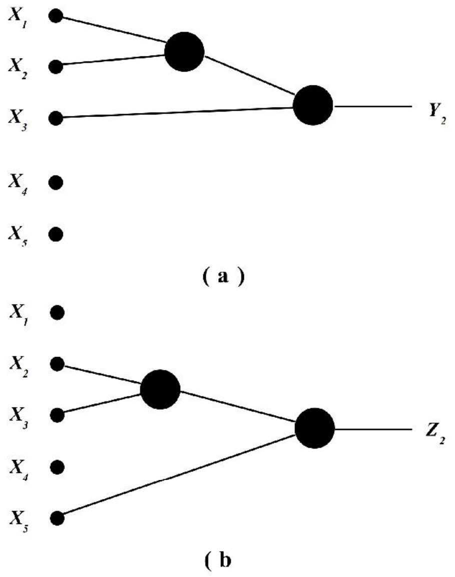

Figure 2. The GMDH-type network topology for mass fraction of acid in (a) alcohol phase (b) oil phase.

We used GMDH model to calculate the mass fractions of the components in alcohol and oil phase. The calculated values are see in Figure 3 and 4 show the experimental points and predicted tie lines from GMDH model for the systems corn oil/oleic acid/5% aqueous ethanol and corn oil/oleic acid/8% aqueous ethanol. The equilibrium diagrams were plotted in triangular coordinates. For representing the pseudo ternary systems in triangular coordinates, ethanol + water was admitted as a mixed solvent [1]. These figures indicate that GMDH model provided a good estimation in both phases.

Table 1. The Overall Composition of Liquid- Liquid Equilibrium Data for the System +solvent(1)+ ethanol (3) + Water (4)] at 298.15 K

water conc. in solvent	overall composition			
	$100\,w_1$	$100\,w_2$	$100\,w_3$	$100\,w_4$
5 wt %	47.98	0.00	49.40	2.63
	47.21	2.53	47.72	2.54
	43.46	4.91	49.02	2.61
	39.25	9.87	48.32	2.57
	35.65	14.52	47.32	2.51
	29.85	19.99	47.62	2.53
8 wt %	49.97	0.00	46.03	4.00
	44.97	5.39	45.67	3.97
	39.78	9.81	46.38	4.03
	35.49	14.59	45.93	3.99
	30.99	19.77	45.30	3.94
12 wt %	50.07	0.00	43.94	5.99
	47.94	2.40	43.70	5.96
	45.85	4.92	43.32	5.91
	41.49	9.65	43.26	5.90
	34.15	14.79	44.93	6.13
	30.04	19.99	43.97	5.99
	24.59	25.06	44.30	6.04
18 wt %	50.35	0.00	40.72	8.94
	48.27	2.42	40.44	8.88
	44.10	4.91	41.81	9.18
	39.94	9.80	41.22	9.05
	34.70	15.08	41.18	9.04
	29.66	20.15	41.16	9.03
	25.22	24.89	40.91	8.97

Figure 3 and 4 presents the fatty acid distribution between the phases. The distribution coefficient and solvent selectivity can be calculated by Eqs (12) and (13) respectively

$$k_i = \frac{w_i^{II}}{w_i^{I}} \qquad (12)$$

$$S = \frac{k_2}{k_1} \qquad (13)$$

The deviations between experimental and predicated compositions in both phases are calculated according to eq (14) and shown in Table 4. These values are compared with the calculated deviations from NRTL and UNIQUAC models [1]. As be shown, GMDH model provided a better estimation against the other models.

$$\Delta w = 100\sqrt{\frac{\sum_{n}^{N}\sum_{i}^{C}\left[\left(w_{i,n}^{\mathrm{I,ex}} - w_{i,n}^{\mathrm{I,calc}}\right)^2 + \left(w_{i,n}^{\mathrm{II,ex}} - w_{i,n}^{\mathrm{II,calc}}\right)^2\right]}{2NC}} \qquad (14)$$

where N is the total number of tie lines, C is the total number of components. w is the mass fraction, the subscripts i, n are component and tie line, respectively and the superscripts I and II stand for oil and alcoholic phases, respectively; ex and calc refer to experimental and calculated concentrations.

In this work, we designed a GMDH model for different water concentration in solvent from all point of data set. For a better comparison with NRTL and UNIQUAC models [1] that were presented for systems with different water concentration, one can design four GMDH models for systems with 5%, 8%, 12%

and 18% aqueous ethanol. It is obvious that each GMDH model uses four inputs (X_2, ..., X_5) that are the mass fractions of the components according to Table 1.

3. Conclusions

In this study, a GMDH model designed using the experimental liquid- liquid equilibrium data for system Corn Oil + Oleic Acid + Ethanol + Water at 298.15 K [1]. The LLE data are predicted by GMDH model and then compared with the experimental data. Despite the complexity of the studied system, GMDH model allows a good prediction of phase equilibrium. Also the global deviation of the proposed model were lower than 0.57% in relation to the experimental data and the calculated data from NRTL an UNIQUAC models. GMDH model may be suitable to use in place of conventional methods predicting of LLE. The quality of the model is related to the quality of data used for the training of the model. For a better comparison it needs to design an independent GMDH model for each water concentration in solvent that can be studied in future works.

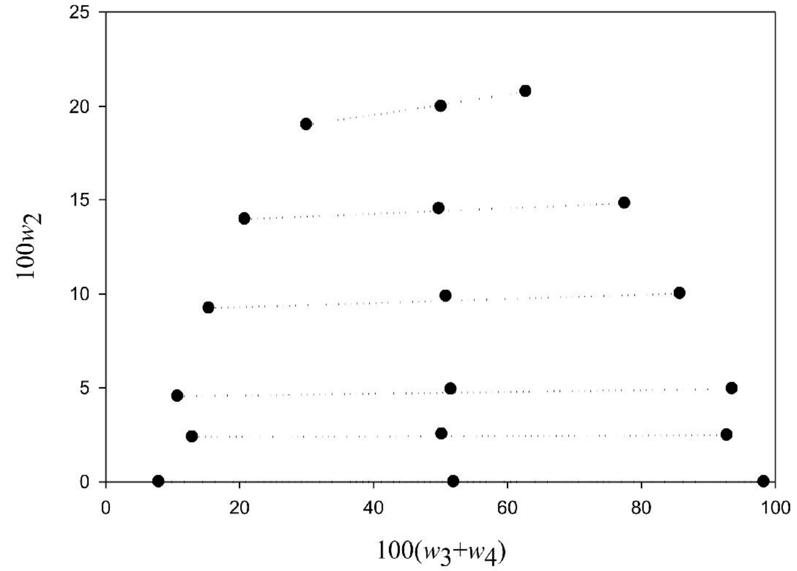

Figure 3. System of cyclohexane +ethanol + water at 298.15 K: (•) experimental; (...) GMDH.

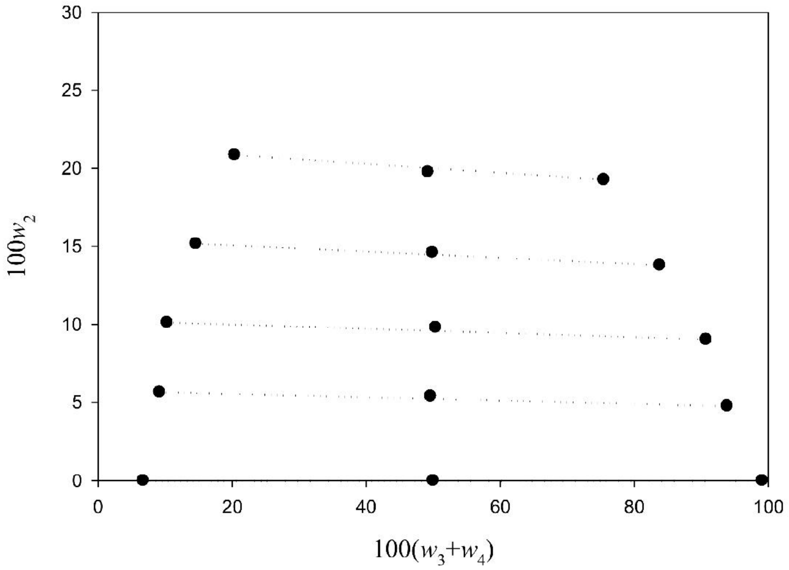

Figure 4. System of cron oil (1) + oleic acid (2) + 8% aqueous solvent [ethanol (3) + water (4)] at 298.15 K: (●) experimental; (...) GMDH.

Reference

[1] C.B. Goncalves, E. Batista, A.J.A. Meirelles, "Liquid- Liquid Equilibrium data for system cron oil+oleic acid+ethanol+water at 298.15 K," *J. Chem. Eng. Data*, 47, 416- 420, 2002.

[2] P.R.B. Guimaraes, C. McGreavy, "Flow of information through an artificial neural network," *Comput. Chem. Eng.*, 19 (S1), 741–746, 1995.

[3] R. Sharma, D. Singhal, R. Ghosh, A. Dwivedi, "Potential applications of artificial neural networks to thermodynamics: vapor–liquid equilibrium predictions ," *Comput. Chem. Eng.*, 23, 385–390, 1999.

[4] Sh. Urata, A. Takada, J. Murata, T. Hiaki, A. Sekiya, "Prediction of vapor–liquid equilibrium for binary systems containing HFEs by using artificial neural network," *Fluid Phase Equilib*, 199, 63–78, 2002.

[5] S. Ganguly, *Comput. Chem. Eng.*, 27, 1445–1454, 2003.

[6] S. Urata, A. Takada, J. Murata, T. Hiaki, A. Sekiya, "Prediction of vapor–liquid equilibrium for binary systems containing HFEs by using artificial neural network ," *Fluid Phase Equilib*, 199, 63–78, 2002.

[7] S. Mohanty, "Estimation of vapour liquid equilibria of binary systems, carbon dioxide–ethyl caproate, ethyl caprylate and ethyl caprate using artificial neural networks," *Fluid Phase Equilib*, 235, 92–98, 2005.

[8] A.G. Ivakhnenko, "Polynomial theory of complex systems," *IEEE Trans. Syst. Man Cybern,* **1,** 364–78, 1971.

[9] S.J. Farlow, *Self- Organizing Method in Modeling: GMDH Type Algorithm* , Dekker, New York, 1984.

[10] G.C. Onwubolu, “Data Mining using Inductive Modelling Approach,” International Workshop on Inductive Modelling- IWIM2007, 78- 86, 2007.
[11] H.R. Madala, A.G. Ivakhnenko, *Inductive Learning Algorithms for Complex Systems Modeling*, CRC Press Inc., Boca Raton, 1994.
[12] N. Nariman- Zadeh, A. Darvizeh, M.E. Felezi, H. Gharababaei, “Polynomial modelling of explosive compaction process of metallic powders using GMDH-type neural networks and singular value decomposition,” *Modelling Simul. Mater. Sci. Eng.*, 10, 727–744, 2002.
[