Journal of Applied Sciences

ISSN 1812-5654

science alert

ANSInet
an open access publisher
http://ansinet.com

Journal of Applied Sciences 8 (15): 2687-2694, 2008
ISSN 1812-5654

A Novel Intelligent Method of Experiment Design for Modeling

[1]S. Jafari, [2]H.R. Abdolmohammadi, [3]H. Eliasi, [3]M.B. Menhaj and [4]M.R. Rajati
[1]Department of Biomedical Engineering, Amirkabir University of Technology,
424 Hafez Ave., Tehran 15875-4413, Iran
[2]Department of Electrical Engineering, Iran University of Science and Technology,
Narmak, Tehran 16846-13114, Iran
[3]Department of Electrical Engineering, Amirkabir University of Technology,
424 Hafez Ave., Tehran 15875-4413, Iran
[4]Department of Electrical Engineering, K.N. Toosi University of Technology,
Tehran 16315-1355, Iran

Abstract: The aim of this study is to provide an experiment design method for modeling and function approximation. Modeling real-life systems is extremely of interest nowadays. Models could be useful in analysis of systems and help us understand their behavior. From a new point, models could be classified into three classes: black box models, gray box models and white box models. Our idea is related to black box modeling. Proper performance of a black box model depends on structure of the model as well as the data used to determine its parameters. Although one of the important factors affecting the richness of the dataset is the number of data, increasing the number of data points is limited in real problems. For instance gathering data from many systems imposes spending lots of time and cost. In this study, inspired by honey bee algorithm, we have designed a method which enriches the datasets for a known number of data, in comparison to other conventional data extraction methods. In such a method, after extracting some data by grid method, the other data points are extracted according to an intelligent analysis on available data. The results illustrate the efficiency of the proposed method.

Key words: Data extraction, modeling, evolutionary algorithms

INTRODUCTION

In many areas of science, system modeling is an important task. A model is a useful tool for system analysis and helps the designer obtain a better scope of the behavior of the system. Furthermore, a model enables us to simulate and predict the behavior of a system.

In engineering applications, models are especially needed for analysis and design of systems. For instance, advanced technologies for controller design, optimization, supervision and surveillance, fault detection and so on are based on models of the systems.

From a viewpoint, models could be divided into three groups: white box, black box and gray box.

Our idea is related to black box modeling. Black box models rely on experimental data and need no a-priori knowledge. The parameters and the structure of the model may have no relation with the actual structure of the system.

It is well-known that a model is always an approximation of the system's real relationships and its response does not always correspond to the real responses of the system. This is more sensible in black-box models.

Actually, we cannot test the correctness of the model's responses for all inputs. All efforts are performed to increase the probability of proper performance of the model.

It is well-known that the proper performance of a model is dependent on the model structure and the data used to tune the model parameters. The richness of the training data to represent the system appropriately enhances the accuracy of the model.

One of the factors which influence the richness of the data is the number of the data points. However, in real-world problems, there are limitations on the increment of the number of data points. Acquiring data from many systems and processes requires lots of time and cost which are serious limits on increasing the number of data

Corresponding Author: Sajad Jafari, Department of Biomedical Engineering, Amirkabir University of Technology,
424 Hafez Ave., Tehran 15875-4413, Iran Tel: +989357874169 Fax: +982166468186

points. For instance, one can point to modeling the mechanical characteristics of an alloy in terms of its ingredients (Datta and Banerjee, 2006), estimation of the amount of mineral reservoirs (Tercan and Karayigit, 2001), modeling the dying properties of materials in the textile industry (Senthilkumar, 2006; Daneshvar *et al*., 2006; Keong *et al*., 2004; Dai *et al*., 2004; Hebbar *et al*., 2006; Khamis *et al*., 2006; Elkamel, 1998; Turkoglu *et al*., 1999).

Sometimes, it is necessary to model the system with an available set of data. This is not of interest of in this study. The idea is to obtain the proper set of data, which represents the system, appropriately.

Suppose, that the limitations permit us just to obtain M data points. We can obtain the data points with the following methods:

- Selection of M data points in the domain of interest randomly
- Griding each dimension of the input
- Griding each dimension of the inputs and selecting the data randomly in each interval

The problem with all the above three methods is the batch nature of acquiring the data and there is no intelligence in obtaining them. On the other hand, it seems that when the data points are bring extracted, our knowledge of the system behavior increases and we could better judge the nature of the system.

The idea is using such kind of knowledge to enrich the procedure of data extraction to do this, inspired by the honey-bee algorithm (Seifipour and Menhaj, 2001), we have designed a method which analyzes the available data set and selects the best data point appropriately.

THE PROPOSED ALGORITHM

At the first step, we determine the number of the needed data points (M) according to the existing constraints (for example costs, time and so on). Then, we obtain a number of the data (N) according to the grid method to initialize the algorithm. Inspired by the honey bee algorithm, one of the data points should be selected as queen. The new data will be generated around the queen. To do this, we need a fitness function to choose the queen. It seems that the data in which the gradient is maximum is a suitable candidate. On the other hand, those part of the input space in which the number and density of data are less should be considered for data extraction. According to this, we selected the following fitness function:

$$FF(x_i) = \left|\sum_{j=1}^{H}(f(x_i) - f(x_j))\right| \times \prod_{j=1}^{H} d_j \quad (1)$$

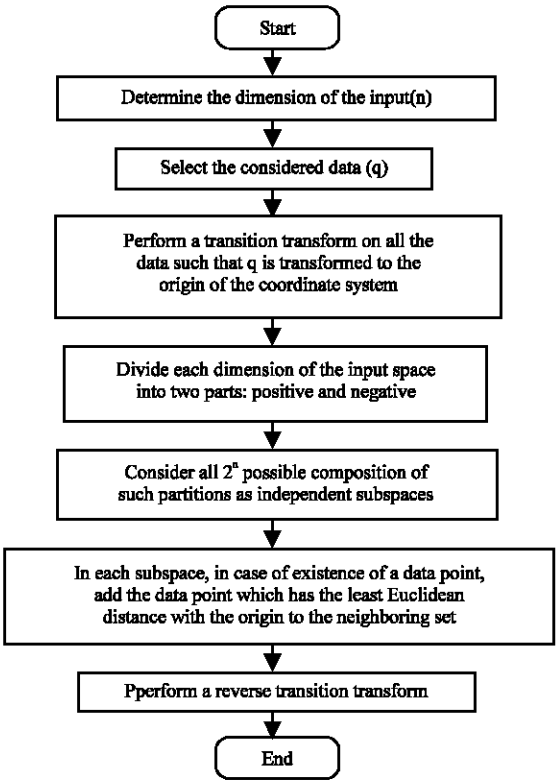

Fig. 1: Neighbor selection algorithm's flowchart

In which $(x_i, f(x_i))$ is the data point whose fitness is to be evaluated. H is the number of the data in x_i's neighborhood, f is the unknown function to be modeled and d_j is the Euclidean distance between x_i and its j^{th} neighbor x_j. The method of determination of a neighbor for each point in the input space is shown in Fig. 1.

It is notable that we could consider any other criterion and constraint in data selection in the fitness function. Although it is considered in the fitness function, the data could be avoided from being very dense in some regions of the input space by a tabu list. After the selection of the queen, according to the relative fitness function, the best individual is selected from its neighborhood as its mate. Considering the fact that the neighboring data $(x_i, f(x_i))$ which has the most distant $f(x_i)$ from the queen is a suitable candidate to breed the queen, we consider:

$$RFF(x_j) = \left|(f(x_j) - f(q))\right| \times d_j \quad (2)$$

In which x_i is the point of which the relative fitness function is to be calculated. Q is the queen and d_j is the Euclidian distance between x_j and q. the offspring is created by the crossover of the queen and its mate. The crossover could be performed by many methods. We used the arithmetical mean as the crossover operator.

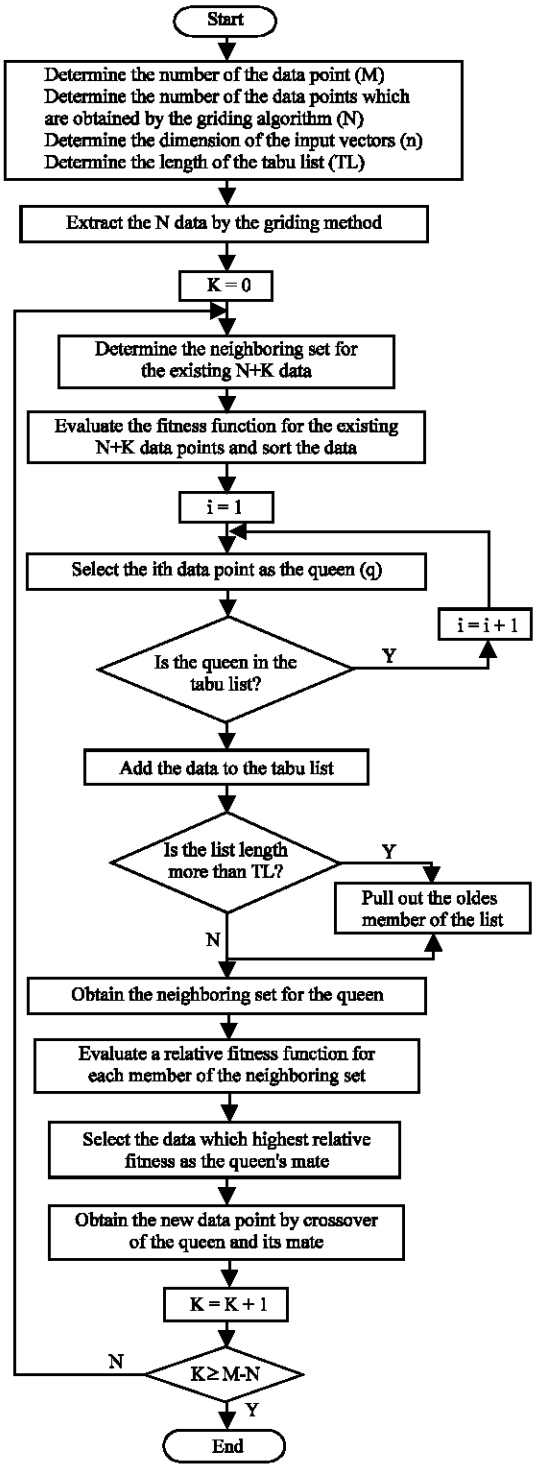

Fig. 2: Flowchart of the proposed algorithm

This algorithm continues until the remainder M-N data points are generated. Figure 2 shows the flowchart of the proposed algorithm.

SIMULATION RESULT

To verify the algorithm, we extracted some data points to approximate the functions of Table 1 and 2. We employed four methods: random, grid, grid-random and the proposed method. On the other hand, from each function, we obtained a huge bulk of test data points (100 M) to examine the ability of the methods in providing rich data.

To approximate the function of Table 1, we used the following methods:

- Training on MLP neural network with 5 neurons in the first layer and 3 neurons in the second layer and 1 neuron in the output layer with tangent-sigmoid stimulus functions in the first and second layer and a linear stimulus function in the output layer. This structure has been obtained by try-and-error and is not necessarily the optimal structure, but it is adequate for our purpose which is comparison of the methods

For each method for data point extraction, we trained the network 20 times and calculated the mean of network's mean square error for the test data.

- Piecewise Cubic Hermite Interpolating polynomial

The error criterion is MSE (Mean Square Error). It is seen that the proposed method is usually better than the other methods (Table 3).

Table 1: Single-variable functions used for data extraction

Function	Case
$f_1(x) = 10e^{-2\|x\|}$	1
$f_2(x) = 10\sin(x)$	2
$f_3(x) = \begin{cases} \frac{1}{0.1+x} & x \geq 0 \\ \frac{1}{-0.1+x} & x < 0 \end{cases}$	3
$f_4(x) = \begin{cases} \frac{1}{2}\sin(0.2\pi x) & x \geq 3.5 \text{ or } x \leq 2.5 \\ 4\sin(5\pi x) & 2.5 < x < 3.5 \end{cases}$	4
$f_5(x) = \begin{cases} \frac{1}{0.1+\|x\|} & x \geq 0 \\ \frac{-1}{0.1+\|x\|} & x < 0 \end{cases}$	5

Table 2: A two-variable function used for data extraction

Function	Case
$f_6(x,y) = (x+12)e^{-0.5\sqrt{x^2+y^2}}$	1
$f_7(x,y) = 10(e^{-\sqrt{(x+5)^2+(y+5)^2}} - e^{-\sqrt{(x+5)^2+(y-5)^2}} - e^{-\sqrt{(x-5)^2+(y+5)^2}})$	2

Table 3: The simulation results for the single-variable functions

				Modeling error for the test data (MSE)			
Function	M	N	Modeling method	Grid method	Random method	Grid-random method	Proposed method
f_1	20	2	Neural network	0.6294	0.7841	0.7248	0.4077
			PCHIP	0.6104	0.2705	0.3643	0.0015
f_2	15	10	Neural network	24.5895	34.1137	37.7052	30.3321
			PCHIP	1.1285	25.2516	2.0788	3.1754
f_3	20	8	Neural network	0.9025	0.8772	1.4727	0.7826
			PCHIP	0.6420	0.8148	2.7119	0.0684
f_4	30	10	Neural network	1.1046	6.1944	0.8635	0.5922
			PCHIP	0.7891	1.0935	0.6208	0.3335
F_5	21	2	Neural network	0.7354	0.9153	0.7632	0.3466
			PCHIP	0.7005	0.8276	0.5892	0.0022

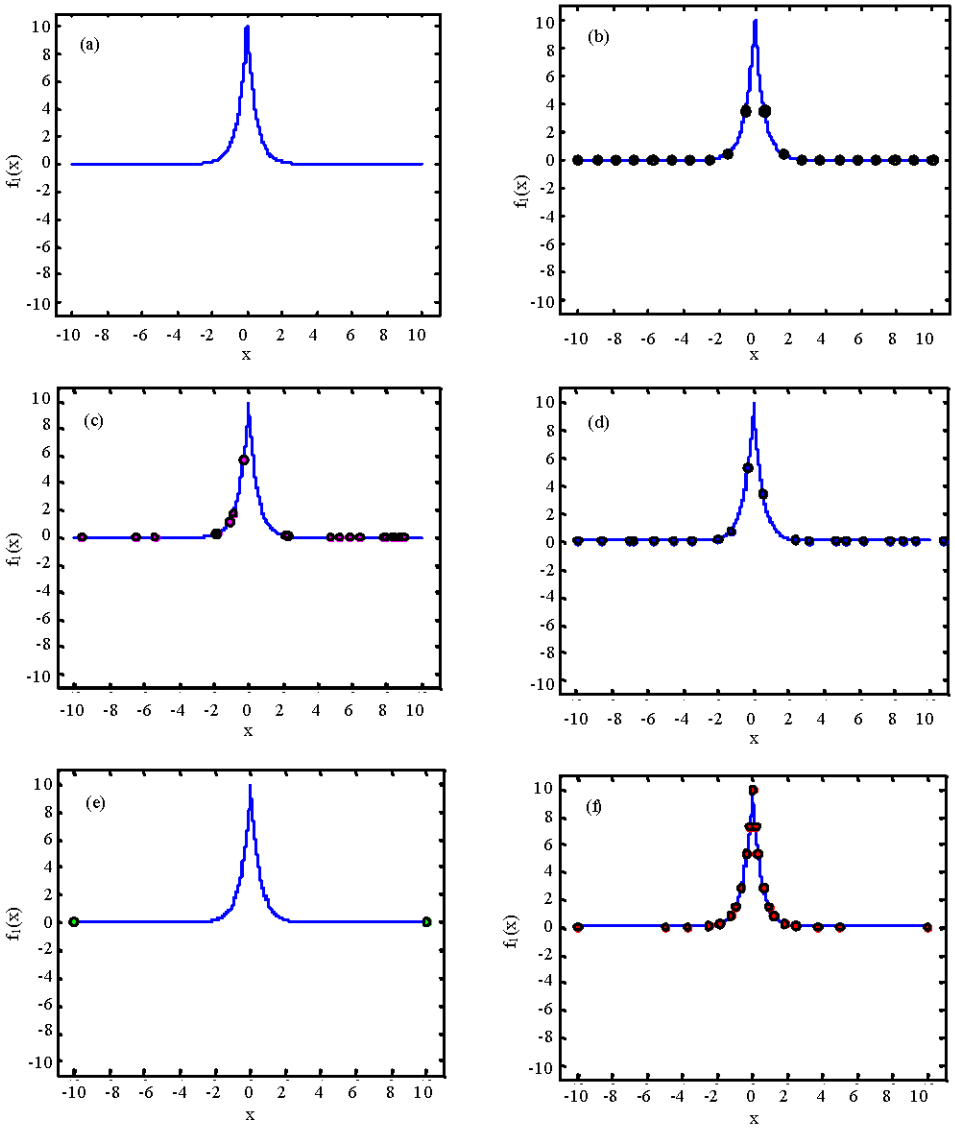

Fig. 3: (a) The diagram of the function f_1, (b) data extracted from f_1 by grid method, (c) data extracted from f_1 randomly, (d) data extracted from f_1 by the grid-random method, (e) extracted data from f_1 by the grid method applied in proposed method as initial data and (f) data extracted from f_1 by the proposed method

Table 4: The simulation result for the two-variable function

				Modeling error for the test data (MSE)			
Function	M	N	Modeling method	Grid method	Random method	Grid-random method	Proposed method
F_6	36	16	Neural network	0.8493	1.7977	0.9813	0.5353
			TCI	0.3457	0.5618	0.6065	0.0236
F_7	49	9	Neural network	1.0268	1.9854	1.2135	0.7812
			TCI	0.5925	0.7346	0.9001	0.0554

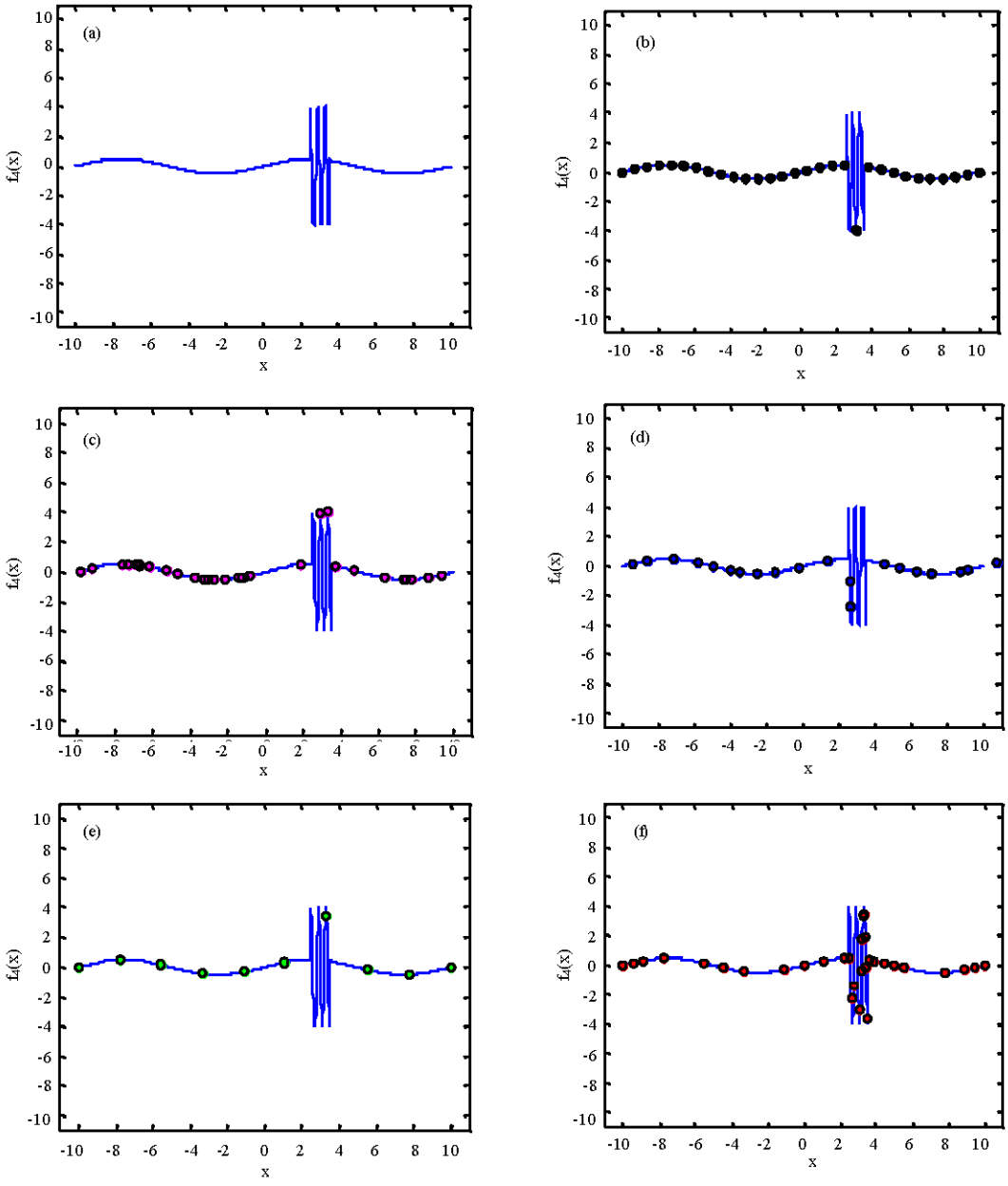

Fig. 4: (a) The diagram of the function f_4, (b) data extracted from f_4 by the grid method, (c) data extracted from f_4 randomly, (d) data extracted from f_4 by the grid-random method (e) extracted data from f_4 by the grid method applied in proposed method as initial data and (f) data extracted from f_4 by the proposed method

To approximate the function shown in Table 2 we used two methods:

- A three layer MLP neural network with 10 neurons in the first layer and 6 neurons in the second layer and tangent-sigmoid stimulus functions and 1 neuron in the output layer with a linear stimulus function
- Triangle-based Cubic Interpolation (TCI)

It is easily seem that the error due to the proposed method is less than the other methods (Table 4).

For example in Fig. 3-5 the data extracted by the proposed method are shown and compared to the data extraction by other methods.

Figure 6 shows more examples, which demonstrate efficiency of the proposed method.

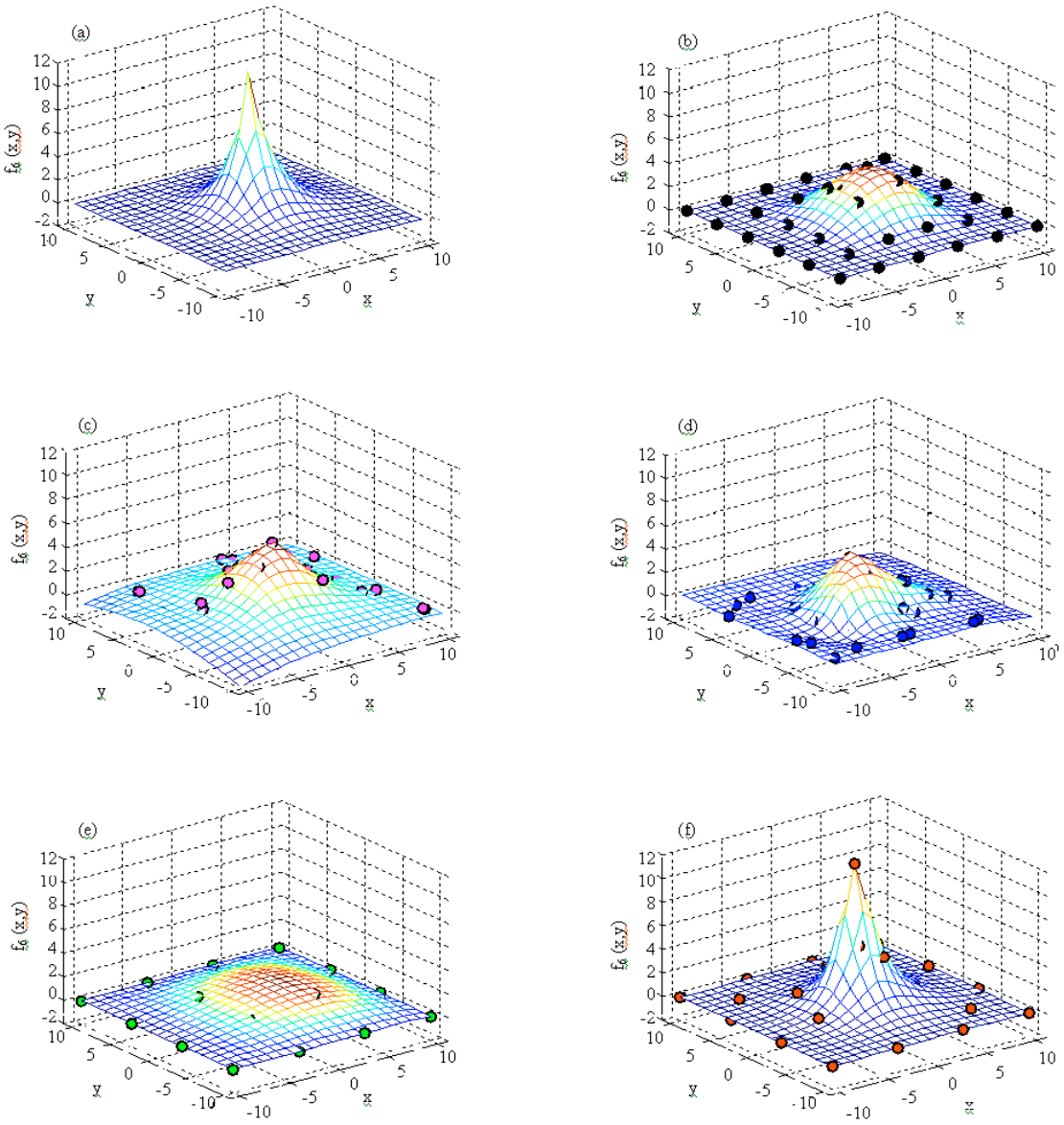

Fig. 5: (a) The diagram of the function f_6, (b) data extracted from f_6 by the Grid method, (c) data extracted from f_6 randomly, (d) data extracted from f_6 by the grid-random method, (e) extracted data from f_6 by the grid method applied in proposed method as initial data and (f) data extracted from f_6 by the proposed method

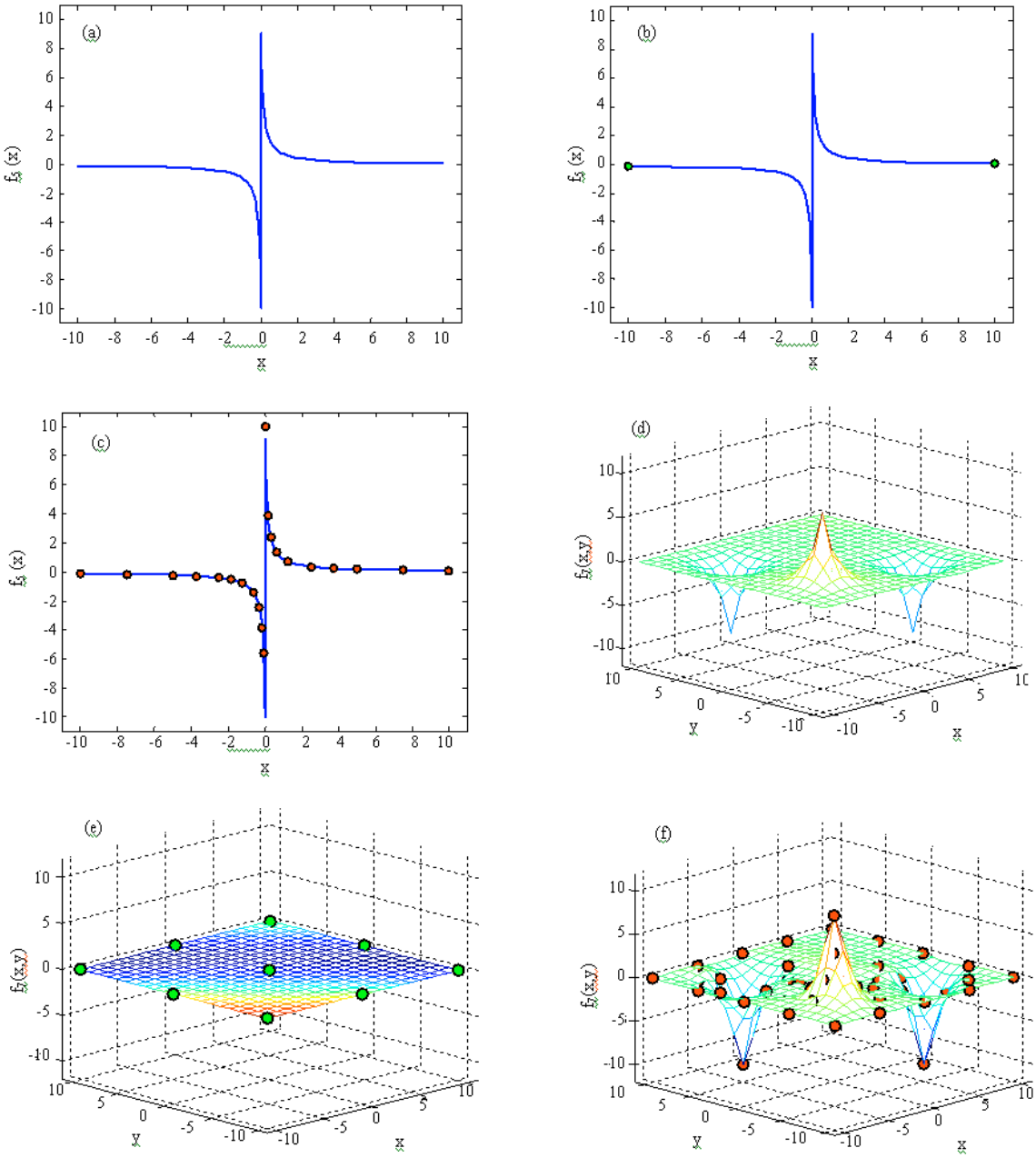

Fig. 6: (a) The diagram of the function f_5, (b) extracted data from f_5 by the grid method applied in proposed method as initial data, (c) data extracted from f_5 by the proposed method, (d) the diagram of the function f_7, (e) extracted data from f_7 by the grid method applied in proposed method as initial data and (f) data extracted from f_7 by the proposed method

CONCLUSION

In this study a novel method is presented for data extraction for function approximation. The proposed method is explained completely and its efficiency is demonstrated theoretically. The algorithm is verified by simulation results. More efficient fitness functions and other evolutionary operators should be investigated in the forthcoming contributions. Other intelligent algorithm such as neural networks should be examined for data extraction in the future.

REFERENCES

Dai, Q.X., Z.Z. Yuan, M.X. Luo and X.N. Cheng, 2004. Numerical simulation of Cr2N age-precipitation in high nitrogen stainless steels. Mater. Sci. Eng., 385: 445-448.

Daneshvar, N., A.R. Khataee and N. Djafarzadeh, 2006. The use of artificial neural networks (ANN) for modeling of decolorization of textile dye solution containing C.I. basic yellow 28 by electro coagulation process. J. Hazard. Mater., 137: 1788-1795.

Datta, S. and M.K. Banerjee, 2006. Mapping the input-output relationship in HSLA steels through expert neural network. Mater. Sci. Eng., 420: 254-264.

Elkamel, A., 1998. An artificial neural network for predicting and optimizing immiscible flood performance in heterogeneous reservoirs. Comput. Chem. Eng., 22: 1699-1709.

Hebbar, A., M. Mechmache D. Ouinas and A. Berras, 2006. Application of the experimental designs on the modeling of the combustion's parameters. J. Applied Sci., 6: 2599-2604.

Keong, K.G., W. Sha and S. Malinov, 2004. Artificial neural network modeling of crystallization temperatures of the Ni–P based amorphous alloys. Mater. Sci. Eng., 365: 212-218.

Khamis, A., Z. Ismail, K. Haron and A.T. Mohammed, 2006. Neural network model for oil palm yield modeling. J. Applied Sci., 6: 391-399.

Seifipour, N. and M.B. Menhaj, 2001. A GA-based algorithm with a very fast rate of convergence. 1st Edn. London, UK., pp: 185-193 .

Senthilkumar, M., 2006. Modeling of CIELAB values in vinyl sulphone dye application using feed-forward neural networks. Dyes Pigments, 75: 356-361.

Tercan, A.E. and A.I. Karayigit, 2001. Estimation of lignite reserve in the Kalburcayiri field, Kangal basin, Sivas, Turkey. Int. J. Coal Geol., 47: 91-100.

Turkoglu, M., I. Aydin, M. Murray and A. Sakr, 1999. Modeling of a roller-compaction process using neural networks and genetic algorithms. Eur. J. Pharm. Biopharm., 48: 239-245.