Expectation Maximization Clustering Algorithm for Use in Estimating Projects Costs

M. Bouabaz*. M. Belachia*.
M. Mordjaoui. B. Boudjema*****

**Civil Engineering Department, LMGHU Laboratory, University of 20August, Skikda, Algeria, Algeria (Tel: 21338701674; e-mail: mbouabaz@ hotmail.fr).*
** Civil Engineering Department, LMGHU Laboratory, University of 20August, Skikda, Algeria, (e-mail:Belachia@yahoo.fr)*
*** Electrical Engineering Department, LRPCSI Laboratory, University of 20August, Skikda, Algeria, (e-mail: Mordjaoui_mouradr@yahoo.fr)*
**** Physics Department, LRPCSI Laboratory, University of 20August, Skikda, Algeria, (e-mail:Boudjema_b@yahoo.fr)*

Abstract: In this paper, fuzzy logic approach is employed for predicting projects costs. Expectation maximization clustering algorithm was used, in order to optimize the number of clusters which is important in model simulation. The clusters given by the expectation maximization algorithm has lead to the development of fuzzy rules. The results indicate that the use of fuzzy rules to predicting projects costs has reduced the uncertainty of estimate, which in turn the accuracy was improved.

Keywords: Fuzzy clustering, Expectation maximization algorithm, Optimization, Simulation, Validity indices, Model identification.

1. INTRODUCTION

The concept of fuzzy logic is derived from the theory of fuzzy sets. The theory of fuzzy sets was developed by Zadeh (1965). It is ranked among the methods of artificial intelligence. The method of fuzzy logic provides a way for cope with problems arising from unexpected situations. It is a means for solving hard problems, by determining a mathematical model that describes the system behavior, known as an unsupervised learning method.
The proposed model consists of two parts: Firstly. Optimization is the process to determine the number of clusters from data input-output, which respectively serves for subsequent use. The second part involves the extraction of fuzzy rules.

In this paper, we evaluate the use of the expectation maximization clustering algorithm in modeling and estimating projects costs.

1.1 The fuzzy clustering algorithm

The Takagi-Sugeno was the first model developed in 1985. This model can effectively represent complex nonlinear systems using fuzzy sets. The clustering technique is an essential method in data analysis and pattern recognition. Fuzzy clustering allows natural grouping of data in a large data set and provides a basis for constructing rule-based fuzzy model. It is a partitioning method of data into subsets or groups based on similarities between the data (Takagi et al., 1985). The representation of fuzzy rules for the Takagi-Sugeno model takes the form :

$$R_i: \quad \text{If} \quad x_1 \text{ is } A_{i,1} \text{ and } \ldots..\text{and } x_p \text{ is } A_{i,p} \qquad (1)$$
$$\text{Then } y_I = a_{i0} + a_{i1}x_1 + \ldots.a_{ip}x_p$$

where, R_i is the rule number, x_j is the j-th input variable, A_{ij} is the fuzzy set of the j-th input variable in the i-th rule, y_i are output of the i-th fuzzy rule.

1.2 The fuzzy c-means algorithm

Fuzzy C-means algorithm (FCM) is a fuzzy clustering technique which is different from C-means that uses hard partitioning. The fuzzy c-means uses fuzzy partitioning in which a data point can belong to all clusters with different grades between 0 and 1. The FCM is an iterative algorithm that aims to find cluster centers that minimize the objective function.

$$J_{FCM}(Z;\Phi,V) = \sum_{i=1}^{c}\sum_{k=1}^{N}(\mu_{ik})^m D_{ikA}^2 \qquad (2)$$

where $V = [v_1, v_2, \ldots, v_c]$, v_i are the clusters centers to be determined. $\phi = \{\mu_{ik}\}$ is a fuzzy partition matrix; (μ_{ik}) is a membership degree between the i^{th} cluster and k^{th} data which is subject to conditions (3).

$$\mu_{ik} \in [0,1] \quad , \quad 0 < \sum_{k=1}^{N}\mu_{ik} < N, \quad \sum_{i=1}^{c}\mu_{ik} = 1 \qquad (3)$$

1.3 The expectation maximization algorithm

The EM algorithm was proposed by Abonyi (2003). It is an extension of the algorithm of Gath and Geva (1989), with a covariance matrix has nonzero diagonal elements, which creates an error in the projection of these elements. The methodology of the used algorithm is described as follow:

The partition matrix is expressed by:

$$\mu_{ik} = \frac{1}{\sum_{j=1}^{c} (D_{ikA}^2 / D_{jkA}^2)^{1/(m-1)}} \tag{4}$$

$$1 \le i \le c, \quad 1 \le k \le N$$

The prototypes (center) are:

$$\mathbf{v}_i^x = \frac{\sum_{k=1}^{N} \mu_{i,k}^{(l-1)} \mathbf{x}_k}{\sum_{k=1}^{N} \mu_{i,k}^{(l-1)}} \tag{5}$$

The standard deviation of the Gaussian membership functions is:

$$\sigma_{i,j}^2 = \frac{\sum_{k=1}^{N} \mu_{i,k}^{(l-1)} \left(x_{j,k} - v_{j,k}\right)^2}{\sum_{k=1}^{N} \mu_{i,k}^{(l-1)}} \tag{6}$$

The local model parameters are extracted as follows:

$$\theta_i = \left(\mathrm{X}_e^T \Phi_i \mathrm{X}_e\right)^{-1} \mathrm{X}_e^T \, \Phi_i y \tag{7}$$

where Φ is the weights matrix having the membership degrees defined by:

$$\Phi_i = \begin{bmatrix} \mu_{i,1} & 0 & \cdots & 0 \\ 0 & & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mu_{i,N} \end{bmatrix} X = \begin{bmatrix} X_1^T \\ . \\ . \\ X_N^T \end{bmatrix} y = \begin{bmatrix} y_1 \\ . \\ . \\ y_N \end{bmatrix}$$

The extended matrix X_e is given by; $X_e = [X \quad 1]$
And the prior probability is expressed by:

$$\alpha_i = \frac{1}{N} \sum_{k=1}^{N} \mu_{i,k} \tag{8}$$

The weights on rules are expressed as follow:

$$w_i = \prod_{j=1}^{n} \frac{\alpha_i}{\sqrt{2\pi\sigma_{i,j}^2}} \tag{9}$$

with a distance norm given by the following expression:

$$\frac{1}{D_{i,k}^2} = w_i . \exp\left(-\frac{1}{2} \frac{\left(x_{j,k} - v_{i,j}\right)^2}{\sigma_{i,j}^2} \right) . \frac{1}{\sqrt{2\pi\sigma_i^2}} . \exp\left(-\frac{\left(y_k - f_i(\mathrm{x}_k, \theta_i)\right)^T \left(y_k - f_i(\mathrm{x}_k, \theta_i)\right)}{2\sigma_i^2} \right) \tag{10}$$

where $f_i(\mathrm{x}_k, \theta_i)$ is the model consequents.

The proposed algorithm is summarized as follow

Let $Z=\{ z_{k1}, z_{k2},\ldots, z_{kn} \}^T$
Select the number of clusters c>1
Select the weithning exponent (**m** =2)
and the termination criterion (**ε**>0)
Initialize the partition matrix such as (**3**)
Start
1: Compute the clusters centers using (**5**)
2: Compute fuzzy covariance matrix by (**6**)
3: Compute (7, 8, and 9)
4: Compute the distances using (**10**)
5: Update the partition matrix using (**4**)
If stopping criterion $|U^{(l)} - U^{(l-1)}| \le \varepsilon$ satisfied then stop
Otherwise $l \leftarrow l+1$ and go to step 2
End

2. VALIDATION

2.1 Indices

It is important to determine the number of clusters for use in simulation. For this, different indices for validation have been proposed by Bezdek (1975) in data clustering. This can be done by the partition coefficient (PC) and the partition entropy (PE).

The partition coefficient Measure the ammount of overlapping between clusters.

$$PC(c) = \frac{1}{N} \sum_{i=1}^{c} \sum_{k=1}^{N} (\mu_{ik})^2 \tag{11}$$

The partition entropy measure the fuzzyness of the cluster

$$PE(c) = -\frac{1}{N} \sum_{i=1}^{c} \sum_{k=1}^{N} \mu_{ik} \log(\mu_{ik}) \tag{12}$$

where $PC(c) \in [1/c, 1]$, $PE(c) \in \left[0, \log_a\right]$, with an increase of c, the values of *PC* and *PE* are

decreased/increased, respectively. The above mentioned cluster validity indices are sensitive to fuzzy coefficient m. When $m \to 1$, the indices give the same value for all c. When $m \to \infty$, both *PC* and *PE* exhibit significant knee at c=2. The number corresponding to a significant knee is selected as the optimal number of clusters.

2.2 The performance error

To evaluate the performance of the Fuzzy models, we use the following criteria proposed by Bezdek. (1975). they are used for evaluating the output of the model.

The root mean squared error is expressed as follow:

$$RMSE = \sqrt{\frac{1}{N}\sum_{k=1}^{N}(y_k - \hat{y}_k)^2} \qquad (13)$$

where *y* is the output of the process and *y'* is the output of the model.

The variance accounted-for is expressed by:

$$VAF = 100\ \% \left[1 - \frac{\text{var}(y - \hat{y})}{\text{var}(y)}\right] \qquad (14)$$

Where n is the number of projects and i is the project number (i = 1… n), and *y'* is the actual output

3. APPLICATION OF FUZZY LOGIC TO COST ESTIMATION

The production of an accurate estimate for estimating projects costs is a challenging task for the estimator, at the early stage of a project.
In an attempt to overcome the problem, a soft computing method has been used to solve the problem of uncertainties in estimating, and construct accurate model for predicting the final cost of a project.

3.1 The model development

The present model has been developed in three phases. Firstly: It consists of the determination of the number of cluster. Secondly: The learning phase and finally the testing phase.

3.2 The data

Data used, comes from a research report (Bouabaz et al. 2008). A wide range of project contracts made on cost-significance work packages were used for modelling purposes

3.3 The clustering

The clustering consists of the selection of the number of clusters, which depends on the partition coefficient and the classification entropy.
The clustering map by the expectation maximisation algorithm is shown in figure 1.

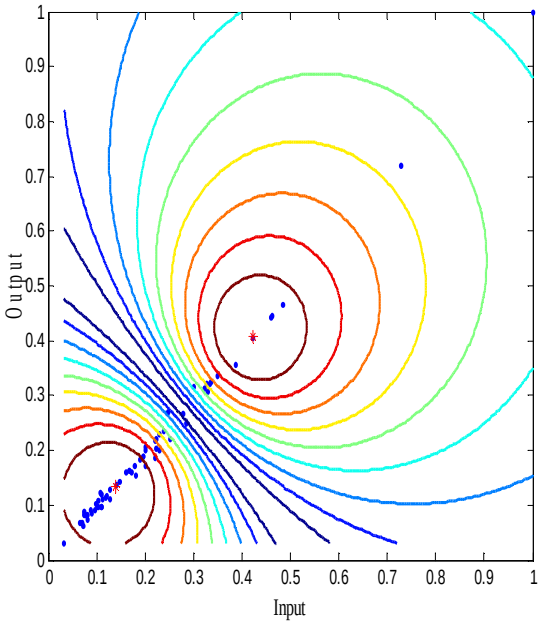

Fig 1. The clustering map using expectation maximisation algorithm.

As seen from figures 1 and 2 we can deduce that the clustering number is equal to2 clusters.

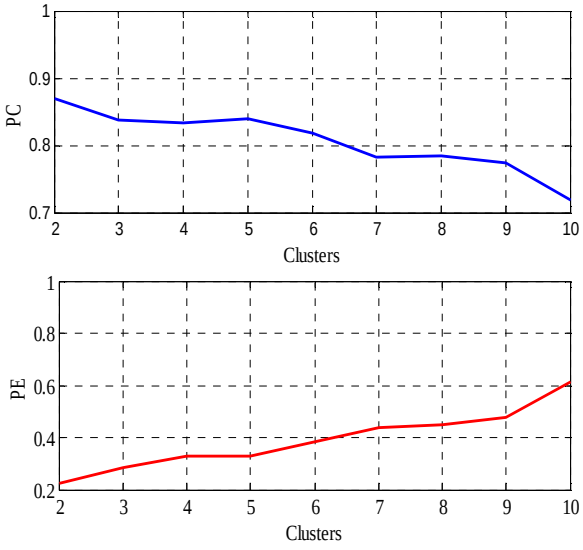

Fig 2. The results for PC and PE.

Table 1 gives the values for PC and PE at optimization stage.

Table 1. Summarize the clustering results.

Indices	Clusters								
	2	3	4	5	6	7	8	9	10
PC=	0.869	0.837	0.832	0.839	0.818	0.781	0.783	0.774	0.719
CE=	0.224	0.284	0.328	0.327	0.385	0.436	0.449	0.475	0.611

Table 2. The Values for the clusters centres

rules	y(k-1)	u
R1	1.61×10^{-1}	1.69×10^{-1}
R2	4.20×10^{-1}	4.28×10^{-1}

3.4 The model simulation

The proposed model was developed utilizing a set of projects. Historical data employed comes from a research report done in UK. The developed model was generated from 68 data samples using Matlab toolbox (Abonyi J. 2003, Abonyi et al. 2005) in a micro-computer. It has 1 input and 1 output. The training was stopped when the variance accounted-for (VAF) reached the maximum percentage value of 99.5304 in an elapsed time of 1.335000 seconds. The termination tolerance of the clustering algorithm was 0.01. The training error (RMSE) is 0.0107. The simulation model at learning phase is shown graphically on figure 3.

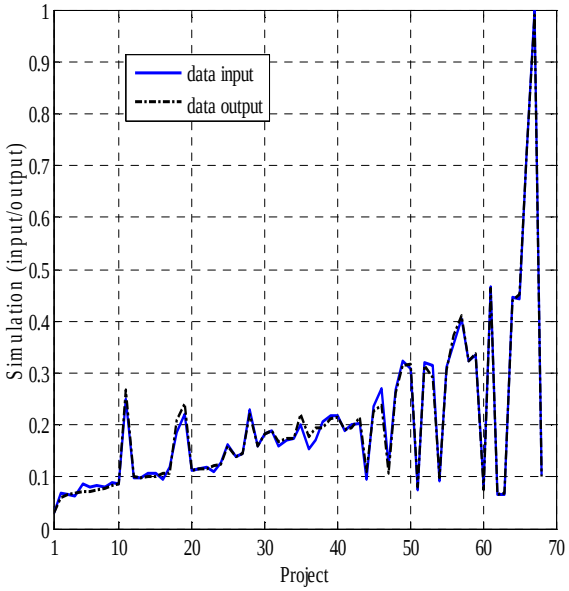

Fig 3. Simulated model at the learning phase.

The membership functions of the actual data versus simulated data obtained by the projection of the clusters by the expectation maximization algorithm are shown graphically.

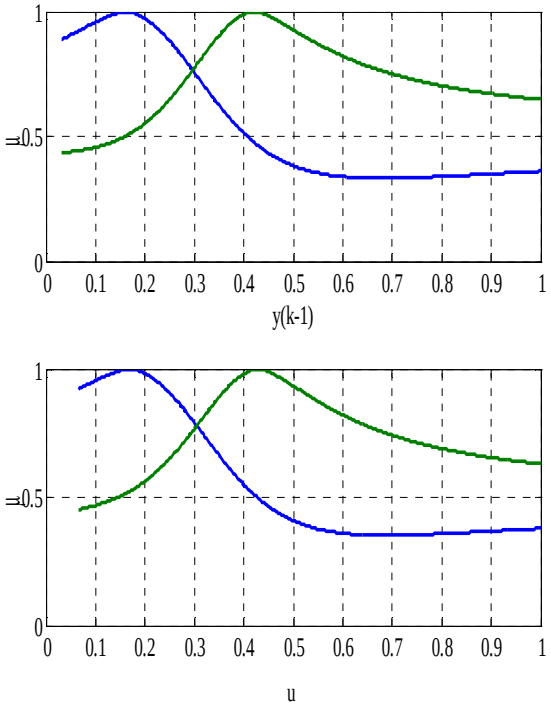

Fig 4. Show the membership functions of the model

Table 3. Fuzzy rules obtained by expectation maximisation algorithm

Rule 1: **If** *y(k-1)* ***is*** A_{11} **and** *u* **is** A_{12} **Then** $y(k) = 4.50.10^{-2}\, y(k-1) + 9.23.10^{-1} u + 1.04.10^{-3}$
Rule 2: **If** *y(k-1)* **is** A_{21} **and** *u* **is** A_{22} **Then** $y(k) = 5.27.10^{-3}\, y(k-1) + 1.04.10^{0} u + 270.10^{-2}$

3.5 The testing model

A testing phase was investigated on the adopted rules in order to determine the accuracy of the model. Some flattering results are shown in table 4.

Table 4. Results of testing rules

Project N°	Value of cswp's (£)	Simulated fuzzy model (£)	Actual bill value (£)	Cpe (%)
1	21 402	26 662	26 753	0.000
2	47 158	57 931	57 775	- 0.267
3	37 520	47 145	47 451	0.650
4	29 668	37 086	37 794	1.909
5	44 500	54 884	55 152	0.489
6	39 400	50 407	49 663	-1.475
7	57 898	72 373	71 702	-0.925
8	91 023	112 374	115 301	2.605
9	62 890	80 481	78 833	- 2.046
10	110 264	137 241	140 700	2.520
Mean error				0.345
Standard deviation				1.617

Figure 5 shows the plot of the results given by the obtained rules at testing phase.

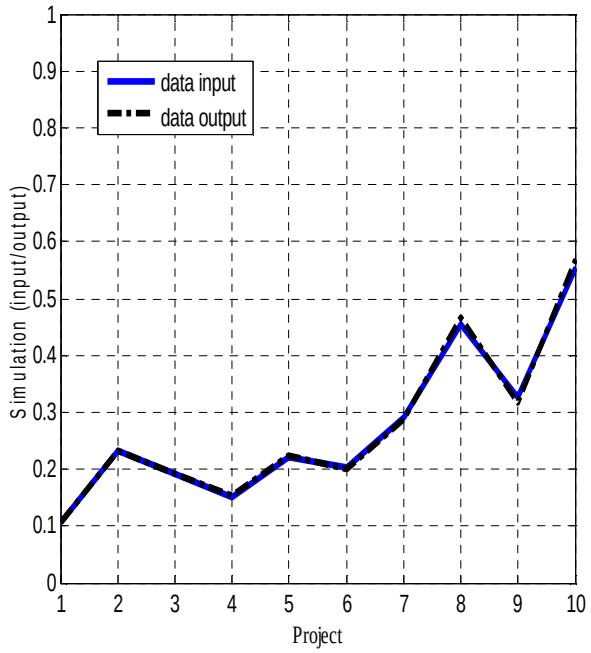

Fig 5. Plot of simulated values on actual at testing phase

4. Conclusion

As a conclusion fuzzy clustering approach seems to reveal promising results in modelling, and forecasting projects costs. A modified fuzzy clustering algorithm based on the expectation maximization algorithm was used to construct a fuzzy model.

The simulation results on projects contracts illustrate the accuracy of the model

References

Abonyi. J (2003). *Fuzzy Modeling Identification for Control. Birkhäuser, Boston.*

Abonyi. J Balasko. B, and Balazs. L (2005). Fuzzy Clustering and Data Analysis Toolbox for Use With Matlab.

Bezdeck, J C. (1975). Cluster Validity With Fuzzy Sets, *Cybernetics* **3**(3),pp 58-73.

Bezdeck. J C.; Ehrlich, R,; Full, W (1984). FCM: Fuzzy C-means Algorithm. *Computers and Geoscience*, **10**(2-3):pp191-203.

Bouabaz, M and Hamami. M (2008). A Cost Estimation Model for Repair Bridges Based On Artificial Neural Network. *American Journal of Applied Sciences* 5 (4) pp 334-339

Gath I and Geva, A.B (1989). Unsupervised Optimal Fuzzy Clustering. *IEEE Trans Pattern and Machine Intell*, vol. **11** (7), pp 773-781.

Takagi. T. and M. Sugeno (1985) Fuzzy Identification of Systems and its Applications to Modeling and Control, *IEEE Transactions on Systems, Man, and Cybernetics* **15**(1), pp 116-132.

Zadeh, L. A. 1965.Fuzzy sets. *Information and Control*. **8**, pp 338-353