MODIFICATION OF CONTINUOUS AND BINARY ICU DATA IN FCRM MODELS

Mohd Saifullah Rusiman[1], Robiah Adnan[2], Efendi Nasibov[3] and Kavikumar Jacob[4]
[1, 4]Department of Mathematics, Faculty of Science, Technology and Human Development, Universiti Tun Hussein Onn Malaysia, 86400 Parit Raja, Batu Pahat, Johor, Malaysia

[2]Faculty of Science, Universiti Teknologi Malaysia

[3]Faculty of Science and Arts, Dokuz Eylul University, Turkey

[1]saifulah@uthm.edu.my, [2]robiaha@utm.my, [3]efendi.nasibov@deu.edu.tr, [4]kavi@uthm.edu.my

This research is an attempt to present a proper methodology in data modification by using analytical hierarchy process (AHP) technique and fuzzy c-mean (FCM) model. The continuous data were built from binary data using analytical hierarchy process (AHP). Whereas, the binary data were created from continuous data using fuzzy c-means (FCM) model. The models used in this research are fuzzy c-regression models (FCRM). A case study in scale of health at an intensive care unit (ICU) ward using the AHP, FCM model and FCRM models was carried out. There are six independent variables involved in this study. There are four cases considered as a result of using AHP technique and FCM model toward independent data. After comparing the four cases, it was found that case 4 appeared to be the best model, having the lowest mean square error (MSE). The original data have the MSE value of 97.33, while the data of case 4 have MSE by 83.48. This means that the AHP technique can lower the MSE, while the FCM model cannot lower the MSE in modelling scale of health in the ICU. In other words, it can be claimed that the AHP technique can increase the accuracy of modelling prediction.

Keywords: Analytical hierarchy process (AHP), fuzzy c-means (FCM) model, fuzzy c-regression models (FCRM), mean square error (MSE)

1. INTRODUCTION

Regression model has become one of the standard tools in data analysis since the mathematical equation from its analysis could explain the relationship between the dependent and independent variables. It is available in computer packages, easy to interpret and has been widely used in applied sciences, economic, engineering, computer, social sciences and other field. Fuzzy modelling has become popular for the past few years because it describes complex systems better. The fuzzy c-mean (FCM) model proposed by Bezdek in 1981 develops hyper-spherical-shaped clusters. In contrast, the fuzzy c-regression models (FCRM) proposed by Hathaway and Bezdek [1], develop hyper-plane-shaped clusters. Analytical hierarchy process (AHP) has been proposed by Thomas L. Saaty in 1977 in handling factor weights due to a lack of historic information. It has been widely used in decision making, since it includes the

natural feelings of human beings.Many researchers employ AHP technique in handling data mining problem [2].

The intensive care unit (ICU) plays an important role in the medical care sector not only for the critically ill, who makes up 5% of inpatients, but also in terms of generating a major contribution of health care funds. The United States health care industry makes up 15-20% total hospital cost. In 1968, the first ICU in Malaysia was established. Intensive care has then developed rapidly and it is now available in all tertiary care hospitals and selected secondary care hospitals. The National Audit on Adult Intensive Care Units in Malaysia in 2002 is modeled on the UK experience in 1994 and coordinated by a national committee comprising of senior intensive care specialists in the Ministry of Health. This audit unit develops a national database to assess fundamental aspects of intensive care functions within a hospital. The clinical indicators developed by ACHS (The Australian Council on Healthcare Standard) are useful tools for clinicians to flag potential problems and areas for improvement [3].

Currently, there was common method used in ICU involves logistic regression [4]. Only Pilz and Engelmann [5] did a basic fuzzy rule to determine the medical decision in ICU. This work inspires us to do work in fuzzy model into ICU area that could give a challenge to this study. The first research on mortality rates in Malaysian ICU has been done at a general hospital in Ipoh, involving only a logit model [3]. The second research is continued by Mohd Saifullah Rusiman et al. [6] on the analysis of logit, probit and linear probability models. As a comparison among the three models, logit model has been appeared to be the best model.

The objective of this research is to explore data modification using AHP technique and FCM model in scale of health at the ICU. The other objective is to make a comparison among the beginning data (without any method), AHP technique, FCM model or any combination of methods which are applied to data in order to find the best model. So, we can make recommendation based on this data mining method in predicting scale of health in the general hospital.

2. MATERIAL

In this study, the data were obtained from the intensive care unit (ICU) of a general hospital in Johor . The data obtained were classified as a cluster sampling. It involves 1311 patients in the ICU within the interval of January, 2001 to August, 2002. The dependent variable is scales of health or score of SAPS II discharge from hospital (s2sdisc). There are six independent variables considered in this study which are sex, race, organ failure (orgfail), comorbid diseases (comorbid), mechanical ventilator (mecvent) and score of SAPS II admit (s2sadm). The s2sdisc and s2sadm scores are 15 accumulated values for heart rate, blood pressure, age, body temperature, oxygen pressure, urine result, urea serum level, white blood count, potassium serum level, sodium serum level, bicarbonate serum level, bilirubin level, glasgow coma score, chronic illness and type of admittance as proposed by Le Gall et al. [7].

3. METHODOLOGY

3.1 AHP technique

The AHP technique is a complete decision making process that permits more complete consideration of multi-factors/criteria. The AHP procedure involves three steps as;

Step 1: Establish the decision hierarchy

In this step the decision maker must identify the overall decision, the factors that must be weighted or used to make the decision and the alternative choices from which a decision it to be made. Once these are identified they are placed in a decision hierarchy.

Step 2: Compute the weighted of alternatives

In this step the decision maker or expertise must compare each alternative with all other alternatives for one factor at one time. The rating measure scale used to rate the alternatives on a range from 1 to 9 as it relates to each of the factors. The weighted or probabilities obtained from a paired comparison matrix, summing to 1.

Step 3. Compute the weighted of factors

In this step the decision maker uses the previously determined comparison ratings to compute a set of priorities for the individual factors. This involves several small computation sub-steps where the probabilities or weighted obtained from a paired comparison matrix with the total of one [2].

3.2 FCM model

In FCM clustering, based on the Dunn [8] and Bezdek [1] algorithm, we have to minimise the criterion J in (1),

$$J = \sum_{j=1}^{c} \sum_{i=1}^{N} u_{ij}^{w} d_{ij}^{2} \, , \, w > 1 \tag{1}$$

where u_{ij} is the membership values, d_{ij}^{2} is $\left\| x_i - \sum_{i=1}^{N} u_{ij}^{w} x_i \Big/ \sum_{i=1}^{N} u_{ij}^{w} \right\|^2$ or the Euclidean distances, N is the number of objects, c is the number of clusters and w is the weight or fuzzifier. In order to minimize (1), we have to;

(a). Fix the value of c. Initialise membership values **U** or u_{ij} at random. Choose the termination tolerance $\delta > 0$.

(b). Update Euclidean distances, d_{ij} for given **U** by computing the weighted averages for each group.

(c). Update membership values, $u_{ij} = \dfrac{1}{\sum_{k=1}^{C} \left(\dfrac{d_{ij}}{d_{ik}} \right)^{\frac{2}{w-1}}}$, for $w > 1$

$$u_{ij} = \begin{cases} 1 & \text{if } d_{ij} = \min(d_{ik}) \\ 0 & \text{otherwise} \end{cases} \quad \text{for } w = 1 \tag{2}$$

(d). Calculate the criterion J in (1) and check for convergence. If $|J_{old} - J_{new}| < \delta$ stop the iteration, else go to step (b).

3.3 FCRM models

There are no conditions needed in FCRM models. Based on the algorithm in Hathaway & Bezdek [1], Abonyi & Feil [9] and Kung & Su [10], we have to,

(a) Fix the number of cluster c, $c \geq 2$. Choose the termination tolerance $\delta > 0$. Fix the weight, w, $w > 1$ (a common choice in practice is to set $w = 2$) and initialise the initial value for membership function matrix, $\mathbf{U}^{(0)}$ satisfying (4).

(b) Estimate $\theta_1,..., \theta_c$ simultaneously by modifying the FCM algorithm. If the regression functions $f_i(x;\theta_i)$ are linear in the parameters θ_i, the parameters can be obtained as a solution of the weighted least squares,

$$\theta_i = [\mathbf{X}_b^T \mathbf{W}_i \mathbf{X}_b]^{-1} \mathbf{X}_b^T \mathbf{W}_i \mathbf{Y} \tag{3}$$

where $\mathbf{X}_b = [\mathbf{X},1]$.

(c) Calculate the objective function:

$$E_w[\mathrm{U},\{\theta_i\}] = \sum_{i=1}^{c}\sum_{j=1}^{d} u_{ij}^w E_{ij}[\theta_i] \tag{4}$$

where

(i) u_{ij} is membership degree $(i = 1,\ldots, c\ ; j = 1,\ldots, N)$.

(ii) $E_{ij}[\theta_i]$ is the measure of error with $E_{ij}[\theta_i] = \| Y_j - f_i(X_j;\theta_i) \|^2$. The most commonly used is the squared vector Euclidean norm for $Y_j - f_i(X_j;\theta_i)$.

(d) Do iterations in order to minimize the objective function in (4). Repeat for l = 1, 2,..., ∞ until $\| \mathbf{U}^{(l)} - \mathbf{U}^{(l-1)} \| < \delta$. Next, follow the steps below:

Step 1 : Calculate model parameters $\theta_i^{(l)}$ to globally minimize (4).

Step 2 **:** Update **U** with $E_{ij} = E_{ij}[\theta_i^{(l-1)}]$, to satisfy:

$$u_{ij}^{(l)} = \begin{cases} \dfrac{1}{\displaystyle\sum_{k=1}^{c}\left(\frac{E_{ij}}{E_{kj}}\right)^{\frac{2}{w-1}}}, & \text{for } I_j = \phi \\ 0, & \text{for } I_j \neq \phi \text{ and } i \notin I_j \end{cases} \tag{5}$$

where $I_j = \{ i \mid 1 \leq i \leq c \text{ and } E_{ij} = 0 \}$

until $\| \mathbf{U}^{(l)} - \mathbf{U}^{(l-1)} \| < \delta$.

The mean square error (MSE) is used as follow,

$$MSE = \frac{1}{N} \sum \left(Y_i - \hat{Y}_i \right)^2 \tag{6}$$

where Y_i denotes the real data, $\hat{Y}$ represents the predicted value of Y_i and N is the number of data.

4. DATA ANALYSIS

4.1 AHP technique

The AHP technique is applied to organ failures variable (orgfail). This independent variable only has two binary data, that is, patients who have or do not have organ failures. This technique will fuzzify the binary data of organ failures to be a continuous data within the interval [0, 1].

Organ failures are divided into 6 types which are: (*A*) Respiratory failure, (*B*) Cardiovascular failure, (*C*) Neurological failure, (*D*) Renal failure, (*E*) Hepatic failure and (*F*) Haematological failure. Referring to the expert physicians in the general hospital, they stated that *B* and *D* have a twice higher probability that contribute to high mortality if compared to the *A* and *F*. In fact, the *A* and *F* have a twice higher probability if compared to the *C* and *E*. However, *B* and *D* have the same weightage. The same weightage are also given to the *A* and *F*. *C* and *E* also receive the same weightage. The paired comparison matrix and probabilities (weighted) are shown in Table 1.

Table 1. The paired-comparison matrix and weighted for organ failures

	B	*D*	*A*	*F*	*C*	*E*	Total	Weighted
B	1	1	2	2	4	4	14	0.2857
D	1	1	2	2	4	4	14	0.2857
A	½	½	1	1	2	2	7	0.1429
F	½	½	1	1	2	2	7	0.1429
C	¼	¼	½	½	1	1	3.5	0.0714
E	¼	¼	½	½	1	1	3.5	0.0714

4.2 FCM model

In order to get categorical data of s2sadm with ‘1’ and ‘2’ coded, we have to cluster s2sadm data based on FCM clustering algorithm. The data for cluster 1 with 860 data ranges from 0 to 43 whereas the data for cluster 2 with 443 data ranges from 44 to 126. This is the same as the cluster given by the doctors who indicated that the s2s score over 43 is classified as a bad condition.

4.3 FCRM models

In this study, there are four cases considered as in Table 2 as a result of using AHP and FCM model toward independent data. The four cases involves six variables with different combination of variable types in each case were considered in order to find the best model using FCRM models. The variables involved are sex (x_1 is binary), race (x_2 is category), orgfail (x_3 is binary or continuous), comorbid (x_4 is binary), mecvent (x_5 is binary) and s2sadm (x_6 is binary or continuous). Case 3 is the beginning data without any modification being carried out.

Table 2 : Different case of multivariate data (Y vs x_1, x_2, x_3, x_4, x_5, x_6)

Case	1	2	3	4
x_1	B	B	B	B
x_2	Ca	Ca	Ca	Ca
x_3	B	Co	B	Co
x_4	B	B	B	B
x_5	B	B	B	B
x_6	B	B	Co	Co
MSE for MLR(SV)	632.14 (VA)	526.40 (VA)	498.29 (VB)	463.10 (VA)
MSE for FCRM(AV)	116.05	114.71	98.28	84.01
MSE for FCRM(SV)	121.92(VA)	97.29(VA)	97.33(VB)	83.48(VA)

Note:
B:Binary, Ca:Category, Co:Continuous, AV:All variables SV:Significant variables(VA, VB)
VA: 4 Variables chosen - x_1, x_3, x_4 & x_6 VB: 5 Variables chosen - x_1, x_3, x_4, x_5 & x_6

There are four cases considered as a result of combination cases with/without using AHP technique and/or FCM model toward independent data. Table 3 shows that case 4 is the best case with the lowest MSE, that is, when x_1 is binary, x_2 is category, x_3 is continuous, x_4 is binary, x_5 is binary and x_6 is continuous. The MSE value for FCRM models for case 4 is 84.01 (all variables) and 83.48 (significant variables - x_1, x_3, x_4, x_6). The MSE value for significant variables shows better result. The MSE value for case 3 (original data) is 97.33, while the MSE value for case 4 is 83.48. The MSE values for the other cases are 97.29 and 121.92. In conclusion, case 4 is the best case in which data modification involves only the orgfail variable. These chosen models (y vs x_1, x_3, x_4, x_6) are represented in (7) with two clusters.

Cluster 1

R^1 : IF x_1 is A_1^1 and x_3 is A_3^1 and x_4 is A_4^1 and x_6 is A_6^1

$$\text{THEN } y^1 = 2.4644x_1 + 12.8113x_3 + 4.6925x_4 + 0.1721x_6 + 61.2967$$

Cluster 2

R^2 : IF x_1 is A_1^2 and x_3 is A_3^2 and x_4 is A_4^2 and x_6 is A_6^2

$$\text{THEN } y^2 = 1.2257x_1 - 1.8245x_3 + 4.0093x_4 + 0.4788x_6 - 4.8764 \quad (7)$$

5 CONCLUSION

In this research, data modifications were done to the case study in ICU where the binary data (s2sadm variable) were built from continuous data using FCM model, whereas the continuous data (orgfail variable) were created from binary data using AHP technique. There are four cases considered as a result of combination cases

with/without using AHP technique and/or FCM model toward independent data. After comparing the four cases, it was found that case 4 appeared to be the best model, having the lowest MSE of 83.48, while the original data have the MSE value of 97.33. This means that the AHP technique can lower the MSE value, while the FCM model cannot lower the MSE in modelling scale of health in the ICU. In other words, it can be declared that the AHP technique can increase the accuracy of modelling prediction and should be used as a reference in hospitals to improve data accuracy.

References

[1] Hathaway, R. J. & Bezdek, J. C. (1993). Switching Regression Models and Fuzzy Clustering. *IEEE Transactions on Fuzzy Systems*, 1: 195–204.

[2] Saaty, T. L., Peniwati, K. & Shang, J. S. (2007). The Analytic Hierarchy Process and Human Resource Allocation: Half the story. *Mathematical and Computer Modelling*, 46 (7-8): 1041–1053.

[3] The Committee for National Audit on Adult Intensive Care Units (2002). *Protocol* : *National Audit on Adult Intensive Care Units*.

[4] Colpan, A., Akinci, E., Erbay, A., Balaban, N. & Bodur, H. (2005). Evaluation of Risk Factors for Mortality in ICUs: A Prospective Study from a Referral Hospital in Turkey. *American Journal of Infection Control*, 33: 42-47.

[5] Pilz, U. & Engelmann, L. (1998). Integration of Medical Knowledge in an Expert System for Use in Intensive Care Medicine. *Fuzzy and Neuro-Fuzzy Systems in Medicine*, 2: 290-315.

[6] Rusiman, M. S., Mohd Daud, Z. & Mohamad, I. (2004). The Comparison between Logit, Probit and Linear Probability Model toward Mortality Rate at ICU General Hospital. *Jurnal Statistika Universitas Islam Bandung*, 4: 129-138.

[7] Le Gall, J-R., Lemeshow, S., Saulnier, F (1993). A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study. *The Journal of the American Medical Association*, 24: 2957-2963.

[8] Dunn, J. C. (1973), A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters, *Journal of Cybernetics*, 3: 32-57.

[9] Abonyi, J. & Feil, B. (2007). *Cluster analysis for data mining and system identification*, USA: Springer.

[10] Kung, C. C. & Su, J. Y. (2007). Affine Takagi-Sugeno fuzzy modelling algorithm by fuzzy c-regression models clustering with a novel cluster validity criterion, *IET Control Theory Application*, 1: 1255–1265.