*i*Journals: International Journal of Software & Hardware Research in Engineering (IJSHRE)
ISSN-2347-4890
Volume 9 Issue 3 March 2021

Estimate the Missing Value in the Relational Database by the Integrated Multi-disciplines: Computational Intelligence and Statistics

Author: Shin-Jye Lee[1]; Ching-Hsun Tseng[2]; Hsueh-Cheng Liu[1]

Affiliation: *Institute of Management of Technology, National Chiao Tung University, Hsinchu, Taiwan [1]; Department of Computer Science, University of Manchester, United Kingdom [2]*

E-mail: camhero@gamil.com[1]; hank131415go61@gmail.com[2]; stdm11528@gmail.com[3]

DOI: 10.26821/IJSHRE.9.3.2021.9316

ABSTRACT

Nowadays, there are so many methods trying to find out an optimal solution for estimating the missing value in the relational database with a reliable accuracy performance. Due to the purpose, this research proposal proposes two integrated methodologies trying to estimate the missing value in the relational database with a very high accuracy rate, and the relevant theories include Fuzzy Theory, Genetic Algorithms, the method of Least Squares and Simple Linear Regression Model. Meanwhile, these theories can be integrated to work into more than one methodology. Also, each methodology gets various results by their peculiarities, and which will lead to their unique results respectively. Moreover, the proposed methodologies can not only overcome the drawbacks of inadvertently ignored problems, but also can make a very high estimated accuracy rate of this achievement.

Keywords: Null Value Estimation, Computational Intelligence

1. INTRODUCTION

Being highly reliable, Computation Intelligence is one of popular subjects applied at various fields in the industrial engineering, and Data Mining plays an important role in finding the pattern by Computational Intelligence algorithms. In addition, effectively estimating the missing value in the relational database has also been considered. To attain a reliable performance with a good accuracy, more elaborated methods are designed by either reinforcing or integrating the existing methods. Therefore, how to develop advanced methods and use them as a foundation to discover the hidden information has become an increasing issue. Also, there exit several ways to improve the problem and make the achievement optimal, and one of which is work it out by integrating multi-disciplines. This research aims to develop a kind of integrated methodology with Fuzzy Theory, Genetic Algorithms, the method of Least Squares and Simple Linear Regression Model. In order to try and understand what are at stock, I'll focus my research on the following main topics.

The structure of this initial research proposal is arranged as follows: In Section 2, briefly review the basic concepts and application of Fuzzy Theory. In Section 3, briefly introduce the basic concepts and process of Genetic Algorithms. In Section 4, briefly review the method of Least Squares and its functions used in this proposal. In Section 5, briefly introduce the basic concepts of Simple Linear Regression Model. In Section 6, present two methodologies for computing the estimated value with a good accuracy. Finally, this work is concluded in Section 7.

2. FUZZY THEORY

In 1965, Lotfi A. Zadeh proposed the theory of fuzzy sets [1], and the knowledge of fuzzy logic has therefore been developed gradually. Also, a fuzzy system is constructed based on the fuzzy if-then rules. To verify the data set and then sort it by precisely logical order, the main purpose of membership functions is distinguishing the certain data set from various kinds of data sets by transforming from general data set into fuzzy set, and then places it at the appropriate level.

Afterwards, in accordance with the fuzzified results by membership functions, the Fuzzy IF-THEN rules have been structured. As a whole, a special feature of Fuzzy Theory is its unique transparency – good understanding as a result of clear membership functions. Most important of all, predict the highly precise information by insufficient data is the superexcellent uniqueness of Fuzzy Theory.

2.1 Basic Configuration of Fuzzy Systems

Fuzzy IF-THEN rules is one of essential concepts in Fuzzy Theory. It's structured on which factors lead to which actions ignited by Fuzzy Systems. For example, for safety, if the fire sensor detects the density of the smoke is higher than that of regular standard in the air, then the fire alarm works immediately. Meanwhile, the Fuzzy IF-THEN rule of this example is described in the following form (1):

IF the density of the smoke is higher than 20%, THEN the fire alarm works immediately (1)

For carrying Fuzzy IF-THEN Rules out, Fuzzy System is a significant system working out the whole procedure from a factor to an action. Also, the basic configuration of fuzzy systems is shown in Fig. 1. [2]. A basic configuration of Fuzzy System includes Fuzzifier, Defuzzifier, Fuzzy Rule Base and Fuzzy Inference Engine, and every component plays its necessary role in the whole Fuzzy System. As shown in Fig. 1, first, the Fuzzifier transforms a real-valued variable (x in U) into a fuzzy set, and then the Fuzzy Inference Engine based on Fuzzy logic maps the fuzzy set with Fuzzy Rule Base. Final, the fuzzy sets come out by Fuzzy Inference Engine will be transformed into a real-valued variable (y in V) by Defuzzifier.

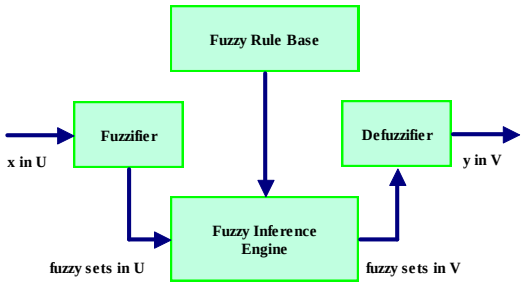

Fig 1:Basic configuration of fuzzy systems [2]

2.2 Fuzzy Memberships Functions

Transparency is a unique feature in Fuzzy Systems, and the concept of membership functions is one of significant concepts making Fuzzy Systems simplicity. As a result of the feature of simplicity, it's not difficult to understand the detailed information by membership functions. Basically, each element would be transformed into a fuzzy set in the membership functions. Also, each fuzzy set is associated with a membership value (between 0 and 1) in membership functions, and the sum of the entire fuzzy sets is 1 absolutely. Let U be the universe of discourse, and μA(x) be the membership functions of a fuzzy set A. Therefore, the definition of the fuzzy set A is represented as follows [3]:

$$A = \{(x, \mu A(x) \mid x \in U\} \tag{2}$$

$$A_i(x) = \begin{Bmatrix} 1, x \in A \\ 0, x \notin A \end{Bmatrix} \tag{3}$$

Moreover, two types of membership functions are defined, comprising continuous membership functions and discrete membership functions. Also, the differentiation between these two types membership functions depends on the universe of discourse U it's been belonged. In case the U is continuous, it would be continuous membership functions. Otherwise, the U is discrete, and it would be discrete membership functions. Meanwhile, the definition of continuous membership function is represented as follows [3]:

$$A = \int_U \mu A(x) / x \tag{4}$$

Also, the definition of discrete membership function is represented as follows [3]:

$$A = \sum_U \mu A(x) / x \tag{5}$$

Meanwhile, a significant definition of membership functions, the value of the sum of the entire fuzzy sets is 1 absolutely, and which can be defined as follows:

$$Y = f(x) = \sum_{i=1}^{n} Ai(x) Yi = 1 \tag{6}$$

Whereas the equation is provided at (6), the concept of complement of membership functions can be deduced from (6) [3]. In the equation (7), μĀ(x) is the complement of μA(x). In another word, Ā is the complement of the fuzzy set A, and the sum of A and Ā is 1.

$$\mu A(x) + \mu \bar{A}(x) = 1 \tag{7}$$

$$\mu \bar{A}(x) = 1 - \mu A(x) \tag{8}$$

The chart of membership functions for A and Ā is shown in the Fig. 2 [3]. In the Fig. 2, the definition of the relationship is A(μA(x)) + Ā(μĀ(x)) =1. When A(μA(x)) increases, meanwhile Ā(μĀ(x)) decreases. On the other hand, Ā(μĀ(x)) increases, and then A(μA(x)) decrease at the same time. However, these two values change, the sum of A(μA(x)) and Ā(μĀ(x)) is 1 eternally.

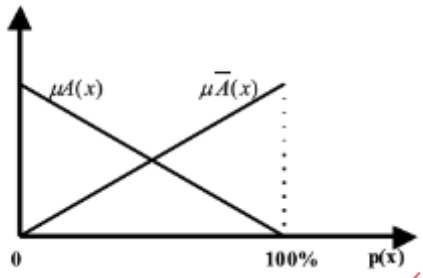

Fig 2:Membership functions for A(μA(x)) and Ā(μĀ(x)) based on the percentage of A(μA(x)) + Ā(μĀ(x)) =1

3. GENETIC ALGORITHMS (GA)

Genetic Algorithms was proposed by Holland in 1975 [4], and it is been applied as an adaptive algorithm for finding global optimal solution, which can be also considered as a computational mode of natural evolutionary systems [5]. In Genetic Algorithms, the main steps are: (a) Format of a Chromosome; (b) Calculation of the Fitness Degree; (c) Selection Operations; (d) Crossover Operations; (e) Mutation Operations. Moreover, for each generation, the Genetic Algorithms keep performing the Selection Operations, Crossover Operations and Mutation Operations until the Genetic Algorithms converge [6]. In another word, the GA system keeps working recursively until the best chromosome has been created. Definitely, the essential goal of GA is to pick the best chromosome out in the whole procedure of generation.

4. SIMPLE LINEAR REGRESSION MODEL (SLRM)

According to the explanation of Oxford Dictionary of Computing, Simple Linear Regression Model is a technique of Statistics that is concerned with fitting relationships between a dependent variable and one or more independent variables [7]. In another word, it can be used to mine the relationship between two variables, so it also plays a unique role in Data Mining. In a Simple Linear Regression Model (SLRM), the variable y_i equals to a linear function of another variable x_i plus a random term ε_i. Moreover, the formula of SLRM can be represented by $y_i = \alpha + \beta x_i + \varepsilon_i$, where y_i is dependent variable, x_i is independent variable, ε_i is disturbance or error, and α and β are regression coefficients or regression parameters [8]. Most important of all, the purpose of SLRM is to make the error minimize, and get a higher accuracy in the whole research.

5. THE METHOD OF LEAST SQUARES

The method of Least Squares is also one of prevalent methods to minimize the error rate at a variety of disciplines, and has been applied in many fields nowadays. According to the explanation of Oxford Dictionary of Computing [9], the method of Least Squares focuses on estimating parameters in a model by minimizing the sum of squares of differences between observed and theoretical values of a variable [9], and its purpose aims to make the estimated value simplicity. Further, the equation of the method of Least Squares can be calculated as follows [10]:

$$S = \sum_{i=1}^{n} [f(x_i) - y_i]^2 \tag{9}$$

where S means the sum of the squared errors, $f(x_i)$ presents the observed value of a variable, and y_i presents the theoretical values of a variable.

Moreover, in case the goal would like to come out the average squared errors, the equation can be calculated as follows:

$$A(S) = Min1/n\sum_{i=1}^{n}[f(x_i) - y_i]^2 \quad (10)$$

$$A(S) = Min1/n\sum_{i=1}^{n}[\frac{y_i - \sum_{t=1}^{k} f(x_i)}{y_i}]^2 \quad (11)$$

where *A(S)* means the average squared errors. (5.2) and (5.3) show two types of *A(S)*.

However, the weakness of the method of Least Squares is: The data set is farther away the least squares regression line, the inaccuracy is higher; On the contrary, the data set is closer to the least squares regression line, the accuracy is higher.

6. METHODOLOGIES OF ESTIMATING THE MISSING VALUE IN THE RELATIONAL DATABASE

Assume that there are missing values in the relational database, and the purpose of these methodologies aims to estimate the missing value with a good accuracy. Here, two methodologies have been introduced to estimate the missing value in the relational database with a reliable performance. Meanwhile, the procedure should be respectively described as follows:

A. Methodology I

Values will be fuzzified and then come Fuzzy rules out by Fuzzy Systems first, and then compute with the method of Least Squares. Final, the values will be computed with Simple Linear Regression Model. Therefore, the simply regular steps can be described as follows:

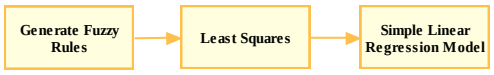

Fig 3: The procedure of Methodology I

(1) Generate Fuzzy Rules: In the light of values provided by the relational database, the values will be fuzzified and then come Fuzzy rules out by Fuzzy Systems. Definitely, the task of this step is transformed the real-valued variable to the regular fuzzy sets by using membership functions, and the definition can be described as follows:

$$\Delta = f(X,...,Z) = \sum_{i=1}^{l}...\sum_{k=1}^{n} A_i(x)...A_k(z)\Delta_{i...k} \quad (12)$$

where Y means the entire fuzzy rules, X,…,Z present the attributes of membership functions from X to Z, $\Delta_{i...k}$ presents each tuple with respect to the relational database. Totally, the number of fuzzy rules is the product of X to Z.

(2) Compute Fuzzy sets with the method of Least Squares: After Fuzzy Rules are generated, the values will be computed with the method of Lease Squares. Meanwhile, the task of this step calculates the fuzzy sets with the method of Least Squares, and its equation is defined in (10). Further, the equation based on (10) can be evolved as follows:

$$S = \sum_{i=1}^{n}[f(x_i) - y_i]^2$$

Base:

Evolution:

$$(Y - \Phi\theta)^2 = (Y - \Phi\theta) \times (Y - \Phi\theta)^T$$

=

$$YY^T - Y\Phi^T\theta^T - Y^T\Phi\theta + \Phi\theta\Phi^T\theta^T$$

$$= YY^T - 2Y\Phi^T\theta^T + \Phi\Phi^T\theta^2$$

Then evaluate the value of θ by differentiation:

$$f'(\theta) = 2\Phi\Phi^T\theta - 2\Phi^T Y = 0$$
$$2\Phi\Phi^T\theta = 2\Phi^T Y$$
$$\Phi\Phi^T\theta = \Phi^T Y$$
$$(\Phi\Phi^T)^{-1}\Phi\Phi^T\theta = (\Phi\Phi^T)^{-1}\Phi^T Y$$

$$\theta = (\Phi\Phi^T)^{-1}\Phi^T Y$$

Thus, it can be defined as follows:

$$\theta = (\Phi\Phi^T)^{-1}\Phi^T Y \quad (13)$$

where θ means the output variable, Φ presents the matrix ofinput values,

and *Y* presents the theoretical output values.

(3) Calculate the advanced Weighted Fuzzy Rules by Simple Linear Regression Model: An optimal equation or best weights of attributes will be come out by computing with Simple Linear Regression Model in the final step. The goal of this step is to find out the optimal regression line to close in the absolutely accurate value in the relational database, and the least squares regression line [10] has been used for evaluating the optimal value. Also, the equation and definition of the least squares regression line can be defined as follows:

The least squares regression line for $\{(x_1, y_1),...,(x_n, y_n)\}$ is given by:

$$Y = f(x) = ax + b \qquad (14)$$

where x_i and y_i mean observed and theoretical values of a variable, the definition of *a* and *b* is respectively calculated as follows:

$$a = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

$$b = \frac{1}{n}(\sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} x_i)$$

This equation presents the optimal regression line for the final solution evaluating the estimated value with a very high estimated accuracy rate.

B. **Methodology II**

Values will be fuzzified and then come Fuzzy rules out by Fuzzy Systems first, and then generate weighted Fuzzy rules by Genetic Algorithms. Final, the values will be computed with Simple Linear Regression Model. Therefore, the simply regular steps can be described as follows:

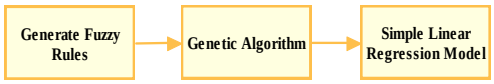

Fig 4:The procedure of Methodology II.

(1) Generate Fuzzy Rules: According to the values provided by the relational database, the values will be fuzzified and then come Fuzzy rules out by Fuzzy Systems. Also, the definition and method have been given in Methodology I.

(2) Generate Weighted Fuzzy Rules by Genetic Algorithm: After Fuzzy Rules has been come out, the values will be tuned better weights of attributes of Fuzzy Rules by GA systems. Also, the process of this step is described as follows:

a) Format of a Chromosome
b) Calculation of the Fitness Degree
c) Selection Operations
d) Crossover Operations
e) Mutation Operations.

Repeat step (c), step (d) and step (e) until the best chromosome has been created.

(3) Calculate the advanced Weighted Fuzzy Rules by Simple Linear Regression Model: An optimal equation or best weights of attributes will be come out by computing with Simple Linear Regression Model in the final step. Also, the definition and method have been given in Methodology I.

As a whole, the estimated accuracy rate of Methodology I is better than that of Methodology II a little bit. However, here presents different thinking to work this research project out by providing these two methodologies.

7. CONCLUSION

In this initial research proposal, two innovative methodologies have been presented for trying to work this research out, integrating Fuzzy Theory, Genetic Algorithms, the method of Least Squares and Simple Linear Regression Model. Also, certain

theories of Computational Intelligence or other disciplines may be adopted in this research at any conditions, and may possible to be applied in certain states in particular. However, there are still many good methods which can be discovered, and which will lead to their unique results. When the methodology works as the base of this research proposal, the new requirements or ideas may be discovered in the future. Basically, the methodology runs recursively until an approximate optimization has been generated, and afterwards the coming advanced or relevant researches may be discovered as the base of this research proposal.

8. REFERENCES

[1] Zadeh, L. A., (1965), Fuzzy Sets, Inform. Control, vol. 8, pp. 338-353.

[2] Wang, L., (1997), A Course in Fuzzy Systems and Control, Upper Saddle River, N.J.: Prentice Hall PTR, pp.7.

[3] Wang, L., (1997), A Course in Fuzzy Systems and Control, Upper Saddle River, N.J.: Prentice Hall PTR, pp.22-29.

[4] Holland, J. H., (1975), Adaptation in Natural and Artificial Systems, MA: MIT Press.

[5] Mitchell, M., (1996), An Introduction to Genetic Algorithms, Cambridge, Mass: MIT Press.

[6] Chen, S., (2003), Generating Weighted Fuzzy Rules From Relational Database Systems for Estimating Missing Values Using Genetic Algorithms, IEEE Transactions on Fuzzy Systems, Vol. 11, No. 4, pp. 495-505.

[7] Illingworth, V., (1996), A Dictionary of Computing, 4th ed., Oxford University Press, pp. 416.

[8] Jong, J. F., (2004), The Simple Linear Regression Model, Academia Sinica, http://www.sinica.edu.tw/~metrics/Pdf_Note/03simple.pdf#search='simple%20linear%20regression%20model', accessed: 01/01/06

[9] Illingworth, V., (1996), A Dictionary of Computing, 4th ed., Oxford University Press, pp.271.

[10] Larson, R. E., Hostetler, P. R. and Edwards, B. H., (1998), Calculus with analytic geometry, 6th ed., Boston : Houghton Mifflin, pp. 896-897.

