International Journal of Computer Information Systems and Industrial Management Applications (IJCISIM)
ISSN: 2150-7988 Vol.1 (2009), pp.249-257
http://www.mirlabs.org/ijcisim

A Dynamic Modular Method for Estimating Null Values in Relational Database Systems

Shin-Jye Lee and Hui-Shin Wang

Abstract— **With the development of the fuzzy system, a lot of sophisticated methods based on the fuzzy system try to do the relational database estimation with a highly accuracy of the approximation by constructing a great diversity of mathematical models. In order to achieve a high-reliability performance with less complication as far as possible, this paper presents a modular method for estimating null values in the relational database system, and which is constructed based on a simple fuzzy learning algorithm. However, there exists a conflict between the degree of the interpretability and the accuracy of the approximation in a general fuzzy system. Thus, how to properly make the best compromise between the accuracy of the approximation and the degree of the interpretability in the entire system is a significant study of the subject. Due to achieve the best compromise practically, the proposed method does not only integrate advantages of fuzzy system and the method of least squares, but also introduce a new criterion, differential rate, to enhance the accuracy of the approximation with a highly accuracy of this achievement.**

Index Terms—**Fuzzy sets, function approximation, relational database estimation.**

I. Introduction

Data Mining, one of popular fields in Computational Intelligence, which mainly extracts knowledge from data, plays an important role in advanced Computational Intelligence. In Data Mining, processing the relational database estimation or function approximation by constructing a great diversity of modular methods based on the automatic fuzzy system has also been emphasized on and researched at any time. In particular, one of the unique features of the fuzzy system is its well-understanding interpretability. To get a highly estimated accuracy of the approximation and the well-understanding interpretability of the fuzzy system together, more complex methods are trying to carry this achievement out by reinforcing the strength of existing methods. However, there still exists a conflict between the interpretability and the accuracy of the approximation in a general fuzzy system. If the degree of the interpretability increases, then the accuracy rate of the approximation decreases. In opposition, if the degree of the interpretability decreases, then the accuracy rate of the approximation increases. Thus, how to make the best compromise between the accuracy of the approximation and the degree of the interpretability in the entire fuzzy system is a significant study of the subject. Furthermore, the predetermined partitioning method before constructing an automatic fuzzy system is also a way of achieving a better performance in the overall system, because most clustering algorithms have always been oriented to solve pattern recognition problems by constructing initial models for the rudiment of function approximation. Due to the reason, the development of the advanced methods and then carry out which as a foundation to capture useful knowledge become increasingly important issues.

This paper proposes a modular method for trying to process high-reliability relational database estimation, and the structure of the proposed method can be composed of three phases, comprising partition determination, automatic fuzzy system generation, and relational database estimation. Basically, in the rudimentary phase, in order to try to attain a well-understanding interpretability in the entire system, the predetermined partitioning method before constructing an automatic fuzzy system is developed based on the concept of output-oriented clustering, and the reason is that it's significantly helpful for the technique of output-oriented clustering to decrease the degree of the dimensionality while doing the partition and then constructing a system. Meanwhile, in the second phase, the main purpose is construct the automatic fuzzy system generating from data in the relational database system according to the resultant clusters generated by the prior phase. Basically, the type of membership function adopted in this phase is triangular membership function and the membership function is shaped by using the calculation of α-cut. Also, in the final phase, processing the relational database estimation is the principal purpose since the fuzzy sets have been generated, and which can be carried out by computing the method of least squares.

Further, there is a new criterion, differential rate, which has been developed for bringing a slightly positive performance in this research work. Basically, the new criterion, differential rate, is used to examine the performance of the current used methods of function approximation by calculating the differential rate between the forecasting model and the original model in the relational database system after the whole procedure of the current used method has been completed. Moreover, the differential rate cannot only be regarded as an examiner checking the performance of the current used methods of function approximation, but also can be applied to enhance the estimated accuracy of the approximation of the testing example in the relational database system. Interestingly, in addition to the above discussions, a subtle relationship between the α-cut and the differential rate has been measured and then discovered in this research work. Therefore, the purpose of the measurement is dedicated to try to discover the optimal rang of the degree of

Shin-Jye Lee is with the School of Computer Science, University of Manchester, Kilburn Building, Oxford Road, Manchester M13 9PL, UK (email: S.Lee@cs.man.ac.uk).

Hui-Shin Wang is with The University of Texas at Austin, Austin, Texas 78712, U.S.

α-cut in particular. By other means, it could be decreasing the uncertainty of the distribution of data set in the relational database system before constructing the membership function at the rudiment of the fuzzy system and furthermore enhancing the reliability of the fuzzy system by optimizing the degree of the sensitivity of the system, in case the optimal range of the degree of α-cut has been definitely discovered.

The rest of this paper is organized as follows. In Section II, the basic concept of fuzzy sets, the application of least squares, the criterion of differential rate, and the relevant components worked in this research work are briefly introduced. In Section III, the sound concept and the detailed procedure of the proposed method are described. In Section IV, we use a variety of simulation examples to illustrate the proposed method and then compare the results of the proposed method with those of other methods. Finally, the paper is concluded in Section V.

II. Preliminaries

In 1965, *Zadeh* proposed the concept of fuzzy sets [1], and fuzzy theory has had therefore been structured and developed gradually. Basically, the theory of fuzzy set is developed based on Fuzzy *IF-THEN* rules with discrete or continuous membership functions, and it can effectively perform on a variety of systems in particular, including linear and non-linear systems.

The well-interpretability is a unique feature of the fuzzy system, and the concept of the membership function is one of essential features making fuzzy system simplicity. As a result of the feature of the simplicity, it's not difficult to understand the detailed information by fuzzy membership function. Basically, each element or data set would be transformed into fuzzy set by calculating with fuzzy membership function in the fuzzy system. Also, each fuzzy set is associated with the membership value (between 0 and 1.0), and the sum of the value of the fuzzy set is 1.0 absolutely. Let U be the universe of discourse, and $\mu_A(x)$ be the membership function of the fuzzy set A. Therefore, the definition of the fuzzy set A is represented as follows [2]:

$$A = \{(x, \mu A(x) \mid x \in U\} \qquad (1)$$

$$A_i(x) = \begin{Bmatrix} 1, x \in A \\ 0, x \notin A \end{Bmatrix} \qquad (2)$$

So far as the type of fuzzy membership function is concerned, there are two types of membership functions, including continuous membership function and discrete membership function respectively. Also, the differentiation between these two types of membership functions depends on the universe of discourse U it has been belonged to. In case the type of the universe of discourse U is continuous, the type of the membership function would be the continuous membership function. On the contrary, if the type of the universe of discourse U is discrete, the type of the membership function would be the discrete membership function. Hence, the definition of the continuous membership function is represented as follows [2]:

$$A = \int_U \mu A(x) / x \qquad (3)$$

Also, the definition of the discrete membership function is represented as follows [2]:

$$A = \sum_U \mu A(x) / x \qquad (4)$$

Moreover, another significant definition of fuzzy membership function is the sum of the value of the fuzzy set is 1.0 absolutely, and which can be defined as follows:

$$Y = f(x) = \sum_{i=1}^{n} Ai(x)Yi = 1 \qquad (5)$$

Based on (5), the concept of complement of fuzzy membership function can be deduced from (5) [2]. In equation (6), $\mu\bar{A}(x)$ is the complement of $\mu A(x)$. In another word, $\bar{A}$ is the complement of the fuzzy set A, and the sum of the value of A and $\bar{A}$ is 1.0 totally.

$$\mu A(x) + \mu \bar{A}(x) = 1 \qquad (6)$$

$$\mu \bar{A}(x) = 1 - \mu A(x) \qquad (7)$$

The relationship between A and $\bar{A}$ in the fuzzy membership function is illustrated in Figure 1. [2]. In Figure 1, the definition of the relationship is $A\ (\mu A(x)) + \bar{A}\ (\mu\bar{A}(x)) = 1.0$ absolutely. When $A\ (\mu A(x))$ increases, then $\bar{A}\ (\mu\bar{A}(x))$ decreases with the synchronization. On the other hand, when $\bar{A}\ (\mu\bar{A}(x))$ increases, then $A\ (\mu A(x))$ decreases with the synchronization. Definitely, the sum of the value of $A\ (\mu A(x))$ and $\bar{A}\ (\mu\bar{A}(x))$ is 1.0 completely. However, there exist certain exceptions as well, when the membership functions irregularly overlap to each other simultaneously. In order to understand how to deal with this kind of situation, the concept of fuzzy normalization will be briefly introduced in the following content.

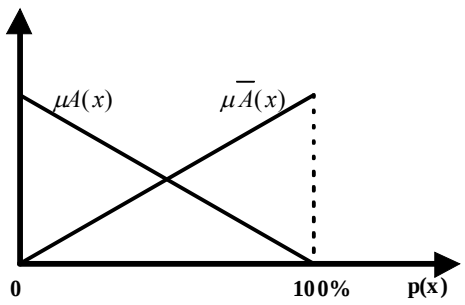

Fig. 1. The relationship between A (μA(x)) and Ā (μĀ(x))

A. Fuzzy Normalization

Basically, not do all membership functions regularly overlap to each other, but there also exist membership functions irregularly overlapping to each other. For example, as illustrated in Figure 2, the sum of the value of $A_1(x_0)$, $A_2(x_0)$ and $A_3(x_0)$ is supposed to exceed the regular degree of the membership function and whose regular value is 1.0 absolutely. In order to solve the problem, the definition of fuzzy normalization can convert the irregular degree of the membership function into the regular degree of the membership function, and an one-input variable comprising multi-membership functions can be defined as follows [14]:

$$\Delta = \sum_{i=1}^{l} \cdots \sum_{k=1}^{n} \left[\frac{A_i(x) \ldots C_k(z)}{\sum_{i=1}^{l} \cdots \sum_{k=1}^{n} A_i(x) \ldots C_k(z)} \right] y_{i \ldots k} \qquad (8)$$

where Δ is the fuzzy set, $A_i(x) \ldots C_k(z)$ presents the degree of the corresponding membership of the fuzzy set respectively, and $y_{i \ldots k}$ presents each tuple of the fuzzy set.

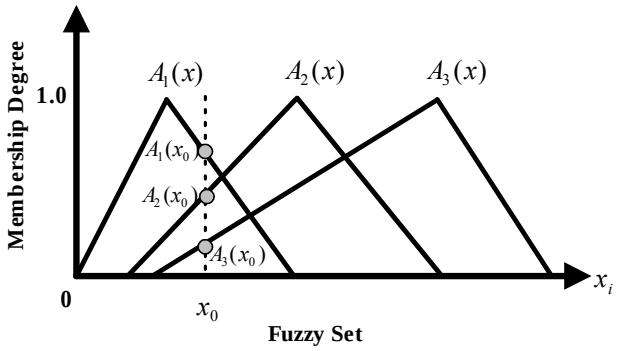

Fig. 2. An example of irregular membership function

B. The α-cut of Fuzzy Set

An α-cut of a fuzzy set *A* is a crisp set that A_α contains all the elements in the universe of discourse *U* that have membership values in *A* greater than or equal to *α*, and it can be defined as follows [2]:

$$A_\alpha = \{x \mid \mu A(x) \geq \alpha, x \in U\} \qquad (9)$$

where $\alpha \in [0,1]$.

Basically, any fuzzy set can be regarded as a family of fuzzy sets, and it can be defined by the notion of α-cuts of fuzzy sets. A complete membership function can be divided into levels by the α-cuts of fuzzy sets. Definitely, the lower the level of *α*, the more elements are admitted to this α-cut. Also, it can be defined as follows [3]:

$$\text{If } \alpha_1 > \alpha_2 \text{ then } A_{\alpha 1} \subset A_{\alpha 2}. \qquad (10)$$

Based on (10), the relationship of fuzzy sets as a family of its α-cut can be illustrated as Figure 3. According to the chart, the relationship of the fuzzy sets of its α-cuts is $\alpha_1 > \alpha_2 > \alpha_3$, so the implicit relationship of the fuzzy sets of its α-cuts is $A_{\alpha 1} \subset A_{\alpha 2} \subset A_{\alpha 3}$.

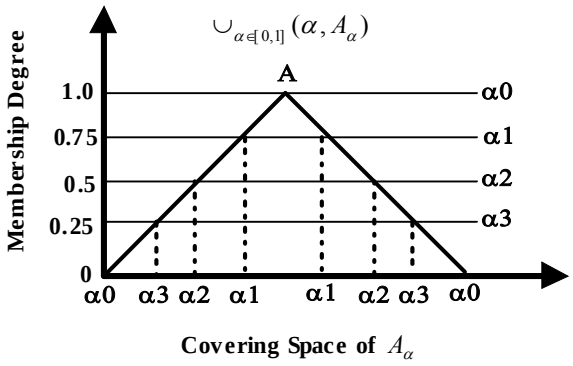

Fig. 3. The relationship of fuzzy set as a family of its α-cuts

C. The Method of Least Squares

The method of least squares is a popular method of estimating parameters in a model by minimizing the sum of squares of differences between observed and theoretical values of a variable [4]. Basically, the purpose of the method of least squares is to make the observed value simplicity and accuracy, and the calculation of the method of least squares can be defined as follows [5]:

$$S = \sum_{i=1}^{n} [f(x_i) - y_i]^2 \qquad (11)$$

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 value of a variable.

D. The Criterion of Differential Rate

One of main contributions of the phase of relational database estimation is that it can effectively optimize the estimated value by processing the advantages of function approximation, such as the application of the method of least squares, making the estimated value more precise with high accuracy of the approximation. Generally, most methods mainly put focus on improving the reliability of the training example, but spend less consideration on reinforcing the feasibility of the testing example. Hence, in order to get the positive performance on the accuracy of the approximation of testing example simultaneously, the criterion of differential rate has therefore been introduced. Also, the differential rate can be definitely defined by calculating the average difference between the original output value of output variable of the training example and the estimated output value of output variable of the training example in the relational database system. Therefore, due to calculate the differential rate

between the original output value of output variable of the training example and the estimated output value of output variable of the training example in the relational database system, the criterion of differential rate D^r can be represented by

$$D^r = Avg\left(\sum_{i=1}^{n}\left(Est.V_i \Big/ Ori.V_i\right)\right) * \omega \qquad (12)$$

where $Est.V_i$ presents the estimated output value of output variable of the training example, $Ori.V_i$ presents the original output value of output variable of the training example, *n* is the number of output value of output variable in the relational database system, and ω presents the weighted value of the approximation.

According to (12), D^r is the differential rate between the original output value of output variable of the training example and the estimated output value of output variable of the training example in the relational database system, and whose range is basically no limitation as a result of the unpredictable performance of the initially estimated accuracy of the approximation of the used methods of function approximation. In principle, the general range of the differential rate D^r is probably located between -1.0 and 1.0 by repeated experiments and which is dependent on the performance of the used methods of function approximation. However, the differential rate D^r cannot be only applied to enhance the estimated accuracy of the approximation of the testing example in the relational database system, but also can be regarded as an examiner checking the performance of the current used methods of function approximation. If the value of D^r is being evaluated more far away from 1.0, it could indicate a higher differential rate existing between the forecasting model and the original model in the relational database system after processing the relational database estimation completely. By other means, the opportunity of a better accuracy of the approximation may be available by improving or reinforcing the current used methods of function approximation. Moreover, the weighted value of the approximation could be properly tuned in case a more precise approximation is being required and available, and the default weighted value of approximation is being set 1.0 initially.

Totally, in addition to examine the reliability of the current used methods by the differential rate D^r, the estimated accuracy of the approximation of the testing example in the relational database system can be slightly improved by properly applying the value of the differential rate D^r as a result of understanding the practically differential rate between the forecasting model and the original model in the relational database system in advance. Further, the application of the differential rate D^r on enhancing the estimated accuracy of the approximation of the testing example will be represented by (23) afterwards.

E. The Advanced Application of the Measurement between Alpha-Cut and Differential Rate

As introduced in the above, the α-cut is applied to adjust the degree of the sensitivity to the system under observation before constructing the fuzzy system, and the differential rate is used to examine the performance of the current used method after the corresponding algorithm of function approximation has been entirely converged and also enhance the estimated accuracy of the approximation of the testing example in the relational database system positively. Interestingly, there is a subtle relationship between α-cut and differential rate D^r, and which may support the positive performance on decreasing the uncertainty of the distribution of data set before constructing the membership function and hence enhancing the reliability of the fuzzy system by discovering the optimal degree of the sensitivity of the system. Therefore, the overall performance of the system can be slightly improved by discovering the optimal rang of the degree of α-cut, and which can be probably obtained by observing the relationship between α-cut and differential rate. In order to discover an optimal degree of the sensitivity of the system by measuring the relationship between α-cut and differential rate, a variety of experiments constructed on the proposed algorithm have been simulated for discovering the optimal range of the degree of α-cut. Meanwhile, based on the basic criterion of differential rate, if the value of the differential rate is more close to the value 1.0, the accuracy of the approximation (MAPE%, MSE, and RMSE) could be higher with the synchronization. On the contrary, if the value of the differential rate is more far away from the value 1.0, the accuracy of the approximation (MAPE%, MSE, and RMSE) could be lower with the synchronization. Therefore, it presents a direct relationship undoubtedly. Generally, each method gets its own specific range of the degree of α-cut, and there is no very precise range of the degree of α-cut for general methods. Basically, the optimal range of the degree of α-cut for the proposed method is around the value 0.7 probably. Furthermore, according to a variety of simulation experiments, including multi-tones non-linear single-input and single-output (MTNLSISO) function, non-linear single-input and single-output function (NLSISO), linear multi-inputs and single-output function (LMISO), non-linear dynamic system, and the real world relational database system, the value of the differential rate can reach the closest position to the value 1.0 whilst the value of α-cut is being set between 0.3 and 0.8 probably.

III. METHODOLOGY

As described in the introduction, the method is structured on a modular methodology, including partition determination, automatic fuzzy system generation and relational database estimation successively. Assume that there is a null value in the relational database system, and the purpose of the proposed method is to process the relational database estimation with a highly accuracy of the approximation by integrating the advantages of the automatic fuzzy system and relational database estimation simultaneously. In the phase of

partition determination, in order to get a well-understanding interpretability in the system by decreasing the uncertainty of the complication, the predetermined partitioning method based on the technique of output-oriented clustering is simply processed by partitioning all the data set in the relational database system equally. In the phase of automatic fuzzy system generation, the membership functions as well as rules are automatically generated by the fuzzy system according to the resultant clusters generated by the prior phase and are not given by experts in advance or pre-defined in the relational database system. In the phase of relational database estimation, this phase puts focus on processing the relational database estimation with the fuzzy sets generated by the prior phase by utilizing the advantage of the method of least squares. Finally, a new criterion, differential rate, is being used to examine the reliability of the proposed method by calculating the differential rate between the original model and the forecasting model in the relational database system after the algorithm has been completely converged, and which can be also used to enhance the estimated accuracy of the approximation of the testing example in the relational database system afterwards. Moreover, the detailed concepts and procedure of the proposed method can be described in the following content

A. Partition Determination

A predetermined partitioning behavior is essentially required before constructing fuzzy membership functions and generating fuzzy rules, because it decides which kind of partition will be used to construct fuzzy membership functions automatically. A proper technique of partition can not only reduce the useless rule, but also can simplify the rule-base to achieve the rule-base optimization as far as possible. In order to make a well-understanding interpretability in the fuzzy system by decreasing the uncertainty of the complication, the partition method based on the technique of output-oriented clustering is simply processed by partitioning all the data set in the relational database system equally, Basically, the simple partition algorithm determines the value of k first, and the value of k means how many partitions will be grouped in the beginning of the whole procedure. After the value of k has been determined, the partition algorithm will start generating k partitions equally based on the output values of the output variable in the relational database system.

1) Determine the value of K first

Interpretability is the unique feature in the fuzzy system, so the determination of the value of K depends on how the interpretability of the fuzzy system is originally required. Moreover, if the value of K increases, then the degree of the interpretability of the fuzzy system would decrease simultaneously as a result of complicated partition; if the value of K decreases, then the degree of the interpretability of the fuzzy system would increase simultaneously as a result of simple partition. In other words, more partitions possess the possibility of making a low degree of the interpretability in the fuzzy system, and fewer partitions possess the possibility of making a high degree of the interpretability in the fuzzy system. By other means, the number of partition initially determines the degree of the interpretability of the fuzzy system and the accuracy of the approximation of the relational database estimation afterwards.

2) Make the partition

Before making the partition, the original data set in the relational database system needs to be sorted in ascending order according to the output value of the output variable in the relational database system. Since the value of k has been determined, all the sorted data set in the relational database system will be partitioned into k partitions equally based on the total number of the sorted data set in the relational database system.

Therefore, the regular steps of the partition method can be stated as follows:

Step 1: Determine the value of k first.

Step 2: Sort the data set P in ascending order based on the output value of the output variable in the relational database system and gets the sorted data set P' in the relational database system.

$$P \xrightarrow{ascending\_order} P' \qquad (13)$$

Step 3: Partition the sorted data set P' in the relational database system into k partitions equally based on the total number of the sorted data set P' in the relational database system.

$$P' = \{(P_1', ..., P_m', ..., P_k') \mid 1 \le m \le k\} \quad (14)$$

where k is the number of the partition, and m presents one of k partitions, where $1 \le m \le k$. Also, P_1' is the first partition in the sorted data set P', P_m' is the mth partition in the sorted data set P', and P_k' is the last partition in the sorted data set P'.

Step 4: Simple output-oriented clustering complete.

B. Generate Fuzzy Membership Functions based on the α-cuts of Fuzzy Sets

To construct the triangular membership function, the calculation of the triplet (left vertex, centroid, right vertex) of the triangular membership function based on the α-cuts of fuzzy sets is one of the significant processes in the phase of automatic fuzzy system generation. Basically, the left vertex, the centroid, and the right vertex of the triplet (left vertex, centroid, right vertex) of each membership function can be simply represented by (a_i, b_i, c_i), and the value of the triplet (a_i, b_i, c_i) of the triangular membership function can be calculated as follows [13]:

$$b_i = \frac{X_{\min} + X_{\max}}{2} \qquad (15)$$

where $X_{\min}$ is the value of the minimum unit or data set of each partition, and $X_{\max}$ is the value of the maximum unit or data set of each partition. According to (15), the value of the

left vertex a_i and that of the right vertex c_i can be calculated by

$$a_i = b_i - \frac{b_i - X_{\min}}{\alpha} \qquad (16)$$

$$c_i = b_i + \frac{X_{\max} - b_i}{\alpha} \qquad (17)$$

where α presents the threshold value of the α-cut of fuzzy set, and $\alpha \in [0,1]$.

According to the value of the triplet (a_i, b_i, c_i) of the triangular membership function, the membership function can be calculated as follows:

$$\mu A_i(x) = \begin{cases} \dfrac{x - a_i}{b_i - a_i}, a_i \le x \le b_i, \\ \dfrac{c_i - x}{c_i - b_i}, b_i \le x \le c_i, \\ 0, otherwise. \end{cases} \qquad (18)$$

C. Generate Fuzzy Rules and Fuzzy Sets

Basically, fuzzy rules can be come out after fuzzy membership functions have been generated, and the total number of fuzzy rules can be defined as follows [8]:

$$\sum_{j=1}^{r} \prod_{i=1}^{n} T_j(X_i) \qquad (19)$$

where *r* is the number of output fuzzy sets of the output variable *Y*, and $T_j(X_i)$ is the number of input fuzzy sets of the input variable X_i.

In the light of the real value of data set provided by the relational database system and the fuzzy membership functions come out by the previous step, the data set will be fuzzified and then come fuzzy set out. Basically, the purpose of this step is to transform the real-valued variable into the regular fuzzy set by calculating with fuzzy membership functions, and the definition of the fuzzy set can be described as follows [14]:

$$\Delta = \sum_{i=1}^{n} \sum_{j=1}^{m} A_j(x_i) Y_j \qquad (20)$$

where Δ means the fuzzy set, $A_j(x_i)$ means the antecedent part of the fuzzy rule and presents the respective degree of the corresponding fuzzy rule of input variable in the relational database system as well, Y_j means the consequent part of the fuzzy rule and presents the corresponding output value of output variable in the relational database system as well, where $1 \le j \le m$, $1 \le i \le n$, *m* is the number of fuzzy rule, and *n* is the number of example in the relational database system.

D. Compute Fuzzy Sets with the Method of Least Squares

After fuzzy sets have been come out by the automatic fuzzy system, the values of fuzzy sets will be processed with the method of lease squares. Basically, the purpose of this step is to do the relational database estimation by processing the fuzzy sets with the method of least squares. As mentioned in (11), the adopted equation evolving from (11) for processing fuzzy sets with the method of least squares can be represented by

$$\theta = (\Phi\Phi^T)^{-1}\Phi^T Y \qquad (21)$$

where θ means the evaluated output variable values of fuzzy sets by the method of least squares, Φ presents the entire input variable values of fuzzy sets, and *Y* presents the original output variable values in the relational database system.

Furthermore, the estimated output value of output variable in the relational database system can be calculated by

$$V = \Phi * \theta \qquad (22)$$

E. Calculate the Differential Rate

It's an additional step in the entire algorithm, and the main purpose of this step does not only examine the reliability of the current used methods by calculating the differential rate between the estimated output value of output variable of the training example and the original output value of output variable of the training example in the relational database system, but also try to optimize the estimated accuracy of the approximation of the testing example in the relational database system simultaneously. Meanwhile, the differential rate between the estimated output value of output variable of the training example and the original output value of output variable of the training example in the relational database system was calculated using (12). Moreover, the differential rate used to apply to enhance the estimated accuracy of the approximation of the testing example in the relational database system can be represented by

$$V_3 = V_2 * D^r \qquad (23)$$

where V_2 means the estimated output value of output variable of the testing example in the relational database system.

Thus, the detailed procedure of the proposed method can be stated as follows:

Step 1: Determine the value of *k* first.

Step 2: Sort the data set *P* in ascending order based on the output value of the output variable in the relational

database system and gets the sorted data set *P'* in the relational database system by using (13).

Step 3: Partition the sorted data set *P'* in the relational database system into *k* partitions equally based on the definition of (14).

Step 4: Simple output-oriented partition complete.

Step 5: Calculate the fuzzy membership function based on the concept of α-cut according to (15), (16), (17), and come fuzzy set out by using (18).

Step 6: Automatic fuzzy system complete.

Step 7: Processing the relational database estimation by computing the fuzzy set with the method of least squares by using (21).

Step 8: The estimated output value of output variable in the relational database system was calculated using (22).

Step 9: Relational database estimation complete.

Additional Step 1: Examine the reliability of the current used methods by using (12).

Additional Step 2: Reinforcing the estimated accuracy of the approximation of the testing example in the relational database system by using (23).

IV. Simulation

In order to understand the reliability of the performance of the proposed algorithm, comparisons of the simulation results between other methods and the proposed method are described in the following content by three classical simulation examples and one real world relational database system example.

A. Non-Linear Single-Input-Single-Output Function Approximation

The samples of simulation can be generated by

$$y = 0.6\sin(\pi x) + 0.3\sin(3\pi x) + 0.1\sin(5\pi x) \qquad (24)$$

where $x \in [-1.1]$, and 100 samples are randomly generated within the definition domain.

The comparison of the accuracy in RMSE between Pedrycz's method [6] and the proposed method are reported in TABLE I. In the comparison, the accuracy in RMSE of the proposed method is better than that of Pedrycz's method, and the proposed method uses fewer clusters and rules/parameters than Pedrycz's method.

B. Two-Dimensional Non-Linear Function Approximation

The samples of simulation can be generated by

$$y = f(x_1, x_2) = (1 + x_1^{-2} + x_2^{-1.5})^2 \qquad (25)$$

where $x_1 \in [1.5]$, $x_2 \in [1.5]$ and 50 samples are randomly generated within the definition domain.

The comparison of the accuracy in MSE between Sugeno and Yasukawa's method [7], Wu and Chen's method [8] and the proposed method are reported in TABLE II. In the comparison, the accuracy in MSE of the proposed method is much better than that of Sugeno and Yasukawa's method, and the parameters of the proposed method are significantly fewer than those of Sugeno and Yasukawa's method. Meanwhile, the accuracy in MSE of the proposed method is much better than that of Wu and Chen's method, and the parameters of the proposed method are fewer than those of Wu and Chen's method.

C. Non-Linear Dynamic System

The samples of simulation can be generated by

$$y(k) = g(y(k-1), y(k-2)) + u(k) \qquad (26)$$

where $g(y(k-1), y(k-2)) = \dfrac{y(k-1)y(k-2)[y(k-1)-0.5]}{1+y^2(k-1)y^2(k-2)}$

$u(k) = \sin\left(\dfrac{2\pi t}{25}\right)$, $y(0) = y(1) = 0$, $t \in [1.200]$, and 200 samples are randomly generated within the definition domain.

The comparison of the accuracy in MSE between others' methods [9] [10] [11] and the proposed method are reported in TABLE III. In the comparison, the proposed method obtains better accuracy in MSE than other methods.

D. A Real World Database

Finally, a second-hand cars database [12] of the real world data set is used for performing the simulation, and each tuple in the relational database can be identified by three input variables (*Style*, *Year* and *C.C.*) and one output variable (*Price*). The comparison of the accuracy in MAPE% between Wu and Chen's method [8] and the proposed method are reported in TABLE IV. In the comparison, the proposed method uses same rules with Wu and Chen's method but obtains better accuracy in MAPE% than Wu and Chen's method.

TABLE I
COMPARISON RESULTS WITH METHODS IN [6] AND THE PROPOSED METHOD

Method	RMSE / Clusters / Parameters		
Pedrycz's Method [6]	0.180/6/18	0.150/8/24	0.147/10/30
	0.144/12/36	0.192/9/27	0.140/12/36
	0.123/15/45	0.114/18/54	0.174/12/36
	0.140/16/48	0.108/20/60	0.100/24/72
	0.149/15/45	0.136/20/60	0.102/25/75
	0.092/30/90	0.141/18/54	0.102/24/72
	0.097/30/90	0.061/36/108	
Proposed Method	0.146/6/18		

TABLE II
COMPARISON RESULTS WITH METHODS IN [7] [8] AND THE PROPOSED METHOD

Method	Neurons (Rules)	Parameters	MSE
Sugeno and Yasukawa [7]	6	65	0.079
Wu and Chen[8]	9	19	0.162
Proposed method	9	18	0.012

TABLE III
COMPARISON RESULTS WITH METHODS IN [9] [10] [11] AND THE PROPOSED METHOD

Method	Number of rules	MSE
GG-TLS [9]	12	3.7E-4
GG-LS [9]	12	3.7E-4
EM-TI [9]	12	2.4E-4
EM-NI [9]	12	3.4E-4
Wang [10]	28	3.3E-4
Wang [11]	20	6.8E-4
Proposed method	27	1.0E-5

TABLE IV
COMPARISON RESULTS WITH METHODS IN [8] AND THE PROPOSED METHOD

Method	Number of rules	MAPE%
Wu and Chen[8]	18	0.13423
Proposed method	18	0.10775

V. CONCLUSION

So far as a perfect modular method developed based on the fuzzy system, it should probably comprise the following features, including the well-understanding interpretability, low-degree dimensionality, highly reliability, highly accuracy of the approximation, less computational cost, and maximum performance. However, it is extremely difficult to meet all of these conditions above. In order to reach the optimal achievement as far as possible, this research work tries to effectively perform the advantages of the proposed method and improve the weaknesses of existing methods practically. In this paper, we have proposed a modular method to estimate null values in the relational database system by constructing automatic fuzzy system, and which integrates advantages of fuzzy system and the technique of function approximation simultaneously. Inasmuch as attaining a high accuracy of the approximation by the existing methods of function approximation are not enough, an innovative criterion, differential rate, based on the development of function approximation is hereupon introduced. The concept of differential rate is a new criterion to make a positive performance on function approximation by calculating the differential rate between the original model and the forecasting model in the relational database system. Due to these developments, the proposed method can effectively achieve a better performance on the relational database estimation with the highly reliability.

ACKNOWLEDGMENT

The author would like to thank Dr. Xiao-Jun Zeng for his thoughtful supervising, and especially appreciate the great courage brought from his girlfriend Ms. Hui-Shin Wang.

REFERENCES

[1] L. A. Zadeh, "Fuzzy sets," *Inform. Control*, vol. 8, pp. 338-353, 1965.
[2] L. X. Wang, *A Course in Fuzzy Systems and Control*, Upper Saddle River, N.J.: Prentice Hall PTR, 1997.
[3] W. Pedrycz, *Fuzzy Sets Engineering*, Boca Raton: CRC Press, 1995.
[4] V. Illingworth, *A Dictionary of Computing*, 4th ed., Oxford University Press, 1996.
[5] R. E. Larson, P. R. Hostetler, and B. H. Edwards, *Calculus with analytic geometry*, 6th ed., Boston: Houghton Mifflin, 1998.
[6] W. Pedrycz, "Linguistic models as a framework of user-centric system modelling," IEEE Transaction on System Man and Cybernetics, Part A, vol. 36, no. 4, pp. 727-745, 2006.
[7] M. Sugeno and T. Yasukawa, "A fuzzy-logic based approach to qualitative modelling," IEEE Transaction on Fuzzy Systems, vol. 1, no.1, pp. 7-31, 1993.
[8] T. P. Wu and S. M. Chen, "A New Method for Constructing Membership Functions and Fuzzy Rules from Training Examples," IEEE Transaction on System Man and Cybernetics, Part B: Cybernetics, vol. 29, no. 1, pp. 25-40, 1999.
[9] J. Abonyi, R. Babuska and F. Szeifert, "Modified Gath-Geva fuzzy clustering for identification of Takagi-Sugeno fuzzy models," IEEE Transaction on Systems, Man and Cybernetics, Part B, vol. 32, no. 5, pp. 612–621, 1999.
[10] L. Wang and J. Yen, "Extracting fuzzy rules for system modeling using a hybrid of genetic algorithms and Kalman filter," Fuzzy Systems and Sets, vol. 101, no.3 , pp. 353–362, 1999.
[11] J. Yen and L. Wang, "Simplifying fuzzy rule-based models using orthogonal transformation methods," IEEE Transaction on Systems, Man and Cybernetics, Part B, vol. 29, no.1, pp. 13–24, 1999.
[12] A Subset of the Collection of the Secondhand Cars. [Online]. Available: http://www.eauto.com.tw/older/servlet/OldCarSearch. (Jan. 1. 2006)
[13] S. J. Lee and X. J. Zeng, "A modular method for estimating null values in relational database systems," *in Proceedings of the Eighth International Conference on Intelligent Systems Design and Applications*, vol. 2, Kaohsiung, Taiwan, Republic of China, 2008, pp.415-419.
[14] S. J. Lee, X. J. Zeng, and H. S. Wang, "Generating automatic fuzzy system from relational database system for estimating null values," *Cybernetics and Syst.*, vol. 40, no. 6, pp. 528-548, 2009.

Author Biographies

Shin-Jye Lee received the BS degree from the Department of Computer Science and Information Management in Providence University, Taiwan, in 1998, and got a MSc(Eng) in Advanced Software Engineering from the Department of Computer Science in the University of Sheffield, U.K., in 2001. He was ever a network engineer and a system engineer successively in Fujitsu, Taipei Branch, Taiwan, and a supportive helpdesk engineer in Microsoft, Taipei Branch, Taiwan, 2002-2005. Currently, he is a PHD candidate at the School of Computer Science in the University of Manchester, U.K. His research interests include learning and identification of fuzzy systems, computational intelligence and computer networking.

Hui-Shin Wang received the BA degree from the Department of Mass Communication in Tamkang University, Taiwan, in 2002, and got a MA in Chinese Literature from the Department of Chinese Literature in the National Chengchi University, Taiwan, in 2004. She was ever a PhD student at the Department of Chinese Language and Literature in Fudan University, China, in 2004, and also was a PhD student at the Department of Educational Psychology in the University of Texas at Austin, U.S., in 2007. She has been working as a senior high school's teacher in Taipei, Taiwan. Currently, she is the only research assistant of Shin-Jye Lee. Her research interests include the modern poems of Chinese literature, educational psychology and human computer interaction.