

TEMPERATURE PREDICTION USING HEURISTIC DATA MINING ON TWO-FACTOR FUZZY TIME-SERIES

Adesh Kumar Pandey¹, Dr. V. K. Srivastava², A.K. Sinha³
^{1,2,3}Krishna Institute of Engineering & Technology, Ghaziabad, U.P, India
adeshpandey@kiet.edu

Dr. S. A. M. Rizvi⁴
⁴Jamia Millia Islamia, New Delhi, India

ABSTRACT

The Data mining technology is widely used in various fields. Temperature forecasting is also the domain area where data mining is used. Accuracy is always an issue in forecasting and it will remain a matter of the concern forever. This paper presents data mining model based on heuristic search, which leads to more accurate solution of forecasting problems. The suggested model is implemented on the problem of temperature forecasting with the two-factor time-variant fuzzy time series algorithm given by Shyi-Ming Chen [1].

Keywords: Data mining, heuristic, temperature prediction, main-factor fuzzy time series, second-factor fuzzy time series, two factor fuzzy time series, and summarization.

1 INTRODUCTION

The current evolution of data mining functions and products are the result of years of influence from many disciplines, including databases, information retrieval, statistics, algorithms and machine learning. The data mining technique is used in various areas like forecasting e-government [2], customer relation management [3], and clinical guidelines [4].

Fuzzy time series is also used by various researchers to achieve the accuracy in forecasting. Song and Chissom [8] successfully implemented the concept of fuzzy sets having linguistic variables presented by Zadeh [11, 12] to overcome the drawback of traditional forecasting methods. Song and Chissom [9, 10] implemented their time invariant and time variant models on the historical time series data of student enrollments of university of Alabama.

Chen and Hwang [1] proposed a two factor time variant fuzzy time series model. Chen and Hwang [1] implemented their algorithm on the 1996 temperature data collected in Taipei. The daily average temperature and daily cloud density are two factors used for the forecasting in two –factors time-variant fuzzy time series. The variation in daily average temperature is used as main–factor fuzzy time series and daily cloud density is used as second-factor fuzzy time series. Chen and Hwang [1] compared the two factor fuzzy time series with

one factor fuzzy time series forecasting method.

The paper proposes a heuristic data mining model, which makes the concept of the data mining more useful for the forecasting problems. The paper also implements the suggested model on 1996 temperature data collected in Taipei, Chen [1], for the prediction of daily average temperature.

2 THE CONCEPTS OF DATA MINING AND FUZZY TIME SERIES

In view of making our study self explanatory, some basic definitions and properties of fuzzy time series, data mining, and heuristic found in [1 – 7] are presented here:

Definition 2.1: Data mining is defined as finding hidden information in a database. Alternatively, it has been called exploratory data analysis, data driven discovery, and deductive learning. The data mining model can be either predictive or descriptive in nature. Some of the most common data mining tasks under each type of model are shown in Fig. 1. A predictive model makes a prediction about values of data using known results found from different data. A descriptive model identifies patterns or relationships in data. Unlike the predictive model, a descriptive model serves as a way to explore the properties of the data examined, not to predict new properties.

Definition 2.2: Summarization maps data into subsets with associated simple descriptions.

Summarization is also called characterization or generalization. It extracts or derives representative information about the database; this may be accomplished by actually retrieving portions of the database. Alternatively, summary type information (such as mean of the numeric attribute) can be derived from the data.

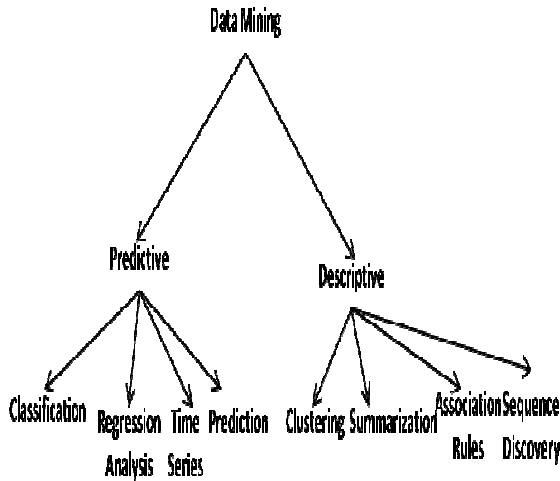


Figure 1: Data mining models and tasks

Definition 2.3: A heuristic is a method that might not always find the best solution but is guaranteed to find the good solution in reasonable time. Generate and test method is the simplest search method. The algorithm for generate and test is as follows:

- a) Generate a possible solution which can either be point in the problem space or a path from the initial state.
- b) Test to see if this possible solution is real solution by comparing the state reached with the set of goal states.
- c) If it is a real solution then return , else repeat step “a”.

Definition 2.4: A fuzzy set is a class of object with a grade of membership. Let U be the universe of discourse with $U = \{\mu_1, \mu_2, \mu_3, \dots, \mu_n\}$ where μ_i are possible linguistic values of U , then a fuzzy set of linguistic variables A_i of U is defined by

$A_i = \mu_{A_i}(\mu_1)/\mu_1 + \mu_{A_i}(\mu_2)/\mu_2 + \dots + \mu_{A_i}(\mu_n)/\mu_n$
 Where μ_{A_i} is the member ship function of the fuzzy set A_i , such that $\mu_{A_i}: U \rightarrow [0, 1]$. If μ_j is the member of A_i , then $\mu_{A_i}(\mu_j)$ is the degree of belonging of μ_j to A_i .

Definition 2.5: Chen [1] proposed a two factor fuzzy time series since any event may be influenced by many factors. Assume that fuzzy time series $F(t)$ and $G(t)$ are the factors of the forecasting problems. If only $F(t)$ is used to solve the forecasting problem,

then it is called a one-factor fuzzy time series. If both $F(t)$ and $G(t)$ are used to solve the forecasting problems, then it is called a two-factor fuzzy time series.

Definition 2.6: Chen [1] proposed new model based on two factor fuzzy time series. Assume that $Y(t)$ ($t = \dots, 0, 1, 2, \dots$) is a subset of R and is the universe of discourse. Let $F(t)$ and $G(t)$ ($t = 1, 2, \dots$) be two factor fuzzy time series on $Y(t)$, where $F(t) = \{\mu_{11}(t), \mu_{12}(t), \dots, \mu_{1n}(t)\}$, $G(t) = \{\mu_{21}(t), \mu_{22}(t), \dots, \mu_{2n}(t)\}$, $\mu_{1i}(t)$ is a fuzzy set on $Y(t)$, $\mu_{2i}(t)$ is a fuzzy set on $Y(t)$, and $1 \leq i \leq n$. Assume that we want to forecast $F(t)$, then $F(t)$ and $G(t)$ are called the main factor fuzzy time series and second factor fuzzy time series of two factor time-variant fuzzy time series model respectively.

Definition 2.7: Window basis size ‘w’ denotes the number of previous observations used to generate the prediction. On the basis of size of window, criterion vector $C(t)$ and the operation matrix $O^w(t)$ at time t are expressed as follows:

$$C(t) = F(t) = [C_1 \ C_2 \ \dots \ C_m] \quad (1)$$

$$O^w(t) = \begin{bmatrix} f(t-1) \\ f(t-2) \\ \vdots \\ f(t-w) \end{bmatrix} = \begin{bmatrix} O_{11} & O_{12} & \dots & O_{1m} \\ O_{21} & O_{22} & \dots & O_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ O_{(w-1)1} & O_{(w-1)2} & \dots & O_{(w-1)m} \end{bmatrix} \quad (2)$$

Where $f(t-1)$ is the fuzzified variation of the main factor fuzzy time series $F(t)$ between time $t-1$ and $t-2$, m is the number of elements in the universe of discourse, C_j and O_{ij} are crisp values, $0 \leq C_j \leq 1$, $0 \leq O_{ij} \leq 1$, $1 \leq i \leq w-1$, and $1 \leq j \leq m$.

Definition 2.8: The second vector used by Chen [1] is $S(t)$, which is composed by the second-factor fuzzy time series $G(t)$. The $S(t)$ is described by Chen [1] as follows:

$$S(t) = g(t-1) = [S_1 \ S_2 \ \dots \ S_m] \quad (3)$$

Where $S(t)$ is the second-factor vector at time t , $g(t-1)$ is the fuzzified data of the second factor fuzzy time series $G(t)$ at time $(t-1)$, m is the number of elements in the universe of discourse, $S_i \in [0, 1]$ and $1 \leq i \leq m$. $R_{ij} = O_{ij} \times S_j \times C_j$, $1 \leq i < w-1$, and “ \times ” is the multiplication operator.

$$R(t) = \begin{bmatrix} O_{11} \times S_1 \times C_1 & O_{12} \times S_2 \times C_2 & \dots & O_{1m} \times S_m \times C_m \\ O_{21} \times S_1 \times C_1 & O_{22} \times S_2 \times C_2 & \dots & O_{2m} \times S_m \times C_m \\ \vdots & \vdots & \ddots & \vdots \\ O_{(w-1)1} \times S_1 \times C_1 & O_{(w-1)2} \times S_2 \times C_2 & \dots & O_{(w-1)m} \times S_m \times C_m \end{bmatrix}$$

$$R(t) = \begin{bmatrix} O_{11} & O_{12} & \dots & O_{1m} \\ O_{21} & O_{22} & \dots & O_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ O_{(w-1)1} & O_{(w-1)2} & \dots & O_{(w-1)m} \end{bmatrix}$$

$$R(t) = \begin{bmatrix} R_{11} & R_{12} & \dots & R_{1m} \\ R_{21} & R_{22} & \dots & R_{2m} \\ \dots & \dots & \dots & \dots \\ R_{(n-1)1} & R_{(n-1)2} & \dots & R_{(n-1)m} \end{bmatrix} \quad (4)$$

Definition 2.9: According to Chen [1] the fuzzified forecasted variation $f(t)$ between time t and time $t-1$ can be described for $R(t)$ as follows:

$$f(t) = \begin{bmatrix} \text{Max}(R_{11}, R_{21}, R_{(n-1)1}) \\ \text{Max}(R_{12}, R_{22}, R_{(n-1)2}) \\ \dots \\ \text{Max}(R_{1m}, R_{2m}, R_{(n-1)m}) \end{bmatrix} \quad (5)$$

Assume $F(t)$ is the main factor fuzzy time series and $G(t)$ is the second factor fuzzy time series, $C(t)$ is the criterion vector on $F(t)$, $O^w(t)$ is an operation matrix on $F(t)$, and $S(t)$ is a second factor vector on $G(t)$. The fuzzy relationship matrix is as follows:

$$R(t) = O^w(t) * S(t) * C(t)$$

Where $f(t)$ is the fuzzified variation of the fuzzy time series $F(t)$ between time t and time $t-1$.

3 HEURISTIC DATA MINING MODEL FOR TEMPERATURE PREDICTION

The proposed data mining model with the application of heuristic search is shown in Fig.2

The proposed Heuristic Model is implemented on the data used by Chen [1], which is given in Table I. The Heuristic Data Mining Model is applied for the temperature prediction in various steps as follows:

Step 1) Data Selection: The target data for the temperature prediction is same as used by Chen [1], given in Table I.

Step 2) Preprocessing: The concept of fuzzy time series is applied for the preprocessing of the target data. The fuzzification of three factors: daily average temperature (factor-1), Variation of daily average temperature (factor-2), and daily cloud density (factor-3) is done during the preprocessing.

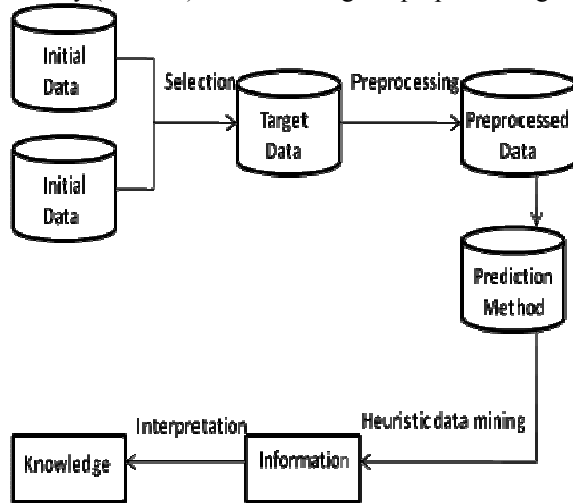


Figure 2: Heuristic Data Mining Model

In the fuzzification of factor-1 the universe of discourse is $U = [\mu_1, \mu_2, \dots, \mu_7]$, where $\mu_1 = [23.00, 24.23]$, $\mu_2 = [24.23, 25.46]$, $\mu_3 = [25.46, 26.69]$, $\mu_4 = [26.69, 27.91]$, $\mu_5 = [27.91, 29.14]$, $\mu_6 = [29.14, 30.37]$, $\mu_7 = [30.37, 31.60]$. We consider seven fuzzy sets which are as follows:

A_1 = (very low temperature), A_2 = (low temperature), A_3 = (average temperature), A_4 = (moderate temperature), A_5 = (high temperature), A_6 = (very high temperature), A_7 = (extremely high temperature). Fuzzy sets A_1, A_2, \dots, A_7 in the universe of discourse U are defined as follows:

$$A_1 = \frac{1}{\mu_1} + \frac{0.5}{\mu_2} + \frac{0}{\mu_3} + \frac{0}{\mu_4} + \frac{0}{\mu_5} + \frac{0}{\mu_6} + \frac{0}{\mu_7}$$

$$A_2 = \frac{0.5}{\mu_1} + \frac{1}{\mu_2} + \frac{0.5}{\mu_3} + \frac{0}{\mu_4} + \frac{0}{\mu_5} + \frac{0}{\mu_6} + \frac{0}{\mu_7}$$

$$A_3 = \frac{0}{\mu_1} + \frac{0.5}{\mu_2} + \frac{1}{\mu_3} + \frac{0.5}{\mu_4} + \frac{0}{\mu_5} + \frac{0}{\mu_6} + \frac{0}{\mu_7}$$

$$A_4 = \frac{0}{\mu_1} + \frac{0}{\mu_2} + \frac{0.5}{\mu_3} + \frac{1}{\mu_4} + \frac{0.5}{\mu_5} + \frac{0}{\mu_6} + \frac{0}{\mu_7}$$

$$A_5 = \frac{0}{\mu_1} + \frac{0}{\mu_2} + \frac{0}{\mu_3} + \frac{0.5}{\mu_4} + \frac{1}{\mu_5} + \frac{0.5}{\mu_6} + \frac{0}{\mu_7}$$

$$A_6 = \frac{0}{\mu_1} + \frac{0}{\mu_2} + \frac{0}{\mu_3} + \frac{0}{\mu_4} + \frac{0.5}{\mu_5} + \frac{1}{\mu_6} + \frac{0.5}{\mu_7}$$

$$A_7 = \frac{0}{\mu_1} + \frac{0}{\mu_2} + \frac{0}{\mu_3} + \frac{0}{\mu_4} + \frac{0}{\mu_5} + \frac{0.5}{\mu_6} + \frac{1}{\mu_7}$$

In case of Chen [1] the fuzzification is done month by month. In the fuzzification of factor-2 the universe of discourse is $U = [\mu_1, \mu_2, \dots, \mu_7]$ as given by Chen [1] for the month of June, where $\mu_1 = [-2.2, -1.6]$, $\mu_2 = [-1.6, -1.0]$, $\mu_3 = [-1.0, -0.4]$, $\mu_4 = [-0.4, 0.2]$, $\mu_5 = [0.2, 0.8]$, $\mu_6 = [0.8, 1.4]$, $\mu_7 = [1.4, 2.0]$.

In this paper the fuzzification has been done for all 122 days of four month. In the fuzzification of factor-2 the universe of discourse is $U = [\mu_1, \mu_2, \dots, \mu_7]$, where $\mu_1 = [-2.41, -1.78]$, $\mu_2 = [-1.78, -1.15]$, $\mu_3 = [-1.15, -0.52]$, $\mu_4 = [-0.52, 0.11]$, $\mu_5 = [0.11, 0.74]$, $\mu_6 = [0.74, 1.37]$, $\mu_7 = [1.37, 2.0]$. Chen [1] consider seven fuzzy sets which are B_1 = (very big decrease), B_2 = (big decrease), B_3 = (decrease), B_4 = (no change), B_5 = (increase), B_6 = (big increase), B_7 = (very big increase), where the fuzzy sets B_1, B_2, \dots, B_7 in the universe of discourse U are defined as follows:

$$B_1 = \frac{1}{\mu_1} + \frac{0.5}{\mu_2} + \frac{0}{\mu_3} + \frac{0}{\mu_4} + \frac{0}{\mu_5} + \frac{0}{\mu_6} + \frac{0}{\mu_7}$$

$$B_2 = \frac{0.5}{\mu_1} + \frac{1}{\mu_2} + \frac{0.5}{\mu_3} + \frac{0}{\mu_4} + \frac{0}{\mu_5} + \frac{0}{\mu_6} + \frac{0}{\mu_7}$$

$$B_3 = \frac{0}{\mu_1} + \frac{0.5}{\mu_2} + \frac{1}{\mu_3} + \frac{0.5}{\mu_4} + \frac{0}{\mu_5} + \frac{0}{\mu_6} + \frac{0}{\mu_7}$$

$$B_4 = \frac{0}{\mu_1} + \frac{0}{\mu_2} + \frac{0.5}{\mu_3} + \frac{1}{\mu_4} + \frac{0.5}{\mu_5} + \frac{0}{\mu_6} + \frac{0}{\mu_7}$$

$$B_5 = \frac{0}{\mu_1} + \frac{0}{\mu_2} + \frac{0}{\mu_3} + \frac{0.5}{\mu_4} + \frac{1}{\mu_5} + \frac{0.5}{\mu_6} + \frac{0}{\mu_7}$$

$$B_6 = \frac{0}{\mu_1} + \frac{0}{\mu_2} + \frac{0}{\mu_3} + \frac{0}{\mu_4} + \frac{0.5}{\mu_5} + \frac{1}{\mu_6} + \frac{0.5}{\mu_7}$$

$$B_7 = \frac{0}{\mu_1} + \frac{0}{\mu_2} + \frac{0}{\mu_3} + \frac{0}{\mu_4} + \frac{0}{\mu_5} + \frac{0.5}{\mu_6} + \frac{1}{\mu_7}$$

In the fuzzification of factor-3 the universe of discourse is $U = [\mu_1, \mu_2, \dots, \mu_7]$. According to Chen [1] $\mu_1 = [90, 100]$, $\mu_2 = [75, 90]$, $\mu_3 = [60, 75]$, $\mu_4 = [45, 60]$, $\mu_5 = [30, 45]$, $\mu_6 = [15, 30]$, $\mu_7 = [0, 15]$. Chen [1] consider seven fuzzy sets which are $C_1 =$ (very very cloudy), $C_2 =$ (very cloudy), $C_3 =$ (more or less cloudy), $C_4 =$ (cloudy), $C_5 =$ (little cloudy), $C_6 =$ (very little cloudy), $C_7 =$ (very very little cloudy) for the second factor fuzzy time series. According to Chen [3], C_1, C_2, \dots, C_7 fuzzy sets are defined on universe of discourse U , $U = \{\mu_1, \mu_2, \dots, \mu_7\}$ as follows:

$$C_1 = \frac{0}{\mu_1} + \frac{0}{\mu_2} + \frac{.5}{\mu_3} + \frac{1}{\mu_4} + \frac{1}{\mu_5} + \frac{1}{\mu_6} + \frac{1}{\mu_7}$$

$$C_2 = \frac{0}{\mu_1} + \frac{1}{\mu_2} + \frac{1}{\mu_3} + \frac{1}{\mu_4} + \frac{1}{\mu_5} + \frac{1}{\mu_6} + \frac{1}{\mu_7}$$

$$C_3 = \frac{.5}{\mu_1} + \frac{1}{\mu_2} + \frac{1}{\mu_3} + \frac{1}{\mu_4} + \frac{1}{\mu_5} + \frac{1}{\mu_6} + \frac{1}{\mu_7}$$

$$C_4 = \frac{1}{\mu_1} + \frac{1}{\mu_2} + \frac{1}{\mu_3} + \frac{1}{\mu_4} + \frac{1}{\mu_5} + \frac{1}{\mu_6} + \frac{1}{\mu_7}$$

$$C_5 = \frac{1}{\mu_1} + \frac{1}{\mu_2} + \frac{1}{\mu_3} + \frac{1}{\mu_4} + \frac{1}{\mu_5} + \frac{1}{\mu_6} + \frac{.5}{\mu_7}$$

$$C_6 = \frac{1}{\mu_1} + \frac{1}{\mu_2} + \frac{1}{\mu_3} + \frac{1}{\mu_4} + \frac{1}{\mu_5} + \frac{.5}{\mu_6} + \frac{0}{\mu_7}$$

$$C_7 = \frac{1}{\mu_1} + \frac{1}{\mu_2} + \frac{1}{\mu_3} + \frac{1}{\mu_4} + \frac{.5}{\mu_5} + \frac{0}{\mu_6} + \frac{0}{\mu_7}$$

Step 3) Prediction Methods:

a) Method-1: In this method factor -1(daily average temperature) is used as main factor fuzzy time series and factor-3 (daily cloud density) as second factor fuzzy time series for the for the temperature prediction by the algorithm given by Chen [1] as follows:

Let window base be 3 and we want to predict temperature of 19 September 1996 by using Method-1. The operation matrix $O^3(t)$, where $t =$ September 19, 1996, is calculated below.:

$$O^3(t) = \begin{bmatrix} \text{fuzzy data of factor - 1 of Sep.17} \\ \text{fuzzy data of factor - 1 of sep.16} \end{bmatrix}$$

$$= \begin{bmatrix} A_3 \\ A_2 \end{bmatrix}$$

$$= \begin{bmatrix} 0 & 0 & 0 & .5 & 1 & .5 & 0 \\ 0 & 0 & 0 & .5 & 1 & .5 & 0 \end{bmatrix}$$

The criterion vector $C(t)$, where $t =$ September 19, 1996, is calculated on the basis of (1) as follows:

$$C(t) = [\text{fuzzy data of factor - 1 of Sep.18}]$$

$$= [A_2]$$

$$= [0 \ 0 \ 0 \ .5 \ 1 \ .5 \ 0]$$

The second factor vector $S(t)$, where $t =$ September 19, 1996), is calculated as per (3) as given below:

$$S(t) = [\text{fuzzy data of factor - 3 of Sep.18}]$$

$$= [C_2]$$

$$= [1 \ 1 \ 1 \ 1 \ 1 \ 1 \ .5]$$

The relation matrix R (September 19, 1996) is calculated according to (4) as shown below:

$$R(t) = \begin{bmatrix} 0 & 0 & 0 & .25 & 1 & .25 & 0 \\ 0 & 0 & 0 & .25 & 1 & .25 & 0 \end{bmatrix}$$

The fuzzified forecasted value $f(t)$, where $t =$ September 19, 1996, is calculated on the basis of (5).

$$f(t) = [0 \ 0 \ 0 \ .25 \ 1 \ .25 \ 0]$$

The defuzzification is done by combined method Chen [1]. The maximum membership value of $f(\text{September 19, 1996})$ is 1 which occurs at u_5 , where the midpoint of u_5 is 28.52. Thus the forecasted temperature of September 19, 1996 is 28.52.

b) Method-2: In method-2, factor -2(the variation of daily average temperature) is used as main factor fuzzy time series and factor-3(daily cloud density) as second factor fuzzy time series for the for the temperature prediction by the algorithm given by Chen [1].

Let window base be 3 and we want to predict temperature of 19 September 1996 by using Chen method. The operation matrix $O^3(t)$, where $t =$ September 19, 1996, is calculated as per (2) as shown below:

$$O^3(t) = \begin{bmatrix} \text{fuzzy data of factor - 2 of Sep.17} \\ \text{fuzzy data of factor - 2 of sep.16} \end{bmatrix}$$

$$= \begin{bmatrix} B_5 \\ B_2 \end{bmatrix}$$

$$O^3(t) = \begin{bmatrix} 0 & 0 & 0 & .5 & 1 & .5 & 0 \\ .5 & 1 & .5 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Again criterion vector $C(t)$, where $t =$ September 19, 1996, is calculated on the bases of (1) as follows:

$$C(t) = [\text{fuzzy data of factor - 2 of Sep.18}]$$

$$= [B_4]$$

$$= [0 \ 0 \ .5 \ 1 \ .5 \ 0 \ 0]$$

The second factor vector $S(t)$, where $t =$ September 19, 1996), is calculated below as per (3):

$$S(t) = [\text{fuzzy data of factor - 3 of Sep.18}]$$

$$= [B_2]$$

$$= [1 \ 1 \ 1 \ 1 \ 1 \ 1 \ .5]$$

The relation matrix $R(t)$, where $t =$ September 19, 1996, is calculated below according to (4) :

$$R(t) = \begin{bmatrix} 0 & 0 & 0 & .5 & .5 & 0 & 0 \\ 0 & 0 & .25 & 0 & 0 & 0 & 0 \end{bmatrix}$$

The fuzzified variation $f(t)$, where $t =$ September 19, 1996, is calculated below on the basis of (5).

$$f(t) = [0 \ 0 \ .25 \ .5 \ .5 \ 0 \ 0]$$

Again defuzzification is done by combined method Chen [1]. The maximum membership value of $f(\text{September 19, 1996})$ is .5 which occurs at u_4 and u_5 , where the midpoint of u_4 is -0.205(m_1) and midpoint of u_5 is 0.425(m_2) . The forecasted variation (mid point of m_1 and m_2) is 0.11. The forecasted data is

equal to the forecasted variation plus the actual data of the last day of the factor-1 fuzzy time series. So the forecasted temperature of September 19, 1996 is equal to $28.10 + 0.11 = 28.21$.

c) Summarization method: we have taken the average of forecasted value by method-1 and method-2. The forecasted value by summarization method comes 28.36. The average forecasting errors and error variance by method-1, method-2, and summarization method at different windows for all four month are shown in Table 2.

In this paper a program in C language is written to implement the algorithm given by Chen [1], which calculates forecasted values and errors on various windows for all four months.

Step 4) Heuristic Data Mining: Now there are three methods of temperature forecasting, method-1, method-2, and summarization method. Now generate and test algorithm is applied as follows:

The acceptable solution (goal state) will be the solution which meets the requirement of data mining task and have the minimum error variance. The solution space has three solutions, method-1, method-2 and Summarization method.

- i. Let initial solution be the method-1. The error variance of this method 4.06 and it comes under the predictive data mining task.
- ii. The method-2 has error variance 7.03 and it also comes under the predictive data mining.
- iii. The summarization method has error variance 3.14 and it comes under summarization task of data mining.
- iv. The summarization is goal state.

Step 5) Information: In this interpretation and analysis of the result has been done as follows:

Table 2 contains the average percentage forecasting error and error variance of predicted temperature by different methods at different windows for the month of June, July, August and September. The average error shows the variation of predicted value and the actual value of temperature. The error variance is measure of dispersion of a set of forecasted values around their mean value. The mathematical expectation of the average squared deviations from the mean. From the Table 2, it is clear that in summarization method the percentage of average forecasting error is always minimum for all four months at every window like 2, 3, 4, 5, 6, 7, and 8. Similarly the error variance is also minimum for the summarization method for all four months at every window. Table 3 contains the error variance of predicted temperature of four months at windows 2, 3, 4, 5, 6, 7, and 8 by method-1, method-2 and summarization method. Again it is minimum for the summarization method. Analysis form Table 2 and Table 3 suggest that summarization method produces better forecasted values.

4 CONCLUSION

In this paper a data mining model based on heuristic search is proposed and the proposed model is used for temperature forecasting. According to the proposed model summarization method is the solution and it produces better and accurate forecasting results of temperature then method-1 and method-2. The paper also suggest that heuristic data mining model produces better results in forecasting problems.

5 REFERENCES

- [1] Shyi-Ming Chen, "Temperature Prediction Using Fuzzy Time Series," IEEE transaction on systems, Man and Cybernetics-Part B: Cybernetics, Vol.30, No. 2, April 2000.
- [2] Yilei Wang , Hui Pan and Tao Li, "The Data Mining of the E-Government on the basis on fuzzy logic," IEEE International Conference on Integration Technology March 20-24,2007 Shenzhen,China.
- [3] Yun Chen, Chuan Fu, Hanhong Zhu, "A Data Mining Approach to Customer Segment Based on Customer Value," Fifth International Conference on Fuzzy Systems and Knowledge Discovery,2008, IEEE Computer Society.
- [4] Reza Sherafat Kazemzadeh and Kamran Sartipi, "Incorporating Data Mining Application into Clinical Guildelines,"19th IEEE Symposium on Computer-Based Medical System (CBMS'06).
- [5] Margaret H. Dunham, "Data Mining Introductory and Advanced Topics," Pearson Education,Page No. 3,4,5, and 8.
- [6] Elaine Rich and Kevin Knight, "Artificial Intelligence, Second Edition," Tata McGraw-Hill Edition, Page No. 41, and 64.
- [7] Gongemi yu and Zvi Schwartz, "Forecasting Short Time-Series Tourism Demand with Artificial Intelligence Models," Journal of Travel Research, Vol. 45 , November 2006.
- [8] Q. Sang, B. Chissom, Fuzzy Time series and its Models, Fuzzy sets and System 54 (1993) 269-277
- [9] Q. Sang, B. Chissom, Forecasting enrollments with fuzzy time series – Part I, Fuzzy sets and systems 54 (1993)
- [10] Q. Sang, B. Chissom, Forecasting enrollments with fuzzy time series – Part III, Fuzzy sets and systems 64 (1994)
- [11] L.A. Zadeh, Fuzzy set, Fuzzy Set Information and control 8 (1965)338-353
- [12] L.A. Zadeh, The concept of a linguistic variable and its application to approximate reasoning. Part I. Information Science 8 (1975)199-249

TABLE 1: Historical data of the daily average temperature and daily cloud density (Unit: %) form June 1996 to September 1996 in Taipei (Unit: °C) [2]

Day / Month	June			July			August			September		
	Temp. (Unit:°C)	Variation of Temp. (Unit:°C)	Cloud Density	Temp. (Unit:°C)	Variation of Temp. (Unit:°C)	Cloud Density	Temp. (Unit:°C)	Variation of Temp. (Unit:°C)	Cloud Density	Temp. (Unit:°C)	Variation of Temp. (Unit:°C)	Cloud Density
1	26.1		36	29.9	-0.3	15	27.1	0.2	100	27.5	-0.2	29
2	27.6	1.5	23	28.4	-1.5	30	28.9	1.8	78	26.8	-0.7	53
3	29	1.4	23	29.2	0.8	26	28.9	0	68	26.4	-0.4	66
4	30.5	1.5	10	29.4	0.2	34	29.3	0.4	44	27.5	1.1	50
5	30	-0.5	13	29.9	0.5	24	28.8	-0.5	56	26.6	-0.9	53
6	29.5	-0.5	30	29.6	-0.3	28	28.7	-0.1	89	28.2	1.6	63
7	29.7	0.2	45	30.1	0.5	50	29	0.3	71	29.2	1	36
8	29.4	-0.3	35	29.3	-0.8	34	28.2	-0.8	28	29	-0.2	76
9	28.8	-0.6	26	28.1	-1.2	15	27	-1.2	70	30.3	1.3	55
10	29.4	0.6	21	28.9	0.8	8	28.3	1.3	44	29.9	-0.4	31
11	29.3	-0.1	43	28.4	-0.5	36	28.9	0.6	48	29.9	0	31
12	28.5	-0.8	40	29.6	1.2	13	28.1	-0.8	76	30.5	0.6	25
13	28.7	0.2	30	27.8	-1.8	26	29.9	1.8	50	30.2	-0.3	14
14	27.5	-1.2	29	29.1	1.3	44	27.6	-2.3	84	30.3	0.1	45
15	29.5	2	30	27.7	-1.4	25	26.8	-0.8	69	29.5	-0.8	38
16	28.8	-0.7	46	28.1	0.4	24	27.6	0.8	78	28.3	-1.2	24
17	29	0.2	55	28.7	0.6	26	27.9	0.3	39	28.6	0.3	19
18	30.3	1.3	19	29.9	1.2	25	29	1.1	20	28.1	-0.5	39
19	30.2	-0.1	15	30.8	0.9	21	29.2	0.2	24	28.4	0.3	14
20	30.9	0.7	56	31.6	0.8	35	29.8	0.6	25	28.3	-0.1	3
21	30.8	-0.1	60	31.4	-0.2	29	29.6	-0.2	19	26.4	-1.9	38
22	28.7	-2.1	96	31.3	-0.1	48	29.3	-0.3	46	25.7	-0.7	70
23	27.8	-0.9	63	31.3	0	53	28	-1.3	41	25	-0.7	71
24	27.4	-0.4	28	31.3	0	44	28.3	0.3	34	27	2	70
25	27.7	0.3	14	28.9	-2.4	100	28.6	0.3	29	25.8	-1.2	40
26	27.1	-0.6	25	28	-0.9	100	28.7	0.1	31	26.4	0.6	30
27	28.4	1.3	29	28.6	0.6	91	29	0.3	41	25.6	-0.8	34
28	27.8	-0.6	55	28	-0.6	84	27.7	-1.3	14	24.2	-1.4	59
29	29	1.2	29	29.3	1.3	38	26.2	-1.5	28	23.3	-0.9	83
30	30.2	1.2	19	27.9	-1.4	46	26	-0.2	33	23.5	0.2	38
31				26.9	-1	95	27.7	1.7	26			

TABLE 2: Error variance & Average Percentage error of temperature prediction at different window by applying three different methods

Window	Month	Method-1		Method-2		Summarization Method	
		Error Variance	%Average Error	Error Variance	%Average Error	Error Variance	%Average Error
2	June	5.84	2.91	4.49	3.09	3.59	2.16
	July	4.10	2.48	4.90	3.33	2.76	1.91
	August	4.19	3.03	6.37	3.72	2.75	2.30
	September	6.22	3.00	5.94	3.17	3.56	2.03
3	June	4.34	2.72	6.05	3.45	3.01	2.01
	July	3.68	2.56	8.19	3.66	2.99	1.97
	August	4.04	2.86	8.71	3.75	2.55	2.14
	September	6.00	3.15	6.95	3.15	3.43	1.85
4	June	4.50	2.71	6.05	3.47	3.09	2.03
	July	3.49	2.63	7.64	3.78	2.83	2.04
	August	4.04	2.86	8.76	3.69	2.54	2.16
	September	6.20	3.07	6.47	3.32	3.88	2.01
5	June	4.44	2.87	5.98	3.53	3.19	2.16
	July	3.98	2.70	8.73	3.85	3.43	2.10
	August	4.04	2.86	8.33	3.69	2.50	2.19
	September	6.59	3.23	6.52	3.68	3.94	2.19
6	June	4.00	3.08	5.99	3.68	3.00	2.34
	July	4.39	2.77	8.35	3.93	3.84	2.21
	August	4.07	2.58	8.48	3.62	2.39	1.93
	September	6.20	3.14	7.07	3.88	4.13	2.26
7	June	3.76	3.26	6.08	3.76	2.98	2.47
	July	4.18	2.71	8.75	3.97	3.91	2.18
	August	4.23	2.58	8.35	3.65	2.47	2.02
	September	6.20	3.14	6.85	3.87	4.03	2.29
8	June	3.95	3.26	5.75	3.93	2.89	2.58
	July	4.15	2.70	8.57	3.93	3.84	2.17
	August	4.23	2.58	8.35	3.65	2.47	2.02
	September	6.20	3.14	6.87	3.91	4.03	2.29

TABLE 3: The error variance of forecasted values for all four months

Method	Method-1	Method-2	Summarization Method
Error variance	4.60	7.03	3.14