Implementation of Fuzzy Inference System for Classification of Dengue Fever on the villages in Malang

Samigun Handoyo\textsuperscript{1*} and Heni Kusdarwati\textsuperscript{1}

\textsuperscript{1}Department of Statistics, Faculty of Mathematics and Natural Science, Brawijaya University Malang

*Corresponding author: samistat@ub.ac.id

Abstract. Dengue fever is a disease that must be watched out and early preventive measures are taken so the spread of this disease can be reduced. An early preventive dengue fever can be controlled by a mathematical model. The article proposes a model of fuzzy inference system with optimal fuzzy rule bases generated by fuzzy c-means and optimized by ordinary least square (OLS). The developing system is done by forming a data structure in the form of input-output pairs. Factors that influence the number of dengue fever cases are used as the system input and the number of dengue fever cases as the system output. Based on the input-output pairs, the fuzzy rule bases is generated by using the fuzzy c-means method. The consequent part of the rule bases is optimized by the OLS method to produce the optimal rule bases. The resulted system is able to predict the level of dengue fever in the villages with an accuracy of 90% and can be used to predict the level of dengue fever in a village by inputting factors that influence the level of dengue fever.

Keywords: classification, dengue fever, fuzzy inference system, prediction, rule bases.

1. Introduction

The development of models that are able to classify objects accurately or be able to predict future values correctly is the main topic and is still an open problem for researchers in both the fields of statistics and the field of computer science. The development of a hybrid model for the purpose of object classification by comparing statistical and neural network (NN) methods, among others, is carried out by Widodo and Handoyo [1] applying Logistic regression and Support Vector Machine, while Nugroho et al. [2] applied logistic regression and Learning Vector Quantization. On the other hand, Kusdarwati and Handoyo [3] combined discrete wavelet transform as a preprocessing method on the input layer of the Radial Base function Neural Network that is used for non-stationary time series predictions [3]. The NN-based modeling above is very sensitive to the setting of NN parameter values (epoch numbers, learning rates, and mean squared errors) which are both subjective and sensitive.

Fuzzy Inference System (FIS) modeling is very dependent on the Fuzzy Rules Bases (FRB) that originally came from experts. The availability of abundant data have led to the compilation of the FRB based on the patterns in the data, so the role of experts in making the FRB can be replaced. Initially the generating FRB based on pairs of input-output data was carried out by Wang and Mendel [4] known as the lookup table scheme method. Handoyo and Marji [5] carried out Ordinary Least Square (OLS) optimization to determine the consequent part of the FRB generated by the Wang and Mendel method.
for time series predictions. The number of rules on the FRB produced by the Wang and Mendel method are quite many so it takes a long processing time. The formation of the FRB with the clustering method will result the number of rules equal to the number of clusters. Handoyo et al. [6] applied the Fuzzy C-means (FCM) method to generate a FRB on datasets that are causal in nature to predict the regional minimum wage in East Java province. The use of the FCM method produces an efficient rule bases, but the use of the Gaussian membership function on the rule bases is only the optimal in mean parameter but the spread parameter is deterministic (not optimal).

In the study, the Fuzzy Rule Bases (FRB) will be generated using FCM with the optimal parameters of the Gaussian membership function, i.e. the spread parameter is calculated based on partition matrix elements optimized with OLS. The optimal FRB obtained will be applied to the Fuzzy Inference System (FIS) to classify dengue fever status of the areas or villages in Malang.

2. Research Method

The goal of FIS is to map values in the input space into values in the output space through 4 components namely Fuzzification, Fuzzy Rule Bases (FRB), Inference Engine, and Defuzzification. The most important FIS component is the FRB which originally comes from an expert or operator who has rich experience in his field. The fuzzification component has the task of changing crisp values into linguistic values and degrees of membership. Inference engine component is method for making inference based on FRB, the result of the inference which is still in the form of fuzzy set is converted into crisp value through the defuzzification component [7].

According to Handoyo and Marji [5], FIS output was generated through an inference process divided into two parts. The first one is to calculate the activation level of each rule in FRB proceeded by aggregating the output of each rules. The second part is carried out defuzzification process of the aggregation result to be transformed into crisp values. Mathematically, it can be formulated as follows:

\[ Rule^l : \]

\[ IF \ (x_{t1} \ is \ A^1_1) \ AND \ (x_{t2} \ is \ A^1_2) \ ... \ AND \ (x_{tk} is \ A^1_k) \ THEN \]

\[ y^l_t = \theta^l_0 + \theta^l_1 x_{t1} + \cdots + \theta^l_k x_{tk} \]  \hspace{1cm} (1)

where \( x_{tk} \) is the value of the \( k \)-th input variable, \( y^l_t \) is the local output (output generated by the \( l \)-th rule), \( \theta^l_0, \theta^l_1, \ldots, \theta^l_k \) are the consequent parameters of rule \( l \), and \( A^l_k \) is the linguistic value (fuzzy set) of the \( k \)-th input variable for the first rule represented by a membership function \( \mu_{A^l_k}(x_{tk}) \).

In the research, the Gaussian membership function (GaussMF) is used in the fuzzification process which is expressed as follows:

\[ \mu_{A^l_k}(x_{tk}) = \exp \left( -\frac{1}{2} \left( \frac{x_{tk} - c^l_k}{\sigma^l_k} \right)^2 \right) \]  \hspace{1cm} (2)

where \( c^l_k \) and \( \sigma^l_k \) are respectively the parameters of the center (mean) and spread (standard deviation) of gaussMF. The means parameter from GaussMF is obtained by Fuzzy c-means method, while spread parameters are calculated by OLS optimization. According to Bezdek et al. [8], the FCM algorithm was summarized as follows:

1. Set input data in the matrix data structure
2. Specific some parameters needed in the application of FCM method i.e. number of clusters, weight factor, maximum iteration, and stopping condition.
3. Generate the initial elements of partition matrix \( U_0(\mu_{0ik}) \) for \( i = 1, 2, \ldots, c \) and \( k = 1, 2, \ldots, n \). Based on the initial partition matrix, then compute the degree of membership of each object in the each cluster uses the formula follows:

\[ \mu_{ik} = \frac{\mu_{0ik}}{Q_i} , \text{ where } Q_i = \sum_{j=1}^{n} \mu_{0ik} \]  \hspace{1cm} (3)

4. Calculate the center of the first cluster \( (v_{1i}) \) with the formula:

\[ v_{ij} = \frac{\sum_{k=1}^{n} (\mu_{ik})^w \times x_{ki}}{\sum_{k=1}^{n} (\mu_{ik})^w} \]  \hspace{1cm} (4)
5. Calculate the value of the objective function \( P_t \) at the \( t \)-th iteration by using equation follows:

\[
P_t = \sum_{k=1}^{n} \sum_{c=1}^{c} (\mu_{ik})^2 D(x_k, v_i)^2
\]

(5)

6. Compute the new partition matrix using the following formula:

\[
\mu_{ik} = \frac{\sum_{j=1}^{m} (x_{kj} - v_{ij})^2}{\sum_{i=1}^{m} \sum_{j=1}^{m} (x_{kj} - v_{ij})^2}
\]

\( \xi \)

(6)

7. Stop the iteration if \(|P_t - P_{t-1}| < \xi \) or \( t > t_{\text{max}} \), in other conditions, iteration is continued with \( t = t + 1 \) and repeat from step 4.

The main output of FCM are the V matrix which the elements are the center of the clusters and the partition matrix U, which the elements are the degree of membership of each record in each cluster. The luster center will be used as the mean parameter of Gaussian membership function (GaussianMF), while the partition matrix is used as the basis for calculating the spread parameters of GaussianMF.

The following will be explained how to calculate the spread parameters of the proposed method. Assume that each element of the partition matrix equals to the membership degree of GaussMF where expressed in equation (7) as follows:

\[
\mu_{ht} = \exp \left( -\frac{1}{2} \left( \frac{x - c}{\sigma} \right)^2 \right)
\]

(7)

If there are as many as \( k \) attributes in each object, then the equation (8) can be expanded to become as follows:

\[
\mu_{ht} = \prod_{i=1}^{k} \exp \left( -\frac{1}{2} \left( \frac{x_i - c_i}{\sigma_i} \right)^2 \right)
\]

(8)

We multiply the natural logarithm to both sides of equation (9) and substitute of

\[
2 \ln(\mu_{ht}) = y, -\frac{1}{\sigma_i} = b_i \text{ and } (x_i - c_i)^2 = x_i, \text{ we get the equation } (10) \text{ as follows}
\]

\[
y = b_1x_1 + b_2x_2 + \cdots + b_kx_k
\]

(9)

If there are observed values of \( i = 1, 2, \ldots, n \), finally the equation (10) can be expressed in the matrix notation as follows:

\[
\begin{bmatrix}
Y_1 \\
Y_2 \\
\vdots \\
Y_n
\end{bmatrix} =
\begin{bmatrix}
x_{11} & x_{12} & \cdots & x_{1k} \\
x_{21} & x_{22} & \cdots & x_{2k} \\
\vdots & \vdots & \ddots & \vdots \\
x_{n1} & x_{n2} & \cdots & x_{nk}
\end{bmatrix}
\begin{bmatrix}
b_1 \\
b_2 \\
\vdots \\
b_k
\end{bmatrix}
+ \begin{bmatrix}
\xi_1 \\
\xi_2 \\
\vdots \\
\xi_n
\end{bmatrix}
\]

(10)

\[
Y = X \theta + \varepsilon \text{ so that, } \varepsilon = Y - X \theta
\]

(11)

By using OLS optimization which minimizes sum square of error through first derivative we obtained the result that \( \theta = (X^TX)^{-1}X^TY \) where referring to equation (11), furthermore there is yeilded \( \theta = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_k \end{bmatrix} \), so that be obtained \( \sigma_i^2 = -\frac{1}{b_i} \). The optimal spread of Gaussian membership function is obtained as

\[
\text{spread} = \text{sqrt} (abs \left( -\frac{1}{b_i} \right)).
\]

The system output is \( \hat{y}_k \) (the value of crips) which is the predicted value on the \( t \)-observation calculated by using equation (12) as follows:

\[
\hat{y}_i = \sum_{i=1}^{l} \hat{a}_i y_i = \sum_{i=1}^{l} \hat{a}_i (\theta_0^i + \theta_1^i x_{i,1} + \cdots + \theta_p^i x_{i,p})
\]

And
\[ \bar{a}_t \equiv \frac{a_t^l}{\sum_{t=1}^L a_t} = \frac{\prod_{k=1}^p \mu_{A_k}(x_{t,k})}{\sum_{t=1}^L \prod_{k=1}^p \mu_{A_k}(x_{t,k})} \]

The equation (12) is a defuzzification process with weighted average method. The weights used in the Takagi-Sugeno model are normalized fire strength \( \bar{a}_t \), consequent part \( y_t \) is a constant number for FIS Takagi-Sugeno of zero order, or in the form of a linear equation for FIS Takagi-Sugeno of one order [6].

3. Result and Discussion

The effect of linearly increasing deterministic trends and holiday variations are made in the model of linear deterministic trend holiday variations.

Factors that influence the number of dengue fever cases are used as the system input and the number of dengue fever cases as the dependent variable that will categorized in binary class.

| Variables                                      | Max. | Min. | average |
|------------------------------------------------|------|------|---------|
| Number of dengue fever cases as Y              | 13   | 0    | 2.72    |
| Flick Free Numbers (%) as X1                   | 87.38| 80.87| 84.13   |
| Number of dengue fever cases in the previous year as X2 | 33   | 0    | 7.18    |
| Average Rainfall (mm) as X3                    | 293.54| 215.12| 241.34  |
| Average Number of Rainy Days (day) as X4       | 13   | 11   | 12.04   |
| Population density (person/km\(^2\)) as X5     | 33120| 1018 | 11288   |

The observation of 56 villages in the city of Malang to determine the village class or category against dengue fever outbreaks. Based on the number of cases of dengue fever, a village is classified as a village prone to dengue fever and villages that are not prone to dengue fever. Distribution of datasets into two parts training and testing with three composition of training and testing pairs, namely 60: 40, 70: 30, and 80: 20 will be used as data to build models and data to validate the model.

In this study many clusters are determined to be 5, so the number of rules obtained is also the same as 5 where this value is seen as a moderate number of rules. The following table 2 gives the results of the cluster center from grouping the data with a proportion of 70%.

| Cluster | X1     | X2     | X3     | X4     | X5     |
|---------|--------|--------|--------|--------|--------|
| 1       | 83.2143| 3.467497| 229.9797| 11.95024| 3742.655|
| 2       | 84.6842| 7.427957| 229.3838| 12.23011| 21110.88|
| 3       | 82.80766| 4.700519| 223.0097| 12.30902| 32100.81|
| 4       | 84.93505| 11.16192| 242.9167| 12.03566| 15034.2|
| 5       | 84.527 | 7.540302| 254.2247| 12.09273| 8599.932|

In the Table 2, columns 2 to 6 are cluster centers of the variables 1 to 5. It appears that there are 5 values for each variable, which means each variable is divided into 5 linguistic values determined by ranking each column. The values in each column will be in the form of a linguistic value of 1 to 5 stating that 1 = Small 2 (S2), 2 = Small 1 (S1), 3 = Normal (N), 4 = Big 1 (B1), and 5 = big 2 (B2). Table 3 states the results of ranking each variable.
Based on linguistic values in table 3, finally, it can be compiled the Fuzzy Rule Bases (FRB) as follows:

\textbf{Rule1.} IF (X1 is S1 and X2 is S2 and X3 is N and X4 is S2 and X5 is B2 ) THEN (Y_t is C1)

\textbf{Rule2.} IF (X1 is B1 and X2 is N and X3 is S1 and X4 is B1 and X5 is N ) THEN (Y_t is C2)

\textbf{Rule3.} IF (X1 is S2 and X2 is S1 and X3 is S2 and X4 is B2 and X5 is B1 ) THEN (Y_t is C3)

\textbf{Rule4.} IF (X1 is B2 and X2 is B2 and X3 is B1 and X4 is S1 and X5 is S1 ) THEN (Y_t is C4)

\textbf{Rule5.} IF (X1 is N and X2 is B1 and X3 is B2 and X4 is N and X5 is S2 ) THEN (Y_t is C5)

We can interpret the FRB above, for example the 1st rule is

a. Antecedence part:

\begin{itemize}
  \item X1 input is Big
  \item input X2 is Small2, and input X3 is Normal, and input X4 is Small2, and input X5 is Big2
\end{itemize}

b. Consequence part:

\begin{itemize}
  \item output Y is Constant 1.
\end{itemize}

The constant value in the consequent part of each rule is determined in a certain way, in this case using Ordinary Least squared optimization as in Handoyo and Marji (2018) [5]. The first rule form and membership functions can be stated as follows:

\[
IF \left( \text{is } \mu_{S1} = \exp \left( -\frac{1}{2}\frac{(x - cs1)^2}{sgs1} \right) \right) \text{ and } \mu_{S2} = \exp \left( -\frac{1}{2}\frac{(x - cs2)^2}{sgs2} \right) \text{ and } \mu_{N} = \exp \left( -\frac{1}{2}\frac{(x - cn)^2}{sgn} \right) \text{ and } \mu_{B2} = \exp \left( -\frac{1}{2}\frac{(x - cB2)^2}{sgB2} \right) \right)
\]

On each rule there is a Gaussian membership function whose denominator part of the membership function is the spread parameter of the Gaussian membership function. The spread parameter values for each linguistic value of each rule are determined by OLS based on the elements of the partition matrix. The following is the value of the degree of membership in the partition matrix for the first 6 records.

\begin{table}[h]
\centering
\caption{The Linguistic value resulted of ranking each variable}
\begin{tabular}{llllll}
Number & X1 & X2 & X3 & X4 & X5 \\
\hline
1 & 2 & 1 & 3 & 1 & 5  \\
2 & 4 & 3 & 2 & 4 & 3  \\
3 & 1 & 2 & 1 & 5 & 4  \\
4 & 5 & 5 & 4 & 2 & 2  \\
5 & 3 & 4 & 5 & 3 & 1  \\
\end{tabular}
\end{table}

\begin{table}[h]
\centering
\caption{Value of membership on each cluster of the first 6 records}
\begin{tabular}{lllllll}
Record & cls1 & cls2 & cls3 & cls4 & cls5 \\
\hline
1 & 0.014959 & 0.139546 & 0.010977 & 0.796683 & 0.037835  \\
2 & 0.05334 & 0.054088 & 0.010497 & 0.609816 & 0.272259  \\
3 & 0.046085 & 0.015075 & 0.003789 & 0.074439 & 0.860611  \\
4 & 0.002164 & 0.971865 & 0.007007 & 0.014942 & 0.004022  \\
5 & 0.005302 & 0.029614 & 0.003026 & 0.947429 & 0.014629  \\
6 & 0.019648 & 0.269445 & 0.017044 & 0.646607 & 0.047255  \\
\end{tabular}
\end{table}
The results of the FIS implementation in the proportion of 70%: 30% for training data: testing data can be presented visually where plots between actual values and predicted values both in training data and testing data are given in Figure 1 as follows:

![Plot Training Data](image1)

![Plot Testing Data](image2)

**Figure 1.** Plot Actual versus Predicted on proportion training of 70% and proportion testing of 30%.

The plot of the training data from 45 villages shows that each village can be predicted or classified correctly, that is, the target class is the same as the prediction class, whereas in the data testing there is one village that is wrongly predicted, namely the village in the 14th record. This event actually also occurs in the proportion of other training and testing data, both at the proportions of 60:40 and 80:20, the model successfully predicts with 100% accuracy in the training data and finds one record that is incorrectly classified in the testing data. The performance of the FIS model on data testing is presented in table 5 as follows:

| Proportion between Training data and Testing data | APER | Accuracy |
|--------------------------------------------------|------|----------|
| 60%:40%                                          | 8.70%| 91.30%   |
| 70%:30%                                          | 11.76%| 88.24%   |
| 80%:20%                                          | 16.67%| 83.33%   |

Based on table 5, it can be seen that the FIS performance on testing data varies, but this difference is actually due to the relatively small number of observations, which are only 56 villages. The fact is that in all proportions of data testing (40%, 30%, and 20%) there is only one village that is incorrectly classified. The results of the above performance cannot make generalizations that the smaller the proportion of data testing, the lower the accuracy will be obtained. But in essence the FIS model developed in this study has a high performance that can achieve 91% accuracy.

**4. Conclusion**

In the paper the FIS model has been successfully developed which has a number of rules in FRB of 5 rules, that is generation results using fuzzy c–means clustering, which is carried out by OLS optimization in both the part of rules consequent and spread parameters of the Gaussian membership function. The resulting FIS model has 100% accuracy in predicting training data, while FIS performance in testing data can reach 91%.

**References**

[1] Widodo, A., and Handoyo, S. (2017). The Classification Performance Using Logistic Regression And Support Vector Machine (Svm). *Journal Of Theoretical & Applied Information Technology*. Vol. 95(No. 19) pp. 5184-5193.
[2] Nugroho, W. H., Handoyo, S., and Akri, Y. J. (2018). An Influence of Measurement Scale of Predictor Variable on Logistic Regression Modeling and Learning Vector Quantization Modeling for Object Classification. *International Journal of Electrical and Computer Engineering (IJECE)*, Vol. 8(No. 1) pp. 333-343.

[3] Kusdarwati, H., and Handoyo, S. (2018). System for Prediction of Non Stationary Time Series based on the Wavelet Radial Bases Function Neural Network Model. *Int J Elec & Comp Eng (IJECE)*. Vol. 8(No. 4) pp. 2327-2337

[4] Wang, L.X., and Mendel, J.M. (1992). Generating fuzzy rules by learning from examples. *IEEE Transactions on systems, man. and cybernetics*. Vol. 22(No. 6) p. 1414-1427.

[5] Handoyo, S., and Marji. (2018). The Fuzzy Inference System with Least Square Optimization for Time Series Forecasting. *Indonesian J Elec Eng & Comp Science (IJEECS)*. Vol.11(No. 3) pp. 1015-1026.

[6] Handoyo, S., Purwanto, I.N. and Jie, F. (2018). The Fuzzy Inference System with Rule Bases Generated by using the Fuzzy C-Means to Predict Regional Minimum Wage in Indonesia. *International J. of Opers. and Quant. Management (IJOQM)*. Vol. 24(No. 4) pp. 277-292.

[7] Handoyo, S., and Prasojo, A.P.S. (2017). Applied Fuzzy System with R Software (Sistem Fuzzy Terapan dengan Software R). UB Press. Malang.

[8] Bezdek, J.C., Ehrlich, R., and Full, W. (1984). FCM: The fuzzy c-means clustering algorithm. *Computers & Geosciences*. Vol. 10(No. 2-3) pp. 191-203.