1K2-IOS-1b-7

A Data Mining Framework for Building Dengue Infection Disease Model

Daranee Thitiprayoonwongse[1] Prapat Suriyaphol[2] Nuanwan Soonthornphisaj[*1]

[*1] Department of Computer Science, Faculty of Science, Kasetsart University, Bangkok Thailand
[*2] Bioinformatics and Data Management for Research Unit, Office for Research and Development, Faculty of Medicine, Siriraj Hospital, Mahidol University, Bangkok Thailand

Dengue infection is a virus-caused diseased that is spread by mosquitoes. Currently, there is no specific treatment, moreover there is no vaccine to prevent the people from the disease. Symptoms of this infection show rapid and violent to patients in a short time. Two classification problems are explored in this study, which are the Dengue Classification Problem and the Day of defervescence of fever (day0) detection problem. Two data mining models, fuzzy logic and decision tree, are studied in this work. We propose to use the knowledge obtained from the decision tree with the fuzzy logic approach in order to obtain better performance. The experimental result shows that using fuzzy logic approach for Dengue Classification is a suitable method. We get 97.39% of the average accuracy from fuzzy logic approach. Consider the day0 detection problem, using the attributes obtained from the dengue patient can provide the early warning on one day before the day0 with the sensitivity as 71.11% by using fuzzy logic.

Key-word: Data Mining, Dengue infection, the day of defervescence of fever, fuzzy logic.

1. Introduction

Dengue infection is one of the major health problems in Thailand. According to World Health Organization, Dengue is divided into 4 types which are DHF I, DHF II, DHF III and DHF IV, respectively [WHO 1999]. Symptoms of dengue infection show rapid and violent to patients in a short time. In this paper, we obtain 2 sources of datasets which contain DF, DHF I, DHF II and DHF III. The data was collected from the first visit of patient until the date of discharge. The number of patients is 524 patients from Srinagarindra Hospital and 477 patients from Songklanagarind Hospital. We propose forty-eigth attributes as a set of meaningful attributes. We selected a decision tree learning as an approach to find the set of informative attributes in order to classify the type of dengue infection. Moreover we selected a fuzzy reasoning approach to classify the type of dengue infection to compare the performance.

Another objective of this research is to provide the early warning for the day of defervescence of fever which is called day0. The day0 date is the critical date of Dengue patients that some patients face the fatal condition. Therefore the target class is day-1 (a day before the day0 date). Therefore we use a set of attributes found in the decision tree as an attribute in fuzzy reasoning and we generate the fuzzy rules from the set of rules obtained from the decision tree.

2. Related Work

The problem of dengue classification was explored by Tanner, et al and Faisal, et al. The team of Tanner applied decision tree approach to classified patients into 4 levels which were Probable dengue, Likely dengue, Likely non-dengue and Probable non-dengue. Their dataset contained 1,012 patients from the EDEN study and 188 patients from Vietnam. They found 6 significant features which were platelet count, white blood cell count, body temperature, hematrocrit, absolute number of lymphocytes and absolute of neutrophils. They obtained only 84.7 % correctness. [Tanner 2008].

The team of Faisal's research [Faisal 2010] used a combination of the self-organizing map (SOM) and multilayer feed-forward neural networks (MFNN) to predict the risk of dengue patients. They classified patients into 2 groups which were low risk and high risk using three criteria. These criteria were platelet counts less or equal than 40,000 cell per mm^3, hematocrit greater than or equal to 25% and aspartate aminotransferase or alanine aminotransferase rose by fivefold the normal upper limit. They used only examples from Day0 until Day2 (Day2 refers to 2 days after the day of defervescence of fever). Finally, they got only 70% correctness.

Fatimah Ibrahim et al. [Ibrahim 2005] predicted the day of defervescence of fever (day0) from 252 dengue patients (4 DF and 248 DHF). They used Multi-Layer Perceptrons (MLP) and got 90% correctness. However they just used eight signs and symptoms in their research. In this paper, seven significant attributes from decision tree and the fuzzy logic approach were employed to detect the day of defervescence of fever.

3. Method

The raw data obtained from two hospitals in north eastern and southern of Thailand was prepared before applying to the algorithm.

3.1 Datasets

Our datasets were obtained from Srinagarindra Hospital, Khon Kaen, Thailand (KK) and Songklanagarind Hospital, Songkla, Thailand (SK). KK data consists of 524 patients. There are 240 DF, 116 DHF I, 118 DHF II and 50 DHF III. In SK dataset, we get 477 patients that consists of 248 DF, 106 DHF I, 111 DHF II and 12 DHF III. No DHF IV is found in both data sets therefore

[1]Contact: Nuanwan Soonthornphisaj, Department of Computer Science, Faculty of Science, Kasetsart University, Bangkok, Thailand. Email: fscinws@ku.ac.th

we will discard this dengue class. These attributes consists of 26 numerical attributes, 21 categorical attributes and one class attribute.

We excluded noisy data such as outliers in some attributes. Then, we replaced missing value with the mean of the same attributes. In addition, we picked up the set of suitable attributes and created new features to represent some data pattern. Then we transformed some attribute values in order to qualify the requirement of the algorithms. In this paper, we got 48 attributes from the clinical and laboratory data [Thitiprayoonwongse 2011].

Table 1. Feature set obtained after preprocessing step.

Attribute name	Meaning
JE vaccine	Received JE vaccine
URI	Upper respiratory tract infection
Bleeding	Bleeding
hematocrit _max	Maximum value of hematocrit concentration
hematocrit _min	Minimum value of hematocrit concentration
AST_max	Maximum value of AST
AST_min	Minimum value of AST
AST_avg	Average value of AST
ALT_max	Maximum value of ALT
ALT _min	Minimum value of ALT
ALT _avg	Average value of ALT
temperature_max	Maximum of temperature
temperature _min	Minimum of temperature
sbp _dbp_avg	The difference between sbp and dbp
liver_size_avg	Average size of grown liver
hematocrit_max_dx	Maximum value of hematocrit concentration
hematocrit_min_dx	Minimum value of hematocrit concentration
hematocrit_avg_dx	Average value of hematocrit concentration
white_blood_cell_max	Maximum of WBC (x1000)
white_blood_cell _min	Minimum of WBC (x1000)
white_blood_cell _avg	Average of WBC (x1000)
platelet_max	Maximum of platelet count (x1000) by machine
platelet_min	Minimum of platelet count (x1000) by machine
platelet_avg	Average of platelet count (x1000) by machine
protein_avg	Average value of protein in liver
albumin_avg	Average value of albumin
globurin_avg	Average value of globulin
ratio_albumin_avg	Average value of ratio between albumin and globulin
quantity_max_found	Maximize quantity value of tourniquet test
pulse_pre_min_found	Minimum of different pressure value evidence
rash_found	Rash on skin evidence
itching_found	Itching related to rash evidence
bruising_found	Bruising evidence
diarrhea_found	Diarrhea evidence
uri_found	Upper reparatory infection evidence
abdominal_found	Abdominal pain
dyspnea_found	Evidence of dyspnea
ascites_found	Evidence of ascites
juandice_found	Evidence of jaundice
liver_tenderness	Evidence of liver tenderness
liver_found	Evidence of grown liver
lymph_found	Evidence of lymph node enlargement
injected_found	Injected conjunctive evidence
atypical_lymp_found	Atypical lymphocyte evidence
Effusion_Result	Effusion evidence
leakage	Evidence of plasma leakage
shock	Evidence of shock
dx	Class

3.2 Decision Tree Approach

Decision tree is an inductive learning approach. It is a powerful and popular tool for classification and prediction. The attractiveness of decision trees is due to the fact that, in contrast to neural networks, decision trees represent rules. Rules can readily be expressed so that humans can understand them.

A decision tree can be used to classify an example by starting from the root node and moving through it until a leaf node is found which provides the classification of the instance.

The requirements to do data with decision tree are 1) a set of predefined classes, 2) Attribute-value description: object or case must be expressible in terms of a fixed collection of properties or attributes, and 3) Sufficient training data.

The algorithm constructs a tree which consists of a set of informative attributes. These attributes are qualified by the gain ratio since they can reduce the entropy of the classes. Consider the entropy equation (see equation 1). For the multiclass problem, entropy equation is defined as shown in equation 2. Finally the gain value is calculated in equation 3.

$$\text{Entropy}=\frac{-P}{P+N}\log_2\frac{P}{P+N}-\frac{N}{P+N}\log_2\frac{N}{P+N} \quad (1)$$

Where S is the training data set, P is the number of positive class and N is the number of negative class.

$$\text{Entropy}(S)=\sum_{i=1}^{c}-p_i\log_2 p_i \quad (2)$$

Note that S is the training data set, p_i is a ratio of class i compare with all data, and c is the number of class.

$$\text{Gain}(S,A)=\text{Entropy}(s)-\sum_{v\in\text{Values}(A)}\frac{|S_v|}{S}\text{Entropy}(S_v) \quad (3)$$

Note that S is the prior data set before classified by attribute A, $|S_v|$ is the number of examples those value of attribute A are v, $|S|$ is the total number of records in the data set.

For the decision tree approach, we used a software namely WEKA to learn from the training set.

3.3 Fuzzy Logic

Fuzzy logic is a form of probabilistic logic. It evaluates with reason that is approximate rather that fixed value. In traditional logic theory, the binary sets have true or false but fuzzy logic variables may have a degree value of truth value that ranges between 0 and 1. The reasoning of fuzzy logic is similar to human reasoning.

Fuzzy inference is the process of formulating the mapping from a given input to an output using fuzzy logic. The mapping then provides a basis from which decisions can be made, or patterns discerned. The process of fuzzy inference involves all of the pieces that are described in the previous sections: Membership Functions, Logical Operations, and If-Then Rules. Fuzzy inference systems have been successfully applied in fields such as automatic control, data classification, decision analysis, expert systems, and computer vision.

There are 5 steps for fuzzy inference system which are 1) fuzzify inputs 2) apply fuzzy operators 3) apply implication method 4) aggregate all output and 5) defuzzify.

In the first step, given the input attributes, the system determines the degree to each of the appropriate fuzzy sets via membership functions.

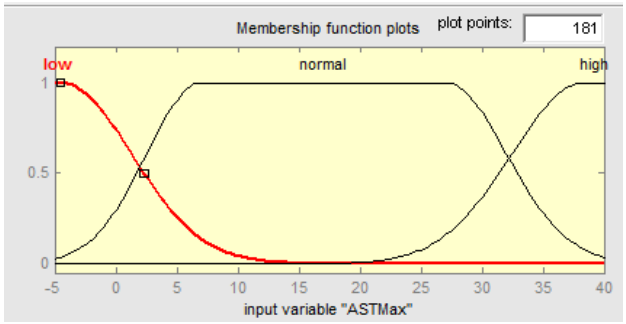
In the second step, we obtained the degree of inputs which were each part of the antecedent. We applied fuzzy operator when there is more than one condition in the rule. Note that In this work, we apply the set of rules obtained from the decision tree.

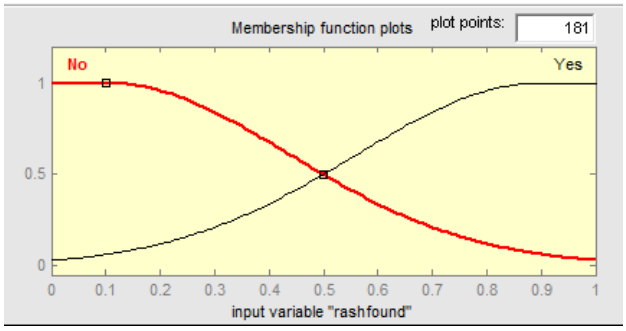
In the Apply Implication Method step, we can assign weight to each rule (a number between 0 and 1). This step is defined as the shaping of the fuzzy sets.

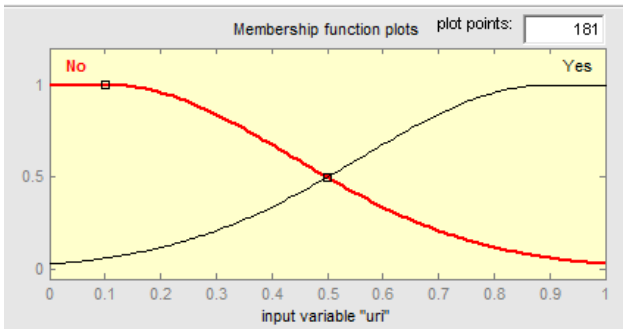
In the forth step, we aggregated the outputs of each rule and combine them into a single fuzzy set in preparation for final step.

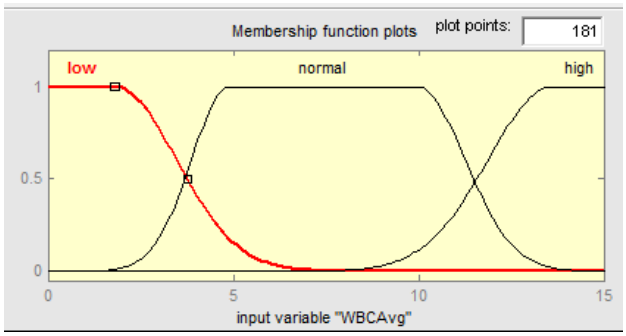
For the defuzzification step, the centroid is calculated which correspond to the center of area under the curve. In this paper, we test our approach with the data which contains 968 instances and seven attributes. We create membership functions are built on Guassian distribution curve with the normal range value in each input.

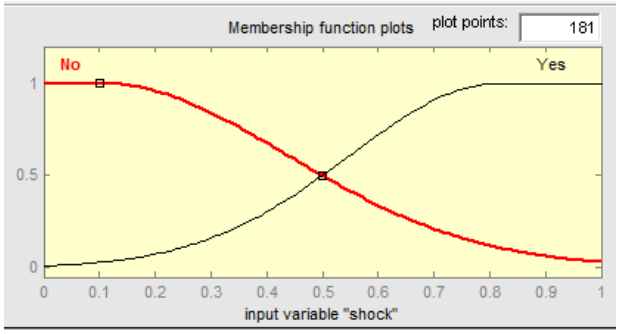
We provide the detail of each membership function used in Dengue Classification Problem and Day0 Detection Problem. Note that the range of each feature was obtained from expert. For the Dengue Fever Classification Problem, we propose to use 16 attributes as shown in Fig 1.

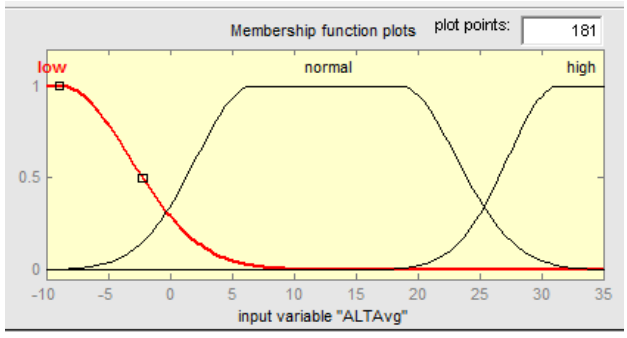


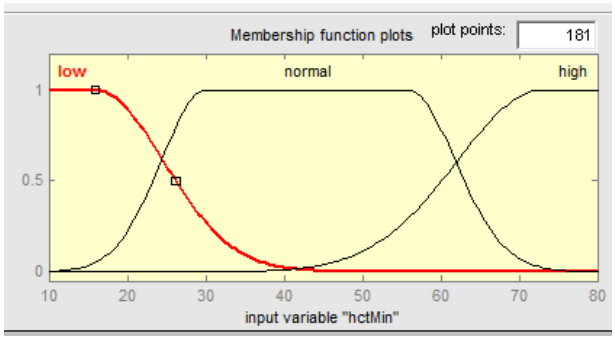


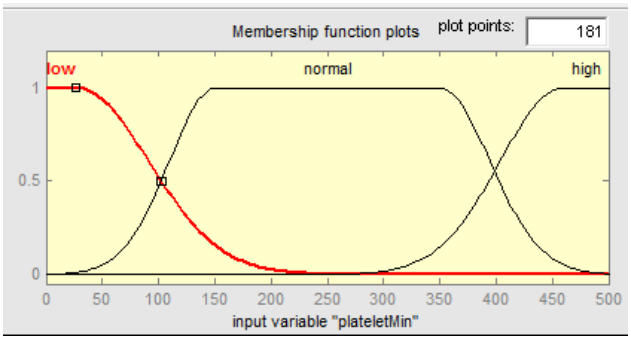


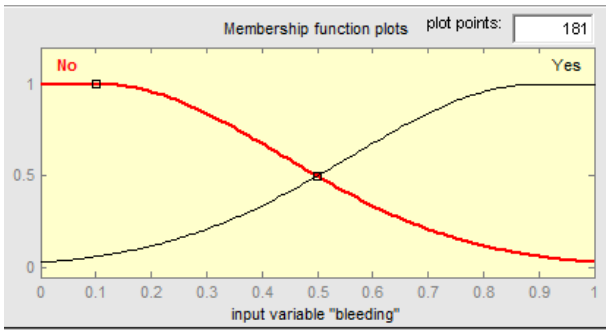


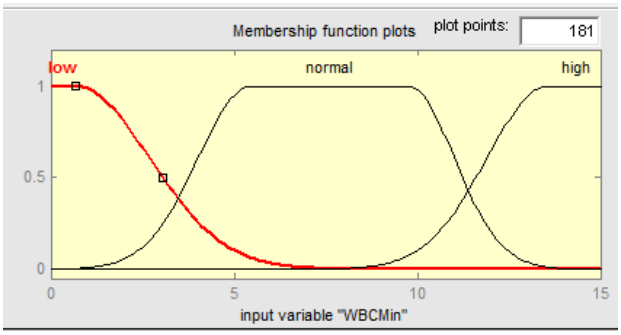


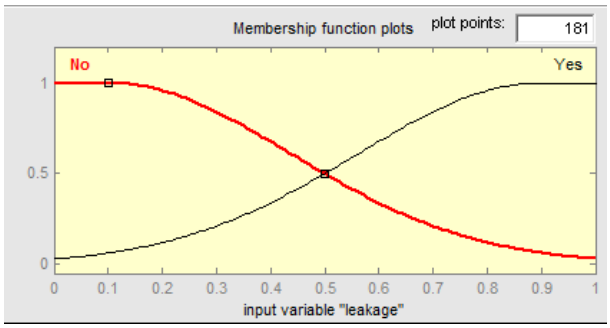


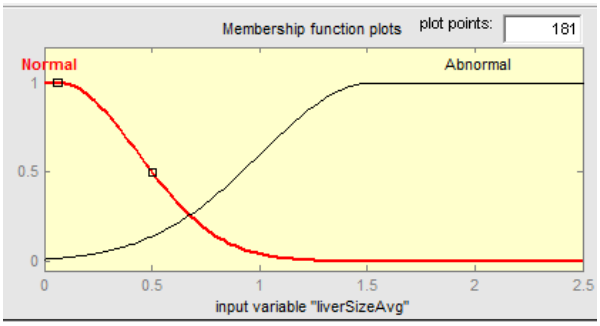

Membership function plots
plot points:
181
low
normal
high
input variable "hctMin"

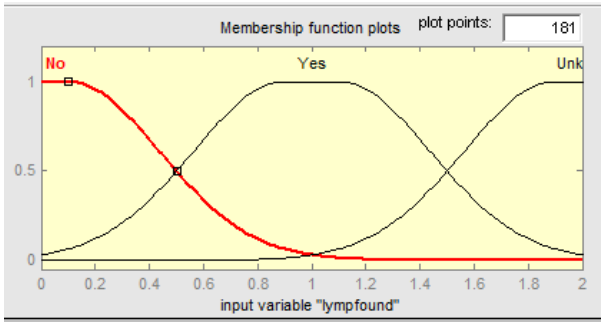

Membership function plots
plot points:
181
low
normal
high
input variable "plateletMin"

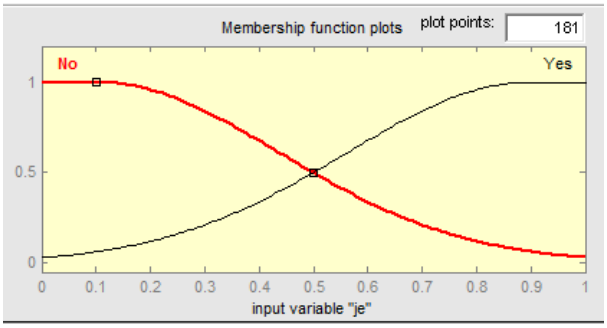

Membership function plots
plot points:
181
No
Yes
input variable "bleeding"

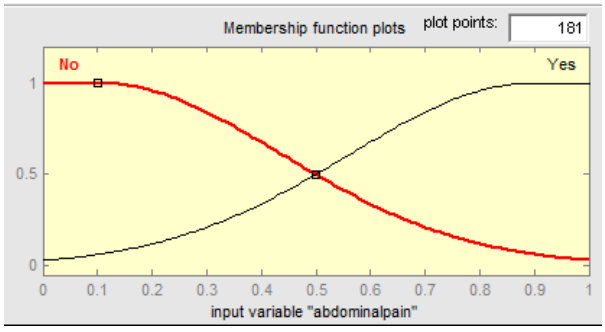

Membership function plots
plot points:
181
low
normal
high
input variable "WBCMin"

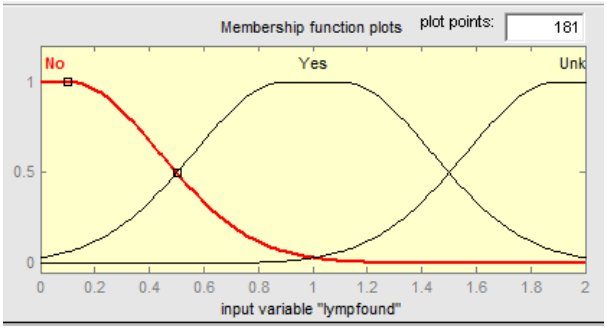

Membership function plots
plot points:
181
No
Yes
input variable "leakage"


Membership function plots
plot points:
181
Normal
Abnormal
input variable "liverSizeAvg"


Membership function plots
plot points:
181
No
Yes
input variable "lympfound"


Membership function plots
plot points:
181
No
Yes
input variable "je"


Membership function plots
plot points:
181
No
Yes
input variable "abdominalpain"


Membership function plots
plot points:
181
No
Yes
input variable "lympfound"

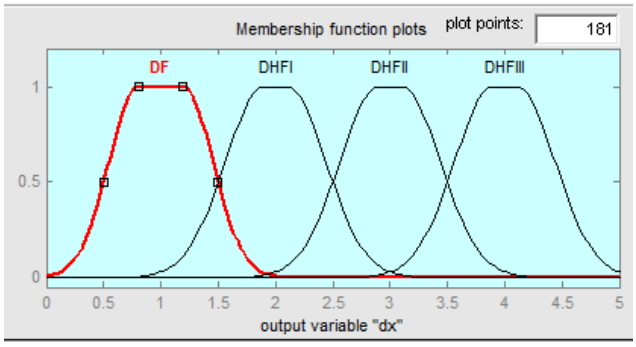
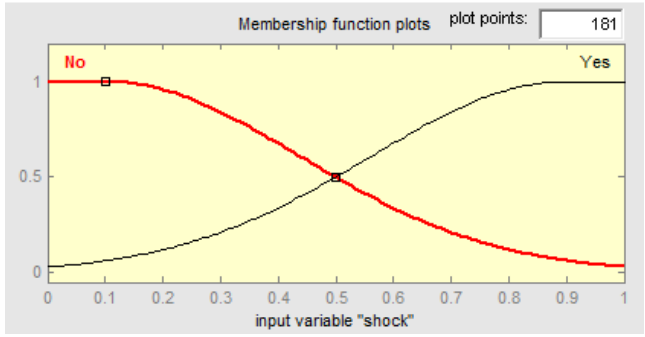
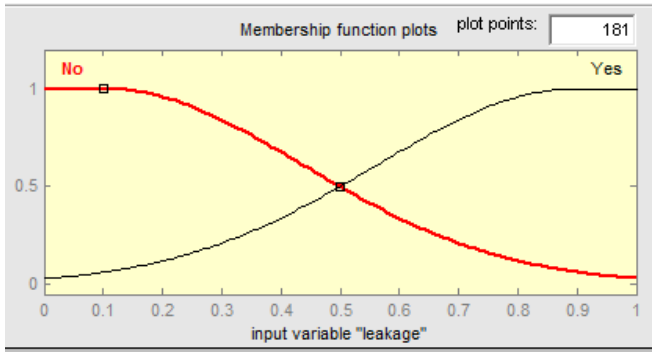
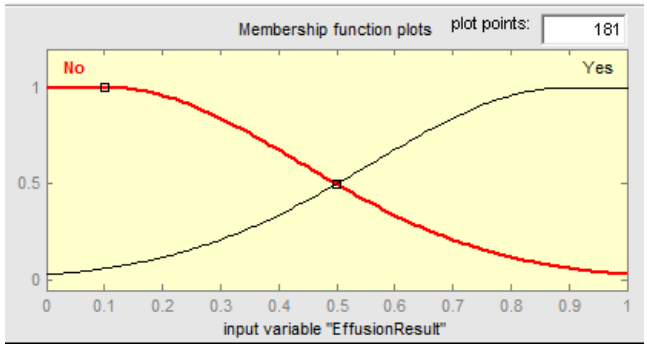


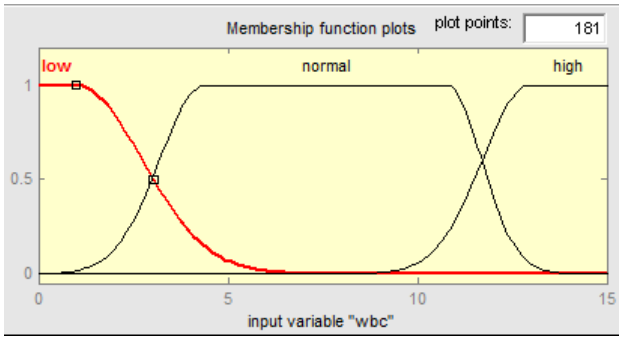
Fig. 1 Membership Functions of each input in Dengue classification

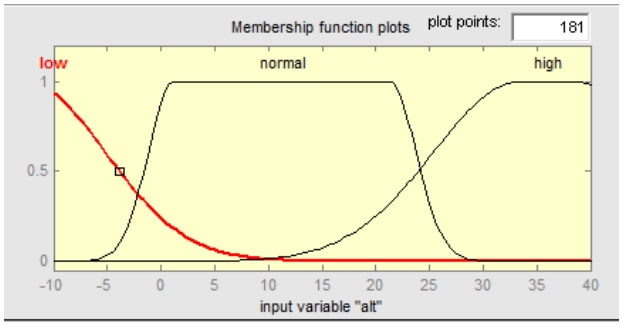
For the Day0 Detection Problem, we propose to use 7 attributes as shown in Fig 2.

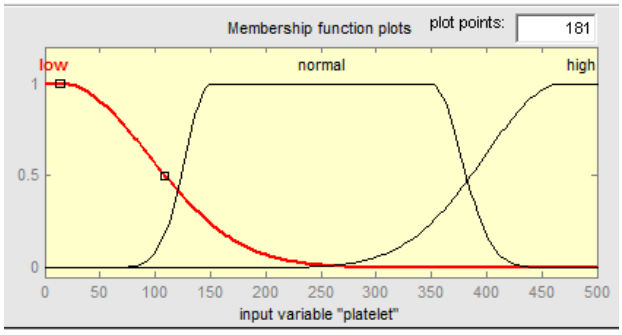


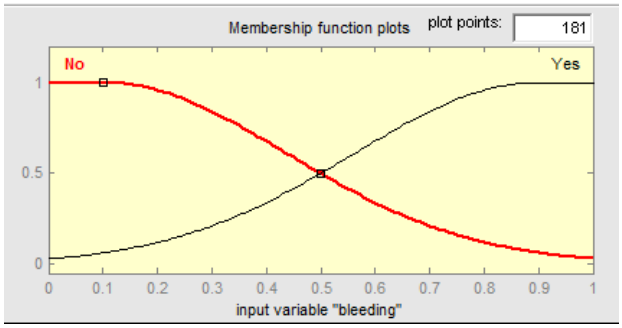












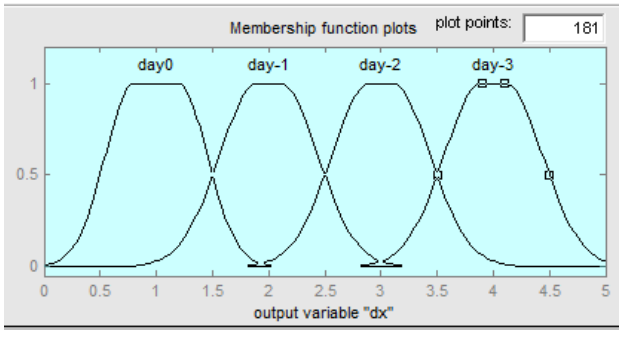


Fig. 2 Membership Functions of each input in Day0 detection problem

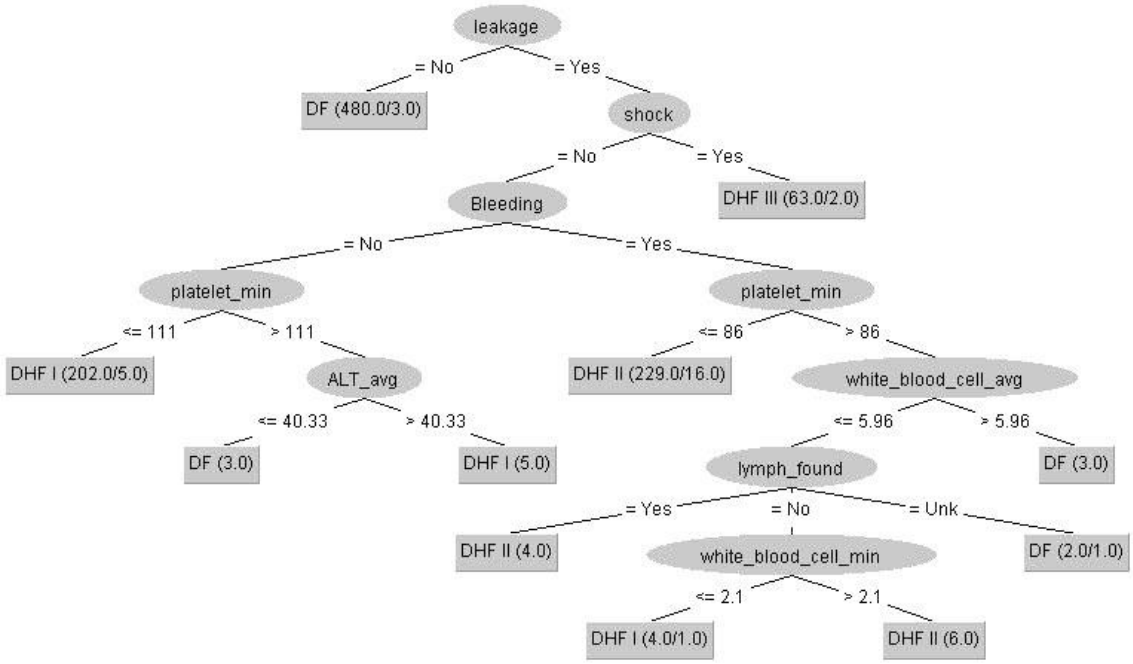


Fig. 3 Decision Tree for Dengue Classification Problem

4. Performance Measurement

We use sensitivity, specificity and accuracy as performance measures. Three equations are defined as following.

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (4)$$

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (5)$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (6)$$

Where TP is the number of True Positives, TN is the number of True Negatives instance, FP is the number of False Positives and FN is the number of False Negative.

4.1 Result on Dengue Classification Problem

In this experiment, we applied fuzzy logic and decision tree algorithm in order to compare the performance of both algorithms. We obtained the decision tree as shown in Fig. 3.

We found 8 significant attributes needed to classify patients. These attributes were **leakage ::** leakage of plasma in blood, **shock ::** shock evidence found during treatment period, **Bleeding ::** bleeding evidence found, **platelet_min ::** the minimum of platelet count, **ALT_avg ::** the average of alanine aminotransferase value, **white_blood_cell_avg ::** the average number of white blood cell in laboratory test, **lymp_found ::** evidence of lymphocyte node enlargement, and **white_blood_cell_min ::** the minimum number of white blood cell in laboratory test.

There were three rules found for the DF patients. If there was no leakage evidence, he/she would be diagnose as DF. The second rule, if there were leakage evidence, no bleeding, the minimum of platelet count more than 111 and the average number of ALT less than or equal 40.33, those patients would be diagnose as DF. However, if there were bleeding evidence, the minimum of platelet count more than 86 and the average number of white blood cell more than 5.96, those patients would be diagnose as DF.

For DHF I, there were three rules. If there were leakage evidence, no shock evidence, no bleeding evidence and the minimum of platelet count less than or equal 111, those patients would be diagnose as DHF I. If the minimum of platelet count more than111 and the average number of ALT more than 40.33, those patients would be diagnose as DHF I. The third rule, there were bleeding evidence, the minimum of platelet count more than 86, the average number of white blood cell less than or equal 5.96, no lymphocyte node enlargement and the minimum number of white blood cell less than or equal 2.1, those patients would be diagnose as DHF I.

Consider DHF II class; there were three rules. The patients would be diagnosed as DHF II if there was leakage evidence and no shock evidence, and bleeding evidence was found and the minimum of platelet count was less than or equal 86 then those patients would be diagnosed as DHF II. The second rule, if the minimum of platelet count was more than 86, the average value of white blood cell less than or equal 5.96 and lymphocyte node enlargement evidence was found, then those patients would be diagnosed as DHF II. The third rule, if there were no lymphocyte node enlargement evidence and the minimum number of white blood cell more than 2.1, then the patients would be diagnosed as DHF II.

For DHF III class, there was only one rule found. The patient would be diagnosed as DHF III if they found leakage evidence and shock evidence.

We found that the minimum of platelet count, the average number of ALT and the minimum number of white blood cell were correlated with classes. We obtained two attributes which were not correlated with the classes. These attributes were the evidence of lymphocyte node enlargement and the average number of white blood cell. These results also conform to the decision tree.

We found that the decision tree classified patients in each class. As shown in Table 2, the overall accuracy of this model was 96.7 %.

Next, we propose 8 significant attributes to be inputs in fuzzy logic approach. These attributes are shock, leakage, bleeding, the minimum of platelet count, the average of ALT, the minimum of white blood cells, the average of white blood cells and evidence found of lymphocyte node enlargement. The data contained 1001 dengue patients.

As shown in Table 2, Fuzzy logic approach gets 97.39% of overall accuracy.

Table 2 Performance of Fuzzy Logic and Decision Tree on Dengue Classification Problem

Method	Class	Sens(%)	Spec(%)	Acc (%)	Overall Acc(%)
Fuzzy Logic	DF	97.75	99.37	98.55	**97.39**
	DHF I	92.79	96.63	95.77	
	DHF II	90.39	97.51	95.87	
	DHF III	98.39	99.78	99.69	
Decision Tree	DF	97.34	98.56	97.95	96.70
	DHF I	89.64	98.44	96.46	
	DHF II	95.63	96.84	96.56	
	DHF III	98.39	99.55	99.48	

4.2 Day0 Detection Problem

We used daily data obtained from dengue patients without missing value records. We selected the data obtained on the date before day0 and day0 date. The training set was labeled into 4 classes which were day0, day-1, day-2, and day-3. Note that "day0" referred the day of defervescence, whereas "day-1" referred to 1 day before day0 and so on.

First, we selected only significant attributes which had correlation value more than 0.5 from the third experiment. There are bleeding appearance, the number of white blood cell, platelet count, the number of ALT, shock evidence and leakage evidence.

Table 3 The normal range of membership functions

Attribute Name	Normal Range
White Blood Cell (wbc)	4-11 x 10*3 /ul
Platelet count	150-440 x 10*3 /ul
Systolic BP – Diastolic BP	25-35
Albumin	3.5-5.5 g/dl
Hematocrit (HCT)	37-52 %

After that, we applied decision tree approach in order to extract the knowledge and assigned the rules in fuzzy logic approach. We got 64.8% of accuracy as shown in Table 4. The decision tree contained 62 rules to classify the day of defervescence of fever.

The Decision Tree algorithm can correctly classification the day-1 class with the accuracy at 51.56%. For the day-2 patients, we obtained 23.08% correctness. We obtained 64.8 % of overall accuracy.

Next step, we create membership function and assigned the rules. We used Fuzzy Logic Toolbox in MATLAB in order to detect the day0. First, we took the inputs and determine the degree to each of the fuzzy sets via membership functions. We used Gaussian distribution curve of membership function for input and output variables. Then, we created a set of rules by using the rules from decision tree. We transformed the rules of decision tree to the rules of fuzzy. Finally, we obtained the accuracy as shown in Table 4. Note that the sensitivity of day -1 was 71.11% which was better than accuracy from the decision tree. Overall accuracy was also increased to 65.45%.

Table 4 Performance of Fuzzy Logic and Decision Tree on Day0 Detection Problem

Method	Class	Sens(%)	Spec(%)	Acc (%)	Overall Acc(%)
Fuzzy logic	day 0	50.86	75.13	64.88	**65.45**
	day -1	71.11	45.37	54.75	
	day -2	3.21	96.92	81.82	
	day -3	2.00	99.67	94.63	
Decision Tree	day 0	61.61	57.92	59.82	64.80
	day -1	51.56	54.66	53.43	
	day -2	23.08	85.74	71.11	
	day -3	10.00	96.91	88.79	

5. Discussion and Conclusion

This study proposes a hybrid method which combines a decision tree and fuzzy logic approach in order solve 2 well-known problems of Dengue infection. We found that better performance could be achieved by using the combination of these two approaches.

Since the mechanism of decision tree explores the set of informative attributes. We found some significant features that conform with the knowledge of physician. These attributes are shock evidence, leakage evidence, bleeding evidence and the minimum of platelet count.

For the Day0 detection problem, we use the significant features obtained from the decision tree. Then, we create decision tree with J48 and got the knowledge of day0 detection. The knowledge was created the rules for fuzzy logic approach. We obtained better accuracy in day-1which was 71.11% (see Table 4. for details). Note that the class day-1 is the target class because the physician needs to know in advance before the fatal condition occurs on the patient. However, the performance of day-2 and day-3 class is very low for fuzzy logic because missing value is found in the original data. We know that fuzzy logic takes advantage from the continuous value through the

member ship function. From the medicine literature [Green, 1999] the main feature that determines the Dengue severity is the plasma leakage which can be measured in term of the Pleural effusion index. However most patients from the dataset were not test with this measure.

Acknowledgements

This research was supported by the grant from Kasetsart University Research and Development Institute (KURDI). We would like to thank Department of Computer Science, Faculty of Science and Kasetsart University for their support.

References

[Ibrahim 2005] F. Ibrahim, M. N Taib, W. A. B. Wan Abas, C. G. Chan and S. Sulaiman, A novel dengue fever (DF) and dengue haemorrhagic fever (DHF) analysis using artificial neural network (ANN), Computer Methods and Programs in Biomedicine, No.79, 2005, pp. 273-281.

[Faisal 2010] T. Faisal, F. Ibrahim and M.N. Taib, A noninvasive intelligent approach for predicting the risk in dengue patients, Expert Systems with Application,Vol.37, No.3, 2010, pp. 2175-2181.

[Green 1999] S. Green, D. W. Vaughn, S. Kalayanarooj, S. Nimmannitya, S. Suntayakorn, A. Nisalak, A. L. Rothman, and F. A. Ennis, Elevated Plasma Interleukin-10 Levels in Acute Dengue Correlate With Disease Severity, Journal of Medical Virology, 1999.

[Roger Jang 1997] JSR Jang and N. Gulley, MATLAB: Fuzzy Logic Toolbox User's Guide, The MathWorks Inc., 24 Prime Park Way, Natick, Mass. 1997.

[Tanner 2008] L. Tanner, M. Schreiber, J.G. Low, A. Ong, T. Tolfvenstam, Y.L. Lai, L.C. Ng, Y.S. Leo, L. Thi Puong, S.G. Vasudevan, C.P. Simmons, M.L. Hibberd and E.E. Ooi, Decision Tree Algorithms Predict the Diagnosis and Outcome of Dengue Fever in the Early Phase of Illness, PLoS Neglected Tropical Disease, Vol.2, 2008.

[Thitiprayoonwongse 2011] D. Thitiprayoonwongse, P. Suriyaphol, and N. Soonthornphisaj, Data Mining on Dengue Virus Disease, 13th International Conference on Enterprise Information Systems (ICEIS 2011), No.1, 2011, pp. 32-41.

[WHO 1999] World Health Organization, Guideline for Treatment of Dengue Fever/Dengue Haemorrhagic Fever, 1999.