Data Mining of Dengue Infection Using Decision Tree

DARANEE THITIPRAYOONWONGSE[1], PRAPAT SURIYAPHOL[2] AND NUANWAN SOONTHORNPHISAJ[1*]
[1]Department of Computer Science, Faculty of Science, Kasetsart University
BANGKOK, THAILAND
g521440162@ku.ac.th, fscinws@ku.ac.th
[2]Bioinformatics and Data Management for Research Unit Office for Research and Development
Siriraj Hospital, Mahidol University
BANGKOK, THAILAND
sipur@mahidol.ac.th

Abstract: - Dengue infection is an epidemic disease typically found in tropical region. Nowadays, the experts need to know the set of features on dengue infection in order to correctly classify the patients since these classes require different treatment. Our temporal dataset consists of clinical data and laboratory data. The data was collected from the first visit of patient until the date of discharge. We obtained 2 sources of datasets from different regions of Thailand which are Srinagarindra Hospital and Songklanagarind Hospital. Each dataset consists of more than 400 attributes. To accomplish the knowledge discovery task, we consider to employ decision tree as a data mining tool. We propose a set of meaningful attributes from the temporal data. Our experiments are divided into 4 parts. The first two experimental results show the useful knowledge to classify dengue infection from Srinagarindra Hospital's dataset and Songklanagarind Hospital's dataset, respectively. Each set of knowledge is tested by different dataset to make sure that the test data was a real unseen data. The third experimental results show the useful knowledge when we integrated 2 datasets. Another objective of this research is to detect 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 physicians need to predict day0 in order to treat the patients. They expect to have an intelligent system that can trigger the day0 date of each patient.

Key-Words: - Data Mining, Dengue infection, the day of defervescence.

1 Introduction

Dengue Fever causes by a bite from mosquitoes carrying dengue virus. Symptoms of dengue infection show rapid and violent to patients in a short time. Dengue infection is divided into 4 types which are DHFI, DHF II, DHF III and DHF IV, respectively [6]. In this paper, we obtained 2 sources of datasets. The total number of patients is 524 patients from Srinagarindra Hospital and 477 patients from Songklanagarind Hospital. We selected a decision tree learning as an approach to find knowledge in order to classify the type of dengue infection. We propose forty-eight attributes as a set of meaningful attributes [1].

We selected decision tree for 4 experiments. The first two experimental results show knowledge for each type of dengue infection. The third experimental results show knowledge from all sources. In the forth experiment, we predicted the day of defervescence of fever or day0.

2 Related Work

There are some researchers who work on dengue classification such as Tanner, et al and Tarig, et al. The team of Tanner classified 1,200 patients using decision tree approach. They found 6 significant features and they got 84.7 % correctness [3]. A combination of the self-organizing map (SOM) and multilayer feed-forward neural networks (MFNN) was employed for the risk prediction of dengue patients in the Tarig's research. They clustered patients into 2 groups which are low risk and high risk using three criteria [5]. They used only examples from Day0 until Day2 (Day2 refers to 2 days after the day of defervescence of fever).On the other hand, they got only 70% correctness.

Fatimah Ibrahim et al. [2] 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.

* corresponding author

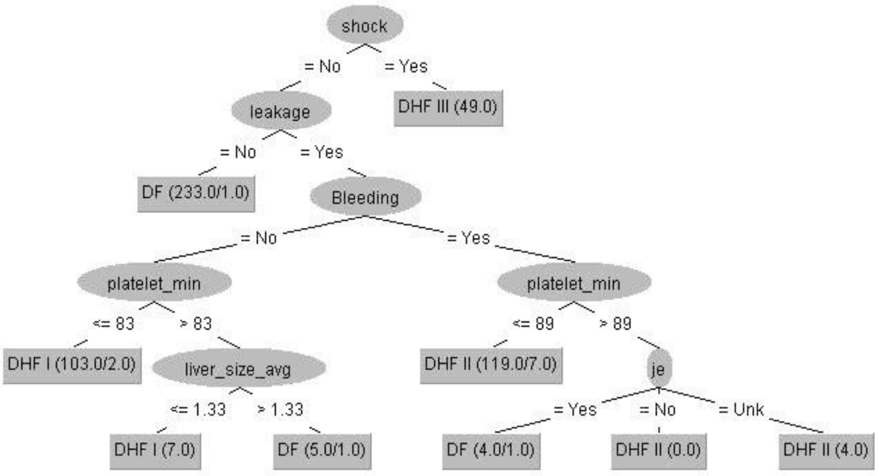

Fig.1 Decision tree of the first experiment.

3 Decision Tree Approach

Decision tree learning is a supervised learning method. The algorithm constructs a tree which consists of a set of selected 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.

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

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

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

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.

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

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.

In this study, we used WEKA software in order to learn from the training set. We changed the value of some parameters in order to know more knowledge and get more accuracy. That parameter was “confidenceFactor” which was used for pruning step (smaller values incur more pruning).

4 Experimental Results

We need to preprocess data followed by the mining step [4]. We excluded noisy data such as outliers in some attributes. Then, we replaced missing value with the mean of some numeric 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 used 48 attributes [1].

We set up 4 experiments. In the first three experiments, we would like to find knowledge in order to classify type of dengue infection. For forth experiment, we would like to predict the day of defervescence with the data before day0 date. We applied decision tree approach to all experiments.

Note that we use sensitivity, specificity and accuracy as performance measures.

4.1 Datasets

We divided dataset into 2 datasets. The total number of patients is 524 patients from Srinagarindra Hospital, Khon Kaen, Thailand (KK). There are 240 DF, 116 DHF I, 118 DHF II and 50 DHF III. In the dataset of Songklanagarind Hospital, Songkla, Thailand (SK), we obtained 477 patients that consists of 248 DF, 106 DHF I, 111 DHF II and 12 DHF III. There is no patient who was diagnosed as DHF IV in both datasets. These attributes consists of 26 numerical attributes, 21 categorical attributes and one class attribute.

4.2 The First Experiment

In the first experiment, we classified on KK's dataset. We set confidenceFactor value to 0.4. We obtained the decision tree as shown in Fig.1

We found 6 significant attributes which are needed to classify patients. These attributes were **shock** – shock evidence found during treatment period, **leakage** - leakage of plasma in blood, **Bleeding** – bleeding evidence found, **platelet_min** – the minimum of platelet count, **liver_size_avg** – the average number of liver size and **je** –JE vaccine.

There are three rules found for the DF patients. The first rule states that ***if*** *there is no shock and no leakage evidence* ***then*** *the patients would be diagnose as DF.* The second rule states that ***if*** *there is leakage evidence and there is no bleeding evidence, and the minimum number of platelet count is more than 83 and the average number of liver size more than 1.33* ***then*** *the those patients would be diagnose as DF.* The third rule states that ***if*** *bleeding evidence is found and the minimum of platelet count is more than 89 and the patient got JE vaccine* ***then*** *those patients would be diagnose as DF.*

There are two rules found for the DHF I patients. The first rule states that ***if*** *there is no shock and leakage evidence and no bleeding evidence and the minimum of platelet count is less than 83* ***then*** *those patients would be diagnose as DHF I.* The second rule states that ***if*** *the minimum of platelet count more than 83 and the average number of liver size less than or equal 1.33* ***then*** *those patients would be diagnose as DHF I.*

There are two rules found for DHF II. The patients would be diagnose as DHF II if there were no shock evidence and leakages evidence, bleeding evidence and the minimum of platelet count is less than 89. However, if the minimum of platelet count is more than 89 and there is no evidence of JE vaccine injection then the patient would be diagnose as DHF II.

For DHF III class, the rule states that the patient would be diagnose as DHF III if there was shock evidence.

We got three significant attributes which were shock evidence, leakage evidence and bleeding evidence. In the first experiment, the minimum of platelet count and the average number of liver size were correlated with the classes. If the minimum of platelet count was increased, the level of dengue fever was non-severe type. On the other hand, if the average number of liver size was increased, the level of dengue fever was severe type. Note that, JE vaccination attribute didn't relate with classes. The results conform to the decision tree.

Table 1 Performance of the first experiment using 10-fold cross validation.

Class	Sens(%)	Spec(%)	Acc (%)	Overall Acc(%)
DF	97.08	97.47	97.29	97.6%
DHF I	93.10	98.50	97.29	
DHF II	95.76	97.99	97.48	
DHF III	98.00	100.00	99.80	

Table 1 shows the performance measured in terms of sensitivity (Sens.), specificity (Spec.) and the average accuracy (Acc.) We found that the decision tree classified patients in DHF III with 98.00 % on sensitivity value. The first experimental results show interesting attributes which were platelet_min, liver_size_avg and JE vacination. Our accuracy rate was 97.6%.

To evaluate the potential of knowledge obtained from the decision tree, we also test the knowledge with the unseen test set using SK's dataset. We obtained the accuracy as shown in Table 2.

Table 2 Performance of 1st experiment measure on unseen testset

Class	Sens(%)	Spec(%)	Acc (%)	Overall Acc(%)
DF	97.98	94.91	96.55	97.2%
DHF I	85.85	99.44	96.34	
DHF II	91.89	96.92	95.73	
DHF III	100.00	98.87	98.90	

We found that the decision tree can completely classified patients in DHF III with 100 % on sensitivity value. We obtained high accuracy when we cross check with different dataset. That means our knowledge can be applied on real data.

From the experimental results, the decision tree algorithm was learned from KK dataset and was tested using SK dataset. We got 97.2% for accuracy.

4.3 The Second Experiment

In the second experiment, we used SK's dataset. The dataset also consists of 4 classes. We set confidenceFactor value to 0.3. We obtained the decision tree as shown in Fig.2.

We found 9 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, **abdominal_pain** – abdominal pain evidence found during treatment peroid, **rash_found** – rash evidence found during treatment period, **uri** - the appearance of upper respiratory infection, **hematocrit_min_d05** – the minimum

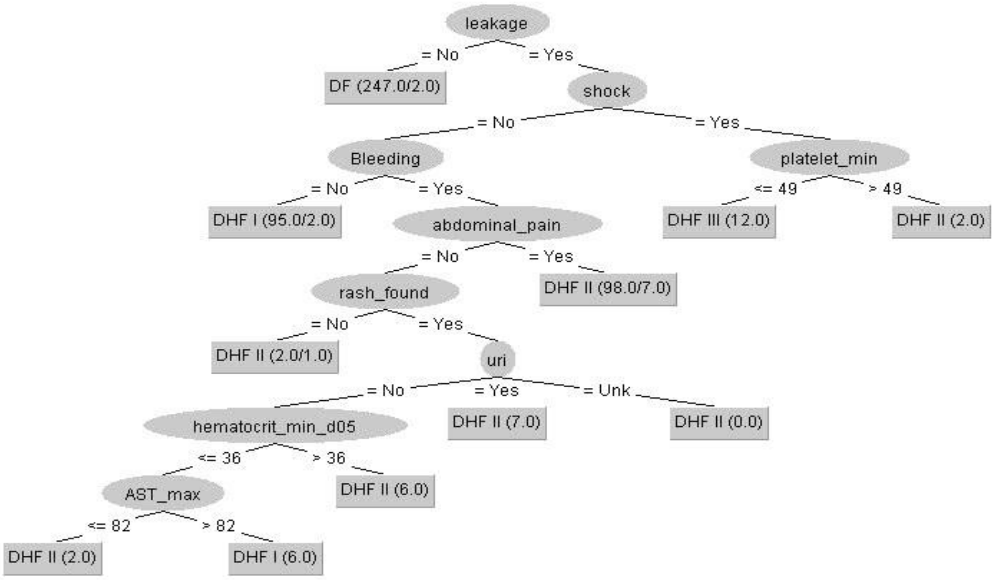

Fig.2 Decision tree of the second experiment

number of hematocrit test during treatment peroid and **AST_max** – the maximum value of aspartate aminotransferase.

There was only one rule found for the DF patients. The rule states that ***If*** *there is no leakage evidence,* ***then*** *the patients would be diagnosed as DF.*

There were two rules found for the DHF I patients. The first rule states that ***if*** *there is leakage evidence and no shock evidence and no bleeding evidence* ***then*** *those patients would be diagnosed as DHF I.* The second rule states that ***if*** *there were bleeding evidence, no abdominal pain, rash evidence found, no upper respiratory infection, the minimum number of hematocrit is less than 36 and the maximum value of AST is more than 82 then those patients would be diagnosed as DHF I.*

There were six rules found for DHF II class. The patients would be diagnose as DHF II if there were leakage evidence, shock evidence, and the minimum of platelet count more than 49. The second rule, if there were no shock evidence, bleeding evidence and they had abdominal pain, then those patients would be diagnosed as DHF II. The third rule, if there was no abdominal pain and no rash evidence, the patients would be diagnosed as DHF II. But if rash evidence and URI were found, then those patients would be diagnosed as DHF II. They also diagnose as DHF II if there were no URI and the minimum number of hematocrit more than 36. However, if that number less than 36 and the maximum value of AST less than 82, they would be diagnose as DHF II.

For DHF III class, there was only one rule found. The patient would be diagnose as DHF III if they found leakage evidence and shock evidence and the minimum of platelet count is less than or equal to 49

Table 3 Performance of the second experiment

Class	Sens(%)	Spec(%)	Acc (%)	Overall Acc(%)
DF	98.79	99.05	98.91	96.6%
DHF I	90.57	97.81	96.18	
DHF II	90.99	96.70	95.37	
DHF III	91.67	99.55	99.34	

We found that the decision tree classified patients in DF with 98.79 % on sensitivity value. In addition, we selected KK's dataset to be the test set. We obtained the accuracy as shown in Table 4. For the second experiment, there were 6 interested attribute. These attributes were platelet_min, abdominal_pain, rash_found, uri, hematocrit_min_d05 and AST_max. We obtained 96.6% of accuracy.

Table 4 Performance of 2nd experiment (on KK dataset)

Class	Sens(%)	Spec(%)	Acc (%)	Overall Acc(%)
DF	96.67	99.62	98.21	95.7%
DHF I	93.97	96.97	96.29	
DHF II	94.07	95.50	95.17	
DHF III	82.00	100.00	98.21	

Consider the unseen data test, if we used SK as a training set and used KK as a test set, we got 95.7%.

In the second experiment, the minimum of platelet count was still related with classes. The maximum of AST related with low correlation

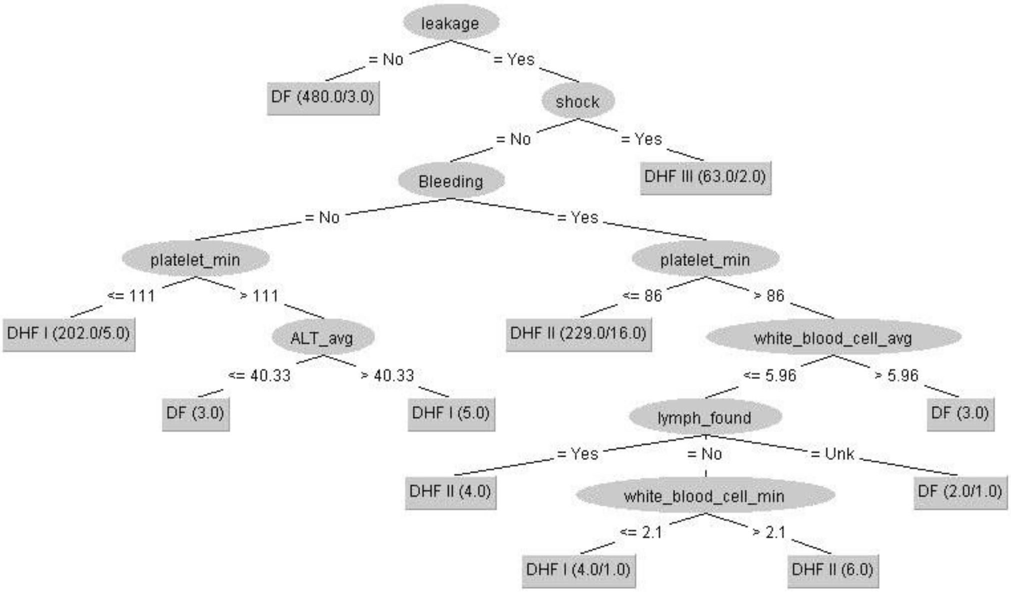

Fig.3 Decision tree of the third experiment.

coefficient value. When we compared with decision tree in the second experiment, AST_max node was a criterion which after hematocrit_min_05 node. So that refer AST_max also didn't relate with classes. There were 4 attributes which didn't relate with classes. There were evidence of abdominal pain, rash found evidence, upper respiratory infection appearance and the minimum number of hematocrit test in the laboratory test.

4.4 The Third Experiment

This experiment, we merged 2 datasets and used decision tree algorithm in order to extract the total knowledge of dengue infection in Thailand. 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, 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.

In the third experiment, 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. This results also conform to the decision tree.

Table 5 Performance of third experiment

Class	Sens(%)	Spec(%)	Acc (%)	Overall Acc(%)
DF	97.34	98.56	97.95	96.7%
DHF I	89.64	98.44	96.46	
DHF II	95.63	96.84	96.56	
DHF III	98.39	99.55	99.48	

We found that the decision tree classified patients in each class. As shown in Table 5. The overall accuracy of this model was 96.7 %.

4.5 The Forth Experiment

This experiment, we focus on day0 prediction. We used daily data obtained from dengue patients without missing value records. We selected the data before day0 and day0 date. The training set was labeled into 5 classes which were day0, day-1, day-2, day-3 and day-4. Note that "day0" referred the day of defervescence, where as "day-1" referred to 1 day before day0 and so on. We obtained the accuracy as shown in Table 6.

Table 6 Performance of third experiment

Class	Sens(%)	Spec(%)	Acc (%)	Overall Acc(%)
day-4	0.00	97.79	96.09	67.8
day-3	5.00	94.02	87.01	
day-2	40.91	82.20	73.18	
day-1	49.32	64.50	58.62	
day0	68.36	67.11	67.79	

The algorithm can correctly classification the day-1 patients with the accuracy rate 58.62%. For the day-2 patients, we obtained 73.18% correctness. We obtained 67.79% of accuracy. The overall accuracy was 67.8%.

5 Conclusion

This study proposed to use a decision tree approach to classify dengue patients from two datasets. We got an accuracy as 97.6%, 96.6 % from the first experiment and the second experiment, respectively. After we explored the correlation of attributes found in the decision trees, we found some significant features which were platelet_min, liver_size_avg, ALT_avg and white_blood_cell_min.

We applied the unseen data in order to test knowledge in the first two experiments. We found that the first decision tree completely classified patients in DHF III with 100% (see Table 2 for details). Because the training set of KK contained more examples compared to those of SK dataset. The accuracy of both experiments in unseen test set were more than 95%.

In the experiment of day0, we obtained very low accuracy in day-4 and day-3 (see Table 6 for details). We found that the tree is overfit. The experimental results shown that the decision tree approach didn't suit this task thus we think we should to select a new classification approach in the future works.

Acknowledgement
This research is supported by Kasetsart University Research and Development Institute (KURDI) under Grant no. 130.55.

References:

[1] 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.

[2] 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.

[3] 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.

[4] N. Soonthornphisaj, *Artificial Intelligence*, Chulalongkorn University Printing House, 2010.

[5] 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.

[6] World Health Organization, *Guideline for Treatment of Dengue Fever/Dengue Haemorrhagic Fever*, 1999