วารสารวิจัยสหวิทยาการไทย
Journal of Thai Interdisciplinary Research
Volume 11, Number 2, Pages 53 - 61

Knowledge discovery on dengue fever using data mining techniques

Nuanwan Soonthornphisaj

Department of Computer Science, Faculty of Science
Kasetsart University, Bangkok 10900, Thailand

Abstract

Dengue fever is a tropical disease mostly found in Asia including Thailand. The report from Ministry of Public Health showed that in 2015 there were 60,000 dengue patients in Thailand.The research aims to analyze the treatment record obtained from Dengue fever patients to find the feature set that can classify the Dengue severity. Several questions from Thai physicians are investigated using association rules. Some important markers such as the size of grown liver, the level of Aspartate aminotransferase and Alanine aminotransferase are studied. The decision tree, fuzzy logic are applied to classify dengue severity. The performance of data mining techniques are compare with World Health Organization criteria.

Keywords: Data mining, decision tree, association rule, fuzzy logic

บทคัดย่อ

ไข้เลือดออกเป็นโรคเขตร้อนที่พบมากในแถบเอเชีย รวมถึงประเทศไทย จากรายงานของกระทรวงสาธารณสุขพบว่า ในปี ค.ศ.2015 มีจำนวนผู้ป่วยด้วยโรคไข้เลือดออกมากถึง 60,000 ราย งานวิจัยนี้มีวัตถุประสงค์เพื่อวิเคราะห์ข้อมูลการรักษาผู้ป่วย เพื่อค้นหาลักษณะเด่นของโรคไข้เลือดออก งานวิจัยนี้ต้องการตอบโจทย์ทางการแพทย์โดยใช้ขั้นตอนวิธีกฎความสัมพันธ์ และต้นไม้ตัดสินใจ ตัวบ่งชี้โรคที่สำคัญเช่นอาการตับโต และระดับของเอนไซม์ AST และ ALT ได้ถูกศึกษาในงานวิจัยนี้ ต้นไม้ตัดสินใจและฟัซซี่ลอจิกได้ถูกนำมาทดลองเพื่อประเมินความแม่นยำในการทำนายระดับความรุนแรงของโรคไข้เลือดออก โดยได้ทำการศึกษาเปรียบเทียบกับเกณฑ์ขององค์การอนามัยโลก

คำสำคัญ: เหมืองข้อมูล, ต้นไม้ตัดสินใจ, กฎความสำคัญ, ฟัซซี่ลอจิก

1. Introduction

Dengue fever is a mosquito-borne tropical disease caused by the dengue virus Symptoms include fever, headache muscle joint pains, and skin rash that is similar to measles. In a small proportion of cases, the disease develops into the life-threatening dengue hemorrhagic fever, resulting in bleeding, low levels of blood platelets and blood plasma leakage. Some cases develops into dengue shock syndrome, where low blood pressure occurs. The magnitude of the dengue problem has increased dramatically and has extended geographically to many previously unaffected areas. It was then, and remains today, the most important arthropod-borne viral disease of humans [1]. Dengue is the most rapidly spreading viral disease in the world. In the last 50 years, incidence has increased 30-fold with increasing geographic expansion to new countries and, in the present decade, from urban to rural settings. An estimated 50 million dengue infections occur annually and approximately 2.5 billion people live in dengue endemic countries.

Epidemic dengue is a major public health problem in Thailand where dengue is a leading cause of hospitalization and death in children. In Thailand, dengue is reported from all four regions: Northern, Central, North-Eastern and Southern.

Dengue has a wide spectrum of clinical presen-

E-mail: fscinws@ku.ac.th

tations, often with unpredictable clinical evolution and outcome. While most patients recover following a self-limiting non-severe clinical course, a small proportion progress to severe disease, mostly characterized by plasma leakage with or without haemorrhage. The group progressing from non-severe to severe disease is difficult to define, but this is an important concern since appropriate treatment may prevent these patients from developing more severe clinical conditions. An appropriate treatment, and the decision as to where this treatment should be given (in a health care facility or at home) are influenced by the case classification for dengue. This is even more the case during the frequent dengue outbreaks worldwide, where health services need to be adapted to cope with the sudden surge in demand. Changes in the epidemiology of dengue, as described in the previous sections, lead to problems with the use of the existing World Health Organization guidelines.

Symptomatic dengue virus infections were grouped into three categories: undifferentiated fever, dengue fever (DF) and dengue hemorrhagic fever (DHF). DHF was further classified into four severity grades, with grades III and IV being defined as dengue shock syndrome (DSS). There have been many reports of difficulties in the use of this classification [2, 3, 4]. The study findings confirmed that, by using a set of clinical and/or laboratory parameters, one sees a clear-cut difference between patients with severe dengue and those with non-severe dengue [5]. However, for practical reasons it was desirable to split the large group of patients with non-severe dengue into two subgroups -- patients with warning signs and those without them. It must be kept in mind that even dengue patients without warning signs may develop severe dengue.

This research aims to use various data mining techniques to answer the following questions.

1) Which attributes can be used to categorize the dengue fever patients?

2) What are the patterns for the dengue fever severity grading?

3) How to predict the day of defervescence?

4) Does the hematomegaly is the indicator todifferentate the type of dengue?

5) Can the dengue fever occur in the patient who had the Japanese Encephalitis: JE vaccine injected?

6) Is the level of Aspartate aminotransferase (AST) always higher than the Alanine aminotransferase (ALT) or not?

2. Related works

The overall goal of the medical data mining is to extract knowledge from a database and transform it into an understandable structure for new knowledge discovery. A lot of raw data in the form of surveys taken from different hospitals and diagnosis laboratories in Chennai and Tirunelveli from India were used as the data sets used in theexperiment. It consists of 5,000 samples with 29 symptoms associated with the disease. Each sample consists of few measurements with label that denotes its symptoms. Of these, some were dengue positive and some were dengue negative though the symptoms seemed to be like dengue positive. Support vector machines (SVM) was applied as a classifier to classify the dengue positive and dengue negative patients [6]. They found that SVM model can differentiate dengue positive and negative cases. The total accuracy for training data for SVM was 90.3% in both cases.

Decision Tree (DT), Artificial Neural Network (ANN), and Rough Set Theory (RS) was studied to find the best model for Malaysian dengue outbreak detection dengue classification [7]. They concluded that the selection of attributes used in the study is more appropriate than those in previous researches. On the other hand, although DT and ANN are well known methods in dengue outbreak domain, the significant selection of attributes enable the algorithms gain the highest accuracy.

Rao and Kumar [8] developed a new computational intelligence-based methodology that predicts the diagnosis of dengue in real time, minimizing the number of false positives and false negatives. One of their methodologies is by using decision tree method that employs boosting for generating highly accurate decision rules. The predictive models developed using their methodology are found to be more accurate than the state-of-the- art methodologies used in the diagnosis of the dengue fever. Tanner *et al.* [9] also employed decision tree algorithm in predicting the outcome of the dengue fever in the early phase. They use a C4.5 decision tree classifier for analysis of

all clinical, haematological, and virological data. The accuracy of the model produced is 84% which can differentiate dengue from non-dengue febrile illness. This study shows a proof-of-concept that decision algorithms using simple clinical and haematological parameters can predict diagnosis and prognosis of dengue disease, a finding that could prove useful in disease management and surveillance.

A non-invasive prediction of the day of defervescence of fever in dengue patients using artificial neural network using Multilayer Feed Forward Neural Networks was studied by Ibrahim *et al.* [10].

3. Data mining techniques

Data Mining comprises techniques and algorithms, for determining interesting patterns from large datasets. There are currently hundreds of algorithms that perform tasks such as frequent pattern mining, clustering, and classification.

Decision tree is an algorithm that generates a tree representing the model of classes from training data. The algorithm is attractive because it can transform to the understandable set of rules. Each node in the tree is an attribute that is the best splitter because it can reduce the diversity of the training set by the greatest amount. The well-known decision tree proposed by Quinlan [11] namely C4.5 uses Gain ratio to avoid the bias caused by attribute having larger number of values.

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

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.

$$\mathrm{GainRatio}(S, A) = \frac{Gain(S, A)}{SplitInfo(S, A)}$$

where $\mathrm{SplitInfo}(S, A)$ is the information due to the split of S on the basis of the value of the categorical attribute A.

Association Rule mining using Apriori algorithm [12] is a well-known algorithm used in data mining. It finds interesting associations and/or correlation relationships among large set of data items. Association rules show attribute value conditions that occur frequently together in a given dataset in the form of ‘‘if-then" statements.

Data Clustering can be considered as the most important unsupervised learning problem. An output cluster is a collection of objects which are ‘‘similar" between them and are ‘‘dissimilar" to the objects belonging to other clusters. Clustering algorithms can be applied in many fields, such as Biology, Marketing, etc. K-mean clustering is the most widely used algorithm for data clustering which is a partitioning method that separates instances into k groups. K-means uses an iterative algorithm that minimizes the sum of distances from each object to its cluster centroid, over all clusters. This algorithm moves objects between clusters until it is converged to the minimum distance. The result is a set of clusters that are as compact and well-separated as possible.

Correlation analysis is another statistics tools for finding the relationships among the variables. The relationship represents in term of coefficients that measure the degree of correlation. The most common of several coefficients is the Pearson correlation coefficient, which is sensitive only to a linear relationship between two variables.

4. Experimental result data preparation

The total number of 258 patients was obtained from Siriraj Hospital, Bangkok, Thailand. The data set consists of 128 DF, 65 DHF I, 52 DHF II and 13 DHF III. The set of attributes consists of clinical attributes and hematological attributes. There are totally 48 attributes (26 numerical attributes, 21 categorical attributes and one class attribute).

Attributes in Table 1 were recorded during the first visit of the patient. Some attributes were preprocessed such as Bleeding. The Bleeding value was determined from any evidences found from spontaneous petechiae, ecchymosis, gum, nose, vomiting, stool.

During the treatment period, nurses and physicians followed the symptoms as shown in Tables 2 and 3. Temporal attributes are summarized in terms of maximum, minimum and average values [13].

Table 1. Attributes obtained from the first visit of patients.

Attribute	Type	Meaning
JE vaccine	Categorical	Received JE vaccine
URI	Categorical	Upper respiratory tract infection
Bleeding	Categorical	Bleeding

Table 2. Numerical Attributes obtained during the treatment period.

Attribute	Meaning
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

There are three commonly used performance measurements including sensitivity, specificity and accuracy as defined in (1), (2) and (3), respectively. The sensitivity is referred as the true positive rate, and the specificity as the true negative rate. The accuracy of classifiers is the percentage of correctness of outcome among the test sets.

$$\text{sensitivity} = \frac{TP}{TP+FP} \quad (1)$$

$$\text{specificity} = \frac{TN}{TN+FN} \quad (2)$$

$$\text{accuracy} = \frac{TP+TN}{TP+FP+TN+FN}. \quad (3)$$

Experiment 1: Which attributes can be used to categorize the dengue fever patients?

Data set: The data set is obtained from 1001 patient consisting of 4 dengue types which are Dengue Fever patient (DF: 488), dengue hemorrhagic fever type I (DHF I: 222), dengue hemorrhagic fever type II (DHFII : 229) and dengue hemorrhagic fever type III (DHF III: 12).

Table 3. Categorical Attributes obtained during the treatment period (categorical values).

Attribute	Meaning
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	Atyp lymphocyte evidence
Effusion_Result	Effusion evidence
leakage	Evidence of plasma leakage
shock	Evidence of shock
dx	Class

Result: The decision tree algorithm is applied for the feature selection process and it is found that the plasma leakage, the shock occurrence, the bleeding, the number of platelet, the level of ALT, the number of white blood cell, lymphadenopathy are the potential feature sets that can categorize the dengue patients.

After the feature selection, the fuzzy logic approach is tried to see the classification performanceas shown in Table 4. The experimental result shows that Fuzzy logic outperforms Decision tree with the 97.94% of accuracy.

Experiment 2: What are the patterns for the dengue fever severity grading?

Result: The result obtains from decision tree reveals the pattern for each of dengue fever severity as follows:

The pattern of DF comprises 4 patterns which are 1) If there is no evidence of plasma leakage then the patients should be classified as the Dengue fever (DF). Or 2) if the plasma leakage occurs but no bleeding and the number of platelet count is greater than 111,000 cells/μl and the average level of ALT is $\leq$ 40.33 U/L. Or 3) if the plasma leakage occurs, the bleeding is found, even the platelet count is less than the normal range (86,000 cells/μl) but no shock evidence then the patients should be classified as the Dengue fever (DF). Or 3) if the plasma leakage occurs, the bleeding is found even the platelet count is less than the normal range (86,000 cells/μl) and the white blood cell count is less than normal range (5,960 cells/μl) but no evidence of lymphadenopathy is found then the patients should be classified as the Dengue fever (DF).

The main characteristics of DHF I are that there are no evidence of plasma leakage, no shock evidence but if the patient has the level of ALT a bit higher than the normal range ($\geq$ 40.33 U/L) or the number of platelet count is a bit less than normal range ($\leq$ 111,000 cells/μl) then the patientwould be classified as the DHF I. However if the bleeding occurs for those patient and the number of white blood cells is less than the normal range ($\leq$ 5,960 cells/μl) and no evidence of lymphadenopathy is found then the patientwould be classified as the DHF I as well.

For the DHF II, the evidence of plasma leakage and the bleeding occurrence are the main indicator. In case that the patient has shocked then he/she will be classified as DHF III.

Table 4. The classification performance of Decision Tree and Fuzzy Logic.

Method	Class	Sensitivity (%)	Specificity (%)	Accuracy (%)	Overall Accuracy (%)
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	
Fuzzy Logic	DF	97.75	99.39	98.57	**97.94**
	DHF I	91.89	98.44	96.98	
	DHF II	96.51	97.25	97.08	
	DHF III	98.39	99.78	99.69	

Table 5. The performance of Day 0 prediction.

Method	Class	Sensitivity (%)	Specificity (%)	Accuracy (%)	Overall Accuracy (%)
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	
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	

Table 6. The co-occurrence of hematomegaly and the degree of dengue severity.

Dengue severity	DF	DHF I	DHF II	DHF III	Total
No. of patients	488	222	229	62	1001
No. of hematomegaly	333	180	206	59	778
(%)	68.24%	81.08%	89.96%	95.16%	77.72%

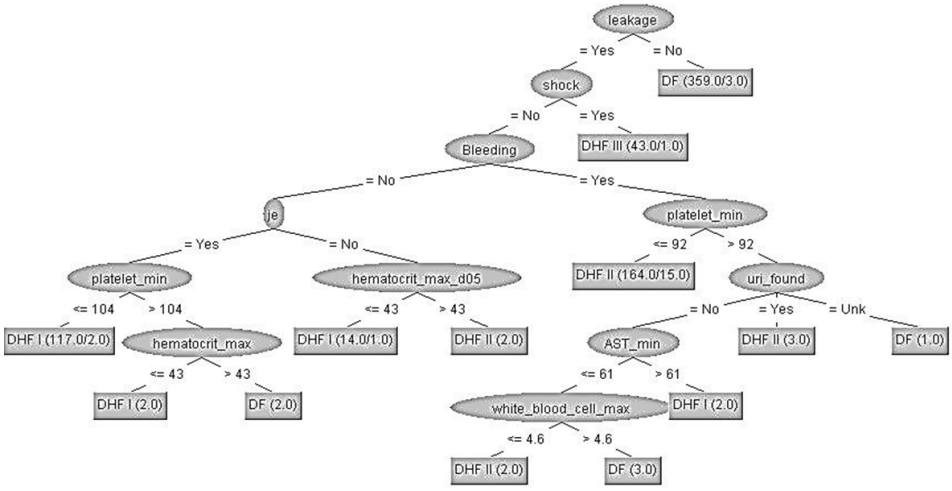

Table 7. Number of patients who have AST > ALT.

Class	DF	DHF I	DHF II	DHFIII	DHF
Number of patients	487	222	229	62	513
AST > ALT(patients)	411	213	226	62	501
AST > ALT(%)	84.39	95.95	98.69	100	97.66

Table 8. Correlation analysis of attributes obtained from experiment 1.

Attributes	Correlation	output
leakage	0.8456	+high
shock	0.5377	+medium
Bleeding	0.3719	+low
platelet_min	-0.5853	-medium
ALT_avg	0.1941	+very low
wbc_avg	0.0842	+very low
lymp_found	-0.0905	-very low
wbc_min	-0.0785	-very low

Table 9. The false negative value obtained from decision tree.

class	No. of patients	Classified by Decision Tree	No of misclassified	False Negative (%)
DF	488	482	13	2.66
DHF I	222	211	23	10.36
DHF II	229	243	10	4.37
DHF III	62	65	1	1.61

Experiment 3: How to predict the day of defervescence (Day 0) ?

Result: The feature selection was done using decision tree then fuzzy logic was tried to find the performance of the prediction. However the tree over fitting problem is occurred therefore the correlation is used for filter out the features that has the correlation coefficients ≤ 0.5. and repeat the decision tree learning process. Finally the fuzzy logic is applied to compare the classification performance as shown in Table 5.

The feature sets that can be used to identify the Day 0 are the bleeding evidence, the number of white blood count, the number of platelet, the value of ALT, the leakage evidence, the shock evidence and the effusion index.

Experiment 4: Does the hematomegaly is the indicator to differentiate the type of dengue? The decision tree obtained from experiment 1 is further analyzed to see the co-occurrence of hematomegaly and the degree of dengue severity (DF, DHF I, DHF II, DHF III).

It is found that the hematomegaly can be found in both dengue fever and all degrees of dengue hemorrhagic fever. Therefore the hematomegaly evidence is not the indicator for the severity degree of dengue patients.

Experiment 5: Can the dengue fever occur in the patient who had the Japanese Encephalitis: JE vaccine injected?

Result: The preprocessing is done to delete the patient's record in which the JE information is not found, therefore the number of patients is reduced to 714 records. Then the decision tree is used as a learning algorithm followed by the association rule mining using Apriori algorithm. Note that the minimum Support and minimum confidence are set as 0.1 and 0.9.

The patterns obtained from the decision tree reveals that patients with JEvaccine injected can also infected by dengue virus. The results obtained from Apriori algorithm also confirm that JE vaccine injection cannot prevent the dengue virus infection. (88.25% of DHF are JE vaccine injected patients)

Experiment 6: The level of Aspartate aminotransferase: AST is always higher than the Alanine aminotransferase: ALT or not.

Result: The new logical feature is created namely AST_ALT which means that the level of AST is higher than that of ALT. The patient's records are rearranged into 2 classes: DF (487) and DHF (513). Then Apriori algorithm is performed to recheck the answer. The result confirms that Dengue virus typically causes the higher level of AST over ALT.

Table 10. The False Positive value using WHO Criteria.

Class	No. of patients	Classified by WHO	No of misclassified	False Negative (%)
DF	488	819	8	1.64
DHF I	222	137	156	70.27
DHF II	229	41	198	86.46
DHF III	62	0	57	91.94
DHF IV	0	1	0	-
Non Dengue	0	3	0	-

Table 11. The correlation analysis between attributes and class (experiment 5).

Attributes	Correlation	output
leakage	0.845613	+high
shock	0.537681	+medium
Bleeding	0.371913	+medium
je	-0.06766	-very low
platelet_min	-0.58527	-medium
hct_max	0.326651	+low
hct_max_d05	0.264204	+very low
uri_found	-0.06759	-very low
AST_max	0.239272	+very low
wbc_max	0.130681	+very low

5. Discussion

All attributes found in the decision tree (experiment 1) are reprocessed to see the correlation between these attributes and the class as shown in Table 8.

The correlation coefficients show that the plasma leakage and the shock evidence affect the Dengue severity. Furthermore the reduced number of platelet count induces the more severity in Dengue patients. The data mining results obtain in this work are compare with the criteria launched by WHO in terms of False Negative value (see Table 9).

The Dengue severity classification using Decision tree and WHO shows that decision tree can classify the dengue severity better than that of WHO for Class DHFI, II, III. However for DF class, WHO criteria is more suitable than decision tree.

For Day 0 problem, it is found that the dataset is lack of information since most patients visit the physician when the decease has already progressed. Moreover, the size of grown liver affects the Dengue severity as well. In order to investigate the effect of JE vaccine, the patient's record with unknown information about JE vaccine are excluded. The decision tree shows the appearance of JE attribute, therefore the attribute is reprocessed to see the correlation. The correlation coefficient of JE shows that JE vaccine can not prevent the patient from dengue infection.

Acknowledgements: This research was funded by KURDI, Kasetsart University. I would like to thank Dr. Prapat Suriyaphol from Bioinformatics and Data Management for Research Unit, Mahidol University for giving comments that make this research a great success.

References

[1] World Health Organization. (2009). Dengue: guidelines for diagnosis, treatment, prevention and control, New ed. Geneva: World Health Organization.

[2] Guha-Sapir, D. & Schimmer, B. (2005). Dengue fever: new paradigms for a changing epidemiology. **Emerging Themes in Epidemiology**, 2:1.

[3] Deen, J. L., Harris, E., Wills, B., Balmaseda, A., Hammond, S. N., Rocha, C., Dung, N. M., Hung, N. T., Hien, T. T., & Farrar, J. J. (2006). The WHO dengue classification and case definitions: Time for a reassessment. **Lancet**, 368, 170-173.

[4] Rigau-Perez, J. (2006). Severe dengue: the need for new case definitions. **Lancet Infectious Diseases**, 6, 297-302.

[5] World Health Organization, Supra note 1, p. 11.

[6] Shameem Fathima, A. & Manimeglai, D. (2012). Predictive analysis for the Arbovirus - Dengue using SVM classification. **International Journal of Engineering and Technology**, 2 (3), 521-527.

[7] Tarmizi, N. D. A., Jamaluddin, F., Bakar, A. A., Othman, Z. A., Zainudin, S., & Hamdan, A. R. (2013). Malaysia Dengue Outbreak Detection Using Data Mining Models. **Journal of Next Generation Information Technology**, 4 (6), 96-107.

[8] Rao, V. & Kumar, M. (2012). A new intelligence - based approach for computer-aided diagnosis of dengue fever. **Information Technology in Biomedicine**, 16 (1), 112-118.

[9] Tanner, L., Schreiber, M., Low, J. G., Ong, A., Tolfvenstam, T., Lai, Y. L., Ng, L. C., Leo, Y. S., Thi Puong, L., Vasudevan, S. G., Simmons, C. P., Hibberd, M. L., & Ooi, E. E. (2008). Decision tree algorithms predict the diagnosis and outcome of dengue fever in the early phase of illness. **PLoS Negl. Trop. Dis.**, 2, p.196.

[10] Ibrahim, F., Taib, M. N., Abas, C., Guan, C., & Sulaiman, S. (2005). A novel dengue fever (DF) and dengue haemorrhagic fever (DHF) analysis using artificial neural network (ANN). **Computer Methods and Programs in Biomedicine**, 79 (3), 273-281.

[11] Quinlan, J. R. (1987). Simplifying the decision tree. **International Journal of Man-Machine Studies**, 27, 221-234.

[12] Agrawal, R. & Srikant, R. (1994). Fast algorithms for mining association rules in large databases. **In Proc. 20th International Conference on Very Large Data Bases**, p.487-499, Santiago, Chile.

[13] Thitiprayoonwongse, D., Suriyaphol, P., & Soonthornphisaj, N. (2011). Data mining on Dengue Virus Disease. **ICEIS**, 1, 32-41.