ISBN 978-981-11-0008-6
Proceedings of 2016 6th International Workshop on Computer Science and Engineering (WCSE 2016)
Tokyo, Japan, 17-19 June, 2016, pp. 371-375
doi: 10.18178/wcse.2016.06.059

Knowledge Discovery on Dengue Patients Using Data Mining Techniques

Nuanwan Soonthornphisaj [+] and Daranee Thitiprayoonwongse

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

Abstract. Dengue fever represents the leading cause of deaths in Thailand. The statistical data obtained from Ministry of Public Health, Thailand reveals that there are 60,000 dengue patients in 2015. Several questions from Thai physicians are investigated using data mining techniques.This paper presents the research work related to knowledge discovery using data mining technique. The objective is to analyze data obtained from Dengue fever patients in Thailand in order to know the set of indicators that can classify the Dengue severity. 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 and Apriori algorithms are applied to discover new knowledge. The performance of data mining techniques are compare with World Health Organization criteria.

Keywords: data mining, decision tree, fuzzy logic, Apriori.

1. Introduction

A person infected by the dengue virus develops severe flu-like symptoms. Severe dengue is a potentially deadly complication due to plasma leaking, fluid accumulation, respiratory distress, severe bleeding, or organ impairment [1]. The warning signs to look out for occur 3-7 days after the first symptoms in conjunction with a decrease in temperature (below 38 ℃/ 100 ℉). Note that it is referred as day0 in this work. A wide spectrum of clinical presentations is reported ranging from non-severe to severe disease. However, the group of patients 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.Therefore the dengue classification is needed in order to provide an appropriate treatment for patients. 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 useof this classification [2]-[4]. The study findings confirmed that, by using a set of clinicaland laboratory data, one sees a clear-cut difference between patients with severe dengue and those with non-severe dengue.

A research related to dengue patient was reported in [5]. This research tried to classify the patients into 2 categories; dengue positive and dengue negative. The data collection was obtained in the surveys taken from different hospitals and diagnosis laboratories in India. Five thousand patients with 29 symptoms associated with the disease were collected. Support vector machines (SVM) was applied to classify the dengue positive and dengue negative patients. The total accuracy for training data for SVM was 90.3% in both cases.

The dengue outbreak model was study using Decision Tree, Artificial Neural Network, and Rough Set Theory classification [6]. A real-time system aimed for dengue diagnosis was done by [7] in order to minimize the number of false positives and false negatives. Tanner [8] employed decision tree in predicting

[+] Corresponding author. Tel.: +6681 9282461.
E-mail address: fscinws@ku.ac.th.

the outcome of the dengue fever in the early phase. The accuracy of the model produced is 84% which can differentiate dengue from non-dengue febrile illness. Moreover, a non-invasive prediction of the day of defervescence of fever in dengue patients using artificial neural network was studied by [9]

2. Data Mining Techniques

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 anattribute 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 [10] namely C4.5 uses Gain ratio to avoid the bias caused by attribute having larger number of values.

$$Gain(S, A) = Entropy(s) - \sum_{v \in Values(A)} \frac{|S_v|}{S} Entropy(S_v) \quad (1)$$

$$\text{GainRatio(S, A)} = \frac{Gain(S,A)}{SplitInfo(S,A)} \quad (2)$$

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 dataset.Where SplitInfo(S,A) is the information due to the split of S on the basis of the value of the categorical attribute *A*.

In data mining, Apriori is a classic algorithm for learning association rules. It is designed to operate on dataset containing records or transactions (for example, collections of items bought by customers).The main point of the algorithm is to extract useful information from large amounts of data. The algorithm aims to find the rules which satisfy both a minimum support threshold and a minimum confidence threshold.

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. 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.

3. Experimental Result

The total number of 1,001patients was obtained from Siriraj Hospital, Bangkok, Thailand. The dataset consists of 488 DF, 222 DHF I, 229 DHF II and 62 DHF III. The type ofattributes consists are clinical and hematological attributes. There are totally 48 attributes. The first set of attribute were collected on the first visit of the patient. The second set of attribute were obtained during the treatment period. Temporal attributes are summarized in terms of maximum, minimum and average values [11].Three commonly used performance measurements including sensitivity, specificity and accuracy are used in all experiments.

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

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

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

Result: Decision Tree and Fuzzy Logic approach were performed to see the classification performance as shown in Table 1.The experimental result shows that Fuzzy logic outperforms Decision tree with of accuracy of97.94%. The result obtains from decision tree (Figure 1) reveals the pattern for each of dengue fever severity as follows:1)If there is no evidence of plasma leakage then the patients should be classified as the Dengue fever (DF). 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 ≤ 40.33 U/L. 3) if the plasma leakage occurs, the bleeding is found, even the platelet count is less than the normal range (86,000cells/µl) but no shock evidence then the patients should be classified as the Dengue fever (DF). 4) if the plasma leakage occurs, the bleeding is found even the platelet count is less than the normal range (86,000cells/µl) and the white blood cell count is less than normal range (5,960cells/µl) but no evidence of lymphadenopathy is found then the patients should be classified as the Dengue fever (DF).

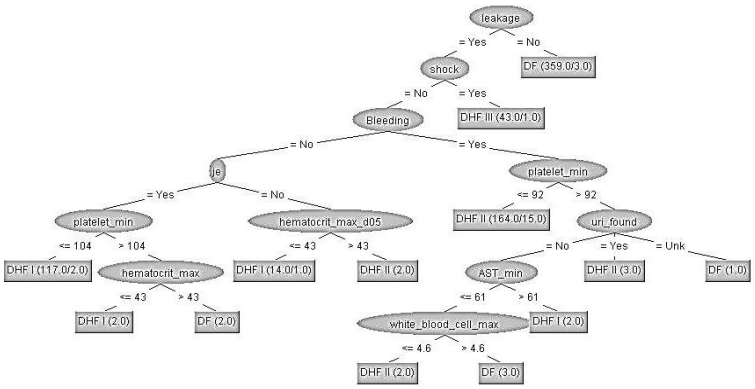

Fig. 1: Decision tree obtained from Experiment 2.

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 (≥40.33 U/L) or the number of platelet count is a bit less than normal range (≤111,000 cells/µl) then the patientwould be classified as the DHF I. However if the bleeding occurs for those patients and the number of white blood cells is less than the normal range (≤ 5,960 cells/µl) and no evidence of lymphadenopathy is found then the patientwould be classified as the DHF I as well. For DHF II, the evidence of plasma leakage and the bleeding occurrence are the main indicator. In case that thepatient has shocked then he/she will be classified as DHF III.

TABLE I: THE CLASSIFICATION PERFORMANCE OF DECISION TREE AND (FUZZY LOGIC)

Class	**Sensitivity** **DT (Fuzzy)**	**Specificity** **DT (Fuzzy)**	**Accuracy** **DT (Fuzzy)**
DF	97.34 (**97.75**)	98.56 (**99.39**)	97.95 (**98.57**)
DHF I	89.64 (**91.89**)	98.44 (98.44)	96.46 (**96.98**)
DHF II	95.63 (**96.51**)	96.84 (**97.25**)	96.56 (**97.08**)
DHF III	98.39 (**98.39**)	99.55 (**99.78**)	99.48 (**99.69**)

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

Result: The feature selection is done using decision tree then fuzzy logic areapplied to find the performance of the prediction. However the tree overfitting 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 2. The feature sets that can be used to identify 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.

TABLE II: PERFORMANCE OF DECISION TREE AND FUZZY LOGIC ON DAY0 PROBLEM

Class	**Sensitivity** **DT (Fuzzy)**	**Specificity** **DT (Fuzzy)**	**Accuracy** **DT (Fuzzy)**
day 0	**61.61** (50.86)	57.92 (**75.13**)	59.82 (**64.88**)
day -1	51.56 (**71.11**)	**54.66** (45.37)	53.43 (**54.75**)
day -2	23.08 (3.21)	85.74 (**96.92**)	71.11 (**81.82**)
day -3	10.00 (2.00)	96.91 (**99.67**)	88.79 (**94.63**)

Experiment 4: Does the hematomegaly is the indicator to differentiate the type of dengue?

Result: The decision tree obtained from experiment 2 is further analyzed to see the co-occurrence of hematomegaly and the degree of dengue severity. 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.

TABLE III: 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%

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

Result: The preprocessing is done to delete the patient's record in which the JE information is unknown, 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.

TABLE IV: THE CORRELATION ANALYSIS BETWEEN ATTRIBUTES AND CLASS (EXPERIMENT 5)

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

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 V: NUMBER OF PATIENTS WHO HAVE AST > ALT

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

4. Discussion

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 6).

TABLE VI: THE FALSE NEGATIVE VALUE OBTAINED FROM DECISION TREE AND WHO CRITERIA

class	**No. of patients**	**Decision Tree False Negative (%)**	**WHO False Negative(%)**
DF	488	2.66	1.64
DHF I	222	10.36	70.27
DHF II	229	4.37	86.46
DHF III	62	1.61	91.94

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.

5. Acknowledgements

This research was funded by KURDI, Kasetsart University. We would like to thank Dr. Prapat Suriyaphol from Bioinformatics and Data Management for Research Unit, Mahidol University for research cooperation.

6. References

[1] World Health Organization, 2015. Dengue Control, retrieved December, 5, 2015 from http://www.who.int/denguecontrol/human/en/

[2] D. Guha-Sapir and B. Shimmer, 2005. Dengue fever: new paradigms for a changing epidemiology. Emerging Themes in Epidemiology. (http://www.ete-online.com/content/2/1/1).

[3] J. Deen *et al.*, 2006. The WHO dengue classification and case definitions: time for a Reassessment. *Lancet*, 2006, 368, p. 170--173.

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

[5] Shameem Fathima, A. and Manimeglai, D., 2012. Predictive Analysis for the Arbovirus-Dengue using SVM Classification, *International Journal of Engineering and Technology*, Vol.2 No. 3, p. 521-527.

[6] D.A.Tarmizi*et.al*., 2013. Malaysia Dengue Outbreak Detection Using Data Mining Models, *Journal of Next Generation Information Technology*, Vol 4, No6, p. 96-107.

[7] V. Rao and M. Kumar, 2012. A new intelligence-based approach for computer-aided diagnosis of dengue fever.*Information Technology in Biomedicine*, Vol. 16, No. 1, 2012.

[8] L. Tanner *et al.*, 2008. Decision Tree Algorithms Predict the Diagnosis and Outcome of Dengue Fever in the Early Phase of Illness, *PLoSNegl. Trop. Dis.* 2, p.196.

[9] F. Ibrahim, *et al.* 2005. A novel dengue fever (DF) and dengue haemorrhagic fever (DHF) analysis using artificial neural network (ANN).*Computer methods and programs in biomedicine*, Vol. 79, No. 3, p. 273–281.

[10] J.R. Quinlan. Simplifying the decision tree. *International Journal of Man-Machine Studies*, Vol. 27, 1987, p. 221-234.

[11] D. Thitiprayoonwongse, P. Suriyapol and N. Soonthornphisaj, 2011. Data Mining on Dengue Virus Disease, *ICEIS* , p. 32-41