Journal of Fundamental and Applied Sciences

ISSN 1112-9867

Available online at http://www.jfas.info

Research Article

Special Issue

DENGUE FATALITY PREDICTION USING DATA MINING

N. F. Rahim, S. M. Taib* and A. I. Z. Abidin

Faculty of Science and Information Technology, UniversitiTeknologiPETRONAS, Bandar Seri Iskandar, Perak, Malaysia

Published online: 10 November 2017

ABSTRACT

Dengue fever, a mosquito-borne tropical disease caused by the dengue virus is life-threatening. In Malaysia, although necessary control measures have been carried out, the number of dengue fever cases keeps increasing. Among the measures, dengue vector control appears to be the most effective way to control the spread of the dengue virus particularly in Malaysia. The aim of this research is to study the current implementation of dengue outbreak control in Malaysia and predict dengue fever cases using data mining techniques. Real data on dengue fever and weather are collected from the Ministry of Health in its Perak Tengah district office and Perak Meteorological office respectively. Different data mining classification techniques are applied onto these data with the performance of each technique is measured. The results highlight the best performance among techniques used.

Keywords: data mining; prediction; dengue; classification.

Author Correspondence, e-mail: shakita@utp.edu.my

doi: http://dx.doi.org/10.4314/jfas.v9i6s.52

1. INTRODUCTION

Dengue fever (DF) and dengue fever (DHF) pose a major health concern in Malaysia. Since 1995, the number of dengue cases reported showed an increasing trend until 2002 except for a

slight drop in 1999-2001. There was a dramatic increase in the number of dengue cases from 2012 until 2016 as shown in Fig. 1. In 2012, a total of 21,900 cases were reported and this figure shot up to 120,836 cases in 2015. About 366 cases of dengue death were recorded in 2015 [1].

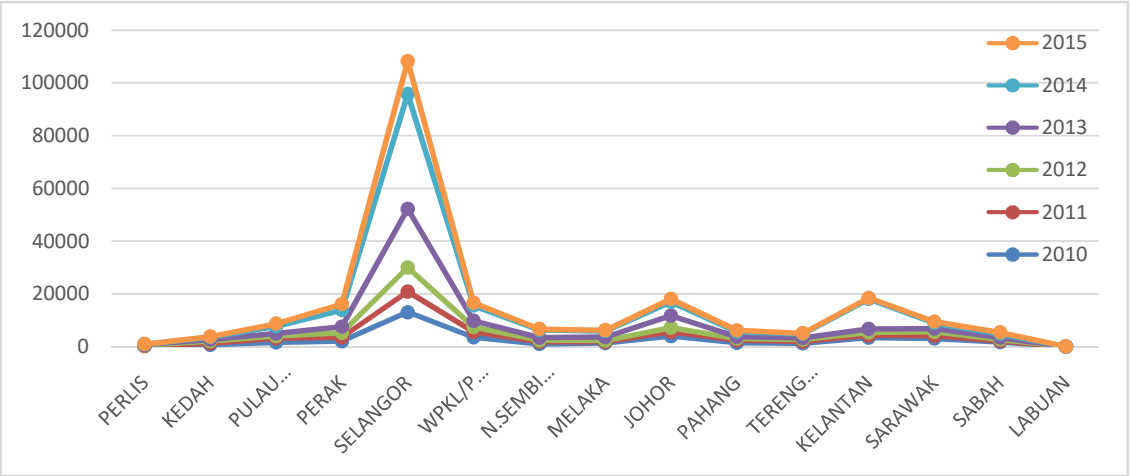

Fig.1. Malaysia dengue cases by state 2010-2015 [1]

From 2010 to 2015, Selangor is recorded as the highest with 108294 dengue cases were reported. From this number 169 fatality cases are recorded. Fig. 2 shows number of cases versus number of fatality from 2012 to 2015. The higher number of cases, the higher number of fatality is reported.

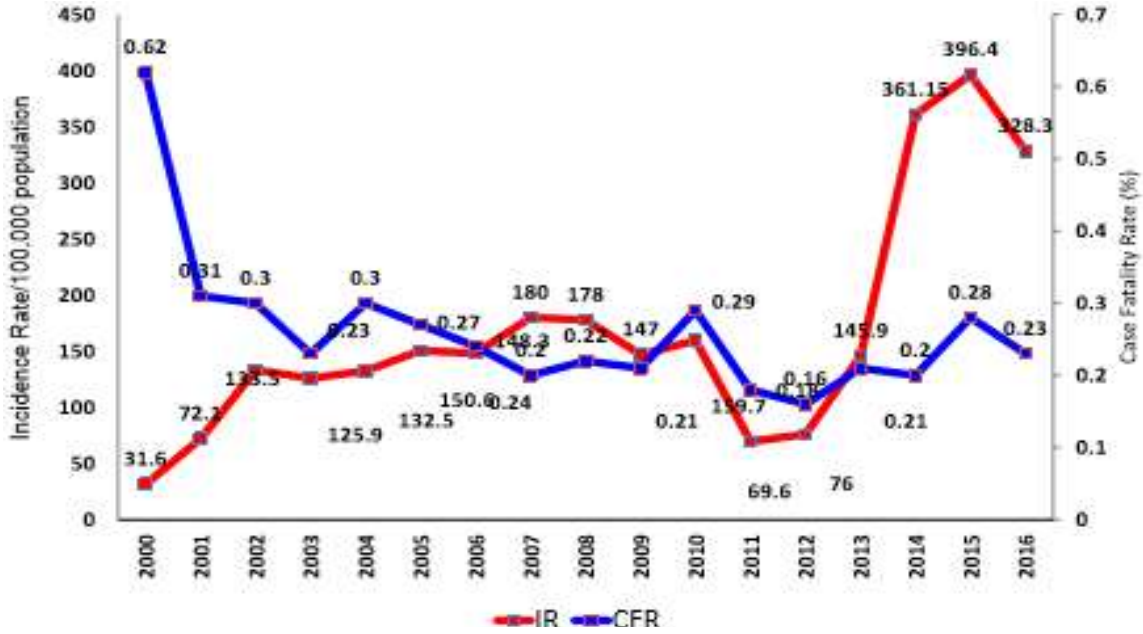

Fig.2. Dengue incidence and case fatality rate [1]

Despite the increase in DF and DHF cases recorded, no vaccine, medicine or control measure proven to be effective in tackling the outbreak. In Malaysia, dengue vector control is still considered the most effective way to control and prevent the transmission of DF and DHF virus. Malaysia's current dengue vector relies greatly on thermal or ULV fogging.However, the approach falls short of expectation [2].

Many researches from various studies using different techniques were done to analyze the best and effective way in controlling and managing dengue cases. The objective of this paper is to predict and analyze different classification techniques to determine the outbreak of dengue fever especially in Malaysia's Perak state based on fatality record. The performance of different classification techniques is being compared. Difference data mining techniques are applied onto dengue and weather data parameter.

2. BACKGROUND OF STUDY

The increase in dengue cases reported in Malaysia is very threatening. At present, there is yet a specific antiviral drug for DF and DHF treatment. Even though the World Health Organization (WHO) has developed dengue vaccine, the vaccine introduced was too complicated [3]. This happened due to several concerns such as high risk of severe disease through antibody-dependent enhancement [4] and the possible threats of new serotype, DENV-5[5].

Since actions will only be taken once the cases are reported, the chances to spread the DF and DHF virus are high. This study aims to have a preventive action before some DF and DHF cases actually occur. A research done by [6] stated that that the current indices used for dengue are not sensitive and accurate to forecast the outbreaks. In this study, weather data parameter will be used to perform the prediction analysis using different data mining techniques.

2.1. Factors of Dengue Cases

The AedesAegyptus mosquito is the primary vector of dengue. This virus is transmitted to humans through the bites of infected female mosquitoes [7]. After virus incubation stage which normally takes around 4-10 days, an infected mosquito is capable of transmitting the virus for the rest of its life.Human host or infected symptomatic or asymptomatic human also

known as the main carrier and multiplier of the virus which is a source of the virus form uninfected mosquito [7]. Infected individuals can transmit the infection for 4-5 days and maximum of 12 days via Aedes mosquito after the first symptoms detected.The occurrence of epidemic peak also contributed to the other key factors of outbreak transmission. Changes in climate such as temperature, increased rainfall and relative humidity were also highlighted as the most influential driving forces of dengue transmission [8]. A study by [9] of Selangor in Malaysia, in [10] of Cambodia and by [11] of Taiwan concluded that weather is an effective predictor for dengue cases and outbreak. Based on the previous studies done, it can be concluded that the pattern of dengue transmission is influenced by a very complex factor. There are at least five majors highlighted in influencing the transmission which include the dengue virus, human as host, environment condition such as cleanliness, the vectors and its behavior and changes in climate [8].

2.2. Current Preventive Action for Dengue

As the current practice in Malaysia, strategy for dengue prevention and control were focused on five elements [12, 13]. Such elements are:

- Vector control which based on the principles of integrated vector management
- Active disease surveillance based on a comprehensive health information system
- Emergency preparedness
- Capacity building and training
- Vector control research

Despite the threatening status of dengue cases, only vector control is effective to reduce or prevent the dengue virus transmission [3]. For the current practice, both larva and adult mosquito are under the vector control program. In larva control programs, the strategies involved environmental management, source reduction, larvicides usage such as temephos (Abate), house and environment inspection and enforcement of Destruction of Disease-Bearing Insect Act 1975. While for adult mosquito control, fogging will be carried out once viral cases are reported [7].

In handling and controlling the outbreak, urbanization, population growth and human behavior are factors that cause control of dengue transmission even more difficult. Hence, in 2011, the Ministry of Health, Malaysia implemented its integrated strategy for dengue

prevention and control program under the National Dengue Strategic Plan (NDSP). There are several strategies highlighted including dengue surveillance, National Cleanliness Policy and Integrated Vector Management (IVM), Social Mobilization and Communication for Dengue and Dengue Research [1].

From WHO perspective [3], at present, the main method to control or prevent the transmission of dengue virus is to combat vector mosquitoes through environmental management, improved community participation and mobilization for sustained vector control, emergency vector-control measures during outbreaks and active monitoring and surveillance to determine effectiveness of control interventions.

3. DATA MINING IN HEALTHCARE

Data Mining (DM) application in healthcare is still new compared to in business and marketing. According to [14], there are several factors that have motivated the use of DM application in healthcare. They highlighted the existence of medical insurance fraud and abuse, led healthcare insurers to attempt in reducing their losses. Besides, the growth and massive amounts of data collected and generated from healthcare transactions also contribute in the application of DM in healthcare. Difficulty in processing and analyzed these complex and voluminous data has force healthcare related organization and practitioner investing in DM application. Apart from that, in [15] in his study stated that successful mining applications have been implemented in the healthcare area can be classified through three distinct perspectivesnamely Hospital Infection Control, Ranking Hospitals and Identifying High Risk Patients. The developed systems are able to enhance infection control, strengthen the identification of high-risk patient and reduce the cost of treatments.

3.1. Data Mining for Dengue Analysis

A number of previous works have proposed data mining to predict dengue disease or discover the patterns either from dengue patient data or demographic dengue data. The dengue patient data includes clinical, symptoms and features of dengue infection. While the demographic data contains demographic features such as epidemic, incubation period, type of outbreak, repetition case and fatality. Table 1 shows the description of related previous works that performed different types of data mining techniques in their prediction.

Data mining techniques were applied to predict dengue disease and fever, to predict the dengue risk or to classify dengue outbreak and dengue infection. Most of the dataset used in previous works are from dengue patients' data collected from hospitals and health authority. Some of researcher choose one data mining technique but modify the algorithm to fit with dengue data as well as to achieve better prediction accuracy. However, there are researchers that proposed hybrid model that combines different types of data mining techniques. Based on the results from the previous works, combining different techniques can improve prediction/classification accuracy compared to a single technique.

Table 1. Data mining for dengue

Research	Technique	Data	Result
Dengue Infection Classification [16]	Decision Tree	Clinical data	Due pooraccuracy, this model is not suitable to predict the abatement of a fever case.
Dengue Fever Prediction [17]	Classification	Features of infection	Naïve Bayes and J48 are thetop performance classifier techniques
Dengue Disease Prediction [18]	Decision Tree, Support Vector Machine (SVM), Fisher Filtering	Clinical Data	The decision tree model was effectively by ranking the input attributes according to their relevance.
Dengue Outbreak Classification [19]	Hybrid model – Decision Tree, Neural Network and Rough Set	Demographic Dengue data	The hybrid model achieves high accuracy and generates interesting and interpretable rules. Thus it outperformed other models.
Predicting the risk in	Self-organizing	Dengue patient data	The prediction accuracy

dengue patients [20]	map and Neural Network	of the model needs to be improved.

4. METHODS

Predictive methods are often used when the attributes can be subdivided into two groups: input and output attributes. Predictive algorithms are used togenerate models or rules to predict continuous or discrete target values given input data [16]. Two types of prediction can be recognized are: classification and regression. Typical algorithms of classification include neural networks, decision trees, instance based learning, etc. While for regression models, it transforms the space of input attributes into a real-valued domain [21].

For this study, an open source data mining tool with visual programming and python as scripting is used for testing and execution tool. Six techniques of classification are used in this study to predict the dengue outbreak and fatality. The six techniques aredescribed in Table 2.

Table 2. Description of data mining techniques

Techniques	Description
Naïve Bayes	Fast to train for single scan, fast to classify, not sensitive to irrelevant features, can handles real and discrete data [22].
Decision Tree (ID3)	A decision structure consists of nodes that represent the tested attributes and the outgoing branches correspond to all the possible outcome of the tested nodes. Suitable for predictive and descriptive model [23].
Logistic Regression	Generates coefficients of a formula to predict a logit transformation of the probability of presence of the tested attributes [24].
Support Vector Machine	A model that finds a hyperplane to divide a dataset into two classes. It can be employed for both classification and regression [25].
CN2	It uses entropy as it searches heuristicsto generate an ordered list of rules. It is designed based on the ideas of ID3 and

	quasi-optimal (AQ) algorithms [26].
Random Forest	A collective learning technique that generates many single learners by creating a set of random data for building an ID3. Each node is split using a best amongst the subset of predicators randomly chosen at that node [27].

5. DATASET

In this study, sample data set of historical dengue cases and weather data set are collected from the available and relevant online sources such as data.gov.my which is provided by Malaysian Administrative Modernization and Management Planning Unit (MAMPU). Weather parameter data of temperature, maximum temperature, minimum temperature, rainy days and amount of rainfall for Perak state is collected directly from Malaysian Meteorological Department. As shown in Table 1, the collected data has almost 312 entries of dengue cases occurred in Perak state from 2010 to 2015. Information also obtained from reliable webpage such as Ministry of Health (MOH) [28]and dengue information portal [1].

Spatio-temporal predictions are used in this study. All the selected variables need to fit the same spatio-temporal scale. Spatio-temporal data can be defined as an object that has at least one spatial and one temporal property. Whereas the spatial property refers to location and geometry of the object, temporal property is the time stamp or time interval for which the object is valid [29].The spatio-temporal scale used in this work was selected based on the distribution of the dengue data from 2010 to 2015: the chosen temporal scale was one week and the chosen spatial distribution was one state, Perak. Table 3 shows the sample data collected from selected sources and accuracy is measured based on testing dataset.

Table 3. Sample data for dengue analysis

Week	Cases	Rainfalls	Max Temp	Min Temp	Mean Temp	Fatality
week40	133	427.2	32.374	23.774	26.768	0
week41	105	223.081	32.854	23.953	27.319	1
week42	116	223.081	32.854	23.953	27.319	2
week43	115	223.081	32.854	23.953	27.319	1
week44	111	464.4	32.067	23.713	26.71	0
week45	130	223.081	32.854	23.953	27.319	0
week46	237	223.081	32.854	23.953	27.319	1

6. RESULTS AND DISCUSSION

The accuracy of each prediction model is calculated based on a measure of a predictive model that represents the proportionate number of times that the model is correct when applied to data. Training dataset are verified in Orange Data Mining tool with NB, D3, SVM, CN2, RF and LR Technique where the outcome is summarized in Table 4.

Table 4. Comparison table

Techniques	Error Rate	Accuracy
Naïve Bayes	0.28	0.72
Decision Tree	0.06	0.94
Logistic Regression	0.07	0.93
Support Vector Machine	0.07	0.93
CN2	0.04	0.96
Random Forest	0.05	0.95

The highest accuracy is 0.96 measured from CN2, while the lowest accuracy is 0.72 using Naïve Bayes. The other four classification techniques yield similar accuracy between 0.93 and 0.95. Based on the comparison, it can be concluded that CN2 techniques is the best techniques. The rules from CN2 technique is extracted for dengue fatality and outbreak prediction. The extracted rules are represented in Fig. 3.

From the extracted rule result, it can be concluded that the high number of dengue cases recorded have the medium rate potential in having fatality case. However, this situation also affected by number of rainfalls received at particular time. Based on the rules, it can be concluded that if rainfall is frequent, the potential of having fatality is still influenced by the number of dengue cases recorded. If the number of dengue cases is low and rainfall recorded is high, no fatality is predicted. It also highlighted that number of minimum temperature also has significance impact on rate of fatality predicted.

IF cases>=249.0 THEN Fatality=med
IF cases<=47.0 AND rainfalls>=241.8 THEN Fatality=no
IF mean temp>=27.319 AND max temp>=32.948 THEN Fatality=no
IF max temp>=33.141 THEN Fatality=low
IF min temp>=23.96 THEN Fatality=low
IF mean temp>=27.319 AND mean temp>=27.35 THEN Fatality=no
IF cases<=66.0 AND cases>=66.0 THEN Fatality=low
IF rainfalls<=227.8 AND rainfalls>=227.8 THEN Fatality=low
IF rainfalls<=358.4 AND rainfalls>=358.4 THEN Fatality=low
IF rainfalls>=223.081 AND rainfalls>=223.4 THEN Fatality=no
IF cases<=47.0 AND rainfalls>=223.081 AND cases>=47.0 THEN Fatality=low
IF cases<=78.0 THEN Fatality=no
IF cases>=108.0 THEN Fatality=no
IF TRUE THEN Fatality=low

Fig.3. CN2 rules

7. CONCLUSION

Results show the performance of CN2 is more accurate compared to other techniques. Prediction capacity of CN2 by extracting the rules can help health organization to perform advanced prediction, thus assisting in developing an early warning system and prevention program based on weather data. The association between dengue and weather data varies by locality, thus an early warning system is best implemented locally, e.g. a system per district. Further study on the weather and the population of the dengue mosquito using local ovitrap index that records the collection of eggs laid by the mosquitoes is recommended. This approach has a potential in developing a much accurate model to predict dengue fatality and outbreak.

8. ACKNOWLEDGEMENTS

We thank Dr. Nor Samsiah, Perak Tengah Ministry of Health Officer for assistance with the dengue case surveillance data and helpful insight discussions and Ms. Noor Azura Ismail, Officer, National Climate Centre, Malaysian Meteorological Department for the required meteorological data.

9. REFERENCES

[1] iDengue. Malaysian dengue information. Kuala Lumpur: Malaysian Remote Sensing Agency, Ministry of Science, Technology and Innovation and Ministry of Health Malaysia, 2017

[2] Song Q O. Dengue vector control in Malaysia: A review for current and alternative strategies. SainsMalaysiana, 2016, 45(5):777-785

[3] World Health Organization (WHO). Global strategy for dengue prevention and control 2012-2020. Geneva: WHO, 2012

[4] Webster D P, Farrar J, Rowland J S. Progress towards a dengue vaccine. The Lancet Infectious Diseases, 2009, 9(11):678-687

[5] Paul B, Tham WL. Controlling dengue: Effectiveness of biological control and vaccine in reducing the prevalence of dengue infection in endemic areas. Health, 2016, 8(1):64-74

[6] Tham AS. Issue and challenges inAedes surveillance and control. In Workshop Proceeding Behavior Intervention in Dengue Control of Malaysia, 2000, pp. 15-23

[7] World Health Organization (WHO). Dengue: Call for urgent interventions for a rapidly expanding emerging disease. Technical paper, Geneva: WHO, 2011

[8] Mudin RN. Dengue incidence and the prevention and control program in Malaysia. International Medical Journal Malaysia, 2015, 14(1):5-10

[9] Cheong YL, Burkart K, Leitão, PJ, Lakes T. Assessing weather effects on dengue disease in Malaysia. International Journal of Environmental Research and Public Health, 2013, 10(12):6319-6334

[10] Choi Y, Tang CS, McIver L, Hashizume M, Chan V, Abeyasinghe RR, Huy R. Effects of weather factors on dengue fever incidence and implications for interventions in Cambodia. BMC Public Health, 2016, 16(1):1-7

[11] Wu PC, Guo HR, Lung SC, Lin CY, Su HJ. Weather as an effective predictor for occurrence of dengue fever in Taiwan. ActaTropica, 2007, 103(1):50-57

[12] Guzman MG, Halstead SB, Artsob H, Buchy P, Farrar J, Gubler DJ, Hunsperger E, Kroeger A, Margolis HS, Martínez E, Nathan MB, Pelegrino JL, Simmons C, Yoksan S, Peeling RW. Dengue: A continuing global threat. Nature Reviews Microbiology, 2010, 8(12):7-16

[13] Zahari CD. Workshop in proceeding on behaviour intervention in dengue control in Malaysia. Pulau Pinang: Centre for Drug and research and School of Communication, UniversitiSains Malaysia, 2001

[14] Koh H C, Tan G. Data mining applications in healthcare. Journal of healthcare information Management, 2011, 19(2):64-72

[15] Obenshain M K. Application of data mining techniques to healthcare data. Infection Control and Hospital Epidemiology, 2004, 25(8):690-695

[16] Thitiprayoonwongse D A, Suriyaphol P R, Soonthornphisaj N U. Data mining of dengue infection using decision tree. In Latest Advances in Information Science and Applications, 2012, pp. 154-159

[17] Shaukat K, Masood N, Mehreen S, Azmeen U. Dengue fever prediction: A data mining problem. Journal of Data Mining in Genomics and Proteomics, 2015, 6(3):1-5

[18] Arun K P M, Chitra DB, Karthick P, Ganesan M, Madhan AS. Dengue disease prediction using decision tree and support vector machine. SSRG International Journal of Computer Science and Engineering, 2017, 2017(Special Issue March):60-63

[19] Tarmizi N D, Jamaluddin F, Bakar A A, Othman Z A, Hamdan A R. Classification of dengue outbreak using data mining models. Research Notes in Information Science, 2013, 12:71-75

[20] Faisal T, Ibrahim F, Taib M N. A noninvasive intelligent approach for predicting the risk in dengue patients. Expert Systems with Applications, 2010, 37(3):2175-2181

[21] Zaitseva E, Kvassay M, Levashenko V, Kostolny J. Introduction to knowledge discovery in medical databases and use of reliability analysis in data mining. In IEEE Federated Conference on Computer Science and Information Systems, 2015, pp. 311-320

[22] Hemalatha I, Varma G S, Govardhan A. Sentiment analysis tool using machine learning

algorithms. International Journal of Emerging Trends and Technology in Computer Science, 2013, 2(2):105-109

[23] Quinlan JR. Induction of decision tree. Machine Learning, 1986, 1(1):81-106

[24] Strano M, Colosimo BM. Logistic regression analysis for experimental determination of forming limit diagrams. International Journal of Machine Tools and Manufacture, 2006, 46(6):673-682

[25] Cortes C. Support vector networks. Machine Learning, 1995, 20(3):273-297

[26] Clark P, Niblett T. The CN2 induction algorithm. Machine Learning, 1989, 3(4):261-283

[27] Breiman L. Random forests. Machine Learning, 2001, 45(1):5-32

[28] Ministry of Health (MoH). Official portal-Ministry of Health Malaysia. Putrajaya: MoH, 2017

[29] Rao K V, Govardhan A, Rao K V C. Spatiotemporal data mining: Issues, tasks and applications. International Journal of Computer Science and Engineering Survey, 2012, 3(1):39-52

How to cite this article:
Rahim N F, Taib S M, Abidin A I Z. Dengue fatality prediction using data mining. J. Fundam. Appl. Sci., 2017, *9(6S), 671-683.*