Sains Malaysiana 46(2)(2017): 255–265
http://dx.doi.org/10.17576/jsm-2017-4602-10

Feature Selection Algorithms for Malaysian Dengue Outbreak Detection Model

(Pemilihan Ciri Algoritma untuk Model Pengesanan Wabak Denggi)

HUSAM, I.S., ABUHAMAD, AZURALIZA ABU BAKAR, SUHAILA ZAINUDIN*,
MAZRURA SAHANI & ZAINUDIN MOHD ALI

ABSTRACT

Dengue fever is considered as one of the most common mosquito borne diseases worldwide. Dengue outbreak detection can be very useful in terms of practical efforts to overcome the rapid spread of the disease by providing the knowledge to predict the next outbreak occurrence. Many studies have been conducted to model and predict dengue outbreak using different data mining techniques. This research aimed to identify the best features that lead to better predictive accuracy of dengue outbreaks using three different feature selection algorithms; particle swarm optimization (PSO), genetic algorithm (GA) and rank search (RS). Based on the selected features, three predictive modeling techniques (J48, DTNB and Naive Bayes) were applied for dengue outbreak detection. The dataset used in this research was obtained from the Public Health Department, Seremban, Negeri Sembilan, Malaysia. The experimental results showed that the predictive accuracy was improved by applying feature selection process before the predictive modeling process. The study also showed the set of features to represent dengue outbreak detection for Malaysian health agencies.

Keywords: Feature selection; dengue outbreak; knowledge discovery from databases; nature-based algorithms; outbreak detection

ABSTRAK

Demam denggi merupakan penyakit bawaan nyamuk yang wujud di merata dunia. Pengesanan wabak denggi bermanfaat sebagai satu usaha praktikal mengawal penyebaran penyakit ini dengan menyediakan pengetahuan untuk meramal kejadian wabak yang seterusnya. Penyelidikan lepas telah dijalankan untuk memodel dan meramal pengesanan wabak denggi menggunakan pelbagai teknik perlombongan data. Penyelidikan ini bertujuan untuk mengenal pasti ciri yang meningkatkan ketepatan ramalan wabak denggi menggunakan tiga algoritma pemilihan ciri; particle swarm optimization *(PSO),* genetic algorithm *(GA) dan* rank search *(RS). Berdasarkan ciri yang dipilih, tiga teknik permodelan ramalan (J48, DTNB dan Naive Bayes) dijalankan untuk peramalan wabak denggi. Set data yang digunakan dalam penyelidikan ini diperoleh dari Jabatan Kesihatan Awam, Negeri Sembilan, Malaysia. Keputusan kajian menunjukkan bahawa ketepatan ramalan meningkat apabila proses pemilihan ciri dijalankan sebelum proses permodelan. Kajian ini turut menghasilkan set ciri baru untuk mewakilkan pengesanan wabak denggi untuk agensi berkaitan kesihatan di Malaysia.*

Kata kunci: Algoritma berasaskan alam; pemilihan ciri; penemuan ilmu dari pangkalan data; pengawalan wabak; wabak denggi

INTRODUCTION

Dengue fever is considered as one of the most common and life-threatening diseases worldwide (Edelman 2007). Guy (2008) indicated that 50 to 100 million people are infected per year worldwide by dengue fever with half a million life threatening cases. Dengue fever and dengue hemorrhagic fever is a serious arboviral infection spread by *Aedes aegypti* and *Aedes albopictus* mosquitoes (WHO 2010). The first dengue fever case in Malaysia was reported in 1902 (Skae 1902) in Pulau Pinang. Earlier records documented that dengue cases were first reported in Asia in 1770 (Beltz 2011). The rapid outbreak of dengue is a serious problem. Hombach (2007) cautioned that more than 2.5 billion people are at high risk of infection in more than 100 endemic countries throughout the tropical and subtropical climate zones. Therefore, this study proposed data mining approach to study dengue fever time-series transition data to discover the relations and the factors that might halt the rapid spread of this disease.

Many researchers have identified related factors to predict the next dengue outbreak. The early detection of dengue outbreak has been given the highest priority in previous works with the predictive accuracy achieved in the range of 70-80% (Long et al. 2010; Mousavi et al. 2011). Past research attempted to model the problem based on the historical data of dengue transitions using different classification techniques in order to build dengue outbreak predictive models. This research harness existing feature selection algorithms to explore combinations of features that contain interesting knowledge for outbreak detection. To date, not many researches has considered selecting the best features (attributes) and remove the irrelevant

and redundant features by employing feature selection algorithms. Most research in feature selection concluded that irrelevant and redundant features could negatively affect the accuracy of any model. Hence, conducting feature selection process before the predictive modeling task can enhance the accuracy of the model and reduce computational and memory demands. Identifying the major features of dengue outbreak would not only help to achieve better prediction accuracy but also help in understanding the root of the problem.

Related Work

Dengue Outbreak Detection

Understanding time-series transmission data of dengue fever will lead to prediction of dengue outbreaks in the very early stage. Therefore, the relevant authorities can take swift actions to prevent its spread. Research is not confined to only predicting dengue outbreak but also identifying factors that contribute to better prediction of its occurrences. Andrick et al. (1997) included different time-lags of weather features into their model to examine the importance of these factors in the feature selection process. This research shows a real link between dengue outbreaks and weather features. Fu et al. (2007) considered climactic factors such as rainfall, humidity and temperature with time lags in a support vector machine (SVM) to model dengue fever outbreak. Early detection of infectious dengue patients is important to prevent local transmission in areas where the vector is present and active. Dengue outbreak can spread rapidly thus leading to high morbidity and mortality rates. It is important to detect an outbreak as early as possible and control the spread of disease among the population at risk.

Outbreaks can be defined as the occurrence of any disease cases higher than is expected in a particular area during specific period of time. Barbazan et al. (2002) and Runge-Ranzinger et al. (2008) defined dengue outbreak as the occurrence of dengue cases one significant deviation (SD) more than the average. Talarmin et al. (2000) defined dengue outbreak as number of reported cases with two SD above the baseline for the period of non-epidemic week. Using geospatial modeling, Seng et al. (2005) define dengue outbreak as the occurrence of more than one case in a specific geographical area where the onset date between cases is less than 14 days. Adopting the Apriori frequent mining approach, Long et al. (2010) concluded that high volumes of records are not critical for outbreak detection since the complexity of the existing attribute values can determine the potential dengue outbreaks. The Control and Prevention of Vector Borne Diseases Program, Ministry of Health Malaysia adopts the operational definition for outbreak of dengue as the occurrence of more than one case of dengue in same locality, within the same incubation period of the first case or index case of dengue notified to the authorities. This definition is used for the District Health Office to implement control measures immediately, in order to prevent further transmission of the dengue virus or local transmission of dengue virus in the same locality of outbreak (Public Health Department, Malaysia).

Outbreak detection can be considered as a deviation detection or classification problem. Past research has shown that subset of features could represent the data more efficiently and leads to better understanding (Vainer et al. 2011). Datasets with irrelevant and redundant information could lead to a complicated learning process and less accurate results (Tuv et al. 2009). Feature selection process facilitates at understanding the correlation among attributes and defines the most related attributes to a given problem or task.

Outbreak detection or predictive modeling can be done without any prior feature selection. For example, Buckeridge et al. (2005) suggested using previous data from the surveillance system to build outbreak detectors based on different classification method. Husin and Salim (2008) employ artificial neural networks with back propagation algorithm and non-linear regression models to predict the next dengue outbreak for 2 dengue datasets and rainfall dataset with variation in terms of time and location. Bakar et al. (2011) proposes a predictive model based on multiple rule-based classifiers to detect dengue outbreak using a dataset containing 8505 dengue patient records with 134 attributes. Moreover, Long et al. (2010) present an interesting study of pattern mining in outbreak detection using Apriori model which shows promising results in this domain.

Nevertheless, feature selection has the potential to improve classification result as proven by Mousavi et al. (2013). Mousavi et al. (2013) proposed negative selection algorithm, a variant of artificial immune system (AIS) for detecting dengue outbreak and resulted in 79.85% accuracy for 8505 dengue patient records with 12 attributes. This work aimed to find the most related attributes and features which lead to better dengue outbreak detection by applying feature selection process. The necessity of applying feature selection process is high since the patient dataset contains several irrelevant attributes. This work has contributed to better dengue outbreak detection by adding seasonal data that produced a model with large dimensionality. In the Malaysian context, the seasonal data is data related to the annual rainy season. The idea to include data related to weather such as rainfall, humidity and temperature has also been suggested by Husin and Salim (2008) as a possible extension of their research.

Feature Selection

Feature selection process is described as the process of detecting relevant features and removing irrelevant features in order to best represent the data. A proper representation of data can enhance the inductive learner, not only in the speed but also in the generalization capacity and the simplicity of the induced model. A reduced number of features also means less measurement cost, less memory

space and possibly better understanding of the domain. There are several issues that can restrain feature selection process, such as the existence of noise in data collection, related and interacted features, multiple class labels and number of samples (Saari et al. 2011). Feature selection has been applied successfully in several real-world applications, such as intrusion detection (Bolón-Canedo et al. 2011), information retrieval (Egozi et al. 2008) and text categorization (Gomez et al. 2012). These research proved the benefits that feature selection can bring to the learning and modeling process (Guyon & Elisseeff 2003). Researchers agree that the efforts in feature selection must be focused on investigating and determining the proper method in specific problem domain. Zhang et al. (2008) apply reliefF and mRMR as a combination of different algorithms from the same approach (filters), whereas El Akadi et al. (2011) apply mRMR filter and GA wrapper using filters and wrappers approaches. On the other hand, other works combined feature selection with other processes, such as combining spatial dimension reduction, spatio-temporal features extraction, and feature selection for hurricane severity classification (Vainer et al. 2011). Tuv et al. (2009) proposed feature selection with ensembles, artificial variables and redundancy elimination. Moreover, some studies have adopted other strategies, such as Sun et al. (2006) who proposed reinterpreting existing algorithms to be utilized as feature selection. Chidlovskii and Lecerf (2008) and Loscalzo et al. (2009) proposed new feature selection methods to handle unsolved problems such as multi-classification problem and identifying intrinsic feature groups from a small training set. A collection of feature selection techniques can be applied for the same problem to ensure a better result as shown in Bolón-Canedo et al. (2012). All these methods address the feature selection problem from different aspects.

Liu and Yu (2005) provide a survey of feature selection techniques and some guidelines of selecting the proper feature selection technique, presenting an outstanding way for building an intelligent feature selection system. On the experimental perspective, many studies investigated different feature selection methods, such as the studies presented in Molina et al. (2002).

Partial discharge (PD) measurements is used for assessing the integrity of insulation systems. Sharkawy et al. (2011) applied particle swarm optimization (PSO) technique to select features from measured PD dataset. The fused features from both acoustic and electrical signals were used to classify the particles in the oil, giving an overall classification accuracy of 100%. In the PSO algorithm, each individual is called a 'particle' and it is subject to a movement in a multidimensional space that represents the search space. Particles have memory to retain their previous states. There is no restriction for particles to share the same point in the search space yet in any case, their individuality is preserved. Each particle's movement is the composition of an initial random velocity and two randomly weighted influences which are: The individuality, the tendency to return to the particle's best previous position; and the sociality, the tendency to move towards the neighborhood's best previous position.

GA is inspired by the process of natural selection. GA begins with a sample set of potential solutions which then evolves toward a set of more optimal solutions. Within the sample set, solutions that are poor tend to die out while better solutions evolve and propagate their advantageous traits to introduce more solutions into the set with greater potential. The total set size remains constant since the old solutions are replaced with new ones. The random mutation process guarantees that a set would not stagnate. GA tends to work better than traditional optimization algorithms because it is less prone to local optima. GA does not make use of single-point transition rules to move from one single instance in the solution space to another. Instead, GA takes advantage of an entire set of solutions spread throughout the solution space, all of which are experimenting upon many potential optima. Wu et al. (2009) proposed a wavelet transformation for data preprocessing before employing a support vector machines (SVM)-based genetic algorithm to select the most important features to predict dengue outbreak. After which, regression based on SVM was used to perform forecasting of the model. The performance of the proposed method shows that MSE is minimal (< 0.01) and the correlation coefficient is within 2% of each other. These empirical results suggest that the proposed model is capable of producing relatively reliable predictions for up to 2 years ahead and able to construct a stable model instance using only 5 years of data.

Unlike PSO and GA, RS performs evaluations on single attributes instead of an explicit best feature subset and the cutoff point in the ranking is chosen via cross-validation process. Que and Tsui (2011) proposed a spatial clustering algorithm, rank-based spatial clustering (RSC) that detects rapidly infectious but non-contagious disease outbreaks. RSC consistently outperforms the other algorithms in terms of detection timeliness while having comparable detection powers.

METHODS

Not all features are related to the outbreak problem so they might negatively affect the accuracy of the predictive modeling task. The time-series dengue dataset used in this research includes clinical and weather factors. This work aimsed to apply feature selection process to dengue data to determine the most relevant and related features to dengue outbreak and to investigate the impact of applying feature selection process on the predictive accuracy of dengue outbreak detection.

OUTBREAK DETECTION METHODS

Due to the fact that irrelevant features negatively affect the prediction accuracy of dengue outbreak, this research proposes a feature selection process before modeling the problem. Using dengue time-series transition data, several feature selection techniques can be applied to select the

best features in order to increase the predictive accuracy of the detection model. Identifying the major features of dengue outbreak will help to achieve better accuracy and help health authorities to understand the problem better. To achieve this purpose, two experiments are conducted (Figure 1). Experiment I deals with feature selection process where feature subsets are evaluated based on the accuracy achieved by artificial neural network (ANN) model to detect dengue outbreaks. Whereas, experiment II deals with the predictive modeling task of dengue outbreak detection based on three predictive modeling techniques namely decision tree (J48), naïve bayes decision tree (DTNB) and ANN (Kumar et al. 2005).

DATASET DESCRIPTION

The dengue data used in this research is obtained from the Seremban District Health Office, Negeri Sembilan, Malaysia. This data includes records of 6082 dengue cases with 20 attributes (Table 1). The data covers 7 years of patients' records from 2003 until 2010 accompanied with local rainfall data as suggested by experts from the Health Office. The rainfall data were maximum temperature value, minimum temperature value, average temperature value, humidity value, rainfall and month. The seasonal data and the patients' data were matched by year and week. Seasonal data were also provided by the Public Health Department, Ministry of Health, Negeri Sembilan.

The attributes of this dataset are shown as follows: year, week of the year, cumulative week from year 2003 till year 2010, number of dengue fever for current week, number of dengue hemorrhagic fever for current week, total number of cases in current week, maximum temperature value, minimum temperature value, average temperature value, humidity value, rainfall, month, age, sex, race, work, address, district office in charge, district and outbreak. Outbreak has 3 possible values; TKW (Not in outbreak area), DKW (Near to outbreak area) and MWB (Might in outbreak). These indicate the location of the patient whether the patient is, near or maybe in an outbreak area.

FEATURE SELECTION FOR DENGUE DATA

In this work, all the data pre-processing work has been done separately using statistical and expert verification. Then the dataset is used for feature selection and predictive modeling tasks. This experiment adopts wrapper feature selection approach using ANN as feature subsets evaluator in terms of dengue outbreak detection accuracy. Three different feature selection algorithms are applied for this task to examine the selected features and their impact on the accuracy of problem modeling. This phase applied three feature selection techniques to select the best feature in dengue dataset in terms of dengue outbreak detection accuracy. PSO, GA and RS were applied to select the best features based on the accuracy of ANN model built for dengue outbreak detection. The feature selection process is shown in Figure 2. ANN algorithm is used to validate the performance of the feature selection algorithms. PSO, GA and RS were employed to select the features from the dataset. Then, the selected features from each algorithm

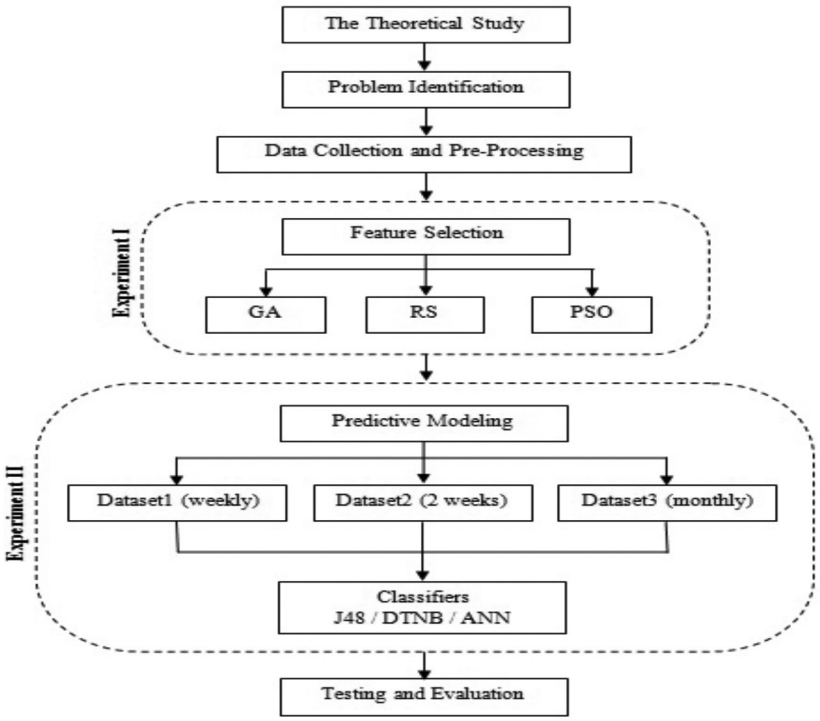

FIGURE 1. The Methods

TABLE 1. View from the used dengue dataset

YEAR	WEEK	ACC_ WEEK	DF	DHF	CASE	TEMP_ MAX	TEMP_ MIN	TEMP_AVE	HUMID	RAINFALL	Month	Age	Gender	Race	Work	Address	Office	District	Outbreak
2003	1	1	100	9	109	31.5125	22.45	25.72395833	77.25208333	50.92857143	JAN	MIDDLE	L	INDIA	STUDENT	AMPANGAN	MPS	SN	TKW
2003	1	1	100	9	109	31.5125	22.45	25.72395833	77.25208333	50.92857143	JAN	ADULT	P	MALAY	HOUSEWIFE	RASAH	MPS	SN	TKW
2003	1	1	100	9	109	31.5125	22.45	25.72395833	77.25208333	50.92857143	JAN	CHILD	L	MALAY	CHILD	AMPANGAN	MPS	SN	DKW
2003	1	1	100	9	109	31.5125	22.45	25.72395833	77.25208333	50.92857143	JAN	ADULT	L	MALAY	WORKER	AMPANGAN	MPS	SN	TKW
2003	1	1	100	9	109	31.5125	22.45	25.72396	77.25208	50.92857	JAN	ADULT	L	INDIA	STUDENT	RASAH	MPS	SN	TKW
2003	1	1	100	9	109	31.5125	22.45	25.72396	77.25208	50.92857	JAN	MIDDLE	L	INDIA	STUDENT	RANTAU	MPS	SN	MWB
2003	1	1	100	9	109	31.5125	22.45	25.72396	77.25208	50.92857	JAN	ADULT	P	MALAY	WORKER	AMPANGAN	MPS	SN	DKW
2003	1	1	100	9	109	31.5125	22.45	25.72396	77.25208	50.92857	JAN	CHILD	L	MALAY	CHILD	RASAH	MPS	SN	TKW
2003	1	1	100	9	109	31.5125	22.45	25.72396	77.25208	50.92857	JAN	MIDDLE	P	CHINESE	NON	RASAH	MPS	SN	DKW
2003	1	1	100	9	109	31.5125	22.45	25.72395833	77.25208333	50.92857143	JAN	MIDDLE	L	MELAYU	Pelajar	AMPANGAN	MPS	SN	DKW

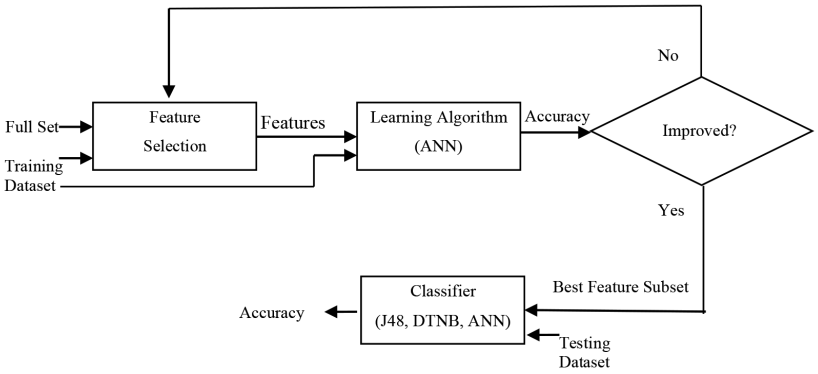

FIGURE 2. Feature selection process: Feature subsets evaluation based on ANN

were passed on to the ANN learning algorithm that will evaluate the selected features based on accuracy. If the selected features increased the accuracy, then the features were selected, else other features will be selected and tested again by ANN.

All the above is performed in a wrapper approach. The wrapper approach is used for feature selection algorithms (in this case, PSO, GA and RS) because wrapper methods apply an induction algorithm (in this case, ANN) to evaluate the merit of selected feature. The rationale of wrapper methods is that the induction algorithm (ANN) provides a better accuracy than using separate data measures or characteristics (El Akadi et al. 2011). Wrapper methods usually achieve better results than filters because they are adjusted to particular interaction between training data and the induction algorithm.

PREDICTIVE MODELING

The performance of the selected features for outbreak detection is further evaluated by employing three classification algorithms. The aim was to investigate the best classifier that fits the dengue dataset. Three algorithms chosen are the J48, DTNB and ANN representing structural, statistical and machine learning approaches for classification (Kumar et al. 2005). These three approaches were chosen since each approach has its own strength for the classification.

Past works indicate that decision tree algorithms (J48) are capable of achieving an outstanding performance in different real world applications. ANN provides a robust method to capture non-linear relationships among patterns and variables (Falavigna 2012). For classification tasks, naïve Bayes has shown outstanding results even in comparison with complex methods (Liangxiao et al. 2009). Many researchers concluded that naïve Bayes is efficient computationally and outperforms other techniques in many applications in terms of accuracy and performance (Baharudin et al. 2010; Godec et al. 2010). The class for predictive model is set as O (outbreak) in case of the occurrence of more than one case in specific geographical area above the average of the previous period according to the used definition of dengue outbreak as described in the previous section, otherwise class value will be N (no outbreak).

EXPERIMENTS AND RESULTS

This section explains experimental results from Experiment I for feature selection algorithms in dengue dataset and Experiment II for predictive model development for selected features.

EXPERIMENT I: FEATURE SELECTION

According to Nemati et al. (2010), the parameters of population-based algorithms have been set as follows: The population size to 30 and the number of iterations to 50. GA and RS parameters have been set as most of previous works recommended in terms of feature selection task. The crossover probability in GA has been set to 0.7 whereas the mutation probability has been set to 0.3. In PSO, the acceleration constants c_1 and c_2 have been set equally to 1 and the inertia weight has been set from 0.4 to 1.4. The velocity of particles *V* has been adjusted to be from 1 to 17. As shown in Table 2, the number of features selected by PSO is 8 while by GA is 9 and RS is 4. Both GA and PSO select almost similar set of features as RS.

The evaluation of feature subsets is based on the accuracy of dengue outbreak detection using neural networks after 1000 learning iterations. The modeling of dengue data is based on multi-layer perceptron (MLP-ANN) with back-propagation algorithm. The parameters of back-propagation algorithm have been as suggested by Toth et al. (2000). Toth et al. (2000) suggested that the number of hidden layer neurons be set equally as the input layer neurons as this will result in better forecasting results. Alpha value in the back-propagation has been set to 1.0 and the momentum value to 0.2. After all the settings are adjusted, the dengue cases records provided as input features into feature selection algorithm with percentage splits to provide a general model of the data represented.

The experimental results of the feature selection algorithms on dengue outbreak data using ANN is shown in Table 3 (average mean squared error (MSE)) and Table 4

TABLE 2. The features selected by different algorithms

FS method	GA	PSO	RS
Feature/#Feature	9	8	4
YEAR	×	×	
ACC_WEEK	×	×	×
DF	×	×	×
DHF	×		
CASE	×	×	
TEMP_MIN	×	×	×
TEMP_AVE	×	×	
RAINFALL	×	×	
RACE		×	×
ADDRESSID	×		

TABLE 3. Average MSE for ANN, GA-ANN, PSO-ANN and RS-ANN

Partitions		ANN	GA-ANN	PSO-ANN	RS-ANN
90	10	0.1924	0.0177	0.0081	0.3004
80	20	0.3198	0.0491	0.2129	0.3224
70	30	0.242	0.0258	0.0516	0.2562
60	40	0.2718	0.0405	0.0285	0.293
50	50	0.2478	0.0589	0.0624	0.3042
40	60	0.2837	0.0868	0.0482	0.2263
30	70	0.3099	0.1108	0.0132	0.2924
20	80	0.3437	0.1309	0.125	0.3085
10	90	0.4362	0.1965	0.2726	0.3308
Average MSE		0.294144	0.079667	0.091389	0.292689

TABLE 4. Average accuracy for ANN, GA-ANN, PSO-ANN and RS-ANN

Partitions		ANN	GA-ANN	PSO-ANN	RS-ANN
90	10	91.01	100	100	86.84
80	20	90.1	99.83	94.57	84.21
70	30	89.25	100	99.72	91.55
60	40	88.34	99.87	99.95	85.64
50	50	89.23	99.63	99.53	86.57
40	60	91.18	98.95	99.69	93.45
30	70	80.27	98.4	99.5	88.81
20	80	77.52	97.88	98.21	86.8
10	90	77.43	95.1	90.24	84.25
Average Accuracy		86.04	98.85	97.93	87.57

(average accuracy). Referring to Tables 3 and 4, the ANN column refers to the approach that did not use feature selection (FS). The other columns refer to the approach that used FS. The measurement values were evaluated further using a Paired t-test using p= 0.05.

Based on the significance test at p=0.05 for MSE values in Table 3, the difference between ANN vs GA-ANN and ANN vs PSO-ANN is significantly different. However, the difference between ANN vs RS-ANN is not significantly different.

Analyzing the significance test at p=0.05 for Accuracy values for ANN vs GA-ANN and ANN vs PSO-ANN (Table 4), the means in both comparisons are significantly different at p=0.05. However, for ANN vs RS-ANN, the means for both groups are not significantly different at p=0.05. This result confirms the ability of GA-ANN and PSO-ANN which are nature-inspired algorithms to optimize the feature selection process and produce the most presentative features for outbreak prediction. RS-ANN which ranks single features could not match the optimization prowess of PSO-ANN and

GA-ANN. The results also proved that using feature selection improves the classification accuracy on the training dataset.

PSO-ANN has achieved the highest average accuracy at 98.85% compared to GA-ANN and RA-ANN. PSO-ANN achieved this using only 8 features compared to 9 features selected by GA-ANN. It indicated that PSO manages to obtain maximum features with high accuracy that best represent that knowledge contained in the dataset. PSO-ANN managed to select the best features from dengue data since the purpose of this experiment is to determine the best representation of dengue data selected by each algorithm (Table 2).

EXPERT EVALUATION OF PSO FEATURES

The selected features by PSO are presented to a group of experts from the Public Health Department, Seremban and a group of three researchers from the Faculty of Health Science, Universiti Kebangsaan Malaysia. They confirm that all the selected features are significant and most related to dengue outbreak. Eight selected features by PSO that gives the best model are YEAR, ACC_WEEK, DF, CASE, TEMP_MIN, TEMP_AVE, RAINFALL and RACE.

For the past 30 years, there has been a dramatic global re-emergence of dengue fever with expanding geographic distribution of both the viruses and the mosquito vectors, resulting in increased epidemic activity and the emergence of dengue hemorrhagic fever (Gubler 2008). The relationship between vector-borne disease and climatic changes has been well documented globally (Reiter 2001), including in Malaysia (Ambu et al. 2003). This has verified the importance of features such as YEAR, ACC_WEEK and CASE. Rain water provides ample opportunities for the Aedes mosquitoes to breed in man-made containers which abound within human settlements (Li et al. 1985). After bouts of rain, the pools of stagnant water provide ample breeding sites for mosquitoes that indicate the importance of TEMP_MIN, TEMP_AVE, and RAINFALL.

The mosquito vectors become infected when they feed on humans during the usual five-day period of viraemia. The virus passes from the mosquito intestinal tract to the salivary glands after an extrinsic incubation period, a process that takes approximately 10 days and is most rapid at high ambient temperatures (WHO 2009). Therefore the incubation period (ACC_WEEK), DF, and CASE play major role in the dengue outbreak problem. A warm and ambient temperature is conducive to the mosquito's gonotrophic life cycle (Delatte et al. 2009; Patz & Reisen 2001) and to viral replication (Patz & Reisen 2001) indicate the significance of the temperature and the rainfall.

Dengue was found to be more prevalent in densely populated urban areas (Chong 2010; Nyamah et al. 2010) and a strong positive spatial association between prevalence of dengue fever and population distribution (Husin & Salim et al. 2005). Race-related susceptibility to dengue has been observed in a few studies and merits further investigation (Guha-Sapir & Schimmer 2005). In Asia, two studies report racial differences in disease incidence. A 15-year study of the epidemiology of dengue reports a significantly higher incidence of DHF among Chinese compared to Malaysian males (Shekhar & Huat 1992). This finding is supported by a six-year surveillance data study in Singapore, which found the race-specific morbidity rate among the Chinese to be three times that of the Malays and 1.7 times that of Indians (Goh 1997). This explains the importance of RACE selected by the PSO algorithm.

The experiment and expert justifications show that feature selection process facilitates the improvement of the accuracy in the modeling task. It also shows the capability of PSO-ANN to select the best features in dengue data. The eight features selected by PSO-ANN will be further analyzed in the next experiment to investigate their impact on other predictive modeling techniques.

EXPERIMENT II: PREDICTIVE MODELS DEVELOPMENT

This experiment concentrates on the predictive modeling (classification) task of dengue outbreak detection based on the selected attributes from PSO-ANN. Three different classification techniques for these tasks which are J48, DTNB and ANN using percentage splits. J48 classification scored accuracy of 100% for 4 models in 1 week dataset, 5 models in 2 weeks dataset and 3 models in 1 month dataset. DTNB scored 100% for 3 models in 1 week dataset, 4 models in 2 weeks dataset and 2 models in 1 month dataset. ANN produced only 1 model in 1 week dataset that scores 100% accuracy. J48 and DTNB are better at capturing the relation among features in for outbreak detection compared to ANN.

COMPARATIVE ANALYSIS

Three different predictive modeling techniques namely J48, DTNB and ANN have been applied to model the problem (Kumar et al. 2005). These techniques have different approaches to model the problem such as structural modeling, statistical modeling or neural network. As shown in the previous sections, the predictive modeling task has been applied using different definitions of dengue outbreak in terms of timeframe basis. Thus, this comparison is shown in the same manner. After comparing the accuracy achieved by each predictive modeling technique in each fold, the main conclusion of these experiments will be discussed. Table 5 shows the comparison of the three used techniques in term of the accuracy achieved in each fold using 1 week, 2 week and 1 month definitions of dengue outbreak.

As shown in Figure 5, J48 (Decision Tree) outperforms the DTNB and ANN and achieve 100% accuracy in the most folds. ANN has the worst performance in this experiment as the results showed. However, most of the studies in this area conclude that different training settings would lead to different results. This might open a future direction to investigate the influence of different settings on the final results of ANN performance modeling dengue data. It is also obvious that in J48 and DTNB, the accuracy in early folds is the best due to the partitioning manner of training

data. When the training subset of data is larger than the testing subset, it facilitates the discovery of relations among variables and enhances the modeling task especially using structural and statistical approaches of data modeling. This explains achieving the best accuracy by J48 and DTNB in some folds wherein the training data is larger than the testing data. The same explanation of result goes for the experiment using 2-weeks definition of dengue outbreak. For the 2 weeks' data, ANN, J48 and DTNB each scored 100% accuracy with J48 and DTNB scoring 100% in 4 models and ANN with 1 model. This is in accordance with the definition of dengue outbreak as the occurrence of more than one case in a specific geographical area where the onset date between cases is less than 14 days (Seng et al. 2005). In the domain of dengue outbreak, most experts suggest and support this 2 weeks definition. In terms of the performance of predictive modeling techniques, J48 has achieved the best average accuracy of 99.86% for 2 weeks (Table 5). DTNB has also achieved average accuracy of 99.29% whereas ANN scored the lowest average accuracy (97.71%). For the 1 month dataset of dengue outbreak, the average accuracy achieved by different techniques are ANN (87.44%), J48 (98.61%) and DTNB (99.02%)

PERFORMANCE OF PSO-BASED TECHNIQUE WITH OTHER DENGUE DETECTION

Table 6 shows the comparative results from three related works using the dengue dataset. All dataset use the Malaysian dengue data with similar original features. The experiments have been conducted under different feature scheme, however the comparison indicate the applicability of using feature selection method that improves the modeling of dengue outbreak detection. The comparison includes the results from this research with three methods namely negative selection algorithm (NSA) by Mousavi et al. (2011), multiple attribute value (MAV) by Long et al. (2010) and cumulative sum (CUSUM) used by Long et al. (2010) to compare their result. PSO generates the least number of features with the highest accuracy and able to optimize the number of features thus producing the highest detection accuracy.

CONCLUSION

Using dengue historical data, this work facilitates at capturing the relations and patterns within the data and enhancing the predictive accuracy of dengue outbreaks. Unlike most previous works, this work determines the most related attributes of dengue data before modeling the problem by applying feature selection process. The experimental results showed the advantages of using feature selection process to enhance the accuracy of the modeling task. Among the three feature selection algorithms that have been applied as feature selection techniques, PSO has reached the best accuracy revealing new set of features to represent dengue data. These features are: year, cumulative week from year 2003 till year 2010, number of dengue fever for current week, minimum temperature value, average temperature value, rainfall and race. Using wrapper approach to evaluate the feature subsets by employing multi-layer perceptron (MLP) neural network has helped to select to proper features

TABLE 5. Average accuracy for ANN, GA-ANN, PSO-ANN and RS-ANN

Model	1 week			2 weeks			1 month		
	ANN	J48	DTNB	ANN	J48	DTNB	ANN	J48	DTNB
1	98.02	100	100	100	100	100	90.29	100	100
2	97.86	100	99.91	94.57	100	100	88.4	100	100
3	96.76	100	100	99.72	100	100	89.36	100	99.83
4	98.47	100	100	99.95	100	100	89.59	99.87	99.87
5	98.78	99.96	99.7	99.53	100	99.83	80.85	99.76	99.9
6	96.57	99.97	98.35	99.69	99.94	99.69	88.7	98.79	99.91
7	97.36	99.97	98.09	99.5	99.97	99.78	87.22	98.52	99.92
8	95.12	98.66	95.41	98.21	99.97	97.3	87.17	98.15	99.63
9	95.46	98.06	87.94	90.24	98.86	97.04	85.38	92.36	92.1
Average Accuracy	97.16	99.62	97.71	97.93	99.86	99.29	87.44	98.61	99.02

TABLE 6. Comparison of results from this research and previous works

Algorithm	NSA	MAV	CUSUM	PSO-J48	PSO-DTNB	PSO-ANN
# Features	12	13	13	8	8	8
Accuracy	79.85%	73.1%	67.3%	100%	100%	100%

that directly related to the predictive accuracy of dengue outbreak detection.

ACKNOWLEDGEMENTS

This work is supported by the Exploratory Research Grant Scheme (ERGS/1/2011/STG/UKM/02/49) and in collaboration with Negeri Sembilan State Health Department, Ministry of Health, Malaysia.

REFERENCES

Ambu, S., Lim, L.H., Sahani, M. & Bakar, A.B. 2003. Climate change-impact on public health in Malaysia. *Environ Health Focus* 1: 13-21.

Andrick, B., Clark, B., Nygaard, K., Logar, A., Penaloza, M. & Welch, R. 1997. Infectious disease and climate change: Detecting contributing factors and predicting future outbreaks. *Geoscience and Remote Sensing, 1997. IGARSS '97. Remote Sensing - A Scientific Vision for Sustainable Development* ***4:*** 1947-1949. *IEEE International*.

Bakar, A.A., Kefli, Z., Abdullah, S. & Sahani, M. 2011. Predictive models for dengue outbreak using multiple rulebase classifiers. *Electrical Engineering and Informatics (ICEEI), 2011 International Conference*, Bandung. pp. 1-6.

Barbazan, P., Yoksan, S. & Gonzalez, J.P. 2002. Dengue hemorrhagic fever epidemiology in Thailand: Description and forecasting of epidemics. *Microbes Infect.* 4: 699-705.

Beltz, L.A. 2011. *Emerging Infectious Diseases: A Guide to Diseases, Causative Agents, and Surveillance*. New York: John Wiley & Sons. pp. 315-322.

Bolón-Canedo, V., Sánchez-Maroño, N. & Alonso-Betanzos, A. 2012. An ensemble of filters and classifiers for microarray data classification. *Pattern Recognition* 45: 531-539.

Bolón-Canedo, V., Sánchez-Maroño, N. & Alonso-Betanzos, A. 2011. Feature selection and classification in multiple class datasets: An application to KDD Cup 99 dataset. *Expert Systems with Applications* 38: 5947-5957.

Buckeridge, D.L., Burkom, H., Campbell, M., Hogan, W.R. & Moore, A.W. 2005. Algorithms for rapid outbreak detection: A research synthesis. *Journal of Biomedical Informatics* 38: 99-113.

Chidlovskii, B. & Lecerf, L. 2008. Scalable feature selection for multi-class problems. 2008. *Proceedings of the 2008 European Conference on Machine Learning and Knowledge Discovery in Databases - Part I* (ECML PKDD '08), Walter Daelemans, Bart Goethals, and Katharina Morik (Eds.). Springer-Verlag, Berlin, Heidelberg. pp. 227-240.

Chong, C. 2010. Scenario of dengue in Malaysia. Paper presented at *Europe-South East Asia Symposium on Dengue,* 5-6 August 2010, Ministry of Health, Malaysia.

Delatte, H., Gimonneau, G., Triboire, A. & Fontenille, D. 2009. Influence of temperature on immature development, survival, longevity, fecundity, and gonotrophic cycles of *Aedes albopictus*, vector of chikungunya and dengue in the Indian Ocean. *Journal of Medical Entomology* 46: 33-41.

Edelman, R. 2007. Dengue vaccines approach the finish line. *Clin. Infect.* 45(Suppl. 1): S56-S60.

El Akadi, A., Amine, A., El Ouardighi, A. & Aboutajdine, D. 2011. A two-stage gene selection scheme utilizing MRMR filter and GA wrapper. *Knowledge and Information Systems* 26: 487-500.

Fu, X., Liew, C., Hung, T., Goh, H. & Lee, G. 2007. Time-series infectious disease data analysis using SVM and genetic algorithm. In *IEEE Congress on Evolutionary Computation*. pp. 1276-1280.

Goh, K. 1997. Dengue-a re-emerging infectious disease in Singapore. *Annals of the Academy of Medicine Singapore* 26(5): 664-670.

Gomez, J.C., Boiy, E. & Moens, M.F. 2012. Highly discriminative statistical features for email classification. *Knowledge and Information Systems* 31(1): 23-53.

Gubler, D.J. 2008. Dengue viruses. In Encyclopedia of Virology. 3rd ed., edited by *Mahy, B.W.J. & van Regenmortel, M.H.V. Boston: Academic Press.* pp. 5-14.

Guha-Sapir, D. & Schimmer, B. 2005. Dengue fever: New paradigms for a changing epidemiology. *Emerg. Themes. Epidemiol.* 2(1): 1-10.

Guy, B. & Almond, J.W. 2008. Towards a dengue vaccine: Progress to date and remaining challenges. Comparative Immunology. *Microbiology and Infectious Diseases* 31(2-3): 239-252.

Guyon, I. 2003. An introduction to variable and feature selection. *Journal of Machine Learning Research* 3: 1157-1182.

Hombach, J. 2007. Vaccines against dengue: A review of current candidate vaccines at advanced development stages. *Revista Panamericana de Salud Pública* 21(4): 254-260.

Husin, N.A. & Salim, N. 2008. A comparative study for back propagation neural network and non-linear regression models for dengue outbreak prediction. *Jurnal Teknologi Maklumat* 20(4): 97-112.

Hussin, N., Jaafar, J., Naing, N.N., Mat, H.A., Muhamad, A.H. & Mamat, M.N. 2005. A review of dengue fever incidence in Kota Bharu, Kelantan, Malaysia during the years 1998-2003. *Southeast Asian J. Trop. Med. Public Health* 36(5): 1179-1186.

Li, C., Lim, T., Han, L. & Fang, R. 1985. Rainfall, abundance of Aedes aegypti and dengue infection in Selangor, Malaysia. *Southeast Asian J. Trop. Med. Public Health* 16(4): 560-568.

Liu, H. & Yu, L. 2005. Toward integrating feature selection algorithms for classification and clustering. *IEEE Trans. on Knowl. and Data Eng.* 17(4): 491-502.

Long, Z., Abu Bakar, A., Razak Hamdan, A. & Sahani, M. 2010. Multiple attribute frequent mining-based for dengue outbreak. In *Proceedings of the 6th International Conference on Advanced Data Mining and Applications: Part I (ADMA'10)*, edited by Longbing Cao, Yong Feng and Jiang Zhong. Berlin, Heidelberg: Springer-Verlag. pp. 489-496.

Loscalzo, S., Yu, L. & Ding, C. 2009. Consensus group stable feature selection. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '09)*. New York: ACM. pp. 567-576.

Mousavi, M., Bakar, A.A., Zainudin, S. & Awang, Z. 2013. Negative selection algorithm for dengue outbreak detection. *Turkish Journal of Electrical Engineering & Computer Science* 21: 2345-2356.

Nemati, S. & Basiri, M. 2010. Particle swarm optimization for feature selection in speaker verification. *Applications of Evolutionary Computation*. Lecture Notes in Computer Science 6024: 371-380.

Nyamah, M., Sulaiman, S. & Omar, B. 2010. Categorization of potential breeding sites of dengue vectors in Johor, Malaysia. *Tropical Biomedicine* 27(1): 33-40.

Patz, J.A. & Reisen, W.K. 2001. Immunology, climate change and vector-borne diseases. *Trends in Immunology* 22(4): 171-172.

Que, J. & Tsui, F.C. 2011. Rank-based spatial clustering: An algorithm for rapid outbreak detection. *Journal of the American Medical Informatics Association* 18(3): 218-224.

Reiter, P. 2001. Climate change and mosquito-borne disease. *Environ. Health Perspect.* 109(Suppl 1): 141-161.

Research, S. P. f., Diseases, T. i. T., & Diseases, W. H. O. D. o. C. o. N. T. (2010). Dengue Bulletin. 34.

Runge-Ranzinger, S., Horstick, O., Marx, M. & Kroeger, A. 2008. What does dengue disease surveillance contribute to predicting and detecting outbreaks and describing trends?. *Tropical Medicine & International Health* 13: 1022-1041.

Saari, P., Eerola, T. & Lartillot, O. 2011. Generalizability and simplicity as criteria in feature selection: Application to mood classification in music. *IEEE Transactions on Audio, Speech, and Language Processing* 19(6): 1802-1812.

Seng, S.B., Chong, A.K. & Moore, A. 2005. Geostatistical modelling, analysis and mapping of epidemiology of Dengue Fever in Johor State, Malaysia. Presented at the *17th Annual Colloquium of the Spatial Information Research Centre (SIRC 2005: A Spatio-temporal Workshop*). pp. 109-123.

Shekhar, K.C. & Huat, O.L. 1992. Epidemiology of dengue/ dengue hemorrhagic fever in Malaysia - A retrospective epidemiological study 1973-1987. Part I: Dengue hemorrhagic fever (DHF). *Asia Pac. J. Public Health* 6(3): 126-133.

Skae, F. 1902. Dengue fever in Penang. *Br. Med. J.* 2(2185): 1581-1582.

Sun, Y., Babbs, C. & Delp, E. 2005. A comparison of feature selection methods for the detection of breast cancers in mammograms: Adaptive sequential floating search vs. genetic algorithm. *IEEE-EMBS 2005*. 27th Annual International Conference, Shanghai. pp. 6532-6535.

Talarmin, A., Peneau, C., Dussart, P., Pfaff, F., Courcier, M., de Rocca-Serra, B. & Sarthou, J. 2000. Surveillance of dengue fever in French Guiana by monitoring the results of negative malaria diagnoses. *Epidemiol. Infect.* 125(1): 189-193.

Toth, E., Brath, A. & Montanari, A. 2000. Comparison of short-term rainfall prediction models for real-time flood forecasting. *Journal of Hydrology* 239(1-4): 132-147.

Tuv, E., Borisov, A., Runger, G. & Torkkola, K. 2009. Feature selection with ensembles, artificial variables, and redundancy elimination. *J. Mach. Learn. Res*. 10: 1341-1366.

Vainer, I., Kraus, S., Kaminka, G.A. & Slovin, H. 2011. Obtaining scalable and accurate classification in large-scale spatio-temporal domains. *Knowledge and Information Systems* 29(3): 527-564.

World Health Organization. 2009. Research SPF, Diseases TIT, Diseases WHOD, Epidemic WHO and P. Alert, *Dengue, Guidelines for Diagnosis, Treatment, Prevention and Control*.

Wu, Y., Lee, G., Fu, X., Soh, H. & Hung, T. 2009. Mining weather information in dengue outbreak: Predicting future cases based on wavelet, SVM and GA. *Advances in Electrical Engineering and Computational Science*. Netherlands: Springer. pp. 483-494.

Zhang, Y., Ding, C. & Li, T. 2008. Gene selection algorithm by combining reliefF and mRMR. *BMC Genomics* Supp1 2: S27.

Husam I.S. Abuhamad, Azuraliza Abu Bakar & Suhaila Zainudin*
Center for Artificial Intelligence Technology
Faculty of Information Science and Technology
Universiti Kebangsaan Malaysia
43600 UKM Bangi, Selangor Darul Ehsan
Malaysia

Mazrura Sahani
Faculty of Health Sciences
Universiti Kebangsaan Malaysia, Jalan Raja Muda Abd Aziz
50300 Kuala Lumpur, Wilayah Persekutuan
Malaysia

Zainudin Mohd Ali
Public Health Department, Ministry of Health, Jalan Rasah
70300 Seremban, Negeri Sembilan Darul Khusus
Malaysia

*Corresponding author; email: suhaila.zainudin@ukm.edu.my

Received: 11 March 2016
Accepted: 8 June 2016