The Classification Performance using Support Vector Machine for Endemic Dengue Cases

N I Nordin1, N Mohd Sobri2, N A Ismail3, S N Zulkifli4, N F Abd Razak5 and M Mahmud6

1,2,3,4,5,6 Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA Kelantan, Bukit Ilmu, Machang, Kelantan, Malaysia

Email: ilanie@uitm.edu.my

Abstract. Dengue fever (DF) and the potentially fatal dengue haemorrhagic fever (DHF) are continue to be a crucial public health concern in Malaysia. This paper proposes a prediction model that incorporates Support Vector Machine (SVM) in predicting future dengue outbreak. Datasets used in the undertaken study includes data on dengue cases provided by the Health Department in Kelantan, Malaysia. Data scaling were applied to normalize the range of features before being fed into the training model. In this regard, SVM models built on the basis of three different kernel functions including Gaussian radial basis function (RBF), polynomial function and linear function. The SVM with RBF kernel function was superior to the other techniques because it obtains the highest prediction accuracy of 85%. The polynomial is an alternative model that can achieve a high prediction performance in terms of sensitivity (76%) and specificity (87%).

1. Introduction
Dengue fever (DF) is one of the most important infectious diseases worldwide, particularly in areas with tropical and sub-tropical climates [1-3]. In Malaysia, dengue fever (DF) remains to be a public health dilemma causing 50 to 100 million cases annually, incorporating more than 100 endemic countries in the world [4]. It was first recorded in Malaysia in 1902 before haemorrhagic fever (DHF) came out in 1962 [5]. The main factors contributing to the spread of the disease include stagnant water in containers, pots and drains, global warming and unexpected rainfall. In addition, the increase in universal trade and tourism also contributes to the expansion of dengue cases [6]. Rachata et al. [6] indicated that 80% of dengue cases contribute to the elevated breeding rates for Aedes. Thus, there have been several numbers of campaign undertaken by Malaysian government, especially Ministry of Health in educating community in checking their premises and destroy all breeding sites. Nevertheless, there are still no vaccines available for DF. Hence, serious efforts are essential to control and prevent this disease from becoming pervasive [7]. Obviously, the current precaution steps that have been carried out such as awareness campaign, education to the community and others are not adequate. Thus, other effort needs to be identified and these include the ability to predict future dengue outbreak [8].

There are several techniques, such as the autoregressive integrated moving average (ARIMA), the K-H model, the artificial neural network (ANN) and a Poisson regression analysis that have been utilized in predicting of future dengue outbreak [8]-[10]. Achieving high accuracy in prediction is important as it can lead to a well-planned precaution programmer. Different training and learning techniques...
approaches may lead to different prediction accuracy. In this research, we applied the SVM with the linear kernel and RBF kernel to produce high prediction accuracy in the dengue outbreak.

2. Literature Review

Study on predicting dengue haemorrhagic fever (DHF) in Thailand has been done by [9]. In the study, an automatic prediction system for DHF is proposed by utilizing entropy technique and Artificial Neural Network (ANN). Entropy is used to extract the relevant information that affects the prediction accuracy. Yusof and Mustaffa [11] have done a study in predicting future dengue outbreak. They proposed a prediction model that incorporates Least Square Support Vector Machines (LS-SVM). The data sets used includes data on dengue cases and rainfall level collected in five districts in Selangor, Malaysia. Decimal Point Normalization was used to pre-process the data before being fed into the training model. The result shows that in term of prediction accuracy and computational time, the LS-SVM prediction model outperformed the Neural Network model. This result is based on the prediction of unseen data. By applying entropy technique, it would yield a better result as the entropy technique produces 85.92% accuracy while only 78.16% when entropy technique is not applied.

In another study, Fathima and Hundewale [12] examined a various data mining technique to predict disease-Dengue. The analysis was used the real time data set taken from specialty hospitals and diagnostic laboratories. They used SVM and Naïve Bayes Classifier techniques to test the performance. The random forest classifier with its associated Gini feature importance allows identifying small sets of parameters to be used for diagnostic purposes in clinical practice. This involves obtaining the smallest possible set of symptoms that can achieve decent predictive performance for dengue disease. They combined both the approaches and evaluate the classifiers performance. The results showed that Naïve Bayes is less perform compared to SVM that outperform well.

They are also several other studies that used different data mining approach to predict dengue disease. Yan Wu et al [13] used a Wavelet transformation for data pre-processing before applying a SVM-based Genetic Algorithm to select the most important features. The results showed an improvement in the prediction performance. Xiuju Fu et al [14] used SVM classifier for analysing the time series dengue data and genetic Algorithm (GA) to determine the subsets and time-lags of climatic factors as effective factors influencing the spread of dengue. As a result, the GA-based on SVM classifier could improve the classification accuracy significantly. Rahim et al [15] have done a study on dengue fatality prediction using different data mining technique. They used Naïve Bayes, Decision Tree, Logistic Regression, SVM, CN2 and Random Forest. The results show the performance of CN2 is more accurate compared to other techniques. The study done by Kamran Shaukat et al [16] used Naïve Bayesian (NB), REP Tree, Random Tree, J48 and SMO as a classification technique. They conclude that NB and J48 are the top performance classifier.

Dengue outbreak prediction in Malaysia using the ANN was proposed by Husin et al [17]. The experiments showed that the ANN-based model outperformed a nonlinear regression model, as measured by the mean square error (MSE). However, the ANN needs to determine how many neurons are required for a task, and this is another issue that affects whether the ANN is optimized. A new machine learning method, e.g., the SVM, has become more attractive to researchers in the field. SVM is an efficient approach for nonlinear classification problem solving.
3. Support Vector Machine (SVM)

A Support Vector Machine (SVM) is a linear marginal classifier, which can also be extended to nonlinear cases. Support Vector Machines are robust and commanding tools to categorize data according to Fung et al., [18]. It generally categorizes data into two parts dividing it by a hyperactive plane. It is an intricate model that uses multidimensional surfaces to outline the relationship between features and outcomes. In spite of it being complex, it can be applied to real world problems where classification and prediction are essential [19-21].

Hasti and Tibshirani [22] said that SVM was a method to make predictions in both classification and regression cases. This method works to find the optimal separator function (hyperplane) that can separate datasets into two different classes or categories. The concept of SVM is as follow: input vectors \( x \) are mapped to a very high dimension feature space \( z \) through some nonlinear mapping \( \varphi(x) \), \( z = \varphi(x) \). In this space, an optimal separating hyperplane is constructed. For a given training dataset with \( n \) samples \((x_1,y_1), (x_2,y_2), \ldots (x_n, y_n)\) where \( x_i \) is a feature vector in a \( d \)-dimensional feature space \( \mathbb{R}^d \) and \( y_i \in \{1, +1\} \) is the corresponding class label. The task is to find a classifier with a decision function \( f(x) = w^T x + b \), where \( w \) represents the weight vector and \( b \) is the bias. The hyperplane is a linear separator that divides space into two parts which can separate data set by maximizing margins. Finding of the best hyperplane done by maximizing the margin or the distance between two objects from different classes. For a linearly separable problem, the equation of a linear SVM can be written as:

\[
f(x) = \sum_{i=0}^{N} \alpha_i y_i x_i^T \cdot x + \beta_0 \tag{1}
\]

where \( x_i \) is the instant with label \( y_i \), \( \alpha \) is Lagrange multiplier and \( \beta_0 \) is bias. For non-linearly separable problem, the above equation can be modified for kernel SVM as:

\[
f(x) = \sum_{i=0}^{N} \alpha_i y_i K(x_i, x) \cdot x + \beta_0 \tag{2}
\]

Here \( N \) represents the number of support vectors, whereas \( K(x_i, x) \) is the kernel function.

3.1. Predictive Model Construction Based on SVM

Several kernel functions are used with the SVM. For instance, the linear kernel is the simplest kernel function as in equation (3), where \( c \) is a constant.

\[
k(x, y) = x \cdot y + c \tag{3}
\]

The polynomial kernel as in equation (4) is a classical method for non-linear modelling and is well fitted for problems when all the training data are normalized where \( d \) is an integer number.

\[
k(x, y) = (x \cdot y)^d \tag{4}
\]

The radial basis function (RBF) kernel is one of the most powerful kernels as in equation 5, where \( \sigma^2 \) is a tuning parameter which associated with RBF function.

\[
k(x, y) = \exp\left[\frac{||x-y||^2}{2\sigma^2}\right] \tag{5}
\]

These three kernels will be deployed in the dengue predictive model, and the highest accuracy model will be selected.

3.2. Predictive Model Performance Criteria

Sensitivity, specificity, and accuracy were used to assess the effectiveness of the proposed model. Sensitivity indicates the fraction of high-risk dengue outbreak areas that are correctly classified, whereas
specificity shows the fraction low-risk areas that are correctly classified. These three measures are computed by the following equations:

\[
\text{Sensitivity} = \frac{TP}{TP+FN} \\
\text{Specificity} = \frac{TN}{TN+FP} \\
\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}
\]

where:
- \(TP\) represents True positive values’ count
- \(TN\) represents True negative values’ count
- \(FP\) represents False positive values’ count
- \(FN\) represents False negative values’ count

Ideally, a perfect predictor would be described as 100% sensitive (i.e., predicting all areas from the high-risk group as being high risk) and 100% specific (i.e., not predicting any areas from the high-risk group being low risk). Three indices, mean absolute error (MAE), mean square error (MSE) and root mean square error (RMSE) were used as measures of predicting accuracy.

4. Methodology
In order to develop the prediction model and later evaluate the effectiveness, the flow of process as depicted in figure 1 is applied.

![Figure 1. Implementation Phase](image)

4.1. Raw Dataset
In this paper, we used endemic dengue cases dataset provided by the Health Department in Kelantan, Malaysia to evaluate the performance for SVM. This dataset comprises of 4 attributes which are as follows,
- \(Y\) = probability of deaths
- \(X1\) = total dengue fever cases
- \(X2\) = average temperature (°C)
- \(X3\) = rainfall (mm)

In total, 84 samples of different ages and genders have been used in this study. Out of 84 samples, 31 were dengue positive and 53 were negative. The dataset are of collection from 5 locations (i. Kota Bharu ii. Machang iii. Bachok iv. Gua Musang and v. Jeli) from which 4 patients died in 2017.
4.2. Data Scaling
Data scaling refers to the method used to normalize the range of features or independent variables of a data mining model. This approach can ensure that data are not overwhelmed by each other in terms of distance measures [23]. We used a standardization technique as in equation (6) to rescale both the training and testing data:

\[ x' = \frac{x - \text{mean}(f)}{\text{std.dev}(f)} \]  

(9)

where \( x' \) is a standardized value, \( x \) is an original value, and mean and std.dev are the value of each feature \( f \), respectively. The results of each model will be assessed by sensitivity, specificity, and accuracy. Finally, the model that best optimizes those measures is selected.

Using the prepared data sets, a model of experiment is performed. The data proportion applied is, 70% for training and 30% for testing data. The project is coded in R using e1071 package [24-26].

4.3. Model Construction
The performance of an algorithm has been evaluated by computing confusion matrix under different conditions. Confusion matrix is basically \( N \times N \) matrix, where each column represents an event in the predicted class, whereas each row represents an event in the actual class. The confusion matrix has been generated with three different kernel functions i.e. linear, polynomial and RBF. Initially, Gaussian radial basis kernel with the scaling factor of 1 and 2 has been applied. Later on the model has been tested with polynomial kernel function of order 1 and 2 and linear kernel, respectively. Based on these confusion matrices in tables 1, 2 and 3, the accuracy, specificity and sensitivity percentage of each method are calculated and displayed in table 5. The overall results for the model with different kernel functions are given in tables 4 and 5.

| Table 1. Confusion Matrix for Linear Function |
|---------------------------------------------|
| Class 1 | Class 2 |
| Class 1 | 27 | 11 |
| Class 2 | 4 | 42 |

| Table 2. Confusion Matrix for Polynomial Function |
|-----------------------------------------------|
| Class 1 | Class 2 |
| Class 1 | 28 | 10 |
| Class 2 | 3 | 43 |

| Table 3. Confusion Matrix for RBF Function |
|------------------------------------------|
| Class 1 | Class 2 |
| Class 1 | 25 | 10 |
| Class 2 | 6 | 43 |

4.4 Predicted Result Evaluation
This section discusses results obtained from the conducted experiments. Figure 2 shows a graph of comparison between male and female for dengue cases in five locations (i. Kota Bharu ii. Machang iii. Bachok iv. Gua Musang and v. Jeli). Location 1 shows the highest dengue cases with two died in 2017. The parameter sets with the lowest values of MSE, MAE and RMSE were selected for best performance. Table 4 compares three accuracy indices of different kernel function. Those results indicate that the SVM with RBF kernel function performs better than linear and polynomial kernel function in terms of MSE, MAE and RMSE.
Figure 2. Comparison between male and female for dengue cases in five location.

| Table 4. Prediction Result |
|---------------------------|
|                      | Linear | Polynomial | RBF   |
|-------------------------|--------|------------|-------|
| MSE                     | 0.1    | 0.2049     | 0.1125|
| MAE                     | 0.2303 | 0.2579     | 0.221 |
| RMSE                    | 0.3162 | 0.4526     | 0.3154|

The overall results for the model with different kernel functions are given in Table 5. The performance of a model is usually evaluated in terms of accuracy, sensitivity and specificity. The SVM with RBF kernel function was superior to other kernel function with highest accuracy measures (85%) followed by polynomial (82%) and linear (81%) respectively. However, the polynomial is an alternative model that can achieve a high prediction performance in terms of specificity (87%) and sensitivity (76%).

| Table 5. The Performance of Predictive Model |
|---------------------------------------------|
|                | Linear | Polynomial | RBF |
| True Positive | 27     | 28         | 25  |
| False Positive| 4      | 3          | 6   |
| True Negative | 42     | 43         | 43  |
| False Negative| 11     | 10         | 10  |
| Accuracy      | 81%    | 82%        | 85% |
| Specificity   | 86%    | 87%        | 86% |
| Sensitivity   | 73%    | 76%        | 74% |

5. Discussion
When the SVM is chosen as a prediction technique for dengue cases, we need to first consider what kernel function is the most suitable for the dengue data. It has been suggested that using nonlinear kernel functions could achieve a better performance than linear kernel functions. Therefore, the SVM with RBF kernel is used as the kernel function of SVM because it can effectively handle instances in which the relationship between the independent and dependent variables is nonlinear. In addition, the RBF kernel has less complexity than the polynomial kernel because it has less hyperparameters, which influences the complexity of model selection. This experiment demonstrates the performance of prediction obtained from the RBF kernel compared with the polynomial and linear kernels. However,
Hsu et al [25] suggested that the RBF kernel will not improve the performance of classification when the number of features is very large; in that case, the linear kernel is preferred. Selecting the most suitable SVM kernel parameters for a predictive model is another challenging task of this research.

6. Conclusion
This paper proposed a model for predicting future dengue outbreak incorporating SVM. From the results, it shows that SVM with RBF kernel function is capable to obtain good generalization ability compared to linear and polynomial kernel, thus improving the prediction accuracy and performance. On the other hand, future research is necessary of this study. In the actual prediction scenario, there are several numbers of factors having their influence on dengue outbreak, such as humidity, temperature and cloudiness[10]. Furthermore, to determine turning parameter is the vital of predict study. A hybrid algorithm could be applied to obtain the optimal parameter, thus improve the prediction accuracy and performance.

This proposed model is expected to benefit the community where with better prediction model, it will be useful to the government, specifically Ministry of Health in arranging plans and identifying initiatives needed for the purpose to strengthen dengue control. Besides, it will also give advantage to the people in understanding and taking precaution steps in preventing dengue epidemic from becoming pervasive.

References
[1] Gibbons RV, Vaughn DW 2002 Dengue: an escalating problem. BMJ. 324 1563–1566. pmid:12089096
[2] Campbell-Lendrum D and Corvalan C 2007 Climate change and developing-country cities: Implications for environmental health and equity. J Urban Health. 84 109–117.
[3] Racloz V, Ramsey R, Tong S and Hu W 2012. Surveillance of dengue fever virus: A review of epidemiological models and early warning systems. PLoS Negl Trop Dis. 6: e1648.pmid:22629476
[4] Gusmao G., S. C. S. Machado, and M. A. B. Rodrigues, 2009. A new algorithm for segmenting and counting aedes aegypti in ovitraps, Proc. IEEE Annual International Conference in Engineering in Medicine and Biology Society (EMBC) 6714- 6717.
[5] Ang K T and Singh S 2001 Epidemiology and New Initiatives in the Prevention and Control of Dengue in Malaysia Dengue Bulletin 25 7-14.
[6] Cuddehe M., 2009 Mexico fights rise in dengue fever World Report 374.
[7] Tan Kah B., L. Koh Hock, and Y. Teh Su, 2009 Modeling Dengue Fever Subject to Temperature Change Proc. Sixth International Conference in Fuzzy Systems and Knowledge Discovery (FSKD) 61-65.
[8] Husin N. A., N. Salim, and A. R. Ahmad, 2008. Modeling of dengue outbreak in Malaysia: A comparison of Neural Network and Nonlinear Regression Model Proc. International Symposium in Information Technology (ITSim) 1-4.
[9] Rachata N, Charoenkwan P T, Yooyativong K, Chammongthal C, Lursinsap, and Higuchi K 2008 Automatic Prediction System of Dengue Haemorrhagic-Fever Outbreak Risk by Using Entropy and Artificial Neural Network Proc. International Symposium in Communications and Information Technologies(ISCIT) 210-214.
[10] Wu Y, Lee G, Fu X, and Hung T, 2008 Detect Climatic Factors Contributing to Dengue Outbreak based on Wavelet, Support Vector Machines and Genetic Algorithm Proc. World Congress on Engineering 1947-1949.
[11] Yusof Y and Mustaffa, Z 2011 Dengue outbreak prediction: A least squares support vector machines approach. International Journal of Computer Theory and Engineering, 3(4), 489.
[12] Fathima S and Hundewale N 2011 Comparison of classification techniques-SVM and naive bayes to predict the Arboviral disease-Dengue. 2011 IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBM) 538-539. IEEE.

[13] Wu Y, Lee G, Fu X, Soh H, and Hung T 2009 Mining weather information in dengue outbreak: predicting future cases based on wavelet, SVM and GA. In Advances in Electrical Engineering and Computational Science (pp. 483-494). Springer, Dordrecht.

[14] Fu X, Liew C, Soh H, Lee G, Hung T, and Ng L C 2007. Time-series infectious disease data analysis using SVM and genetic algorithm. In 2007 IEEE Congress on Evolutionary Computation (pp. 1276-1280). IEEE.

[15] Rahim N F, Taib S M, and Abidin A I Z 2017 Dengue fatality prediction using data mining. Journal of Fundamental and Applied Sciences 9(6S) 671-683.

[16] Shaukat K, Masood N, Mehreen S, and Azmeen U 2015 Dengue fever prediction: A data mining problem. Journal of Data Mining in Genomics & Proteomics.

[17] Husin NA, Salim N, Ahmad AR. 2008 Modeling of dengue outbreak prediction in Malaysia: A comparison of Neural Network and Nonlinear Regression Model. International Symposium on Information Technology; pp. 1–4

[18] Fung M, Glenn and Mangasarian O L 2005 Multicategory proximal support vector machine classifiers Machine learning 59 77-97.

[19] Lantz B 2013 Machine learning with R. Packt Publishing Ltd.

[20] Meyer D and Wien F T 2015 Support vector machines. The Interface to libsvm in package e1071. Aug 5.

[21] Meyer D 2014 Support Vector Machines the Interface to libsvm in package e1071 Paper available at http://cran.rproject.org/web/packages/e1071/vignettes/svmdoc.pdf.

[22] Hastie T and Tibshirani R 1998 Classification by pairwise coupling Advances in neural information processing systems.

[23] Yusof Y and Mustaffa Z 2011 Dengue outbreak prediction: A least squares support vector machines approach. IJCTE 3 489–493.

[24] Dimitriadou E, Hornik K, Leisch F, Meyer D and Weingessel A 2009 MF Leisch. Package ‘e1071’. RSoftwarepackage,availableathttp://cran.rproject.org/web/packages/e1071/index.html Jan 6.

[25] Hsu C, Chang C and Lin C, 2003 A practical guide to support vector classification. Department of Computer Science, National Taiwan University; 1–16.