Dengue Outbreak Prediction: A Least Squares Support Vector Machines Approach

Yuhanis Yusof and Zuriani Mustaffa

Abstract—**Dengue fever (DF) and the potentially fatal dengue hemorrhagic fever (DHF) continue to be a crucial public health concern in Malaysia. This paper proposes a prediction model that incorporates Least Squares Support Vector Machines (LS-SVM) in predicting future dengue outbreak. Data sets used in the undertaken study includes data on dengue cases and rainfall level collected in five districts in Selangor. Data were preprocessed using the Decimal Point Normalization before being fed into the training model. Prediction results of unseen data show that the LS-SVM prediction model outperformed the Neural Network model in terms of prediction accuracy and computational time.**

Index Terms— **Decimal Point Normalization, Dengue fever, Least Squares Support Vector Machines, Support Vector Machines.**

I. Introduction

To date, dengue fever (DF) remains to be a public health dilemma in Malaysia. It is one of the most universally disseminate insect-born virus infection, causing 50 to 100 million cases annually, incorporating more than 100 endemic countries in the world [1]. In Malaysia, it was first reported in 1902 before dengue hemorrhagic fever (DHF) came out in 1962 [2]. Nine years later, in 1971, both of the diseases were instituted as notifiable diseases, where under the Prevention and Control of Infectious Disease Act 2000, it is obligatory to all medical practitioners to report cases, either confirmed or suspected of both DF and DHF to the nearest health office (DHO) within 24 hours.

Dengue infection is predominant in urban areas where 61.8% of the country's population lives, as compared to only 34% back in 1980 [2]. The disease keeps on increasing year by year. In 2009, dengue cases have more than doubled since the early of the year, with 3, 211 reported nationwide from 1-17 of January, compared to about half of the figure reported, which is 1, 514 in the same period in a year before [3]. Till the end of the year, there was a record total of 49, 335 cases [4]. The main factors that contribute to the spread of the disease are such as stagnant water collected in containers, pots and drains, global warming and also unpredictable rainfall. Besides that, the increasing of universal commerce and tourism also contribute to the expanding of dengue cases [5].

Manuscript received March, 2011; revised July 18, 2011.

Y. Yusof is with the College of Arts and Sciences, University Utara Malaysia. Phone: +604-928 4623; fax: +604-928 4753; （e-mail: Yuhanis@uum.edu.my).

Z. Mustaffa is with the College of Arts and Sciences, University Utara Malaysia. (e-mail: zuriani.m@gmail.com)

Reference [3] reported that 80% of the areas with dengue cases contribute to the high rates of breeding sites for Andes. Thus, there have been several numbers of campaigns 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 [6]. Obviously, current precaution steps that have been carried out such as awareness campaigns, education to the community and others are not adequate. Thus, other effort needs to be identified and this includes the ability to predict future dengue outbreak [7].

There are several approaches that have been utilized in predicting of future dengue outbreak [7]-[9]. Achieving high accuracy in prediction is important as it can lead to a well planned precaution programmer. Different training and/ or learning techniques approaches may lead to different prediction accuracy. Hence, it is significance to identify mechanisms that are able to produce high prediction accuracy in dengue outbreak.

The work presented in this paper incorporates the Least Squares Support Vector Machines in predicting dengue outbreak for five districts in Selangor, Malaysia. Prediction accuracy obtained in the undertaken experiment is compared against the one implemented using Artificial Neural Network model. This paper is organized as follows: literature review on existing work is provided in Section II while Section III describes LS-SVM in detail. Results and discussion is presented in Section IV and Section V summarizes the results and draws a general conclusion.

II. Literature Review

Reference [7] has presented four architectures for predicting dengue outbreak incorporating Neural Network (NNM) and Nonlinear Regression (NLRM) models. The study was done using dataset of dengue cases and rainfall level for five districts in Selangor. The data were from 2004 to 2005. From the undertaken experiment, it is shown that NNM yields better output compared to NLRM, in all architectures, and from the four proposed architectures, the last architecture perform finer result.

Study on predicting dengue hemorrhagic fever (DHF) in Thailand has been done by [8]. 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. Later, the supervised neural network is applied to predict future DHF outbreak. Result obtained

revealed that, 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.

Reference [9] studied on Wavelet transformation for data pre-processing before implementing Support Vector Machines (SVM)-based Genetic Algorithm in analyzing and predicting the dengue outbreak. The evaluation and fitness of entire individual in the population is carried out before entering next processes. The model was developed based on data collected in Singapore, from 2001 to 2006. From their study, they found that, for predicting, Support Vector Regression (SVR) performed better compared to a simple linear regression and also more reliable, even with the present of over-fitting.

Other than being used in diseases prediction, LS-SVM has also been used in taxation [10], electricity load [11], water quality[12] and many more. In [10], Genetic Algorithm-Support Vector Machine (GA-SVM) is utilized in predicting tax gross. In the developed model, GA is implemented to automatically tune the SVM parameters. The research was using tax gross data sets in China, from 1990-2001. In order to improve the operation speed and generalization performance in the proposed model, datasets used is normalized before training process. The empirical results revealed that the proposed approach yields better results as compared to artificial neural network (ANN) and grey model (GM) in tax gross prediction.

Reference [11] discussed the LS-SVM hybrid with bacterial colony chemo taxis (BCC) for short term load prediction. BCC is a novel category of bionic algorithm and was applied in the research to determine or optimize parameters of LS-SVM since finding appropriate parameters are very important for learning performance and generalization ability of LS-SVM. By applying BCC-LSSVM in the study, the parameters are automatically tuned by BCC. From the results, it showed that the proposed method proved that BCC-LS-SVM can automatically tune the parameters with high recognition rate and fast convergence rate. Finally, it able to achieve higher prediction accuracy with faster speed compared to neural network and LS-SVM with grid search.

Drinking water source quality prediction in Linux River, Guangzhou is presented by [12] where the study incorporated LS-SVM combined with Particle Swarm Optimization (PSO). The aim of the research is to solve the problems in Back Propagation (BP) which is reported to be slow to converge and easy to reach extreme minimum value. PSO is utilized to tune the LS-SVM parameters automatically, thus improve the efficiency of the results obtained. From the experiments conducted, PSO-SVM outperformed BPNN and ARIMA model in terms of prediction accuracy.

Prediction using ANN has presently reported to be considerable success in process prediction [13], however, besides its merits in establishing nonlinear models, ANN is also reported to suffer some problems. A few important issues must be put into consideration before applying the ANN models such as the network structure, learning rate parameter and also the normalization techniques for the input vector [14].The structure of ANN is very complex where it needs many tuning parameters and difficult to select network architecture in determining the number of hidden neuron. The learning speed of NN is reported to be comparatively slow [15] and easily stuck in local minima. As, NN adopts Empirical Risk Minimization (ERM), where in order to obtain generalization, it needs to reduce training error [16]; such an approach would yield a bad generalization performance. On the other hand, LS-SVM applies Structural Risk Minimization (SRM) where the generalization is obtained by minimizing the upper bound of generalization error rather than the training error [17]. LSSVM improves both training time and accuracy in comparison with other competitor prediction approach [15]. Thus, it is suggested that LS-SVM offers the ability to overcome demerits of NN model.

III. Least Squares Support Vector Machines

Support Vector Machines (SVM) is efficient approach for solving problems in nonlinear classification and regression [15]. It was originated by Bolster, Guymon and Vapnik [18]. It is a new machine learning method based on statistical learning theory. They have solved the over fitting problems, curse of dimensionality and local minimum which can be found in ANN. Presently, SVM has been proved to provide better solutions to boundary compared to Neural Network and reported to be good in generalization performance and has been implemented in numerous fields [16].

LS-SVM which stands for Least Squares Support Vector Machines is a reformulation of the SVM algorithm. It has been proposed by Sunken and Vandewalle [19] for the purpose to solve short term load prediction problem. LSSVM is able of performing a faster training process in huge scale problem compared to the standard SVM's. As a modified version of SVM, LS-SVM applies equality constraint instead of inequality constraint that has been used in SVM to obtain a linear set of equations [17], which simplifies the complex calculation and easy to train [11].

According to [13], the Least Squares Support Vector Machines (LS-SVM) prediction model generates outstanding performance in simulation and practical results, compared to Radial Basis Function (RBF) neural network predictor and Back Propagation (BP) neural network predictor.

The standard framework for LS-SVM estimation is based on the primal-dual formulation [11]. Given the dataset [10] $\{x_i, y_i\}_{i=1}^{N}$, the aim is to estimate a model of the form [19]:

$$y(x) = w^T \emptyset(x) + b + e_i \quad (1)$$

where $x \in R^n$, $y \in R$, and $\varphi(.): R^n \rightarrow R^{n_h}$ is a mapping to a high dimensional feature space. The following optimization problem is formulated [19]:

$$\min_{w,b,e} J(w,e) = \frac{1}{2} w^T w + \gamma \frac{1}{2} \sum_{i=1}^{N} e_i^2$$
$$\text{Subject to } y_i = w^T \varphi(x_i) + b + e_i, \quad i=1, 2\ldots,N. \quad (2)$$

With the application of Mercer's theorem [18] for the

kernel matrix $\boldsymbol{\Omega}$ as $\Omega_{ij} = K(x_i, x_j) = \varphi(x_i)^T \varphi(x_j)$, $i, j=1,..,N$ it is not required to compute explicitly the nonlinear mapping $\varphi(.)$ as this is done implicitly through the use of positive definite kernel functions K [19].
From the Lagrangian function:

$$\zeta(w,b,e;\alpha) = \frac{1}{2}w^T w + \gamma \frac{1}{2}\sum_{i=1}^{N} e_i^2 - \sum_{i=1}^{N}\alpha_i (w^T \varphi(x_i) + b + e_i - y_i) \quad (3)$$

where $\alpha_i \in R$ are Lagrange multipliers. Differentiating (3) with w, b, e_i and α_i, the conditions for optimality can be described as follow:

$$\begin{cases} \dfrac{d\zeta}{dw} = 0 \rightarrow w = \sum_{i=1}^{N}\alpha_i \varphi(x_i) \\ \dfrac{d\zeta}{db} = 0 \rightarrow \sum_{i=1}^{N}\alpha_i = 0 \\ \dfrac{d\zeta}{de_i} = 0 \rightarrow \alpha_i = \gamma e_i, i = 1,...,N \\ \dfrac{d\zeta}{\alpha_i} = 0 \rightarrow y_i = w^T \varphi(x_i) + b + e_i \end{cases} \quad (4)$$

by elimination of w and e_i, the following linear system is obtained [19]:

$$\begin{bmatrix} 0 & 1^T \\ y & \Omega + \gamma^{-1} I \end{bmatrix} \begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ y \end{bmatrix} \quad (5)$$

with $y = [y_1,...,y_N]^T$, $\alpha = [\alpha_1,...,\alpha_N]^T$. The resulting LS-SVM model in dual space becomes:

$$y(x) = \sum_{i=1}^{N}\alpha_i K(x, x_i) + b \quad (6)$$

Usually, the training of the LS-SVM model involves an optimal selection of kernel parameters and regularization parameter. Several kernel functions, viz. Gaussian radial basis function (RBF) Kernel, linear Kernel and quadratic Kernel are available. For this project, the RBF Kernel will be used which is expressed as:

$$K(x, x_i) = e^{-\frac{\|x - x_i\|^2}{2\sigma^2}} \quad (7)$$

where σ^2 is a tuning parameter which associated with RBF function.

IV. Methodology

In order to develop the prediction model and later evaluate it's effectiveness, flow of process as depicted in Fig. 1 is applied.

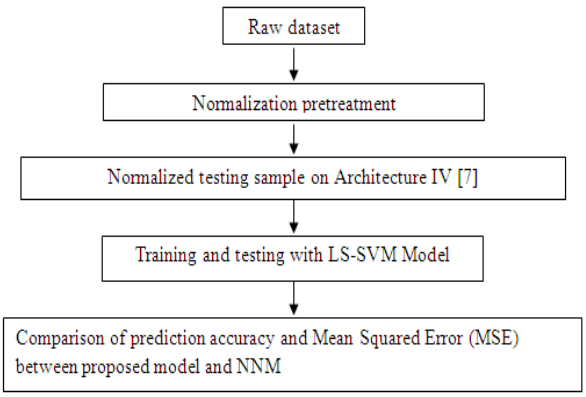

Fig. 1. Implementation Phase

A. Data Preparation and Pre-process

The data sets used in this project are as the one reported in [7] and this includes data on the following:

i. Dengue fever (DF) cases
ii. Neighborhood dengue cases
iii. Total of Rainfall

The data sets are of collection from five districts in Selangor, from 2004-2005. Each of the sample data set consist of 104 samples, which represent 52 weeks in 2 years, five variables (locations: Sepang, Hulu Selangor, Hulu Langat, Klang, Kuala Selangor) and 520 data (104 samples x 5 locations).

B. Normalization Process

All of the input and output data were normalized before training and testing processes in order to ensure data are not overwhelmed by each other in terms of distance measure. In this work, Decimal Point Normalization was applied. Using this approach, the data is normalized by moving the decimal point of values of the attribute. The formula used for Decimal Point Normalization is as equation 8. Examples of data before and after the normalization are depicted in Table 1 and 2 respectively.

$$v' = (v/10^j) \quad (8)$$

where;
v' = New value
v = Old value
j = The smallest integer such that **Max**(|v'|)<1.

TABLE I. Input Before Decimal Point Normalization

Original Input							
1	0	1	1	6	0	1	4
0	1	1	0	0	1	4	7
1	1	0	1	1	4	7	0

TABLE II. Input After Decimal Point Normalization

Normalized Input							
0.01	0	0.01	0.01	0.06	0	0.01	0.04
0	0.01	0.01	0	0	0.01	0.04	0.07
0.01	0.01	0	0.01	0.01	0.04	0.07	0

V. Experiment Setup

Using the prepared data sets, a model of experiment is performed. The data proportion applied is as stated below:

i. Training – 70% (350 data for training)

ii. Testing – *30% (150 data for testing)*

For the purpose to build an LS-SVM model (by using the RBF kernel), two tuning parameters are required, which are γ and $\sigma 2$. The first, γ, is the parameter for regularization, determining the trade-off between the fitting error minimization and smoothness of the estimated function whiles the latter, $\sigma 2$, is the kernel function parameter. The LS-SVM model is developed using LS-SVMlab Toolbox which is can be obtained in [20].

The undertaken experiment was conducted in Mat lab platform on Intel Atom processor, CPU N450 @ 1.66GHz, 982 MHz with 1 GB of RAM in Windows XP environment.

VI. Results And Discussion

This section discusses results obtained from the conducted experiments.

TABLE 3 Testing Prediction Results: LS-SVM vs. NNM

LOCATION	MSE		ACCURACY	
	LS-SVM	NNM	LS-SVM	NNM
Sepang	0.0021	0.0163	91.07	46.77
Klang	0.0015	0.0222	87.52	52.63
H. Selangor	0.0016	0.0139	87.38	54.01
H. Langat	0.0261	0.0894	90.28	91.13
K. Selangor	0.0003	0.0236	77.93	83.38
Average	6.32×10^{-3}	0.0331	86.84	65.58

The testing results of LS-SVM and NNM models are shown in Table 3. It is noted from the table that the LS-SVM model performs better than NNM in all locations in terms of MSE and accuracy. The average MSE obtained for LS-SVM model is 6.32 x 10-3 while for NNM is 0.0331. In terms of prediction accuracy, LS-SVM generated approximately 86.84% while only 65.58% was achieved by NNM.

Fig. 2 shows a graph comparison between LS-SVM, NNM and the target output for dengue cases in one of the district, Sepang. The testing processes involves from week 69 to week 97 which are 30% from the samples. The hyperparamaters of LS-SVM were set to γ= 300 and σ= 20. The MSE produced by LS-SVM is 0.0021 as compared to 0.0163 obtained using NNM.

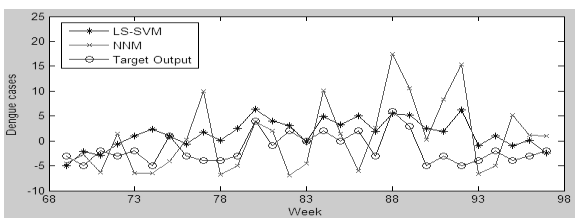

Fig. 2. Testing prediction graph for dengue cases in Sepang: LSSVM vs. NNM vs. target output

Result comparison for the district of Kuala Selangor between LS-SVM, NNM and target output is depicted in Fig. 3. The hyper-parameters for LS-SVM were set to 1000 for γ and 300 for $\sigma 2$.. This has resulted a MSE of 0.0003 for LS-SVM and 0.0236 by NNM.

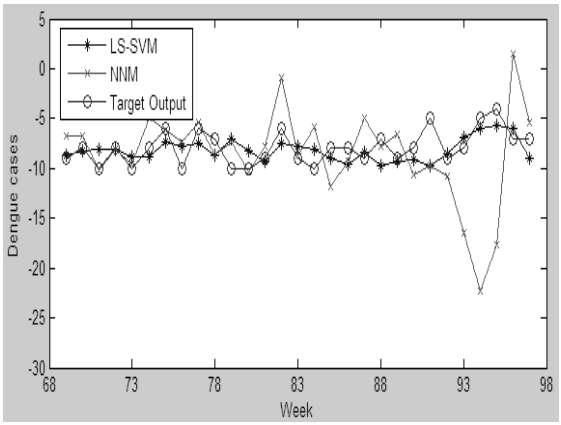

Fig. 3. Testing prediction graph for dengue cases in Kuala Selangor: LS-SVM vs. NNM vs. target output

Even though LS-SVM model is unable to accurately (accuracy higher than 90%) predict the unseen data, but still it does not over fit the training data especially for location in Sepang and Kuala Selangor, as depicted in Fig. 2 and Fig. 3.

Reference [15] reported that the learning speed of NN is comparatively slow. This is proven in this study where LSSVM model computes the results within 0.13694 seconds where else NN took about 452.94 seconds to train data for each location. Computation time is crucial especially when dealing with large size of dataset. Thus, in future, if dataset is large, it is suggested to use LS-SVM compared to NN.

VII. Conclusion

This paper proposed a model for predicting future dengue outbreak incorporating LS-SVM. From the undertaken experiments, it shows that LS-SVM capable to obtain good generalization ability compared to NNM, thus improving the prediction accuracy and MSE.

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 [9]. Furthermore, to determine tuning parameter is the vital of predict study. A hybrid algorithm could be applied to obtain the optimal parameter, thus improve the prediction accuracy and MSE.

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] G. Gusmao, S. C. S. Machado, and M. A. B. Rodrigues, "A new algorithm for segmenting and counting aedes aegypti aedes in ovitraps,"Proc. IEEE Annual International Conference in Engineering in Medicine and Biology Society (EMBC), 2009, pp. 6714-6717.

[2] K.T., Ang and S. Singh, "Epidemiology and New Initiatives in the Prevention and Control of Dengue in Malaysia," Dengue Bulletin, vol. 25, pp. 7-14, 2001.

[3] H. Yusop, "Dengue Cases Doubled," in *The Sun* Putrajaya, 2009.

[4] H. Hassan, "Worst Dengue Outbreak," in *The Straits Time* Putrajaya, 2009.

[5] M. Cuddehe, "Mexico fi ghts rise in dengue fever," *World Report,vol. 374, 2009*.

[6] B. Tan Kah, L. Koh Hock, and Y. Teh Su, "Modeling Dengue Fever Subject to Temperature Change," Proc. Sixth International Conference in Fuzzy Systems and Knowledge Discovery (FSKD '09), pp. 61-65.

[7] N. A. Husin, N. Salim, and A. R. Ahmad, "Modeling of dengue outbreak in Malaysia: A comparison of Neural Network and Nonlinear Regression Model," Proc. International Symposium in Information Technology (ITSim 2008), 2008, pp. 1-4.

[8] N. Rachata, P. Charoenkwan, T. Yooyativong, K. Chamnongthal, C. Lursinsap, and K. Higuchi, " "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 2008), 2008, pp. 210-214.

[9] Y. Wu, G. Lee, X. Fu, and T. Hung, "Detect Climatic Factors Contributing to Dengue Outbreak based on Wavelet, Support Vector Machines and Genetic Algorithm," Proc. World Congress on Engineering, 2008, pp. 1947-1949.

[10] L. Sheng, C. Zhong-jian, and Z. Xiao-bin, "Application of GA-SVM time series prediction in tax forecasting," Proc. *2nd IEEE International Conference* in *Computer Science and Information Technology (ICCSIT 200* pp. 34-36.

[11] L. Yang, "Short-Term Load Forecasting Based on LS-SVM Optimized by BCC Algorithm," Proc. 15th International Conference in Intelligent System Applications to Power Systems (ISAP '09), 2009, pp. 1-5.

[12] Y. Xiang and L. Jiang, "Water Quality Prediction Using LS-SVM and Particle Swarm Optimization," Proc. *Second International Workshop* in *Knowledge Discovery and Data Mining (WKDD 2009)*, pp. 900-904.

[13] C. Qisong, W. Yun, and C. Xiaowei, "Research on Customers Demand Forecasting for E-business Web Site Based on LS-SVM," Proc. International Symposium in Electronic Commerce and Security, 2008 ,pp. 66-70.

[14] T.Jayalakshmi, Dr.A.Santhakumaran, "Statistical for Normalization and Back Propogation for Classification", *International Journal of Computer Theory and Engineering* (IJCTE), Vol.3(1): 89-93, 2011, pp. 89-93

[15] M. Afshin, "Application of least squares support vector machines in medium-term load forecasting," Canada: Ryerson University (Canada), 2007, p. 46.

[16] X. Ying and Z. Hua, "Water supply forecasting based on developed LS-SVM," Proc. *3rd IEEE Conference* in *Industrial Electronics and Applications (ICIEA 2008)* pp. 2228-2233.

[17] J. Wu and D. Niu, "Short-Term Power Load Forecasting Using Least Squares Support Vector Machines (LS-SVM)," Proc. Second International Workshop in Computer Science and Engineering (WCSE '09), 2009, pp. 246-250.

[18] V. N. Vapnik, *the Nature of Statistical Learning Theory 2nd ed.* New York, 1995.

[19] M. Espinoza, J. A. K. Suykens, and B. De Moor, "Load Forecasting Using Fixed-Size Least Squares Support Vector Machines,"*Computational Intelligence and Bioinspired Systems*, 2005, pp. 1018-1026

[20] S. J. A. K. Pelkmans K., Van Gestel T., De Brabanter J., Lukas L., Hamers B., De Moor B., Vandewalle J., "LS-SVMlab: A Matlab/C Toolbox for Least Squares Support Vector Machines," ESAT-SISTA, K. U. Leuven, Leuven, Belgium 2002.

Yuhanis Yusof is a senior lecturer in College of Arts and Sciences, University Utara Malaysia. She obtained her PhD in 2008 from Cardiff University in the area of Computer Science focusing on information retrieval in digital libraries. She also has a Master Degree in Computer Science from University Saints Malaysia and a Bachelor of Information Technology from University Utara Malaysia. Her research interest is broadly in data analysis and management for large scale computing. This includes data mining (discovering patterns of interest from data), data warehousing, information retrieval and software reuse.

Zuriani Mustaffa obtained her B. Sc (Hons) in Computer Science (Software Engineering) from University Teknologi Malaysia in 2003. She just completed her M. Sc (Information Technology) from University Utara Malaysia. Her research interests are machine learning and artificial intelligence.