2010 International Conference on Business and Economics Research
vol.1 (2011) © (2011) IACSIT Press, Kuala Lumpur, Malaysia

A Comparison of Normalization Techiques in Predicting Dengue Outbreak

Zuriani Mustaffa
College of Arts and Sciences
Universiti Utara Malaysia
Sintok, Kedah
zuriani.m@gmail.com

Yuhanis Yusof
College of Arts and Sciences
Universiti Utara Malaysia
Sintok, Kedah
yuhanis@uum.edu.my

Abstract— **In Malaysia, dengue fever (DF) and the potentially fatal dengue hemorrhagic fever (DHF) remain to be a significant public health concern. Higher rainfall and unconcern attitude in the community were some of the factors that contribute to the increase of dengue cases. As number of dengue cases is increasing rapidly in Malaysia, more work need to be done in order to prevent this situation become critical. This includes work on predicting future dengue outbreak. This paper investigates the use of three normalization techniques in predicting dengue outbreak; Min-Max, Z-Score and Decimal Point Normalization. These techniques are incorporated in the LS-SVM and Neural Network (NNM) prediction model respectively. Comparisons of results are made based on prediction accuracy and mean squared error (MSE). Results obtained indicate that the LS-SVM is a better prediction model as compared to the NNM.**

Keywords-Least Squares Support Vector Machines; Support Vector Machines; Neural Network Model; Dengue fever

I. INTRODUCTION

In recent decades, dengue fever (DF) has emerged as one of the critical global public health concern, specifically in the tropical countries. Presently, it is one of the most universally disseminate insect-born virus infection, giving rise to 50 to 100 million cases annually, incorporating more than 100 endemic countries in the world [1]. DF is commonly found in tropical and sub-tropical countries. It is an acute febrile viral disease, regularly presented with some symptoms such as headache, bone or joint and muscular paints and rash [2]. A critical percentage of DF patients lead to a more serious form of disease, which is dengue hemorrhagic fever (DHF) [3]. According to World Health Organization, from 50-100 million cases of DF and 250-500 thousand cases of DHF, 24, 000 among them died due to the infection [4].

Unfortunately, up until now, there are still no vaccines available for DF. Hence, serious efforts are required to control and prevent this disease from become pervasive[5]. 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. There are several approaches that have been used to forecast future dengue outbreak [6], [7], [8].

In 2008, [6] have proposed four architectures for predicting dengue outbreak using Neural Network (NNM) and Nonlinear Regression (NLRM) models. The study was done using two datasets, where the first dataset was on dengue cases collected from five districts in Selangor and the second dataset includes rainfall data supplied by the Malaysian Meteorological Service. The data were from 2004 to 2005. From their experiment, it is shown that NNM produced better output compared to NLRM, in all architectures, and from the four proposed architectures, the last architecture performs finer result.

Study on predicting dengue hemorrhagic fever (DHF) in Thailand has been done by [7]. In their research, they proposed an automatic prediction system for DHF by applying entropy technique and Artificial Neural Network (ANN). Entropy is used to extract the relevant information that gives influence to the accuracy for the prediction process. Later, the supervised neural network is applied to predict future DHF outbreak. They concluded that by applying entropy technique, it would generate a better result as the entropy technique produces 85.92% accuracy while only 78.16% when entropy technique is not applied.

Reference [8] has proposed the 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.

Prediction using ANN has presently said to be considerable success in process prediction [9], nevertheless, besides its merits in establishing nonlinear models it also reported to cause some difficulties. In terms of structure, ANN is very complex. It requires many parameters to be tuned and difficult to select network architecture and in determining the number of hidden neuron. The training of NN is reported to be comparatively slow [10] and it also easy to get stuck into local minima. Other than that, NN applies Empirical Risk Minimization (ERM), which reduces the

learning error that usually yields a bad generalization performance [11]. 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. LS-SVM improves both training time and accuracy in comparison with other competitor prediction approach [10]. Thus, from the studies that have been done, the breakthrough of LS-SVM has become as a technique to improve solution and overcome demerits by NN.

The successful of machine learning normally rely on the quality of the data that they work on [12]. Thus, data pre-processing which includes data normalization is important where it may improve the accuracy and achieves the best performance for the tested data set. Realizing the importance of data pre processing in mining algorithms, [13] presented a different normalization techniques which was experimented on ID3 methodology. The normalization techniques used are Min-Max Normalization, Z-Score Normalization and Decimal Scaling Normalization. Dataset used is HSV dataset which was taken from the UCI repository. The empirical results indicated that Min-Max Normalization produced the best result with the highest accuracy, least complexity and shortest in learning speed.

This project proposes a prediction of dengue outbreak in Malaysia by implementing Least-Square Support Vector Machines (LS-SVM). Data fed into the model is pre-processed by normalizing it using the Min-Max, Z-Score and Decimal Point techniques respectively. This paper is organized as follows: a brief introduction to LS-SVM is given in Section II and Section III describes the methodology implemented in this study. Results and discussion is presented in Section IV and Section V summarizes the results and draws a general conclusion.

II. LEAST SQUARES SUPPORT VECTOR MACHINES

LS-SVM which stands for Least Squares Support Vector Machines is reformulations to the original SVM algorithm. It has been proposed by Suykens and Vandewalle [14] for the purpose to solve short term load prediction problem. LS-SVM able to do a faster training process in the huge-scale problem compared to standard SVM's. As a modified version of standard SVM, LS-SVM apply equality constraint instead of inequality constraint that has been used in SVM to obtain a linear set of equations [15], which it simplify the complex calculation and easy to train [16].

According to [9], the reformulation of standard SVM, Least Squares Support Vector Machines (LS-SVM) predicting model presented 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 [16]. Given the dataset $\{x_i, y_i\}_{i=1}^{N}$, the aim is to estimate a model of the form [14]:

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

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 [14]:

$$\min_{w,b,e} J(w,e) = \frac{1}{2} w^T w + \gamma \frac{1}{2} \sum_{i=1}^{N} e_i^2 \quad (2)$$

Subject to $y_i = w^T \varphi(x_i) + b + e_i$,
$i=1, 2\ldots,N$.

With the application of Mercer's theorem [17] 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*[14].
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:

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

By elimination of *w* and e_i, the following linear system is obtained [14]:

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

with $y = [y_1, \ldots, y_N]^T$, $\alpha = [\alpha_1, \ldots, \alpha_N]^T$. 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 is available. For this project, the RBF Kernel is 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.

III. METHODOLOGY

A. Data Preparation and Pre-Process

The data sets used in this project are as the one reported in [6] and this includes the following:

i. Dengue fever (DF) cases data set
ii. Neighborhood dengue cases data set (the same data set as DF cases data set)
iii. Rainfall data set

The data sets are data obtained for 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 sets (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 that data are not overwhelmed by each other in terms of distance measure. In this work, three types of normalization techniques were applied separately; Min-Max Normalization [6], Z-Score Normalization [12] and Decimal Point Normalization [13]. The purpose of having three different normalization techniques is to identify the best normalization to be used in this project.

The formula used for the min-max normalization is as equation (8):

$$X_n = \frac{X_0 - X_{\min}}{X_{\max} - X_{\min}} \quad (8)$$

where,

X_n = new value for variable X
X_0 = current value for variable X
X_{min} = minimum value in data set
X_{max} = maximum value in data set

In Z-Score Normalization, usually it is useful when the actual minimum and maximum of attribute A are unknown.

$$v' = ((v - \bar{A})/\sigma_A) \quad (9)$$

where;

v' = New value
v = Old value
$\bar{A}$ = Mean of attribute A
σ = Standard deviation of attribute A

For Decimal Point Normalization, the data is normalized by moving the decimal point of values of attribute A.

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

where;

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

C. 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 σ^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, σ^2, is the kernel function parameter. The LS-SVM model is developed using LS-SVMlab Toolbox which is can be obtained in [18].

This project is experimented in Matlab platform on Intel Atom processor, CPU N450 @ 1.66GHz, 982 MHz with 1 GB of RAM in Windows XP environment.

IV. RESULTS AND DISCUSSION

Data transformation such as normalization may improve the accuracy and efficiency of mining algorithms involving neural networks, SVM, *k*-Nearest Neighborhood, and clustering classifiers. All three normalization techniques applied are compared against each other. Table I reveals results obtained when LS-SVM is fed with data transformed using three different normalization techniques, respectively. The purpose is to see the effect of different normalization techniques in terms of MSE and accuracy.

TABLE I. MIN-MAX NORMALIZATION VS. Z-SCORE NORMALIZATION VS. DECIMAL POINT NORMALIZATION: LS-SVM

LS-SVM						
Approach	*Min-Max Normalization*		*Z-Score Normalization*		*Decimal Point Normalization*	
	MSE	*Acc (%)*	*MSE*	*Acc (%)*	*MSE*	*Acc (%)*
Sepang	0.0035	88.78	0.1243	61.82	0.0021	91.07
Klang	0.0091	82.15	0.5821	66.84	0.0015	87.52
H.Selangor	0.0121	84.98	0.4302	75.49	0.0016	87.38
H. Langat	0.0665	87.75	2.18	70.80	0.0261	90.28
K.Selangor	0.0021	66.34	0.1066	57.97	0.00034	77.93
Average	0.0187	82.00	0.6846	66.58	6.328×10^{-3}	86.84

Data depicted in Table I is obtained upon completing the testing process in experiment dealing with 70/30 data proportion. From the table, it is obvious that Decimal Point Normalization produces the lowest MSE compared to other techniques. This is followed by Min-Max Normalization where the MSE average is 0.0187 and 0.6846 for Z-Score Normalization. In terms of prediction accuracy, Decimal Point Normalization produced highest accuracy in all experiment locations with the average is 86.84%, followed by Min-Max with 82% and, Z-Score Normalization with 66.58%.

On the other hand, the Decimal Point Normalization produces the highest computational time where it took 0.16354 seconds to complete the experiment. This is followed by Z-Score with 0.1379 and Min-Max is identified to be the fastest, reporting only 0.13694 seconds. Even though computational time for Decimal Point is the highest, nevertheless the difference is very small, which is only 0.0266 seconds differs from Min-Max Normalization.

TABLE II. MIN-MAX NORMALIZATION VS. Z-SCORE NORMALIZATION VS. DECIMAL POINT NORMALIZATION: NNM

NNM						
Approach	*Min-Max Normalization*		*Z-Score Normalization*		*Decimal Point Normalization*	
	MSE	*Acc (%)*	*MSE*	*Acc (%)*	*MSE*	*Acc (%)*
Sepang	0.0121	52.52	0.2287	36.82	0.0163	46.77
Klang	0.0269	48.14	1.4154	42.29	0.0222	52.63
H.Selangor	0.0462	46.14	1.4256	33.00	0.0139	54.01
H. Langat	0.0965	78.20	1.8398	46.75	0.0894	91.13
K.Selangor	0.0471	44.77	0.3163	25.81	0.0236	83.38
Average	0.0458	53.95	1.0452	36.93	0.0331	65.58

Table II shows the results obtained from experiments conducted on three normalization methods by using NNM. From the tabulated results, it shows that Decimal Point Normalization produce the lowest MSE compared to the other two normalization techniques, with the average of 0.0331. This is followed by Min-Max Normalization which stated 0.0458 and finally, Z-Score Normalization, 1.0452. For the prediction accuracy, Decimal Point Normalization yields the highest accuracy in all locations; with the average of the accuracy 65.58%, followed by Min-Max Normalization, 53.95% and lastly, Z-Score Normalization, 36.93%. Such a result is similar to the one obtained using LS-SVM, hence suggesting that Decimal Point Normalization is the best technique in this work.

Nevertheless, even though Decimal Point technique produced good results in both MSE and accuracy, it generated the highest computational time. It took about 551.7379 seconds to complete the training as compared to Z-Score Normalization with 420.2293 seconds. The Min-Max Normalization generated the least time, reporting only 452.9432 seconds.

V. CONCLUSION

This paper has presented the use of three normalization techniques in predicting dengue outbreak using LS-SVM and NNM model. . From the undertaken experiments, it is suggested that LS-SVM and NNM can achieved better accuracy and MSE by using Decimal Point Normalization compared to the other two techniques (Min-Max and Z-score) . In addition, it also learned that that LS-SVM is competent in producing better results compared to NNM. Such a result supports existing work undertaken in the area of LS-SVM [8, 9, 11]. Nevertheless, the results that are produced are case dependent.

REFERENCES

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

[2] F. Ibrahim, M. N. Taib, W. A. B. W. Abas, G. Chan Chong, and S. Sulaiman, "A novel approach to classify risk in dengue hemorrhagic fever (DHF) using bioelectrical impedance analysis (BIA)," *Instrumentation and Measurement,* vol. 54, 2005, pp. 237-244.

[3] J. C. Tay and P. Tan, "Finding Intervention Points in the Pathogenesis of Dengue Viral Infection," Proc. IEEE 28th Annual International Conference in Engineering in Medicine and Biology Society (EMBS '06), 2006, pp. 5315-5321.

[4] Z. Hua, G. Shuji, Z. Wei, C. Lidan, Z. Hao, P. Liang, and C. Hong, "Bioinformatics Analysis of the Envelope Glycoprotein and Construction of Infectious RNA Transcripts of Dengue Virus," Proc. Frontiers in the Convergence of Bioscience and Information Technologies (FBIT 2007), 2007, pp. 256-260.

[5] 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), 2009, pp. 61-65.

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

[7] N. Rachata, P. Charoenkwan, T. Yooyativong, K. Chamnongthai, 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.

[8] 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.

[9] 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.

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

[11] 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), 2009, pp. 900-904.

[12] S. B. Kotsiantis, D. Kanellopoulos, and P. E. Pintelas, "Data Preprocessing for Supervised Learning," *Computer Science,* vol. 1, 2006, pp. 1306-4428.

[13] L. Al Shalabi and Z. Shaaban, "Normalization as a Preprocessing Engine for Data Mining and the Approach of Preference Matrix," Proc. International Conference in Dependability of Computer Systems (DepCos-RELCOMEX '06), 2006, pp. 207-214.

[14] M. Espinoza, J. Suykens, and B. Moor, "Fixed-size Least Squares Support Vector Machines: A Large Scale Application in Electrical Load Forecasting," Computational Management Science, vol. 3, 2006, pp. 113-129.

[15] 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.

[16] 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.

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

[18] 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