Simulation of Dengue Outbreak Prediction

Nor Azura Husin, Naomie Salim and Ab Rahman Ahmad
Faculty of Computer Science and Information System
University Technology Malaysia
81310 UTM, Skudai, Johor Bahru, Johor, Malaysia
Email address: nazurahusin1112@yahoo.com

Abstract___ Neural Network Model (NNM), Hidden Markov Model (HMM) and Regression Model (RM) are developed to predict the spread of dengue outbreak in Malaysia. The case study covered dengue cases data from Selangor, which include seven mukims and eight administrative districts in year of 2004 and 2005. Specific criteria concerned are location, time (weeks) and intensity of dengue cases. Critical discussion of some previous studies upon the performance of each approach reveals that NNM has several advantages over the two other models in prediction although some limitations are observed and this indicated that NNM might have the better predict of dengue outbreak. However, these models will be further studied by measuring their Root Mean Square Error (RMSE) to identify the best prediction model.

Keywords___ Neural Network, Hidden Markov, Regression, Dengue outbreak prediction.

1. Introduction

Dengue fever (DF) and the potentially fatal dengue haemorrhagic fever (DHF) continue to be an important public health problem in Malaysia. It has been epidemic in Malaysia for a long time (Ghee, 1993). In 1998, about 26,240 of dengue fever cases were recorded by the Ministry of Health, Vector Borne Disease Control Section (VBDC). Malaysia was reported to have higher case fatality rates (4.67%) compared with the neighboring countries like Thailand and Indonesia, with the case fatality rates of 0.3% and 0.5%, respectively. According to Gubler (1998), Malaysia has a good laboratory- based surveillance system; however, it is basically a passive system and has a little predictive capability. Problem may occur if one waits for laboratory confirmation of the case before notification. Delay in notification may lead to delay in control measure, which will further lead to occurrence of outbreaks, since dengue needs optimum time of management as the transformation of DF into severe form of dengue are only takes a very short period (WHO, 1985). One of the solutions is to implement a simulation of dengue spread in Malaysia, with emphasis on an early prediction of dengue outbreak (Gubler, 2002). It may improve public health problem in Malaysia since the accurate and well-validated simulation to predict the dengue outbreak is needed to enable timely action by public health officials to control such epidemics and mitigate their impact on human health (McConnell *et al.*, 2003). Therefore, simulation of dengue outbreak prediction that incorporates location, time and intensity (DF and DHF) are needed to help and produce a prediction for early identified the specific location, temporal and intensity of dengue outbreak accurately and rapidly (Nor, 2005). Unfortunately, no such study has been done to predict the dengue outbreak in Malaysia so far and there have been insufficient discussion about the suitable model to predict the future dengue outbreak. Therefore, several prediction models those have the ability to predict incorporate with the requisite criteria (time, location, intensity) will be counted in this study. From studies done upon prediction (Mooney *et a.l* (2002), Rath *et al.* (2002), Lee *et al.* (1998), McGauley (2003), Vandegriff *et al.* (2004)), Neural Network Model (NNM), Hidden Markov Model (HMM) and Regression Model (RM) are selected in this study to investigate the best method to simulate the prediction of future dengue outbreak.

2. Methodology

2.1 Problem Identification

At the first place, information and resources related to the study are collected by using the right techniques. This information must be obtained from the authorized and verified resources in order to fulfil the standard and particular criterion. Next stage is to identify problem to be solved, aims and

objectives and also the scope of the research. All this information may gives an overview of the study and provide the idea to solve the problem.

2.2 Collection of data

Data concerned in this study is dengue cases data from year of 2004 to 2005 that received from State Health District (SHD). These data are divided by weeks (each year consist of 52 weeks), intensity (DF and DHF) and locations, which include seven mukim and eight administrative districts of Selangor; Gombak, Majlis Perbandaran Selayang (MPS), Majlis Perbandaran Subang Jaya (MPSJ), Majlis Perbandaran Petaling Jaya (MPPJ), Majlis Perbandaran Shah Alam (MBSA), Hulu Langat, Majlis Perbandaran Kajang (MPKj), Majlis Perbandaran Ampang Jaya (MPAJ), Majlis Perbandaran Klang (MPK), Klang, Kuala Langat, Kuala Selangor, Hulu Selangor, Sepang and Sabak Bernam (Department of Statistics Malaysia, 2005).

2.3 Model Implementation

2.3.1 Neural Network Model

A neural network is a powerful data-modeling tool that is able to capture and represent complex input/output relationships. The most common neural network model is the Multilayer Perceptron (MLP) and Radial Basis Functions (RBF). However, MLP is chose in this study to predict dengue outbreak.

In this techniques, the researcher is prefer to use NeuCom © Commercial V0.81 software which developed by the Knowledge Engineering and Discovery Research Institute. Prediction process using NNM can be dividing into 3 steps, building the neural network structure, learning processes and testing process.

Building the NNM

Basically, the steps involved to build neural network are as the following: -

1- Define number of input node.
2- Define the form of training and testing.
3- Define the architecture
    a. Number of hidden layer
    b. Number of neuron in each hidden layer
4- Select the Learning rate (α) and Momentum rate (β).
5- Initialize the network with random weight
6- Decide the structure of NN
7- Determine the convergence criteria
8- Define the stop criteria for learning
9- Define the number of output neurons.

Training/ Learning Process

The process of learning and training involves forward propagation and backward propagation. Firstly, initial values of weights (θji, θkj), bias (wji, wkj) and minimum error ($Emin$) are set.

Testing Process / Validating Process

Testing process is carried out to validate the network on new or unknown data sets, called the testing set. This is a process, which is performed after training the network. A properly build and trained network which can yield the best performance on the validation samples would be the best accurate model. If the validate sample outputs are not acceptable then a new network is to be built undergoing the whole process of learning and testing.

2.3.2 Hidden Markov Model

HMM were first described in a series of statistical paper by Leonard E. Baum and other authors in second half of the 1960s. One of the first applications of HMM was speech recognition or optical character recognition, staring in the middle of the 1970s. In the second half of 1980s, HMM

began to be applied to the analysis of biological sequences, in particular *Deoxyribonucleic Acid* (DNA).

The HMM is tested for dengue outbreak prediction on a time series of DF and DHF intensity from year 2004 to 2005. This model was implemented in MATLAB software as it provides easy operation software by using Statistics Toolbox that includes fine functions for analysing hidden Markov models.

When applying the model from the data, there are two main objectives:

i. Learning the model from the data that is estimating the model parameters from the data. This objective solve by the forward-backward or Baum-Welch algorithm (Baum *et al*., 1970). The algorithm estimates the parameters of a HMM by Expectation Maximization (EM) to carry out the expectation steps efficiently
ii. The updated parameters are then passed to the backward algorithm or Viterbi algorithm (Forney, 1973), which extracts the best likely state sequence that generated the observed time series.

2.3.3 Regression Model

Regression is used to study relationship between interval-ratio variables in which a single dependent (criterion) variable regresses with one or several independent (predictor) variables (Bahaman *et al.*, 1999).

Regression method is a statistical method for determining the relationship between one or more independent variables and a single dependent variable. Various types of regression are studied before, but in this study, Multiple Linear Regression is chosen. In this technique, the researcher is preferred to use NeuCom © Commercial V0.81 software. When applying the model from the data, there are several steps. Each step is break down into the following:

i. Takes dataset that has 15 input variable and 1 output variable from 104 samples. The dataset may need to normalise first.
ii. Generate regression formula that approximates linearly the data samples.
iii. After the regression formula is generated from the data, it can be used as a prediction model for new input.
iv. Determine error between the approximated and the desired output values.
v. Test data on the same or difference dataset.

2.3 Measurement of model performance

At the first place, discussion of previous studies on the performance of NNM, HMM and RM is done to get an overview of the strengths and weaknesses of each approach. Then, collected data will be used to test the performance of these models to predict the spread of dengue outbreak of year 2004 and 2005. Then, the comparison of prediction performance between HMM, NNM and RM techniques are done by measuring their Root Mean Square Error (RMSE) as this method have been applied in various studies before (Masters (1994), Roliana (2001), Jastini (2003)). The technique that produced the least RMSE then will be choosing to simulate the dengue outbreak prediction. The RMSE method can be described as follow:

$$\text{RMSE} = \sqrt{\text{MSE}}$$

MSE is mean square error, which can be $\frac{\text{SSE}}{\text{N}}$ expressed as

SSE is the summation of square error and can be expressed as $\sum (e_i)^2$, while N is number of data.

3. Result and Discussion

A comparative study of NNM, HMM and RM has been carried out in order to identify strengths and weaknesses of each approach. The criteria of performance considered are application, effectiveness, theoretical and technique and ease of use. Prediction model with the highest performance then will be selected in order to simulate the dengue outbreak prediction. RM is mostly implied to predict diseases outbreak over the other two prediction models (Pfeiffer (1997), Mooney (2002)). HMM is rarely seen in study of disease outbreak (Rath, 2003), since the model is most commonly used in speech recognition, comparing protein sequence and timing of trading in financial markets (Schrodt, 1997). However, no study upon disease outbreak prediction has done before by using NNM as it was usually applied in market predictions, meteorological and network traffic forecasting (Edwards (1997), Patterson (1993), Bengio (1995)).

Some observations have been made by Hinich (1997) on the limited utility of standard linear forecasting models. He noted that stochastic linear model such as linear regression model provides very poor predictions if the system is highly auto regression and linear estimates are predicted on the stochastic disturbance term of process being independent. Nevertheless, NNM can be implied to solve a non-linear problem since these techniques are already adopted as numerical prediction and is practicality proven. Eftekhar *et al.* (2005) mentioned that RM is a commonly accepted statistical tool, which can generate excellent models. Its popularity may be attributed to the interpretability of model parameters and ease of use, although it has some limitation. He was also noted that NNM are appealing for a number of reasons, namely, they seem to learn without supervision, they can be created by workers with very little mathematical model building experience, and software for building NNM is now readily available. However, in a study done by Schrodt (1997) indicated that there is a clear interpretation to each of the parameters of *A* (transition probabilities) and *B* (observed symbol probabilities) metrics in HMM technique, which allow them to be interpreted substantively. This contrasts the NNM techniques that have a very diffuse parameter structure. These results are differing from the several early application of NNM in medicine, which reported an excellent fit of the NNM to a given set of data. The impressive results usually were derived from over fitted models, where too many free parameters were allowed. While, the `result showed that the RM has less potential for over fitting primarily because the range of functions it can model is limited (Eftekhar *et al.*, 2005).

According to Mooney (2002) and Kirkwood (1998), NN is more suitably applied to the problem like an open algorithm, or the problem from which a situation changes in time, compared to RM. They indicated that RM is become less reliable at the extreme end of the range of the source data on which it is based. While, HMM is stochastic rather that deterministic and it is specifically designed to deal with noisy and with indeterminate time (Schrodt, 1997). The limitation of NNM observed is that standardized coefficients and odd ratios corresponding to each variable cannot be easily calculated and presented as they are in RM. NNM analysis generates weights, which are difficult to interpret as they are affected by the program used to generate them. This lack of interpretability at the level of individual variables (predictors) is one of the most criticized features in NNM (Eftekhar *et al.*, 2005). However, Eftekhar *et al.* (2005) did mentioned that NNM is rich and flexible non-linear system, which show robust performance in dealing with noisy or incomplete data and have the ability to generalize from the input. They also claims that the NNM may be better suited than other modeling systems to prediction outcomes when the relations between the variables are complex, multidimensional and non-linear. Besides, NN is probably have the biggest potential within general purpose control as it is believed that their ability to model a wide class of systems in many applications can reduce time spent on development and offer a better performance than can be obtained with conventional techniques (Norgaad *et al.*, 2001). Furthermore, there are some theoretical advantages comparing a predictive NNM over RM. One such advantage is that NNM allows the inclusion of a large number of variables. Another advantage of the NNM approach is that there are not many assumptions that need to be verified before the models can be constructed (Eftekhar *et al.*, 2005).

4. Conclusion

As a conclusion, the critical discussion demonstrated that NNM is the most reliable model over the other two models to predict the spread of dengue outbreak. Although RM is widely used on disease outbreak prediction, the comparison had done shows that the NNM is likely to have the best dengue outbreak prediction. However, the effectiveness of the model must be confirmed and therefore, in the next phase, an analysis on test output will be conducted to obtain results. Comparison of

prediction performance between HMM, NNM and RM techniques will be done by testing data of dengue cases in Selangor from year 2004 to 2005 and measuring their Root Mean Square Error (RMSE). The model, which produced the least RMSE, will be selected to simulate the dengue outbreak prediction.

References

Bahaman, A.S. Turiman, S. (1999) "Statistical for Social Research with Computer Application." Universiti Putra Malaysia, Serdang.

Baum, L.E. Petrie, T. Soules, G. and Weiss, N. (1970). "A Maximization Technique Occuring in the Statistical Analysis of Probabilitic Functions of Markov Chains. Annals of Mathematical Statistics." 41:164-171.

Bengio, S.F. Fessant, F. Collobert, D. (1995). "A Connectionist System for medium-Term Horizon Time Series Prediction." In Proc. the Intl. Workshop Application Neural Networks to Telecoms." 308-315.

Edward, T. Tansley, D.S.W. Davey, N. Frank, R.J. (1997). "Traffic Trends Analysis using Neural Networks. Proceedings of the International Workshop on Applications of Neural Networks to Telecommunications 3." 157-164.

Eftekhar, B. Mohammad, K. Ardebili, H.E. Ghodsi, M. and Ketabchi, E. (2005). "Comparison of Artificial Neural Network and Logistic Models for Prediction of Mortality in Head Trauma Based on Initial Clinical Data." BMC Medical Informatics and Decision Making, 5:3 doi:10.1186/1472-6947-5-3.

Forney, G. D. (1973). "The Viterbi Algorithm". Proceedings of the IEEE. 61:268-278.

Ghee, L.K. (1993). "A Review of Disease in Malaysia." Pelanduk Publication.

Gubler, D.J. (1998). "Dengue and Dengue Haemorrhagic Fever." Clin Microbiol Rev, 113: 480-496.

Gubler, D.J. (2002). "How Effectivelly is Epidemiological Surveillance used for Dengue Programme Planning and Epidemic Response?." Dengue Bulletin, Volume 26.

Hinich, M. (1997). Forcasting Time Series." Paper presented at the 14th Summer Conference on Political Methodology. Columbus, Ohio.

Jastini Mohd Jamil. (2003). "Pengkelasan Terhadap Data Pra-Pendiskretan dan Pasca- Pendiskretan Menggunakan Set Kasar dan Rambatan Balik: Satu Perbandingan." Universiti Teknologi Malaysia: Tesis Sarjana.

Kirkwood, B.R. (1998). "Correlation and Linear Regression. Chapter 9 in Essentials of Medical Statistics." Blackwell Science Ltd. Oxford. 57-64.

Lee, S. Cho, S. Wong, P.M. (1998). "Journal of Geographic Information and Detection Analysis." 2: 233-242.

Masters, T. (1994). "Practical Neural Network Recipes in C++." New York: Academic Press.

McConnell K.J. Gubler, D.J. (2003). "Guidelines of the Cost-Effectiveness of Larval Control Programs to Reduce Dengue Transmission in Puerto Rico." Rev Panam Salud Publica. 14:1.

Mooney, J.D. Holmes, E. Christie, P. (2002). "Real Time Modelling of Influenza Outbreak- A Linear Regression Analysis." Euro Surveill. 7:12 pp 184- 187.

Norgaard, M. Ravn, O. Poulsen, N.K. Hansen, L.K. (1999). "Neural Network for Modelling and Control of Dynamic Systems: A Practitioner's." Springer-Verlag, London. pp 4-11.

Patterson, D.W. Chan, K.H. Tan, C.M. (1993). "Time Series Forecasting with Neural Nets: A Comparative Study." Proc. the International Conference on Neural Network Applications to Signal Processing." 269-274.

Pfeiffer, D.U. Duchateau, L. Kruska, R.L. Ushewokunze-Obatolu, U. Perry, B.D. (1997) "A Spatially Predictive Logistic Regression Model For Occurrence of Theileiriosis Outbreaks in Zimbabwe." Proceeding of 8th Symposium of the International Society for Veterinary Epidemiology and Economics, Paris, France. Special Issues of Epidemiologie et Sente' Animale, 31-32, 12.12.1-3.

Rath, T.M. Carreras, M. Paola, S. (2003) "Automated Detection of Influenza Epidemics with Hidden Markov". Technical Report, Department of Computer Science. University of Massachusetts at Amherst.

Roliana Ibrahim (2001). "Carian Corak Kelas Data Indeks Komposit BSKL Dalam Perlombongan Data Menggunakan Model Rambatan Balik, "Universiti Teknologi Malaysia: Tesis Sarjana.

Roslina Salleh @ Sallehuddin (1999). "Penggunaan Model Rangkaian Neural Dalam Peramalan Siri Masa Bermusim." Universiti Teknologi Malaysia: Tesis Sarjana.

Schrodt, P. A. (1997August). "Early Warbning of Conflict in Southern Lebanon Using Hidden Markov Models", American Political Science Association.

Vandegriff, J. Wagstaff, K. Ho, G. Plauger, J. (2005). "Forecasting Space Weather: Predicting Interplanetary Shocks using Neural Networks." Advances in Space Research. 2323-2327

WHO. (1985). "Viral Haemorrhagic Fevers: General Research Needed and Recommendations for Viral Haemorrhagic Fevers". World Health Organization, Geneva.