Thunderstorm Forecasting by using Artificial Neural Network

Ahmad Faiduddin Bin Ali
Faculty of Electrical Engineering
Universiti Teknologi MARA
40450 Shah Alam
Email: ahmadfa_8@yahoo.com.my

Abstract - **Thunderstorm is a form of weather characteristic containing strong wind, lightning, heavy rain and sometimes snow or hail. It can be associated with a cloud type of cumulonimbus. Depending on its type, thunderstorm has great potential to produce serious damage to human life and property. Therefore, there are many sophisticated instrument used to record weather data such as Doppler radar and satellite. By using these data, based on the statistical, mathematical or soft computing technique can be done in order to predict the occurrence the weather characteristic. This project presents the application of artificial neural network (ANN) in forecasting the thunderstorm occurrence in Shah Alam based on the meteorological data. Therefore, several three-layer feed-forward back-propagation ANNs were developed in Matlab and each network was evaluated by using cross-validation technique. A network with the best performance in terms of its R-value was selected as the best design. Thus, the forecasting of thunderstorm occurrence can be successfully done.**

Keyword – artificial neural network, cross validation, thunderstorm forecasting

I. INTRODUCTION

Thunderstorm is one of the global phenomena that can occur anywhere in the world at anytime. It is also known as electrical storm, lightning storm or hailstorm. This storm is a form of weather characteristic containing strong wind, lightning, heavy rain and sometimes snow or hail [1]. Although thunderstorm is generally very short-lived phenomena, it has great potential to produce serious damage to human life and property such as lightning, damaging straight-line wind, large sized hail, heavy precipitation and flooding [2].

Basically, thunderstorm can be associated with the cloud called cumulonimbus cloud. Cumulonimbus cloud can be categorized into family D of cloud which developed in vertical way. It can be recognized by very tall and large cloud appearance. This cloud usually formed from cumulus cloud which grows vertically instead of horizontally [3]. Therefore, some of the parameter that produces this cloud is used as the input for neural network based model.

Artificial neural network (ANN) is a branch of artificial intelligence (AI). It is a mathematical model or computational model that tries to simulate the structure or functional inspired by biological neural network [4]. It consists of an interconnected group of artificial neurons and processes information using connectionist approach to computation. In most cases, ANN is an adaptive system that changes its structure based on external or internal information that flows through the network during the learning phase. Therefore, ANN is suitable for prediction and forecasting tasks.

ANN has several advantages. As stated in [5], the advantages of the ANN are performing task that is a linear program cannot and does not need to be reprogrammed. Besides, ANN can be implemented in any application without any problem. But, ANN also comes with several disadvantages such as it needs training to operate and require high processing time for large network. It also needs to be emulated due to different architecture with microprocessor.

This project presents the application of ANN in forecasting the thunderstorm occurrence in Shah Alam based on the meteorological data. In order to determine the best network, several feed-forward back-propagation of ANN designs were constructed using Matlab and each network was evaluated by using cross validation technique. ANN model with the best performance in terms of its R-value was selected as the best design. Thus, the forecasting of thunderstorm occurrence can be successfully done.

II. METHODOLOGY

The research design for developing the forecasting of thunderstorm occurrence consists of two stages. They are data collection and development of the ANN.

A. Data Collection

There are two types of data be used in designing the ANN, the meteorological data which acts as the input and thunderstorm occurrence as the target output. These data is taken at city of Shah Alam which cover

the whole month of April because of highest number of thunderstorm occurs in this month for every year. Besides, the data used is hourly data which be collected from Malaysian Meteorological Services (MMS). Meteorological parameters such as pressure, moisture difference, wind and type of cloud were chosen as they play important role in shaping the occurrence of thunderstorm.

1) Pressure

Pressure is one the important parameters that determine the weather condition. This is due to variation of pressure for certain time will lead to wind. Therefore, the pressure measurement is needed to predict the atmosphere phenomena.

2) Moisture Different

Moisture different is the estimation of the content of the water of the atmosphere. That's mean the different between dry bulb temperature (T) and dew-point temperature (T_d) is actually the moisture different. Furthermore, the moisture different is inversely proportional with relative humidity which mean when moisture different is high, the relative humidity is low and vice-versa. Moisture different is important because it transfers the heat from lower level to higher level. Therefore, dry bulb temperature, dew point temperature and relative humidity must be considered as an input for ANN model.

3) Wind

Wind has the ability to impact surface transportation. This ability is one of the factors that cause various hazardous weather conditions Therefore, the measurement of speed of wind must be considered in order to forecasting the thunderstorm occurrence.

4) Type of Cloud

Cloud type is use to determine either associated with thunderstorm or not. Cumulonimbus cloud is the one that associated with the thunderstorm. Therefore, type of cloud is important to be considering as input for ANN because it shows the existence of thunderstorm.

B. Development of the ANN

From Figure 1, ANN is a network that formed in three layers. The first layer is the input layer where the selected parameter comes into ANN and followed by hidden layer and output later which the output is produced. The input for ANN is the meteorological data while the output is the thunderstorm occurrence. Each layer consists of one or more nodes, represented in this diagram by the small circles. The lines between the nodes indicate the flow of information from one node to the next.

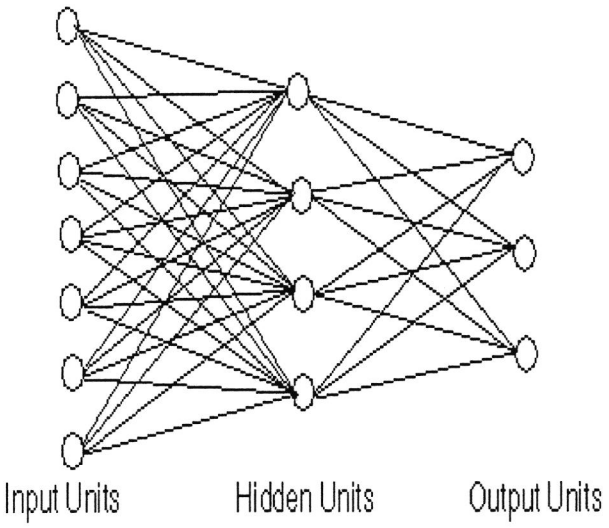

Figure 1: Schematic Diagram of ANN

In order to determine the best network, several feed-forward back-propagation of ANN designs were constructed and each network was evaluated by using cross validation technique. Cross validation is a model evaluation method that estimates the generalization error based on re-sampling.

1) Training Process

In order to determine the best network, each network must be trained and tested to estimate its performance configuration. The best result in terms of its R-value is selected as the best ANN model. This could be done by using cross validation technique which estimates the generalization error based on re-sampling. It involves partitioning a sample of data into complementary subset where analysis is performed on one subset (training set) and validating the analysis on the other subset (evaluation set).

The cross validation technique can be divided into two types. They are the holdout cross validation and K-fold cross validation. For holdout technique, available data is split into two sets called training set and testing set respectively. The testing set is held out and not been looked during training process. That means, holdout technique use the training set only. Then, the training set is subdivided into training set and evaluation set. The training set is used to train the ANN model while evaluation set is used to evaluate the ANN model by estimating its mean absolute error (MSE). This process is repeated for each ANN model. The advantage of this technique is that it avoids the overlap between training and testing data and takes shorter time to compute the ANN model.

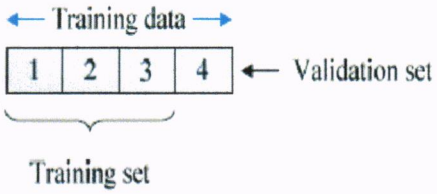

Figure 2: Partitionig of Data for K=4

In K-fold cross validation, the training data is partitioned into K subset. Of the K subset, a single subset is retained as the evaluation set while the remaining subsets are used as training set as shown in Figure 2.

The training set is used to train the ANN model while evaluation set is used to evaluate the ANN model by estimating its mean absolute error (MSE). This process is then repeated K times with each of the K subset used once as the evaluation set. Then, the averaged of the each K is been obtained to produce a single estimate for ANN model. This process then is done for each ANN model in order to determine the best network.

2) Testing Process

Testing process is the process that measure the performance of the trained model. It can be measured by performing a linear regression analysis between model output and target ouuput. The regression coefficient, R with value close to one shows that there is a strong correlation between both outputs. Then, the best model is trained once again by using the whole training data and testing data. It is suppose, this best model should be able to forecast the thunderstotm occurrence by measuring its R-value. Figure 3 shows the whole process of development of the ANN in flowchart.

III. RESULT AND DISCUSSION

The data were selected from meteorological based on city of Shah Alam which cover the whole month of April. There were 720 patterns selected, 480 were utilized for training the model and 240 were employed for the testing process. The meteorological will act as the input data while the thunderstorm occurrence data as the target output. Therefore, the developed ANN model receives the input data and target output data to produce the network output.

Several three-layer feed-forward back-propagation ANN models were developed in Matlab. The network configuration such as number of neuron and transfer function were determined heuristically. The learning rate and momentum constant were chosen as 0.5 and 0.9 respectively. These values were kept constant for each model design at this stage in order to observe the effect of different number of neuron and transfer function.

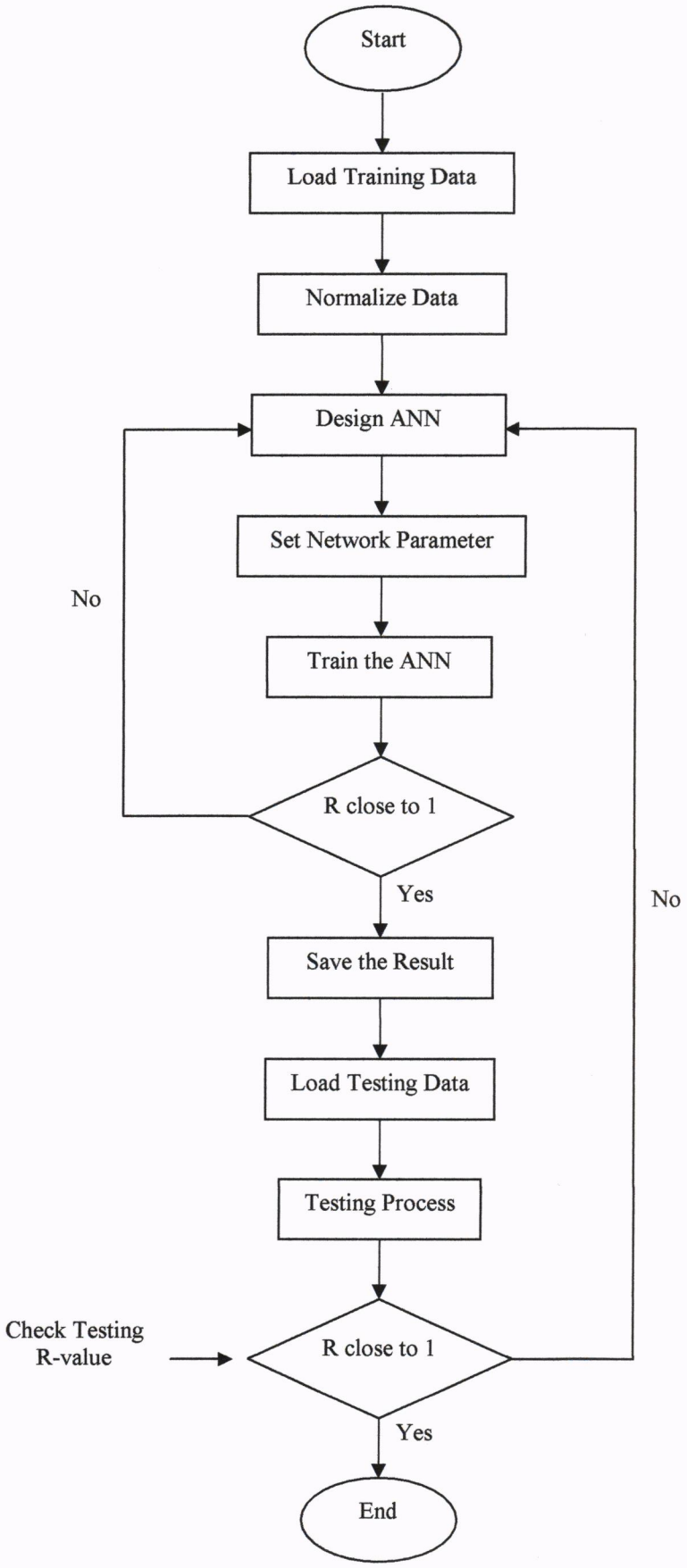

Figure 3: ANN Algorithm

In holdout cross validation, the training data is split into training and evaluation sets. The training set is used to train the ANN models while the evaluation set is used to evaluated the ANN models. For K-fold validation, the training data is split into four equally sets. One of these sets is used as evaluation set while the remaining sets as the training set. Then, the same process as in holdout cross validation are done in order to determine the best ANN model.

The comparison between holdout cross validation and K-fold cross validation depend on their R-value. The ANN model with the highest R-value is selected as the best network. This is due to R-value which closed to one shows that there is a strong correlation between target outputs and network outputs while R-value close to zero indicates otherwise.

From Table 1, the best R-value for holdout is 0.8349 while for K-fold is 0.8655. Based on these, it can be concluded that the ANN model trained using K-fold cross validation gives better result that holdout cross validation. It can also be seen that the best ANN model for forecasting the thunderstorm occurrence was a [8,5,1] configuration with tansig, logsig and purelin as the transfer function trained using Levernberg Marquardt technique. Figure 4, 5, 6, 7 and 8 show the best result in terms of R-value for holdout cross validation and R-value for each partition for K-fold cross validation respectively.

TABLE 1: Comparison Between Holdout and K-fold Cross Validation

No. of Neuron	Transfer Function	R-value	
		Holdout Validation	K-fold Validation
10, 8, 1	Logsig Logsig Purelin	0.8349	0.7998
10, 8, 1	Logsig Tansig Purelin	0.8314	0.8203
10, 8, 1	Tansig Logsig Purelin	0.8297	0.8202
10, 8, 1	Tansig Tansig Purelin	0.7646	0.8381
8, 5, 1	Logsig Logsig Purelin	0.7430	0.8313
8, 5, 1	Logsig Tansig Purelin	0.7701	0.8294
8, 5, 1	Tansig Logsig Purelin	0.8221	0.8655
8, 5, 1	Tansig Tansig Purelin	0.7458	0.8653

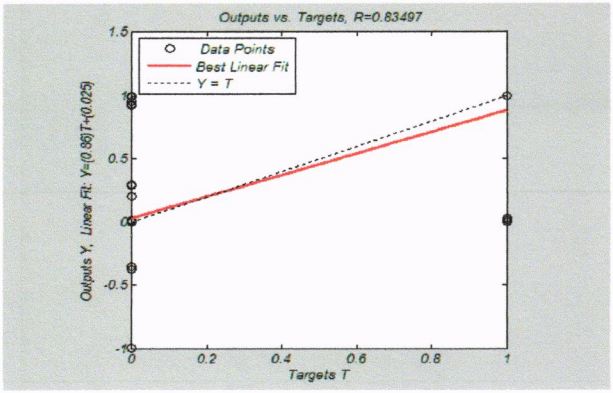

Figure 4: The Best R-value for Holdout Cross Validation

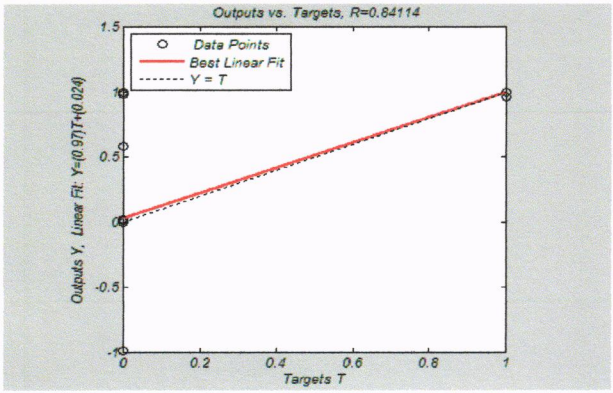

Figure 5: The R-value for Partitioning for K=1

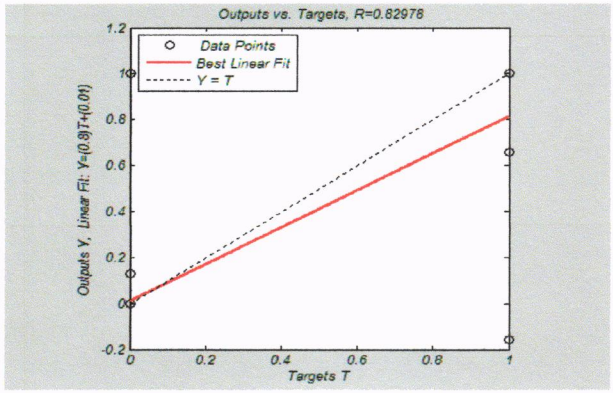

Figure 6: The R-value for Partitioning for K=2

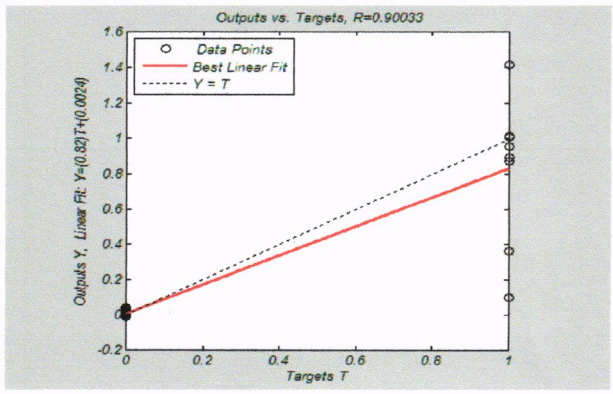

Figure 7: The R-value for Partitioning for K=3

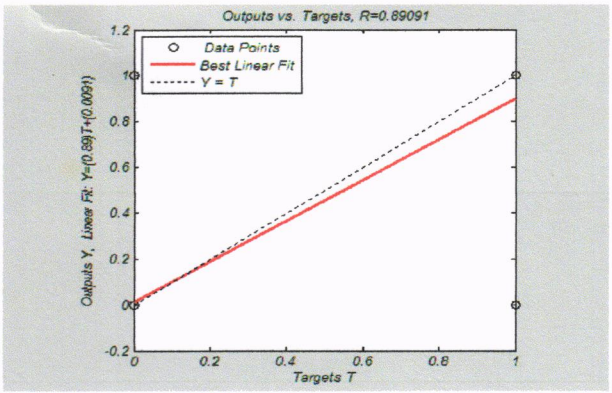

Figure 8: The R-value for Partitioning for K=4

After the cross validation technique, the best ANN model is trained using both training and evaluation sets as the training data and evaluated by using the testing data. For this time, the learning rate (lr) and momentum constant (mc) were varied between 0 to 1 in order to improve the performance of the ANN model.

Table 2 and 3 show that the R-value varies when the values of the learning rate and momentum constant were varied. The best performance configuration for forecasting the thunderstorm occurrence are 0.8 for learning rate and 0.2 for momentum constant. The R-value obtained by this model was 0.8907. Table 4 shows the best properties for ANN model in order to forecast the thunderstorm occurrence while Figure 9 shows the best R-value for this properties.

TABLE 2: Performance of Momentum Constant (mc)

Learning Rate (lr)	Momentum Constant (mc)	R-value
0.9	0.1	0.8307
	0.2	0.8877
	0.3	0.8346
	0.4	0.8033
	0.5	0.8377
	0.6	0.8433
	0.7	0.8033
	0.8	0.8868
	0.9	0.8441
	1.0	0.8776

TABLE 3: Performance of Learning Rate (lr)

Momentum Constant (mc)	Learning Rate (lr)	R-Value
0.2	0.1	0.7996
	0.2	0.8455
	0.3	0.8152
	0.4	0.8191
	0.5	0.8069
	0.6	0.8583
	0.7	0.8237
	0.8	0.8907
	0.9	0.8783
	1.0	0.8564

Table 4: Properties to Developed Network for Thunderstorm Forecasting

ANN Properties	Properties
Network Configuration	[8, 5. 1]
Transfer Function	Tansig, Logsig, Purelin
Learning Rate	0.8
Momentum Constant	0.2
Training Technique	Levernberg-Marquardt
Epochs	100
Regression Coefficeint, R	0.8907
Accuracy	89.07%
Training Pattern	480
Testing Pattern	240

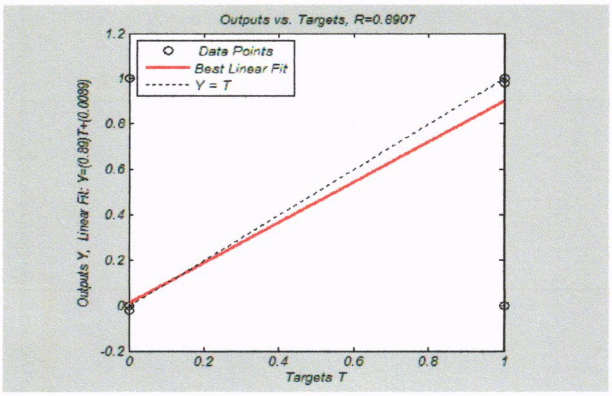

Figure 9: The Best of R-value for ANN Model for Thunderstorm Forecasting

IV. CONCLUSION

This project mainly presents an approach to forecast the thunderstorm occurrence based on the meteorological data. Several three-layer feed-forward back-propagation networks were developed by using Matlab. Then, cross validation technique was employed in order to determine the best ANN design by evaluating its mean absolute error. The result shows that the K-fold cross validation produced a better result in terms of the R-value than holdout cross validation. The optimal values of momentum constant and learning rate were found to 0.2 and 0.8 respectively. They were obtained by testing all possible values within the range of 0 to 1. Therefore, the forecasting of thunderstorm occurrence can be successfully done. In future, it is recommended to use large data in order to improve the thunderstorm forecasting. It is either by adding some new parameter such as low pressure area and vorticity for input data or increase the number of data by using data for a year.

ACKNOWLEDGEMENT

The author would lke to express its gratitude to Cik Dalina Bt. Johari, project supervisor for her valuable advice, ideas and critical guidance throughtout the preparation of this project. Besides, the author also would like to express its appreciation to Malaysian Meteorological Services (MMS) for the supply of the meteorological data.

REFERENCES

[1] S. Choudry, S. Sitra and H. Chakraborty, "A Connectionist Aapproach to Thunderstorm Forecasting".

[2] Ed. Richard C. Dorf, Principe, J.C., "Artificial Neural Networks," The Electrical Engineering Handbook, Boca Raton: CRC Press LLC, 2000.

[3] J. Schneider and A. W. Moore, "A Locally Weighted Learning Tutorial using Vizier 1.0," 1997.

[4] D. Johari, T.K.A Rahman, I. Musirin, "ANN Model Selection and Performance Evaluation for Lightning Prediction System".

[5] S.N. Sivanandam, S. Sumathi and S.N. Deepa, "Introduction to Neural Networks using MATLAB 6.0", Feed Forward Network, pp.184-219. Tata McGraw Hill Publishing Company Limited, 2006.

[6] J. Schneider and A. W. Moore, "A Locally Weighted Learning Tutorial using Vizier 1.0," 1997

[7] Elia Erwani Binti Hassan, " An Application of Artificial Neural Network on Short Term Load Forecasting using Back Propagation Algorithm" pp 40, 1998

[8] D. W. McCann, "A Neural Network Short-Term Forecast of Significant Thunderstorms," *Weather and Forecasting*, vol. 7, pp. 525-534, 1992.

[9] Ron Kohavi "A Study of Cross Validation and Bootstrap for Accuracy Estimation and Model Selection", International Joint Conference on Artificial Intelligent (IJCAI), 1995.

[10] Dasgupta D., "Artificial Neural Networks and Artificial Immune System Simulation and Differences", Proc. of the IEEE SMC, pp 873-878, 1997.