Runoff Prediction with a Combined Artificial Neural Network and Support Vector Regression

Ratiporn Chanklan, Nuntawut Kaoungku, Keerachart Suksut, Kittisak Kerdprasop, and Nittaya Kerdprasop

***Abstract*—Water is an important part of our daily lives: food, manufacture, agriculture, etc. When water is not enough for all population, it leads to many undesirable impacts including drought, famine and death. The solution to this problem is the good management of water resources. The management of water resources is planning and designing of projects related to water. The runoff prediction is one major part of planning. It is a complex process and it also needs an adequate modeling technique for accurate prediction. Therefore, we propose to use combined algorithms to improve prediction performance. Our combination includes the two powerful methods: Artificial Neural Network (ANN) and Support Vector Regression (SVR). The root mean square error (RMSE) and the correlation coefficient (R) are two criteria that we use to evaluate the model performance regarding the comparison between actual runoff and the prediction made by our model. We also compare performance of our model against the other algorithms: Linear Regression, ANN, and Support Vector Machines. The comparison results show that our proposed method shows the best performance and the combined model is also quite accurate on predicting the peak runoff values during heavy rain season.**

***Index Terms*—Runoff prediction, artificial neural network, support vector regression, Mun Basin.**

I. Introduction

Currently, people are experiencing both direct and indirect impacts from droughts such that it results in water usage restrictions, more frequent forest fires, reduced crop yields, loss of livestock, and many others [1]. On the contrary, people also experience floods in larger areas year by year. The efficient management of water resources is one way to protect the two water problems: flood and droughts. A knowledge to make accurate runoff prediction can obviously help planners and water management policy makers to know in advance the water volume as either enough or not enough for the demand of people use and also can improve efficiency on flood and drought control.

Artificial neural network (ANN) shows good performance on predicting outcomes from many complex processes and recognizing patterns [2]. It has also been used to create computational model to predict several kinds of outcomes in the hydrology research, such as stream flow forecasting, rainfall-runoff modeling, water quality and management modeling [3], [4]. Besides ANN, Support vector regression (SVR) is another accurate technique that has been recently applied to predict runoff and [5], [6]. SVR has been known to show high performance on learning solution in complex problems [7].

In this work, we propose a new method to predict monthly runoff. In our method, the better performance can be achieved through the combination of ANN and SVR. The results turn out that our model can predict both low and high runoff values. A literature normally uses statistical values such as mean absolute percentage error (MAPE), coefficient of determination (R^2), correlation coefficient (R), and root mean squared error (RMSE) to compare performance of a prediction algorithm. In our work, we adopt two metrics: R and RMSE.

Manuscript received September 19, 2017; revised January 20, 2018. This work was supported by grant from Suranaree University of Technology through the funding of Knowledge and Data Engineering Research Units.

The authors are with the School of Computer Engineering, Suranaree University of Technology (SUT), 111 University Avenue, Muang, Nakhon Ratchasima 30000, Thailand. (corresponding author: R. Chanklan; Tel: +66994696164; e-mail: arc_angle@hotmail.com, nuntawut@sut.ac.th, mikaiterng@gmail.com, kerdpras@sut.ac.th, nittaya@sut.ac.th).

II. Background Theories

A. Artificial Neural Network

ANN has been developed base on operations of the human brain. It is the most widely used tool in hydrology. The network has three layers: input layer, hidden layer and output layer. The input nodes are name node in input layer. The number of attribute in data is equal to the number of input nodes. The hidden nodes are name node in hidden layer, the number of nodes is defined by a user, and this layer can be more than one layer. The output nodes are name node in output layer, the number of nodes is equal to the number of target on data. The network are connected between the nodes with line, and each line has weight. ANN learning is to find proper weight on each line in the network. The proper weight is the one that can best separating training data into the corrected target groups. There exist several architectures of ANN. In this work, we use feed-forward neural networks. The transfer function in the hidden layer is a linear function.

B. Support Vector Regression

SVR works by transforming the input data into a high-dimensional feature space by linear or nonlinear mapping. SVR is the most popular application form of Support Vector Machine (SVM). The intuitive idea of SVR on predicting future data is illustrated in Fig. 1.

The goal of SVR is to find a function $f(x)$ that has at most ε deviation from the actual target value y_i for all the training data [8]. This linear function f is shown in equation 1. A training data is a set of input-target pairs, $\{(x_1, y_1), \ldots, (x_i, y_i)\}$

doi: 10.18178/ijmlc.2018.8.1.660

$\subset X \times R$.

$$f(x) = \langle w, x \rangle + b \quad (1)$$

When a function $f(x)$ is hyperplane, the size of ε is margin, the symbol $\langle \cdot, \cdot \rangle$ is the dot product in X, $b \in R$ and $w \in X$. The hyperplane has small margin in which the SVR has to find it. This small margin tries to keep all the data lying inside as much as possible. The margin can be calculated as in equation 2.

$$minimize\ \frac{1}{2}\|w\|^2\ subject\ to \begin{cases} y_i - \langle w, x_i \rangle + b \leq \varepsilon \\ \langle w, x_i \rangle + b - y_i \leq \varepsilon \end{cases} \quad (2)$$

Through equation 2, it is implicitly assumed that this is a function f that approximates all the pairs (x_i, y_i) with precision. If it is not possible to keep all data inside the margin, the slack variables ξ_i, ξ_i^* can be introduced to solve the problem. This can be stated according to the equation 3.

$$minimize\ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^*)$$
$$subject\ to \begin{cases} y_i - \langle w, x_i \rangle + b \leq \varepsilon + \xi_i \\ \langle w, x_i \rangle + b - y_i \leq \varepsilon + \xi_i^* \end{cases} \quad (3)$$

The constant $C > 0$ directs the choice between the flatness of f and the amount up to which deviations larger than ε are tolerated.

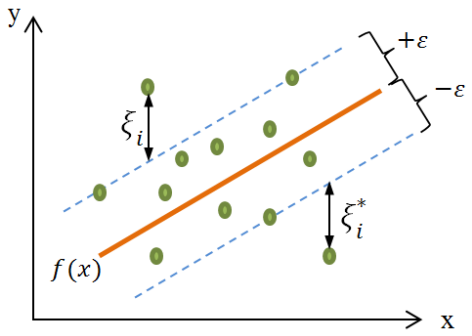

Fig. 1. Example of linear support vector regression.

C. Correlation Coefficient

The correlation coefficient is denoted by R and its numerical value is between -1 and 1. The plus sign means the correlation of the two variables (x and y) moves in the same direction, whereas the minus sign infers opposite direction. The magnitude expresses the strength of the relationship; the higher is the stronger regardless of the sign. No relationship is the magnitude that closes to 0. The correlation coefficient can be computed using equation 4.

$$R = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{\left[n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right]\left[n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2\right]}} \quad (4)$$

where x_i is the value of variable x or predicted value (when i = 1, 2, … , n), y_i is the value of variable y or actual value, and n is the total number of samples.

D. Root Mean Squared Error

The Root Mean Squared Error is denoted by RMSE and used to measure performance of the model. The RMSE can be calculated as in equation 5.

$$\text{RMSE} = \sqrt{\left(\frac{\sum(T_i - O_i)^2}{N}\right)} \quad (5)$$

where N is the number of all data, O_i is the predicted value by the model, and T_i is the actual value, i = 1,2, …, N.

III. Materials and Methods

Our study area is Mun Basin (Fig. 2), the largest basin in the North-eastern region of Thailand.To build a predictive model, we use temperature data from the National Statistical Office (http://www.nso.go.th), monthly rainfall, runoff and the number of rainy days data from the Meteorological Department (http://www.hydro-4.com) and Normalized Difference Vegetation Index (NDVI), from the NOAA STAR (http://www.star.nesdis.noaa.gov). This research use RStudio as the analysis tools in our experiments. The runoff data are from two sources: M145 and M173 station.

Fig. 2. The study area: Mun Basin, Thailand.

M145 station locates at Ban Wang Takhian, Amphur Pak Chong, in Nakhon Ratchasima Province, Thailand. We use the 18-year data during 1998 to 2015. We spilt the 13-year data (1998-2015) to be training data and the remaining 5-year data (2011-2015) are for testing the model performance.

M.173 station locates at Ban Non Sa-at, Amphur Chokchai, in Nakhon Ratchasima Province, Thailand. The training data is a ten-year period (2002-2011) and the test data is the four-year period (2012-2015).

Then, we use rainfall, runoff, the number of rainy days, temperature and NDVI with the lagging time 1-month and 2-month. These data are input into two learning algorithms: ANN and SVR. Then, the average rainfall is lagged 1-month (average rainfall$_{t-1}$) to select the best subset of data from the initial training set for appropriate algorithm. The average rainfall can be calculated as in equation 6.

$$\text{average rainfall}_{t-1} = \frac{\sum_{i=1}^{N} a}{n} \quad (6)$$

Where a is the average monthly rainfall in lagged 1-month over the training years, N is the number of training data (when i=1, 2, .., N), and n is the number of years from training data.

Our combined prediction model has a flow as shown in Fig. 3. The selection of either ANN or SVR model is based on the amount of rain. If the rainfall in lagged 1-month (rainfall$_{t-1}$) has value less than average rainfall$_{t-1}$, then apply the SVR

model because based on our observation this kind of model is good at prediction on normal or drought situation. But if the rainfall in lagged 1-month (rainfall$_{t-1}$), has a value higher than the average rainfall$_{t-1}$, then apply the ANN model that is good in predicting the near-flooding situation.

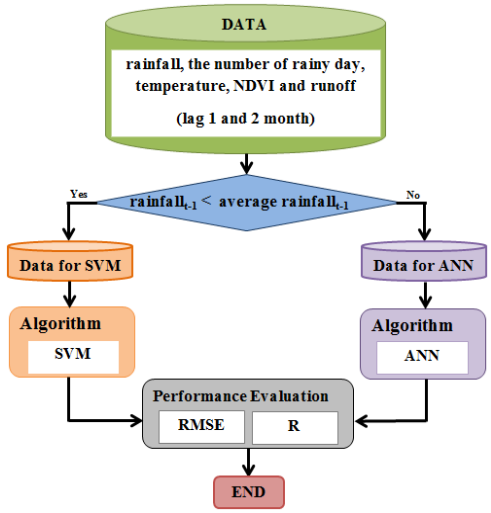

Fig. 3. The modeling process for runoff prediction.

The final step is the performance evaluation using the two statistical values: Correlation Coefficient (*R*) and Root Mean Squared Error (RMSE). The R and RMSE are computed from experimentation with the separate test data.

IV. Experimental Results

TABLE I: Runoff Prediction Performance from ANN Model

Station	ANN topology	R	RMSE
M145	**10-1-1**	**0.67**	**8.17**
	10-2-1	0.59	8.79
	10-3-1	0.66	8.20
	10-4-1	0.43	10.66
	10-5-1	0.51	9.68
	10-6-1	0.66	8.28
	10-7-1	0.62	8.78
	10-8-1	0.48	10.48
	10-9-1	0.64	8.53
	10-10-1	0.59	9.39
M173	10-1-1	0.51	65.40
	10-2-1	0.48	66.13
	10-3-1	0.47	87.51
	10-4-1	0.43	68.66
	10-5-1	0.37	114.24
	10-6-1	**0.66**	**58.73**
	10-7-1	0.55	79.42
	10-8-1	0.52	80.62
	10-9-1	0.47	97.30
	10-10-1	0.65	69.74

The train and test data have ten input variables (rainfall, runoff, the number of rainy day, temperature and NDVI with lagged 1-month and 2-month). We use the initial training set input into ANN and SVR to create a combined model and then test this model with the test data. The results of applying ANN and SVR alone are shown in Tables I and II, respectively. The result from our proposed combined model is presented in Table III.

TABLE II: Runoff Prediction Performance from SVR Model

Station	Kernel	R	RMSE
M145	linear	0.58	10.17
	sigmoid	0.14	29.20
	polynomial	0.45	10.37
	radial basis function	**0.67**	**8.98**
M173	**linear**	**0.54**	**61.66**
	sigmoid	0.04	135.05
	polynomial	0.36	69.62
	radial basis function	**0.55**	**62.84**

TABLE III: Runoff Prediction Performance from Proposed Model

Station	ANN topology	Kernel	R	RMSE
M145	10-6-1	linear	0.68	8.09
	10-6-1	sigmoid	0.63	8.61
	10-6-1	polynomial	0.68	8.11
	10-6-1	**RBF**	**0.69**	**8.01**
M173	10-8-1	linear	0.71	52.09
	10-8-1	sigmoid	0.71	52.42
	10-8-1	polynomial	0.71	52.38
	10-8-1	**RBF**	**0.71**	**51.99**

Our design of ANN architecture is based on the suggestion "the optimal size of the hidden layer is usually between the size of the input and size of the output layers" [9]. We thus set the hidden layer to be in the range 1-10 as shown in Table I. The results reveal that at the M145 station the best network topology is 10-1-1 (*R*=0.67, RMSE=8.17). At the M173 station, the best topology is 10-6-1 (*R*=0.66, RMSE=58.73).

On building the SVR model, we set parameters as follows: cost=1, gamma=1/(data dimension)), degree=3 and coefficients of the support vector =0 for each kernel. The best runoff prediction at the M145 station is the radial basis kernel (*R*=0.67, RMSE=8.98). At the station M173, the linear kernel (*R*=0.54, RMSE=61.66) performs almost as good as the radial basis function (*R*=0.55, RMSE=62.84).

When we combine the power of both ANN and SVR algorithms, it shows clearly good performance. In our proposed method, we use either ANN or SVR depending on the amount of rain in each data instance. We therefore report both the ANN architecture and the kernel function in Table III. The best results are the one highlighted in red bold font. We also show graphical comparisons of ANN and SVR methods for the station M145 (Fig. 4) and the station M173 (Fig. 5). The actual versus predicted runoff values using our proposed combination ANN-SVR method in both stations is presented in Fig. 6.

Figs. 4 and 5 clearly reveal strength of the ANN and SVR models on predicting runoff values. It can be noticed that the ANN performs well on some peak values, but perform poorly on some low runoff values at the M145 and M173. On the contrary, the SVR is good at predicting runoff amount during the water shortage situation. But when runoff is excessive, the SVR model performs quite poor at both the M145 and M173 stations.

It is actually based on these experimental observation that we thus design the proposed model combining the strength from each model using raining amount as a decision criterion.

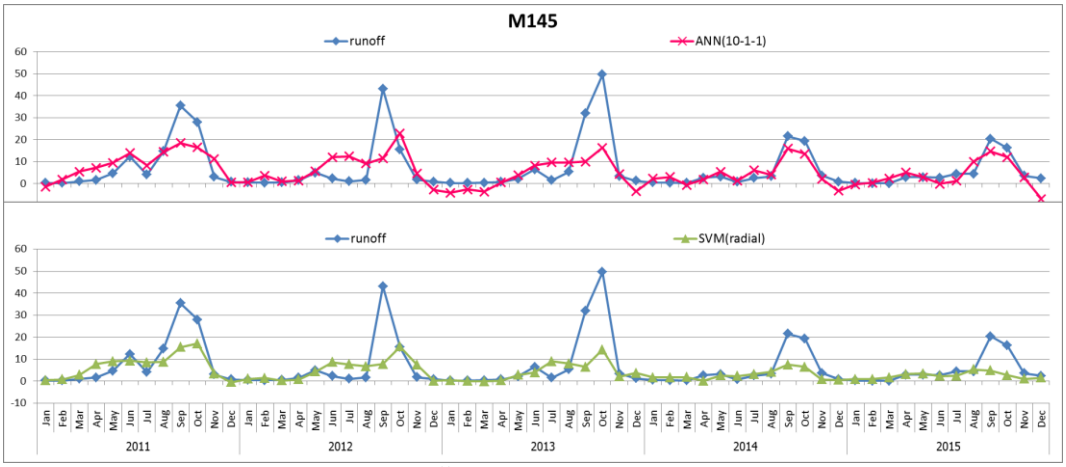

Fig. 4. The predicted and actual runoff values made by ANN and SVR models at the M145 station.

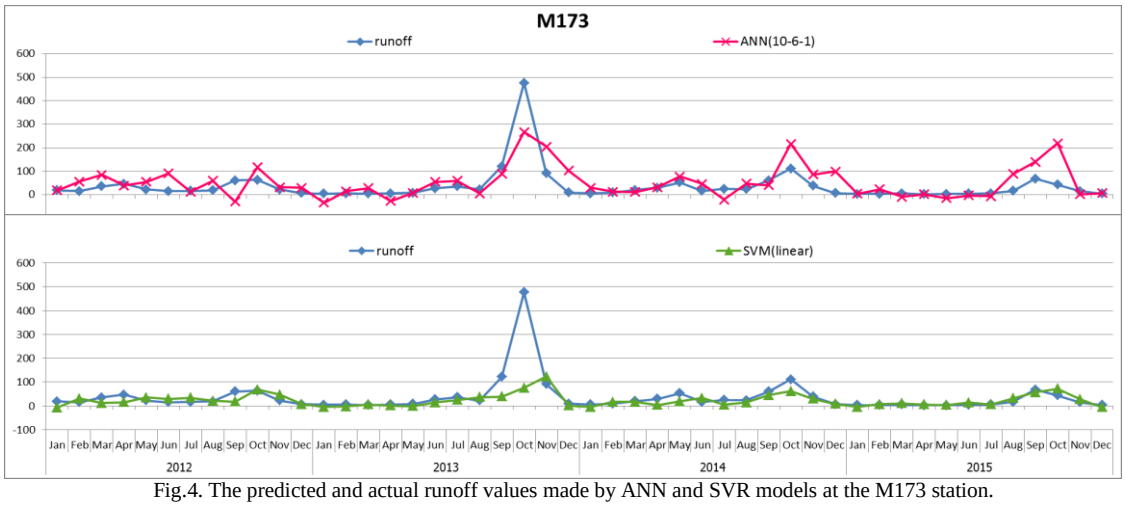

Fig.4. The predicted and actual runoff values made by ANN and SVR models at the M173 station.

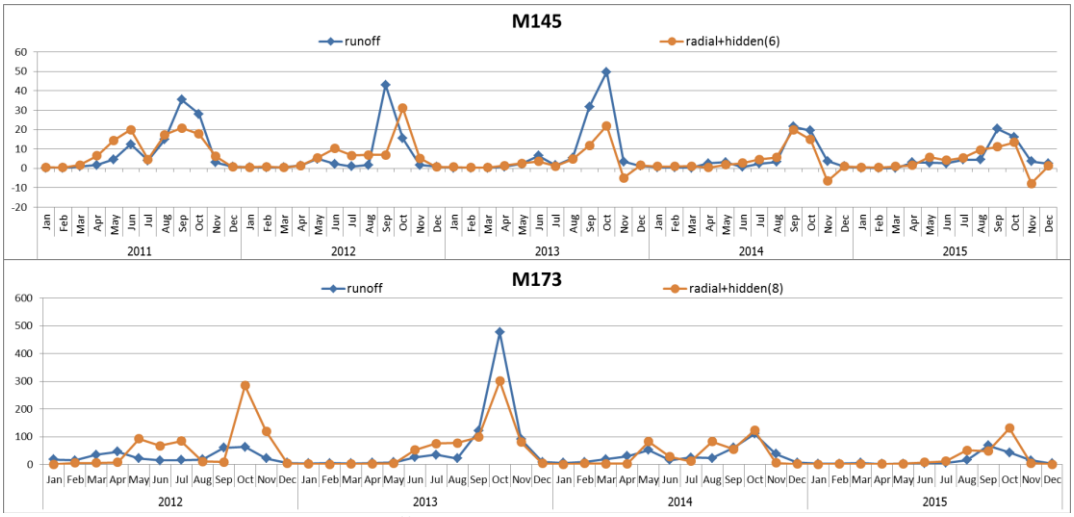

Fig.6. The predicted and actual runoff values at theM145 and M173 stations made by our proposed method.

Our proposed method use average monthly rainfall to select a subset of data for each algorithm with the intuitive idea that rainfall causes runoff. It turns out that our model performs the best in both shortage and excessive rainfall. The proposed method can also find a good architecture for hidden node layer of the ANN.

In addition, we also create model base on Linear Regression (LR). The prediction results are however poor (at the M145 station, R=0.67, RMSE=8.17; at the M173 station, R=0.51, RMSE=65.46). Therefore, we can conclude from these experimental results that our proposed method is the best combined model to predict runoff (at the M145 station, R=0.69, RMSE=8.01; at the M173 station, R=0.71, RMSE=51.99).

V. Conclusion

In this work, we propose a runoff prediction method that combine the advantages of ANN and SVR to predict runoff (rainfall$_t$). The strength of ANN is its high accuracy on predicting runoff when the amount of rainfall is high. The strength of SVR is on the contrary that it is good at the low

amount of rainfall. We thus combine these observed strengths.

On the combination step, we use average accumulative rainfall with lagged 1-month (average rainfall$_{t-1}$) to select a subset of data from the initial training set, then split the initial training set for ANN and SVR to create models. To use our combined method to predict runoff, the decision criteria for choosing either ANN or SVR model is the amount of rainfall on the previous month. If this amount is higher than our threshold value, choose the ANN model; otherwise, choose the SVR model.

From the experimental results, the proposed method shows good efficiency to predict runoff when it is compared against other technique including ANN, SVM and LR. These comparisons are based on R and RMSE metrics using test data from the two stations in the Mun Basin of Thailand.

REFERENCES

[1] S. Subak, "Climate change adaptation in the U.K. water industry: Managers' perceptions of past variability and future scenarios," *Water Resources Management*, vol. 14, pp.137–156, January 2000.

[2] T. A. Sezin and P. A. Johnson., "Precipitation-Runoff Modeling using Artificial Neural Networks and conceptual models," *Journal of Hydrologic Engineering*, vol. 5, no. 2, pp. 156-161, April 2000.

[3] A. W. Minns and M. J. Hall, "Artificial neural networks as rainfall-runoff models," *Hydrological Sciences Journal*, vol. 41, no. 3, pp. 399-417, June 1996.

[4] J. Morshed and J. J. Kaluarachchi, "Application of artificial neural network and generic algorithm in flow and transport simulations," *Advances in Water Resouces*, vol. 22, no. 2, pp. 145-158, 1998.

[5] S. M. V Choubey, S. Pandey, & Shukla, J. "An Efficient Approach of Support Vector Machine for Runoff Forecasting," *International Journal of Scientific & Engineering Research*, vol. 5, no. 3, pp. 158-166, March 2014.

[6] C. L. Wu, K. W. Chau, and Y. S. Li., "River stage prediction based on a distributed support vector regression," *Journal of hydrology*, vol. 358, no. 1-2, pp. 96-111, 2008.

[7] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and Computing*, vol. 1, no. 3, pp. 199–222, 2004.

[8] F. Granata, R. Gargano and Giovanni de Marinis, "Support vector regression for rainfall-runoff modeling in urban drainage: A comparison with the EPA's storm water management model," *Water*, vol.8, no. 3, p. 69, 2016.

[9] *Introduction to Neural Networks with Java*, Heaton Research, Heaton Research, Inc., St. Louis, 2008.

R. Chanklan is currently a doctoral student with the School of Computer Engineering, Suranaree University of Technology (SUT), Thailand. She received her bachelor degree in computer engineering from SUT in 2013, master degree in computer engineering from SUT in 2014. Her current research of interest includes data classification, data mining application, and artificial intelligence.

N. Kaoungku is currently a lecturer at School of Computer Engineering, SUT, Thailand. He received his doctoral degree, master dgree, and bachelor degree in computer engineering from SUT, in 2015, 2013, and 2012, respectively. His current research includes data mining, knowledge engineering, and semantic web.

K. Suksut is currently a doctoral student with the School of Computer Engineering, SUT, Thailand. He received his bachelor degree in computer engineering from SUT in 2011, master degree in computer engineering from SUT in 2013. His current research of interest includes data mining, genetic algorithm, and imbalanced data classification.

K. Kerdprasop is an associate professor and chair of the School of Computer Engineering, SUT. He received his bachelor degree in mathematics from Srinakarinwirot University, Thailand, in 1986, master degree in computer science from the Prince of Songkla University, Thailand, in 1991, and doctoral degree in computer science from Nova Southeastern University, U.S.A., in 1999. His current research includes data mining, artificial intelligence, and computational statistics.

N. Kerdprasop is an associate professor at the School of Computer Engineering, SUT. She received her bachelor degree in radiation techniques from Mahidol University, Thailand, in 1985, master degree in computer science from the Prince of Songkla University, Thailand, in 1991, and doctoral degree in Computer Science from Nova Southeastern University, U.S.A, in 1999. Her research of interest includes knowledge discovery in databases, artificial intelligence, and intelligent databases.