# A Hybrid Modeling Technique to Predict Runoff

Ratiporn Chanklan, Kedkarn Chaiyakhan, Kittisak Kerdprasop, and Nittaya Kerdprasop

***Abstract*—The management of water resources is important to prevent water problems: floods and water shortages. The foreknowledge allows time for officials to sufficient preparation to deal with the problem. This study aims to determine the appropriate weight for predicting runoff from the merge of runoff prediction results from two algorithms: Artificial Neural Network and Support Vector Regression with linear regression modeling. In this paper, we compare the runoff predictive performance of the three algorithms: Linear Regression, Artificial Neural Network, and Support Vector Regression. We use remote sensing data, which are the Normalized Difference Vegetation Index (NDVI) obtained from the NOAA STAR. The ground station rainfall, runoff, the number of rainy days and temperature data in Mun basin, Thailand, are obtained from the Meteorological Department. We evaluate the model performance using two statistical values: Correlation Coefficient and Root Mean Squared Error. Experimental results confirm the best performance of our proposed method.**

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

## I. Introduction

Currently, people are experiencing a natural disaster such as floods and water shortages. The strength of such disaster increases every year and causes much damage. A proper management of water resources is one way to protect flood and water shortage problems. Predicting runoff can help to make decision, planning and management of water resources. Runoff is amount of water in the river caused by the rain that fell in the catchment area then flows into the river. The foreknowledge regarding amount of runoff as either excessive or shortage can be useful for estimating the demand for use and planning to fix or deal with floods and water shortages.

Predicting runoff is a very complex process and it also needs an appropriate modeling technique for accurate prediction. Artificial Neural Network (ANN) is a tool that has been used to create model to predict runoff. It has the ability to simulate both linear and non-linear relationships, without any prior assumptions as most traditional statistical methods. ANN has been successfully used for predicting runoff and the method was widely adopted in hydrology [1], [2]. In addition, some researches have suggested the support vector regression (SVR) as an alternative algorithm for predicting runoff effectively. SVR showed the best performance as reported in [3], [4]. Rainfall lag time values are also used to consider for building a model to predict runoff [5], [6]. Runoff lag time values are also proposed to predict runoff [7]. A comparison on efficiency in the literature normally uses statistical values such as Correlation Coefficient, Coefficient of Determination, Root Mean Squared Error (RMSE), Mean Absolute Percentage Error. In this work, we employ the two measures: Correlation Coefficient and RMSE.

In the model building process, we use a hybrid modeling technique from Artificial Neural Network and Support Vector Regression. At the first step, we find cluster among rainfall and runoff data using k-means clustering. Then, we compute predictor importance to select data (runoff, rainfall, the number of rainy days, temperature and cluster data) that are appropriate for creating a predictive model. Then, we use runoff prediction results from Artificial Neural Network and Support Vector Regression as inputs for Linear Regression to make a final runoff prediction.

Manuscript received February 15, 2017; revised April 19, 2017. This work was supported by grant from Suranaree University of Technology through the funding of Knowledge and Data Engineering Research Units.

R Chanklan is with the School of Computer Engineering, Suranaree University of Technology (SUT), 111 University Avenue, Muang, Nakhon Ratchasima 30000, Thailand (e-mail: arc_angle@hotmail.com).

K. Chaiyakhan is with the Computer Engineering Department, Rajamangala University of Technology Isan, Nakhon Ratchasima, Thailand (e-mail: kedkarnc@hotmail.com).

K. Kerdprasop is with the School of Computer Engineering and the Head of Knowledge Engineering Research Unit, SUT, Thailand (e-mail: kerdpras@sut.ac.th).

N. Kerdprasop is with the School of Computer Engineering and the Head of Data Engineering Research Unit, SUT, Thailand (e-mail: nittaya@sut.ac.th).

## II. Background Theories

### A. Artificial Neural Network

Artificial Neural Network (ANN) is a mathematical method with the basic idea to make a computer machine having the ability to think like humans. ANN has connected multiple computational nodes to form a network. The network has three main levels: input layer, hidden layer, and output layer. The input nodes are node in the input layer. In the input layer, the number of nodes is equal to the number of attributes (or features, independent variables). The hidden nodes means nodes in the hidden layer. The number of nodes in the hidden is defined by a user. There can be more than one layer in the hidden layer of the network. The output nodes are nodes in the output layer. In the output layer, the number of nodes is equal to the number of data groups (or target, dependent variable). The hidden nodes, output nodes, and the structure of the internal calculation are shown in Fig. 1.

In the network, each line has specified weight: $W = [w_1, w_2, ..., w_R]$. Input data are vector of elements: $p = [p_1, p_2, ..., p_R]$, in which R is number of attributes in input data. The network works by multiplying the input data to weight on each edge, summing results from each incoming edge of the node, and summing a bias (b) of that node. The result (n) has to be transformed through a transfer function to obtain the final computation result of each node, denoted as a.

The main objective of ANN learning is the search for proper weight on each edge in the network such that the

DOI: 10.7763/IJMO.2017.V7.559

network is most accurate on separating training data to a correct class. The weight of the edge is uncertain but the network can adjust weight values by teaching it to recognize a pattern of data. It adjusts weight if the output from the neural network is incorrect. The weight adjustment is iterated in a back propagation manner until the network has a small error or it reaches an acceptable threshold.

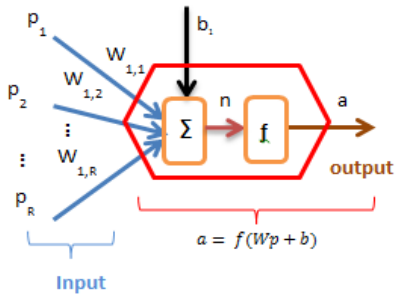



Fig. 1. Neural unit with incoming input edges and a bias.

### B. Support Vector Regression

Support vector regression (SVR) has been developed from support vector machines (SVM). The algorithm can estimate the target by a linear equation [8], as shown in equation 1.

$$f(\vec{x}) = \vec{w} \cdot \vec{x} + b \tag{1}$$

When b is a threshold value and w is a weight vector. It finds hyperplane which has small margin and tries to keep all the data within the margin and allows some data to be outside the margin. The margin can be calculated as in equation 2.

$$\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^*) \tag{2}$$

When $\xi$ is slack variable such that $\xi_i, \xi_i^* \geq 0$ and N is number of train data. The margin within the scope can be computed with equations 3 and 4.

$$y_i - (\vec{w} \cdot \vec{x} + b) \leq \varepsilon + \xi_i \tag{3}$$

$$y_i - (\vec{w} \cdot \vec{x} + b) \leq -\varepsilon - \xi_i^* \tag{4}$$

Then linear Support Vector Regression is as shown in equation 5.

$$y = \sum_{i=1}^{N}(\alpha_i - \alpha_i^*) \cdot \langle x_i, x \rangle + b \tag{5}$$

where $\langle x_i, x \rangle$ is the inner product of two vectors in the feature space.

### C. K-means Clustering

K-means clustering is a method to partition n objects into k clusters in which each object belongs to the cluster with the nearest central point, or centroid [9]. The number of *k* cluster can be defined by user. K-means steps are as follows:

1) Set number of clusters, denoted as *k*
2) Random cluster centers (centroid) in each *k* group
   Measure the distance between the data D = ($x_1$,$y_1$) and Centroid = ($cx_1$,$cy_1$) according to the Euclidean distance function, computed as in equation 6.

$$\text{Euclidean distance} = \sqrt{(x_1 - cx_1)^2 + (y_1 - cy_1)^2} \tag{6}$$

3) Assign data to their group based on the closest distance between the data and the cluster center.
4) Calculate the mean of all data in each cluster and set it to be the new centroid. Suppose there are two data points in group A = (($x_1$,$y_1$), ($x_2$,$y_2$)), the new centroid can be calculate using the mean of all data in the group, as shown in equation 7.

$$\text{centroid} = \left(\frac{x_1+x_2}{2}, \frac{y_1+y_2}{2}\right) \tag{7}$$

5) Repeat steps 3, 4 and 5 until all the centroids do not change.

### D. Predictor Importance

Predictor importance can be determined by computing the reduction in variance of the target attributable to each predictor using a sensitivity analysis. Predictors are ranked according to the sensitivity measure [10] defined as in equation 8.

$$S_i = \frac{V_i}{V(Y)} = \frac{V(E(Y|X_i))}{V(Y)} \tag{8}$$

where $Y$ is target, $X_i$ is predictor ranging from 1,...,$k$. The number of predictors is $k$. Model for $Y$ based on predictors $X_1$through $X_k$ and $V(Y)$ is the unconditional output variance. Predictor importance is then computed as the normalized sensitivity using equation 9.

$$VI_i = \frac{S_i}{\sum_{j=1}^{k} S_j} \tag{9}$$

where $S_i$ is the proper measure of sensitivity to rank the predictors in order of importance for any combination of interaction and non-orthogonality among predictors. $VI_i$ is the estimate of the conditional variances computing by the multidimensional integrals in the space of the input factors using Monte Carlo method.

### E. Correlation Coefficient

The Correlation Coefficient is denoted by R. It is a statistical value to find relationships between two variables (x and y). The coefficient is a numerical value between -1 and 1. There are three possible types of relationship: zero correlations means no relationship, positive correlations is same direction relationship, and negative correlation is the kind of inverse relationship. The correlation coefficient can be calculated using equation 10.

$$R = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{\left[n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right]\left[n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2\right]}} \tag{10}$$

where $x_i = (x_1, x_2, \ldots, x_n)$ are the values of variable $x$, $y_i$ is the value of variable $y$, $n$ is the total number of samples.

The correlation coefficient that is closer to 1 or -1 represents high level of relationship. The coefficient that is close to 0 represents low level or even no relationship.

### F. Root Mean Squared Error

Root Mean Squared Error (RMSE) is used to check errors in the prediction of the model. If the RMSE has a small value that means the prediction model is efficient. The RMSE metric used for model evaluation can be calculated as in equation 11.

$$\text{RMSE} = \sqrt{\left(\frac{\sum(T_i - O_i)^2}{N}\right)} \quad (11)$$

where $N$ is the number of all data, $O_i$ is the value of prediction, $T_i$ is the actual data.

## III. Materials and Methods

The study area in this work is Mun basin, a largest basin in the North-eastern region of Thailand (Fig. 2). We use Normalized Difference Vegetation Index (NDVI), which is a remote sensing data obtained from the NOAA STAR (http://www.star.nesdis.noaa.gov), monthly rainfall, runoff and the number of rainy days data from the Meteorological Department (http://www.hydro-4.com), and temperature data from National Statistical Office (http://www.nso.go.th). This research use IBM SPSS Modeler 14.1 as analysis tool in our experiments. We use two data sources from the Mun basin: M145 and M173 stations.

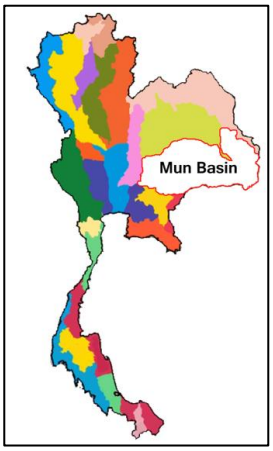



Fig. 2. The study area: Mun Basin, Thailand.

The M145 (Lamphra Phloeng) station locates at Ban Wang Takhian, Tambun Wang Katha, Amphoe Pak Chong, in Nakhon Ratchasima province, Thailand. The 13-year data during 1998 to 2010 have been used as training data. The 5-year data from 2011 to 2015 are to be used as test data.

The M.173 (Moon River) station locates at Ban Non Sa-at, Tambun Tha Yiam, Amphoe Chokchai, in Nakhon Ratchasima province, Thailand. The 10-year training data are those during the periods 2002 to 2011. The 5-year data between 2012 to 2015 are used as test data.

Prepare step: We apply k-means to find clusters from rainfall and runoff data. The appropriate k clusters have been judged from the Silhouette Coefficient. Then, we use predictor importance to select a subset of potential features from the initial set of features including rainfall, runoff, the number of rainy day, temperature, NDVI, and cluster identifier. Data are also lagged 1-month and 2-month. These data are input into two learning algorithms: Artificial Neural Network and Support Vector Regression. Because we use the data from different sources, we thus have to find different sets of parameters that are appropriate to predict runoff in different locations. We later use runoff prediction results from Artificial Neural Network and Support Vector Regression as inputs in Linear Regression, which is a final model to predict runoff in this work.

In the subsequent step, we use test data to test model performance. We evaluate performance using two statistical values: Correlation Coefficient and Root Mean Squared Error. The modeling process of this paper is shown in Fig. 3.

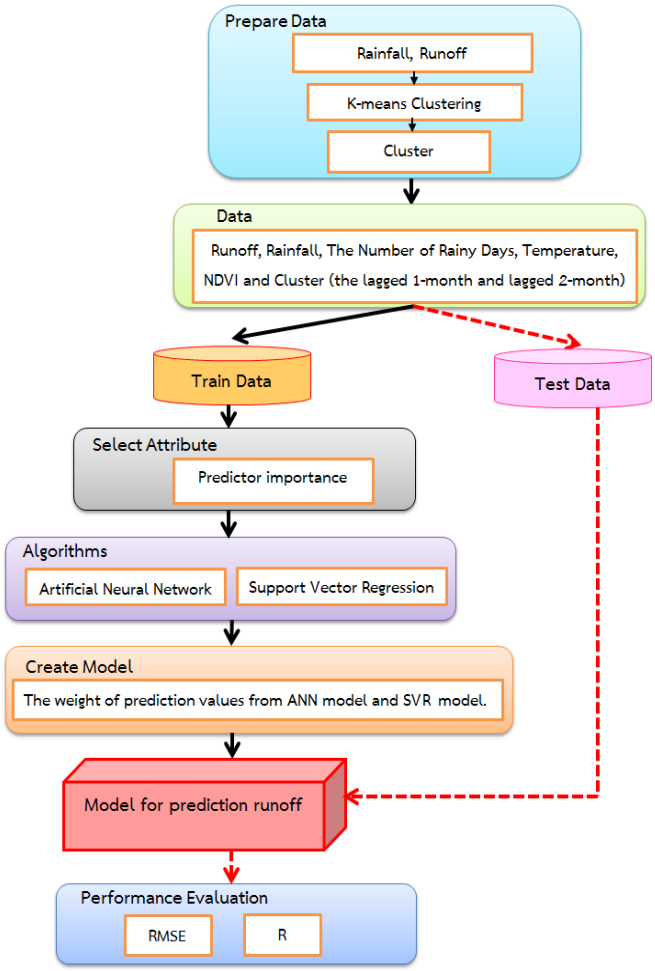



Fig. 3. The modeling process for runoff prediction.

## IV. Experimental Results

We predict runoff with test data using the model proposed in Figure 3 and compare the results against the other three models including Artificial Neural Network, Support Vector Regression, and Linear Regression. We experiment with two data sets from the Mun basin: M145 and M.173 stations.

We firstly explore correlation between runoff and other attributes (rainfall, the number of rainy days, NDVI, and temperature) and the computed coefficients are shown in Table 1. The results show that runoff and rainfall has quite strong positive relationship in both stations. Number of rainy days at the M145 station shows positive relationship to the runoff, but it shows no relationship to runoff at M175 station. The NDVI shows low relationship to runoff at both stations (0.326 and 0.41). Temperature show negative relationship to runoff at both locations (-0.014 and -0.08).

In our proposed method to runoff prediction modeling, we introduce new attribute, that is the cluster identifier, to improve the performance of prediction. From the experimental results as shown in Table 1, we thus use only the runoff and rainfall attributes for clustering with k-means and select the appropriate *k* clusters based on the Silhouette Coefficient values. After that, we use predictor importance as a metric to select top important features and use Artificial Neural Network and Support Vector Regression to predict runoff. From our proposed method, we combine the results from these two models and generate a final prediction using linear regression. The runoff prediction results are presented in Table II.

The models to predict runoff at the M145 station have been built based on the k-means clustering process with k = 5 and Silhouette Coefficient = 0.64. The R values in all models is about 0.6. Our proposed method yields the best R value at 0.67. RMSE values of all models are in the range from 8 to10. The best model based on RMSE metric is our proposed model (RMSE = 8.34). Note that for R measure, the higher is the better. But for the RMSE metric, the lower is the better.

TABLE I: THE RESULTS OF CORRELATION COEFFICIENT ANALYSIS

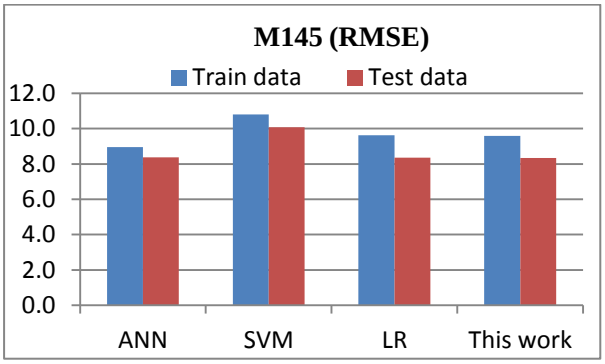
| Attribute | M145 | M173 |
|---|---|---|
| Rainfall | 0.670 | 0.54 |
| Day of Rainy | 0.502 | 0.08 |
| NDVI | 0.326 | 0.41 |
| Temperature | -0.014 | -0.08 |

TABLE II: RUNOFF PREDICTION PERFORMANCE AT M145 AND M173 STATIONS

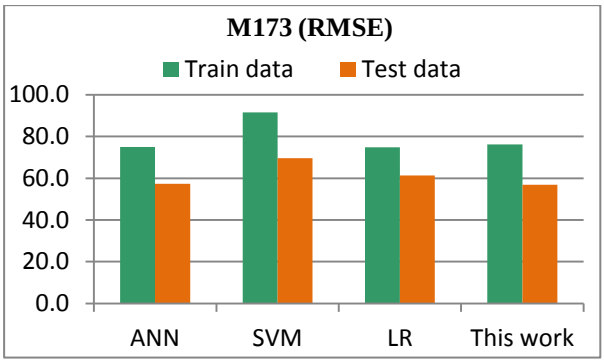
| Station | Model | R | RMSE |
|---|---|---|---|
| M145 | Artificial Neural Network | 0.66 | 8.38 |
| | Support Vector Regression | 0.64 | 10.09 |
| | Linear Regression | 0.66 | 8.35 |
| | This work | 0.67 | 8.34 |
| M173 | Artificial Neural Network | 0.58 | 57.41 |
| | Support Vector Regression | 0.42 | 69.61 |
| | Linear Regression | 0.58 | 61.33 |
| | This work | 0.59 | 56.94 |

The models to predict runoff at the M173 station have been also built based on the k-means clustering process with k = 5 and Silhouette Coefficient = 0.692. The R values in all models is ranging from 0.4 to almost 0.6. Our proposed method yields the highest R value at 0.59. RMSE values of all models are in the range from 56 to70. Our proposed method performs the best with the lowest RMSE value at 56.94.

The performances of all models based on RMSE and R values are also graphically compared and shown in Figures 4 and 5. In the comparison figures, we show RMSE and R values evaluated on both training data and test data. This is for assessing the over-fitting problem of the models. A model is called over-fitting if it performs well on the training data, but poorly perform on the separate set of test data. Currently, there is no agreement regarding how difference between train-test performance should be considered over-fitting. We, therefore set temporary train-test performance not exceeding 35% as non over-fitting. The models based on ANN, SVM, LR, and our proposed one are non over-fitting.

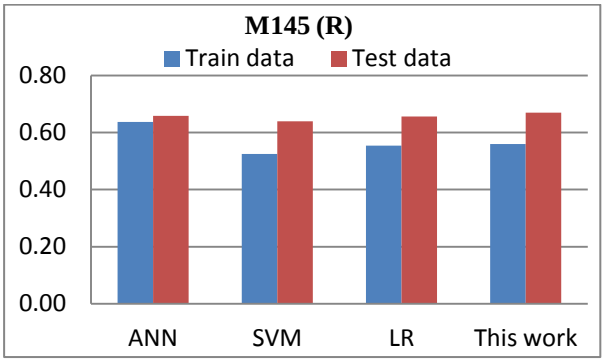



(a) RMSE of the M145 station.

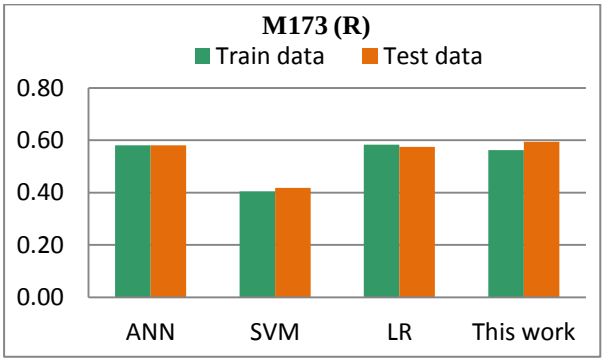



(b) RMSE of the M173 station.

Fig. 4. The RMSE values assessed in train data and test data.




(a) R of the M145 station.




(b) R of the M173 station

Fig. 5. The R values assessed in train data and test data.

It can be noticed from Fig. 5 that ANN model tends to fairly fit model to both train and test data. But our proposed method yields higher R values on the test data than the train data. This may be the reason for our approach being better than others on predicting runoff in the test dataset.

## V. Conclusion

In this work, we propose a hybrid modeling technique to predict runoff from combining results from Artificial Neural Network and Support Vector Regression models. On the combination step we adjust weights from the two models by means of Linear Regression. It is the final output from linear regression that to be used as our runoff prediction. From the experimental results the proposed hybrid model shows good efficiency to predict runoff when it is compared with other single learner technique including Artificial Neural Network, Support Vector Regression, and Linear Regression. The comparison is based on the correlation coefficient and RMSE metrics using data from the two stations in the Mun basin of Thailand. Based on the two metrics, our proposed model show the best performance over other three techniques. We also perform experimentation on both training data and test data to check over-fitting problem. The results confirm that all models are non over-fit to the training data.

## References


[1] M. A. Kaltech, "Rainfall-runoff modeling using artificial neural network modeling and understanding," *Caspian Journal of Environmental Sciences*, vol. 6, pp. 153-158, 2008.

[2] P. S. Kumar, T. V. Praveen, and M. A. Prasad, "Artificial neural network model for rainfall-runoff-a case study," *International Journal of Hybrid Information Technology*, vol. 9, no. 3, pp. 263-272, 2016.

[3] H. Chu, J. Wei, T. Li, and K. Jia, "Application of support vector regression for mid- and long-term runoff forecasting in 'yellow river headwater' Region," *Procedia Engineering*, vol. 154, pp. 1251–1257, 2016.

[4] F. Granata, R. Gargano, and G. D. 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. 69, 2016.

[5] N. Sajikumar and B. S. Thandaveswara, "A non-linear rainfall–runoff model using an artificial neural network," *Journal of Hydrology*, vol. 216, pp.32–55, 1999.

[6] A. Agarwal and R. D. Singh, "Runoff modeling through back propagation artificial neural network with variable rainfall–runoff data," *Water Resources Management*, vol. 18, pp.285–300, 2004.

[7] A. Dorum, A. Yarar, M. F. Sevimli, and M. Onüçyildiz, "Modelling the rainfall–runoff data of susurluk basin," *Expert Systems with Applications*, vol. 37, no. 9, pp. 6587-6593, 2010.

[8] A. Farag and R. M Mohamed, *Regression Using Support Vector Machines: Basic Foundations*, Technical Report, University of Louisville, Louisville, 2004.

[9] J. A. Hartigan, A. Manchek, and A. Wong, "Algorithm AS 136: A k-means clustering algorithm," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 28, no. 1, pp. 100-108, 1979.

[10] A. Saltelli, "Making best use of model evaluations to compute sensitivity indices," *Computer Physics Communications*, vol. 145, no. 2, pp. 280–297, 2002.


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

**K. Chaiyakhan** is a lecturer with the Computer Engineering Department, Rajamangala University of Technology Isan, Nakhon Ratchasima, Thailand. She received her bachelor degree in computer engineering from Rajamangala University of Technology Thanyaburi in 1998, the master degree in computer engineering from King Mongkut's University of Technology Thonbuti in 2007 and doctoral degree in computer engineering from Suranaree University of Technology, Thailand in 2016. Her current research includes image classification and image clustering.

**K. Kerdprasop** is an associate professor and chair of the School of Computer Engineering, Suranaree University of Technology, Thailand. He received his bachelor degree in mathematics from Srinakarinwirot University, Thailand, in 1986, the 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, functional and logic programming languages, computational statistics.

**N. Kerdprasop** is an associate professor at the School of Computer Engineering, Suranaree University of Technology, Thailand. She received her bachelor degree in radiation techniques from Mahidol University, Thailand, in 1985, the 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. She is a member of ACM and IEEE Computer Society. Her research of interest includes knowledge discovery in databases, artificial intelligence, and intelligent databases.