International Journal of u- and e- Service, Science and Technology
Vol.8, No.3 (2015), pp.145-154
http://dx.doi.org/10.14257/ijunesst.2015.8.3.13

The Study on the Accuracy of Classifiers for Water Quality Application

Rosaida Rosly[1], Mokhairi Makhtar[2], Mohd Khalid Awang[3], M Nordin A Rahman[4] and Mustafa Mat Deris[5]

[1,2,3,4]Faculty of Informatics and Computing
University of Sultan Zainal Abidin, Terengganu, Malaysia
[5]Faculty of Information Technology and Multimedia
University of Tun Hussein Onn, Johor, Malaysia
[1]rosaidarosly@gmail.com, {[2]mokhairi, [3]khalid, [4]mohdnabd}@unisza.edu.my,
[5]mmustafa@uthm.edu.my

Abstract

Dirty water is the world's biggest health risk. When water from rain roads into rivers, it picks up toxic chemicals, dirt, trash and disease-carrying organisms along the way. Many of our water resources lack basic protections, making them vulnerable to pollution from factory farms and industrial plants. Due to that, a classification model is needed to present the quality of the water environment. In this paper, the data mining techniques are used in this research by applying the classification method for water quality application. Various classifiers were studied in order to find the most accurate classifier for the dataset. This paper presents the comparison of accuracies for the five classifiers (NB, MLP, J48, SMO, and IBk) based on a 10-fold cross validation as a test method with respect to water quality from the datasets of Kinta River, Perak Malaysia. This study also explores which classifier is suitable to classify the dataset. The selected attributes used in this study were: DO Sat, DO Mgl, BOD Mgl, COD Mgl, TS Mgl, DO Index, AN Index, SS Index, Class, and Degree of pollution. The data consisted of 166 instances and obtained from the East Coast Environmental Research Institute (ESERI) of Universiti Sultan Zainal Abidin (UniSZA). The result of MLP and IBk performed better than other classifiers for Kinta River dataset because these classifiers showed the highest accuracy with the same percentage of 91.57%. In the future, we will propose the multiclassifier approach by introducing a fusion at a classification level between these classifiers to get a higher accuracy of classification.

Keywords*: classifiers, water quality, classification method, feature selection*

1. Introduction

Data mining is increasingly practiced in science to draw out information from the enormous data sets generated by modern experimental and observational methods. It is used to find new, hidden, or unexpected patterns in data. Now, more than one organizations are using data mining techniques [1]. Many research work in data mining have gone into improving the predictive accuracy by applying the data mining techniques. For instance, data mining includes analysis and prediction of data as it contains of more than just data collection and data management. In order to manage the data, there are several data mining techniques that can be used. One of them is the classification method, which is widely known as compared to regression with its ability to process wider variety of information [2]. Classification is a machine learning method that is used in predict data instances[2]. It is also one of the data mining tasks, considering that many of the classification problems do not occur in only one application areas[3]. Data mining

ISSN: 2005-4246 IJUNESST

methods include techniques which evolve from artificial intelligence, machine learning, statistics, and so on. The data mining categorisation is shown in Table 1.

Table 1. Data Mining Categorisation

Author	Data Mining Categorisation
Fayyad, et. al. (1996)	Classification, Regression, Clustering, Summarization, Dependency modeling, Link analysis, Sequence analysis
Han et. al. (1996)	Association, Generalization, Classification, Clustering, Similarity search, Path traversal pattern
Berry (1997)	Classification, Estimation, Prediction, Affinity grouping, Clustering, Description

Classification maps a data item into one of several predefined categorical classes. For instance, decision tree, neural network, and some probability approaches are often used in performing this function. There are two steps to implement classification function. Firstly, a classification model is built to describe a predetermined set of classes or concepts, and secondly, the model is used for classification.

In this paper, water quality was chosen as an application field to demonstrate the proposed approach in achieving high accuracy. It is important to note that water is essential and there is a strong link between water and health. Besides that, safe water is key to life and having poor quality water not only damages the environment and kills wildlife, but it can also sicken and kill people. The effect of human activities on water quality can be seen in terms of the ecosystem and the water usage[1]. Furthermore, the quality of water determines the level of both the health of human beings and the ecosystem. About a billion people do not have a reliable constant supply of safe water, where two to four million deaths a year are attributed to unsafe water [4]. Due of the situation, classifying the accuracy of water quality is very important. The people, lab instruments, and sensors have been part of studies in order to monitor water quality. However, the cost and time are expensive [5].

This paper presents the comparison of accuracies for five classifiers. There are NB, MLP, J48, SMO, and IBk. Three different aspects such as complexity, overfitting, and performance are considered in the selection process of the classifier. Other than that, in order to classify water quality data, several machine learning methods like decision tree and artificial neural networks (ANNs) have been used [5]. In order to handle large data set, the new classification algorithms need to be designed[6].

The rest of this paper is organised as follows: Section 2 presents the related work on classification model, feature selection approach, water quality classification and performance evaluation criteria. The research framework is explained in Section 3, followed by the details of used datasets in Section 4. This paper continues with Section 5 reporting the experimental results and lastly, Section 6 summarises this paper.

2. Related Work

The predictive model such as the decision tree can help people to acquire a target value through the classification and analysis[7]. Decision tree is a model that is both predictive and descriptive. For instance, a decision tree displays relationships found in the training data, in which the prediction of water quality becomes a very complicated issue due to the complexity and diversity. The prediction of water quality becomes effective when it comes up with an improved decision tree learning method[8]. The decision tree also has an advantage in water quality classification. For example, the potential water quality among times can be easily identified by the set of terminal nodes in the tree that has better water quality data classification, and then the user can focus on the specific data described by those nodes. Thus, the decision tree is capable to be constructed effectively as compared to other methods [8].

Some researches use decision tree learning for water prediction. In [9], they applied decision trees method into construction models to qualitatively predict Phaeocystisglobosa blooms in the Dutch coastal waters. In [10], they used the decision tree model to predict the level of chlorophyll on the next day from Online Monitoring Station. Two applications of the decision tree learning were introduced in simultaneous predictions of multiple physic-chemical properties and past physic-chemical properties of the river water from biological properties[11].

Moreover, Artificial Neural Networks (ANNs) has become the main focus of many scientific disciplines, and one of them is water quality [5]. ANNs has been widely used in solving environmental problems including water resources modeling and management problems [12]. ANNs is an excellent predicting tool for WQI and very useful for helping decision makers as part of the Juru River management measures [12]. In [5], they implemented a multilayer perception (MLP) neural network using the levenberg – macquardt (LM) algorithm to classify water quality of canals in Bangkok. As a result, the classification trees performed better than the multilayer perceptron neural network. Multilayer perceptron neural network achieve a high accuracy multilayer perception rate at 96.52% in classifying the water quality of Dusit District canal in Bangkok. The usage of multilayer perceptron (MLP) is very famous in solving a number of different problems [13].

In [14], they used naïve Bayes methods for the diatoms classification. They used this technique to assess relationships between the diatoms and the indicators of the environment, at the same time classified into one of the water quality classes (WQCs). In [15], the Least Squares Support Vector Machine (LS-SVM) theory were proposed in order to improve the accuracy of water quality retrieval. Furthermore, the experimental results showed that LS-SVM achieved good performance with the lower of complexity. ANN and Support Vector Machine (SVM) display optimal training performances and generalization in many fields of application [16]. To appropriately solve the classification and regression issues, support vector machine can transform the learning process into a convex quadratic planning problem to get a global optimisation by using the rule of minimum structure risk [17].

IBk is an implementation of the k-nearest-neighbor (kNN) classifier. Each case is considered a point in multi-dimensional space. The classification is done based on the nearest neighbor. Other than that, the value of ‘k’ for nearest neighbor can be different. This determines how many cases must be considered as neighbor to decide how to classify an unknown instance [3]. Moreover, the k-Nearest Neighbor was also used in classifying drinking water quality (DWQ) [18]. In this research, using different classifiers to compare which classifiers can achieve a higher accuracy of data by applying k-folds cross validation.

Applying the feature selection in this study may improve the performance of classification algorithm and allow us to better understand the domain. By using the

feature selection, it produces higher accuracy [19]. The filters, wrappers, and embedment are feature subset selection approaches that perform the feature selection process as an integral part of a machine learning algorithm [20]. The feature selection in supervised learning has been well studied, where the main goal is to find a feature subset that produces higher classification accuracy [21]. Guidelines for applying feature selection methods are given based on data types and domain characteristics [22]. In reducing the dimension of the problem and enhancing the performance of classifiers, a greedy feature selection procedure based on multidimensional mutual information is applied [23]. Two methods such as dimensionality reduction and feature subset selection are important components in classification or regression problems as they reduce the attribute space of a feature set [20].

In increasing the level of confidence of classification and the generalization performance, we need to choose the set of classifiers, from a population of high accurate classifiers, with lowest inaccuracy among its members [24]. The collective behavior of a set of classifiers can produce useful information to improve system performance [25].

We used data mining tools, WEKA for this experiment. Nowadays, WEKA widespread its used in both academic and business [26]. It is a collection of machine learning and data mining algorithm. The algorithms can either be applied directly to a dataset or called from our own Java code. WEKA contains tools for data preprocessing, classification, regression, clustering, association rules, and visualisation [27].

The classification of water quality is based on water quality index [28]. The relationship between WQI and water qualification is shown in Table 2.

Table 2. Relationship between WQI and Water Qualification

Degree of Pollution	WQI range
Clean	81-100
Slightly Polluted	60-80
Very Polluted	0-59

The visualisation tool such as confusion matrix is used to present the accuracy of the classifiers in classification [29]. The entries in the confusion matrix have the following meanings in the context of our research:

- *a* is the number of correct predictions for clean class,
- *b*, *c* is the number of incorrect predictions for clean class,
- *e* is the number of correct predictions for slightly polluted class,
- *d*, *f* is the number of incorrect predictions for slightly polluted class,
- *i* is the number of correct predictions for very polluted class,
- *g*, *h* is the number of incorrect predictions for very polluted class,

Table 3. Confusion Matrix

		Predicted		
		Clean	*Slightly Polluted*	*Polluted*
Actual	*Clean*	*a*	*b*	*c*
	Slightly Polluted	*d*	*e*	*f*
	Polluted	*g*	*h*	*i*

The accuracy (*Acc*) is the measurement of the total number of predictions that were correct. It is determined using equation 1.

$$Acc = \frac{a + e + i}{a + b + c + d + e + f + g + h + i} \quad (1)$$

3. Research Framework

Figure 1 represents the single classifier framework for Water Quality Classification. It consists of four processes: input, feature selection, data mining process, and output.

In data input, the data are referred to the components of numeric and categorical datasets. The data will be prepared and filtered to improve the quality of data and time to clear the noise at the same. The next process, a feature selection process, will be implemented on the water quality information criterion dataset in increasing the classification accuracy by eliminating noise features. It means that redundant and irrelevant features are ignored. By using the classification approach, the different classifiers are used to achieve the best result of classifying water quality at Kinta River, Perak. Lastly, the predicted class will be determined.

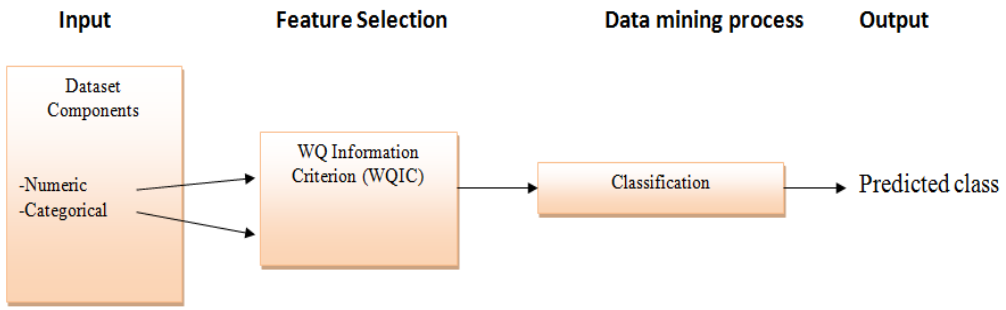

Figure 1. Classifier Framework

4. Dataset Description

The Water Quality of Kinta River, Perak datasets from the ESERI were used to predict the accuracy of data. The description's detail of dataset is shown in Table 4. The classification of water quality is based on water quality index [28].

Table 4. Description of the Water Quality Dataset

Dataset	Water Quality of Sg Kinta, Perak Malaysia
No. of Attributes	10
No. of instances	166
No. of classes	3

Dataset consisted of ten attributes (DO Sat, DO Mgl, BOD Mgl, COD Mgl, TS Mgl, DO Index, AN Index, SS Index, Class, and Degree of pollution). The dataset also consisted instances with a set of numerical and categorical features. Other than that, the datasets contained 166 of instances and for class, they were divided into three categories (clean, slightly polluted, and polluted).

5. Experimental Results

To evaluate the propose model, a number of experiments were performed. Figure 2 shows the comparison of accuracies for the five classifiers used based on a 10-fold cross validation as a test method without the feature selection namely, NB, J48, MLP, SMO, and IBk. From the figure, SMO achieved the higher accuracy of 84.94% followed by NB obtaining 81.33% better than MLP, J48, and IBk.

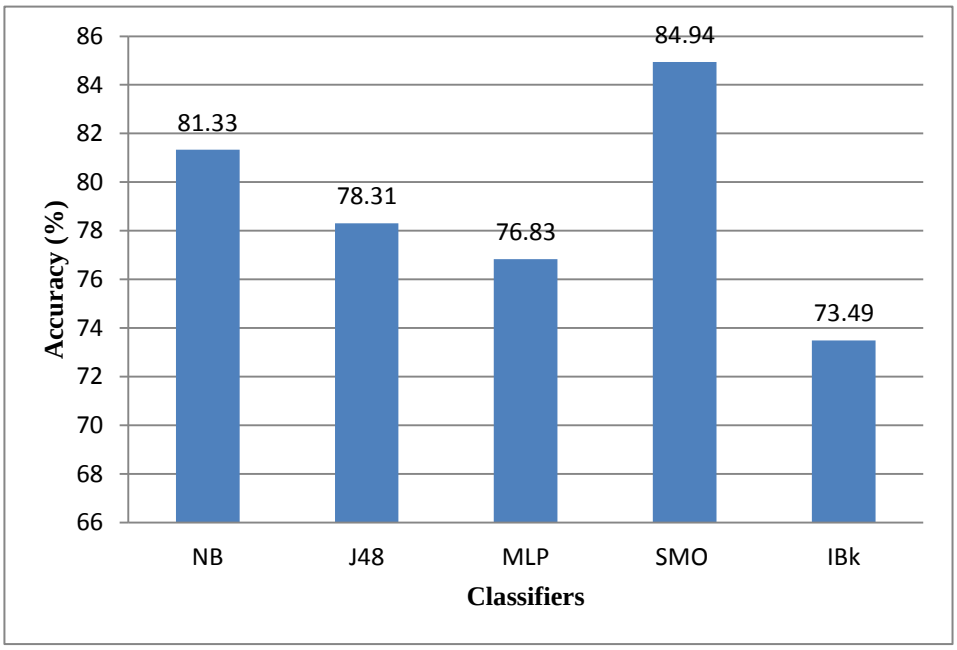

Figure 2. Single Classifier in WQ Dataset without Feature Selection

Figure 3 shows the comparison of accuracies for the five classifiers based on a 10-fold cross validation as a test method with the feature selection: NB, J48, MLP, SMO, and IBk. MLP and IBk achieved the highest accuracy with the same percentage (91.57%) followed by SMO (88.55%), which indicated a result better than that produced by NB and J48.

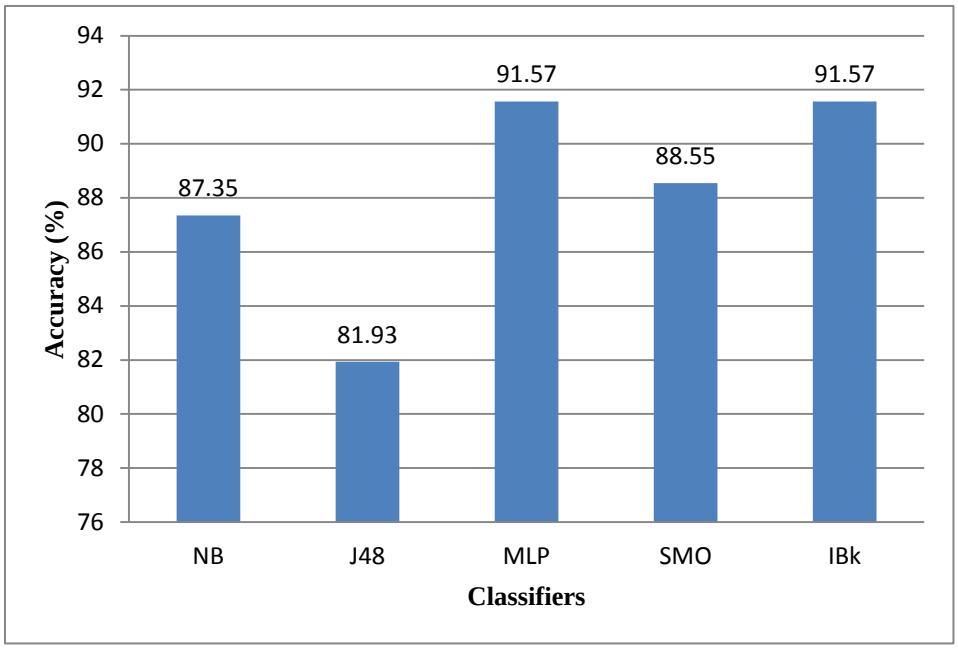

Figure 3. Single Classifier in WQ Dataset with Feature Selection

6. Conclusions

The experimental results in WQ dataset show that MLP and IBk achieved the higher accuracy than other classifiers. These classifiers are indeed suitable for classifying datasets but we also require another approach to find the most accurate classifier in order to improve the higher accuracy of dataset. In future we will propose our multiclassifier approach in improve the accuracy of datasets using a fusion classification level between these classifiers.

Acknowledgement

This work is partially supported by UniSZA and KPM (Grant No. FRGS/2/2013/ICT07/UNISZA/02/2).

References

[1] S. G, "Water Quality Prediction Using Data Mining techniques : A Survey", Int. J. Eng. Comput. Sci., vol. 3, no. 6, **(2014)**, pp. 6299–6306.
[2] T. N. Phyu, "Survey of Classification Techniques in Data Mining", Computer (Long. Beach. Calif)., vol. 1, **(2009)**, pp. 18–20.
[3] C. L. Devasena, "Classification Of Multivariate Data Sets Without Missing Values Using Memory Based Classifiers – An Effectiveness Evaluation," Int. J. Artif. Intell. Appl., vol. 4, no. 1, **(2013)**.
[4] A. Chandrasekhar, "Water quality is as important for ecosystems as for people," 2014. [Online]. Available: http://www.teebweb.org/water-quality-is-as-important-for-ecosystems-as-for-people/, **(2014)**.
[5] S. Areerachakul and S. Sanguansintukul, "Classification and Regression Trees and MLP Neural Network to Classify Water Quality of Canals in Bangkok , Thailand", vol. 1, no. 1, **(2010)**, pp. 43–50.
[6] Q. He, F. Zhuang, J. Li, and Z. Shi, "Parallel Implementation of Classification Algorithms Based on MapReduce," Comput. Linguist. , **(2010)**, pp. 655–662.
[7] J. Han, M. Kamber, and J. Pei, Data Mining Concepts and Techniques, Third Edit. 2012.
[8] H. Liao and W. Sun, "Forecasting and Evaluating Water Quality of Chao Lake based on an Improved Decision Tree Method", vol. 2, **(2010)**, pp. 970–979.
[9] Q. Chen and A. E. Mynett, "Predicting Phaeocystis globosa bloom in Dutch coastal waters by decision trees and nonlinear piecewise regression", Ecol. Modell., vol. 176, no. 3–4, **(2004)** , pp. 277–290.
[10] L. Jinsuo and H. Tinglin, "Data mining on forecast raw water quality from online monitoring station based on decision-making tree", NCM 2009 - 5th International Joint Conference on INC, IMS, and IDC, **(2009)**.

[11] H. Blockeel, S. Džeroski, and J. Grbovi, "Simultaneous prediction of multiple chemical parameters of river water quality with TILDE," Princ. Data Min. Knowl. Discov. , **(1999)**, pp. 32–40.
[12] M. Fahmi, M. Nasir, H. Juahir, N. Roslan, I. Mohd, and N. A. Shafie, "Artificial Neural Networks Combined with Sensitivity Analysis as a Prediction Model for Water Quality Index in Juru River , Malaysia," vol. 1, no. 3, **(2011)**, pp. 1–8.
[13] S. Wechmongkhonkon, N. Poomtong, and S. Areerachakul, "Application of Artificial Neural Network to Classification Surface Water Quality," World Acad. Sci. Eng. Technol., vol. 6, **(2012)**, pp. 205–209.
[14] A. Naumoski and K. Mitreski, "Naïve Bayes technique for diatoms classification with discretised input" ,**(2010),** pp. 21–30.
[15] W. Huang, F. Huang, and J. Song, "An SVM model for Water Quality Monitoring Using Remote Sensing Image", vol. 1, no. 4, **(2010)**, pp. 186–189.
[16] M. Bouamar and M. Ladjal, "Evaluation of the performances of ANN and SVM techniques used in water quality classification" , **(2007)**, pp. 1047–1050.
[17] L. J. C. M. M. Xiaoyan, "Groundwater Quality Assessment Based on Support Vector Machine 1," pp. 173–178.
[18] J. Camejo and O. Pacheco, "Classifier for Drinking Water Quality in Real Time", **(2013)**, pp. q–4.
[19] N. S. Kamarudin, M. Makhtar, S. A. Fadzli, M. Mohamad, F. S. Mohamad, and M. F. Abdul Kadir, "Comparison of Image Classification Techniques using Caltech 101 Dataset", J. Theor. Appl. Inf. Technol., vol. 71, no. 1, **(2015)**.
[20] A. G. K. Janecek, W. N. Gansterer, M. A. Demel, and G. F. Ecker, "On the Relationship Between Feature Selection and Classification Accuracy", **(2008)**, pp. 90–105.
[21] Y. Kim, W. N. Street, and F. Menczer, "Feature Selection in Data Mining", **(1999)**.
[22] M. Dash and H. Liu, "Feature Selection for Classification", vol. 1, **(1997)**, pp. 131–156.
[23] G. Doquire and M. Verleysen, "Feature Selection for Multi-label Classification Problems", **(2011)**.
[24] B. Zeng, Z. Luo, and J. Wei, "Sea Water Pollution Assessment Based On Ensemble of Classifiers" , **(2008),** pp. 241–245.
[25] G. Pirlo, C. A. Trullo, and D. Impedovo, "A feedback-based multi-classifier system", Int. Conf. Doc. Anal. Recognit, **(2009)**.
[26] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software," ACM SIGKDD Explorations Newsletter, vol. 11, **(2009),** p. 10.
[27] S. S. Aksenova, "Machine Learning with WEKA", California, **(2004)**.
[28] Z. Zainudin, "Benchmarking River Water Quality in Malaysia", **(2010)**, pp. 12–15.
[29] G. I. Salama, M. B. Abdelhalim and M. A. Zeid, "Breast Cancer Diagnosis on Three Different Datasets Using Multi-Classifiers", Int. J. Comput. Inf. Technol, vol. 1, no. 1, **(2012)**.

Authors

Rosaida Rosly has recently finished her bachelor degree of science computer from the University Sultan Zainal Abidin (UniSZA), Terengganu, Malaysia. She is currently continues study in master level at UniSZA. Her main research interests include data mining.

Mokhairi Makhtar received his PhD in Computer Science from Bradford University, United Kingdom. He received his MIT from UK Malaysia, BIT from the University of Malaya, and Dip. IT from KUSZA, Malaysia, respectively. He is currently a senior lecturer in the Department of Computer Science at the University of Sultan Zainal Abidin (UniSZA). His current research interests include artificial intelligence, machine learning, and software engineering.

Mohd Khalid Awang received his BS in Computer Science from University of Utara Malaysia (UUM) and MSc in Computer Science from IU Bloomigton, U.S. He is currently a senior lecturer in the Department of Computer Science at the University of Sultan Zainal Abidin (UniSZA). His current research interests include artificial intelligence, machine learning, and E-Learning.

Mohd Nordin Abdul Rahman received his PhD in Computer Science from University of Malaysia Terengganu (UMT), Malaysia. He received his MSc and Bsc in Computer Science from UK Malaysia, respectively. He is currently an assistant professor in the

Department of Computer Science at the University of Sultan Zainal Abidin (UniSZA). His current research interests include software engineering and machine learning.

Mustafa Mat Deris received his PhD from University Putra Malaysia in 2002. He is a professor of computer science in the Faculty of Computer Science and Information Technology, University of Tun Hussein Onn (UTHM), Malaysia. He received his MSc in Computing from the University of Bradford, United Kingdom and Bsc in Mathematics from UPM, Malaysia, respectively. He has successfully supervised six PhD students and published more than 170 papers in journals and conference proceedings. His current research interests include distributed databases, data grid, data mining, and soft computing. He has successfully supervised six PhD students and currently he is supervising nine PhD students and published more than 170 papers in journals and conference proceedings. He has appointed as editorial board member for Journal of Next Generation Information Technology, JNIT, Korea, and Encyclopedia on Mobile Computing and Commerce, Idea Group, USA, ***Guest editor*** of International Journal of BioMedical Soft Computing and Human Science for Special Issue on "Soft Computing Methodologies and Its Applications" a ***reviewer*** of several international journals such as IEEE Transaction on Parallel and Distributed Computing, Journal of Parallel and Distributed Databases, Journal of Future Generation on Computer Systems, Elsevier, Journal of Cluster Computing, Kluwer, and Journal of Computer Mathematics, Taylor & Francis, UK.

