International Journal of Multimedia and Ubiquitous Engineering
Vol. 10, No. 6 (2017), pp. 13-22
http//dx.doi.org/10.14257/ijmue.2017.10.6.02

Domain Specific Predictive Analytics: A Case Study With R

Syed Muzamil Basha, Yang Zhenning , Dharmendra Singh Rajput*,
Iyengar N.Ch.S.N and Ronnie D. Caytiles#

School of Computer Science and Engineering. VIT University, Vellore, T.N., India.
**School of Information Technology and Engineering, VIT University, Vellore, T.N., India*
#Multimedia Engineering department, Hannam University, Daejeon, Korea.
muza.basha@gmail.com, yang.zhenning2015@vit.ac.in,
dharmendrasingh@vit.ac.in, nchsniyengar48@gmail.com, rdcaytiles@gmail.com

Abstract

As part of our research work Predictive analytics, we are interested to perform experiments on the areas, Supply Chain Risk Management, Credit Scoring and Bankruptcy Prediction. When comparing to previous studies on this topic, our research is novel in the following areas. All the experiments carried out in this paper have used three different application specific data repositories that are described in detail in Design and implementation section. Focused on making use of traditional predictive techniques Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA) and compared their performance with respect to Accuracy, Misclassification, Precision, Recall, prevalence and F-Score. we compared the performance of classification algorithms like: Naive Bayes, K Nearest Neighbor with respect to Error in Classification. Analyzed the performance of Model Averaging generation algorithm with respect to average of Markov Blanket size, Neighbourhood size and Branching factor. The main finding from our research is LDA is very good choice when modeling Supply Chain Risk Management, Credit Scoring and Bankruptcy Prediction.

Keywords: *Supply Chain Risk Management, Credit Scoring, Bankruptcy Prediction.*

1. Introduction

Bayesian networks typically represent the joint probability distribution function of all the variables in the network as a product of the smaller conditional probability distributions of each variable by exploiting the expansion rule:

$$p(X \mid \xi) = \sum_{i-1}^{n} p(X \mid Y = y_i, \xi)\, p(y = y_i \mid \xi) \tag{1}$$

The notion of conditional independence must be formally defined since the theory of Bayesian networks relies on it. If we assume that we only have our background knowledge x , we can usually as certain whether two events would be independent or dependent. Suppose we have two events, E1 and E2 about which we know little of their dependence to each other. Now suppose for the moment that we observed a third event, E3, with which we can now conclude the first two events as independent. This concept is known as conditional independence, more formally:

$$p(E_1, E_2 \mid E_3, \xi) = p(E_1 \mid E_3, \xi)\, p(E_2 \mid E_3, \xi) \tag{2}$$

While this method is common with in Bayesian network models, its form is difficult to update if we wish to learn new information. A solution to this is to assume that the probability of an events X_i belongs to X a random variable q. We would then have $P(X_i)$ =

ISSN: 1975-0080 IJMUE

E[deta]. Typically deta is assumed to have as its prior distribution the beta distribution *beta(a_1;a_2)*. When given the new information *X*, we must change our distribution to incorporate the new data. To determine the posterior distribution of deta, Bayes theorem and the likelihood function of binomial sampling are used to find that *Beta(a_1+n_1;a_2+n_2)*, where n_1,n_2 are the number of observations for each outcome. By representing the distribution this way, learning becomes relatively easy. Thus we have our updated belief of X_i as:

$$p(Xi \mid \xi) = E(\phi \mid \xi) = \frac{\alpha_1 + n_1}{\alpha_1 + n_1 + \alpha_2 + n_2} \tag{3}$$

We must note that we took advantage of the implicit assumption that X_i had no parents.

The author Garvey et al. (2015) have designed model for supply chain risk propagation with the following assumption.

1. The complete structure of the supply chain network is known.
2. Total Number of risks indentified are modeled as a binary random entity.
3. All conceivable risks to a supply network for location specific (node/edge) level in the network have taken in to account.
4. With a given set of risks, the regular association among the risks are determined in a procedural and objective manner that helps in constructing an acyclic directed graph.
5. With a given set of risks, the procedure for identifying the regular patterns of risks must be based on the design of the supply network.
6. The data for all risks and conditional probability tables distributions can be identified completely well in advance.
7. The established network results for tractable probabilistic inference.
8. The resulting BN constructed using the risks inherent in a supply chain network using the procedure given below is the best fit network to the data considered.
9. All risks that are dependent on business decisions are static and have only a single distribution.

(Jinqiu Hu *et al.* 2015) considered to design a model for dynamic fault propagation which effects in a petrochemical system. Made an hazard and operability (HAZOP) study to understand the petrochemical system, and derive all the possible deviations and their corresponding potential fault causes are analyzed in detail. The dynamic Bayesian network (DBN) was introduced, to build the fault causal relationships in the complex system. At last the inherent inference mechanism of DBN, can found out accurately possible initial reasons happened in the fault interdependency network when abnormal events occur. (Abroon Qazi *et. al.*, 2017) proposed a Supply Chain Risk Management process that integrating all stages of the risk management process and also, Interdependency between risk sources, risks and mitigation strategies are captured. Introduced a new risk measures for ranking interdependent supply chain risks. Shapley value is used to determine a fair allocation of resources to critical risks. Finally, proposed a method for selecting optimal risk mitigation strategies. (Denise Beaudequin *et. al.*, 2016) have focused on Quantitative microbial risk assessment of wastewater reuse expressed as a Bayesian network and evaluated the major influences in exposure pathways. Finally, performed interactive assessment of risk mitigation measures.

1.1. Credit Scoring

Credit score is a number that represents a summary of individuals' credit history and credit rating. This score has a range(300 - 900) and Maximum (*i.e.*, 900) being the best

score. Individuals with no credit history will have a score of negative one. If the credit history is less than six months, the score will be Zero. CIBIL credit score takes time to build up and usually it takes between eighteen to thirty six months or more of credit usage to obtain a satisfactory credit score. (Joaquin Abellan *et. al.*, 2017) had made a small improvements in credit scoring system to get great profits. Make use of ensembles of classifiers to achieve the better results for credit risk assessment. conducted experiments to show that the credal decision tree classifier is the best one to be used in ensembles. (Yufei Xia *et. al.*, 2017) proposed a novel boosted tree model for credit scoring. This model is proved to outperform several baseline techniques. This model is validated on five different datasets over five performance metrics. The feature importance scores and decision chart enhance model interpretation. (Zhiyong Li *et. al.*, 2017) proposed a Semi-supervised Support Vector Machines for reject inference which uses information of both the accepted and rejected applicants, deals with labelled and unlabelled classes of the outcome. Finally, this model is tested on real consumer loans with a low acceptance rate and achieved high Predictive accuracy compared to traditional methods.

Table 1. Data Repositories

Author	Name of the Case study	Number of Variables	Number of observations
(Sachs et .al 2005)	Bayesian Network Approach in supply chain Management	11	853
(Lauritzen et .al 1988)	Credit Scoring	21	1000
(Ledolter et. al 2013)	Bankruptcy Prediction	3	85

1.2. Bankruptcy Prediction

It is the art of predicting bankruptcy and different measures of financial suffering of public firms'. It is a wide area of finance and accounting research. The significance of the area is relevant to creditors and investors in evaluating the likelihood that a firm may go bankrupt. (Maciej Zięba *et. al.*, 2016) proposed a novel ensemble model for bankruptcy prediction. used Extreme Gradient Boosting as an ensemble of decision trees, and proposed a new approach for generating synthetic features to improve prediction. Finally, evaluated the presented method on real-life data of Polish companies. (Deron Liang *et. al.*, 2016) Combined corporate government indicators and financial ratios for examining bankruptcy prediction. Made a study on seven and five different categories of financial ratios and corporate government indicators. (Hyun-Jung Kim *et. al.*, 2016) have examined the effectiveness an optimized cluster-based under sampling technique. used a Genetic Algorithm -based optimization approach for selecting the appropriate instances. This proposed method is successfully applied to the bankruptcy prediction problem. (Philippe du Jardin *et. al.*, 2017) estimated the degree of stability of firm financial health over several years. used quantization method for firms sharing the same level of stability. compared the performance of developed models to that of traditional models. This method improves the mid-term forecasts when the horizon is higher than two years.

2. Design and Implementation

In Table 1, we have given the complete details of the data repositories used. Here we consider both Numeric and Factor type.

Table 2. Summary of the Bayesian Network Approach

Random/Generated Bayesian network			
Model		undirected graph	Sachs data with catnet in R
Nodes		**11**	**11**
Arcs	Undirected	**7**	**0**
	Directed	**0**	**9**
Average markov blanket size		**1.27**	**1.64**
Average neighbourhood size		**1.27**	**1.64**
Average branching factor		**0.00**	**0.82**
generation algorithm		Model Averaging	Model Averaging
significance threshold		0.85	0.936

In Table 2 and Table 3, we have given the complete details of the summary generated using R on the datasets used in the experiment. In Table 3 the most influenced attributes like Duration, Amount, Installation and Age are considered to be important for credit scoring.

Table 3. Summary of the Credit Scoring

Default	**Duration**	**Amount**	**Installment**	**Age**
Min. :0.0	Min. : 4.0	Min. : 250	Min. :1.000	Min. :19.00
1st Qu.:0.0	1st Qu.:12.0	1st Qu.: 1366	1st Qu.:2.000	1st Qu.:27.00
Median :0.0	Median :18.0	Median : 2320	Median :3.000	Median :33.00
Mean :0.3	Mean :20.9	Mean : 3271	Mean :2.973	Mean :35.55
3rd Qu.:1.0	3rd Qu.:24.0	3rd Qu.: 3972	3rd Qu.:4.000	3rd Qu.:42.00
Max. :1.0	Max. :72.0	Max. :18424	Max. :4.000	Max. :75.00

3. Result

Table 4. LDA: Class Proportions of the Training Set used as Prior Probabilities

N=1000	**Predicted False**	**Predicted True**	
Actual False	TN=669	FP=256	925
Actual True	FN=31	TP=44	75
	700	300	**N=1000**

Table 5. QDA: Class Proportions of the Training Set used as Prior Probabilities

N=1000	**Predicted False**	**Predicted True**	
Actual False	TN=628	FP=221	849
Actual True	FN=72	TP=79	151
	700	300	**N=1000**

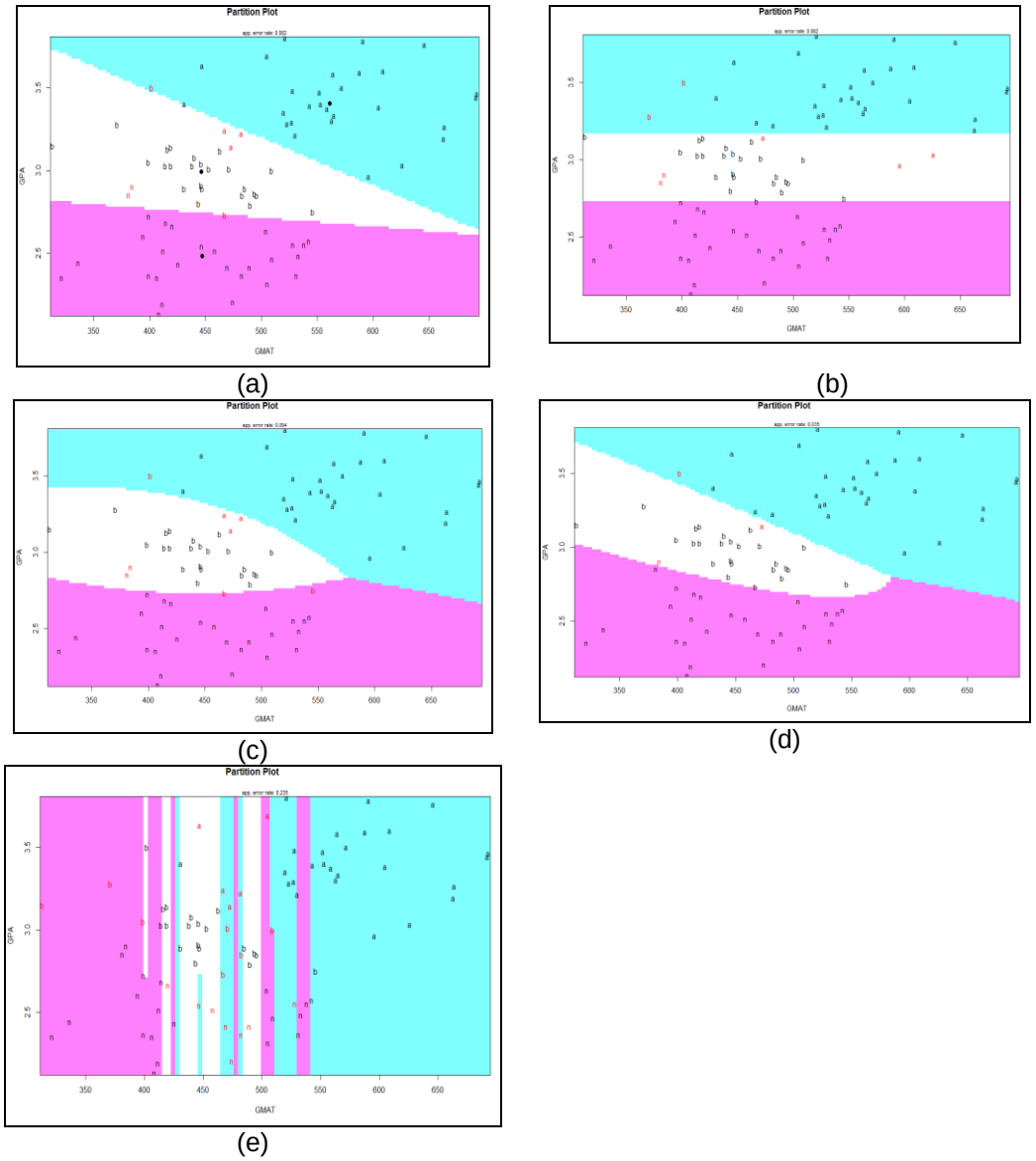

Figure 1. Graphs of Different Machine Learning Models with Classification Error

Table 6. Classification Error of LDA and MLA's

Name of the Method	LDA(a)	Rpart(b)	NB(c)	RDA(d)	SKNN(e)
Classification Error	0.082	0.082	0.094	0.035	0.235

In Table 6, we compared the classification error of LDA with other Machine Learning Algorithms Recursive Partitioning and Regression Trees (Rpart), NaiveBayes (NB), Regularized Discriminant Analysis (RDA), Simple k nearest Neighbours (SKNN). From this values one can easy understand that RDA is having less classification error the same is represented in the Figure 1(d).

.

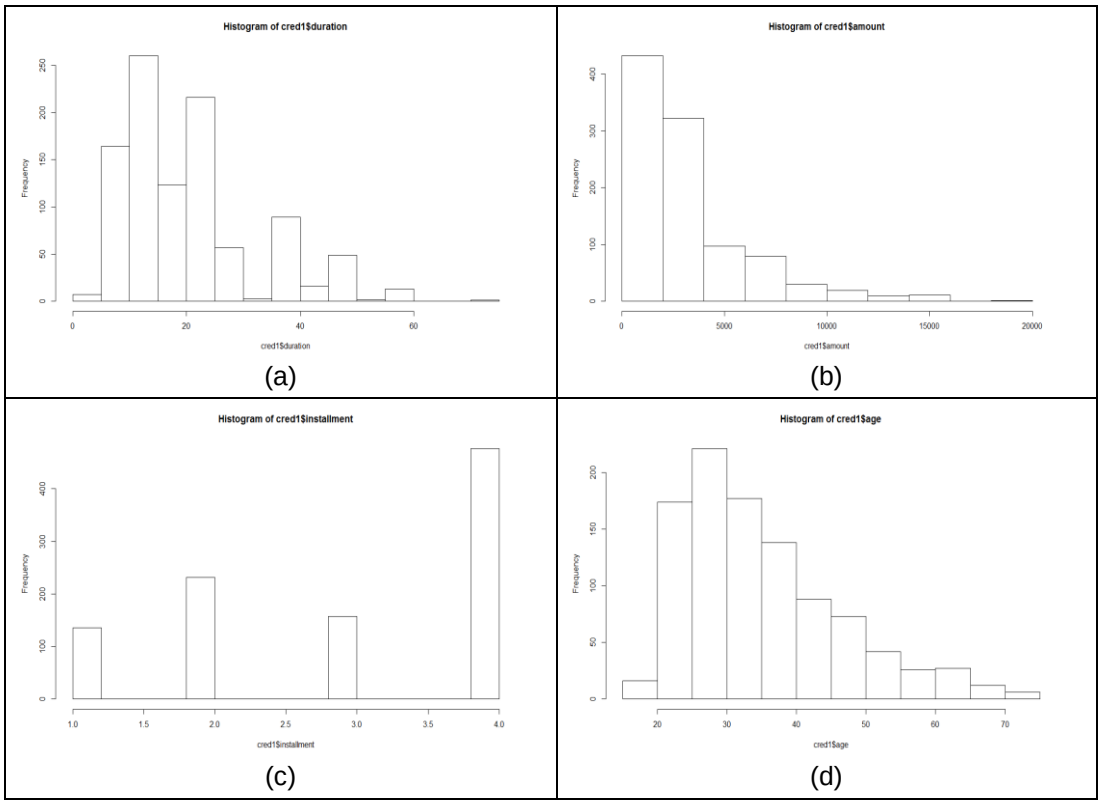

Figure 2. Histogram of Credit Scoring Attributes.

In Figure 2, the histogram plotted for credit with respect to duration, amount, installment and age.

From Table 8, one can understand that both LDA and QDA models have same Mean Error of 0.102 with 100 repetations.

Table 7. LDA and QDA: Comparison of Performance

Parameters	**Formula**	**LDA**	**QDA**
Accuracy(A)	$A = \frac{TP + TN}{Total}$	0.713	0.707
Misclassification Rate(MCR)	$MCR = \frac{FP + FN}{Total}$	0.287	0.293
Recall(R)	$R = \frac{TP}{ActualTrue}$	0.586	0.523
Precision(P)	$P = \frac{TP}{\Pr edictedTrue}$	0.146	0.263
Prevalence(PV)	$PV = \frac{ActualTrue}{Total}$	0.075	0.151
F Score(FS)	$FS = 2 \times \frac{R \times P}{R + P}$	0.234	0.175

Table 8. Mean Error of LDA and QDA with 100 Repetations

	LDA	QDA
Mean Error	0.102	0.102

4. Discussion

In the result section Table 4 and 5 are representing the confusion matrix generated using LDA and QDA methods with 1000 as sample size. The experiment is repeated for 100 times and the Mean Error of both LDA and QDA is plotted. Classification error of different machine learning models are plotted and that can be interpreted as LDA (0.082) is the best compared to other models and the same is plotted in Figure 1. In Table 7 predictive techniques, Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) performance are compared with respect to Accuracy, Misclassification, Precision, Recall, prevalence and F-Score.

5. Conclusion

We would like to conclude that in our research the predictive techniques, Linear Discriminant Analysis (LDA) outperforms Quadratic Discriminant Analysis (QDA) and compared their performance with respect to Accuracy(0.713), Misclassification(0.287), Precision(0.146), Recall(0.586), prevalence(0.075) and F-Score(0.234). The experiment is repeated for 100 times and acquired the Mean Error(0.102) in the both the techniques. Additionally, we compared the performance of classification algorithms like: Naive Bayes, K Nearest Neighbor with respect to Error in Classification(0.082) using LDA and analyzed the performance of Model Averaging generation algorithm with respect to average of Markov Blanket size (1.27 and 1.64), Neighbourhood size(1.27 and 1.64) with Branching factor (0.82).

Acknowledgment

We would like to thank our VIT University for providing us the all the research facilities requirement for publishing Scopes Indexed journals. At the same time we would like to thank all the reviewers, who help us to improve the quality of our paper.

References

[1] FJ. Nogales , J. Contreras , AJ. Conejo, R. Espínola, "Forecasting next-day electricity prices by time series models", IEEE Transactions on power systems, vol. 17, no.2, **(2002)**, pp. 342-8.
[2] Abroon Qazi, John Quigley, Alex Dickson, Şule Onsel Ekici, "Exploring dependency based probabilistic supply chain risk measures for prioritising interdependent risks and strategies", European Journal of Operational Research, vol. 259, no. 1, **(2017)**, pp. 189-204.
[3] Denise Beaudequin, Fiona Harden, Anne Roiko, Kerrie Mengersen, "Utility of Bayesian networks in QMRA-based evaluation of risk reduction options for recycled water", Science of The Total Environment, vol. 541, **(2016)**, pp. 1393-1409.
[4] Deron Liang, Chia-Chi Lu, Chih-Fong Tsai, Guan-An Shih, "Financial ratios and corporate governance indicators in bankruptcy prediction: A comprehensive study", European Journal of Operational Research, vol. 252, no. 2, **(2016)**, pp. 561-572.
[5] Garvey, M. D., Carnovale, S., & Yeniyurt, S, "An analytical framework for supply network risk propagation: A Bayesian network approach", European Journal of Operational Research, vol. *243*, no. 2, **(2013),** pp. 618-627.
[6] Hyun-Jung Kim, Nam-Ok Jo, Kyung-Shik Shin, "Optimization of cluster-based evolutionary under sampling for the artificial neural networks in corporate bankruptcy prediction", Expert Systems with Applications, vol. 59, no. 1, **(2016)**, pp. 226-234.
[7] Jinqiu Hu, Laibin Zhang, Zhansheng Cai, Yu Wang, Anqi Wang, "Fault propagation behavior study and root cause reasoning with dynamic Bayesian network based framework", Process Safety and Environmental Protection, vol. 97, **(2015)**, pp. 25-36.

[8] Joaquín Abellán, Javier G. Castellano, "A comparative study on base classifiers in ensemble methods for credit scoring", Expert Systems with Applications, vol. 73, no. 1, **(2017)**, pp. 1-10.
[9] K. Sachs, O. Perez, D. Pe'er, D. A. Lauffenburger and G. P. Nolan, "Causal Protein-Signaling Networks Derived from Multiparameter Single-Cell Data", Medical Science, vol. 308, **(2005)**, pp. 523-529.
[10] Ledolter, J, "Data mining and business analytics with R", John Wiley & Sons publishers, **(2013),** pp. 173-293.
[11] Maciej Zięba, Sebastian K. Tomczak, Jakub M. Tomczak, "Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction", Expert Systems with Applications, vol. 58, no. 1, **(2016)**, pp. 93-101.
[12] Philippe du Jardin, "Dynamics of firm financial evolution and bankruptcy prediction", Expert Systems with Applications, vol. 75, no. 1, **(2017)**, pp. 25-43.
[13] S. Lauritzen, D. Spiegelhalter, "Local Computation with Probabilities on Graphical Structures and their Application to Expert System", Journal of the Royal Statistical Society: Series B (Statistical Methodology), vol. 50, no. 2, **(1988)**, pp. 157-224.
[14] Yufei Xia, Chuanzhe Liu, YuYing Li, Nana Liu, "A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring", Expert Systems with Applications, vol. 78, no. 1, **(2017)**, pp. 225-241.
[15] Zhiyong Li, Ye Tian, Ke Li, Fanyin Zhou, Wei Yang, "Reject inference in credit scoring using Semi-supervised Support Vector Machines", Expert Systems with Applications, vol. 74, no. 1, **(2017)**, pp. 105-114.

Authors

Syed Muzamil Basha, he had his Bachelor of Science in Information Technology at SITAMS, MTech in Information Technology (Networking) at VIT University and currently doing his research at VIT University. His research area are Wireless Sensor Networks, Text Mining and Big Data Predictive Analytics.

Yang Zhenning, he is pursuing M.ScComputer Science at School of Computing Science and Engineering, VIT University, Vellore. His area of interests are Algorithm design an Pattern Recognition, operating Systems and cloud computing

Dharmendra Singh Rajput, working as Associate Professor in the Department of Software and Systems Engineering, School of Information Technology and Engineering, VIT University. His research area are Data Mining and Big Data Predictive Analytics.

N. Ch. S. N. Iyengar (b 1961), he currently Senior Professor at the School of Computer Science and Engineering at VIT University, Vellore-632014, Tamil Nadu, India. His research interests include Agent-Based Distributed Computing, Intelligent Computing, Network Security, Secured Cloud Computing and Fluid Mechanics. He had 30+ years of experience in teaching and research , guided many scholars, has authored several textbooks and had nearly 200+

research publications in reputed peer reviewed international journals. He served as PCM/reviewer/keynotespeaker/Invited speaker for many conferences. He serves as editorial board member for many international journals, reviews papers for many conferences with an interest of serving to the education community.

Ronnie D. Caytiles, he had his Bachelor of Science in Computer Engineering- Western Institute of Technology, Iloilo City, Philippines, and Master of Science in Computer Science– Central Philippine University, Iloilo City, Philippines. He finished his Ph.D. in Multimedia Engineering, Hannam University, Daejeon, Korea. Currently, he serves as an Assistant Professor at Multimedia Engineering department, Hannam University, Daejeon, Korea. His research interests include Mobile Computing, Multimedia Communication, Information Technology Security, Ubiquitous Computing, Control and Automation.

