

A Classification Approach with Different Feature Sets to Predict the Quality of Different Types of Wine using Machine Learning Techniques

Satyabrata Aich¹, Ahmed Abdulhakim Al-Absi², Kueh Lee Hui³, John Tark Lee³ and Mangal Sain⁴

¹Department of Computer Engineering, Inje University, South Korea

²Department of Computer Engineering, Kyungdong University- Global Campus, Gangwondo, South Korea

³Dept. of Electrical Engineering, Dong-A University, South Korea

⁴Department of Computer Engineering, Dongseo University, South Korea

satyabrataaich@gmail.com, absiahmed@kduniv.ac.kr, leehkueh@dau.ac.kr, jtlee@dau.ac.kr, mangalsain1@gmail.com

Corresponding author email id: absiahmed@kduniv.ac.kr

Abstract—In the past few years, with the availability of lot of wine brands it is difficult to identify the good quality wines. Good quality wine depends on the so many important factors such as chemical, scientific as well as technical factors. However in the previous study the researchers always focus on the subjective study to define the quality of wine. The result based on the subjective study takes much time as well as it is not effective compared to the objective study with the analytical methods. In the last few year's machine learning techniques caught lot of attention in every field. Most of the machines learning techniques are able to produce highly accurate result that compels most of the data scientist to implement it in case of predictive analytics. In the past few works related to wine data has been studied using different classifiers, however so far nobody has compared the performance metrics of the different classifiers with different feature sets to predict the quality of different type of wine by considering several factors. In this paper a new approach has been proposed by considering different feature selection algorithm such as Principal Component Analysis (PCA) as well as Recursive Feature Elimination approach (RFE) approach for feature selection and nonlinear decision tree based classifiers for analyzing the performance metrics. We found accuracies ranging from 94.51% to 97.79% with different feature sets using Random Forest classifier. This analysis will help the wine experts to know the important factors to consider while selecting the good quality wine.

Keywords—*machine learning; feature selection; classifiers; performance metrics; wine quality*

I. INTRODUCTION

In the last few years the consumption of wine has been increased because it has positive correlation with the health, specifically it is beneficial to keep in check the variability of heart rate [1]. The increased amount of consumption of wine forced the wine industry to go for some quality assessment test and certification. At the same time they are also aware of the cost which is one of the important factor while maintaining the quality of the wine. Different type of wine has different purposes and the chemical concentration is also different for different types of wine. To maintain the quality with less cost, it is important to know the contribution of different chemical

attributes used in different types of wine [2]. In the past due to lack of technological resources it become difficult for most of the industries to classify the wines based on the chemical analysis as it takes lot of time and also need more money. These days with the advent of the machine learning techniques it is possible to classify the wines as well as it is possible to figure out the importance of each chemical analysis parameters in the wine and which one to ignore for reduction of cost. The performance comparison with different feature sets will also help to classify it in a more distinctive way. In this paper an intelligent approach is proposed by considering recursive feature elimination (RFE) algorithm for feature selection as well as Principal Component Analysis (PCA) based approach for feature selection considering the nonlinear classifiers to predict the quality in red wine as well as the white wine.

The structure of the paper is organized as follows: Section II presents the past work related to this field. Section III describes about the methodologies used for this research work. Section IV describes about the result of feature selection as well as the result of classification. Section V describes about the conclusion and future work.

II. RELATED WORKS

Some of the machine learning approaches has been used in the past works for prediction of price and quality of wine. Some of the past works are mentioned below. Yeo et al used Gaussian process regression and multi-task learning to predict the wine price. They have used historical price of wine data to predict the price of the wine. They found that advanced machine learning technique has the potential for prediction of wine price [3]. Ashenfelter mentioned that the quality and price of wine depends on the weather on which the grapes are created. He derived a price equation using several factors. He found climate change and expert opinion has a major role to play while deciding the wine price [4]. Ribeiro et al have done the prediction of wine vinification, which is one of the ways to measure the wine quality using data mining tools. They have used Decision trees, Artificial Neural Network, and Linear Regression as the data mining techniques to predict the organoleptic parameters from the chemical parameters of

vinification process. They found good accuracies in all the techniques [5]. Lee et al proposed a decision tree based method to predict the wine quality. They compared their approach with the WEKA based data mining tool using three machine learning approaches such as SVM, Bayes Net and Multi Perceptron and they found their proposed method is better compared to other mentioned approaches [6]. The above past work motivated us to try different machine learning approaches with different feature selection algorithm to predict the quality of wine. The feature selection algorithms used in this paper are PCA and RFE. This method will help the performance measure of different classification approach used for the prediction. The classification approaches used are RPART, C4.5, PART, Bagging CART, RF and Boosting C5.0.

III. METHODOLOGIES

The flow chart of the proposed methodology is shown in the fig. 1.

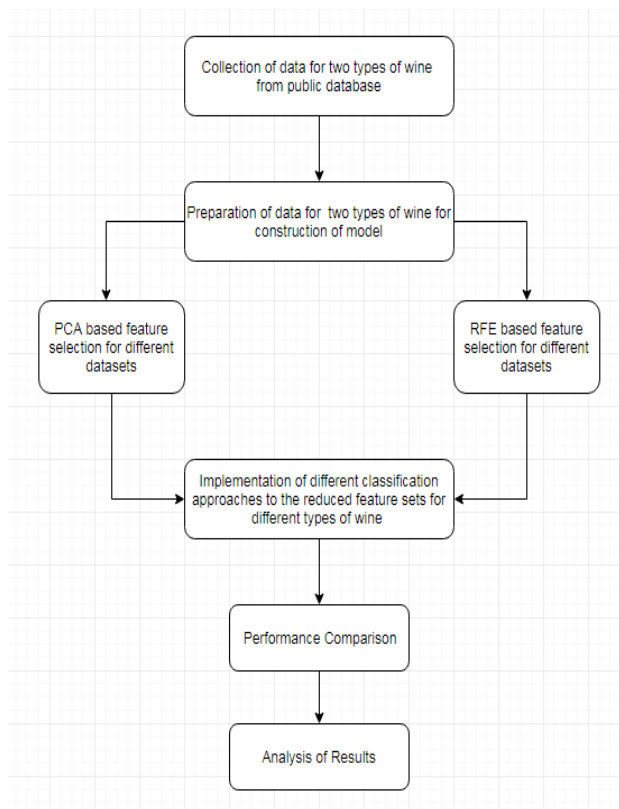


Figure 1. Flow chart of proposed method

A. Data Collection

The wine data set is publicly available in the database of UCI. The two datasets are related to red and white variants of the Portuguese "Vinho Verde" wine. This data set contains the physiochemical variables as well as sensory variables, altogether there are 12 attributes [7].

B. Feature Reduction

The RFE selection method is basically a recursive process that ranks features according to some measure of their importance [8]. Principal Component Analysis (PCA) is a tool that is used for compression of data and extraction of information [9].

C. Classification model

In this paper we have used nonlinear classifier with decision tree for classification of groups are as follows Recursive partitioning decision tree(RPART), C4.5, PART, Bagging classification and Regression tree(Bagging CART), Random Forest and Boosted C5.0.

1) *RPART*: Recursive partition based classifier works basically on the principle of splitting technique. It is called recursive because it keeps on splitting until some stopping criterion has reached [10].

2) *C4.5*: It is basically an improvement over ID3 algorithm. It works on the principle of information entropy. It is also called as statistical classifier [11].

3) *PART*: It is basically a rule based system and it produces some pruned decision tree based on C4.5 and then it try to derive some rule and then remove the instances covered by the rule. It continues until all the instances have been covered by the derived rule [12].

4) *Bagging CART*: Bagging is working in the principle of manipulation of the training sets by taking the average of all training sets. Basically the training sets are created from the original sets by random replacements [13]. CART with bagging is used to improve the performance [14].

5) *Random Forest*: It is basically a collection of random decision trees. Each decision tree in the forest learned from random training sets and random feature sets and then one probability score assigned to each of them and finally overall probability is calculated by taking into account all decision trees [15, 16].

6) *C5.0*: It is the improved version of C4.5 and it has all the advantages over the ancestors [17].

D. Performance Measure Metrics

The parameters used to compare the performance and validations of classifier are as follows: accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV). The sensitivity is defined as the ratio of true positives to the sum of true positives and false negatives. The specificity is defined as the ratio of true negatives to the sum of false positives and true negatives. In our research we have used the Positive predictive value and negative predictive value to check the present and absent of one type of wine. So the PPV is the probability that the one type of wine is present given a positive test result and NPV is the probability that the one type of wine is absent given a negative test result [21]. Accuracy is defined as the ratio of number of correct predictions made to the total prediction

made and the ratio is multiplied by 100 to make it in terms of percentage.

IV. RESULTS AND DISCUSSION

We have divided the data into two groups such as train data and test data. We trained each classifier based on the trained data and predict the power of classifier on the test data. So each classifier able to show all the performance metrics such as accuracy, sensitivity, specificity, PPV, and NPV based on the test data. We have applied all the classification techniques to the RFE based reduced feature sets for two types of wine as well as PCA based reduced feature sets for two types of wine to measures the performance parameter with respect to each classifier. We separated each performance measures with respect to RFE and PCA sets and plot the column plot for better visualization. The results of each performance measure with respect to two feature sets are shown in the figure 2, 3, 4, 5, and 6 respectively for red wine and 7, 8, 9, 10 and 11 for white wine.

A. Comparison of Accuracy for red wine

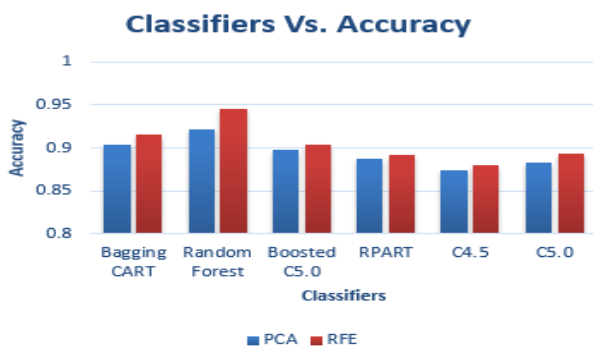


Figure 2. Comparison of Accuracy on PCA and RFE sets

Fig. 2, shows that RFE based feature sets with Random forest shows highest accuracy of 94.51%.

B. Comparison of Sensitivity for red wine

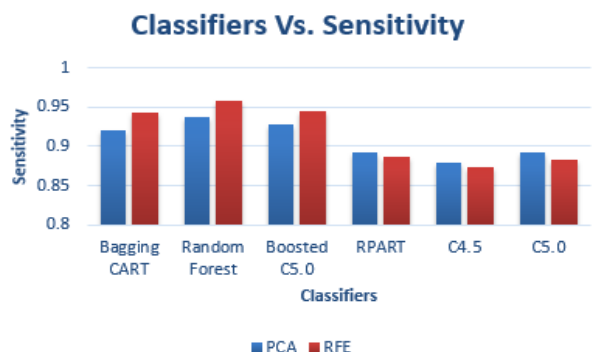


Figure 3. Comparison of Sensitivity on PCA and RFE sets

Fig.3, shows that RFE based feature sets with Random forest shows highest sensitivity of 0.9575.

C. Comparison of Specificity for red wine

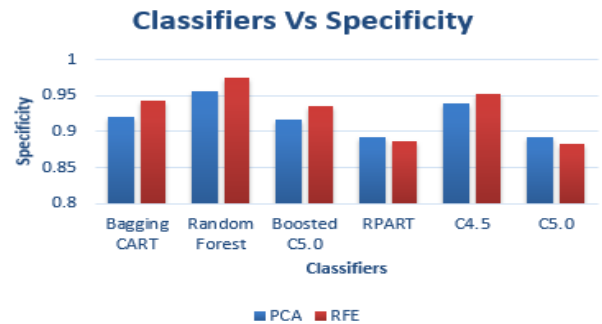


Figure 4. Comparison of Specificity on PCA and RFE sets

Fig. 4, shows that RFE based feature sets with Random forest shows highest specificity of 0.9756.

D. Comparison of PPV for red wine

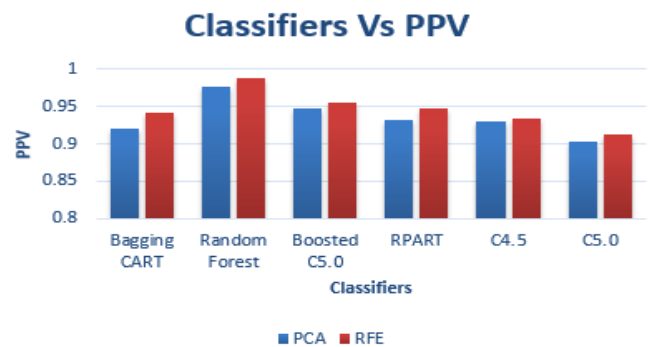


Figure 5. Comparison of PPV on PCA and RFE sets

Fig. 5, shows that RFE based feature sets with Random forest shows highest PPV of 0.9889.

E. Comparison of NPV for red wine

Fig. 6, shows that RFE based feature sets with Random forest shows highest NPV of 0.9891.

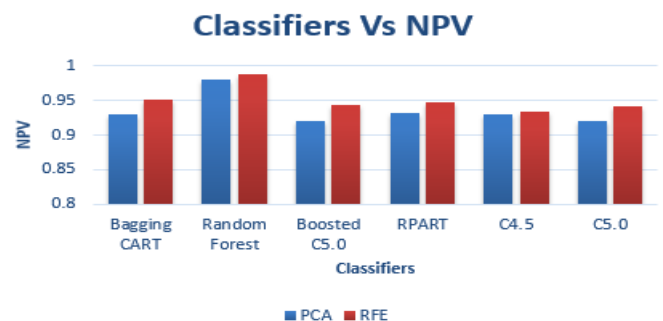


Figure 6. Comparison of NPV on PCA and RFE sets

F. Comparison of Accuracy for white wine

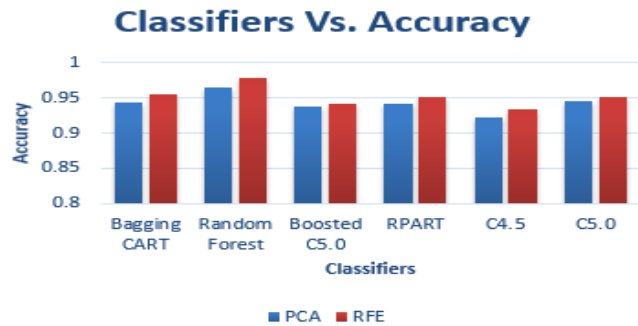


Figure 7. Comparison of Accuracy on PCA and RFE sets

Fig. 7, shows that RFE based feature sets with Random forest shows highest accuracy of 97.79%.

G. Comparison of Sensitivity for white wine

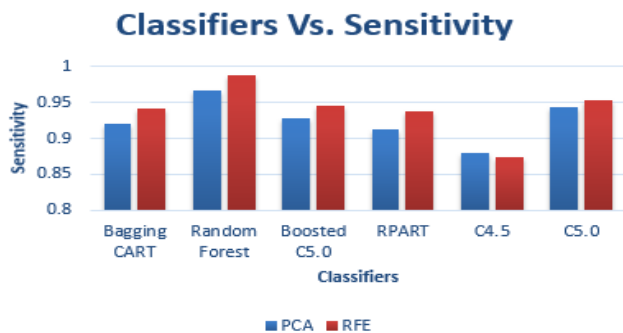


Figure 8. Comparison of Sensitivity on PCA and RFE sets

Fig. 8, shows that RFE based feature sets with Random forest shows highest sensitivity of 0.9882.

H. Comparison of Specificity for white wine

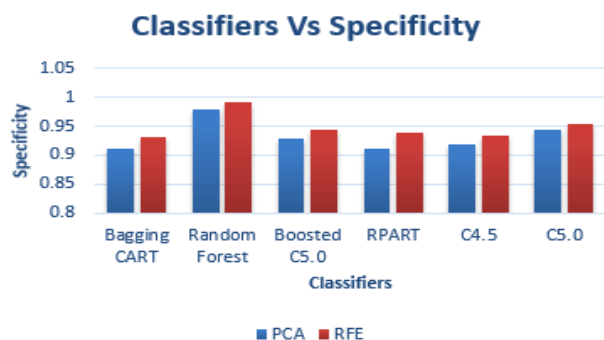


Figure 9. Comparison of specificity on PCA and RFE sets

Fig. 9, shows that RFE based feature sets with Random forest shows highest specificity of 0.9912.

I. Comparison of PPV for white wine

Classifiers Vs PPV

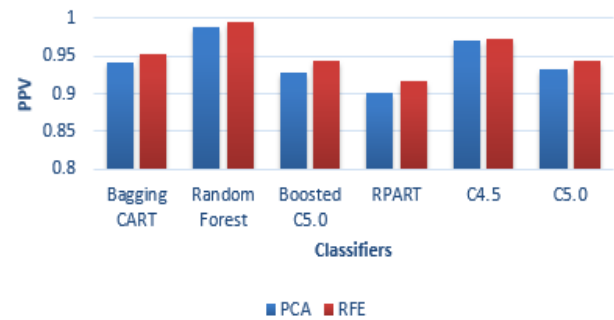


Figure 10. Comparison of PPV on PCA and RFE sets

Fig. 10, shows that RFE based feature sets with Random forest shows highest PPV of 0.9956.

J. Comparison of NPV for white wine

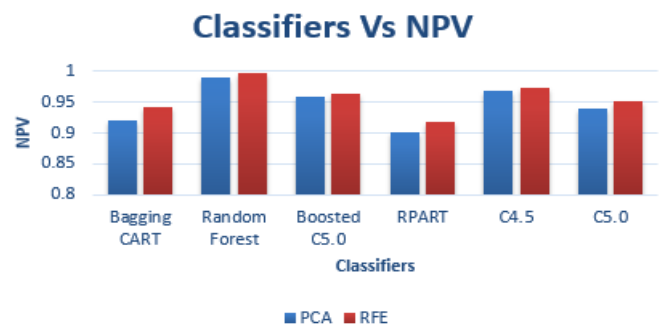


Figure 11. Comparison of NPV on PCA and RFE sets

Fig. 11, shows that RFE based feature sets with Random forest shows highest NPV of 0.9976.

V. CONCLUSION AND FUTURE WORK

This analysis can be opted as an important tool for the prediction of the quality of two types of wine with different feature sets. We have found the important attributes while predicting the quality of white wine as well as quality of red wine by using different feature selection algorithm. In this paper we have used nonlinear classifiers to predict the quality of two types of wines by achieving good classification accuracies ranging from 94.51% to 97.79%. Random forest classifier shows highest accuracy of 94.51% while predicting the quality of red wine with RFE based feature sets at the same time the same classifier shows highest accuracy of 97.79% while predicting the quality of white wine with RFE based feature sets. Overall the RFE based feature sets performed well with the random forest classifier while doing the comparison on the performance metrics compared with the PCA based feature sets. This analysis will help the wine manufactures to shift focus on the important attributes to maintain quality while preparing different types of wine.

REFERENCES

- [1] I. Janszky, M. Ericson, M. Blom, A. Georgiades, J. O. Magnusson, H. Alinagizadeh, and S. Ahnve, "Wine drinking is associated with increased heart rate variability in women with coronary heart disease," *Heart*, 91(3), pp.314-318, 2005.
- [2] V. Preedy, and M. L. R. Mendez, "Wine Applications with Electronic Noses," in *Electronic Noses and Tongues in Food Science*, Cambridge, MA, USA: Academic Press, 2016, pp. 137-151.
- [3] Yeo, M., Fletcher, T. and Shawe-Taylor, J., 2015. Machine Learning in Fine Wine Price Prediction. *Journal of Wine Economics*, 10(2), pp.151-172.
- [4] Ashenfelter, O., 2010. Predicting the quality and prices of Bordeaux wine. *Journal of Wine Economics*, 5(1), pp.40-52.
- [5] Ribeiro, J., Neves, J., Sanchez, J., Delgado, M., Machado, J. and Novais, P., 2009, June. Wine vinification prediction using data mining tools. In *Conference Proceedings, Computing and Computational Intelligence*, Tbilisi, Republic of Georgia (pp. 78-85).
- [6] Lee, S., Park, J. and Kang, K., 2015, September. Assessing wine quality using a decision tree. In *Systems Engineering (ISSE), 2015 IEEE International Symposium on* (pp. 176-178). IEEE.
- [7] P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In *Decision Support Systems*, Elsevier, 47(4):547-553, 2009.
- [8] P. M. Granitto, C. Furlanello, F. Biasioli, and F. Gasperi, "Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products," *Chemometrics and Intelligent Laboratory Systems*, 83(2), pp.83-90, 2006.
- [9] E. R. Hruschka, N. F. Ebecken, "Extracting rules from multilayer perceptrons in classification problems: A clustering-based approach," *Neurocomputing*, 70(2), pp. 384-397, 2006
- [10] Cook EF, Goldman L (1984). "Empiric comparison of multivariate analytic techniques: advantages and disadvantages of recursive partitioning analysis". *Journal of chronic diseases*. 37 (9-10): 721-31.
- [11] S. Vijayarani, and M. Divya, "An efficient algorithm for generating classification rules," *International Journal of Computer Science and Technology*, 2(4), 2011.
- [12] <http://machinelearningmastery.com/non-linear-classification-in-r-with-decision-trees/>, retrieved on 19th August 2017
- [13] L. Breiman, Bagging predictors, *Machine Learning*, 1996, 26(2), pp.123-140.
- [14] X. Sun, Pitch accent prediction using ensemble machine learning. In *Seventh International Conference on Spoken Language Processing*, 2002
- [15] L. Breiman, Random forests. *Machine learning*, 2001, 45 (1), pp.5-32.
- [16] K. Ellis, J. Kerr, S. Godbole, G. Lanckriet, D. Wing, and S. Marshall, "A random forest classifier for the prediction of energy expenditure and type of physical activity from wrist and hip accelerometers", *Physiological measurement*, 35(11), pp.2191-2203, 2014.
- [17] Is See5/C5.0 Better Than C4.5?. 2009. [Online]. Available: <http://www.rulequest.com/see5-comparison.html>
- [18] M. Nilashi, O. bin Ibrahim, H. Ahmadi, and L. Shahmoradi, "An Analytical Method for Diseases Prediction Using Machine Learning Techniques," *Computers & Chemical Engineering*, 106, pp.212-223, 2017.
- [19] <http://www.dtrek.com> Software for Predictive Modelling and Forecasting.
- [20] C.C.Hsu, Y.P.Huang, and K.W.Chang, "Extended Naive Bayes classifier for mixed data," *Expert Systems with Applications*, 35(3), pp.1080-1083, 2008.
- [21] H.B.Wong, and G.H. Lim, "Measures of diagnostic accuracy: sensitivity, specificity, PPV and NPV," *Proceedings of Singapore healthcare*, 20(4), pp.316-318, 2011.



Satyabrata Aich is working as a researcher in the field of computer engineering. He has over four years of teaching, research and industry experience in India and abroad. He has published many research papers in journals and conferences in the realms of Supply Chain Management and data analytics. His research interests are natural language processing, Machine learning, supply chain management, data mining.



Ahmed Abdulhakim Al-Absi is an assistant professor in Department of Computer Engineering (Smart Computing) at Kyungdong University in South Korea. He earned a Ph.D. in computer science from Dongseo University in 2015. He received M.Sc. degree in information technology at University Utara Malaysia in 2011, and B.Sc. degree in computer applications at Bangalore University in 2008. His research interests include Big Data processing, Hadoop, Cloud computing, IoT, Distributed systems, Parallel computing, Bioinformatics, Security, and VANETs.



Kueh Lee Hui is working as an assistant professor at the department of Electrical Engineering, Dong-A University since 2012. She completed her PhD Degrees from Department of Electrical Engineering, Dong-A University, Korea. In 2009 she completed her BS degree in Electronic and Communication, Department of Electronic Engineering, University Malaysia of Sarawak, Malaysia. She also has done MS in 2007 from Malaysia. Her research interests are image processing, face recognition, digital image forensic, intelligent control and control application, power system.

**John Tark Lee**

1979 : BS degree in Electrical Engineering of Dong-A University.

1981 : MS degree in Electrical Engineering of Dong-A University.

1988 : PhD degrees in Electrical Engineering of Chung-Ang University.

1983 ~ 1985 : A researcher in LG Electronic Co. Ltd.

1985 ~ Present : Professor, Department of Electrical Engineering, Dong-A University

Research interests : fuzzy theory, neuron-science, GA, intelligent controller design including nonlinear stability analysis.



Mangal Sain received the M.Sc. degree in computer application from India in 2003 and the Ph.D. degree in computer science in 2011. Since 2012, he has been an Assistant Professor with the Department of Computer Information Engineering, Dongseo University, South Korea. His research interest includes wireless sensor network, cloud computing, Internet of Things, embedded systems, and middleware. He has authored over 50 international publications including journals and international conferences. He is a member of TIIS and a TPC member of more than ten international conferences.