Product Recommendation Based on Content-based Filtering Using XGBoost Classifier

Zeinab Shahbazi, Yung-Cheol Byun*

Jeju National University
Zeinab.sh@jejunu.ac.kr, yungcheolbyun@gmail.com

Abstract

Recommendation systems are a significant part of the machine learning algorithm to recommend related suggestions to the user based on user request. Many online shopping websites that have an acceptable rating information face problem in recommending an item to their users using content-based filtering (CBF) technique. By applying impermanent purchase patterns extracted from sequential pattern analysis (SPA) doesn't make user satisfy from the search result. The objective of this research is XGBoost-based item recommendation using Jeju online shopping mall dataset records to recommend items based on user click information. Based on the output of XGBoost algorithm, we compare the result with other research outputs performance. The proposed CBF recommendation and SPA results successfully shows a better rating than other individual ones.

Keywords: *Recommendation system, XGBoost, Machine Learning, Classification, Content-based filtering*

1. Introduction

There are many types of data in the vast amount available on internet websites in the form of reviews, user feedback, ideas, etc. about any product online shopping that is part of guiding users for a wise decision on their purchases. Furthermore, there are many available blogs which are accessible for users to show their opinion about the purchased item and also, they can review the product [1-3]. The recommendation based on the user's item review and comments can use for decision making.

Information and technology growth has a direct effect on the development of the recommendation system. Recently online shopping become a famous topic among users. The technology of online services develops day by day. This platform permits the user to search for their needs without spending much time and same for sellers to provide the various items for purchasing.

The advantages of this system are for both user and seller. To develop this system more comfortably and professionally, content-based filtering recommendation system considered as a significant way for this process [4-7].

Product recommendation area contains personalization information. However, there is not much research related to high product collaboration that determines to initiate less purchase frequency, high price, estimation basis and take time to decide for purchasing [8].

Recommendation techniques are defined into two categories, content-based filtering and collaborate filtering. Content-based filtering gives the items that are similar to user previous purchase history and collaborate filtering presents similar items with user-selected item [9-11]

Therefore, in this paper, content-based filtering product recommendation using XGBoost machine learning algorithm proposed to recommend items to a user based on user click information. Information collected from user activities and user profile. Remaining of this paper divided as below: Section two presents the literature review of the proposed recommendation system. Section three presents the architectural model of XGBoost-based recommendation. Section four presents the development environment and evaluated the results of the presented recommendation system and finally the conclusion.

2. Literature Review

Recommendation system describes a type of system that recommends suitable items or service to find users interests and not interests. One of the essential aspects of recommendation is how to correctly filter the user behavior that can easily recommend items based on user request [12].

Montaner et al. proposed an advanced classification on an internet-based intelligent recommendation [13]. Applied recommendation systems in this process are content-based filtering. Collaborate filtering and hybrid, which is a combination of the other two methods. Similarity measurement of this system evaluated by naïve Bayesian classifier, cosine similarity, Pearson r correlation, etc.

Table 1 shows the recommendation system based on similarity measurement, recommendation method and recommendation domain.

Table 1. Recommendation system classification

Method	System	Domain	Similarity. M
Content-based Filtering	Let's browse	Web recommendation	Cosine.S
Content-based filtering	Labo Ur	Document Recommendation	NB Classifier
Collaborate filtering	MovieLens	Movie recommendation	Cosine. S
Collaborate filtering	MovieLens	Music recommendation	Pearson r correlation
Collaborate filtering	Re: agent	Email filtering	Nearest. N
Hybrid	News weeder	News recommendation	Cosine. S

The similarity measurements used in the mentioned process face some weaknesses. Moreover, in cosine similarity, the possibility to apply measurement is only possible when the clusters are separated from the original data [14]. Pearson r correlation contains the problem of scalability, which has low speed when the number of users and items increase and similarly face the scarcity issue based on item types [15].

Lee et al. divided the shopping process into four categories. Product sentiment, click information, basket information and finally purchased items records [16]. Similarly, Cho and Kim et al. used the same process with three categories to extract information from the user profile. In the first step, they create a user profile by using the user click information and basket and purchased items [17].

Another essential part of the recommendation is the rating of products. Rating in most of the online shopping websites is based on starts between one to five scores which allows users to share their opinion with others. Product review is also another part of the recommendation that allows to user describe their experience about purchasing an item and review it in the website.

3. System Architecture

The designed recommendation system provides content-based filtering recommended items based on user previous activities and click information. Figure 1 presents the developed system architecture that contains five sections. The first section is representing data architecture that shows the user click information and purchased items in the shopping website. Based on this dataset, XGBoost classifier used to predict the model for the recommendation.

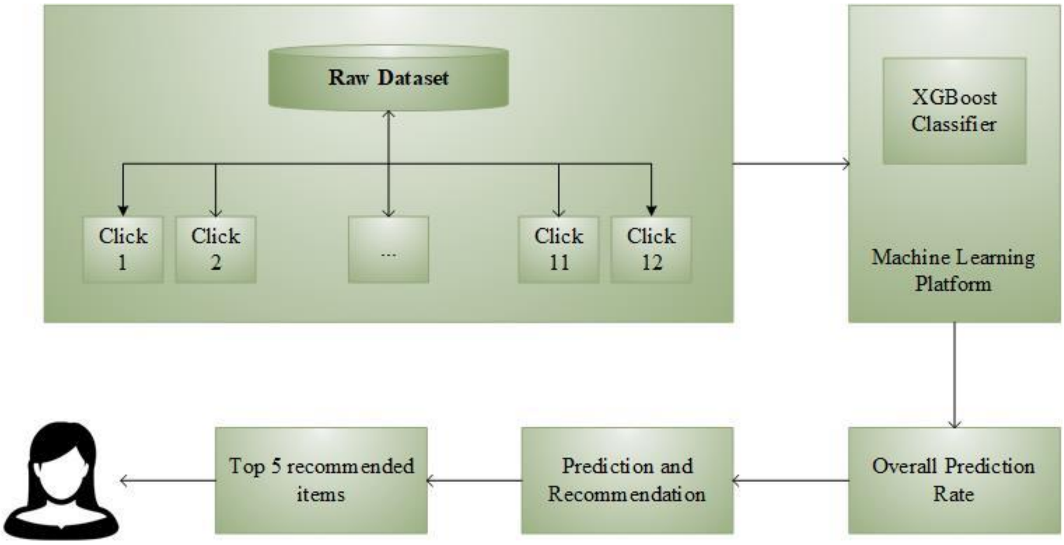

Figure 1. Overall System Architecture

3.1 Recommendation System

Recommendation system is a platform to filter the information and predict the user interest and rate of items. Recommendation applied in a vast area and commonly known in amazon online shopping website or Facebook, Twitter, etc. Recommender system like other systems such as knowledge-based systems contains different parts, and is divided into two categories, content-based filtering and collaborate filtering. Product recommendation system designed to generate opinions for various items and information for users to make the purchasing process more comfortable. There are three ways to create a product recommendation system, user-product relationship, user-user relationship and product-product relationship. Relationship between user and product is based on individual product preferences. The user-user relationship is based on the same situation people, like: same age, same interest, etc. product-product relationship is based on the same supplementary products, like a pen and pencil. Machine learning recommendation techniques are used for the data filtering process.

3.1.1 Content-based Filtering

CBF is extracting the user measures such as user clicks, purchased items, visited pages, the times passed in a website, product categories, etc. Based on this information, customer profile made and this information used to recommend items in this area.

3.1.2 Collaborate Filtering

CF is extracting information based on user behavior and priority and predict the similarity between user with other users. Eg., if the user one order strawberry and user two also order strawberry, then the system recognizes that these users have the same choices and recommend a couple of similar items to them. Figure 2 shows the system architecture flowchart. The first section shows the data pre-processing step, the second section shows the prediction process, and the last section shows the product recommendation result.

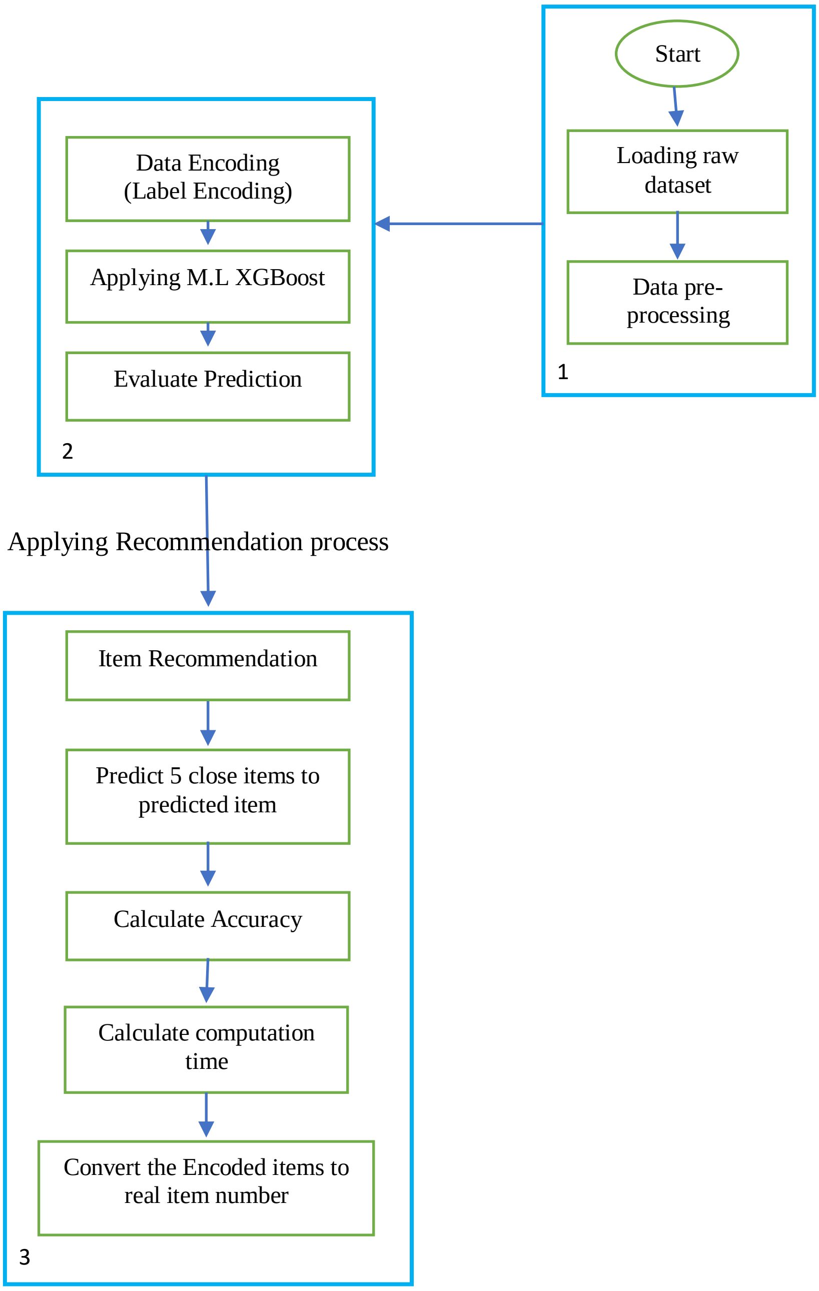
Start
Loading raw dataset
Data pre-processing
1
Data Encoding (Label Encoding)
Applying M.L XGBoost
Evaluate Prediction
2
Applying Recommendation process
Item Recommendation
Predict 5 close items to predicted item
Calculate Accuracy
Calculate computation time
Convert the Encoded items to real item number
3

Figure 2. System Architecture Flowchart

4. Implementation

This section presents the development environment in detail.

4.1 Experimental Setup

Experimental setup shows in Table 2. All experiments and results of the system are carried out using Intel(R) Core (TM) i7-8700 CPU @3.20GHz 3.19 GHz processor with 32 GB memory. XGBoost machine learning algorithm used for recommendation system. Similarly, the library and framework used in the proposed system are Jupyter notebooks. The programming language used in the designing of this system is WinPython–3.6.2.

Table 2. Development Environment of Proposed Recommendation System

Component	Description
Programming language	WinPython 3.6.2
Operating system	Windows 10 64bit
Browser	Google Chrome, opera
Library and framework	Jupyter notebook
CPU	Intel(R) Core (TM) i7-8700 CPU @3.20GHz 3.19 GHz
Memory	32GB
Machine learning algorithm	XGBoost
Distribution Modeling Algorithm	CoreNLP's MaxEnt
Recommendation Method	Content-based Filtering

4.2 Dataset

Dataset used in the proposed recommendation system collected from Jeju online shopping mall records. The total number of dataset records is 10000. The maximum number of click information for each user is 12 clicks, and the minimum number of clicks is 4. To collect the information for recommendation clicks less than four doesn't have enough information for a recommendation.

Table 3. Dataset information

Dataset	Information
Total number of records	10000
Maximum number of clicks	12
Minimum number of clicks	4
Training data	80%
Test data	20%

4.3 Evaluation Metrics

In this section two evaluation metrics defined to show the proposed system functionality. The prediction rate defines as $P^{\wedge} = (x^{\wedge}_{1}, \ldots, x^{\wedge}_{n})$ and true rating defines as $P^{\wedge} = (x^{\wedge}_{1}, \ldots, x^{\wedge}_{n})$.

1. Mean average error: the applied mean average error is to extract the prediction rate from actual rate.

$$D\ (P,\ P^{\wedge}) = ||P-P^{\wedge}||_{1}/n. \qquad (1)$$

2. Mean zero-one test error: the applied mean zero-one error is to show incorrect prediction results by showing one.

$$D\ (P, P^{\wedge}) = |\ \{L{:}x_L \neq x^{\wedge}{}_L\}\ |\ /\ n \qquad (2)$$

Figure 3 shows the products which have more shopping records based on user clicks. In this figure highest 15 purchased product selected.

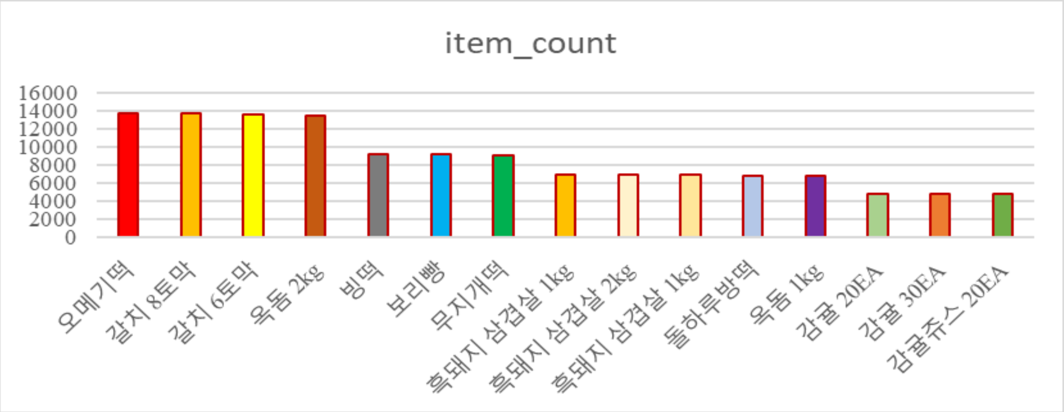

Figure 3. Status of most purchased items

4.4 Comparison and Results

In this section, results are evaluated. As it shows in Table 4, XHBoost classifier used for product recommendation in the purposed system. Based on the selected algorithm, four other machine learning algorithms compared to the recommendation. XGBoost with an accuracy of 89.6% has the highest accuracy compare with Random Forest with the accuracy of 83.24%, Support Vector Machine (SVM) with the accuracy of 78%, K-nearest neighbour with the accuracy of 21% and logistic regression with the accuracy of 14%.

Table 4. Result Comparison of Machine Learning Algorithms

Algorithm	Accuracy
XGBoost	89.6%
Random Forest	83.24%
SVM	78%
Knearest	21%
Logistic Regression	14%

4.4.1 Recognition Rate and Computation Time

Table 5 shows the recognition rate and computation time of the purposed system. The maximum recognition rate is 88% and minimum computation time is 0.135.

Table 5. Recognition Rate and Computation Time Results

Recognition Rate	Computation Time
0.86 %	0.2062
0.88 %	0.179
0.86 %	0.208
0.88 %	0.194
0.87 %	0.195
0.87 %	0.155
0.86 %	0.221
0.88 %	0.144
0.86 %	0.171
0.86 %	0.135

4.4.2 Product Recommendation

In the next step, after prediction process system recommend the five nearest items to predicted item. Table 6 shows the recommendation result. First column presents the predicted values. Second column presents the text values. Third column present the recommended items. Comparing the recommended items with test values shows if the recommended item appears in user click list. To do the processing of recommendation first we used label encoding process to encode our collected data to make it suitable to use in the recommendation system.

Table 6. Five Nearest Recommendation Results

NO.	P_Values	T_Values	Nearest items	True - False
0	766	766	758, 757, 775, 776, 766	True
1	790	789	799, 781, 800, 790, 789	True
2	791	791	783, 782, 800, 781, 791	True
3	758	759	749, 767, 768, 758, 759	True
4	73	144	81, 65, 82, 64, 63	False
5	791	791	783, 782, 800, 781, 791	True

After recommending items to the user, for further processing label encoded items to convert to real item number. Table 7 shows the result of converted item numbers to the real item number.

Table 7. Converting Recommendation Result to Real Item Number

NO.	P_Values	T_Values	Nearest items	Real item Number
0	766	766	776, 766	469979037, 476117141
1	790	789	790, 789	481211266, 481209639
2	791	791	781, 791	480614517, 478761179
3	758	759	749, 767	469978713, 469979180
4	73	144	81, 65	256900419, 256882618
5	791	791	781, 791	480614517, 478761179

4.4.3 XGBoost Classification Error Rate

Figure 4 presents the XGBoost classifier classification error rate. The number of epochs is shown in the down part of the figure, and the left part shows the error rate. In comparison between train and test set error rate is quite high. Figure 5 presents the XGBoost log loss rate in train and test dataset.

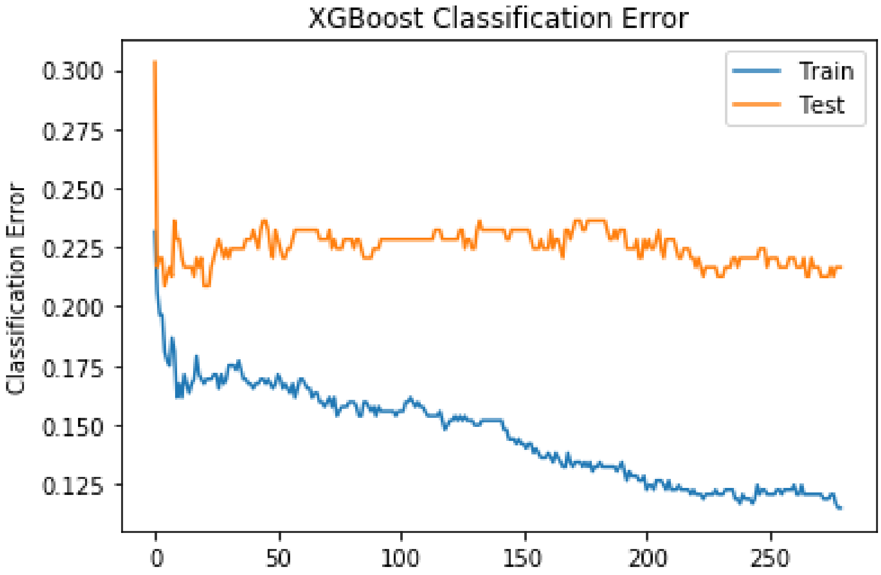

Figure 4. XGBoost Classification Error Graph

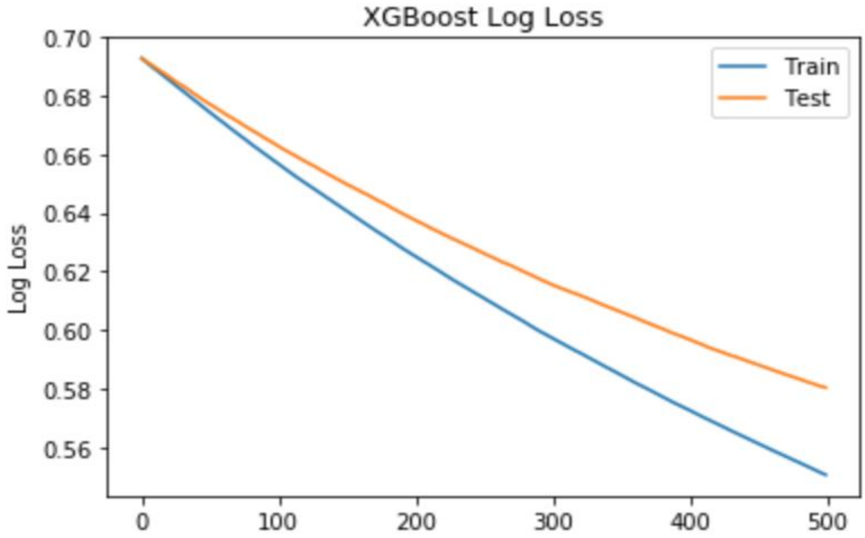

Figure 5. XGBoost Log Loss

Conclusion

Content-based filtering applied successfully for product recommendation on the collected items from Jeju online shopping mall. The system requires user click information and user profile to be available for recommending items to the user. In this process, XGBoost classifier used for classification and prediction process. Experiment and result section present the accuracy and recognition rate of proposed system. Compared with other algorithms the proposed process has higher output in the recommendation system.

Acknowledgements

This work was supported by Korea Institute for Advancement of Technology (KIAT) grant funded by the Korea Government (MOTIE) (N0002327, The Establishment Project of Industry-University Fusion District).

References

[1] Linyuan Lü, Matúš Medo, Chi Ho Yeung, Yi-Cheng Zhang, Zi-Ke Zhang,Tao Zhoua, Recommender systems, Elsevier Journal – Physics Reports 519 (2012) 1–49

[2] Sandra Garcia Esparza, Michael P. O'Mahony, Barry Smyth, Mining the real-time web: A novel approach to product recommendation, Elsevier Journal - Knowledge-Based Systems 29 (2012) 3–11.

[3] "Online Shopping touched new heights in India in 2012". Hindustan Times. 31 December 2012. Retrieved 31 December 2012.

[4] Belkin, N. J., and Croft, W. B. Information filtering and information retrieval: two sides of the same coin. Communications of the ACM, 35, 12, 1992, 29–38.

[5] Lang, K. NewsWeeder: learning to filter netnews. In Proceedings of the Twelfth International Conference on, Machine Learning, 1995.

[6] Mooney, R. J., and Roy, L. Content-based book recommending using learning for text categorization. In ACM SIGIR '99 Workshop on Recommender Systems: Algorithms and Evaluation, Berkeley, CA, 1999.

[7] Pazzani, M., and Billsus, D. Learning and revising user profile: the identification of interesting web sites. Machine Learning, 27, 3, 1997, 313–331.

[8] Blackwell, R. D., Paul, W. M., & James, F. E. (2000). Consumer behavior. Cincinnati, OH: South-Western College Publishing.

[9] Changchien, S. W., Lee, C.-F., & Hsu, Y.-J. (2004). Online personalized sales promotion in electronic commerce. Expert Systems with Applications, 27(1), 35–52.

[10] Balabanovic, M., & Shoham, Y. (1997). Fab: Content-based, collaborative recommendation. Communication of the ACM, 40(3), 66–72.

[11] Cunningham, P., Bergmann, R., Schmitt, S., Traphoner, R., Breen, S., & Smyth, B. (2001). WebSell: Intelligent sales assistants for the world wide web. Kunstliche Intelligenz (KI), 15(1), 28–32.

[12] Cho, Y. H., Kim, J. K., & Kim, S. H. (2002). A personalized recommender system based on web usage mining and decision tree induction. Expert Systems with Applications, 23(3), 329–342.

[13] Montaner, M., Lopez, B., & Rosa, J. L. D. (2003). A taxonomy of recommender agents on the internet. Artificial Intelligence Review, 19, 285–330.

[14] Schalkoff, R. (1992). Pattern recognition: Statistical, structural and neural approaches. New York: Wiley.

[15] Billsus, D., & Pazzani, M. J. (1998). Learning collaborative information filters. International Conference on Machine Learning, 15, 46–54.

[16] Lee, C. H., Kim, Y. H., & Rhee, P. K. (2001). Web personalization expert with combining collaborative filtering and association rule mining technique. Journal of Expert Systems with Application, 21(3), 131–137.

[17] Cho, Y. H., & Kim, J. K. (2004). Application of Web usage mining and product taxonomy to collaborative recommendations e-commerce. Expert Systems with Applications, 26(2), 233–246.

Authors

Zeinab Shahbazi received her B.S. in software engineering from Pooyesh University, IRAN. In March 2017, she moved to Republic of Korea for M.S studies and started working in internet laboratory, Chonbuk National University (CBNU). After completing her master in 2018, she moved to Jeju-do in March 2019 and started working as a Ph.D. research fellow in Machine Learning Laboratory (MLL), Jeju National University. Research interests include artificial intelligence and machine learning, natural language processing, deep learning and data mining.

Yung Cheol Byun received his B.S. from Jeju National University, Korea in 1993, M.S and Ph.D degrees from Yonsei University in 1995 and 2001. He worked as a special lecturer in SAMSUNG Electronics in 2000 and 2001. From 2001 to 2003, he was a senior researcher of Electronics and Telecommunications Research Institute and he promoted to join Jeju National University as an assistant professor in 2003, where he is currently a professor of Department of Computer Engineering. From 2012 to 2014, he had research activities at University of Florida as a visiting professor. His research interests include the areas of pattern recognition & image processing, artificial intelligence & machine learning, security based on pattern recognition, home network and ubiquitous computing, u-Healthcare and RFID & IoT middleware system (Corresponding Author).