International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:7, No:10, 2013

# Recommender Systems Using Ensemble Techniques

Yeonjeong Lee, Kyoung-jae Kim, Youngtae Kim

***Abstract*—**This study proposes a novel recommender system that uses data mining and multi-model ensemble techniques to enhance the recommendation performance through reflecting the precise user's preference. The proposed model consists of two steps. In the first step, this study uses logistic regression, decision trees, and artificial neural networks to predict customers who have high likelihood to purchase products in each product group. Then, this study combines the results of each predictor using the multi-model ensemble techniques such as bagging and bumping. In the second step, this study uses the market basket analysis to extract association rules for co-purchased products. Finally, the system selects customers who have high likelihood to purchase products in each product group and recommends proper products from same or different product groups to them through above two steps. We test the usability of the proposed system by using prototype and real-world transaction and profile data. In addition, we survey about user satisfaction for the recommended product list from the proposed system and the randomly selected product lists. The results also show that the proposed system may be useful in real-world online shopping store.

***Keywords*—**Product recommender system, Ensemble technique, Association rules, Decision tree, Artificial neural networks.

## I. INTRODUCTION

RECENT explosive increase of electronic commerce provides many advantageous purchase opportunities to customers. In this situation, customers who do not have enough knowledge about their purchases may accept product recommendations. Product recommender systems automatically reflect user's preference and provide recommendation list to the users. Thus, product recommender system in online shopping store has been known as one of the most popular tools for one-to-one marketing. However, recommender systems, which do not properly reflect user's preference, cause user's disappointment and waste of time.

In this study, we propose a novel recommender system, which uses data mining and multi-model ensemble techniques to enhance the recommendation performance through reflecting the precise user's preference.

This paper consists of five sections. Section II describes related studies about recommender systems. In Section III, we propose a novel recommender system using ensemble techniques. Section IV introduces experimental data and design of experiments. In Section IV, we also present experimental results. Section V discusses the findings and limitations of our study.

Yeonjeong Lee and Youngtae Kim are with Graduate School of Management, Dongguk University-Seoul, South Korea.

Kyoung-jae Kim is with Business School, Dongguk University-Seoul, South Korea (corresponding author; phone: +82-2-2260-3324; fax: +82-2-2260-3684; e-mail: kjkim@ dongguk.edu).

## II. PRIOR RESEARCH

The most important part of recommender system is recommendation technique and the most representative recommendation techniques are content-based filtering and collaborative filtering. Content-based filtering analyzes item descriptions to establish items that are of particular interest to the user [1]. The advantage of this technique is direct and simple because it only uses item descriptions [2]. On the other hand, collaborative filtering (CF) matches people with similar tastes and recommends products on this basis. The motivation for this technique stems from the idea that people usually accept recommendation from someone with similar taste. CF work by gathering user feedback in the form of ratings for items and exploiting similarities in rating behavior amongst several users in determining how to recommend an item. CF can be divided into neighborhood-based and model based approaches [3].

CF generally shows higher recommendation performance than other techniques when the users give homogeneous ratings to the products. In addition, it also provides better recommendation performance than other techniques when the rating data are enough to validate [4], [5]. Regardless of its success in many prior applications, it has two serious limitations, namely sparsity and scalability problem [6]–[19]. The sparsity problem occurs when available data are insufficient for identifying similar users. It is a major issue that limits the quality of recommendations and the applicability of collaborative filtering [20]. In addition, CF also suffers serious scalability problems, as the numbers of users and items grow.

## III. RECOMMENDER SYSTEMS USING ENSEMBLE TECHNIQUES

In this study, we propose a novel recommender system, which uses data mining and multi-model ensemble techniques to enhance the recommendation performance through reflecting the precise user's preference.

The first step predicts customers who have high likelihood to purchase products in the online shopping store. In this step, we first use logistic regression, decision trees, and artificial neural networks to predict customers who have high likelihood to purchase products in each product group. Then, we combine the results of each predictor using the multi-model ensemble techniques such as bagging and bumping. Leo Breiman initially proposes bagging [21], [22]. Bagging is the abbreviation of "Bootstrap Aggregation" and it composite outputs from several machine learning techniques for raising the performance and stability of prediction or classification. This technique is special form of the averaging method. Bumping is the abbreviation of "Bootstrap Umbrella of Model Parameter", and it only considers the model, which has the lowest error value [23]–[25]. The results show that bumping outperforms bagging and the other predictors except for "Poster" product group. For the "Poster" product group, artificial neural network model performs better than the other models.

In the second step, we use the market basket analysis to extract association rules for co-purchased products. We will extract some association rules according to values of Lift, Support, and Confidence measure.

## IV. Experiments

### A. Experimental Data

In order to validate the usefulness of our proposed recommendation algorithm, we adopt 'empirical validation' that is based on a real-world dataset. The research data are collected from the real-world online shopping store, which deals products from famous art galleries and museums in Korea. The data initially contain 5759 transaction data, but finally remain 3167 transaction data after deletion of null data.

TABLE I
Selected Input Variables and Description After Preprocessing

| Variables | Description | Value |
|---|---|---|
| id | Customer ID | Text |
| age | Customer Age | Integer |
| gender | Customer Gender | 1:Female / 2:Male |
| job1 ~ job13 | Artists, Office Jobs, Educational Jobs, House Wives, Unemployed, Administrative Jobs, Marketers, Others, Special Jobs, Sales, Service Jobs, Students, Technicians | |
| zip1 ~ zip38 | Dobong-Gu, Dongdaemun-Gu, Dongjak-Gu, Youngdeungpo-Gu, Chung-Gu, Chungrang-Gu, Jongro-Gu, Kangbuk-Gu, Kangdong-Gu, Kangnam-Gu, Kangseo-Su, Guro-Gu, Kwangjin-Gu, Kwanak-Gu, Mapo-Gu, Rowon-Gu, Sungbuk-Gu, Seocho-Gu, Seodaemun-Gu, Sungdong-Gu, Songpa-Gu, Yangchun-Gu, Yongsan-Gu, Eunpyung-Gu, Kangwon-Do, Kyunggi-Do, Keungsangnam-Do, Keungsangbuk-Do, Jeju-Do, Jeunranam-Do, Jeunrabuk-Do, Chungcheungnam-Do, Chungcheungbuk-Do, Kwangju-Si, Daegu-Si, Daejun-Si, Busan-Si, Ulsan-Si, Incheun-Si | 0: OK / 1: N/A |
| F | Customers of Fashion Product Group | |
| HD | Customers of Home Decoration Product Group | |
| O | Customers of Office Supplies Product Group | |
| P | Customers of Poster Product Group | |

In this study, we transform the categorical variables into dummy variables and exclude outlier data. Finally, we have 8 variables. Table II shows finally selected variables.

We also classify the products into 5 categories including "Fashion", "Home Deco (Home Decoration)", "Office (Office Supplies)", "Poster", "Sculptures" according to the characteristics of products. Sculptures product group, however, has only two transaction data, so we excluded this group from our research. Table II presents the product taxonomy of the research data.

TABLE II
Product Taxonomy of the Research Data

| Tier 1 Product Group | Codes for Tier 2 Product Group | Products |
|---|---|---|
| Fashion | F_keyholder | Key Holders |
| | F_ties | Ties |
| | F_tshirt | T-shirts |
| | F_umbrella | Umbrellas |
| | F_bag | Bags |
| Home Deco. | HD_bath | Bathroom Wares |
| | HD_interior | Interior Properties |
| | HD_magnet | Magnets |
| | HD_tblware | Table Wares |
| Office | O_calendar | Calendars |
| | O_card | Cards |
| | O_cube | Memo Pads |
| | O_diary | Diaries |
| | O_hobby | Hobby Items |
| | O_letter | Letterhead Stationaries |
| | O_note | Notes |
| | O_etc | Other Items |
| Poster | P_picture | Posters |

### B. Experimental Design and Results

We perform two steps of experiments. In first step, we first perform experiments using several classification algorithms including logistic regression, decision tree, and artificial neural networks to predict customers who have high likelihood to purchase products in each product group. We perform above data mining techniques using SAS E-Miner software. In this study, we split datasets into two sets as modeling and validation sets for the logistic regression and decision trees. We also split datasets into three sets as training, test, and validation sets for the artificial neural network model. The validation dataset is equal for the all experiments. Data sets for classification models and experimental purposes are provided in Table III.

TABLE III
Data Sets for Models

| *Classification Model* | *Total* | *Training Set* | *Test Set* | *Validation Set* |
|---|---|---|---|---|
| Logistic Regression | 3167 | 2533 | N/A | 634 |
| Decision Trees | 3167 | 2533 | N/A | 634 |
| Artificial Neural Networks | 3167 | 1901 | 632 | 634 |

Then, we combine the results of each predictor using the multi-model ensemble techniques such as bagging and bumping. The classification results of data mining techniques are presented in Table IV.

TABLE IV
Classification Results for the Each Product Group

| *Classification Models* | *Fashion* | *Home Deco.* | *Office* | *Poster* |
|---|---|---|---|---|
| Logistic Regression | 64.67% | 87.38% | 56.94% | 93.06% |
| Decision Trees | 63.72% | 87.70% | 58.04% | 93.06% |
| Artificial Neural Networks | 63.09% | 87.22% | 57.89% | 93.20% |
| Bagging | 63.81% | 87.40% | 57.63% | 93.10% |
| Bumping | 65.14% | 88.16% | 58.13% | 93.06% |

The results show that bumping outperforms bagging and the other predictors except for "Poster" product group. For the "Poster" product group, artificial neural network model performs better than the other models.

In the second step, we use the market basket analysis to extract association rules for co-purchased products. We extract some association rules according to values of Lift, Support, and Confidence measure. We set the minimum transaction frequency to support associations as 5%, maximum number of items in an association as 4, and minimum confidence for rule generation as 10%. We can extract thirty-one association rules according to values of Lift, Support, and Confidence measure. This study also excludes the extracted association rules below 1 of lift value. We finally get fifteen association rules by excluding duplicate rules. Table V shows the extracted association rules in this study.

TABLE V
EXTRACTED ASSOCIATION RULES

| | *Lift* | *Support(%)* | *confidence(%)* | *Extracted Association Rules*[*] |
|---|---|---|---|---|
| 1 | 2.77 | 1.84 | 19.35 | O_cube ==> HD_magnet |
| 2 | 2.58 | 2.07 | 18 | O_etc ==> HD_magnet |
| 3 | 2.54 | 1.84 | 16 | O_etc ==> O_hobby |
| 4 | 2.51 | 3.6 | 23.86 | O_note ==> O_cube |
| 5 | 1.94 | 1.61 | 14 | O_etc ==> O_card |
| 6 | 1.9 | 3.3 | 28.67 | O_etc ==> O_note |
| 7 | 1.75 | 2.68 | 16.67 | O_diary ==> O_cube |
| 8 | 1.7 | 1.61 | 10.66 | O_note ==> O_hobby |
| 9 | 1.69 | 1.84 | 12.18 | O_note ==> O_card |
| 10 | 1.67 | 4.06 | 26.9 | O_note ==> O_diary |

* "O" denotes Office Supplies, "HD" denotes Home Decoration

Among the fifteen association rules, eleven rules contain association between products in "Office Supplies" product group, one rules include the association between "Office Supplies" and "Fashion" product groups, and other three rules contain association between "Office Supplies" and "Home Decoration" product groups. Finally, the proposed product recommender systems provides list of recommendations to the proper customers.

### C. Model Validation

We test the usability of the proposed system by using prototype and real-world transaction and profile data. For this end, we construct the prototype system by using the ASP, Java Script and Microsoft Access. In addition, we survey user satisfaction for the recommended product list from the proposed system and the randomly selected product lists. The participants for the survey are 173 persons who use MSN Messenger, Daum Café, and P2P services. We evaluate the user satisfaction using five-scale Likert measure. This study also performs "Paired Sample T-test" for the results of the survey. Table VI presents statistical test results for the survey of user satisfaction.

TABLE VI
STATISTICAL TEST RESULTS FOR THE SURVEY OF USER SATISFACTION

| | *T-value* | *Degree of Freedom* | *Significance* |
|---|---|---|---|
| Proposed Recommender System – Random Selection | 7.972 | 173 | 0.000 |

The results show that the proposed model outperforms the random selection model with 1% statistical significance level. It means that the users satisfied the recommended product list significantly. The results also show that the proposed system may be useful in real-world online shopping store.

## V. CONCLUSIONS

In this study, we proposed a novel recommender system that uses data mining and multi-model ensemble techniques to enhance the recommendation performance through reflecting the precise user's preference. This study tested the usability of the proposed system by using prototype and real-world transaction and profile data. The results showed that the proposed system might be useful in real-world online shopping store.

Although this study got better recommendation performance than random selection model, the performance might be raised if we incorporated the more sophisticated data mining models. In addition, although this study showed statistically significant difference between the proposed model and random selection model, we might get more reliable results than this if we collected sufficient samples.

## ACKNOWLEDGMENT

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2010-0025689).

## REFERENCES

[1] M. J. Pazzani, and D. Billsus, "Content-Based Recommendation Systems," *Lecture Notes in Computer Science*, vol. 4321, pp. 325–341, 2007.

[2] K. –L. Wu, C. C. Aggarwal, and P. S. Yu, "Personalization with dynamic profiler," in *Proceedings of the Third International Workshop on Advanced Issues of E-commerce and Web-based Information Systems*, 2001, pp. 12–20.

[3] J. Breese, D. Heckerman, and C. Kadie, "Empirical analysis of predictive algorithms for collaborative filtering," in *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence*, San Francisco, CA, 1998, pp. 43–52.

[4] J. Konstan, B. Miller, D. Maltz, J. Herlocker, L. Gordon, and J. Riedl, "GroupLens: Applying Collaborative Filtering to Usenet News," *Communication of the ACM*, vol. 40, pp. 77–87, 1997.

[5] M. J. Pazzani, "A framework for collaborative, content-based and demographic filtering," *Artificial Intelligence Review*, vol. 13, no. 5-6, pp. 393–408, 1999.

[6] B. M. Sarwar, G. Karypis, J. A. Konstan, and J. Riedl, "Analysis of recommendation algorithms for e-commerce," in *Proceedings of Conference on ACM Electronic Commerce*, 2000, pp. 158–167.

[7] J. K. Kim, Y. H. Cho, W. J. Kim, J. R. Kim, and J. H. Suh, "A personalized recommendation procedure for Internet shopping support," *Electronic Commerce Research and Applications*, vol. 1, pp. 301–313, 2002.

[8] J. K. Kim, J. H. Suh, D. H. Ahn, and Y. H. Cho, "A personalized recommendation methodology based on collaborative filtering," *Journal of Intelligence and Information Systems*, vol. 8, no. 2, pp. 139–157, 2002.

[9] Y. H. Cho, J. K. Kim, and S. H. Kim, "A personalized recommender system based on Web usage mining and decision tree induction," *Expert Systems with Applications*, vol. 23, pp. 329–342, 2002.
[10] J. K. Kim, D. H. Ahn, and Y. H. Cho, "Development of a personalized recommendation procedure based on data mining techniques for internet shopping malls," *Journal of Intelligence and Information Systems*, vol. 9, no.3, pp. 177–191, 2003.
[11] J. W. Kim, S. J. Bae, and H. J. Lee, "Sparsity Effect on Collaborative Filtering-based Personalized Recommendation," *Asia Pacific Journal of Information Systems*, vol.14, no.2, pp. 131~149, 2004.
[12] Y. H. Cho, and J. K. Kim, "Application of Web usage mining and product taxonomy to collaborative recommendations in e-commerce," *Expert Systems with Applications*, vol. 26, pp. 233–246, 2004.
[13] Y. H. Cho, S. K. Park, D. H. Ahn, and J. K. Kim, "Collaborative Recommendations using Adjusted Product Hierarchy: Methodology and Evaluation," *Journal of the Korean Operations Research and Management Science Society*, vol. 29, no. 2, pp. 59–75, 2004.
[14] J. K. Kim, D. H. Ahn, and Y. H. Cho, "A Personalized Recommender System, WebCF-PT: A Collaborative Filtering using Web Mining and Product Taxonomy," *Asia Pacific Journal of Information Systems*, vol. 15, no. 1, pp. 63–79, 2005.
[15] D. Kim, and B. -J. Yum, "Collaborative filtering based on iterative principal component analysis," *Expert Systems with Applications*, vol. 28, no. 4, pp. 823~830, 2005.
[16] G. Adomavicius, and A. Tuzhilin, "Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 6, pp. 734–749, 2005.
[17] K. Kim, and. B. Kim, "Product Recommender System for Online Shopping Malls using Data Mining Techniques," *Journal of Intelligence and Information Systems*, vol.11, no.1, pp. 191~205, 2005.
[18] H. C. Ahn, I. Han, and K. Kim, "The Product Recommender System Combining Association Rules and Classification Models: The Case of G Internet Shopping Mall," *Information Systems Review*, vol.8, no.1, pp. 181~201, 2006.
[19] K. Kim, and H. Ahn, "Collaborative filtering with a user-item matrix reduction technique for recommender systems," *International Journal of Electronic Commerce*, vol. 16, no. 1, pp. 107–128, 2011.
[20] M. Papagelis, D. Plexousakis, and T. Kutsuras, "Alleviating the sparsity problem of collaborative filtering using trust inferences," in *iTrust*, Springer-Verlag, 2005, pp. 224–239.
[21] L. Breiman, "Heuristics of instability in model selection," in *Technical Report*, Statistics Department, University of California at Berkeley, 1994.
[22] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.
[23] R. Tibshirani, and K. Knight, "Model search and inference by bootstrap "bumping"," in *Technical Report*, University of Toronto, 1995.
[24] T. Heskes, "Balancing between bagging and bumping," in *Advances in Neural Information Processing Systems*, Cambridge, MIT Press, 1996, pp. 466–472.
[25] Y. Lee, and S. Kwak, "A study on training ensembles of neural networks: a case of stock price prediction," *Journal of Intelligence and Information Systems*, vol. 5, no. 1, pp. 95–101, 1999.