FUZZY PERSONALIZED SCORING MODEL FOR RECOMMENDATION SYSTEM

Chao-Lung Yang[1], Shang-Che Hsu[1], Kai-Lung Hua[2], Wen-Huang Cheng[3]*

[1]Department of Industrial Management
National Taiwan University of Science and Technology
[2]Department of Computer Science and Information Engineering
National Taiwan University of Science and Technology
[3]Department of Electronics Engineering
National Chiao Tung University

ABSTRACT

In this research, we aim to propose a data preprocessing framework particularly for financial sector to generate the rating data as input to the collaborative system. First, clustering technique is applied to cluster all users based on their demographic information which might be able to differentiate the customers' background. Then, for each customer group, the importance of demographic characteristics which are highly associated with financial products purchasing are analyzed by the proposed fuzzy integral technique. The importance scores across items and customers are generated either on customer groups and individuals. The analysis shows the proposed method is able to differentiate customers based on their demographic and purchasing behaviors. Also, the generated rating matrix can be directly used for collaborative filtering model.

Index Terms— Recommendation System, Fuzzy Integral, Customer Segmentation

1. INTRODUCTION

Recommendation system, in general, can be considered as software which provides the suggestion to link the user's preference or interest on context through information filtering and decision support system based on the collected data [1]. Recommendation techniques can be classified into three main types: collaborative filtering, content-based filtering and hybrid approach according to variation of data usage [2]. As one of the most used recommendation systems, collaborative filtering (CF) aims to compute the similarity among ratings across items and users to suggest the "right" items to customers whose rating preference is "filter-out" by mutual comparison among users [3]. CF method has been applied in traditional financial sectors such as insurance riders, real estate, venture capital, and stock market, and also new developed FinTech startups for peer-to-peer lending or credit evaluation.

As a profound data owner, the financial bank can utilize the components of salary i.e. saving, investment and expenditure to develop the recommendation system [4]. David Zibriczky's review paper also listed multiple applications which have been applied in CF in financial sector [5]. For example, online banking, peer-to-peer lending, customized insurance, real estate, personalized stock recommendation, portfolio management have used CF as the recommendation system. However, based on author's best knowledge, few research work addresses how to convert the immense financial data to data matrix for CF. How to correlated the customer demographic information with their purchasing behaviors is still an open question.

Yang and Phuong proposed a data analysis framework to correlate the two different datasets by combining clustering and classification methods under multiple objective non-dominated sorting genetic algorithm [6]. The framework seems applicable for searching the correlation between demographic and product purchasing datasets. The searching result can be used to identify the features in demographic dataset which are more correlated to the optimal result of classification on purchasing behavior. However, for recommendation system usage, the scoring/rating matrix is still needed which cannot be obtained by Yang and Phuong's work.

In addition to the issue of data preprocessing framework, how to score or how to weight the information of financial data is a key research topic in this area. Based on David Zibriczky's review, fuzzy technique is widely applied in many financial applications such as fuzzy-based clustering for stock recommendations, fuzzy-based expert systems, and fuzzy-based portfolio recommendations [5]. Essentially, the fuzzy scoring on product preference and likelihood can be formed by fuzzy measurement and weighting [7]. Miao et al. utilized fuzzy mapping and neural network to compute recommendation list based on personal preference, common

preference, and expert's domain knowledge [8]. Cao and Li also used fuzzy-based system to transfer the users' qualitative needs into triangular fuzzy scoring to find the ideal recommended products [9]. Cao and Li's work provided questionnaire interface to gather the qualitative needs of users based on users' subjective decision. Obviously, previous research works more focus on analyzing the existing user's subjective preference. Therefore, in this research, we aim to attack the data preprocessing issue of converting the demographic data and purchasing records, very common in financial sector, to data matrix needed for CF computation.

2. METHODOLOGY

In this research, a data framework of generating score of user against product is proposed based on the importance of demographic features and product purchase records. The framework is illustrated in Fig. 1. The two sets of data are needed: user demographic data and product purchase data. First, clustering method is applied to perform customer segmentation based on user demographic information. Then, for each group of user, the importance of demographic feature against product is evaluated. Fuzzy integral method plays an importance role on aggregating the importance across all users in the group. Once the group-level score is computed, the individual-level score is expended by considering the individual personal purchasing record.

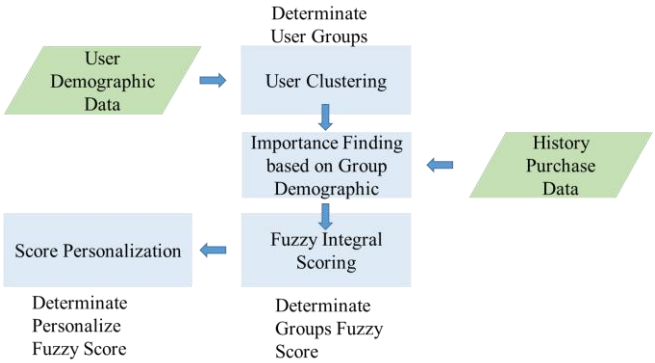

Fig. 1: The process of the proposed fuzzy personalized scoring model.

2.1. Customers Clustering

Assume there are n users and m columns of demographic information such as gender, age, employment, salary range, risk allowance level, credit score, and so on. So, the vector $\boldsymbol{u_i} = (u_{i1}, \dots, u_{im})$ represents demographic information of customer *i*. Thus we can have $\boldsymbol{U}$ matrix to contain Know-Your-Customer (KYC) data:

$$\boldsymbol{U} = \begin{bmatrix} \boldsymbol{u_{11}} & \boldsymbol{u_{12}} & \cdots & \boldsymbol{u_{1m}} \\ \boldsymbol{u_{21}} & \boldsymbol{u_{22}} & \cdots & \boldsymbol{u_{2m}} \\ \vdots & \ddots & \ddots & \vdots \\ \boldsymbol{u_{n1}} & \boldsymbol{u_{n2}} & \cdots & \boldsymbol{u_{nm}} \end{bmatrix} \quad (1)$$

Where u_{ij} identifies the value of *j* demographic information of the *i* customer.

In the proposed framework, first, the customer segmentation is conducted by applying clustering technique. Note that mixed attributes: numeric and categorical data prevail on financial demographic data. Therefore, in this work, k-prototype method, one of famous mixed attribute data clustering methods, is used to cluster customers by their demographic information [10].

In order to perform the clustering task, a function $\mathbf{d}(\boldsymbol{u}, \boldsymbol{v})$ can be defined to measure the similarity among customer $\boldsymbol{u}$ and $\boldsymbol{v}$ across numeric and categorical data columns.

$$\mathbf{d}(\boldsymbol{u}, \boldsymbol{v}) = \sum_{j=1}^{m_{nu}} (\boldsymbol{u_j^{nu}} - \boldsymbol{v_j^{nu}})^2 + \boldsymbol{\gamma} \sum_{j=1}^{m_{ca}} \boldsymbol{u_j^{ca}} \; \boldsymbol{R} \; \boldsymbol{v_j^{ca}} \quad (2)$$

Where $u_{.j}^{nu}$ are all numeric columns, $u_{.j}^{ca}$ represent all categorical columns; γ is a weight for categorical attributes. *R* is a relation between $u_{.j}^{ca}$ and $v_{.j}^{ca}$ and $u_{.j}^{ca}$ *R* $v_{.j}^{ca}$ is defined as (2). The detail of the determination of γ and *R* can be seen in [10].

$$\boldsymbol{u_j^{ca}} \; \boldsymbol{R} \; \boldsymbol{v_j^{ca}} = \begin{cases} \mathbf{1}, \boldsymbol{u_j^{ca}} \neq \boldsymbol{v_j^{ca}} \\ \mathbf{0}, \boldsymbol{u_j^{ca}} = \boldsymbol{v_j^{ca}} \end{cases} \quad (3)$$

After performing customer clustering, the customer groups can be identified as g_i in $\boldsymbol{U_g}$ where g_i is the cluster indicator of user *i* as shown as (3):

$$\boldsymbol{U_g} = \begin{bmatrix} \boldsymbol{u_{11}} & \boldsymbol{u_{12}} & \cdots & \boldsymbol{u_{1m}} & \boldsymbol{g_1} \\ \boldsymbol{u_{21}} & \boldsymbol{u_{22}} & \cdots & \boldsymbol{u_{2m}} & \boldsymbol{g_2} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \boldsymbol{u_{n1}} & \boldsymbol{u_{n2}} & \cdots & \boldsymbol{u_{nm}} & \boldsymbol{g_n} \end{bmatrix} \quad (4)$$

2.2. Feature Importance Finding on Purchase Records

In order to associate the demographic information with the purchasing preference of each customer group, the data aggregation on purchasing is required. For all customers, $\boldsymbol{P}$ matrix can be generated to contain binary indicator of purchasing records on all products, assuming there are *h* products on shelf.

$$\boldsymbol{P} = \begin{bmatrix} \boldsymbol{p_{11}} & \cdots & \boldsymbol{p_{1h}} \\ \vdots & \ddots & \vdots \\ \boldsymbol{p_{n1}} & \cdots & \boldsymbol{p_{nh}} \end{bmatrix} \quad (5)$$

Where $p_{iw} = \begin{cases} 1, \text{ if user } i \text{ purchased prodcut } w \\ 0, \text{if user } i \text{ has not purchased prodcut } w \end{cases}$ and $w \in (1, \dots, h)$.

Then we can combine $\boldsymbol{U_g}$ with $\boldsymbol{P}$ to obtain a matrix which contains customer demographic data, user group, and purchasing history across all products. For each customer group *k* with n_k customers, the combined $\boldsymbol{U_{gp}^k}$ shown in (5) is used to study the influence analysis on which demographic characteristic is highly associated with the purchase behavior within group *k*.

$$\boldsymbol{U_{gp}^{k}} = \begin{bmatrix} \boldsymbol{u_{11}} \cdots \boldsymbol{u_{1m}} & \boldsymbol{p_{11}} \cdots \boldsymbol{p_{1h}} \\ \vdots \ddots \vdots & \vdots \ddots \vdots \\ \boldsymbol{u_{n_k 1}} \cdots \boldsymbol{u_{n_k m}} & \boldsymbol{p_{n_k 1}} \cdots \boldsymbol{p_{n_k h}} \end{bmatrix} \quad (6)$$

As mentioned early in the introduction section, the $\boldsymbol{U_{gp}^{k}}$ can be considered as the combination of two datasets, demographic data and purchase behavior in this case, and the data analysis framework proposed by [6] can be applied. Please note that, in [6], any classification model can be used in the framework. In order to simply the process of analysis, in this research, the Random Forest [11] is used to find importance of demographic data columns.

The equation (7) shows the $\boldsymbol{im_w^k}$ vector containing all importance scores $\boldsymbol{im_w^k}$ of demographic feature $u_{.j}$ where $j \in (1, \ldots, m)$ for product w in customer group k.

$$\boldsymbol{im_w^k} = \left[\boldsymbol{im_w^k}(\boldsymbol{u_{.1}}), \boldsymbol{im_w^k}(\boldsymbol{u_{.2}}), \cdots, \boldsymbol{im_w^k}(\boldsymbol{u_{.m}})\right] \quad (7)$$

2.3. Fuzzy Integral of Importance

We assume the $\boldsymbol{im_w^k}$ is the membership degree of product w when it was purchased. Thus, we can apply fuzzy set theory to convert the importance vector to the fuzzy set defined by [12].

$$\boldsymbol{F_w^k} = \left[\frac{\boldsymbol{im_w^k}(\boldsymbol{u_{.1}})}{\boldsymbol{u_{.1}}} + \frac{\boldsymbol{im_w^k}(\boldsymbol{u_{.2}})}{\boldsymbol{u_{.2}}} + \cdots + \frac{\boldsymbol{im_w^k}(\boldsymbol{u_{.m}})}{\boldsymbol{u_{.m}}}\right] \quad (8)$$

Then we can use fuzzy λ-measure to calculate the importance degree of union between im_w^k columns. The union can be described as (9). The λ can be obtained by solving the unions of the importance measures.

$$im_w^k(u_{.1} \cup u_{.2}) = im_w^k(u_{.1}) + im_w^k(u_{.2}) + \boldsymbol{\lambda} \cdot \boldsymbol{im_w^k}(\boldsymbol{u_{.1}}) \cdot \boldsymbol{im_w^k}(\boldsymbol{u_{.2}}) \quad (9)$$

The fuzzy integral technique is proposed to compute the influence level of demographic feature against purchasing preference of product w. For each group k, the summation of each normalized demographic features j is calculated. C^{kw} is the vector of the summations across all features c_j^{kw} shown in (10).

$$C^{kw} = \left[sum\left(\begin{bmatrix} u_{11} \\ u_{21} \\ \vdots \\ u_{k_j 1} \end{bmatrix}\right), \cdots, sum\left(\begin{bmatrix} u_{1m} \\ u_{2m} \\ \vdots \\ u_{k_j m} \end{bmatrix}\right)\right] = \left[\boldsymbol{c_1^{kw}}, \boldsymbol{c_2^{kw}}, \cdots, \boldsymbol{c_m^{kw}}\right] \quad (10)$$

Where c_i^{kw} is all summary of demographic feature j of group k for product w. In order to apply fuzzy integral, the descending sorting of all c_j^{kw} is generated.

The fuzzy integral can be defined as the equation (11) which compute the summation of importance of all features multiplying by near-by pair-wise c_j^{kw} distance. By considering the importance and the summation of features (weight), the weighted score $\boldsymbol{s_{kw}}$ of group k to product w is computed based on the fuzzy integral theory [13].

$$\boldsymbol{s_{kw}} = \sum_{i=1}^{m} \boldsymbol{im_w^k}(\widehat{\boldsymbol{u_i}} \cup \widehat{\boldsymbol{u_{i-1}}} \cup \cdots \cup \widehat{\boldsymbol{u_1}}) \cdot (\widehat{\boldsymbol{c_i^{kw}}} - \widehat{\boldsymbol{c_{i+1}^{kw}}}) \quad (11)$$

Then, we can obtain the scoring matrix S shown in the equation (12) which contain all score across all product for all customer group.

$$\boldsymbol{S} = \begin{bmatrix} \boldsymbol{s_{11}} & \cdots & \boldsymbol{s_{1h}} \\ \vdots & \ddots & \vdots \\ \boldsymbol{s_{k1}} & \cdots & \boldsymbol{s_{kh}} \end{bmatrix} \quad (12)$$

Where S_{kw} is the fuzzy integral score of user group k against product w.

2.4. Score Personalization

So far, the group-level score can be obtained by fuzzy integral method mentioned above. However, for each product w, we still need the score of user i rather than user group k. Therefore, in this subsection, the score personalization procedure is introduced.

Staring from the user i's purchase record on all product which indicated as $\boldsymbol{P_i} = [\boldsymbol{p_{i1}}, \cdots \boldsymbol{p_{ih}}]$. In this research, $\boldsymbol{P_i}$ is used as a weight to convert group-level score $\boldsymbol{S_k} = [s_{i1}, s_{i2} \cdots, s_{ih}]$ by direct product on $\boldsymbol{f_{ik}} = \boldsymbol{P_i} \otimes \boldsymbol{S_k}$. Thus, $\boldsymbol{f_{ik}}$ is the score vector contains all scores of user i against product w. The F matrix can be further computed to obtain the all users' score for all products. This F can be used as the input matrix of CF method.

$$\boldsymbol{F} = \begin{bmatrix} \boldsymbol{f_{11}} & \cdots & \boldsymbol{f_{1h}} \\ \vdots & \ddots & \vdots \\ \boldsymbol{f_{n1}} & \cdots & \boldsymbol{f_{nh}} \end{bmatrix} \quad (13)$$

3. PRELIMINARY RESULT

In order to evaluate the proposed model, the dataset from Kaggle Santander competition which includes 1.5 year customers demographic and banking information data and their history of purchase record is used [14]. The customer demographic data contains customer's country residence, sex, age, gross income of the household, customer segment level, and so on. That customer information is very common in financial sector. The history purchase record consists of the records of purchasing financial products such as saving account, guarantees, mortgage, and so on. Those products (service) are also representable as the common products in financial institute.

In this paper, the Santander data is used as a test bed to evaluate the proposed fuzzy personalized scoring model. Due to the page limit, two pie charts in Fig. 2-3 which shows the result of scores distribution for three customer groups are presented here as examples. The meanings of products are listed below: Product 1 is Saving Account; Product 2 is Guarantees; Product 5 is Payroll Account; Product 6 is Junior Account; Product 9 is Particular Plus Account, Product 10 is

Short-term Deposits; Product 11 is Medium-term Deposits, Product 12 is Long-term Deposits, Product 15 is Mortgage, Product 22 is Payroll. The percentage indicates the score of a particular product divide by the sum of all scores of products.

As shown in Fig. 2, group "Payroll Oriented" has relative higher scores on products related to payroll functions. It implies that this group of customers is more interested in the fundamental product or service from financial institute. The representative customer from this group can be described as single, young, *with low level gross income of the household.*

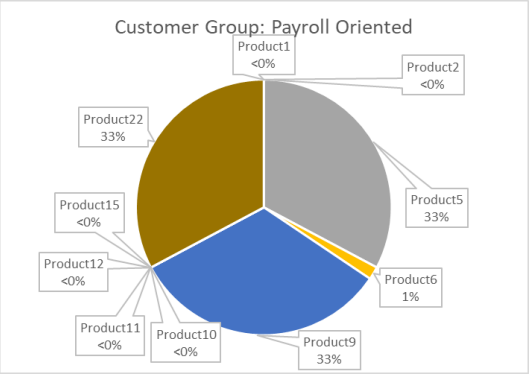

Fig. 2: The scoring proportion of the customer group "Payroll Oriented".

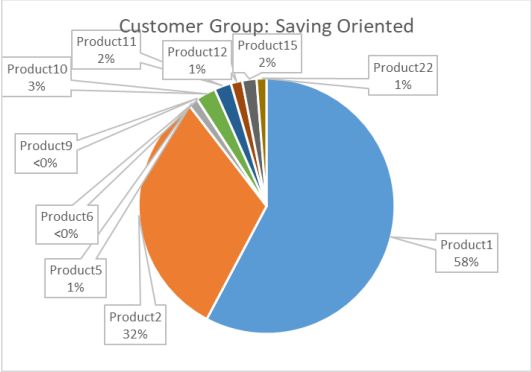

Fig. 3: The scoring proportion of the customer group "Saving Oriented".

From Fig. 3, we can see group "Saving Oriented" has very different pattern comparing with group "Payroll Oriented". Customers in this group are majorly interested in opening a saving and guarantees products which contributes to 80% of scores proportion. Rest of products only account to very few score proportion. The representative demographic information from this group can be described as male with high gross income.

The preliminary results show that the proposed scoring model is able to cluster the customers based on the importance of demographic features correlated to the product purchasing behaviors. For practical perspective in marketing, different group's correlation between demographic and purchasing preference can be used to allocate different marketing campaign. The last but not least, the score matrix of customer-product can be used as *the input of CF for further matrix factorization computation.*

4. CONCLUSION

Collaborative filtering method of recommendation system needs to compute and predict the rating of item-user association. In this research, the data preprocessing framework is proposed to generate the rating or scores of item-user as input of collaborative filtering method based on the customer demographic information and purchasing records. In order to differentiate the customer demographic background, we first cluster all customers to multiple groups. Then, for each group, the random forest method is applied to identify the importance of demographic features which are highly correlated to the decision of purchasing products.

The fuzzy integral method is utilized to aggregate importance of all demographic features and create an importance score against each product. This group-level scores can show the purchasing preference of a particular customer group. In order to calculate individual-level scores, the group-level scores are weighted by considering individual's purchasing historical record. Then, the customer-product matrix can be created as input of any collaborative filtering method.

In this work, the preliminary data analysis was performed based on Kaggle Santander competition data which includes customers demographic and banking information data and their history purchase record. The results of customer segmentation show the proposed method is able to identify each customer group's purchasing preference. More importantly, these purchasing preferences can be correlated to each group's demographic characteristics. By the proposed fuzzy scoring expending model, the importance score of each product for each customer can be generated as input of the collaborative filter method.

In the future work, we will extend the scoring system by considering the timing factor which is sensitive to customer's purchasing decision, especially when purchasing financial products. In addition, using financial data as a test bed, we will integrate the proposed scoring system with tensor factorization to evaluate the accuracy of recommendation system. We will compare the existing scoring model and our proposed method by applying the recommendation model to evaluate the performance.

ACKNOWLEDGMENT

We appreciate the financial support from National Science Council of Taiwan, R.O.C. (Contract No. 107-2218-E-011-014) and "Center for Cyber-Physical System Innovation" from The Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education (MOE) in Taiwan.

5. REFERENCES

[1] F. Ricci, L. Rokach, and B. Shapira, "Introduction to Recommender Systems Handbook," in *Recommender Systems Handbook*, F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, Eds. Boston, MA: Springer US, 2011, pp. 1-35.

[2] P. B. Thorat, R. Goudar, and S. Barve, "Survey on collaborative filtering, content-based filtering and hybrid recommendation system," *International Journal of Computer Applications,* vol. 110, no. 4, 2015.

[3] J. B. Schafer, D. Frankowski, J. Herlocker, and S. Sen, "Collaborative filtering recommender systems," in *The adaptive web*: Springer, 2007, pp. 291-324.

[4] P. K. M. Kanaujia, N. Behera, M. Pandey, and S. S. Rautaray, "Recommendation system for financial analytics," in *2016 International Conference on ICT in Business Industry & Government (ICTBIG)*, 2016, pp. 1-5.

[5] D. Zibriczky, "Recommender systems meet finance: a literature review," in *Proceedings of the 2nd International Workshop on Personalization & Recommender Systems in Financial Services*, Bari, Italy, June 16, 2016, 2016, pp. 3-10.

[6] C.-L. Yang and N. T. P. Quyen, "Data analysis framework of sequential clustering and classification using non-dominated sorting genetic algorithm," *Applied Soft Computing,* 2018.

[7] L.-C. Cheng and H.-A. Wang, "A fuzzy recommender system based on the integration of subjective preferences and objective information," *Applied Soft Computing,* vol. 18, pp. 290-301, 2014/05/01/ 2014.

[8] M. Chunyan, Y. Qiang, F. Haijing, and A. Goh, "Fuzzy cognitive agents for personalized recommendation," in *Proceedings of the Third International Conference on Web Information Systems Engineering, 2002. WISE 2002.*, 2002, pp. 362-371.

[9] Y. Cao and Y. Li, "An intelligent fuzzy-based recommendation system for consumer electronic products," *Expert Systems with Applications,* vol. 33, no. 1, pp. 230-240, 2007/07/01/ 2007.

[10] Z. Huang, "Clustering large data sets with mixed numeric and categorical values," in *Proceedings of the 1st pacific-asia conference on knowledge discovery and data mining,(PAKDD)*, 1997, pp. 21-34: Singapore.

[11] A. Palczewska, J. Palczewski, R. M. Robinson, and D. Neagu, "Interpreting random forest classification models using a feature contribution method," in *Integration of reusable systems*: Springer, 2014, pp. 193-218.

[12] C.-T. Lin and C. S. G. Lee, *Neural fuzzy systems: a neuro-fuzzy synergism to intelligent systems*. Prentice-Hall, Inc., 1996, p. 797.

[13] G.-H. Tzeng and J.-J. Huang, *Multiple Attribute Decision Making*. 2011.

[14] Santander. (2018, Octorber). *Santander Product Recommendation*. Available: https://www.kaggle.com/c/santander-product-recommendation