IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received January 15, 2022, accepted January 24, 2022, date of publication January 28, 2022, date of current version February 8, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3147447

Discover Customers' Gender From Online Shopping Behavior

YUAN AN[1], SIQIAO MENG[2], AND HAO WU[3]

[1]School of Computer Science and Information Technology, Daqing Normal University, Daqing 163712, China
[2]School of Education, City University of Macau, Macau, China
[3]School of Economics and Management, Zhejiang University of Science and Technology, Hangzhou 310023, China

Corresponding author: Yuan An (anyuan@dqnu.edu.cn)

ABSTRACT Gender information is very important for the recommendation system in the online shopping website. However, gender data often face label missing and incorrect labelling problems caused by consumers' unwillingness to actively disclose personal information, which leads to gender estimation results that cannot meet the needs of the product recommendation system. To discover the customers' gender information, we explore the customers' online shopping behavior, especially the items viewed in the shopping session, from the dataset provided by Vietnam FPT Group. The dataset is very imbalanced while the number of female samples is 3× of the male samples. To address the imbalance issue, we cluster the female samples into three subsets and then train a two-layer classifier model to estimate the customers' gender. Experimental results demonstrate that our proposed method could achieve a combined accuracy 78% on average, and takes less than 6 seconds on average. As a data mining model for gender prediction, our approach has a lightweight network structure and less time consumption.

INDEX TERMS Data mining, cluster, gender classification, online shopping.

I. INTRODUCTION

We have witnessed the rapid growth of the online shopping in the recent decade. As COVID-19 hit the world, more and more customers prefer shopping online instead of visiting the stores in person. The online shopping websites make large profits from the customers and also learn from customers' behavior for improving their shopping experience [1]–[4]. One important feature of a customer is the gender information. If the shopping website knows the customer's gender, more accurate recommendation would be performed. For instance, when a customer is searching for a keyboard, the results might be different for different gender. If the customer is a male, his possibility of purchasing a mechanical keyboard is higher than the female customer.

Despite the importance of the gender information, it is hard for the shopping websites to collect customers' gender information because of the privacy issue. Sometimes, the customers even fill in the wrong gender information in the shopping website to protect their own privacy. Given our proposed model, we can precisely estimate a customer's gender, which is very beneficial for improving the performance of the recommendation system in the shopping websites. In addition to the advantage our design brings to the recommendation system, another application of our design is that we can use our model to test the privacy protection mechanism in a shopping website. For example, if a shopping website wants to show the performance of its privacy protection mechanism, it could run our design on their customers' data. If our design could output the correct gender, the website's privacy protection mechanism does not work well. Otherwise, the website is good at protecting the customers' privacy.

Nevertheless, estimating customers' gender from the customers' viewing log is not trivial. The dataset containing gender information is very rare to find in the public sources since the gender information is very sensitive to the customers' privacy. Our design is based on the consumers' viewing log provided by Vietnam FPT Group, which is a leading information and communication enterprise, operating a number of B2B2C services. In general, online shopping log data includes information such as the buyer's product browsing and purchasing activities and the seller's product portfolio. The technical challenges in this paper include (1), the imbalanced data. In the FPT group's dataset, the number of female samples is three times of the male samples. Straightforwardly applying the existing data mining model only results in a

The associate editor coordinating the review of this manuscript and approving it for publication was Massimo Cafaro.

very poor estimation. (2), the size of dataset is small. The FPT group's dataset is only a few Megabytes, containing less than 20000 samples. It is hard to apply some deep learning model on such small size dataset. In this paper, we propose a data mining model, consisting of clustering, decision tree, and random forest models, to overcome these challenges and discover the customers' gender information from the customers' behavior, especially which products are viewed in a shopping session. We observe that despite of the imbalanced male/female samples, we could further divide the female samples into 3 subsets via clustering, indicating the customers' gender is not just related to the customers' physical gender, but also related to the customers' psychological gender. The main contributions of this article are as follows:

- We discover the correlation between personality diversity and gender in online shopping behavior, and explain the characteristics of customer shopping behavior in a specific web browsing log data set. These features are combined into feature combinations as candidate combinations for gender classification.
- We use personality diversity and data visualization to solve the problem of sample imbalance in the FPT group's online shopping behavior dataset. Based on the balanced sample set, the optimal classifier is selected for each layer of the designed gender classification network to improve the performance of gender classification.
- We conducted experiments using a large-scale data set provided by FPT Group and get the estimation accuracy of 78% within less than 6 seconds. The results prove the lightweight and high-efficiency of the proposed gender classification model.

II. RELATED WORK

In the recent years, researchers realize the importance of gender information and believe that gender information is the key factor in solving this contradiction [5], [6]. Empirical research shows that gender, as a typical feature of online shopping behavior that promote customers' demands analysis [7] and personalized recommendations technologies development [8], play an important role in increasing customers satisfaction [9] and online shopping recommendation systems performance [10]. Chen *et al.* [11] analyzed the moderating effect of gender on customers' shopping behavior based a benefit–risk paradigm model, and found that gender has a significant impact on online shopping willingness. Sohab *et al.* [12] studied the moderating effect of consumer cognitive innovation on the influencing factors of iTrust (interpersonal trust) on online purchase intention of new products, and found that gender information is helpful for the product display design of online websites. Lin *et al.* [13] did research on the gender differences of customers' online shopping psychology and behavior, and showed that gender information can promote the improvement and benefits of online shopping websites. Due to the gender information is essential to improve product recommendation performance, some researchers had proposed personalized recommendation algorithms or techniques based on gender information to improve online shopping recommendation systems, for example, Liu *et al.* [14], Karthik and Ganapathy [15], Hammou *et al.* [16], Wu and Yu [17] and Liu and Wei [18]. All these personalized recommendation algorithms or technologies provide many references for online shopping companies to improve their online shopping recommendation systems in time.

It is worth noting that no matter what kind of personalized recommendation algorithm or technology is, it needs to be based on real customers' gender information to play its corresponding effect. This is because based on accurate customer gender [19], it is possible to better discover customer preferences [5], improve the accuracy of product recommendations [6], thereby promoting the development of gender marketing [20], [21], as well as increasing online merchants' Income [22], [23]. For this reason, many researchers have done a lot of research from the perspective of customers' gender information collecting technologies and methods [24]–[26]. Gender information can be collected through questionnaires [27], recruiting volunteers [28], [29] and the information registered by the user [30]. Nonetheless, the gender information collected through these collected methods is far less than enough to contribute to the online shopping recommendation system [31]. However, customers may not want to actively disclose their privacy information [32], so they will ignore or randomly select gender during website registration and set account privacy [33], [34], resulting in incomplete and untrue gender information collected by online shopping system [35]. Therefore, the estimation of customers' gender becomes necessary. Unfortunately, despite the advanced algorithms and technologies for personalized products recommendation are proposed, it is very hard for online shopping companies to change this totally, because the cost of hiring people to check the consumers' gender is unaffordable. Thus, it is urgent to find an effective approach to estimate the true gender of consumers in online shopping recommendation systems. There are also some papers using facial recognition methods to detect the users' gender [57], [58]. However, these methods only work if the users permit the access to the cameras. In contrast, our paper aims at discovering customers' gender from their online shopping behavior, instead of their photos, preserving the customers' privacy.

Despite the lacking of research on mining the customers' gender given the unreliable online shopping system gender data, there are many models based on mining customers online shopping browsing log and purchase log data were proposed to estimate the customers' gender. Zhou *et al.* [36] using the RFMT model to derive 7 characteristic customer clusters from a large dataset retrieved on a global retailer's website, and estimated customers' gender and personalized products preferences by the cluster analysis. Wan *et al.* [37] used large-scale online shopping transaction log modeling to mine consumer personalized preferences for gender estimation. However, their approach mainly relies on the analysis of the users' click behaviors, and ignore the female personality diversity or male personality diversity, and the samples imbalanced issue, which may not be reliable and accurate.

In summary, the research on gender estimation mainly focuses on the impact of customers' shopping behavior and personalized preferences on models [38]–[40]. In order to ensure the effectiveness of the model, the authenticity of gender data in the online shopping system is very critical. Although the gender information can reflect the customers' shopping behavior and preference [41]–[43], it is impossible to distinguish the fake gender information users registered [44]–[47]. The lack or fake of gender data leads to the unreliability of the gender estimation model and the effect that the recommendation system cannot provide consumers with the most needed products [48], [49]. If this continues, it will cause the performance of the online shopping recommendation system to decline, and also affect the economic benefits of e-commerce companies.

III. GIVEN DATA FORMAT AND RESULTS MEASUREMENT

The data we studied is a customers' online product browsing log in a specific time period provided by Vietnam FPT Group. These training data and their corresponding gender labels and test data are all from the PAKDD'15 data mining contest website. In addition, the website also announced the final results and rankings of the competition. In this paper, these data are used as the training set and the test set, and gender estimation is performed on this basis. At the same time, the average combined accuracy of gender estimation can also find the corresponding interval in the competition results published on this website. The data set format can be described in detail as follows:

The training data set contains 11,703 female samples and 3297 male samples, and the corresponding 15,000 gender labels. The number of the products in our dataset is 36634 while the subcategory A has 11 products, the subcategory B has 91 products, the subcategory C has 440 products, and the subcategory D has 36092 products. Its file storage space is 1651 KB. The test data set contains 15,000 samples that lack gender labels. It is stored in a file with a size of 1639 KB. Since the data format of each sample in the two data sets is the same, 4 samples are randomly selected from the training set for display, as shown in Table 1. Each sample represents a customer's viewing session and contains 5 columns data. Specifically, the first 4 columns are "Session ID", "Start Time", "End Time" and "Product IDs", respectively and the last column is "Gender Label". Among them, the "Session ID" column is the session ID, the "Start time" column is the session start time, the "End time" column is the session end time, the "Product IDs" column is the product IDs viewed by a consumer, and the "Gender Label" is the customer's gender. In the "Product IDs" column, there are 4 categories of IDs: The most generalized products are represented by the IDs beginning with 'A'. These product IDs beginning with 'B' and 'C' are the subcategoris and sub-subcategories of the products, respectively. The product IDs start with 'D' are the fourth category, corresponding to individual products. The data used in this paper only shows the items the customers viewed while it is unknown if the item is purchased or not. In addition, some more information about the items, such as price, is also unknown. We are predicting the gender using the minimum information from the customers, indicating the potential of applying our design in more restrict scenarios.

The vectors "predict" and "actual" represent the predicted results of this paper and the truth gender labels, respectively. The variables ACC_m and ACC_f are used to represent the accuracy of predicted male and female, respectively. Next, the integer 0 represents the male label, and the integer 1 represents the female label. Then, the results measurement is followed as:

$$ACC_m(predict, actual) = \frac{|a : predict_a = actual_a = 0|}{|a : actual_a = 0|} \tag{1}$$

and

$$ACC_f(predict, actual) = \frac{|a : predict_a = actual_a = 1|}{|a : actual_a = 1|}. \tag{2}$$

Since the distribution of female labels and male labels is imbalanced, the results of gender prediction will be measured using "Combined Accuracy (CA)". According to the (1) and (2), the definition of combined accuracy is as follows:

$$CA(predict, actual) = \frac{ACC_f(predict, actual)}{2} + \frac{ACC_m(predict, actual)}{2}. \tag{3}$$

In summary, through the research on the data format of the training data set, it is found that each sample data contains the ID of the browsing session, the start time, the end time, and the viewed product IDs. These samples are similar to each other. Directly from the first 4 columns of data, it is difficult to get the same predicted label as the truth gender label. This also means that the correlation between the training sample data and the training sample label is low, which would make the generalization ability of the obtained prediction model low. In addition, in the training sample set, female samples accounted for about 75% while male samples accounted for about 25%, which would lead to sample imbalance and further reduce the generalization ability of the prediction model. To further reduce the impact of sample imbalance on gender prediction results, CA measurement needs to be used to measure the results of gender estimation.

IV. GENDER MINING MODEL SOLUTION

A. FEATURE EXTRACTION AND CANDIDATE FEATURE COMBINATIONS

In order to solve the issue that the correlation between the training sample data and the training sample label is low. Meaningful features should be defined to describe

TABLE 1. Training data format.

Session ID	Start Time	End Time	Product IDs	Gender Label
u10018	2014/12/6 16:09:40	2014/12/6 16 :11:42	A00003/B00022/C00048/D20036/; A00002/B00001/C00010/D18416/	female
u14419	2014/11/15 22:14:07	2014/11/15 22:15:36	A00001/B00031/C00091/D01848/; A00001/B00031/C00091/D03417/; A00001/B00031/C00091/D03009/	male
u17715	2014/11/23 18:50:21	2014/11/23 18:50:28	A00002/B00007/C00023/D10821/	male
u24980	2014/12/22 19:56:57	2014/12/22 19:58:20	A00003/B00012/C00051/D35858/; A00003/B00012/C00051/D35859/; A00003/B00012/C00051/D19736/	female

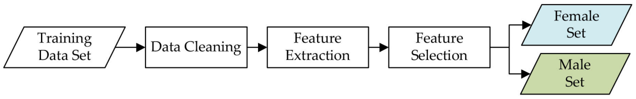


FIGURE 1. Training set generation. The female set and the male set together form the training set.

the given training data set. Since this data set consists of 15000 sessions, in the i-th session, the variable s_i represents the i-th session data, $t_s^{(s_i)}$ and $t_e^{(s_i)}$ denote the start time and end time, respectively. Meanwhile, these view product IDs in this session can be expressed as $product\_ID^{(s_i)} = \{product\_ID_1^{(s_i)}, product\_ID_2^{(s_i)}, \ldots, product\_ID_j^{(s_i)}\}$. These features can be defined as:

Definition 1: $F1$ is referred to being as number of products viewed. Since the number of products viewed in the i-th session can be denoted by $|\, product\_ID^{(s_i)}|$. Then, $F1$ in the i-th session is $|\, product\_ID^{(s_i)}|$.

Definition 2: $F2$ is referred to being as average time spent on each view product. Then, in the i-th session, $F2$ is ($t_s^{(s_i)}$ - $t_e^{(s_i)}$) / $|\, product\_ID^{(s_i)}|$.

Definition 3: $F3$ is referred to being as start time of the session. Then, $F3$ in the i-th session is $t_s^{(s_i)}$.

Definition 4: $F4$ is referred to being as ID of the maximum subcategorized ('B' category) products. Then, in the i-th session, $F4$ is $max\{\, product\_ID_j^{(s_i)} \mid product\_ID_j^{(s_i)}$ *starts with* 'B'}.

Definition 5: $F5$ is referred to being as ID of the maximum sub-subcategorized ('C' category) products. Then, in the i-th session, $F5$ is $max\{\, product\_ID_j^{(s_i)} \mid product\_ID_j^{(s_i)}$ *starts with* 'C' }.

As shown in Fig. 1, first do data cleaning for the training data set. Then, use Definition 1 to Definition 5 for feature extraction, respectively. After that, feature selection is conducted based on the extracted features. Finally, the female set and male set defined by the combination of these selected features constitute the training set.

Then, perform feature selection based on these 5 featured definitions and decide which features to use in our approach. Random forests are an integrated classifier composed of a set of decision tree classifiers [50]. For the given training sample set X, random forests trains K decision tree classifiers, and allows these K decision tree classifiers to participate in voting, and the prediction results of this sample set are determined by majority voting.

In the decision tree generation process, the tree node splits itself into left and right sub-trees according to the selected optimal attribute. The splitting process after comparing other attributes is node splitting. In this paper, the CART (Classification And Regression Tree) algorithm [51] based on Gini coefficient splitting is used to generate each decision tree. Specifically, when a node is split, the CART algorithm first calculates the Gini coefficient of the two subsets after each attribute is split. Then, select the attribute that minimizes the Gini coefficient to split the node into two left and right subnodes. Finally, the decision tree is constructed in the form of recursion. To save the space of the paper, the details of our calculation process of the CART algorithm is demonstrated in Appendix VI.

After CART algorithm is executed, the majority voting method is used to combine all the decision tree classifiers in the random forest obtained. Assuming that the random forests contain K decision tree classifiers, the decision tree classifier is $h_1, h_2, \ldots, h_K$, and the sample x is input to the decision tree classifier and the output is $h_k(x)$. For the customer gender classification task in this article, the decision tree classifier h_k will predict a category tag from the category tag set $\{c_1, c_2, \ldots, c_N\}$. The detailed steps of the random forest is demonstrated in Appendix VI.

After we apply the random forest classifier, we complete the category prediction on the training set X, and we can perform category prediction on the test set. Therefore, the random forest classifier is suitable for gender classification of feature combinations. Because 32 combinations of these 5 features need to be classified by gender, random forester needs to be run to get the gender classification result. These feature combinations with a prediction combined accuracy of more than 50% are selected as candidate feature combinations for further research.

B. CLUSTERING BASED ON PERSONALITY DIVERSITY

Given significant personality overlap among genders [53], and all the customer have diversity of personalities [54], [55], that is to say, different personalities also exist in the female. Therefore, under normal circumstances, there are a large number of personality diversity phenomena among female customers, and these phenomena can be reflected by different

TABLE 2. Combined accuracy for candidate feature combinations.

	F1&F3&F4&F5	F3&F4&F5	F1&F2&F4&F5	F1&F4&F5	F1&F2&F3&F4&F5
CA	0.677	0.660	0.658	0.644	0.643
	F2&F4&F5	**F2&F3&F4&F5**	**F4&F5**	**F1&F2&F5**	**F1&F3&F4**
CA	0.638	0.627	0.612	0.608	0.607
	F1&F2&F4	**F1&F3&F5**	**F1&F2&F3&F5**	**F1&F2&F3&F4**	**F1&F5**
CA	0.603	0.597	0.586	0.582	0.557
	F2&F3&F5	**F2&F3&F4**	**F3&F5**	**F1&F4**	**F3&F4**
CA	0.539	0.536	0.531	0.515	0.514

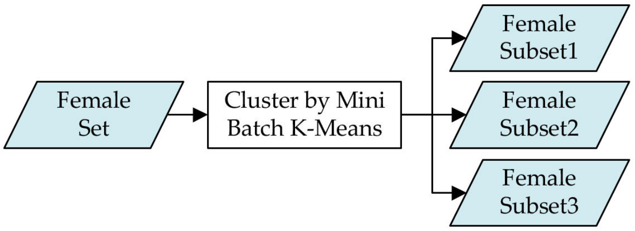


FIGURE 2. Female subsets generation.

feature combinations. In the female set, samples with the same personality as a certain type of personality can naturally gather into a cluster. In the given data set, the female sample is about three times as large as the male sample. If there is a combination of characteristics that can clearly reflect the three personalities, then the female set can naturally be clustered into three clusters. Then count the number of samples in each cluster after clustering. If the number of samples in the three clusters is almost equal, the corresponding feature combination is the feature to be selected. If there are multiple such feature combinations, select the feature combination with the closest number of three cluster samples.

A variant of the traditional K-Means algorithm is the Mini Batch K-Means (MBKM) algorithm, which uses a mini batch of data subsets obtained by random sampling in each iteration to update the centroid. Empirical research shows that this algorithm can increase the calculation speed of the clustering process when the sample volume is big, and effectively reduce the algorithm convergence time [56]. Since the training data set contains 15000 samples, it is a big sample set that can be clustered by this algorithm. The clustering process is shown in Fig. 2.

This paper uses the MBKM method to cluster the female set defined by each feature combination in Table 2, and the corresponding clustering results are shown in Table 3. We split the data sets into 3 clusters because the number of female samples is 3 times of the male samples. To balance the female and male samples in the training set, we consider 3 cluster centers in the subsets generation.

From Table 3, the number of samples in each cluster after the female set defined by the feature combination F3&F4&F5 is clustered are 3797, 4079 and 3827, respectively, and their ratio is close to 1:1:1. In other words, by selecting the feature combination F3&F4&F5, the sample size of each female subset is almost equal to the sample size of the male set. However, after clustering the female set defined by other feature combinations, the proportion of sample size in each cluster is obviously not as good as that of F3&F4&F5. Hence, the imbalance problem of a given training set can be solved preliminarily.

In order to further confirm whether the feature combination of F3&F4&F5 can clearly reflect three clearly and different personalities. Combining Table 2 and Table 3, the three feature combinations with the highest combined accuracy to the fourth highest are F3&F4&F5, F1&F4&F5, F2&F4&F5, and F1&F2&F5, respectively. Then, dram them as 3D, and each corresponding 3D data visualization view is shown in Fig. 3.

Fig. 3 shows that the female set defined by feature combination F3&F4&F5 is clearly divided into 3 subsets that are clearly separated from each other. The female subsets portrayed by other combinations of characteristics are not independent of each other, the interval is not obvious, and they cannot clearly reflect the three independent personalities. Therefore, this paper selects the feature combination F3&F4&F5 as the feature combination.

Then, the female set defined by the feature combination F3&F4&F5 can be average divided into three female subsets. After that, as Fig. 4 depicts, merge the male set and each female subset to generate a balanced training set, and get three such balanced training sets $TS1$, $TS2$ and $TS3$. Finally, the problem of sample imbalance is solved.

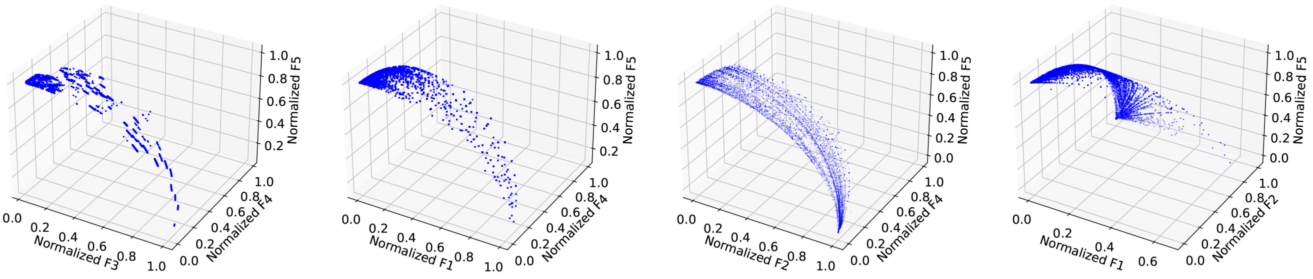
C. A TWO-LAYER GENDER CLASSIFICATION MODEL

Based on the three balanced training sets obtained, a two-layer gender classification network is designed as the gender estimation model. This model is shown in Fig. 5.

The hidden layer of the network consists of three classifiers, and the output layer consists of one classifier. Each classifier in the hidden layer and the output layer can use any typical classification algorithm, such as random forest, SVM, decision tree, and Gaussian NB. On this basis, $C_1^{(1)}$, $C_2^{(1)}$ and $C_3^{(1)}$ represent the optimal classifiers obtained after training with the training subset $TS1$, training with the training subset $TS2$, and training with the training subset $TS3$, respectively. Then use $C_1^{(1)}$, $C_2^{(1)}$ and $C_3^{(1)}$ to form the hidden layer of the network. The output of the hidden layer is used as the input of the output layer, and a classifier $C_1^{(2)}$ with the best classification result is trained to form the output layer of the network. This also means that each layer of classifier has

TABLE 3. Number of samples in each cluster after the sample sets defined by different feature combinations are clustered.

	F1&F3&F4&F5	**F3&F4&F5**	**F1&F2&F4&F5**	**F1&F4&F5**	**F1&F2&F3&F4&F5**
Number of samples	4147&3748&3808	3797&4079&3827	5284&2971&3448	3917&5489&2297	3167&5589&2947
	F2&F4&F5	**F2&F3&F4&F5**	**F4&F5**	**F1&F2&F5**	**F1&F3&F4**
Number of samples	5376&2962&3365	3165&5595&2943	2710&5092&3901	5468&2791&3444	7054&2216&2433
	F1&F2&F4	**F1&F3&F5**	**F1&F2&F3&F5**	**F1&F2&F3&F4**	**F1&F5**
Number of samples	2919&2673&6111	3507&6142&2054	3185&5568&2950	4804&2154&4745	7062&2831&1810
	F2&F3&F5	**F2&F3&F4**	**F3&F5**	**F1&F4**	**F3&F4**
Number of samples	3183&5573&2947	4805&2158&4740	3580&6094&2029	4318&3270&4115	2394&2309&7000
	F2&F4	**F1&F3**	**F1&F2&F3**	**F2&F5**	**F1&F2**
Number of samples	2573&2551&6579	6130&1783&3790	2514&4965&4224	5466&2794&3443	2370&644&8689



(a) F3&F4&F5 (b) F1&F4&F5 (c) F2&F4&F5 (d) F1&F2&F5

FIGURE 3. Data visualization for three feature combinations. (a,b,c,d) the data visualization view corresponding to the top 4 combined accuracy rates in the 3 feature combinations, respectively.

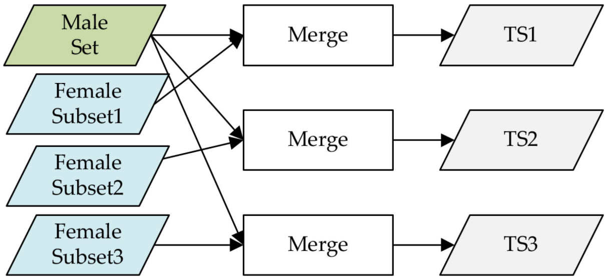


FIGURE 4. Three balanced training sets generation.

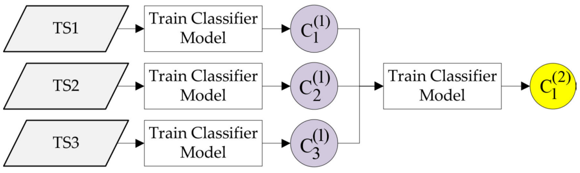


FIGURE 5. Training and generation of a two-layer gender classification model.

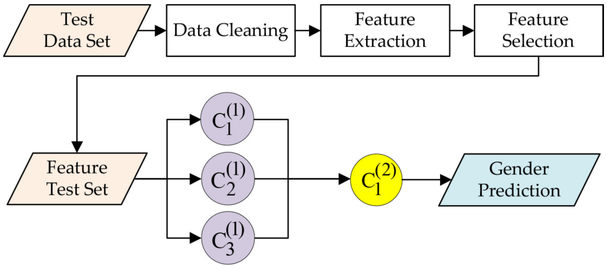


FIGURE 6. Gender prediction.

selected the best classification. Thereby, a two-layer gender classification model is obtained.

Finally, we could use the second-level classifier model to make the final gender decision as shown in Fig. 6.

D. SUMMARY

In summary, the approach of this paper includes the following steps:

1) Feature extraction and selection. We first read the training set file and check whether there are samples with incomplete information. If some samples do not have all four attributes information, we can delete these samples from the training set. Fortunately, all samples in the training set have all 4 attributes. If some samples do not have all 4 attributes, we can delete these samples from the training set. We then conduct feature extraction by Definition 1 to Definition 5 in Section IV-A. After that, select 25 feature combinations as candidate feature combinations;
2) Generate three balanced training subsets. The given training data is unbalanced because the number of female samples is three times that of male samples. Based on personality diversity, we try to use MBKM to find three clusters in the female set according to the relevancy between personality diversity and gender. Then, we find that the combination of F3&F4&F5 can describe the three clusters more clearly than other feature combinations. Therefore, we finally select

F3&F4&F5 combination as the feature combination of our approach. Then, we merge each female subset with male subset to generate a new training subset, so we can get three approximately balanced training subsets TS1, TS2, and TS3.

3) Train and get the gender prediction model. Based on these three training sets, we train the classifiers $C_1^{(1)}$, $C_2^{(1)}$ and $C_3^{(1)}$, respectively and use them as the nodes of the first-layer network. According to the output of each node of the first layer network, we design a new classifier $C_1^{(2)}$ as the second layer network node to make the final gender decision. Considering the irrelevance between the proposed method and these classifiers, representative classification algorithms such as random forests, SVM, decision tree and Gaussian Naive Bayes can be selected as candidate classifiers. In other words, we could select a classifier from the candidate classifiers as any node in each layer, such as $C_1^{(1)}$ uses decision tree and $C_1^{(2)}$ adopts random forests. Therefore, for the combination $C_1^{(1)}\&C_2^{(1)}\&C_3^{(1)}\&C_1^{(2)}$, we can get different combinations of classification algorithms. Then, the combination with the highest gender estimation combined accuracy is selected, and each classification algorithm of the combination is used as the classifier corresponding to each layer in turn. Finally, train the two-layer classifiers network and make the final gender decision.

V. EVALUATION AND DISCUSSION

A. FEATURE COMBINATION SELECTION

Select the appropriate clustering features from the 5 features defined in Section IV-A, which can better reflect the correlation between the training data and gender labels. As Fig. 7 shows, the feature combination F3&F4&F5 selected in this paper can depict the clear three clusters, which further indicates the diversity of female personality and divide the female set into three clusters with approximately the same number of samples. In other words, the clusters 1, 2 and 3 divided by the feature combination F3&F4&F5 contain 3797, 4079 and 3827 samples, respectively. Compared with female clusters classified by other combinations of features, the number of samples in the three clusters of F3&F4&F5 classification is the most balanced. This is because other feature combinations divide a large number of samples into specific subsets, resulting in large imbalance of samples between subsets.

To verify the effect of feature combination selection, we use the MBKM clustering of scikit-learn package in python to cluster three clusters for the female set, and record the number of samples contained in each cluster. According to Table 3, the number of samples between each cluster of the female set may differ by more than 14 times. In other words, the feature combination selection has a direct impact on the number of samples contained in each cluster in the female set. Therefore, the most balanced combination F3&F4&F5 is suitable as the feature combination of our method.

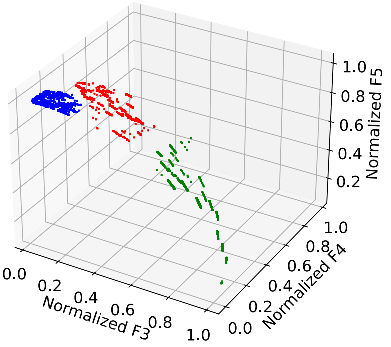


FIGURE 7. Data visualization of female subset clusters.

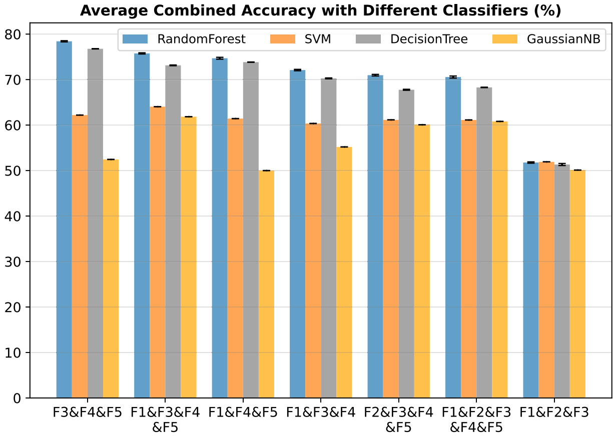


FIGURE 8. Average combined accuracy measurement based on different classifiers.

B. CLASSIFIER SELECTION

In the two-layer classifier network we designed, the first layer has three nodes $C_1^{(1)}$, $C_2^{(1)}$ and $C_3^{(1)}$, and the second layer has one node $C_1^{(2)}$. We could select random forests, SVM, decision tree, Gaussian Naive Bayes, etc. as candidate classifiers among representative classification algorithms. In other words, we can select a classifier from the candidate classifiers as any node in each layer. Therefore, for the combination $C_1^{(1)}\&C_2^{(1)}\&C_3^{(1)}\&C_1^{(2)}$, we can get different combinations of classification algorithms. Then, the combination with the highest gender estimation combined accuracy is selected, and each classification algorithm of the combination is used as the classifier corresponding to each layer node in turn. Finally, the trained two-layer classifier network is used to test the gender prediction performance of our proposed method. Finally, after our model uses the random forest classifier to perform gender classification, select the top 6 feature combinations with the highest average combined accuracy, and then select the feature combination F1&F2&F3, and use them to measure the average combined accuracy of these different algorithms. Then, Fig. 8 shows the result.

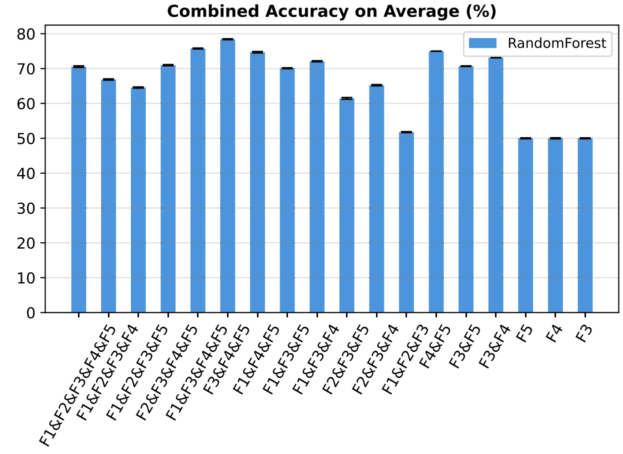


FIGURE 9. Combined accuracy with different feature combinations.

It can be seen from Fig. 8 that on the training subsets generated by the feature combination F3&F4&F5, the random forest classifier has better average combined accuracy than that of the other combinations. In addition, among all four typical classifiers, the average combined accuracy of the decision tree and SVM classifiers are second and third, respectively, and the Gaussian Naive Bayes classifier has the lowest combined accuracy.

C. COMBINED ACCURACY MEASUREMENT

The sample size of women is three times that of men, which leads to the problem of sample imbalance. The feature combination F3&F4&F5 selected in this paper can better solve the problem, while the other combinations cannot solve the problem better. In general, sample imbalance will make gender prediction results more biased towards sample categories with a larger sample size, that is, gender prediction results will be more biased towards female. To improve the balance and credibility of gender prediction results, the combined accuracy (3) is used to evaluate the accuracy of our proposed model.

When the model proposed in this paper uses a random forest classifier, the average combined accuracy of gender prediction is the highest. On this basis, the average gender combined accuracy of the sample set defined by each feature combination is measured, and 18 feature combinations including the first 6 combined accuracy are selected for display. The average combined accuracy result based on each feature combination is shown in Fig. 9. Through comparison, it can be seen that the average combined accuracy of gender prediction based on the feature combination F3&F4&F5 is 78%, which is the best classification effect among all combinations. In addition, it can be observed that with the increase of features, the combined accuracy presents a distribution trend that first increases and then decreases. This also reflects that the number and combination of features have an important influence on the combined accuracy results. For example, the average combined accuracy of a single feature F3 or F4 is always less than that of the two feature combinations F3&F4, while the average combined accuracy of the feature combination F3&F4&F5 is significantly higher than that of F1&F2&F3. We can also find that F3&F4&F5 and F1&F2&F3 have the same number of features, but the average combined accuracy is quite different. Even F1&F2&F3 is lower than F3&F4, F3&F5 and F4&F5. This is because the shopping behavior represented by the feature combination F3&F4&F5 is more accurate in relation to gender than the shopping preference represented by the feature combination F1&F2&F3. The reason for this trend is that the feature combination F3&F4&F5 can cluster the female set into three clusters with the most similar sample numbers and significant separation from each other. This also means that the feature combination F3&F4&F5 can show a clearer personality in terms of gender than other feature combinations.

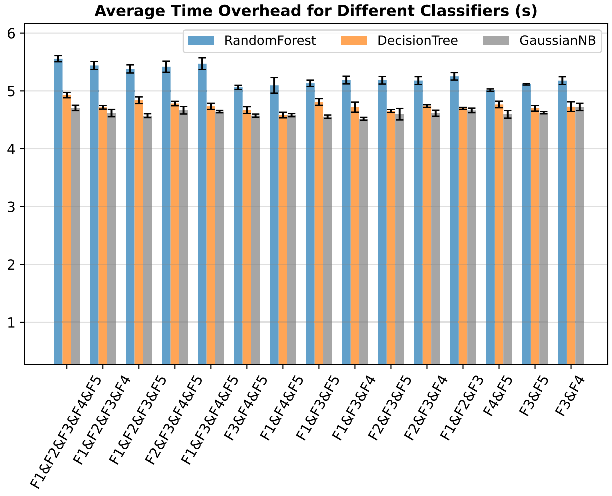


FIGURE 10. Average time overhead for random forest, decision tree and GaussianNB classifiers.

D. TIME OVERHEAD

As we known, the field of e-commerce generates massive amounts of online transaction data all the time. How to dig out useful features from these data in a timely and effective manner for gender prediction is also a challenge to be considered and faced in this article. In other words, our method can extract meaningful features from large-scale data sets in a timely and effective manner, and should be efficient and robust after training a large number of times. Only on this basis can our method be applied to practical and commercial scenarios. Therefore, time overhead is a very critical factor. It can be seen that we need to evaluate the time overhead of different classifiers on different feature combinations. The time cost of this experiment includes two processes of feature extraction and model training. Fig. 10 and Fig. 11 summarize the corresponding average time costs.

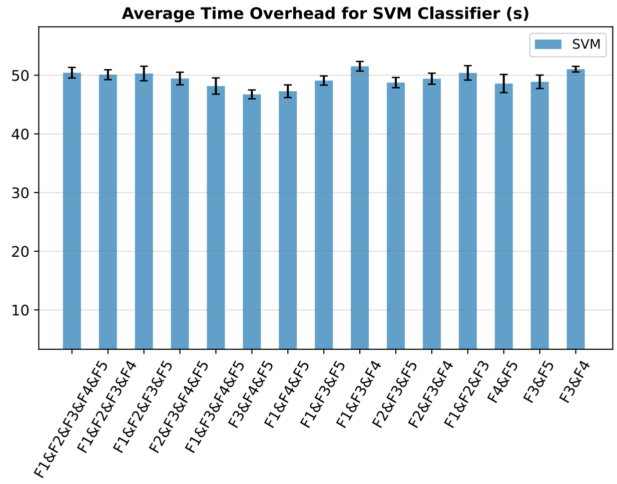


FIGURE 11. Average time overhead for SVM classifier.

Fig. 10 reflects the correlation between the number of features and the time overhead. This is because it takes a certain amount of time to extract a feature from a given data set, and as the number of features to be extracted increases, the time spent will also increase. Meanwhile, the average time-consuming of random forests is very little more than that of the other two classifiers, but it is within an acceptable range. This is because sacrificing average combined accuracy to pursue less average time-consuming is not our first choice. In order to make the average combined accuracy as high as possible while keeping the average time-consuming as short as possible, we can adjust the depth of the tree in the random forest algorithm.

Fig. 11 shows that the correlation between the number of features and the time cost of the SVM classifier is not as good as that of the three classifiers in Fig. 10, and it takes more time. This also reflects that as the number of female samples and male samples increases, the classification lines required in different dimensions will also increase, which increases the complexity and leads to a large number of calculations and increases the time overhead. Among these combinations, the feature combination F3&F4&F5 selected in this paper has the least average time overhead. It also reflects the advantages of lighter weight of the method in this paper. For this paper, if the method does not provide good gender prediction results, then less time overhead is of no research value. In other words, compared with the time overhead, the combined accuracy is more important.

E. DISCUSSION

The customers' gender information is vital to improve the product recommendation performance, and it is significant for the in-depth research of shopping recommendation system. However, the gender information of the existing online shopping recommendation system mainly relies on the gender data users registered. For customers who are unwilling to actively disclose their privacy, the authenticity of the gender information they provide cannot be guaranteed. Meanwhile, the existing gender estimation model based on customers' online shopping transaction log data mainly relies on the analysis of users' click behaviors, ignoring the diversity of female and male personality. Moreover, these gender estimation models may be time-consuming to estimate the gender of large-scale imbalanced data that changes in real time. To this end, a gender estimation model based on personality diversity is proposed, and customer gender is estimated through shopping log data mining. Through our evaluation, we find that the model is more lightweight and efficient.

Up to now, in the face of massive shopping log data, how to process these data in a timely manner and extract features from it for effective customer gender estimation, which has further demonstrated the challenge of estimating customer gender from the massive shopping log data. Therefore, we want to leave the research on how the industry adjusts the feature extraction speed according to the constantly changing scale of online shopping data, thereby affecting the time spent in the customer gender estimation process.

We note that the data we obtained from FPT group website are labeled with 'Male' and 'Female'. Basically all samples have their labels, which are either male or female. For the dataset, which contains just a few samples without label, we just train the model without using these samples. However, training the model using the dataset containing too many not labeled samples if out of the scope of our proposed and we decide to leave this to the future work.

VI. CONCLUSION

This paper introduces a novel approach to mine the customers' gender information from the online product viewing log provided by Vietnam FPT Group. First, we make feature combinations based on the extracted features to reflect the correlation between personality diversity and gender, and select the best feature combination through data visualization. Therefore, we can solve the problem of low correlation between training data and gender labels. Then, using the best feature combination, the female samples are naturally clustered into three subsets equal to the number of male samples. Each female subset and male set generate a balanced training subset. In this way, three balanced training subsets can be obtained. At this point, we can solve the issue of unbalanced training samples. Finally, based on these three balanced training subsets, three independent classifiers are trained as the nodes of the first-layer network. Then train a new classifier as the second layer network node based on the output of the first-layer network. On this basis, a two-layer classifier network can be designed and trained to make the final gender decision. Experimental results on the given data set show that our proposed method can provide accurate prediction results and consume less time. As a data mining model for gender prediction, our method is lightweight and efficient, and can be applied to different actual and e-commerce scenarios.

APPENDIX A DETAILED STEPS OF DECISION TREE MODEL

We demonstrate the detailed calculation of CART algorithm as follows:

1) Calculate the Gini coefficient of the sample set. Suppose that the sample set X contains C categories of samples, and the proportion of each category of samples is $P_i(i = 1, 2, \ldots, C)$. Then the Gini coefficient of the sample set can be expressed as

$$Gini(X) = 1 - \sum_{i=1}^{C} P_i^2. \tag{4}$$

The Gini coefficient $Gini(X)$ reflects the probability of inconsistent class labels when two samples are drawn at random from dataset X. Therefore, the smaller the $Gini(X)$, the higher the purity of the dataset X.

2) Calculate the data set divided by the Gini index of each feature. The decision tree can be built recursively by means of bisection splitting. Each node is split by the CART algorithm, which adopts Gini-index as the split criterion [52]. Suppose that the set F represents all the features in the sample set X, and $F = f^1, f^2, \ldots, f^M$. It can be seen that there are M features in the set F, and any feature $f \in F$. If f used to divide the training set X, M branch nodes will be generated. We use X^m to denote all samples in X whose value is f^m on feature f, and divide X^m into the m-th branch node. In this paper, the features extracted by Definition 1 to Definition 5 can be regarded as different features of the training set, and these features can also be regarded as different attributes. Then the Gini-index of attribute f can be expressed as

$$Gini\_index(X, f) = \sum_{m=1}^{M} \frac{|X^m|}{|X|} Gini(X^m). \tag{5}$$

In the candidate attribute set F, we use the Gini-index of attribute f to divide and score, and find the attribute that makes the Gini-index the smallest after the division as the optimal division attribute, namely

$$f_* = \underset{f \in F}{arg\ min}\, Gini\_index(X, f). \tag{6}$$

3) Recursively build the tree. For the divided decision tree, repeat step 2) until the division cannot be continued or the Gini value is less than the set threshold.
4) Output the final CART decision tree.

APPENDIX B DETAILED STEPS OF RANDOM FOREST MODEL

So as to facilitate prediction, we use an N-dimensional vector $(h_k^1(x), h_k^2(x), \ldots, h_k^j(x), \ldots, h_k^N(x))$ to represent the prediction output obtained after the sample x is input to h_k. Then the logarithmic output $h_k(x) \in \mathbb{R}$, the combination strategy we adopt is majority voting:

$$H(x) = \begin{cases} c_n, & if \sum_{k=1}^{K} h_k^n(x) > 0.5 \sum_{n=1}^{N} \sum_{k=1}^{K} h_k^n(x); \\ reject, & otherwise. \end{cases} \tag{7}$$

where the output of h_k on the category label c_n is $h_k^n(x)$. If a tag has more than half of the votes, random forests will predict the sample x as the tag. Otherwise, the prediction is rejected. In this article, the customer gender classification task requires that the gender prediction results of the sample must be provided, and the majority voting method will degenerate into plurality voting:

$$H(x) = c_{\underset{n}{arg\ max} \sum_{k=1}^{K} h_k^n(x)}. \tag{8}$$

When counting the tags with the most votes, if there are multiple such tags, one is randomly selected as the category tag as the prediction result of the sample.

ACKNOWLEDGMENT

The authors sincerely thank the anonymous reviewers for their valuable and constructive comments. They also thank Phoenixlab for the proofreading.

REFERENCES

[1] W. Nadeem, D. Andreini, J. Salo, and T. Laukkanen, "Engaging consumers online through websites and social media: A gender study of Italian generation Y clothing consumers," *Int. J. Inf. Manage.*, vol. 35, no. 4, pp. 15–64, Aug. 2015.

[2] B. Hasan, "Exploring gender differences in online shopping attitude," *Comput. Hum. Behav.*, vol. 26, no. 4, pp. 597–601, Jul. 2010.

[3] Thomson Reuters Corporation. (2014). *Corporate Responsibility & Inclusion Report 2014*. Accessed: Mar. 15, 2015. [Online]. Available: https://www.thomsonreuters.comt/content/dam/ewp-m/documents/thomsonreuters/en/pdf/corporate-responsibility/2014-cri-report.pdf

[4] M. Zhou, "Gender difference in web search perceptions and behavior: Does it vary by task performance?" *Comput. Educ.*, vol. 78, pp. 174–184, Sep. 2014.

[5] C. McLaughlin, L. Bradley, G. Prentice, E.-J. Verner, and S. Loane, "Gender differences using online auctions within a generation Y sample: An application of the theory of planned behaviour," *J. Retailing Consum. Services*, vol. 56, pp. 1–13, Sep. 2020.

[6] Z. Huang and J. Mou, "Gender differences in user perception of usability and performance of online travel agency websites," *Technol. Soc.*, vol. 66, Aug. 2021, Art. no. 101671.

[7] F. J. Pascual-Miguel, Á. F. Agudo-Peregrina, and J. Chaparro-Peláez, "Influences of gender and product type on online purchasing," *J. Bus. Res.*, vol. 68, no. 7, pp. 1550–1556, Jul. 2015.

[8] K. Wang, T. Zhang, T. Xue, Y. Lu, and S.-G. Na, "E-commerce personalized recommendation analysis by deeply-learned clustering," *J. Vis. Commun. Image Represent.*, vol. 71, Aug. 2020, Art. no. 102735.

[9] S. Thirumalai and K. K. Sinha, "Customization of the online purchase process in electronic retailing and customer satisfaction: An online field study," *J. Oper. Manage.*, vol. 29, no. 5, pp. 477–487, Jul. 2011.

[10] H. Y. Purwantono, A. A. S. Gunawan, H. Tolle, M. Attamimi, and W. Budiharto, "A literature review: Feasibility study of technology to improve shopping experience," *Proc. Comput. Sci.*, vol. 179, pp. 468–479, Jan. 2021.

[11] Y. Chen, X. Yan, W. Fan, and M. Gordon, "The joint moderating role of trust propensity and gender on consumers' online shopping behavior," *Comput. Hum. Behav.*, vol. 43, pp. 272–283, Feb. 2015.

[12] O. Sohaib, K. Kang, and M. Nurunnabi, "Gender-based iTrust in e-commerce: The moderating role of cognitive innovativeness," *Sustainability*, vol. 11, no. 1, p. 175, Dec. 2018.

[13] X. Lin, M. Featherman, S. L. Brooks, and N. Hajli, "Exploring gender differences in online consumer purchase decision making: An online product presentation perspective," *Inf. Syst. Frontiers*, vol. 21, no. 5, pp. 1187–1201, Oct. 2019.

[14] Z. Liu, L. Wang, X. Li, and S. Pang, "A multi-attribute personalized recommendation method for manufacturing service composition with combining collaborative filtering and genetic algorithm," *J. Manuf. Syst.*, vol. 58, pp. 348–364, Jan. 2021.

[15] R. V. Karthik and S. Ganapathy, "A fuzzy recommendation system for predicting the customers interests using sentiment analysis and ontology in e-commerce," *Appl. Soft Comput.*, vol. 108, Sep. 2021, Art. no. 107396.

[16] B. A. Hammou, A. A. Lahcen, and S. Mouline, "APRA: An approximate parallel recommendation algorithm for big data," *Knowl.-Based Syst.*, vol. 157, pp. 10–19, Oct. 2018.

[17] I.-C. Wu and H.-K. Yu, "Sequential analysis and clustering to investigate users' online shopping behaviors based on need-states," *Inf. Process. Manage.*, vol. 57, no. 6, Nov. 2020, Art. no. 102323.

[18] B. Liu and L. Wei, "Machine gaze in online behavioral targeting: The effects of algorithmic 'human likeness' on social presence and social influence," *Comput. Hum. Behav.*, vol. 124, Nov. 2021, Art. no. 106926.

[19] M. Eisend, "Gender roles," *J. Advertising*, vol. 48, no. 1, pp. 72–80, 2019.

[20] H. Kraft and J. M. Weber, "A look at gender differences and marketing implications," *Int. J. Bus. Social Sci.*, vol. 3, no. 21, pp. 1–7, 2012.

[21] S. Bettany, S. Dobscha, L. O'Malley, and A. Prothero, "Moving beyond binary opposition: Exploring the tapestry of gender in consumer research and marketing," *Marketing Theory*, vol. 10, no. 1, pp. 3–28, Mar. 2010.

[22] C. J. Auster and C. S. Mansbach, "The gender marketing of toys: An analysis of color and type of toy on the Disney store website," *Sex Roles*, vol. 67, nos. 7–8, pp. 375–388, Oct. 2012.

[23] V. Amawate and M. Deb, "Antecedents and consequences of consumer skepticism toward cause-related marketing: Gender as moderator and attitude as mediator," *J. Marketing Commun.*, vol. 27, no. 1, pp. 31–52, Jan. 2021.

[24] M. Karpinska-Krakowiak, "Women are more likely to buy unknown brands than men: The effects of gender and known versus unknown brands on purchase intentions," *J. Retailing Consum. Services*, vol. 58, Jan. 2021, Art. no. 102273.

[25] A. Bezirgani and U. Lachapelle, "Online grocery shopping for the elderly in Quebec, Canada: The role of mobility impediments and past online shopping experience," *Travel Behav. Soc.*, vol. 25, pp. 133–143, Oct. 2021.

[26] B. Melović, D. Šehović, V. Karadžić, M. Dabić, and D. Ćirović, "Determinants of Millennials' behavior in online shopping—Implications on consumers' satisfaction and e-business development," *Technol. Soc.*, vol. 65, May 2021, Art. no. 101561.

[27] L. Zhang, Z. Shao, X. Li, and Y. Feng, "Gamification and online impulse buying: The moderating effect of gender and age," *Int. J. Inf. Manage.*, vol. 61, Dec. 2021, Art. no. 102267.

[28] D. Fernandez-Lanvin, J. D. Andres-Suarez, M. Gonzalez-Rodriguez, and B. Pariente-Martinez, "The dimension of age and gender as user model demographic factors for automatic personalization in e-commerce sites," *Comput. Standards Interfaces*, vol. 59, pp. 1–9, Aug. 2018.

[29] C.-W. Liu, A.-Y. Hsieh, S.-K. Lo, and Y. Hwang, "What consumers see when time is running out: Consumers' browsing behaviors on online shopping websites when under time pressure," *Comput. Hum. Behav.*, vol. 70, pp. 391–397, May 2017.

[30] A. Rangaswamy, N. Moch, C. Felten, G. van Bruggen, J. E. Wieringa, and J. Wirtz, "The role of marketing in digital business platforms," *J. Interact. Marketing*, vol. 51, pp. 72–90, Aug. 2020.

[31] G. Aiello, R. Donvito, D. Acuti, L. Grazzini, V. Mazzoli, V. Vannucci, and G. Viglia, "Customers' willingness to disclose personal information throughout the customer purchase journey in retailing: The role of perceived warmth," *J. Retailing*, vol. 96, no. 4, pp. 490–506, Dec. 2020.

[32] C. Van Slyke, C. L. Comunale, and F. Belanger, "Gender differences in perceptions of web-based shopping," *Commun. ACM*, vol. 45, no. 8, pp. 82–86, Aug. 2002.

[33] J. Ge and L. Chen, "The obligation to provide 'non-personalised' search results under the Chinese e-commerce law," *Comput. Law Secur. Rev.*, vol. 41, Jul. 2021, Art. no. 105568.

[34] K. Li, L. Cheng, and C.-I. Teng, "Voluntary sharing and mandatory provision: Private information disclosure on social networking sites," *Inf. Process. Manage.*, vol. 57, no. 1, Jan. 2020, Art. no. 102128.

[35] S. Urbonavicius, M. Degutis, I. Zimaitis, V. Kaduskeviciute, and V. Skare, "From social networking to willingness to disclose personal data when shopping online: Modelling in the context of social exchange theory," *J. Bus. Res.*, vol. 136, pp. 76–85, Nov. 2021.

[36] J. Zhou, J. Wei, and B. Xu, "Customer segmentation by web content mining," *J. Retailing Consum. Services*, vol. 61, Jul. 2021, Art. no. 102588.

[37] M. Wan, D. Wang, M. Goldman, M. Taddy, J. Rao, J. Liu, D. Lymberopoulos, and J. McAuley, "Modeling consumer preferences and price sensitivities from large-scale grocery shopping transaction logs," in *Proc. 26th Int. Conf. World Wide Web*, Perth, WA, Australia, Apr. 2017, pp. 1103–1112.

[38] Z. Zhou, L. Shangguan, X. Zheng, L. Yang, and Y. Liu, "Design and implementation of an RFID-based customer shopping behavior mining system," *IEEE/ACM Trans. Netw.*, vol. 25, no. 4, pp. 2405–2418, Apr. 2017.

[39] M. Kolotylo-Kulkarni, W. Xia, and G. Dhillon, "Information disclosure in e-commerce: A systematic review and agenda for future research," *J. Bus. Res.*, vol. 126, pp. 221–238, Mar. 2021.

[40] R. Bandara, M. Fernando, and S. Akter, "Explicating the privacy paradox: A qualitative inquiry of online shopping consumers," *J. Retailing Consum. Services*, vol. 52, Jan. 2020, Art. no. 101947.

[41] K. Z. K. Zhang, C. M. K. Cheung, and M. K. O. Lee, "Examining the moderating effect of inconsistent reviews and its gender differences on consumers' online shopping decision," *Int. J. Inf. Manage.*, vol. 34, no. 2, pp. 89–98, Apr. 2014.

[42] R. Wakefield, "The influence of user affect in online information disclosure," *J. Strategic Inf. Syst.*, vol. 22, no. 2, pp. 157–174, Jun. 2013.

[43] V. Venkatesh, H. Hoehle, J. A. Aloysius, and H. R. Nikkhah, "Being at the cutting edge of online shopping: Role of recommendations and discounts on privacy perceptions," *Comput. Hum. Behav.*, vol. 121, Aug. 2021, Art. no. 106785.

[44] K. Ariansyah, E. R. E. Sirait, B. A. Nugroho, and M. Suryanegara, "Drivers of and barriers to e-commerce adoption in Indonesia: Individuals' perspectives and the implications," *Telecommun. Policy*, vol. 45, no. 8, Sep. 2021, Art. no. 102219.

[45] R. A. Abumalloh, O. Ibrahim, and M. Nilashi, "Loyalty of young female Arabic customers towards recommendation agents: A new model for B2C E-commerce," *Technol. Soc.*, vol. 61, May 2020, Art. no. 101253.

[46] S. Molinillo, R. Aguilar-Illescas, R. Anaya-Sánchez, and F. Liébana-Cabanillas, "Social commerce website design, perceived value and loyalty behavior intentions: The moderating roles of gender, age and frequency of use," *J. Retailing Consum. Services*, vol. 63, Nov. 2021, Art. no. 102404.

[47] L. Ni, H. Lin, M. Zhang, and J. Zhang, "Hybrid filtrations recommendation system based on privacy preserving in edge computing," *Proc. Comput. Sci.*, vol. 129, pp. 407–409, Jan. 2018.

[48] N. Ameen, S. Hosany, and J. Paul, "The personalisation-privacy paradox: Consumer interaction with smart technologies and shopping mall loyalty," *Comput. Hum. Behav.*, vol. 126, Jan. 2022, Art. no. 106976.

[49] S. Basu, "Personalized product recommendations and firm performance," *Electron. Commerce Res. Appl.*, vol. 48, Jul./Aug. 2021, Art. no. 101074.

[50] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[51] L. Breiman, J. Friedman, R. Olshen, and C. Stone, "Classification and regression trees. Belmont, CA: Wadsworth international group," *Encyclopedia Ecol.*, vol. 57, no. 1, pp. 582–588, 2015.

[52] H. Liu and M. Cocea, "Induction of classification rules by Gini-index based rule generation," *Inf. Sci.*, vol. 436, pp. 227–246, Apr. 2018.

[53] S. N. Bowe and J. A. Villwock, "Does gender impact personality traits in female versus male otolaryngology residents and faculty?" *Amer. J. Surg.*, vol. 220, no. 5, pp. 1213–1218, Nov. 2020.

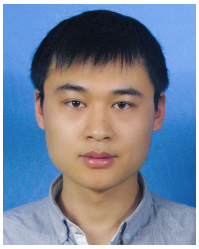
[54] J. Anglim, V. Sojo, L. J. Ashford, A. Newman, and A. Marty, "Predicting employee attitudes to workplace diversity from personality, values, and cognitive ability," *J. Res. Pers.*, vol. 83, Dec. 2019, Art. no. 103865.

[55] O. S. Itani, R. E. Haddad, and A. Kalra, "Exploring the role of extrovert-introvert customers' personality prototype as a driver of customer engagement: Does relationship duration matter?" *J. Retailing Consum. Services*, vol. 53, Mar. 2020, Art. no. 101980.

[56] A. Feizollah, N. B. Anuar, R. Salleh, and F. Amalina, ''Comparative study of k-means and mini batch k-means clustering algorithms in Android malware detection using network traffic analysis,'' in *Proc. Int. Symp. Biometrics Secur. Technol. (ISBAST)*, Kuala Lumpur, Malaysia, Aug. 2014, pp. 193–197.

[57] S. Fekri-Ershad, ''Gender classification in human face images for smart phone applications based on local texture information and evaluated Kullback–Leibler divergence,'' *Traitement Signal*, vol. 36, no. 6, pp. 507–514, Dec. 2019.

[58] K. Karkkainen and J. Joo, ''FairFace: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation,'' in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 1548–1558.



YUAN AN received the B.S. degree in computer science and technology from Liaocheng University, China, in 2006, and the M.S. degree in computer software and theory from Northeast Normal University, China, in 2013.

He is currently working with the School of Computer Science and Information Technology, Daqing Normal University, Daqing, China. His research interests include data mining and digital image processing.

SIQIAO MENG received the B.S. degree in accountancy from the Harbin University of Science and Technology, China, in 2020. She is currently pursuing the M.S. degree in teaching and learning with the City University of Macau, China.

HAO WU received the Ph.D. degree in management science and engineering from the School of Economics and Management, China University of Petroleum, Qingdao, China, in 2021. He is currently working with the School of Economics and Management, Zhejiang University of Science and Technology, Hangzhou, China.