2021 20th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES)

Gender Classification for Online Shopping

Yimeng Zhang
Taiyuan No. 20 High School
Taiyuan, China
E-mail: snowblanche@sina.com

Abstract— **The proliferation of Online shopping has been increasing in the past decades. Different online shopping companies investigate on precise shopping recommendation system based on the customers online viewing log and purchase log data. Even though the online shopping recommendation has been investigated for several years, both industrial and academia could not propose a generalized and efficient model to predict customers' shopping demand. Recently, the customers' gender information attract people's attention since the gender information reflects the customers' shopping behavior and preference. Nevertheless, the gender information collected from online shopping system are neither intact nor fake since customers don't want to leak their privacy. Hence, the estimation of customers' gender becomes critical for the online shopping recommendation system. This paper focuses on gender estimation based on customers' online viewing log collected by the FTP group, a leading information and communication enterprise in Vietnam. Given the imbalanced (population of female is 3 times of male) and ambiguous data, we propose our approach to estimate the gender with 75% accuracy. Specifically, we observe that the female samples naturally form 3 clusters when we select duration of session, number of items viewed, and average time spent on each item as the features. Then, we naturally divide the female set into 3 subsets and merge them with male set to generate the 3 training sets, which don't have imbalance issue. 3 individual models are trained from these 3 training sets and a new classifier is used to make the final decision based on the output of these 3 models. Our experimental results show that we can achieve 75% accuracy while the running time is less than 7 seconds.**

Data Mining; PAKDD Contest; Classifier; Cluster; Imbalance.

I. INTRODUCTION

Online shopping has already become a part of our daily life. Tons of online shopping information including discount, sale and advertisement are sent to our email as well post mailbox every day. The competition among different online shopping companies is getting more and more severe. Hence, the companies providing more convenient shopping experience to customers would become the winner. Hence, these companies spend a lot of money on investigating the customers' shopping behavior and preference to make a better shopping recommendation system, which can provide customers the most needed items. Nevertheless, these companies realize the customers' gender information is very critical in the online shopping recommendation system [1].

Then, a natural question is popped up: can we get customers' gender information? Unfortunately, the answer is no [2]. The reason is easy to understand. We all have the experience to register on some website, where we are asked to select our gender. However, most people ignore that option or randomly select the gender. People don't care about the gender information because (i) they don't want to leak their privacy to the website since they might just purchase the items on this website once [3]. (ii), they don't want to waste time on make the selection. hence, the gender information collected from the website is far less than enough to contribute to the online shopping recommendation system [4]. Therefore, the estimation of customers' gender becomes necessary.

This paper proposes an approach to estimate the gender of customers shopping on FTP group, a leading information and communication enterprise in Vietnam. FPT runs several B2B2C (business-to-business-to-customer) services that provide online shopping sites and mobile applications for small and medium sellers. Transaction data, such as product browsing and purchasing activities, from buyer, and product portfolio, from seller, can be aggregated, to provide more efficient buying and selling experiences. Data mining techniques are applied to predict the optimal organization and display of products that maximize the chance of bringing useful information to user, facilitate the online purchases. Given the online viewing log, we estimate the customers' gender for the FTP group to make a better online recommendation system.

The data we have from FTP group is demonstrated in Table 1, which shows 2 samples in the training data set. Session ID is the ID of each view session while "start" and "end" represent the start time and end time of this session. The "Viewed Items" stores all the items the user views in this session and the "Class Label" is the corresponding gender. More specifically, items are classified to 4 categories: A represent the most generalized items and those starting with 'D' correspond to individual products. The IDs which start with 'B' and 'C' are associated with subcategories and sub-subcategories, respectively. The format of test data set is as same as training data set except for the class label. Both test and training data set contain 15000 samples. we get these data from PAKDD 2015 contest website. The size of training and testing file are 224 and 223KB respectively. We also get the correct class label for the test set. Then, it is very convenient for us to calculate the accuracy of our own result and compare our approach with other public results in the PAKDD 2015 contest ranking list.

As we can see from the Table 1, the relationship between 4 columns and class label are too loose, we need to extract meaningful feature from the data set. The difficulty here is how to define the appropriate features that provides the maximum estimation accuracy. This requires the deep understanding of online shopping

2473-3636/21/$31.00 ©2021 IEEE
DOI 10.1109/DCABES52998.2021.00017

Session Id	Start	End	Viewed Items	Class Label
u10001	11/14/2014 0:02	11/14/2014 0:02	A00001/B00001/C00001/D00001/	Female
u10002	12/12/2014 14:12	12/12/2014 14:12	A00002/B00002/C00002/D24897/	Male

Table 1, Data format.

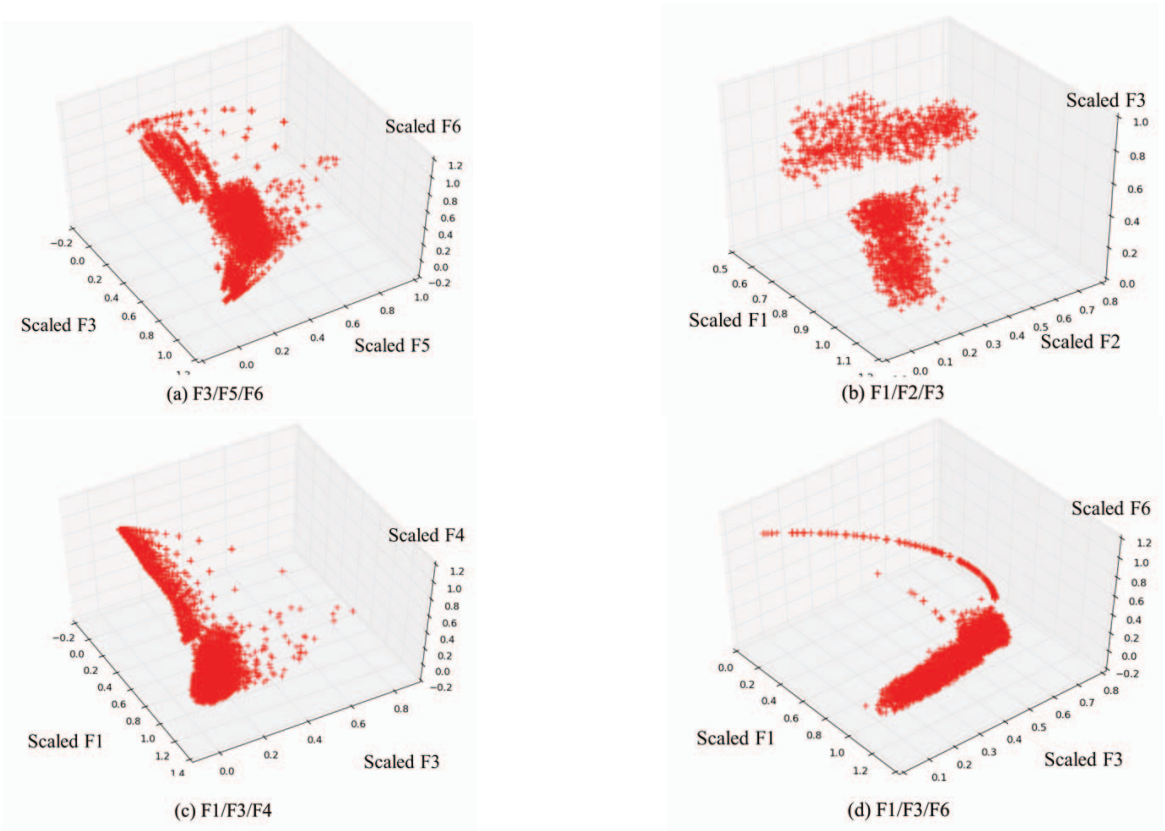

Figure 1, Data visualization for different feature combinations.

behavior and psychologic analysis. Another issue in the dataset is imbalance issue. Training set includes 11703 Female and 3297 Male samples. This imbalanced training dataset causes poor estimation result, which would be shown in section 2.1.

Even though the population of female samples is much larger than the population of male samples, we know that females have different personalities. For example, I am a girl while my personality is not like a typical girl. On contrast, my personality is close to a boy. Then, we can divide the female set into several (3) subsets according to personality. By selecting the feature, the subsets have the similar population which is almost equal to number of male samples. Hence, the imbalance issue is solved. The details are demonstrated in section 2.2 and 2.3.

Our contribution has two folds:

- We extract 6 features from the online viewing log. By applying data visualization on different feature combinations, we select 10 combinations as the feature candidates. We take advantage of personality diversity to further filter out the feature combination candidates and divide the female training set into 3 subsets. Then, we generate 3 balanced sub training sets. Basically, we address the imbalance issue in this dataset.
- We the proposed approach with different classifiers and achieve 75% accuracy while the running time is as short as 7 seconds.

II. DESIGN

In this section, we demonstrate our approach in detail.

A. Feature Definition

We use the sk to denote the k-th session. The start and end of k-th session are represented by ts(sk) and te(sk). The corresponding items are stored in the item set: Item(sk)={ item(sk)1, item(sk)2,… item(sk)i,…}. We have 6 candidate features and their definitions are as follows:

Definition 1: duration of session. Duration of the k-th session, Du(sk)= te(sk)- ts(sk).

Definition 2: number of items viewed. The number of items viewed in the k-th session, N(sk)= | Item(sk)|.

Definition 3: average time spent on each item. The average time spent on each item in the k-th session, Tavg(sk)= Du(sk)/ N(sk).

Definition 4: start time of the session. The time the session is opened, T_s(sk)= ts(sk)

Definition 5: Number of the most generalized (‘A’ category) items. NA(sk)= |{ item(sk)j| item(sk)j starts with “A”}|

Definition 6: Number of the most individual ('D' category) items. ND(sk)= |{ item(sk)j| item(sk)j starts with "D"}|

Let us use an example to show the intuition of these features. As a male, I usually have a list before I shop online. Then, when I view the items, I prefer staying on each webpage for a while because I would like to get more information. So, I think the average time spent on each item for male might be longer than female. The start time of the session maybe important too. Because I would never touch the shopping websites during the daytime. The number of most generalized/individual items shows the customers' shopping behavior and preference.

	F3/F5/F6	F1/F2/F3	F1/F3/F4	F1/F3/F6	F1/F3/F5
Accuracy	0.65	0.63	0.61	0.55	0.54
	F2/F3	F2/F3/F4	F2/F4/F6	F1/F2/F4	F1/F4/F5
Accuracy	0.54	0.53	0.51	0.51	0.50

Table 2, accuracy for different feature combination.

B. *Feature Selection*

Given the 6 featured defined in section 2.1, we need to decide which features are used in our approach. At first, we run the random forest classifier on the combination of these 6 features, which means we need to test 26=64 combinations. Since the space is limited, we show the accuracy of top 10 combinations in table 2. Since these 10 feature combinations provide more than 50% accuracy, all of them are candidates.

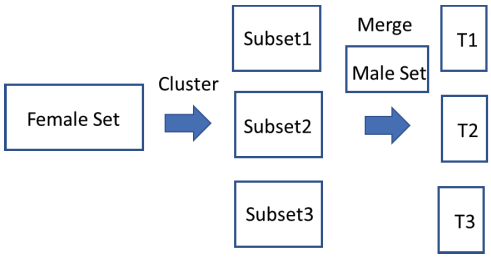

Figure 2, Training set generation.

C. *Clustering based Data Set Cut*

We know that the diversity of personalities exists in all the communities. The female samples should also reflect the personality diversity. We assume the diversity naturally forms several clusters in the female samples. In order to verify this assumption, we plot the female samples defined by top 4 feature combinations in figure 1.

Fortunately, the second highest accuracy combination, i.e., F1/F2/F3, depicts the clear 3 clusters. Since we know the accuracy of these combination is affected by the imbalanced samples, the highest accuracy combination without showing the personality diversity should not be the appropriate the feature combination. Therefore, we select the feature combination F1/F2/F3 as our feature, which shows the clear personality diversity.

Then, we use k-means method to find the 3 clusters. The number of samples in these 3 clusters are 3108, 4925 and 3670 respectively. They are almost balance. Then, as shown in figure 2, we generate 3 training set by merging these 3 female clusters (subsets) with male set. We cut the female data set by clustering method, which smartly solves the imbalance issue.

D. *Training Model*

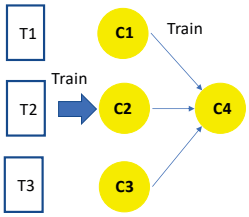

Figure 3, Training Model.

Given the training set generated in section 2.3, we demonstrate our model in this section. As figure 3 shows, we train the 3 model individually on each training set. Then we get classifier C1, C2 and C3. When we estimate the gender for a sample in test set, we input this sample into 3 classifier and get 3 corresponding results. For example, the results are "Female", "Male" and "Male" for C1, C2 and C3 respectively. Then, the question is which one is correct and how to make the final decision? The basic idea is to vote on these 3 results. However, using majority voting doesn't provide any sense. Alternatively, we use another classifier, C4, to smartly make the final decision. We use the output of C1, C2 and C3 as the input and use the true class label as the output to train C4.

The intuition of this approach is the personality of all the females are different and can be classified into several classes. Then, we use these subclasses to generate different training model. When we estimate the gender for the sample in test set, we firstly decide the gender for different female personality. For example, if the sample is a female sample, this stage would decide which personality it belongs to. If this sample is a male sample, this stage would provide the correct result with high confidence. Hence, we decide the personality of this sample first and we decide the gender at the second stage (using C4).

E. *Putting Together*

The whole approach consists of the 3 following steps:

• Step 1, Data cleaning and Feature extraction. Since we select F1/F2/F3 combination as our feature, we read the training set file and check if there exist any samples doesn't have intact information. In other words, we check if some sample doesn't have all the attributes. We find that there are 172 samples don't have all the 4 attributes. This is very normal in the practical data set. Since the number of the incomplete samples (172) is much smaller than the whole number of samples in training set, we can just remove these samples from the training set. Then, F1, F2 and F3 are calculated as their definitions in section 2.1.

• Step 2. We use K-means to find the 3 clusters in female set according to the personality diversity. Then, we merge these 3 clusters with the male set to generate the 3 training sets t1, t2 and T3.

• Step 3. We train 3 models, C1, C2 and C3 on 3 training sets individually. Since our approach is independent of the classifiers, we may use SVM, random forest, decision tree as the classifier. We show the accuracy of the different combination in the section 3. For example, we set C1 to be SVM, C2 to be random forest, C3 to be SVM and C4 to be decision tree. Then, we train the 4 classifiers and compare the result.

Eventually, we select the combination of classifiers with the highest accuracy as our final approach.

III. EVALUATION

We demonstrate the evaluation of our training model in this section. The ground true is already published at PAKDD 2015 competition website. Thus, we know the class label of test data set. Since the data set is imbalanced, we use the "balanced accuracy" to evaluate our approach. The definition of balanced accuracy is BAC (predict, gender)=(ACCm(predict,gender)+ACCf (predict,gender))/2, where ACCm ¬and ACCf are the accuracy of predicted male and female respectively. "predict" and "gender" represent our predicted result and ground truth. The formal definition of ACCm and ACCf are as follows:

$$ACC_m(predict, gender) = \frac{|j: predict_j = genrder_j = Male|}{|j: genrder_j = Male|}$$

$$ACC_f(predict, gender) = \frac{|j: predict_j = genrder_j = Female|}{|j: genrder_j = Female|}$$

A. Feature Selection

We select clustering features among the six features that most closely match the sample numbers of the female and male sets. F1 / F2 / F3 can classify a female sample set into three groups with the most similar number of samples. Through F1, F2, and F3, clusters 1,2,3 have 3108, 4925, and 3670 samples, respectively. On the other hand, the other feature combinations are more biased in a particular cluster, resulting in a large number of samples

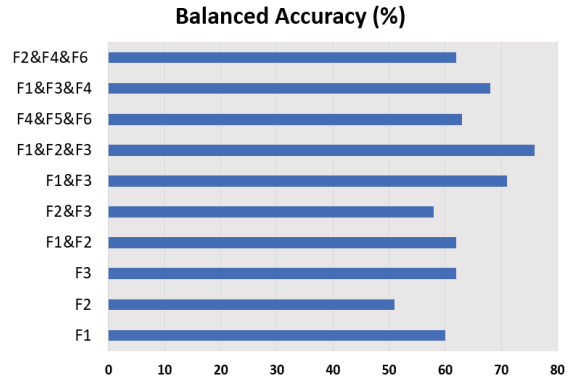

Figure 4, Balanced Accuracy with different feature set. The accuracy depends on the feature set.

B. Balanced Accuracy

We achieve 75% accuracy using Feature 1, 2, and 3. We set the training and test set from PAKDD as ground truth and verify the prediction results of our model. When verifying the accuracy of our model, we measure the accuracy based on the balanced accuracy. This is because our dataset is imbalanced, so if you do it by measuring the existing accuracy, you will get a bias result on the female set. To avoid this, we measure the accuracy through balanced accuracy. Basically, the balanced accuracy measures the predicted results for male and female, and then averages them.

C. Computational Overhead (Running Time)

In order to apply our approach into practical and commercial scenario, the computational overhead is a critical factor since our approach needed to be run for thousands of times on the large-scale data sets. In addition, the running time in this experiment include both feature extraction and training process. We know the online shopping data is dynamically changing every day or every second. If the company uses our approach for the different data collected at different time, the time spent on feature extraction should be considered. Hence, we evaluate the overhead (running time) of different classifier and feature combination. The corresponding results are summarized in table 3.

	Random Forest	SVM	Decision Tree	Gaussia nNB
F1/F2/F3	5s	6s	8s	6s
F4/F5/F6	6s	7s	7s	5s
F1/F3/F4	5s	8s	9s	4s
F1/F3/F6	5s	7s	12s	5s
F1/F3/F5	5s	7s	8s	6s
F2/F3/F4	6s	6s	7s	6s
F1/F2/F4	5s	9s	8s	7s
F1/F4/F5	6s	8s	9s	6s

Table 3, overhead for different classifiers.

The trend in this result is: the larger number of features we use, the more time is consumed. This is reasonable since the feature extraction involves I/Q operation and the more features we need extract, the more time we need to process the file. Random forest consumes the least time because we can control the depth of the tree in the random forest algorithm while the accuracy is not sacrificed. Even though the running time is important, the accuracy is more important. If the approach could not give us a good result, less running time is meaningless.

IV. CONCLUSION

This paper focuses on mining the gender information from the online shopping viewing log provided by FTP company. Given the meaningless raw attributes, we extract 6 meaningful features from the data set. In order to overcome the Female/Male imbalance issue, we take advantage of personality diversity to divide the female set into 3 subsets with similar size. Then, a 2-layer classifiers network is applied to estimate the gender. Our evaluation results show that our model of gender prediction achieves balanced accuracy of 75% while running time is 7 seconds. Even though the winner in PAKDD 2015 contest achieved 87% balanced accuracy, our approach is more light weight and efficient because the winner's method is too complicated.

REFERENCES

[1] Tomson Reuters, Article Title, https://blogs.thomsonreuters.com/answerson/ business-case-gender-parity/.

[2] Van Slyke, Craig, Christie L. Comunale, and France Belanger. "Gender differences in perceptions of web-based shopping." Communications of the ACM 45.8 (2002): 82-86.

[3] Lin, Xiaolin, et al. "Exploring Gender Differences in Online Consumer Purchase Decision Making: An Online Product Presentation Perspective." Information Systems Frontiers (2018): 1-15.

[4] Kim KJ, Ahn H. A recommender system using GA K-means clustering in an online shopping market. Expert systems with applications. 2008 Feb 1;34(2):1200-9