A Resamping Approach for Customer Gender Prediction Based on E-Commerce Data

Duong Tran Duc, Pham Bao Son, Tan Hanh, and Le Truong Thien

Abstract—Demographic attributes of customers such as gender, age, etc. provide the important information for e-commerce service providers in marketing, personalization of web applications. However, the online customers often do not provide this kind of information due to the privacy issues and other reasons. In this paper, we proposed a method for predicting the gender of customers based on their catalog viewing data on e-commerce systems, such as the date and time of access, the products viewed, etc. The main idea is that we extract the features from catalog viewing information and employ the classification methods to predict the gender of the viewers. The experiments were conducted on the datasets provided by the PAKDD'15 Data Mining Competition and obtained the promising results with a simple feature design, especially with the Bayesian Network method along with other supporting techniques such as resampling, cost-sensitive learning, boosting etc.

Index Terms—gender prediction; machine learning; big data; e-commerce; bayesian network.

1. Introduction

Today, many web applications such as e-commerce systems, search engine, online marketing, employ the personalization features to increase user experience. With a good personalized service, the information displayed is customized for each user individually rather than remaining the same for all users. For example, the e-commerce systems can display the promotions or recommend the products which are relevant to the individual visitor rather than the random promotions or products.

The personalization is mainly based on two types of data: historical data (e.g. previous items viewed or purchased) and the demographic attributes of users (e.g. gender, age, education etc.). The historical data can be obtained only if the user has used the system before and has logged in to the system. Therefore, the historical data-based methods are unusable for guests or the new users. The demographic-based methods are useful even when the user has never used the system before. However, this information is not easy to obtain, because the Internet users are often not willing to provide the private information. For that reason, in many cases, the only way to obtain the demographic attributes of users is to predict. The task of author profiling for the anonymous texts are studied for decades [1,2,3], [6], [11], [13], but not all user write something on the system. Another way to predict the demographic information of users is based on their behaviors on the systems, for example the browsing activities, or catalog viewing data [5], [12]. The main advantage of this approach is that the data is available in most cases, because the users must do something on the system such as access the pages, click the items, or browse the catalog.

In this research, we address the problem of predicting the demographic information of users based on their catalog viewing data such as the viewing time/duration, viewed products/categories, etc. We used the popular learning methods such as Support Vector Machine (SVM), Bayesian Network, and Decision Tree to train and test on the datasets provided by FPT Corporation in the PAKDD'15 Data Mining Competition. We also focused on the use of the supporting techniques for dealing with the class-imbalance problem such as resampling, cost-sensitive learning, boosting to improve the accuracy. The results are promising although more types of features need to be investigated to improve the performance.

The organization of the paper is as follows. In section 2, we present the related work on user demographic prediction. Section 3 describes the methods and the system. Section 4 presents the result and discussion. In the section 5, we draw a conclusion and future work.

2. Related Work

Demographic prediction has been studied for a long time. At the early stage, most of researches on this field focused on the authorship studies, which are the tasks of determining or predicting the author characteristics by analyzing the text created by him/her. The methods which the researchers used in these studies are mostly based on the analysis of writing style using various types of features, such as lexical, syntactic, or content-based features. The first study in this field dates back to 19th century when Mendenhall [10] investigated the Shakespeare's plays. But the work which was considered the most thorough study in this field was conducted by Mosteller and Wallace [11] when they

Duong Tran Duc, Tan Hanh are with Posts and Telecommunications Institute of Technology; {ducdt, tanhanh}@ptit.edu.vn.

Pham Bao Son is with University of Engineering and Technology, Vietnam National University, Hanoi; sonpb@vnu.edu.vn.

Manuscript received July 9, 2016; revised October 2 and December 27, 2016; accepted February 10, 2017.

analyzed the authorship of FederalList Papers based on Bayesian statistical analysis of the frequencies of function words. The previous authorship studies often focused on the literature text, such as novel or article. Recently, due to the growth of Internet and the online communication channels, the focus has been moved to the computer mediated communication contents, such as email, blogs, comments, etc. De Vel et al. [3] used 221 features to determine the authorship of emails. Argamon and Argamon et al. [1] investigated the differences in writing style of male and female in the 604 documents from British National Corpus. Schler et al. [15] explored the use of stylometric and content-based features to predict the gender and age of bloggers on the datasets with over 71.000 blog posts from blogger.com. This model achieved the results 80% for gender prediction and 76% for age prediction. Iqbal [6] proposed a method to calculate a ”write print” based on the frequent patterns extracted from the emails to predict to profile of the author. Nguyen et al. [13] conducted a research to predict the gender and age of twitter messages and forum posts using the regression methods with the accuracy around 80%.

Beside the methods based on the analysis of textual data, recently, the researchers investigated the use of user behaviors on the web applications to predict their demographic information. Hu et al. [5] proposed a method to solve the problem of predicting the Internet users' gender and age based on their browsing behaviors. They used the webpage view information as input variables to propagate demographic information of users. The SVM method was employed on the features set consisting of content-based (words from the Webpages) and category-based (hierarchy of web concepts) features and achieved the results of 79.7% on gender and 60.3% on age. Kabbur et al. [7] also investigated the machine learning approaches to predict the demographic attributes of websites using the information from the content and hyperlinked structure. The research of Dong et al. [4] aimed at inferring the users' demographics based on their daily mobile communication patterns. Their study was conducted on a real-world large mobile network of more than 7.000.000 users and over 1.000.000.000 communication records. They used the features set including individual features, friend features, and circle features and achieved the best results of 80% for gender and 70% for age. Ying et al. [17] proposed a prediction framework for predicting users' demographics based on the analysis of their behaviors and environments. They also developed a new method namely Multi-Level Classification Model to solve the imbalanced class problem. Phuong et al. [12] also addressed the problem of predicting the users' gender based on browsing history. They employed the classification-based methods and used the features derived from browsing log data. They added more types of features than the previous works, such as topic-based features, time features, sequential features, and improved results significantly.

In this report, we investigated the use of machine learning methods along with other techniques such as resampling, boosting, cost sensitive learning, etc. to address the problem of predicting the users' gender on the e-commerce system based on the catalog viewing data. The model was trained and tested on the datasets provided in the PAKDD'15 Data Mining Competition and achieved the promising result of 80.8% on the separate test set.

3. Approach

3.1. System Overview

In this work, we developed a system which can take the data from the product viewing logs for the users with known gender, extract the features and class labels to create a training dataset. A model is built from the training dataset using a classification-based method and then can be used to predict the gender of unknown users based on their product viewing activities.

The training data file contains the records which correspond to product viewing logs. A single log contains the information about the product viewing data of a user, such as session start time, session end time, list of products and categories IDs. The class labels for each training sample are male and female. Therefore, the task is a binary classification problem with two labels correspondingly.

In the next sections, we describe the features and the techniques which were used for prediction in detail.

3.2. Features

The feature set we used in this work can be divided into two types: temporal and product-based features.

Temporal features include the features related to timestamp and frequency of viewing activities. The time in day, day of week, holidays, viewing duration, number of products viewed in one session, etc. are the factors that can be used to predict the gender of a customer. We used totally 8 features of this kind as shown in the Table 1.

TABLE 1: The temporal features

Feature name	Description
Day	The day in month (1 to 31)
Month	The month in year (1 to 12)
DayOfWeek	The day in week (Monday to Sunday)
StartTime	Start time of the session
EndTime	End time of the session
Duration	Length of the session
N_Product	Number of products viewed in the session
AverageTime	Average time for viewing one product

Product-based features consist of all the features related to products and categories. In this work, we proposed a simple but effective approach for this kind of features. In each session, user may view multiple products, but many of them view one product only. For the session which contains more than one product viewed, we broke the session into multiple sessions which contain only one product. This can be considered as a resampling approach, in which we create new instances for the dataset by detaching the sessions which contain more than one products to multiple instances which contain only one product. By that way, we created a new dataset in which all the samples have information about only a single product. We built a model for this dataset, and used it to predict the output for all the input data which contains one product. For the input data which contains more than one product, we predict the result for each product separately, and then combine them to calculate the final result. The combination method is also simple: if the number of males predicted is equal or more than number of females, then the final result will be male. Otherwise, the final result is female.

By using this approach, the design for product-related features is quite simple. We used only four features of this type. The Table 2 shows the summary of product-based features which were used in our experiments.

TABLE 2: Summary of product-based features

Feature name	Description
CategoryA_ID	ID of the most general category
CategoryB_ID	ID of the second level category
CategoryC_ID	ID of the third level category
Product_ID	ID of the individual product

Besides the basic product-based features, we also used the product transition features to exploit the relations between products viewed in the same session. For this purpose, we added the features related to the previous and the next product in the same session (if available).

TABLE 3: Summary of product transition features

Feature name	Description
P_CategoryA_ID	ID of the most general category of the previous product
P_CategoryB_ID	ID of the second level category of the previous product
P_CategoryC_ID	ID of the third level category of the previous product
P_Product_ID	ID of the individual product of the previous product
N_CategoryA_ID	ID of the most general category of the next product
N_CategoryB_ID	ID of the second level category of the next product
N_CategoryC_ID	ID of the third level category of the next product
N_Product_ID	ID of the individual product of the next product

3.3. Classification Methods

We used three machine learning algorithms SVM, Random Tree, and Bayesian Network to learn the model. However, because the training data is imbalanced (around 80% of females and 20% of males), we employed some supporting techniques such as Resampling, Cost-Sensitive Learning, Boosting to improve the accuracy.

Machine Learning Algorithms. Support Vector Machine is a learning method having an advantage that it does not require a reduction in the number of features to avoid the problem of over-fitting. This property is very useful when dealing with large dimensions as encountered in the area of text categorization. Bayesian Networks are probabilistic graphical models in which, each node in the graph represents a random variable and the edges between nodes represent probabilistic dependencies among the corresponding random variables. Bayesian Networks are used for modelling beliefs in a number of fields such as bioinformatics, document classification, financial and marketing informatics, etc. Decision Tree is also a commonly used classification method. A Decision Tree is a tree in which each internal (non-leaf) node is labeled with an input feature. The arcs coming from a node labeled with a feature are labeled with each of the possible values of the feature. Each leaf of the tree is labeled with a class or a probability distribution over the classes.

Resampling. Resampling methods are commonly used for dealing with class-imbalance problem. The basic idea is to add or remove some instances to make the dataset become more balanced. Therefore, there are two methods of resampling that are under-sampling (reduce the number of large class instances to close to the small class) and over-sampling (replication of small class instances to close to the large class). According to Kotsiantis et al. [8], the main drawback of under-sampling is that this method can discard the potential data which can be important for the task, while over-sampling may lead to additional computation cost and over-fitting problem in case of random replication. In our experiments, beside the resampling approach by detaching the sessions mentioned above, we also used the spread subsample, which is a random under-sampling method that allows to specify the maximum spread between the rarest and the most common class.

Cost-Sensitive Learning. While resampling is the data-level method, cost-sensitive learning is an algorithm-level method to solve the problem of class-imbalance classification [8]. According to Ling et al. [9], cost-sensitive learning is a method that takes the misclassification cost into the consideration, meaning it treats the different misclassifications differently. In this task, the distribution of the labels is not balanced, the assessment for the solution will be better by using the balanced accuracy measure (the average accuracy within the classes). This is also the evaluation method of the PAKDD'15 Data Mining Competition (will be mentioned in detail in the next section). Because the ratio of male over female class in the training dataset is 1:4, we chose the cost matrix as follow:

TABLE 4: Cost matrix for cost-sensitive learning

	Predict male	Predict female
True male	0.0	4.0
True female	1.0	0.0

Boosting. Boosting is a machine learning ensemble algorithm which is used to convert the weak learners to a strong learner. The boosting algorithm starts with a base learning algorithm and repeatedly feed it with a different distribution or weighting over the training examples. Each time it is called, the base learning algorithm generates a new prediction rule, and after some rounds, the boosting algorithm must combine these rules into a prediction model, which hopefully will be more accurate [16]. There are many boosting algorithms, in which AdaBoost (introduced in 1995 by Freund and Schapire) may be the most famous method. In this work, we used AdaBoost for improving the accuracy of our model.

4. Experiments

4.1. Data

We used the datasets from the PAKDD'15 Data Mining Competition which were provided by FPT Corporation. The data was divided into training and test sets. Each of set contains 15,000 records which correspond to the product viewing logs.

A single log in the training data file is composed of four types of information:

- Session ID
- Start time (including date)
- End time (including date)
- List of product IDs

The list of product IDs contains the IDs of the products which the user has viewed during the session. Because the products may belong to the different categories, the information about the categories is also included in the IDs. Each product ID can be decomposed into four different IDs, from the most general categories (start with "A") to the subcategories (start with "B" and "C") and individual product (start with "D") respectively. An example of a single log is as follow:

u10008, 2014-11-17 19:20:06, 2014-11-17 19:21:54, A00001/B00001/C00001/D00001/; A00001/B00002/C00002/D00002

4.2. Evaluation Metrics

As mentioned earlier, due to the class-imbalance problem, the balanced accuracy measure was used to evaluate the model. Balanced accuracy is defined as an average accuracy obtained on either class and can avoid inflated performance estimates on imbalanced datasets.

$$balanced\ accuracy = \frac{0.5*tp}{tp+fn} + \frac{0.5*tn}{tn+fp} \tag{1}$$

Where *tp* is true positives, *tn* is true negatives, *fp* is false positives, and *fn* is false negatives.

This measure was also used to evaluate the results in the PAKDD'15 Data Mining Competition.

In this work, we report this score along with the accuracy and macro F1 score (the average of F1 scores of male and female) for being convenient in comparing with previous works.

$$accuracy = \frac{tp+tn}{tp+fp+tn+fn} \tag{2}$$

$$F1 = \frac{2pr}{p+r} \tag{3}$$

Where p is precision (defined as the number of correctly predicted cases divided by the number of all predictions of a class) and r is recall (defined as the number of correctly predicted cases divided by the number of all cases of a class)

4.3. Results and Discussion

In order to evaluate the effects of supporting techniques used in this work, we conducted the experiments using each technique separately and all the techniques in combination. However, before all, we tested the model with various learning algorithms, such as Random Tree, Bayesian Network, and SVM, to find out the best one for this problem. Then we used this algorithm for testing on other aspects. The training data and testing datasets are provided separately (each dataset has 15.000 samples). Therefore, our model was created based on the training dataset and tested on a different dataset. Table 5-6 shows the results of our experiments.

TABLE 5: Results of experiments conducted on different algorithms (no supporting techiniques)

	Balanced Accuracy	Accuracy	F1
Random Tree	76.5	85.2	78.3
BayesNet	**77.6**	**85.9**	**78.9**
SVM	74.6	82.1	76.4

As the results shown in table 5, the Bayesian Network outperformed the other methods. Surprisingly, the SVM produced the worst result. SVM is a robust classification algorithm in many cases, including the demographic prediction. However, it may not suitable for this kind of feature design.

From the table 6, we can observe that all the supporting techniques have good effects in improving the performance of the base algorithm. Each technique increased the performance by 1-2%, while the combination of all them increased the performance by more than 3%. It is also interesting when observing that the performance measured by the other metrics such as normal accuracy or macro F1 is decreased or unchanged when applied the supporting techniques. This is reasonable because these techniques aimed at dealing with the class-imbalance problem, so they may improve the balanced accuracy only.

TABLE 6: Results of experiments conducted on the BayesNet with supporting techniques

	Balanced Accuracy	Accuracy	F1
No supporting techniques	77.6	85.9	78.9
Cost-sensitive only	78.7	84.5	78.6
AdaBoost only	78.9	84.2	78.3
Combination of all techniques (no product transition features)	79.8	83.2	78.4
Combination of all techniques and all features	**80.7**	**83.6**	**78.5**

Comparison with the earlier works. The baseline results of the demographic prediction task based on the text analysis is 80% for gender (accuracy). Therefore, the overall accuracy of gender prediction in our work can be considered promising. With the more similar works conducted by Hu et al. [5] and Phuong et al. [12], which predicted the users' gender based on their browsing data, the macro F1 score of our work is also good, although the browsing activities generate more meaningful data for training than product viewing activities.

In comparison with the other solutions from the

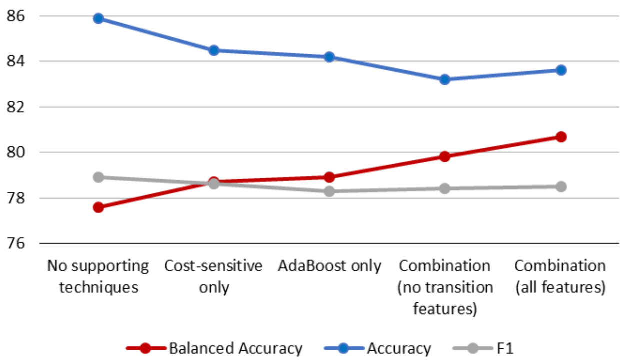

Fig. 1: Experiment results by different metrics

teams participating in PAKDD'15 Data Mining Competition, we achieved the 15^{th} position over 150 teams (we submitted the result without using transition features). The performance of the best team is 87.9%, and the top 10 positions achieved the results from 81%. However, our solution's advantage is that we used a simple feature structure, but still achieved the encourage result. In addition, we don't use any dataset-specific features such as product-prefixes, session alignment between sessions in training and test sets, etc. which were used by most of teams in the competition. Therefore, our method can be considered as a general solution which can be applied on any dataset.

TABLE 7: PAKDD'15 Data Mining Competition ranking for gender prediction

Position	Team	Result
1	frdc	0.8789
2	newolfy	0.8779
3	ws	0.8511
4	sohrab	0.851
5	kimiyoung	0.849
6	dymitrruta	0.848
7	ibayer	0.8407
8	songshuangyong	0.8306
9	stderr	0.8115
10	amy	0.8107
11	gambi	0.8102
12	siyu	0.8053
13	ahwangyuwei	0.8014
14	kkurach	0.7988
15	**duongtranduc**	**0.7978**

5. Conclusion and Future Work

In this study, we investigated a method for predicting the gender of customer based on the product viewing data on the e-commerce systems. We proposed an approach which used a resampling method and a simple features design on the Bayesian Network with various supporting techniques to deal with the problem of class-imbalance, such as cost sensitive learning, and boosting algorithms. The advantage of our resampling method and simple feature design is that it is easier to adapt to the new domain because it uses only basic and no dataset-specific features. Regarding the methodology, we focused on the supporting techniques which were used to improve the performance on the special cases, such as class-imbalance. As shown in the results, all the supporting techniques we employed have effects on the performance of the final prediction.

In the future, this work can be investigated further on the feature set. More type of features such as the particular categories or products should be taken into account to improve the performance, although it may be domain-dependent. We also have plan to collect data from other sources to improve the general performance and expand to other demographic attributes such as age, location, job, and so on.

References

[1] Argamon, S., Koppel, M., Fine, J. and Shimoni, A. (2003). Gender, Genre, and Writing Style in Formal Written Texts, Text 23(3), August

[2] Argamon, S., Koppel, M., Pennebaker, J.W. and Schler, J. (2009). Automatically profiling the author of an anonymous text.Communications of the ACM,52(2), pp.119-123.

[3] De Vel, O., Anderson, A., Corney, M., Mohay, G. M. (2001). Mining e-mail content for author identification forensics. SIGMOD Record 30(4), pp. 55-64

[4] Dong Y, Yang Y, Tang J, Yang Y, Chawla NV. (2014). Inferring User Demographics and Social Strategies in Mobile Social Networks. In: KDD'14. ACM. p. 15-24.

[5] Hu, J., Zeng, H.J., Li, H., Niu, C., Chen, Z. (2007). Demographic prediction based on user's browsing behavior. Proceedings of the 16th international conference on World Wide Web. Pages 151-160.

[6] Iqbal, F., Khan, L.A., Fung, B. and Debbabi, M. (2010). E-mail authorship verification for forensic investigation. InProceedings of the 2010 ACM Symposium on Applied Computing, pp. 1591-1598

[7] Kabbur, S., Han, E.H., Karypis, G., (2010). Content-based methods for predicting web-site demographic attributes. Proceedings of ICDM 2010.

[8] Kotsiantis, S., et al. (2006). Handling unbalanced datasets: A review, GESTS International Transactions on Computer Science and Engineering 30 (1), pp. 25-36

[9] Ling, C.X., Sheng, V.S. (2008). Cost-sensitive learning and the class imbalance problem. In: Sammut C (ed) Encyclopedia of machine learning. Springer, Berlin

[10] Mendenhall, T.C. (1887). The characteristic curves of composition. Science, 11(11), 237-249

[11] Mosteller, F., Wallace, D.L. (1964). Inference and disputed authorship: The Federalist. Reading, MA: Addison-Wesley

[12] Phuong, T.M., Phuong, D.V. (2014). Gender Prediction Using Browsing History. Proceedings of the Fifth International Conference KSE 2013, Volume 1. Pages 271-283.

[13] Nguyen, D., Gravel, R., Trieschnigg, D., and Meder, T. (2013). "How old do you think i am?"; a study of language and age in twitter. Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media

[14] Rangel, F., Rosso, P. (2013). Use of language and author profiling: Identification of gender and age. In Natural Language Processing and Cognitive Science, p. 177.

[15] Schler, J., Koppel, M., Argamon, S. and Pennebaker, J. (2006). Effects of Age and Gender on Blogging. In 43 proceedings of AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs

[16] Schapire, R.E. (2001). The Boosting Approach to Machine Learning: An Overview. Proc. MSRI Workshop Nonlinear Estimation and Classification.

[17] Ying, J.J.C., Chang, Y.J., Huang, C.M., Tseng, V.S. (2012). Demographic prediction based on users mobile behaviors. Mobile Data Challenge

[18] Zhang, C., Zhang, P. (2010). Predicting gender from blog posts. Technical report, Technical Report. University of Massachusetts Amherst, USA.

Duong Tran Duc received the Master degree from University of Leeds, UK in 2004. Currently, he works at Posts and Telecommunications Institute of Technology as a lecturer of Faculty of Information Technology. His research interests are big data, machine learning, and data mining.

Pham Bao Son received the PhD degree from University of New South Wales in 2007. Currently, he is vice rector of University of Engineering and Technology, Vietnam National University, Hanoi. His research interests are natural languge processing, machine learning, and data mining.

Tan Hanh received the PhD degree from Grenoble Institute of Technology, France. Currently, he is vice president of Posts and Telecommunications Institute of Technology. His research interests are machine learning, signal processing, and data mining.

Le Truong Thien received the Master degree from University of Engineering and Technology, Vietnam National University in 2002. His research interests are data mining, robotics, and machine learning.