Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 122 (2017) 1175–1181

Procedia Computer Science

www.elsevier.com/locate/procedia

Information Technology and Quantitative Management (ITQM 2017)

Identifying Subscribers in Freemium E-commerce Model Based on Support Vector Classification

Yongtao Yu[a],Bo Wang[a],Xiaodan Yu[a,] *

[a] *School of Information Technology and Management, University of International Business and Economics,Beijing 100029,China*

Abstract

Advances in information technologies have brought many changes to our lives. Finding free music through online platform rather than buying hard copies offline is one of the most significant changes. Recent researches suggest that as users becoming more engaged with the online content-provider platform, they are more willing to pay for the service or premium service. This study addresses the need of freemium e-commerce identifying potential subscribers. In specific, we propose a novel method, namely support vector classification (SVC), to categorize content viewers into potential subscriber and non-potential subscribers. Our method provides satisfied prediction result and the experiment showed that SVC is a superior method in this kind of task.

Peer-review under responsibility of the scientific committee of the 5th International Conference on Information Technology and Quantitative Management, ITQM 2017.

Keywods: subscriber identifying, music community, SVM, SVC;

1. Introduction

Introduction of the internet has brought rapid development of e-commerce in the past decades. Freemium e-commerce model has shown to be a successful model for building large consumer base for online business. However, for the freemium e-commerce company to make profits, a big challenge is to identifying active approaches so that users' interaction and engagement can generate economic values. Among the many approaches chosen by the e-commerce company, this study focuses on one particular approach, i.e. subscription.

Using users' preferences and behaviors to identify value consumer has been successfully applied in many areas, including: establishing a user credit score model in the financial sector to identify potential default users[1],[2],[3]; recommending potential products to users[4],[5],[6],[7]; recommending advanced services to users by integrating users behaviors in multiple social media platforms [8],[9]; recommending news and articles to mobile device users by learning their touch gestures in historical reading experience. However, there are few studies focus on the music community platform.

* Corresponding author.
E-mail address: yxd.xiaodanyu@gmail.com

1877-0509
Peer-review under responsibility of the scientific committee of the 5th International Conference on Information Technology and Quantitative Management, ITQM 2017.
10.1016/j.procs.2017.11.489

One of exceptional recent studies found that user participation can influence whether or not users are willing to become subscribers. For example, users who engage in the social features provided by the platform are more likely to subscribe to the platform and make the decision sooner. Another findings were that companies can seek design active approach to guide and/or affect users' effort level in their interaction with the online platform. However, it didn't consider both users' preferences and historical web use behaviors can influence the users' decision to subscription.[10]

Therefore, this study focuses on using users' preferences' and behaviors information to predict potential subscribers and user's social attributes for online music platforms. In particular, we chose to use support vector classification (SVC) method for this task. SVC method is suitable for high dimensional small sample classification problem and can avoid over-fitting problem of the traditional statistical methods [11],[12],[13],[14].

Fig.1 shows the research process of this study.

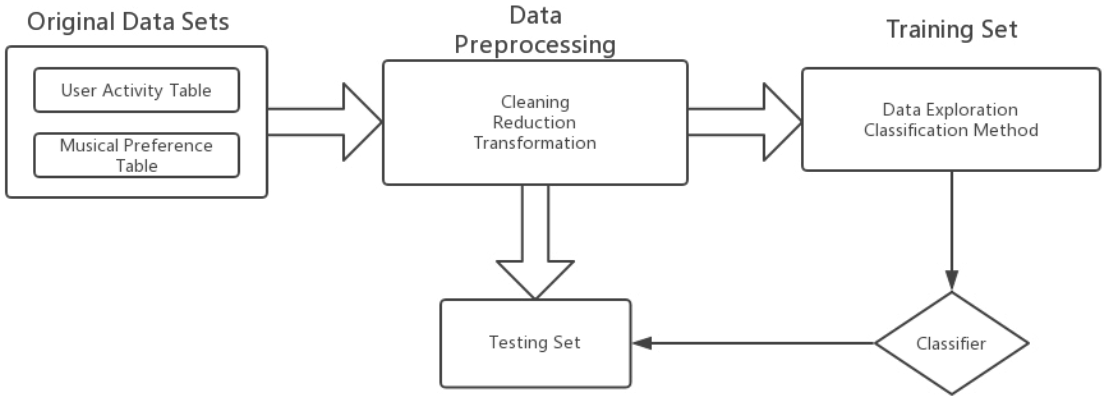

Fig. 1. Overview of the Data Mining for Music community platform subscriber identification

2. Support Vector Classification

The goal of SVC is to construct a partitioning hyperplane $(w^* \cdot x) + b^* = 0$, and to obtain the decision function.

Let the training samples be $T = \{(x_1, y_1), \cdots, (x_l, y_l)\} \in (R^n \times \mathcal{Y})^l$, where the inputs $x_i \in R^n$, and response variables $y_i \in \mathcal{Y} = \{1, -1\}, i = 1, \cdots, l$. Identifying values for parameters ω and b can be done by minimizing the following objective function

$$\min_{\omega, b, \xi} \frac{1}{2} || \omega ||^2 + C \sum_{i=1}^{l} \xi_i \tag{1}$$

$$s.t. y_i((\omega \bullet x_i) + b) \geq 1 - \xi_i, i = 1, \cdots, l, \tag{2}$$

$$\xi_i \geq 0, i = 1, \cdots, l \tag{3}$$

where $\xi = (\xi_1, \cdots, \xi_l)^T$, C>0 is a penalty parameter. The two objectives in function (1) show that it is necessary to minimize $||\omega||^2$ and $\sum_{i=1}^{l} \xi_i$, where the size of the parameter C is used to weigh the two terms. Appropriate penalty parameters C can guarantee the accuracy of the SVC model.

Considering the case when linear functions cannot separate the input patterns in the original space, we need to project the input data into a higher dimensional feature space as is shown in the following formula:

$$\Phi: \begin{array}{c} R^n \rightarrow H \\ x \rightarrow x = \Phi(x) \end{array} \tag{4}$$

Then we express the hyperplane in space H as

$$(\omega^* \bullet \Phi(x)) + b^* = 0 \tag{5}$$

The objective function in space H is given by the following formulation:

$$\min_{\omega, b, \xi} \frac{1}{2} ||\omega||^2 + C\sum_{i=1}^{l} \xi_i \tag{6}$$

$$s.t. y_i((\omega \bullet \Phi(x_i)) + b) \geq 1 - \xi_i, i = 1, \cdots, l, \tag{7}$$

$$\xi_i \geq 0, i = 1, \cdots, l \tag{8}$$

The dual formation of the problem（6）is

$$\min_{\alpha} \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} y_i y_j K(x_i, x_j)\alpha_i\alpha_j - \sum_{j=1}^{l} \alpha_j \tag{9}$$

$$s.t. \sum_{i=1}^{l} y_i\alpha_i = 0 \tag{10}$$

$$0 \leq \alpha_i \leq C, i = 1, \cdots, l \tag{11}$$

where $K(x_i, x_j) = \Phi(x_i) \cdot \Phi(x_j)$,which is inner product. K is called the kernel function.

Then the decision function for x in the space,R^n, is of the following form:

$$f(x) = \mathrm{sgn}(g(x)) \tag{12}$$

where

$$g(x) = \sum_{i=1}^{l} y_i \alpha_i^* K(x_i, x) + b^* \tag{13}$$

3. Experiment

3.1. Data Set

We selected Netease cloud music platform (http: // Music.163.com/#) as a data source. Netease cloud music platform is the largest social community with a social function and content distribution function in China. This platform exposes all user activity information and music history information. Users can access the music source on this platform. They publish the news about music and comment on interactive music or music list(The song is a list of music sources and related information). Users can create their own music lists and collect music lists from others.. They can focus on other users and get other users (fans) attention. Users can pay subscribers by paying, so you can enjoy VIP services. After the users become subscribers, they can get the following privileges: first, access and download the whole platform high quality lossless music; then, get part of the premium album source; in particular get exclusive icon.

This platform shows user activity information such as grade, number of fans, published dynamic situation, and it also contains the user's music list information, including song name, song label (song label shows the characteristics of the user's collection of songs, such as language features: Mandarin, Europe and the United States, singles style features such as Rock, jazz), so we can have a more comprehensive understanding of the platform users.

3.2. Data Preprocessing

Based on the purpose of the study, we obtained the user data from the Netease cloud music platform through the Gooseeker(V8.2.1) crawler software(http://www.jisouke.com/). We count the attributes of the user's collection of songs according to the original platform label. There are over 80 attributes for each record on many aspects, such as music genres, music suitable for various scenes, emotions of the music and etc.

Next step, we did the dimension reduction. First of all, according to previous research [10], user participation will have a significant impact on the user's willingness to pay for enhanced services. Therefore, for the purpose of this study, we selected four representative attributes, i.e. user level, number of fans, the number of other users concerned about and the number of published social messages. Then we choose from the tags that show users' preferences on music. The reduced list of the data attributes are shown in table 1. The total number of data attributes is 31.

Table 1. Description of the 31 dimensions in the datasets.

ID	Dimension	ID	Dimension
1	User level	17	Number of Weird/Independent music list
2	Number of fans	18	Number of Light Music and New Age music list
3	Number of other users who are concerned	19	Number of Jazz and Bossa Nova music list
4	Number of published dynamics	20	Number of music list at night and for noon break
5	Number of Chinese music list	21	Number of music in the morning, in learning,at work, for afternoon tea and at the bar list
6	Number of Western music list	22	Number of music on the subway,behind the wheel, at travel time, at walk time and at sports time.

7	Number of Cantonese music list	23	Number of conciliative, comfortable, excited and happy music list
8	Number of Japanese and Korean music list	24	Number of sentimental,alone and yearning music list
9	Number of Minority Language music list	25	Number of remember reminiscence ,fresh,romantic,sexy,touched and quiet music list
10	Number of Pop and England music list	26	Number of ACG and game music list
11	Number of Rock,Metal, PUNK,Reggae and Post-Rock music list	27	Number of classic,cover version,post 70s,post 80s and post 90s music list
12	Number of Folk Rhyme and country-and-western music list	28	Number of instrumental music,guitar music and piano music list
13	Number of Dance Music list	29	Number of the music list for generation after 00s and for children
14	Number of Electronic Music list	30	Number of campus song,Songs on the Internet and KTV music list
15	Number of R&B/Soul Rap,Blues and Latin music list	31	Number of music list for Soundtrack and Music List
16	Number of Nation Music,World Music,Classical Music and Ancient Music music list		

In the next step, for the purpose of the study, we use factor analysis, i.e. principal component method.to further reduce dimension. We used SPSS software (version 21) to conduct the analysis. Table 2 shows the results of the factor analysis. We chose 20 potential factors, which can account for 88.53% of the variability in all variables.

Therefore, the independent variables in our study were the 20 variables chosen after the factor analysis, and the dependent variables are users' subscribers' status, i.e. VIP membership or regular membership.

Table 2. Dimension Reduction after Factor Analysis

Component	Principle component method for factor analysis		
	Eigenvalues	% of Variance	Cumulative %
1	7.494	24.173	24.173
2	2.279	7.353	31.526
3	2.065	6.662	38.188
4	1.675	5.402	43.590
5	1.479	4.771	48.361
6	1.185	3.822	52.183
7	1.162	3.748	55.931
8	1.109	3.578	59.509
9	0.941	3.036	62.544
10	0.905	2.920	65.465
11	0.876	2.825	68.290
12	0.829	2.673	70.963
13	0.800	5.581	73.544
14	0.749	2.417	75.961
15	0.724	2.335	78.296
16	0.688	2.221	80.516

17	0.663	2.138	82.654
18	0.645	2.081	84.735
19	0.599	1.933	86.668
20	0.577	1.862	88.530

3.3. C-SVC model and result

Our algorithm was written in Libsvm3.22. The experimental environment was: Inter Core I5 CPU, 4GB memory. For the training data set, we selected a total of 350 VIP user samples and 350 non-VIP user samples. For the testing data set, there was a pool which contains1153 samples of mixing VIP and non-VIP users. We built five testing data sets by randomly sampling 50% of the data pool. Each testing data set contains 563 samples. From testing data set 1 to 5, the numbers of VIP users are: 62, 44, 55, 40, and 49 respectively.

Then we first ran the C-SVC on the training set, using the RBF kernel function. Table 3 shows the model accuracy rates on all five testing data sets. Figure 2 shows the TPR (True Positive Rate) and TNR (True Negative Rate).

TPR represents the proportion of users who are correctly identified as potential subscribers in real subscribers. We are more concerned about the size of TPR. It can be seen that TPR of all test sets is above 0.8, and some even reach 0.9. In this way, the model can correctly obtain the user's willingness to purchase subscriber services.

Table 3. The accuracy of SVC model on training set and test datasets

Data set number	Prediction accuracy
Test 1	80.9947%
Test 2	79.7513%
Test 3	79.0409%
Test 4	77.9751%
Test 5	80.9947%
Training Set	82.8571%

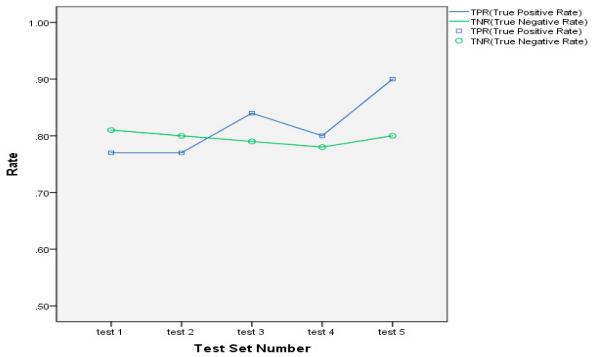

Fig. 2. TPR and TNR of Test SET

4. Conclusion

This study focuses on the freemium e-commerce business model company, particularly the music platform company. The overall objective of this study is to predict whether or not a user will become a subscriber, i.e. a VIP member. The findings of the study will help companies to better design their company strategy to realize the business value of the users. This study combined both social attributes and users preferences of music to predict. Particularly, we used SVC method to do the analysis. Results showed the method is reliable and provide high accuracy rate.

Future research could compare the performance of SVC with the performance of other data mining methods Further, more research is needed in selecting appropriate explanatory variables with theoretical bases. Future research can pay more attention to the user education level, work, identity, personal introduction and other static information. Researchers can also focus on the user music history ranking, music reviews and other dynamic information, which can be more detailed embodiment of the user style.

Acknowledgements

This work was partially supported by National Natural Science Foundation of China (Grant No. 71501044), the Fundamental Research Funds for the Central Universities in UIBE (Grant No: 16YQ07, 14QN03),

References

[1] Drake, Michael S, Quinn, Phillip J., Thornock, Jacob R. Who Uses Financial Statements? A Demographic Analysis of Financial Statement Downloads from EDGAR ACCOUNTING HORIZONS 2017;31:55–68.
[2] Maher Ala'raj , Maysam F. A new hybrid ensemble credit scoring model based on classifiers consensus system approach. Expert Systems With Applications 2016;64:36–55.
[3] A.I. Marqués a, V. García b, J.S. Sánchez. Two-level classifier ensembles for credit risk assessment. Expert Systems with Applications; 2012;39:10916--10922.
[4] Wayne Xin Zhao, Sui Li, Yulan He, Liwei Wang, Ji-Rong Wen, Xiaoming Li: Exploring demographic information in social media for product recommendation. Knowl. Inf. Syst. 49(1): 61-89 (2016)
[5] Wayne Xin Zhao, Sui Li, Yulan He, Edward Y. Chang, Ji-Rong Wen, Xiaoming Li: Connecting Social Media to E-Commerce: Cold-Start Product Recommendation Using Microblogging Information. IEEE Trans. Knowl. Data Eng. 28(5): 1147-1159 (2016)
[6] K. -W. HwangD,ApplegateA,et al. Leveraging Video Viewing Patterns for Optimal Content Placement.J NETWORKING 2012.p 44-58
[7] Girish Punj. The relationship between consumer characteristics and willingness to pay for general online content: Implications for content providers considering subscription-based business models.J Marketing Letters,2015; 26(2): pp 175–186.
[8] Li Kai. Research on the Influence of Social Media User 's Participation on Paid Behavior.D Beijing University of Posts and Telecommunications,2015
[9] A Esfandyari,M Zignani,S Gaito,GP Rossi. User identification across online social networks in practice: Pitfalls and solutions. Journal of Information Science , 2016
[10] Lior Zalmanson,Gal Oestreisher-Singer.Turning Content Viewers Into Subscribers.MITSloan Management Review,2016,57(3):p 10-15
[11] N. Deng, Y. Tian. Support vector machines: Theory, Algorithms and Extensions, Science Press, China, 2009.
[12] Z. Qi, Y. Tian, Y. Shi. Robust twin support vector machine for pattern classification, Pattern Recogn. 2013;46 (1): 305–316.
[13] V. Vapnik, L. Bottou. Local algorithms for pattern recognition and dependencies estimation, Neural Comput. 1993;5 (6): 893–909. doi:10.1162/neco.1993.5.6.893.
[14] Z. Qi, Y. Tian, Y. Shi. Structural twin support vector machine for classification, Knowledge-Based Systems ;2013;43:74–81. doi:10.1016/j.knosys.2013.01.008.