Electronic Commerce Research and Applications 38 (2019) 100887

Contents lists available at ScienceDirect

Electronic Commerce Research and Applications

journal homepage: www.elsevier.com/locate/elerap

Online purchase decisions for tourism e-commerce

Guixiang Zhu[a], Zhiang Wu[b,*], Youquan Wang[b], Shanshan Cao[b], Jie Cao[a,b]

[a] College of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China
[b] Jiangsu Provincial Key Laboratory of E-Business, Nanjing University of Finance and Economics, Nanjing, China

ARTICLE INFO

Keywords:
Purchase prediction
Tourism E-commerce
Clickstream data
Classification
Logistic regression
Semi-supervised learning
Multi-view learning

ABSTRACT

In this study, we consider the purchase prediction problem in the context of e-tourism, an emerging and prevailing application in e-commerce. Although a wide array of studies have been taken on purchase prediction, little analysis has been done on the purchasing behaviors towards tourism products. Also, the design of the corresponding purchase prediction model deserves researchers' full attention. We begin by introducing a real-life e-tourism dataset and constructing a suite of variables based on the detailed current and historical clickstream information. To validate the effectiveness of variables, we then perform a quantitative analysis to address quite a few interesting characteristics of purchase patterns. To predict whether or not a purchase is made for a current visiting session, we present a novel model called co-EM Logistic Regression (co-EM-LR) which combines the semi-supervised learning and the multi-view learning into its procedure. The co-EM-LR model has at least two outstanding merits: (1) it inherits the ease interpretation of the logistic modeling; and (2) it fully exploits both unlabeled data and the compatibility of multiple views to improve the prediction accuracy. Comprehensive experiments demonstrate the proposed co-EM-LR model yields significant prediction performance advantages over five competitive methods. Furthermore, two complementary views can mutually improve the performance with each other and finally offer fast convergence.

1. Introduction

As one of the primitive adopters of Internet, the tourism industry has become one of the most successful and profitable applications in e-commerce (Huang et al., 2017). In the e-tourism environment, tourists increasingly fall back on various online platforms to collect more rich, comprehensive and personalized information for planning their travel. As a result, a large amount of travel data has been readily available to firms, who are eagerly seeking the novel use of data analytic techniques to release the potential for businesses from their data (Navío-Marco et al., 2018). Similar to the other types of e-commerce, the foremost concern of e-tourism is also to understand and predict the online buying behavior, in order to improve the visit-to-purchase conversion rate (Sismeiro & Bucklin, 2004; Chen et al., 2009). As it is common knowledge that even a small improvement in the conversion rate (CR) would be worth millions of dollars to firms (Ayanso & Yoogalingam, 2009; Ludwig et al., 2013). For instance, Ludwig et al. (2013) claimed that only 1% increase in CR can translate into millions of dollars in sales revenues on Amazon.com.

To understand consumer preferences and thus to match them with the most desired products are now more important than ever in online shopping worlds. Existing customized promotion studies (Zhang & Wedel, 2009; Wan et al., 2017) have established the three-stage purchase decision model including (i) purchase incidence, (ii) product choice and (iii) purchase quantify. The first stage is also termed as *purchase prediction* in the literature (Kim et al., 2003; Sismeiro & Bucklin, 2004; Van den Poel & Buckinx, 2005; Mokryn et al., 2019) and it is indeed the fundamental step of the purchase decisions model. This paper offers studies on analyzing and predicting online purchasing behaviors of tourism products by regarding the distinct characteristics of real-life e-tourism dataset. Also, this paper aims to explore how characteristics of tourism products and online behavior of consumers affect buying decisions.

In the marketing and data analytics literature, there has been extensive re-search on the analysis and prediction of online purchasing behavior (Van den Poel & Buckinx, 2005; Pavlou & Fygenson, 2006; Kooti et al., 2016). In these studies, the clickstream constitutes the main component of data obtained from the e-commerce site, and it provides the opportunity for thoroughly understanding customers' online behaviors (Sismeiro & Bucklin, 2004; Van den Poel & Buckinx, 2005). Along this line, a rich set of variables that may determine the purchasing behaviors have been defined, evaluated and used as the feed of prediction models (Sismeiro & Bucklin, 2004; Van den Poel & Buckinx, 2005; Pavlou & Fygenson, 2006), including the demographic of the

* Corresponding author.
E-mail address: zawuster@gmail.com (Z. Wu).

https://doi.org/10.1016/j.elerap.2019.100887
Received 9 April 2019; Received in revised form 9 July 2019; Accepted 11 August 2019
Available online 25 September 2019

customer, detailed browsing behavior, repeat visitation, historical browsing or purchasing behav-ior, etc. Besides the clickstream data, some recent studies (Lu et al., 2010; Kooti et al., 2016) attempted to add external factors (e.g., social contacts or friends, virtual community) into the purchasing behavior analysis. Unlike most of previous studies that focused on e-retailers (e.g., Amazon and Walmart) selling a broad assortment of low-cost products, we used the real-life data from an e-tourism site. Hence, the tasks of understanding and predicting the online purchasing behavior on such intangible, experiential and perishable tourism products are markedly different from those at grocery e-commerce sites. Furthermore, there has been less research focusing on developing more sophisticated learning models to improve the prediction accuracy. Instead, most of existing research utilize general-purpose classification models for addressing the purchase prediction problem. As one of the contributions of our study, we will present a specific model for the task of online purchase prediction, which is built upon the semi-supervised learning and the multi-view learning techniques.

The tourism product usually refers to an integrated package containing a set of necessary travel-related ingredients (Liu et al., 2014), such as the departure/destination city, financial cost, the number of days, transportation, accommodation, and so on. Many studies (Liu et al., 2014; Ge et al., 2014; He et al., 2016) have been performed on personalized recommendation for travel products. Based on the tourism products consumption data provided by an offline travel agency, they have shown that the travel-product recommendations possess distinct characteristics compared with traditional products such as movies, books and groceries. However, little is known about the characteristics of consumer behavior when browsing the e-tourism sites and how these behavioral variables affect customers' purchase decisions. This paper makes multiple contributions in this regard.

- We first introduce a real-life e-tourism dataset obtained from a large tourism e-commerce company in China and then define a suite of effective variables for the task of online purchase prediction. Then, a quantitative analysis is performed to address quite a few interesting characteristics of purchase patterns on the e-tourism data.
- A novel learning model named co-EM Logistic Regression (co-EM-LR) for online purchase prediction is proposed. The co-EM-LR model exploits unlabeled data (i.e., semi-supervised learning) and the compatibility of multiple views (i.e., multi-view learning) to improve the prediction accu-racy. Also, the use of regression within our model can provide the good interpretation of control variables.
- Extensive experiments are conducted to evaluate our proposed model on purchase prediction by using the real-life e-tourism dataset. The experimental results demonstrate the superiority of our methods.

The reminder of this paper is organized as follows. In Section 2, we summarize the related work. Section 3 describes the problem that we study in this article. Section 4 introduces the real-life e-tourism dataset, defines the variable set that will be used in our prediction model, and performs a quantitative analysis on purchase patterns. Section 5 details our proposed purchase prediction model co-EM-LR. The experimental results are presented in Section 6, followed by the conclusion in Section 7.

2. Literature review

In this section, we survey the relevant literature in two streams of research: travel behavior modeling as well as online purchase analysis and prediction.

Travel Behavior Modeling. Recent developments of ubiquitous computing and location-based social networks have given birth to numerous location-based ap-plications within the urban computing. Among them, much of work focuses on understanding users' travel behaviors by mining the human mobility data in daily life. For instance, many researchers attempt to predict the next location a user will visit and further to generate itinerary as a sequence of locations under trip constraints such as time limits, start and end points, etc (Khan et al., 2017; Wen et al., 2017). Meanwhile, quite a few work has been done on developing effective recommendation methods for travel packages (Liu et al., 2014; Ge et al., 2014; He et al., 2016). For instance, by taking the travel cost (i.e., the financial and the time) into the consideration, Ge et al. (2014) provided focused study of matrix factorization used in the cost-aware latent factor models. There are also extensive studies on mining transportation data to estimate passengers' future travel pattern (Zhao et al., 2017a; Liu et al., 2018). For instance, Zhao et al. (2017a) proposed an effective data-mining procedure to better understand the travel patterns of individual metro passengers in Shenzhen, a modern and big city in China. The above studies have repeatedly verified that the travel behavior is very complex and it is influenced by various contextual factors, e.g., cost (Ge et al., 2014), season (Liu et al., 2014) and social relationships (He et al., 2016). In light of this, we explore the online purchasing behavior for travel products by mining the clickstream data sourced from an e-tourism website, which is barely touched in the study of travel-related data mining. Here, we review only a few papers on several representative research directions. Interested reader can refer to Calabrese et al. (2015), Dong et al. (2018) for a more comprehensive review of the previous literature.

Online Purchase Analysis and Prediction. A wide array of studies within this field is available in information systems, economics, computer science and marketing. Many empirical studies target at revealing that what visitors are exposed to, and what they do in a site visit, will affect a visitor will buy online. For example, Moe and Fader (2004) developed an individual probability model to accommodate a variety of visit-to-purchase relationships, where some visits were motivated by planned purchases while others were simply browsing visits. Bucklin and Sismeiro (2003) proposed a model to predict whether a visitor decided to continue browsing or to exit the site, as well as how long the visitor would spend browsing a web page. They also presented a task-completion approach to estimate the user's online shopping behavior (Sismeiro & Bucklin, 2004). Additionally, Moe (2006) applied the empirical two-stage model to Internet clickstream data and modeled the consumer decision process on e-commerce sites as two choice stages: products viewed and products purchased. Similarly, Pavlou & Fygenson (2006) presented a prediction model for two-staged online behaviors: getting information and purchasing products. These work explained the online purchasing behavior as the result of deliberate planning, which is known as theory of planned behavior. The studies presented thus far have established the general causality correlation between visits and purchases.

There is a large number of published studies on investigating how the specific factor will affect the online purchasing behavior. For instance, Lo et al. (2016) studied user activity and purchasing behavior with the goal of building models of time-varying user purchasing intent, which provides a promising starting point in terms of identifying potential purchasers and better understanding their long-term behavior. Ludwig et al. (2013) studied how the affective content and the linguistic style of product reviews influence the conversion rate. Lu et al. (2010) explored how trust in virtual communities affect the purchase decision making. Iwanaga et al. (2016) investigated the relationship between the recency/frequency of customers' page views and the probabilities of their product choices on e-commerce sites. Other factors that have been addressed include social factors (Kooti et al., 2016), search behavior (Schlosser et al., 2006), live chat (Lv et al., 2018), prior purchases (Brown et al., 2003; Morisada et al., 2019) and trusting beliefs (McKnight et al., 2002).

Several studies suggest to construct detailed clickstream variables for predicting the online purchasing behavior. Young Kim & Kim (2004) evaluated the importance of variables from four dimensions, i.e., transaction/cost, incentive programs, site design and interactivity, on the prediction of purchase intentions for clothing products. Van den

Poel & Buckinx (2005) defined variables from four different categories and demonstrated that detailed clickstream variables are the most important ones for the task of online purchase prediction. Furthermore, Liu et al. (2016) created a more comprehensive view of variables including profiles for users, merchants, brands, categories, items and their interactions for predicting the repeating buying behavior. Although these studies shed light on the construction of variables for predictors, the prediction models used so far are limited in some classical models such as Logistic Regression (Van den Poel & Buckinx, 2005), Random Forest (Liu et al., 2016), Gradient Boosting Decision Tree (Volkovs, 2015). In fact, Kim et al. (2003) have compared five combination methods of multiple classification models for online purchase prediction. They showed that the combination of multiple classifiers outperformed the single classifier. Therefore, it would be expected that to design a purchase prediction model using advanced machine learning techniques will result in an increase of the prediction accuracy. Our work, in this respect, is unique to the relevant research on the purchase prediction. That is, we present a novel model that is built by using the semi-supervised learning and the multi-view learning, which is more suitable for the purchase prediction task.

3. Modeling online purchase prediction

To predict and understand online buying behavior typically rely on the browsing behavior by using the page-to-page clickstream data recorded in server log files (Sismeiro & Bucklin, 2004; Van den Poel & Buckinx, 2005). Meanwhile, the consecutive clickstream is usually divided into a series of *sessions*, each of which represents a single visit to the website. From a data perspective, we model the online purchase prediction problem as shown in Fig. 1. The live system aims to predict a user whether to buy in the near future time mainly based on the current session of this user. Furthermore, existing research (Moe & Fader, 2004; Jerath et al., 2011; Iwanaga et al., 2016; Morisada et al., 2019) has repeatedly verified that the recency and frequency of customers' previous purchases and visits are important indicators for forecasting purchases in future. So we should take the recent clickstream into consideration for part of users. As a result, this prediction model is mainly trained with the clickstream data and labels in the observed window, and with recent clickstream and demographics if available for experienced customers.

User Segmentation. Existing literature (Tkaczynski et al., 2009; Berthon et al., 2012; Morisada et al., 2019) have shown the *one-size-fits-all* marketing strategy is far from the best practice. So the customer segmentation is essential for deeply understand and also accurately predict users' purchasing behaviors. In general, users in the observed window can be divided into three disjoint categories, solely according to the browsing records.

- *First-Time Visitors.* This kind of users visit the Website for the first time, which implies none historical information is known about these users except the current clickstream.
- *Ever-Visited Users.* This kind of users have ever visited the Website but did not visit recently, e.g., in recent one month. So some demographic information is available besides the current clickstream.
- *Recent-Visited Users.* This kind of users are very active, i.e., they visited the Website recently. As a result, recent, current clickstream and some demographic information are available.

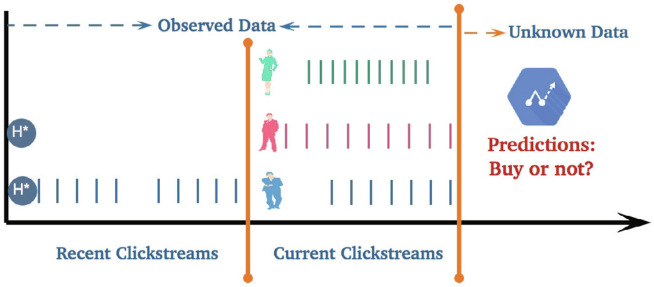


Fig. 1. A general model for the online purchase prediction.

It is noteworthy that the standard for user segmentation is not unique. For instance, Hernández et al. (2010) distinguished users as two groups according to whether they have purchased: potential e-customers who have never purchased and experienced e-customers who have made at least one purchase. Although our above user segmentation is based on previous visits, we will further distinguish recent-visited users as "not purchased" and "purchased" groups for carefully examining their purchasing behaviors in Section 4.3.

4. e-Tourism dataset and analysis

The goal of our research is to investigate the online purchasing behavior on the tourism e-commerce platform, which has been verified to have many different characteristics with other comprehensive e-business platforms (e.g., Amazon, eBay). Hence, in this section, we first introduce a real-life e-tourism dataset used in our study and then construct a suite of effective variables for the task of online purchase prediction. We then perform a quantitative analysis to show some interesting characteristics of purchase patterns on this dataset.

4.1. Data description

The dataset is provided by Tuniu[1], one of the largest online travel agency (OTA) platforms in China. At the time of this study, Tuniu is capable of providing over 1 million tourism products and has provided tourism service for approximately 15 million customers. For these reasons, Tuniu is a compelling setting in which to investigate Internet buying behavior on the tourism e-commerce. This dataset is mainly made up of the page-to-page clickstream data from server logs, which is in fact the common setting within the research on online purchasing behavior analysis (Sismeiro & Bucklin, 2004; Van den Poel & Buckinx, 2005). In our study, we extract the clickstream of three disjoint weeks with different contexts, as listed in Table 1. In particular, D_1 corresponds to a week during summer holidays, D_2 is the last week before China's National Day holiday, and D_3 is a typical week in working day (i.e., slow season for travel). Moreover, each *session* consists of a sequence of pages clicked by a user during a certain period and it is regarded as a sample (i.e., an instance) hereafter in our study. Each session is labeled purchased or not according to whether it contains the booking pages. Thus, the number of purchased sessions and the corresponding conversion rates (CR) are obtained. Not surprisingly, the conversion rates of weeks during or near holidays (i.e., D_1 and D_2) are higher than that of working day (i.e., D_3). As reported by Moe and Fader (2004) and Ludwig et al. (2013), conversion rates average approximately 2%–3% across online retail sites. We can thus believe that the conversion rate on e-tourism sites is quite low, that is, the CR of slow season is only 1.26% (i.e., D_3) and the CR of week near long holiday barely reaches 2.54% (i.e., D_2).

In addition to the clickstream, our dataset also includes a bank of attributes associated with every page. A majority of pages are used to introduce different tourism products such as travel packages, attraction tickets and visa services. For these product pages, the descriptive attributes contain price of the product, departure city of the product, and two tourism product classifications from two angles. In detail, the first product classification is according to the travel region and thus each product is described as local/around tour, domestic short/long haul and oversea short/long haul. Another product classification addresses the

[1] http://www.tuniu.com

Table 1
Statistics of e-Tourism Datasets Used in Our Study.

	Time	#Record	#User	#Session	#Purchase	CR
D_1	1 to 7 Aug., 2012	2,022,633	364,067	431,321	7284	1.69%
D_2	24 to 30 Sept., 2012	1,980,299	341,878	403,032	10,236	2.54%
D_3	1 to 7 Nov., 2012	941,930	190,292	217,692	2731	1.26%

Note: (1) "#Purchase" indicates the number of *sessions* including the buying events;
(2) "CR" means the *conversion rate*:#Purchase/#Session*100%.

travel type including Tuniu special tour, package tour, self-driving tour, selfguided tour, company's package tour, and local-attended tour. Meanwhile, if a list page is returned by the search engine, then the search engine and the corresponding search keywords will also be recorded. Besides, IP address of every session is recorded, and thus the city that a customer lives in can be inferred from the IP address. The above descriptive information plays a vital role in constructing different variables for the prediction task, which will be introduced in Section 4.2.

In our dataset, there are many basic variables exhibiting the heavily-tailed characteristic. Fig. 2 shows the distribution of three variables: the length of sessions, price of all products and purchases made by users. As can be seen, an overwhelming majority of users clicked less than 10 pages. What's more, less than 2% of users have clicked more than 20 pages (Fig. 2(a)). Similarly, existing research (Bucklin & Sismeiro, 2003; Sismeiro & Bucklin, 2004; Van den Poel & Buckinx, 2005; Iwanaga et al., 2016) generally suggests that the number of browsed pages of each session and its dwell time exert the positive impact on purchases. To examine these on our e-tourism data, we employ the Mann-Whitney *U* test (Mudambi & Schuff, 2010) on two groups of samples (i.e., buy and not-buy). According to the results (p-value ≤ 0.001), we find a statistically significant positive association between length (dwell time) of a session and purchases. Most tourism products cost hundreds of dollars (Fig. 2(b)), but customers are likely to buy products around 100 dollars (Fig. 2(c)). Compared with daily-used commodities (the average expenses is around 10 dollars (Kooti et al., 2016)), the tourism products are quite expensive.

4.2. Variable definitions

In our settings, the demographics of users are scarce due to the privacy and resources concerns, which is largely different from a body of existing studies (Kim et al., 2003; Van den Poel & Buckinx, 2005; Lu et al., 2010; Kooti et al., 2016). There also exist a large number of new-visited users of which the unique data is their current clickstream. Hence, it is required to fully exploit the clickstream data to define a rich set of variables, including not only browsing behaviors but also tourism-specific behaviors. Furthermore, based on clickstream in the last month, we can define another set of variables for recent-visited users. Formally, we denote i as the user number and j as the session number. Thus, a variable associated with i indicates it is defined at the user-level (analogically for j at the session-level). Every instance to be predicted is a session, rather than a user, and a user may lead to multiple sessions. However, given a session, we can associate this session with a specific user and then variables defined on this user will be added onto this session.

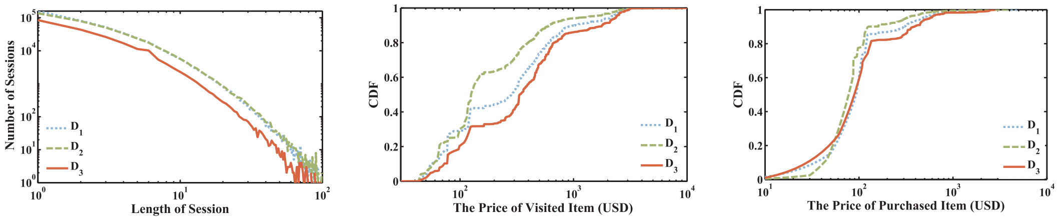
Table 2 summarizes the variables that will be used for the task of online purchase prediction and gives their brief descriptions. The one but last column indicates whether this variable is already used in existing studies. Most of the variables are self-explanatory, whereas for several intricate variables including $Location_j$, $TRegions_j$, $PTypes_j$ and $RPages_j$, we present their computational details in Appendix A. Van den Poel & Buckinx (2005) classified the variables for purchase prediction into several categories: clickstream measures, customer demographics and purchase behavior. Besides these categories, we add a new category called spatio-temporal measures to address some particular feature of tourism products. For example, $Location_j$ measures the distance between the city that a customer lives in and the departure cities that he has clicked, and $Holiday_j$ indicates the interval between current time and the next holiday.

To validate the utility of the variables (Bacharach, 1989), many studies have provided an in-depth analysis on hypothesis testing or statistical significance for the variables (e.g., Van den Poel & Buckinx, 2005; Brown et al., 2003; Lu et al., 2010; Pavlou & Fygenson, 2006). The Mann-Whitney *U* test enjoys great popularity among scientists comparing two independent groups of observations, and tests specifically whether there is a difference between randomized groups in terms of their mean ranks (Bergmann et al., 2000). We employ the Mann-Whitney *U* test on two groups of samples (i.e., buy and not-buy) and show the statistical results in the last column of Table 2. As can be seen, all of variables are statistically significant at least at the 0.05 level.

4.3. Purchase pattern analysis

In this subsection, we perform a quantitative analysis to address quite a few interesting characteristic of purchase patterns on both recency and current clickstream, especially for some domain characteristics in e-tourism. Each group of analysis corresponds to some variable as listed in Table 2. In fact, these numerical results reflect the observed phenomena in e-tourism domain. From this perspective, we validate the selected variables are incapable of providing explanations of observed phenomena, which is a primary criteria for evaluating variables (Bacharach, 1989).

First, we investigate the correlation between purchasing behavior and the user segmentation. As introduced in Section 3, we have divided users into three groups: first-time visitors, ever-visited users and recent-visited users. Also, by collecting the buying behavior recent-visited users, we further divide them into recent-purchased users and recent-not-purchased users. Table 3 shows the conversion rates of every kind of users. We first observe that a majority of users are first-time visitors that account for 52.3% on D_1, 61% on D_2 and 55.6% on D_3, respectively. This situation is somewhat like the cold-start problem addressed in recommendation (Liu et al., 2014; Ge et al., 2014; He et al., 2016): the purchasing behavior prediction is complicated by the lack of abundant information for the first-time visitors. As expected, existing customers (i.e., members) bring about higher conversion rates, which is modeled by the variable $Member_i$. However, it is striking that the



(a) Length of Sessions (b) Price of All Products (c) Price of Purchases

Fig. 2. Distributions of session length and price.

Table 2
. Variable Definitions: ✓ means the variable has been used in previous studies.

Variable			Description		*p*-Value
Variables for Recency Clickstream		Customer demographics			
	1	$Member_i$	Whether user i is member (1 = yes, 0 = no)	✓	***
		Clickstream measures			
	2	$LVDays_i$	Days elapsed since user i's last visit (Sigmoid)	✓	***
	3	$TotV_i$	Number of visits made by user i in last month	✓	***
	4	$PAvgV_i$	Average price in dollars of products browsed by user i in last month	✓	**
	5	$PDevV_i$	Stand. Dev. of price of browsed-products for user i in last month	✓	**
	6	$LDwell_i$	Dwell time in seconds of the last session of user i	✓	*
	7	$DwellAvg_i$	Average dwell time in seconds of user i's sessions in last month	✓	*
		Purchase behavior			
	8	$TotP_i$	Number of purchases made by user i in last month	✓	***
	9	$LPDays_i$	Days elapsed since user i's last purchase (Sigmoid)	✓	***
	10	$MAvgP_i$	Average monetary spending in dollars of user i in last month	✓	***
		Clickstream measures			
Variables for Current Clickstream	11	$PAvg_j$	Average price of products within session j (Log)	✓	***
	12	$PDev_j$	Stand. Dev. of price of products within session j (Log)	✓	**
	13	$Length_j$	Length of session j (Log)	✓	***
	14	$Dwell_j$	Dwell time in seconds of session j (Log)	✓	***
	15	$Search_j$	Which search engine is used in session j (0 = none, 1 = offsite, 2 = onsite)	✓	***
	16	$TRegions_j$	The entropy of travel destination distribution in session j	✓	***
	17	$PTypes_j$	The entropy of travel type distribution in session j	✓	**
	18	$RPages_j$	Percentage of pages containing travel product displays in session j	✓	***
		Spatio-temporal measures	Average price of products within session j (Log)		
	19	$Location_j$	Average semantic Sim. between user i's living city and departure cities of products in session j	✓	***
	20	$Holiday_j$	Number of days between log-time of session j and the latest holiday	✓	**
	21	$Weekend_j$	Whether current time of session j is weekend (1 = yes, 0 = no)	✓	**

Note: (1) "Sigmoid" means the variable is normalized into [0, 1] by Sigmoid function $\frac{1}{1+e^{-x}}$;
(2) "Log" means the variable is taken its logarithm as $\log_{10} x$.
(3) Mann-Whitney U test: * < 0.05; ** < 0.01; *** < 0.001.

Table 3
Conversion Rates of Different Users.

	First-Time Visitors	Ever-Visited Users	Recent-Visited	Users
			Not Purchased	Purchased
$D1$	0.88%	2.55%	2.19%	17.15%
$D2$	1.34%	1.39%	4.47%	17.70%
$D3$	0.61%	1.73%	1.92%	17.60%

conversion rates (CR) of recent-purchased users achieves an exceptionally high rate, i.e., over 17% on three datasets. This indicates that the historically purchasing behavior is strongly correlated to predicting the future purchases, which coincides with the results reported in previous studies (Brown et al., 2003; Iwanaga et al., 2016; Morisada et al., 2019). In contrast, we are fully conscious of that approximately 60% purchases are made by non-members. The primary data in hand for understanding non-members is the current clickstream, where the browsed pages are still very few according to Fig. 2(a). Therefore, how to fully exploit implicit information from current clickstream is very critical to the purchase prediction task.

Second, we consider the search behavior that affects purchasing decisions for current clickstream. Users often utilize search engines to find a needle and whether a user has used search engine is recorded for each current session. We further distinguish the search behavior as onsite search and offsite search. Note that the offsite search indicates one of sources directed to the e-travel site, e.g., from Baidu. The search behavior is modeled by the variable $Search_j$. To quantitatively characterize the effect of search behavior, we adopt two indicator random variable: $X = 1$ means the search behavior is included in a session; and Y is associated with the event in which the purchasing behavior occurs ($Y = 1$), otherwise for $Y = 0$. Then, we are interested in comparing two pairs of conditional probabilities over all samples (i.e., current sessions): $P\ (Y = 1|X =$ versus $P\ (Y = 0|X = 1)$; and $P\ (X = 1|Y = 1)$ versus $P\ (X = 1|Y = 0)$. As can be seen from Table 4, the onsite search behavior is a strong signal of purchases, because over 95% customers using onsite search have bought products and over 20% customers have ever used onsite search during their purchasing process.

Third, we focus on analyzing the recent-visited users in order to examine the recency effects of visiting and purchasing behaviors, corresponding to variables $LVDays_i$ and $LPDays_i$. For this purpose, we extract the set of users who have purchases and count the number of days elapsed since the last visit and purchase. From the CDF distributions as depicted in Fig. 3, we first observe that both distributions are heavy-tailed, implying the recency effects are significant. Moreover, the recency effect of last visit is even stronger, i.e., approximately 90% multi-visited users completed their purchases during their second visits within five days. However, time between purchases grows relatively gradually. This is in contrast to the findings of (Iwanaga et al., 2016; Kooti et al., 2016), which find that the time between purchases of daily-used commodities has bursty dynamics and weekly cycles. When examining the products purchased in a short time interval (e.g., about 60% users perform the second purchase within three days), we find these products are mostly trivial ones with the low cost, such as tours around, attraction tickets, etc. Once a user bought the package tour with high cost, such as domestic long haul and oversea short/long haul, etc. There is a relatively long time interval until the second purchase happens, e.g., approximately 18% second purchases occur after over 11 days.

Fourth, we evaluate a domain-specific factor on purchase decisions, i.e., the distance between the city that a customer lives in (inferred by IP address) and the departure cities that he/she has clicked. This factor corresponds to the variable $Location_j$ of which the computation details are shown in Appendix A. Fig. 4 shows the comparison results of buy and not-buy group. It is obvious that sessions without purchasing behaviors mostly own smaller similarity values (e.g., falling into the intervals (0, 0.25] and (0.25, 0.5]), while sessions with purchasing behaviors mostly own larger similarity values (e.g., falling into the interval (0.75, 1.0]). The impact of variable $Location_j$ accounts for that customers who frequently browse travel products departing from cities near to his living city are likely have strong purchase intents.

Table 4
Impact of Search Behavior on Purchases.

		D_1		D_2		D_3	
		Onsite	Offsite	Onsite	Offsite	Onsite	Offsite
$P(Y\|X = 1)$,	Buy	95.12%	3.74%	96.23%	4.19%	94.78%	4.69%
$Y = \{0, 1\}$	Not-Buy	4.88%	96.26%	3.77%	95.81%	5.22%	95.31%
$P(X = 1\|Y)$,	Buy	21.76%	61.47%	25.61%	53.25%	20.39%	59.65%
$Y = \{0, 1\}$	Not-Buy	1.67%	56.83%	1.78%	49.37%	1.22%	58.89%

Note: “Onsite“ and “Offsite” stand for *onsite search* and *offsite search*, respectively.

Finally, we investigate the impact of regions of travel products that customers have browsed. This factor is expressed by the variable $TRegions_j$ (see computation details in Appendix A). The boxplots in Fig. 5 compare the entropy distributions of regions, where small entropy indicates one customer centers on browsing few regions. As can be seen, the entropy of the buy group has small medians as well as small variances on three datasets. This implies that customers who have strong purchase intents are likely to browse travel products taking his interested regions as destinations. In contrast, customers who extensively browse various travel products hardly make purchase decisions.

5. The purchase prediction model

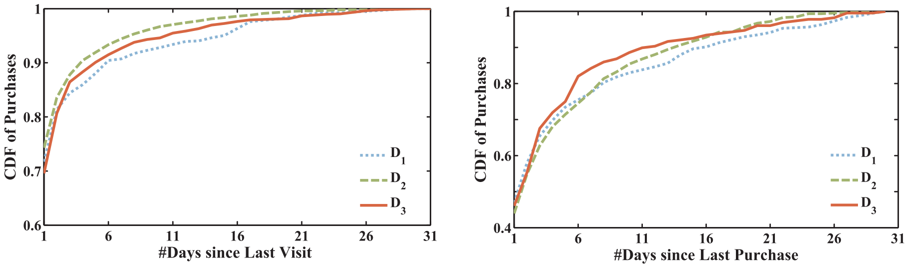
In this section, we present the purchase prediction model named co-EM Logistic Regression (co-EM-LR). In what follows, we first describe the motivation for our co-EM-LR model and present its general settings. Then we derive the inference algorithm based on the semi-supervised learning and multi-view learning approaches.

5.1. Model specification

Most of research (Kim et al., 2003; Van den Poel & Buckinx, 2005; Volkovs, 2015; Liu et al., 2016)on the use of machine learning methods to predict purchasing behavior focused on designing a variety of effective variables. Then, the well-formed classification models were adopted as predictors, e.g., the Logistic Regression model was used by Van den Poel & Buckinx (2005), and multiple classifiers were combined by a genetic algorithm in Kim et al. (2003), etc. However, little attention has been paid to develop advanced learning models making an attempt to fit the characteristics of purchase prediction problem. We adopt Logistic Regression as the base model, because it is a most widely-used model in marketing realm due to the ease of its interpretation on variables (Van den Poel & Buckinx, 2005). Furthermore, we extend the infrastructural Logistic Regression model to handle two-fold difficulties within the purchase prediction problem. Firstly, the labeled users used for training (i.e., the users are labeled whether purchased) are more limited than unlabeled ones, which is likely to reduce the generalization capacity of supervised predictors. So, we would like our model to be able to bootstrap the weak predictor, that is built with an initial small set of labeled instances, by using a large amount of unlabeled data. Secondly, as introduced in Section 4.2, we have two classes of variables: one for current clickstream and the other for recent clickstream. Inspired by the multiview learning (Zhao et al., 2017b), the compatibility between different views (i.e., classes) of variables can be exploited for improving the learning performance. So we might expect two classes of variables can work cooperatively to deliver a consistent decision. Based on the above discussions, we present a novel learning model, called *co-EM Logistic Regression* (co-EM-LR), for purchasing prediction. The co-EM (Fan et al., 2018) is a well-known and conceptually simple model that combines multi-view learning with the probabilistic Expectation Maximization (EM) algorithm. Our co-EM-LR model contributes to combine the discriminative model (e.g., regression) with the generative probabilistic model (e.g., EM). This novel model brings about several advantages: (i) it inherits the ease of model’s interpretation in marketing research; and (ii) its learning procedure is in semi-supervised as well as multi-view style. Mathematically, we denote $\mathscr{D}$ as the data collection, where each instance corresponds to a session in the case of purchase prediction. Thus, the set of labeled instances is denoted as $\mathscr{D}_l = ((x_1, y_1), \cdots, (x_{|\mathscr{D}_l|}, y_{|\mathscr{D}_l|}))$ where the binary valued variable $y_i \in \{1, 0\}$ denotes session i results in purchasing or not, and the remaining set of unlabeled instances is denoted as $\mathscr{D}_u = (x_1^*, \cdots x_{|\mathscr{D}_u|}^*)$. We have $\mathscr{D} = \mathscr{D}_l \cup \mathscr{D}_u$. Recall that two groups of features are constructed for each session in Section 4.2. According to the common setting of the multi-view learning, we hereafter denote V_1 and V_2 as the set of variables for recency and current clickstream, respectively. Any instance x_i can then be decomposed as (x_{i1}, x_{i2}), where $x_{i1} \in V_1$, $x_{i2} \in V_2$. It is noteworthy that the session delivered by the first-time visitor only has the single view V_2. So these sessions in the labeled data will be utilized for training V_2 and their labels will ultimately be determined by V_2.

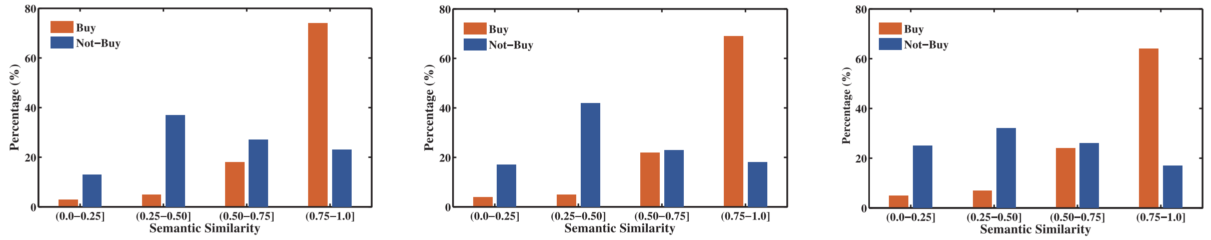
Our goal is to learn a Logistic Regression (LR) function which assigns high values to positive and low values to negative samples in $\mathscr{D}_u$. Let $h_\theta(x_i)$ denote the posterior probability output by the LR classifier governed by parameters $\theta = (\theta_1, \theta_2)$, where θ_1 and θ_2 correspond to the parameters for V_1 and V_2 respectively. If we set x_{i1} as a zero-vector when V_1 is missing for the first-time visitors, the LR function can be defined as

$$h_\theta(x_i) = \frac{1}{1 + \exp(-\theta^{\mathrm{T}} x_i)} = \frac{1}{1 + \exp(-\theta_1^{\mathrm{T}} x_{i1} - \theta_2^{\mathrm{T}} x_{i2})} \tag{1}$$



(a) Last Visit (b) Last Purchase

Fig. 3. Recency effects of last visit and purchase.



(a) D_1 (b) D_2 (c) D_3

Fig. 4. Impact of geographic information on purchases. The semantic similarity measures the distance between the city that a customer lives in and the departure city of tourism product.

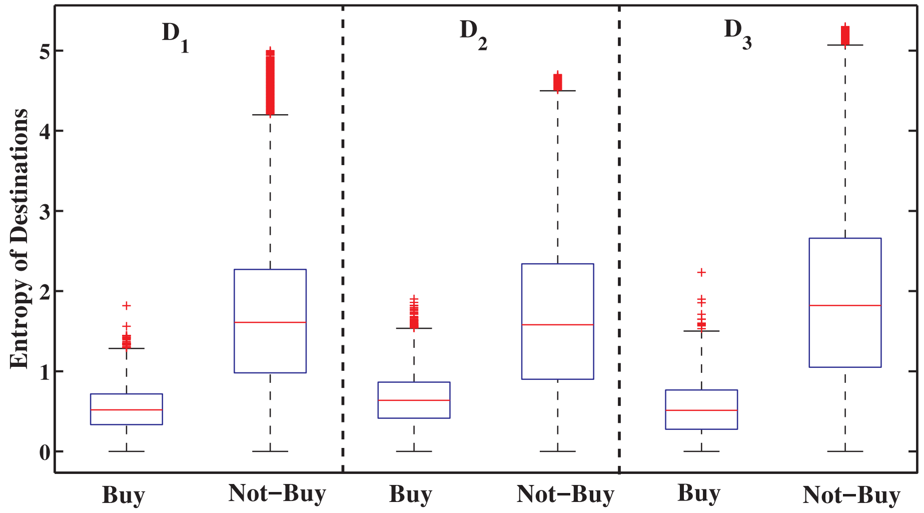


Fig. 5. Impact of travel regions distribution on purchases.

To estimate the class posterior $p(y_i|x_i)$, $y_i \in \{1, 0\}$, we assume a generative model: the decision function values for a class (i.e., $p(h_\theta(x_i)|y_i)$) are assumed to follow a Gaussian distribution N (μ, σ^2). Hence, our model needs to learn two types of parameters: θ within the LR function as well as μ and σ within the generative model.

5.2. Inference and learning

The co-EM-LR model is actually a semi-supervised learning procedure, which benefits from the Gaussian generative model for fully exploiting the unlabeled data. This is very similar to the co-EM model (Fan et al., 2018). In particular, we first build an initial LR classifier based on labeled data, and we can obtain the value of $h_\theta(x_j^*)$ for each unlabeled instance (x_j^*). By picking up a certain percentage of unlabeled instances with the largest $h_\theta(x_j^*)$ values, we obtain a set of positive instances (i.e., sessions being likely to purchasing), denoted as $\mathcal{D}_u^1$. To keep balance of whole data, we set the percentage of positive instances in $\mathcal{D}_u^1$ as equal to the percentage of instances with $y_i = 1$ in the labeled dataset. Thus, the set of negative instances is obtained by $\mathcal{D}_u^0 = \mathcal{D}_u \backslash \mathcal{D}_u^1$. The parameters of the generative model are μ_1, σ_1^2 and μ_0, σ_0^2 for two classes. We can estimate these parameters by

$$\mu_y = \frac{1}{|\mathcal{D}_l^y| + \lambda\,|\mathcal{D}_u^y|}\left(\sum_{x_i\in\mathcal{D}_l^y} h_\theta(x_i) + \lambda \sum_{x_j^*\in\mathcal{D}_u^y} h_\theta(x_j^*)\right), \tag{2}$$

$$\sigma_y^2 = \frac{1}{|\mathcal{D}_l^y| + \lambda\,|\mathcal{D}_u^y|}\left(\sum_{x_i\in\mathcal{D}_l^y} (h_\theta(x_i) - \mu_y)^2 + \lambda \sum_{x_j^*\in\mathcal{D}_u^y} (h_\theta(x_j^*) - \mu_y)^2\right), \tag{3}$$

where the preset hyper-parameter $\lambda \in [0, 1]$ is the weight that controls the importance of unlabeled instances in $\mathcal{D}_u$. For each unlabeled instance x_j^*, we can compute the conditional probability $p(x_j^* \,|\hat{y}_j)$ by using the Gaussian estimator.

$$p(x_j^*|\hat{y}_j) = \mathcal{N}(h_\theta(x_j^*)|\mu_{\hat{y}j}, \sigma^2_{\hat{y}j}), \quad \hat{y}_j \in \{1, 0\} \tag{4}$$

Then, according to the Bayesian theorem, we can infer the desired class probabilities $p(\hat{y}_j \,|x_j^*)$:

$$p(\hat{y}_j \Big| x_j^*; \theta, \mu_{\hat{y}_j}, \sigma^2_{\hat{y}_j}) = \frac{p(x_j^* \Big| \hat{y}_j)p(\hat{y}_j)}{\sum_{y\in\{1,0\}} p(x_j^*|y)p(y)} = \frac{\mathcal{N}(h_\theta(x_j^*) \Big| \mu_{\hat{y}_j}, \sigma^2_{\hat{y}_j})p(\hat{y}_j)}{\sum_{y\in\{1,0\}} \mathcal{N}(h_\theta(x_j^*) \Big| \mu_y, \sigma_y^2)p(y_j)}, \tag{5}$$

where $p(\hat{y}_j)(p(y))$ is the prior probability computed on labeled dataset $\mathcal{D}_l$.

The second component of the co-EM-LR model is how to retrain the LR model based on labeled dataset $\mathcal{D}_l$ and probabilistically labeled dataset $\mathcal{D}_u$ with class probabilities $p(\hat{y}_j \,|x_j^*)$. This is equivalent to building a semi-supervised LR classifier. First, we assign every unlabeled instance a crisp label by $y_j^* = \max_{\hat{y}_j} p(\hat{y}_j \,|x_j^*) - \min_{\hat{y}_j} p(\hat{y}_j \,|x_j^*)$. Second, we define a weight as w_j = $w_j = \max_{\hat{y}_j} p(\hat{y}_j \,|x_j^*) - \min_{\hat{y}_j} p(\hat{y}_j \,|x_j^*)$ which is used to measure the degree of reliability for each unlabeled instance x_j^*, and if $w_j > \eta$, the unlabeled instance x_j^* will be selected for training in the following process. Finally, a pseudo-labeled dataset, denoted as $\mathcal{D}'_u = \langle(x_1^*, y_1^*, w_1), \cdots(x^*_{|\mathcal{D}'_u|}, y^*_{|\mathcal{D}'_u|}, w_{|\mathcal{D}'_u|})\rangle$, is generated according to the selected unlabeled instances in $\mathcal{D}_u$.

In general, given a parameter θ, LR model expresses the posterior probabilities for an instance x_i in a compact form as:

$$p'(y_i|x_i;\, \theta) = (h_\theta(x_i))^{y_i}(1 - h_\theta(x_i))^{1-y_i} \tag{6}$$

Algorithm 1 co-EM-LR

```
Input: $\mathscr{D}_{\mathrm{l}}$: Labeled dataset; $\mathscr{D}_{\mathrm{u}}$: Unlabeled dataset;
Output: The probabilities $p'(y_j\ |x_j^*)$, $y_j \in \{1, 0\}$ for each instance $x_j^* \in \mathscr{D}_{\mathrm{u}}$;
1: Train an initial LR model governed by $\theta_2$ on view $V_2$ of $\mathscr{D}_{\mathrm{l}}$;
2: repeat
3:   for $v = 1$ to 2 do
4:     Obtain $\mathscr{D}_{\mathrm{u}}^1$ by $h_{\theta\bar{v}}(x_j^*)$ on complementary view $V_{\bar{v}}$; $\mathscr{D}_{\mathrm{u}}^0 = \mathscr{D}_{\mathrm{u}} \backslash \mathscr{D}_{\mathrm{u}}^1$;
5:     Estimate $\mu_1$, $\mu_0$, $\sigma_1^2$ and $\sigma_0^2$ according to Eqs. (2) and (3);
6:     Estimate $p(\widehat{y}_j\ |x_j^*)$, $\forall\, x_j^* \in \mathscr{D}_{\mathrm{u}}$ by $\theta_{\bar{v}}$ according to Eq. (5);
7:     Generate a pseudo-labeled dataset $\mathscr{D}'_{\mathrm{u}}$;
8:     Update $\theta_v$ by maximizing Eq. (7) using SGD method with a smoothing factor $\Lambda_j$;
9:   end for
10: until converge
11: return Probability $p'(y_j\ |x_j^*)$, $y_j \in \{1, 0\}$, $\forall\, x_j^* \in \mathscr{D}_{\mathrm{u}}$, computed by Eq. (6);
```

As a result, given labeled dataset $\mathscr{D}_{\mathrm{l}}$ and pseudo-labeled dataset $\mathscr{D}'_{\mathrm{u}}$, to train the semi-supervised LR model can be performed by maximizing the following log-likelihood:

$$l(\theta) = \log(\prod_{x_i \in \mathscr{D}_{\mathrm{l}}} p'(y_i|x_i;\, \theta)^{\wedge i} \prod_{x_i \in \mathscr{D}'_{\mathrm{u}}} p'(y_j^* \big| x_j^*;\, \theta)^{\wedge j}), \tag{7}$$

where $\Lambda_i = 1$ and $\Lambda_j = \lambda * w_j$ are the smoothing factors to measure the contributions of every labeled instance x_i in $\mathscr{D}_{\mathrm{l}}$ and every crisp labeled instance x_j^* in $\mathscr{D}'_{\mathrm{u}}$, respectively. Specifically, the Stochastic Gradient Descent (SGD) (Mandt et al., 2017) method is employed to update parameters for increasing the log-likelihood through one instance at a time. The computational details will be presented in Appendix B.

Algorithm 1 presents a sketch of our co-EM-LR model. Lines 3–9 address the multi-view learning, another important component within our model. Initially, when $v = 1$, we utilize the $\theta_{\bar{v}} = (\bar{v} = 2)$ to split unlabeled dataset $\mathscr{D}_{\mathrm{l}}$ into a positive set and a negative set, since an initial LR model governed by θ_2 has been built in Line 1. Then, we update parameters, i.e., μ and σ^2, of the generative model on the complementary view $V_{\bar{v}}$, and thus compute predicted probabilities of every unlabeled instance. Afterwards, a pseudo-labeled dataset $\mathscr{D}'_{\mathrm{u}}$ is generated. In line 8, we invoke a semi-supervised LR training to update parameters of the current view θ_v based on $\mathscr{D}_{\mathrm{l}}$ and $\mathscr{D}'_{\mathrm{u}}$. In this way, parameters of two views, i.e., θ_1 and θ_2, are updated alternatively with each other. Note that the loop achieves convergence when the maximum difference between new and old θ values is smaller than a threshold (e.g., 0.0001 is set in our experiments).

6. Experimental results

In this section, we first demonstrate the effectiveness of the proposed co-EM-LR model on the real-world travel data. Next, we validate that two complementary views can mutually improve the performance with each other and finally offer fast convergence. Last but not the least, we evaluate the importance of every variable to discover the key factors that affect online purchase decisions.

6.1. Experimental setup

Baseline Methods. The proposed co-EM-LR approach adopts the LR as the base classifier and trains the classifier in a semi-supervised and multi-view learning way. So we first employ the following three methods for the performance comparison.

- *co-EM* (Fan et al., 2018). co-EM is a semi-supervised learning method that assumes a Gaussian generative process on unlabeled data.
- *co-Training* (Yu et al., 2011). co-Training is a multi-view learning method that assumes each example is described by two different and complementary feature sets. Here, the LR model is also used as the base classifier of co-Training.
- *LR*. Logistic Regression is the base classifier of our co-EM-LR model and also is one of the most widely-used classifier models in the purchase prediction (Van den Poel & Buckinx, 2005; Kim et al., 2003).

In addition, we employ two famous ensemble classification models: Random Forest (RF) and Gradient Boosting Decision Tree (GBDT). They often exhibit an excellent performance on various prediction tasks including the purchase prediction (Li et al., 2015; Volkovs, 2015) and other highly-related tasks such as the repeat buyer prediction in e-commerce (Liu et al., 2016). In total, five competitive methods are used in the experiments for the purpose of performance comparisons. Specifically, we use 10-fold cross-validation to provide robust results for different methods. We randomly split D_1, D_2 and D_3 into two parts, respectively, 10% of which as the training set $\mathscr{D}_{\mathrm{l}}$ and the rest set as the test set $\mathscr{D}_{\mathrm{u}}$ (the class size distribution holds). Our co-EM-LR, co-EM and co-Training are implemented in Python by ourselves. We select instances having 10% highest and 10% smallest decision function values from $\mathscr{D}_{\mathrm{u}}$ into $\mathscr{D}_{\mathrm{l}}$ in co-Training. The same as co-EM-LR, we set the threshold for convergence as 0.0001 in the setting of co-Training and co-EM. Besides, LR, RF and GBDT are implemented by scikit-learn[2], a machine learning library for Python. We adopt default parameters setting of scikit-learn in LR, RF and GBDT.

Evaluation Metrics. Since the ground-truth is known, we adopt the widely-used metrics such as precision (P), recall (R), and F-measure (F) for the performance evaluation (Brzezinski et al., 2018). Specifically, we focused on the ability of detectors to recognize the sessions with online purchases, where

$$P = \frac{TP}{TP + FP}, \quad R = \frac{TP}{TP + FN}, \quad F = \frac{2PR}{P + R}, \tag{8}$$

with TP being the number of truly identified sessions with online purchases, TN the number of truly identified sessions without online purchases, FP the number of wrongly identified sessions with online purchases, and FN the number of missed sessions without online purchases. In general, P and R highlight the accuracy and completeness of a classifier, respectively, and F provides a global view.

Parameter Analysis. Here, we study the impact of thehyper-parameter λ on the performance of our co-EM-LR model. Fig. 6 shows the classification performance in terms of F for online purchase prediction given different weights of unlabeled dataset $\mathscr{D}_{\mathrm{u}}$, where η is empirically set to 0.4 to select convincing unlabeled instances for training. We find that $\lambda = 0.6$ is a relatively robust choice, becoming the default setting in our experiments. It validates the important role of unlabeled dataset.

Goodness-of-Fit Test. The goodness-of-fit measures how well our co-EM-LR model fits the training data. Since co-EM-LR employs the Logistic Regression as its base model, the statistical methods designed for evaluating the goodness-of-fit of Logistic Regression are appropriate for evaluating the co-EM-LR model. Here, we select the Hosmer-Lemeshow statistic (Hosmer et al., 1988), a chi-squared test, to access the goodness-of-fit of co-EM-LR. First, we set up the *null hypothesis* assuming that there is no significant difference between the observed and the predicted values. We then randomly sample 10% instances as observed data and compute the significant level p-value on three datasets. We repeat 10 times and takes the median p-value as the final results. As a consequence, all the p-values are more than 0.05 on three datasets. To be specific, the median p-values are 0.138, 0.156 and 0.145 on D_1, D_2 and D_3 respectively. According to Hosmer et al. (1988), the null hypothesis is accepted when p-value > 0.05. In turn, this implies that the co-EM-LR model shows a good fit based on the chi-square of Hosmer-Lemeshow goodness-of-fit statistics.

[2] http://scikit-learn.org/stable/

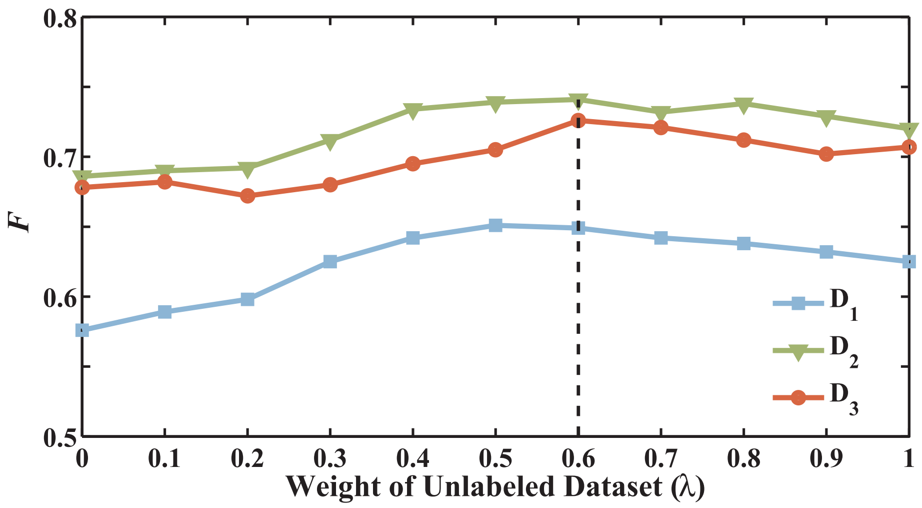


Fig. 6. Impact of the parameter on the co-EM-LR model.

6.2. Overall performance comparison

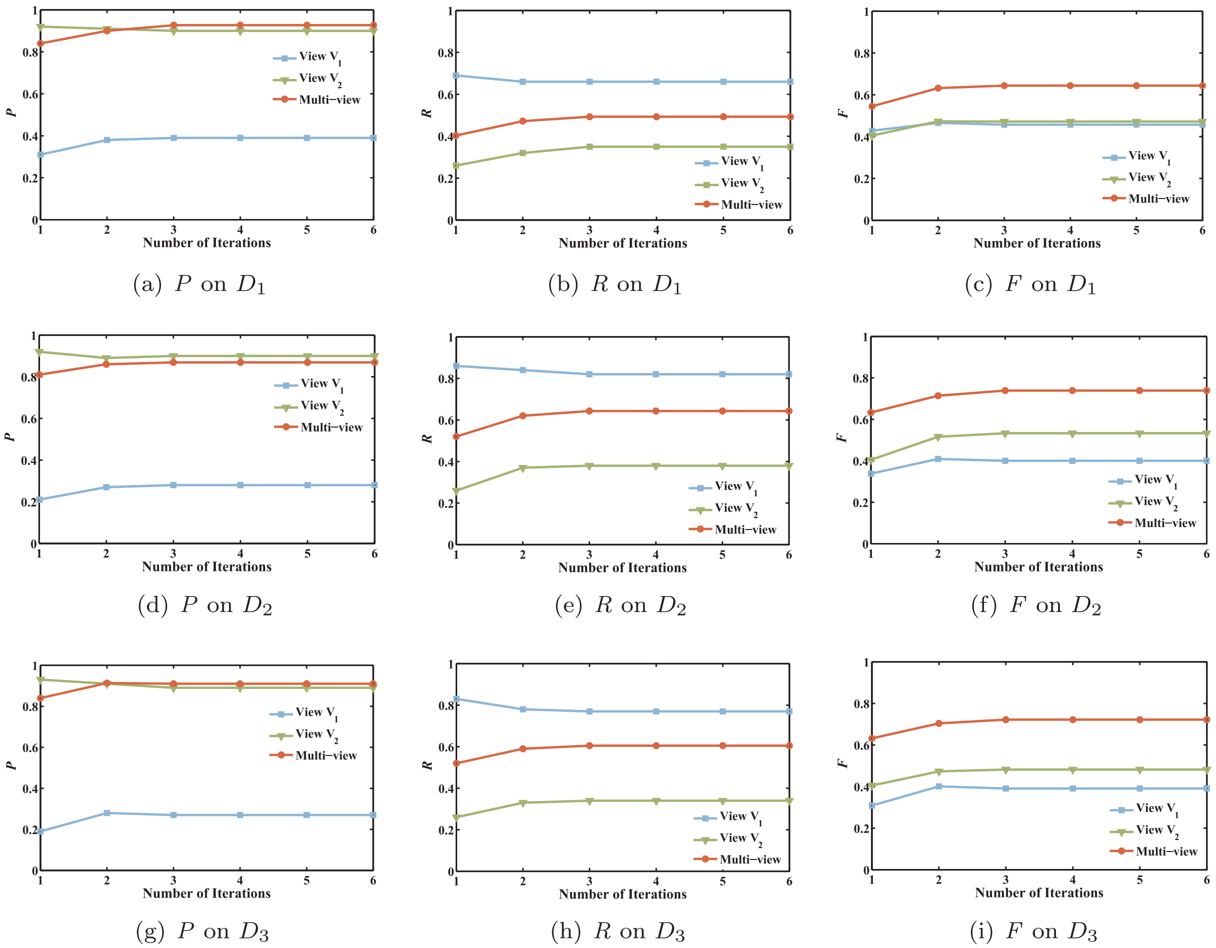
In this subsection, we present the overall performance comparison between the co-EM-LR model and five baseline methods to demonstrate the effectiveness of the co-EM-LR model for online purchase prediction. Table 5 reports the results of overall performance comparison, where the column “Overall” lists the mean metric values and their standard deviations on three datasets and the row “Average” is the mean metric values of all methods on different user segmentations. Note that the mean metric value and its standard deviation is computed on the 10-fold cross-validation (Lo et al., 2016) results.

It is clear to observe that the co-EM-LR is noticeably superior to baseline methods in terms of *R* but slightly inferior to some baseline methods in terms of *P*. However, the global performance of co-EM-LR, indicated by *F*, surpasses all of competitive methods. In particular, our co-EM-LR method makes an improvement rate of 14.6%–28.3% on *R* and of 5.6%–11.5% on *F*, compared with the second-best method on three datasets. We have to stress that *the recall is much more important than the precision* in the application of purchase prediction. Image tens of thousands of customers are visiting the e-tourism website every day and the call center team wants to call them all for improving the visit-to-purchase conversion rate, but it is impossible (Navío-Marco et al., 2018). Hence, it is expected that customers with good chances to be a buyer are always in their selection (Moe, 2006; Pavlou & Fygenson, 2006). From this point of view, the higher *R* values of our co-EM-LR model make it more valuable to the e-tourism platform. Nevertheless, the co-EM-LR model owns an acceptable performance over *P*, i.e., *P* values of co-EM-LR are all more than 85% on three datasets. When we examine the standard deviation of 10-fold cross-validation trails, our co-EM-LR also shows the best of all in stability. Among five baseline methods, two ensemble classifiers RF and GBDT indeed perform better than other baselines. Perhaps this is why they were largely applied to the purchase prediction task (e.g., Kim et al., 2003; Volkovs, 2015). We further look inside the performance on three user groups. As shown by the “Average” rows of Table 5, regardless of any approach, the performance on recent-visited users (U_3) is the winner, followed by ever-visited users (U_2) and first-time users (U_1). The reason for this is pretty obvious: the recent-visited user has more clickstream data which helps

Table 5
Overall Comparison of Purchase Prediction Performance. The highest metric values among six methods are in bold.

Metric	Method	D_1				D_2				D_3			
		U_1	U_2	U_3	Overall	U_1	U_2	U_3	Overall	U_1	U_2	U_3	Overall
P	co-EM-LR	0.857	0.923	0.947	0.927 ± 0.006	0.835	0.861	0.874	0.869 ± 0.006	0.845	0.896	0.913	0.899 ± 0.005
	co-EM	**0.951**	**0.972**	**0.986**	**0.974** ± 0.007	0.876	0.931	0.956	0.920 ± 0.008	0.861	0.894	0.925	0.887 ± 0.009
	co-Training	0.827	0.896	0.926	0.864 ± 0.013	0.766	0.852	0.883	0.843 ± 0.012	0.879	0.943	0.978	0.952 ± 0.011
	LR	0.751	0.894	0.912	0.879 ± 0.028	0.853	0.882	0.913	0.861 ± 0.020	0.793	0.836	0.886	0.882 ± 0.021
	RF	0.891	0.944	0.977	0.962 ± 0.019	**0.948**	0.979	0.985	0.980 ± 0.016	**0.965**	**0.978**	**0.982**	**0.975** ± 0.021
	GBDT	0.912	0.971	0.978	0.969 ± 0.011	0.941	**0.980**	**0.987**	**0.982** ± 0.012	0.948	0.974	0.979	0.973 ± 0.014
	Average	0.865	0.933	0.954	–	0.870	0.914	0.933	–	0.882	0.920	0.944	–
R	co-EM-LR	**0.357**	**0.462**	**0.568**	**0.493** ± 0.005	0.423	**0.547**	**0.719**	**0.643** ± 0.008	0.358	**0.569**	**0.716**	**0.605** ± 0.009
	co-EM	0.181	0.201	0.221	0.196 ± 0.009	0.198	0.291	0.301	0.281 ± 0.008	0.182	0.217	0.230	0.203 ± 0.009
	co-Training	0.303	0.383	0.431	0.356 ± 0.009	**0.542**	0.476	0.578	0.492 ± 0.011	0.298	0.337	0.349	0.321 ± 0.009
	LR	0.271	0.351	0.421	0.336 ± 0.022	0.372	0.461	0.487	0.424 ± 0.019	0.291	0.378	0.418	0.347 ± 0.021
	RF	0.328	0.411	0.441	0.394 ± 0.015	0.365	0.468	0.521	0.465 ± 0.019	0.381	0.473	0.489	0.501 ± 0.018
	GBDT	0.335	0.429	0.481	0.422 ± 0.016	0.391	0.502	0.545	0.501 ± 0.014	**0.427**	0.521	0.553	0.528 ± 0.017
	Average	0.296	0.373	0.427	–	0.382	0.458	0.525	–	0.323	0.416	0.460	–
F	co-EM-LR	**0.509**	**0.618**	**0.717**	**0.644** ± 0.005	0.563	0.670	**0.789**	**0.739** ± 0.006	0.507	**0.697**	**0.809**	**0.723** ± 0.007
	co-EM	0.304	0.333	0.361	0.326 ± 0.007	0.323	0.443	0.458	0.431 ± 0.007	0.300	0.349	0.367	0.330 ± 0.009
	co-Training	0.444	0.537	0.588	0.504 ± 0.012	**0.587**	**0.689**	0.709	0.660 ± 0.011	0.445	0.497	0.514	0.480 ± 0.010
	LR	0.398	0.504	0.576	0.486 ± 0.017	0.518	0.606	0.635	0.568 ± 0.014	0.426	0.521	0.568	0.498 ± 0.012
	RF	0.479	0.573	0.608	0.559 ± 0.018	0.438	0.520	0.581	0.631 ± 0.019	0.429	0.503	0.565	0.635 ± 0.018
	GBDT	0.490	0.595	0.645	0.588 ± 0.012	0.552	0.664	0.702	0.663 ± 0.015	**0.583**	0.679	0.705	0.685 ± 0.015
	Average	0.437	0.527	0.583	–	0.497	0.599	0.646	–	0.448	0.541	0.588	–

Note: (1) U_1, U_2 and U_3 denote first-time visitors, ever-visited users and recent-visited users, respectively.



(a) P on D_1 (b) R on D_1 (c) F on D_1

(d) P on D_2 (e) R on D_2 (f) F on D_2

(g) P on D_3 (h) R on D_3 (i) F on D_3

Fig. 7. Effect of the multi-view learning on co-EM-LR performance.

to construct the complete view of variables.

6.3. Effect of the multi-view learning

In this experiment, we try to reveal the run-time mechanism and validate the effectiveness of multi-view learning inside our co-EM-LR model. To this end, for each iteration of the co-EM-LR, we extract the intermediate values of $\theta = (\theta_1, \theta_2)$ and predict the label of every instance by using Eq. (1). Then, by comparing with ground-truth, we compute values of three evaluation metrics for each iteration. Fig. 7 reports the results on three datasets in terms of P, R and F when the number of iteration is set to 6, where "View V_1" ("View V_2") represents the prediction is only performed by θ_1 (θ_2) and "Multi-view" denotes the prediction is made by θ.

The results of Fig. 7 yield the following conclusions. First, the characteristics of two views are different and complementary. That is, the view V_1 corresponding to recency clickstream often delivers low P values but high R values, which implies the classifier on V_1 tends to label a great many of sessions as "will be purchase". In contrast, the classifier on V_2 corresponding to current clickstream is much more conservative, which tends to produce a list of sessions being likely to purchase once it labels "will be purchase". However, both P and R curves of the multi-view learning are sandwiched between those of V_1 and V_2. This implies that our co-EM-LR model utilizes the multi-view learning (Zhao et al., 2017b)to balance the precision and recall of two single views and ultimately improves the global classification accuracy. Second, our co-EM-LR model converges fastly. All curves of Fig. 7 show a sharp increase in the second iteration, and then remain broadly flat until the convergence. The increase after the first iteration is caused by the complementary effect which are first perceived in the second round, which is consistent with the result of multi-view semi-supervised approach (Fan et al., 2018). In practice, by setting the threshold of the change of θ as 0.0001, our co-EM-LR model usually converges after around 10 iterations. The flat region of Fig. 7 indicates the 10 iterations are more than sufficient.

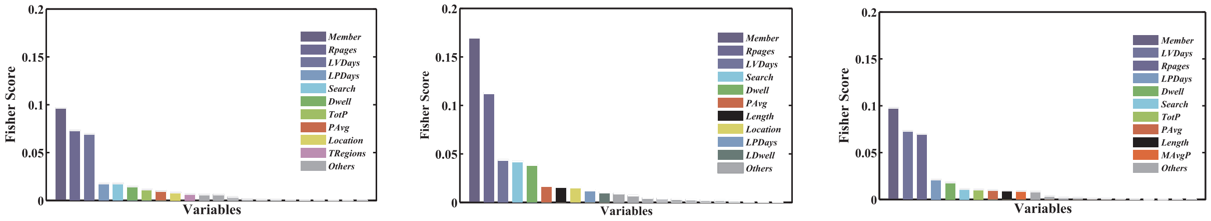
6.4. Importance of variables

Inspired by the study (Young Kim & Kim, 2004), importance of variables could provide managerial implications for future online marketing. Hence, we attempt to evaluate the importance of every variable introduced in Section 4.2, by the impurity measure commonly used in the classification model (Li et al., 2017). To this end, given labeled dataset D_l, then variables set V is regard as the combination of V_1 and V_2 variable sets for all instances in D_l. Here, we adopt *Fisher Score* (Zhang & Parhi, 2018) as the impurity measure for the evaluation of variable set V. Specifically, for the f-th variable, let μ_y^{*f} and σ_y^{*f} denote the mean and standard deviation of the y-th class, respectively. Let μ^{*f} and σ^{*f} denote the mean and standard deviation of dataset D_l corresponding to the f-th variable, respectively. As a result, the Fisher score of the f-th variable is computed as:

$$F(V^f) = \frac{\sum_{y=0}^{1} n_y(\mu_y^{*f} - \mu^{*f})^2}{(\sigma^{*f})^2} \tag{9}$$

where $(\sigma^{*f})^2 = \sum_{y=0}^{1} n_y(\sigma_y^{*f})^2$, and n_y is the number of instances in class y. A higher Fisher score value indicates the variable has more discriminative power (i.e., more important).

Fig. 8 shows evaluation results of all variables by using Fisher score



(a) D_1 (b) D_2 (c) D_3

Fig. 8. The importance of all variables on Fisher score.

on three datasets, where color bars denote top-10 important features, whereas unified gray bars devote other variables having a low level on Fisher score. As can be seen, top-10 important variables on three datasets are very similar. Furthermore, top-3 important variables on three datasets are either "$Mem\text{-}ber_i \rightarrow LVDays_i \rightarrow RPages_j$" or "$Member_i \rightarrow RPages_j \rightarrow LVDays_i$", which indicates that demographics of users, percentage of pages containing travel product dis-plays in current sessions and recent-visited time in recency sessions are the key factors that significantly affect online purchase decision. Besides, current browsing behaviors (e.g., $Search_j$ and $Dwell_j$), recency browsing behavior (e.g., $LPDays_i$), cost of finance (e.g., $PAvg_j$) and users' geographic information (e.g., $Location_j$) also have slight effects on online purchase decisions.

7. Concluding remarks

With the rapid uptake of e-tourism industry, understanding the online purchasing behavior of customers and thus devising the purchase prediction strategies in the case of e-tourism application are of substantial interest to decision makers (Navío-Marco et al., 2018). In this paper, we study this problem by taking full advantage of a real-life e-tourism data provided by a large online travel agency platform in China. We perform a quantitative analysis to address quite a few interesting characteristics of purchase patterns. Based on this analysis, we construct a bank of variables for the current clickstream and recency clickstream, which serve as the feature set of the classification model. Our paper contributes to the existing body of knowledge on the design of purchase prediction models. Specifically, we present an advanced purchase prediction model *co-EM-LR* that combines the semi-supervised learning and the multi-view learning. Also, our co-EM-LR adopts the regression model as its base classifier to provide the good interpretation of control variables. Through the extensive experiments, we find that our co-EM-LR model yields significant prediction performance advantages over five competitive methods. In particular, the co-EM-LR model can remarkably improve the recall which is crucial to increase the marketing opportunities of converting visitors to buyers. This advantage will help to single out potential buyers from a huge number of visitors, and will possess a source of huge value to big e-commerce platforms that require the follow-up of many thousands customers every day.

This study has several limitations and opens up opportunities for further research. First, the prediction model is performed mainly according to the customers' browsing behaviors on product pages. Unlike the traditional e-commerce websites, the content of pages in e-tourism platform is not limited to product descriptions, whereas there are a large number of pages about travel notes and strategies. Also, to understand the content of user-generated travel notes is not a trivial task. Hence, to model such complex browsing behavior and incorporate it into the purchase prediction is an interesting avenue for future research. Second, despite our current work provides a solid foundation on the three-staged purchase decision model (Zhang & Wedel, 2009; Wan et al., 2017), a comprehensive examination of how customers make online purchase decisions on e-tourism websites is still worthy of further studies.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 71571093, Grant 91646204, Grant 71372188, and the National Center for International Joint Research on E-Business Information Processing under Grant 2013B01035.

Appendix A. Definitions of some intricate variables

As a supplement to Table 2, we introduce the definition and computational details of four variables: $Location_j$, $TRegions_j$, $RTypes_j$ and $RelatedPages_j$.

To reduce the financial and time cost, a customer usually choose a departure city near to the city that he/she lives in (inferred by IP address) to start the trip. On this account, we use the structural similarity upon a tree to define a semantic relationship between two cities. In detail, we utilize the hierarchical structure from United Nations geoscheme[3] to construct a geographical tree including all departure cities of product and customers' live cities in the real-life e-tourism dataset. Fig. A.9 shows an illustrative example of this hierarchical structure. Then, the distance between cities is transformed to the similarity of two nodes in this geographical tree.

Definition A.1 ($Location_j$). $Location_j$ is average semantic similarity between user's living city and departure cities of products that the user has clicked in session j. Let C denotes the living city and C_j denotes the set of departure cities extracted from session j.

$$Location_j = \frac{1}{|C_j|}\sum_{C_{jk}\in C_j}\frac{2Depth(C\cap C_{jk})}{Depth(C) + Depth(C_{jk})}, \tag{A.1}$$

where $Depth(\cdot)$ is the depth of a node (the depth of root is 1), and $C \cap C_{jk}$ denotes the last common ancestor of two nodes in the geographical tree.

As introduced in Section 4.1, travel region and travel type are the two angles of classification to describe travel products in the session. Generally speaking, customers who have strong purchase intents are likely to browse travel products taking his interested travel region as destination or interested travel type as target. Hence, we utilize entropy to measure the degree of dispersal or concentration of travel region and travel type

[3] http://en.wikipedia.org/wiki/United_Nations_geoscheme

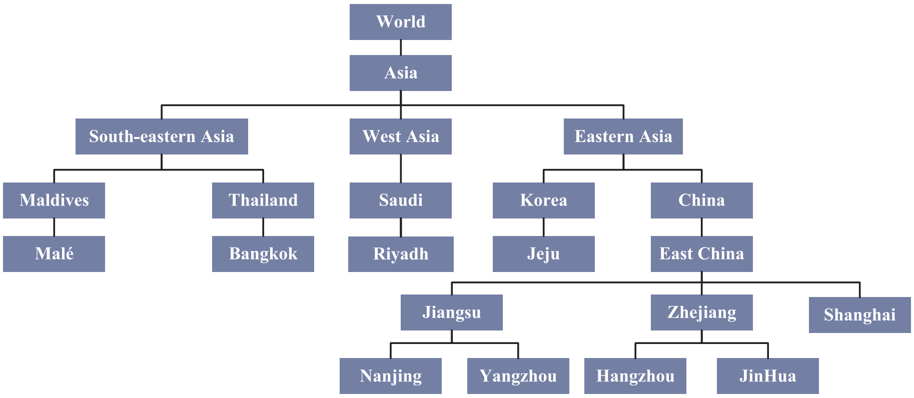


Fig. A.9. Illustration of the geographical tree.

associated with products in the session. **Definition A.2** (*TRegions*$_j$). *TRegions*$_j$ is the entropy of travel region distribution in session j. Let T_j denotes the set of travel regions extracted from session j, and for each travel region t in T_j, travel region distribution π assigns a probability π(t) which is calculated by the ratio between the number of travel region t and the number of all travel regions in T_j. We have:

$$TRegions_j = -\sum_{t\in T_j} \pi(t)\log_2 \pi(t). \tag{A.2}$$

The minimum value 0 of *TRegions*$_j$ indicates that the customer browses the travel products which are in the same travel region in session j. Whereas the bigger value of *TRegions*$_j$ indicates that the customer browses travel products which are in various travel regions in session j. Analogously, *PTypes*$_j$ is defined as the entropy of travel type distribution in session j, and the computational details is consistent with definition of *TRegions*$_j$.

Customers who have strong purchase intents are more likely to browse related product pages to get rich information which can help to make online purchase decisions. In this real-life e-tourism dataset, category page is one type of pages which usually displays a set of travel products according to a specific topic (e.g., Beijing tour, amusement park and sea island, etc). Therefore, apart from product pages themselves, category pages in the session are also important for identifying customers' purchase intents.

Definition A.3 (*RPages*$_j$). *RPages*$_j$ is the percentage of pages containing travel product displays in session j. Let *Category*$_j$ and *Product*$_j$ be the number of category pages and travel product pages in *session j*, respectively. We have:

$$RPages_j = \frac{Category_j + Product_j}{Length_j}, \tag{A.3}$$

where *Length*$_j$ is the length of *session j* as shown in Table 2.

Appendix B. Computational details of maximizing the log-likelihood

SGD is one of the simplest and most popular stochastic optimization methods, which can be used to optimize any convex function based only on a finite sampled training set. Consider a labeled instance x_i in $\mathscr{D}_{\mathrm{l}}$ and a crisp labeled instance x_j^* in $\mathscr{D}'_{\mathrm{u}}$, a loss function $J(\theta)$ in SGD is defined as:

$$J(\theta) = -\frac{1}{n_l}\log\left(\prod_{i=1}^{n_l} p'(y_i|x_i;\theta)\right) - \frac{1}{n_u}\log\left(\prod_{j=1}^{n_u} p'\left(y_j^*\middle|x_j^*;\theta\right)^{\wedge j}\right), \tag{B.1}$$

where, $n_l(1 \le n_l \le |\mathscr{D}_{\mathrm{l}}|)$ and $n_u(1 \le n_u \le |\mathscr{D}'|_u)$ denote the number of stochastic mini-batch instances in $\mathscr{D}_{\mathrm{l}}$ and $\mathscr{D}'$, respectively. We commonly set $n_l = 5\%|\mathscr{D}_{\mathrm{l}}|$ and $n_u = 5\%|\mathscr{D}'_{\mathrm{u}}|$ in this process.

Since the loss function $J(\theta)$ defines a negative correlation with log-likelihood function $l(\theta)$, maximizing $l(\theta)$ could be converted to minimizing $J(\theta)$ by using SGD. Thus, the updated θ at each epoch t is

$$\theta^{t+1} = \theta^t - \alpha\tfrac{\partial J(\theta)}{\partial\theta}$$

$$= \theta^t - \alpha\left(\tfrac{1}{n_l}\textstyle\sum_{i=1}^{nl}(h_\theta(x_i) - y_i)x_i + \tfrac{1}{n_u}\sum_{j=1}^{n_u}(h_\theta(x_j^*) - y_j^*)\wedge_j x_j^*\right),$$

where α is the learning rate to control the magnitude of the changes to parameters. The iterative procedure will be suspended if the value of loss function $J(\theta)$ is less than the threshold s. Here, we set the parameters of SGD as follows:$\alpha = 0.01$ and $s = 0.0001$, respectively.

References

Ayanso, A., Yoogalingam, R., 2009. Profiling retail web site functionalities and conversion rates: a cluster analysis. Int. J. Electron. Commerce 14, 79–114.

Bacharach, S.B., 1989. Organizational theories: some criteria for evaluation. Acad. Manage. Rev. 14, 496–515.

Bergmann, R., Ludbrook, J., Spooren, W.P., 2000. Different outcomes of the Wilcoxon-Mann-Whitney test from different statistics packages. Am. Statist. 54, 72–77.

Berthon, P.R., Pitt, L.F., Plangger, K., Shapiro, D., 2012. Marketing meets Web 2.0, social media, and creative consumers: implications for international 740 marketing strategy. Bus. Horiz. 55, 261–271.

Brown, M., Pope, N., Voges, K., 2003. Buying or browsing? An exploration of shopping orientations and online purchase intention. Eur. J. Mark. 37, 1666–1684.

Brzezinski, D., Stefanowski, J., Susmaga, R., Szczech, I., 2018. Visual-based analysis of classification measures and their properties for class imbalanced problems. Inf. Sci. 462, 242–261.

Bucklin, R.E., Sismeiro, C., 2003. A model of web site browsing behavior estimated on clickstream data. J. Mark. Res. 40, 249–267.

Calabrese, F., Ferrari, L., Blondel, V.D., 2015. Urban sensing using mobile phone network data: a survey of research. ACM Comput. Surv. 47, 20 25:1–25.

Chen, Y.-L., Kuo, M.-H., Wu, S.-Y., Tang, K., 2009. Discovering recency, frequency, and monetary (RFM) sequential patterns from customers' purchasing data. Electron. Commer. Res. Appl. 8, 241–251.

Dong, X., Suhara, Y., Bozkaya, B., Singh, V.K., Lepri, B., Pentland, A.S., 2018. Social bridges in urban purchase behavior. ACM Trans. Intell. Syst. Technol. 9, 1–33.

Fan, A., Chen, L., Chen, G., 2018. A multi-view semi-supervised approach for task-level web search success evaluation. Inf. Sci. 430, 554–566.

Ge, Y., Xiong, H., Tuzhilin, A., Liu, Q., 2014. Cost-aware collaborative filtering for travel tour recommendations. ACM Trans. Inf. Syst. 32, 31 4:1–4.

He, J., Liu, H., Xiong, H., 2016. Socotraveler: travel-package recommenda-tions leveraging social influence of different relationship types. Inf. Manage. 53, 934–950.

Hernández, B., Jiménez, J., Martín, M.J., 2010. Customer behavior in electronic commerce: the moderating effect of e-purchasing experience. J. Bus. Res. 63, 964–971.

Hosmer, D.W., Lemeshow, S., Klar, J., 1988. Goodness-of-fit testing for the logistic regression model when the estimated probabilities are small. Biometrical J. 30, 911–924.

Huang, C.D., Goo, J., Nam, K., Yoo, C.W., 2017. Smart tourism technologies in travel planning: the role of exploration and exploitation. Inf. Manage. 54, 757–770.

Iwanaga, J., Nishimura, N., Sukegawa, N., Takano, Y., 2016. Estimating product-choice probabilities from recency and frequency of page views. Knowl.-Based Syst. 99, 157–167.

Jerath, K., Fader, P.S., Hardie, B.G.S., 2011. New perspectives on customer "death" using a generalization of the Pareto/NBD model. Mark. Sci. 30, 866–880.

Khan, M.U.S., Khalid, O., Huang, Y., Ranjan, R., Zhang, F., Cao, J., Veeravalli, B., Khan, S.U., Li, K., Zomaya, A.Y., 2017. MacroServ: a route recommendation service for large-scale evacuations. IEEE Trans. Serv. Comput. 10, 589–602.

Kim, E., Kim, W., Lee, Y., 2003. Combination of multiple classifiers for the customer's purchase behavior prediction. Decis. Support Syst. 34, 167–175.

Kooti, F., Lerman, K., Aiello, L.M., Grbovic, M., Djuric, N., Radosavljevic, V., 2016. Portrait of an online shopper: Understanding and predicting consumer behavior. In: Proceedings of the 9th ACM International Conference on Web Search and Data Mining. ACM, pp. 205–214.

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R.P., Tang, J., Liu, H., 2017. Feature selection: a data perspective. ACM Comput. Surv. 50, 45 94:1–94.

Li, D., Zhao, G., Wang, Z., Ma, W., Liu, Y., 2015. A method of purchase prediction based on user behavior log. In: Proceedings of the 15th IEEE International Conference on Data Mining Workshop. IEEE, pp. 1031–1039.

Liu, Q., Chen, E., Xiong, H., Ge, Y., Li, Z., Wu, X., 2014. A cocktail approach for travel package recommendation. IEEE Trans. Knowl. Data Eng. 26, 278–293.

Liu, J., Liu, B., Liu, Y., Chen, H., Feng, L., Xiong, H., Huang, Y., 2018. Personalized air travel prediction: a multi-factor perspective. ACM Trans actions on Intelligent. Syst. Technol. 9, 1–30.

Liu, G., Nguyen, T.T., Zhao, G., Zha, W., Yang, J., Cao, J., Wu, M., Zhao, P., Chen, W., 2016. Repeat buyer prediction for e-commerce. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, pp. 155–164.

Lo, C., Frankowski, D., Leskovec, J., 2016. Understanding behaviors that lead to purchasing: a case study of pinterest. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, pp. 531–540.

Lu, Y., Zhao, L., Wang, B., 2010. From virtual community members to C2C e-commerce buyers: trust in virtual communities and its effect on consumers' purchase intention. Electron. Commer. Res. Appl. 9, 346–360.

Ludwig, S., De Ruyter, K., Friedman, M., Brüggen, E.C., Wetzels, M., Pfann, G., 2013. More than words: the influence of affective content and linguistic 37 style matches in online reviews on conversion rates. J. Mark. 77, 87–103.

Lv, Z., Jin, Y., Huang, J., 2018. How do sellers use live chat to influence consumer purchase decision in China? Electronic Commerce Res. Appl. 28, 102–113.

Mandt, S., Hoffman, M.D., Blei, D.M., 2017. Stochastic gradient descent as approximate bayesian inference. J. Mach. Learn. Res. 18, 4873–4907.

McKnight, D.H., Choudhury, V., Kacmar, C., 2002. The impact of initial consumer trust on intentions to transact with a web site: a trust building model. J. Strategic Inf. Syst. 11, 297–323.

Moe, W.W., 2006. An empirical two-stage choice model with varying decision rules applied to internet clickstream data. J. Mark. Res. 43, 680–692.

Moe, W.W., Fader, P.S., 2004. Dynamic conversion behavior at e-commerce sites. Manage. Sci. 50, 326–335.

Mokryn, O., Bogina, V., Kuflik, T., 2019. Will this session end with a purchase? Inferring current purchase intent of anonymous visitors. Electron. Commer. Res. Appl. 34, 100836. https://doi.org/10.1016/j.elerap.2019.100836.

Morisada, M., Miwa, Y., Dahana, W.D., 2019. Identifying valuable customer segments in online fashion markets: an implication for customer tier programs. Electron. Commer. Res. Appl. 33, 100822. https://doi.org/10.1016/j.elerap.2018.100822.

Mudambi, S.M., Schuff, D., 2010. What makes a helpful review? A study of customer reviews on Amazon.com. MIS Quarterly 34, 185–200.

Navío-Marco, J., Ruiz-Gómez, L.M., Sevilla-Sevilla, C., 2018. Progress in information technology and tourism management: 30 years on and 20 years after the internet-revisiting buhalis & law's landmark study about etourism. Tourism Manage. 69, 460–470.

Pavlou, P.A., Fygenson, M., 2006. Understanding and predicting electronic commerce adoption: an extension of the theory of planned behavior. MIS Quarterly 30, 115–143.

Schlosser, A.E., White, T.B., Lloyd, S.M., 2006. Converting web site visitors into buyers: how web site investment increases consumer trusting beliefs and online purchase intentions. J. Mark. 70, 133–148.

Sismeiro, C., Bucklin, R.E., 2004. Modeling purchase behavior at an ecommerce web site: a task-completion approach. J. Mark. Res. 41, 306–323.

Tkaczynski, A., Rundle-Thiele, S.R., Beaumont, N., 2009. Segmentation: a tourism stakeholder view. Tourism Manage. 30, 169–175.

Van den Poel, D., Buckinx, W., 2005. Predicting online-purchasing behaviour. Eur. J. Oper. Res. 166, 557–575.

Volkovs, M., 2015. Two-stage approach to item recommendation from user sessions. In: Proceedings of the 9th ACM International Conference on Recommender Systems Challenge. ACM, pp. 1–3.

Wan, M., Wang, D., Goldman, M., Taddy, M., Rao, J., Liu, J., Lymberopoulos, D., McAuley, J., 2017. Modeling consumer preferences and price sensitivities from large-scale grocery shopping transaction logs. In: Proceedings of the 26th ACM International Conference on World Wide Web. ACM, pp. 1103–1112.

Wen, Y.-T., Yeo, J., Peng, W.-C., Hwang, S.-W., 2017. Efficient keywordaware representative travel route recommendation. IEEE Trans. Knowl. Data Eng. 29, 1639–1652.

Young Kim, E., Kim, Y.-K., 2004. Predicting online purchase intentions for clothing products. Eur. J. Mark. 38, 883–897.

Yu, S., Krishnapuram, B., Rosales, R., Rao, R.B., 2011. Bayesian cotraining. J. Mach. Learn. Res. 12, 2649–2680.

Zhang, Z., Parhi, K.K., 2018. Muse: Minimum uncertainty and sample elimination based binary feature selection. IEEE Trans. Knowl. Data Eng. 1. https://doi.org/10.1109/TKDE.2018.2865778.

Zhang, J., Wedel, M., 2009. The effectiveness of customized promotions in online and offline stores. J. Mark. Res. 46, 190–206.

Zhao, J., Qu, Q., Zhang, F., Xu, C., Liu, S., 2017a. Spatio-temporal analysis of passenger travel patterns in massive smart card data. IEEE Trans. Intell. Transp. Syst. 18, 3135–3146.

Zhao, J., Xie, X., Xu, X., Sun, S., 2017b. Multi-view learning overview: recent progress and new challenges. Inf. Fusion 38, 43–54.