Knowledge-Based Systems 99 (2016) 157–167

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Estimating product-choice probabilities from recency and frequency of page views

Jiro Iwanaga[a], Naoki Nishimura[b], Noriyoshi Sukegawa[c,*], Yuichi Takano[d]

[a] Business Intelligence Deployment Center, NTT DATA Mathematical Systems Inc., 1F Shinanomachi Rengakan, 35 Shinanomachi, Shinjyuku-ku, Tokyo 166-0016, Japan
[b] Product Management Unit, Internet Business Development Division, Recruit Lifestyle Co., Ltd., GranTokyo SOUTHTOWER 1-9-2 Marunouchi, Chiyoda-ku, Tokyo 100-6640, Japan
[c] Department of Information and System Engineering, Faculty of Science and Engineering, Chuo University, 1-13-27 Kasuga, Bunkyo-ku, Tokyo 112-8551, Japan
[d] School of Network and Information, Senshu University, 2-1-1 Higashimita, Tama-ku, Kawasaki-shi, Kanagawa 214-8580, Japan

ARTICLE INFO

Article history:
Received 11 July 2015
Revised 1 February 2016
Accepted 4 February 2016
Available online 12 February 2016

Keywords:
Product choice
E-commerce
Clickstream data
Optimization
Recency
Frequency
Page view

ABSTRACT

This paper investigates the relationship between customers' page views (PVs) and the probabilities of their product choices on e-commerce sites. For this purpose, we create a probability table consisting of product-choice probabilities for all recency and frequency combinations of each customers' previous PVs. To reduce the estimation error when there are few training samples, we develop optimization models for estimating the product-choice probabilities that satisfy monotonicity, convexity and concavity constraints with respect to recency and frequency. Computational results demonstrate that our method has clear advantages over logistic regression and kernel-based support vector machine.

1. Introduction

An increasing number of companies are operating e-commerce sites that offer products or services via the Internet [40]. Customers can compare and purchase a variety of products on such sites without making a trip to a brick-and-mortar store. In addition, companies can exploit the detailed electronic information stored on e-commerce sites to build profitable relationships with customers [28]. In particular, clickstream data, which is a record of a visitor's page view (PV) history, is of demonstrated value for understanding consumer behavior [5,6,18,19,26].

The motivation behind our research is to investigate the relationship between customers' PVs and their product-choice probabilities. In other words, we explore the process of selecting products on the basis of clickstream data [25,42]. Since the information search process reflects the customer's concern about products, the results of our research can be used to help him/her go to a target page on an e-commerce site. Our research will also be useful in forecasting demand for inventory management [17].

The statistical estimation methods can be divided into two categories, i.e., parametric methods and nonparametric methods. Parametric methods typically force an estimation model to be a parametric function. In contrast, nonparametric methods do not assume any particular parametric form and, accordingly, have a high degree of freedom for modeling the relationship between customers' PVs and their product-choice probabilities.

The purpose of this paper is to propose a novel nonparametric method for estimating product-choice probabilities. It has been demonstrated that the recency and frequency of customers' previous purchases are key indicators for forecasting repeat purchases [12,13,20,32,33]. In view of these facts, our method utilizes the recency and frequency of each customer's previous PVs. Specifically, we create a probability table, consisting of product-choice probabilities for all recency and frequency combinations, through the use of past clickstream data. Although this approach can fully reflect the interaction between the frequency and recency of PVs, a probability estimated from a small number of training samples is subject to unavoidable estimation errors.

To resolve this problem, we exploit the properties of recency and frequency of PVs in the maximum likelihood estimation of the product-choice probabilities. To accomplish this, we develop

[*] Corresponding author. Tel.: +818091761821; fax: +81338171681.
E-mail address: sukegawa@ise.chuo-u.ac.jp, yy.n.s.goo@gmail.com (N. Sukegawa).

http://dx.doi.org/10.1016/j.knosys.2016.02.006
0950-7051/© 2016 Elsevier B.V. All rights reserved.

optimization models for estimating the product-choice probabilities that satisfy the monotonicity, convexity and concavity constraints with respect to recency and frequency.

We compare the predictive performance of the proposed method with those of the common methods of binary classification, i.e., logistic regression and kernel-based support vector machine (kernel SVM). Six typical features were used to represent recency or frequency in computational experiments, which thoroughly verified the effectiveness of these features and their combination for predictive purposes.

The advantages of our method are summarized as follows:

1. **Stability.** Our method maintains high predictive performance even for small sample data sets by utilizing the properties of recency and frequency of PVs. Computational results demonstrate that it was very effective even when the number of available training samples was small.
2. **Flexibility.** In sharp contrast to many parametric prediction models, e.g., logistic regression, our nonparametric approach enhances flexibility in modeling the relationship between customers' PVs and their product-choice probabilities. Indeed, in our experiments, our method provided higher predictive performance than the other methods did.
3. **Scalability.** Our method has remarkable scalability in the sense that the size of the probability table does not depend on the data size. Consequently, it dealt with massive amounts of data in the computational experiments. By contrast, the kernel SVM was only applicable to a small-scale data set, because its computation load depended heavily on the number of training samples.

The rest of the paper is organized as follows. The next section makes a brief review of related works. Section 3 develops the optimization models for estimating the product-choice probabilities. Section 4 assesses the effectiveness of our method through computational experiments. Section 5 concludes with a brief summary of our work and a discussion of future research directions.

2. Related works

One of the most active areas of clickstream research has been the analysis of online purchasing behavior of customers on e-commerce sites [6]. According to this purpose, Moe and Fader [24] propose a stochastic model for predicting each customer's purchase probability based on an observed history of visits and purchases. Many other studies use logit/probit modeling and various types of input variables to predict online purchasing behavior [27,29,37,38,41], whereas Boroujerdi et al. [2] apply different classification algorithms, such as decision tree, support vector machine and rule-based method, to predict customer's buying intention. These studies, however, focus on the prediction of customer visits that lead to purchases, and accordingly, they do not assign a purchase probability to each product.

There are a number of studies that analyze online product-choice behavior; however, most of them place emphasis on more detailed data rather than clickstream data. For instance, Chen and Fan [7] improve multiple kernel SVM to handle multiplex data, such as static, time series, symbolic sequential and textual data. Zhang and Pennacchiotti [43] build machine learning models to predict a customer's choice of product categories from their social media profiles. Qiu [31] takes advantage of product reviews and ratings in the support vector regression model. In contrast to these studies, the present paper has a different purpose of investigating the relationship between customers' PVs and their product-choice probabilities from clickstream data.

Table 1
Page view history of customers u_1, u_2 and u_3.

Product	#PVs of (u_1, u_2, u_3)			Recency	Frequency
	Mar. 1	Mar. 2	Mar. 3		
v_1	(1, 0, 2)	(0, 0, 1)	(0, 0, 1)	(1, 0, 3)	(1, 0, 4)
v_2	(0, 3, 0)	(2, 0, 2)	(2, 0, 0)	(3, 1, 2)	(4, 3, 2)
v_3	(0, 0, 1)	(0, 1, 0)	(0, 2, 0)	(0, 3, 1)	(0, 3, 1)
v_4	(3, 1, 0)	(0, 0, 1)	(0, 0, 0)	(1, 1, 2)	(3, 1, 1)
v_5	(0, 1, 1)	(0, 0, 1)	(0, 1, 0)	(0, 3, 2)	(0, 2, 2)

Recommender systems are software tools and techniques providing suggestions for products that may be of use to a customer [35]. A well-known class of recommender algorithms, i.e., collaborative filtering [11,34], recommends products on the basis of the preferences of other customers who have expressed opinions on those products. The primary objective of recommender systems is to help customers find unknown but worthwhile products. By contrast, the present paper focuses on products that individual customers have viewed in the past, and it estimates the probabilities that those products will be purchased. The estimated product-choice probabilities represent a customer's preferences for the products. Since such information is aggregated in collaborative filtering in order to recommend products to customers, our research will be valuable when applying collaborative filtering techniques to e-commerce. Additionally, it is noteworthy that in the context of recommender systems [8,21], the frequency of PVs is used in creating input data, and the recency of PVs is considered implicitly by limiting data acquisition period.

Our optimization models for estimating the product-choice probabilities are new applications of the isotonic regression to the analysis of clickstream data. The isotonic regression, which has its origin in the early works [4,15,16], fits a nonparametric function to given data points under mild shape constraints, such as monotonicity and convexity/concavity. Applications fields include statistics, operations research, and image processing (see Pardalos and Xue [30] and the references therein), and various algorithms have been developed so far [1,9,10,16,23,36,39]. Although Geyer [14] deals with the logistic regression under the monotonicity and convexity constraints, this model is only one-dimensional. To the best of our knowledge, none of the existing studies implement the maximum likelihood estimation method for multidimensional regression with monotonicity, convexity and concavity constraints. Moreover, our optimization models enhance the worth of traditional constrained regression on the forefront of data analysis technology.

3. Proposed method

This section presents our method for estimating the product-choice probabilities of individual customers from clickstream data. The method is based on two effective predictors, i.e., the recency and frequency of a customer's page views (PVs).

3.1. Problem description

As already mentioned, we estimate the product-choice probabilities from the recency and frequency of each customer's previous PVs. Roughly speaking, the *recency* of customer u with respect to product v represents the time of the last visit of customer u to a webpage of product v, while *frequency* represents the number of visits of customer u to a webpage of product v.

Table 1 shows an example of a page view history of the three customers. For instance on March 1st, the webpage of product v_1 was viewed once and twice, respectively by customers u_1 and u_3. Such a PV history is summarized in terms of recency and frequency in Table 1, where they are characterized by the day of the last visit and the number of PVs. Our aim is to analyze the customer's