European Journal of Operational Research 000 (2014) 1–13

Contents lists available at ScienceDirect

European Journal of Operational Research

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

Stochastics and Statistics

Behavior-aware user response modeling in social media: Learning from diverse heterogeneous data

Zhen-Yu Chen[a,*], Zhi-Ping Fan[a,b], Minghe Sun[c]

[a] Department of Information Management and Decision Sciences, School of Business Administration, Northeastern University, Shenyang 110819, China
[b] State Key Laboratory of Synthetical Automation for Process Industries, Northeastern University, Shenyang 110819, China
[c] Department of Management Science and Statistics, College of Business, The University of Texas at San Antonio, San Antonio, TX 78249-0632, USA

ARTICLE INFO

Article history:
Received 30 April 2013
Accepted 11 September 2014
Available online xxx

Keywords:
Data mining
Direct marketing
Response modeling
Social media
Engagement behavior

ABSTRACT

With the rapid development of Web 2.0 applications, social media have increasingly become a major factor influencing the purchase decisions of customers. Longitudinal individual and engagement behavioral data generated on social media sites post challenges to integrate diverse heterogeneous data to improve prediction performance in customer response modeling. In this study, a hierarchical ensemble learning framework is proposed for behavior-aware user response modeling using diverse heterogeneous data. In the framework, a general-purpose data transformation and feature extraction strategy is developed to transform the heterogeneous high-dimensional multi-relational datasets into customer-centered high-order tensors and to extract attributes. An improved hierarchical multiple kernel support vector machine (H-MK-SVM) is developed to integrate the external, tag and keyword, individual behavioral and engagement behavioral data for feature selection from multiple correlated attributes and for ensemble learning in user response modeling. The subagging strategy is adopted to deal with large-scale imbalanced datasets. Computational experiments using a real-world microblog database were conducted to investigate the benefits of integrating diverse heterogeneous data. Computational results show that the improved H-MK-SVM using longitudinal individual behavioral data exhibits superior performance over some commonly used methods using aggregated behavioral data and the improved H-MK-SVM using engagement behavioral data performs better than using only the external and individual behavioral data.

1. Introduction

Mass marketing and direct marketing are two commonly used approaches for product (service) advertising and promotional activities (Bose & Chen, 2009). For direct marketing, a marketing message is delivered to target customers without an intermediary person or indirect media involved (Bose & Chen, 2009). Customer response modeling aims at identifying the target customers who will respond to a specific marketing campaign from the existing customer base (Cui, Wong, & Zhang, 2010; Kang, Cho, & MacLachlan, 2012). With more and more companies adopting direct marketing, customer response modeling has become one of the most effective direct marketing strategies to increase total revenue and decrease marketing cost (Cui, Wong, & Lui, 2006; Kang et al., 2012; Lee, Shin, Hwang, Cho, & MacLachlan, 2010). Because the purpose is to identify customers as possible respondents and non-respondents to a specific marketing campaign (Bose & Chen, 2009; Lee et al., 2010), customer response modeling is a binary classification problem.

For customer response modeling, external and behavioral data are usually used (Bose & Chen, 2009). Customer demographic, geographic and lifestyle data are often obtained from external data vendors (Baecke & Van den Poel, 2011), and thus are called external data. Customer behavioral data including transaction records, feedbacks to marketers, customer reviews and Web browsing records are considered to be the most important data in customer response modeling (Bose & Chen, 2009). Many supervised and semi-supervised machine learning techniques have been proposed for the customer response modeling problem (Lessmann & Voß, 2008). These techniques include artificial neural networks (ANN) (Crone, Lessmann, & Stahlbock, 2006; Kim, Street, Russell, & Menczer, 2005), decision trees (Crone et al., 2006), Bayesian networks (Baesens, Viaene, Van den Poel, Vanthienen, & Dedene, 2002; Cui et al., 2006), logistic regression (Kang et al., 2012), bagging (Ha, Cho, & MacLachlan, 2005), support vector machines (SVM) (Crone et al., 2006; Kang et al., 2012; Lessmann & Voß, 2009) and transductive SVMs (Lee et al., 2010). Moreover, some other techniques including clustering (Kang et al., 2012), sampling (Crone et al., 2006; Kang et al., 2012), sequential

* Corresponding author. Tel.: +86 24 83671630.
E-mail address: zychen@mail.neu.edu.cn (Z.-Y. Chen)

http://dx.doi.org/10.1016/j.ejor.2014.09.008

Please cite this article as: Z.-Y. Chen et al., Behavior-aware user response modeling in social media: Learning from diverse heterogeneous data, European Journal of Operational Research (2014), http://dx.doi.org/10.1016/j.ejor.2014.09.008

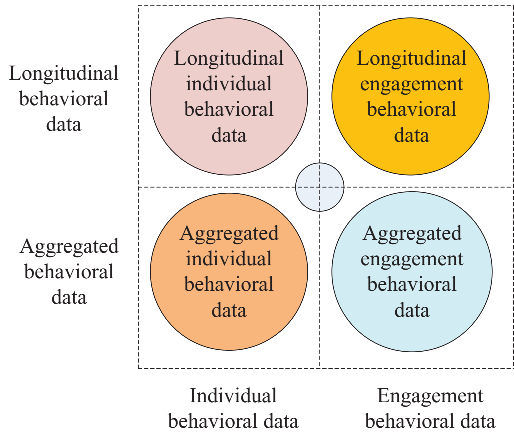

Fig. 1. Different types of behavioral data.

pattern discovery (Chen, Hsu, & Hsu, 2011), feature selection (Cui et al., 2010) and other preprocessing methods (Crone et al., 2006) have been combined with classification techniques to refine the customer base and improve prediction performance.

In the age of Web 2.0, social media sites develop rapidly. Social media refers to a group of online applications allowing the creation and exchange of user-generated contents (Kaplan & Haenlein, 2010). The most popular types of social media include wikis, blogs, microblogs, social networks, video and photo sharing and online communities. They become popular communication tools due in part to the open access of the Internet, the popularity of mobile devices, the availability of the tools and the fast social interactions among users. Social media have increasingly become a major factor influencing the opinions, attitudes and the purchase behavior of customers (Mangold & Faulds, 2009).

User behavioral data generated and collected on social media sites include two categories, *i.e.*, individual behavioral data and engagement behavioral data. Moreover, according to the ways of using the behavioral data in the customer response models, user behavioral data can be classified as longitudinal behavioral data and aggregated behavioral data. Fig. 1 illustrates the different types of behavioral data. For traditional customer response modeling, the longitudinal individual behavioral variables derived from the transactional databases are usually transformed into the aggregated variables such as recency, frequency and monetary (RFM) variables which have been included in most direct marketing datasets and adopted in most response models (Baesens et al., 2002; Crone et al., 2006; Cui et al., 2010).

In comparison with individual behavior, customer engagement behavior, as an emerging concept, focuses on the customers' behavioral manifestation beyond purchase such as electronic word-of-mouth, customer–customer interaction, recommendations, blogging and online reviews (van Doorn et al., 2010). In social media, customer engagement behavior has great effect on the individual purchase decisions (Cheung & Thadani, 2012; Dellarocas, 2003; van Doorn et al., 2010). For example, Dell gained high income by posting offers to its followers on Twitter (Li & Li, 2013). A survey showed that 91 percent of respondents said that they consulted online reviews before purchasing, and 46 percent of respondents believed that the online reviews influenced their purchase decisions (Cheung & Thadani, 2012). Therefore, incorporating engagement behavioral data into customer analytical models is increasingly recognized as a new direction of customer relationship management and direct marketing (Bijmolt et al., 2010).

The aggregated individual behavioral attributes are usually used as predictors in most customer response models. Few existing studies of customer response modeling pay attention to the longitudinal individual and engagement behavioral data which are widely available in the social media databases. In recent years, the analysis of engagement behavior has been used widely in the areas of recommendation and customer churn prediction. Some researchers combined the extended factorization model with other methods such as additive forest, logistic regression and scorecard model to predict the top-N items the customer was most likely to follow using the aggregated customer–customer interaction data (Chen, Liu, et al., 2012; Chen, Tang, et al., 2012; Zhao, 2012). The information of individual customers and a group of customers which have similar characteristics was used in a novel customer profile model for product recommendation (Park & Chang, 2009). For customer relationship management of the telecommunication industry, the customer–customer interaction data have been recognized as important complements to traditional behavioral data. The aggregated engagement behavioral attributes were combined with traditional attributes to predict customer churn (Zhang, Zhu, Xu, & Wan, 2012).

Some researchers recognized that customer purchase behavior varies over time and the use of the longitudinal individual behavioral data can improve prediction performance (Chen, Fan, & Sun, 2012; Liu, Lai, & Lee, 2009). Sequential pattern analysis was combined with collaborative filtering for temporal purchase behavioral data to improve recommendation performance (Cho, Cho, & Kim, 2005; Choi, Yoo, Kim, & Suh, 2012; Huang & Huang, 2009; Liu et al., 2009; Min & Han, 2005). Prinzie and Van den Poel (2006, 2007,2011) incorporated customer purchase sequence into dynamic Bayesian networks and Markov models to predict the next product for a customer to buy. Ballings and Van den Poel (2012) studied the problem of how long the customer historical data should be for customer churn prediction. They suggested that selecting a good length of historical data can decrease computational burden.

For social media, the term Item may represent a specific user, organization, product (service) or event. Examples of events include the appearance of a new term or keyword, the announcement of a new product (service) or activity, or a new price of an existing product (service). The rich behavioral data generated on social media sites can be used for managers to predict user responses to an Item, make marketing policies and allocate marketing resources to influence customer behavior (Power & Phillips-Wren, 2011). For social media, customer response modeling is also called user response modeling, and the two terms are used interchangeably. In this study, customer response modeling taking into consideration of user behavioral, *e.g.*, longitudinal individual and engagement behavioral, data is called behavior-aware user response modeling. However, the large, diverse and heterogeneous data generated on social media sites bring great challenges on behavior-aware user response modeling (Bijmolt et al., 2010; Cao, Ou, & Yu, 2012; Chau & Xu, 2012).

How to deal with diverse heterogeneous data is a challenge. A variety of methods can be used for customer response modeling using external and aggregated individual behavioral data. However, to the best of our knowledge, this study is the first attempt of combining the individual behavioral and the engagement behavioral data, as well as the longitudinal and the external data for user response modeling in social media.

How to deal with large amount of data is another challenge. Social media sites produce large amount of user data. For example, the daily volume of posts mentioning some well-known brands or products such as Google, Microsoft, Sony, iPhone and iPad in Twitter is in the millions (Li & Li, 2013). It is necessary to use marketing intelligence methods to automatically analyze the massive amount of data. The analysis of the massive amount of data requires efficient preprocessing of the data and excellent scalability of the customer response models.

In this study, a hierarchical ensemble learning framework is developed for behavior-aware user response modeling in social media. In the framework, a general-purpose data preprocessing strategy is proposed to transform the large-scale and multi-relational user datasets