Computer Science and Application 计算机科学与应用, 2022, 12(11), 2526-2532
Published Online November 2022 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2022.1211258

Hans汉斯

基于店铺忠诚度和店铺关联性的朴素贝叶斯推荐算法

刘兴林，黄　荣

五邑大学智能制造学部，广东 江门

收稿日期：2022年10月17日；录用日期：2022年11月15日；发布日期：2022年11月22日

摘　要

本文将k-means算法、关联规则和朴素贝叶斯算法结合，提出一种基于店铺忠诚度和店铺关联性的朴素贝叶斯推荐算法。该算法首先使用k-means算法对用户进行店铺忠诚度聚类，再使用关联规则算法和用户的历史支付信息推测出店铺关联性，最后，使用朴素贝叶斯算法对用户进行训练，并利用训练结果对用户进行店铺预测。本文采用天池大数据提供的十五个月支付宝支付日志和浏览日志对该算法进行测试，并验证其可行性。

关键词

店铺忠诚度，店铺关联性，K-Means算法，朴素贝叶斯算法

Naive Bayesian Recommendation Algorithm Based on Store Loyalty and Store Relevance

Xinglin Liu, Rong Huang

Faculty of Intelligent Manufacturing, Wuyi University, Jiangmen Guangdong

Received: Oct. 17th, 2022; accepted: Nov. 15th, 2022; published: Nov. 22nd, 2022

Abstract

Combining k-means algorithm, association rule and naive Bayesian algorithm, this paper proposes a naive Bayesian recommendation algorithm based on store loyalty and store relevance. First, the k-means algorithm is used to cluster the store loyalty. Second, the association rule algorithm and the user's historical payment information are used to infer the store relevance. Finally, the naive Baye-

文章引用: 刘兴林, 黄荣. 基于店铺忠诚度和店铺关联性的朴素贝叶斯推荐算法[J]. 计算机科学与应用, 2022, 12(11): 2526-2532. DOI: 10.12677/csa.2022.1211258

sian algorithm is used to train the user, and the training result is used to predict the store. In this paper, the fifteen-month Alipay payment logs and browsing logs provided by Tianchi Data were used to test the algorithm and verify its feasibility.

Keywords

Store Loyalty, Store Relevance, K-Means Algorithm, Naive Bayesian Algorithm

1. 引言

随着商业发展和网络普及，店铺的支付方式不仅限于现金、银行卡的方式，还有类似支付宝、微信等电子现金转账方式。本文主要针对使用支付宝支付的店铺进行用户忠诚度分析，并根据用户的店铺忠诚度进行商品推荐。黄益国[1]使用 k-means 算法对淘宝用户进行店铺忠诚度分类，并根据该分类和朴素贝叶斯分类对用户进行商品预测，后文对该算法简称为 KBS。但文献[1]中存在以下问题：一是 k-means 算法的用户分类数 k 值，作者暂定为 3，且没有给出选取 k 为 3 的具体理由；二是利用贝叶斯分类判断用户是否购买某件商品时采用的特征，仅考虑了用户店铺忠诚度、购买商品 A 的系列商品的数量以及店铺是否有折扣，却没有考虑到商品之间的关联性；三是数据较为单一化，即只针对一间店铺进行用户忠诚度分析；四是缺少聚类对比实验，无法说明 k-means 聚类结果为最优。因此，本文提出一种基于店铺忠诚度分类和店铺关联性的朴素贝叶斯推荐算法，其主要研究内容如下：一是采用天池大数据提供的 2000 个店铺支付宝支付记录进行用户忠诚度分类；二是使用 k-means 算法对用户店铺忠诚度进行多个 k 值聚类；三是在使用贝叶斯分类判断用户是否购买某件商品时的特征中，增添购买商品 A 前所有可能购买商品的数量，即增添与商品 A 具有关联性的商品数量。

2. 用户店铺忠诚度聚类

黄益国采用的数据是 NALA 专业彩妆护肤名店(后面简称 NL) 2009 年到 2016 年 6 月 30 日有过网购经历的 958 名用户数据。其对用户店铺忠诚度进行 k-means 分类时采用的分类特征有网购历史、是否收藏本店、最近一次消费不超过三年、一年内网购次数、每次平均消费金额和最近一次商品评价。本文的实验数据采用天池大数据提供的 IJCA117_data 数据集，该数据包含支付宝用户从 2015 年 7 月 7 日到 2016 年 10 月 31 日的支付日志和用户浏览日志以及店铺信息，数据具体信息见章节 3.1。根据文献[2] [3] [4] [5]和实验数据信息可将用户 U_i 对店铺 S_j 的忠诚度分类特征设置如表 1 所示。

Table 1. The classification characteristics of the loyalty of customer U_i in store S_j
表 1. 用户 U_i 对店铺 S_j 的忠诚度分类特征

k-means 聚类特征	离散值	说明
用户 ID	U_i	用户 ID
店铺 ID	S_j	店铺 ID
支付宝使用次数(C_1)	$C_1 = c(c \in N^*)$.	用户 U_i 在店铺 S_j 支付及浏览总次数。

Continued

用户平均每次使用金额(C_{21})与店铺平均使用金额(C_{22})的差值(C_2)	$C_2 = C_{21} - C_{22}$ 其中 $C_{21} = \frac{\sum_{k=1}^{N} \text{per\_pay} * \text{pay\_time}_k}{N}$， $C_{22} = \text{per\_pay}$.	pay_time_k 为用户 U_i 在第 k 时间点下在店铺 S_j 的支付次数，N 为用户 U_i 在不同时间点下在店铺 S_j 的支付次数。
平均多长时间购买该店铺商品(C_3)	$C_3 = \frac{\sum_{k=2}^{n} \text{pay\_time}_k - \text{pay\_time}_{k-1}}{n}$.	n 为用户 U_i 在不同时间点下在店铺 S_j 的支付次数， pay_time_k 表示用户 U_i 在第 k 个在店铺 S_j 支付的时间点。
店铺评论数量(C_4)	$C_4 = \text{comment\_cnt}$.	comment_cnt 的数值越大代表评论数量越多
店铺评分(C_5)	$C_5 = \text{score}\left(\text{score} \in \{0,1,2,3,4,5\}\right)$.	score 的数值越大代表评分越高
店铺类型(C_6)	$C_6 = \text{cate\_id}$.	cate_id 的具体表示方法见章节 3.1 数据预处理。
店铺档次(C_7)	$C_7 = \text{shop\_level}\left(\text{shop\_level} \in \{1,2,3\}\right)$.	1 表示低档，2 表示中档，3 表示高档。

其中，C_k 表示聚类的第 k 个特征，pay_time 表示支付时间，per_pay 表示店铺平均支付金额，score 表示店铺评分，comment_cnt 表示店铺被评论的次数，cate_id 表示类别 ID，shop_level 表示店铺档次。

本文算法使用 MATLAB 中自带的 k-means 来实现用户店铺忠诚度聚类，具体步骤如下[6]：

1) 根据表 1 进行数据处理，并存入变量 **user_data** 中，其中，**user_data** 的第 j 行数据表示为：$\text{user\_data}(j,:) = \{\text{user\_id}, \text{shop\_id}, C_1, C_2, C_3, C_4, C_5, C_6, C_7\}$；

2) 开启并行运算 parpool；

3) for i = 3:6；

a) 调用 k-means 算法，[**idx**,**C**,**sumd**,**D**] = kmeans (**user_data**(:,3:end), i,'MaxIter', 10000, 'Display', 'final', 'Replicates', 100)，其中，**idx** 表示用户店铺忠诚度类别，**C** 表示 i 类的各类质心，**sumd** 表示类内所有点与该类质心点距离之和，**D** 表示每个点到所有质心的距离，**user_data**(:,3:end)表示截取原始数据中除用户 ID 和店铺 ID 之外的所有用户店铺忠诚度分类特征，i 表示类别数，('MaxIter',10000)表示最大迭代次数为 10000，(' Display','final')表示显示最终迭代结果，('Replicates',100)表示运算重复次数为 100。

b) 按照类别数值越大忠诚度越高整理计算结果，并输出到 **user_shop_class**，其中，**user_shop_class** 的第 j 行数据格式为：$\textbf{user\_shop\_class} = \{\text{user\_id}, \text{shop\_id}, \text{user\_class}\} = \{\textbf{user\_data}(:,1{:}2), \textbf{idx}\}$。

4) end；

5) 关闭并行运算。

3. 基于店铺忠诚度的朴素贝叶斯推荐算法

文献[1]中判断消费者是否购买产品 A 的方法是利用贝叶斯分类算法计算在 a1、a2 和 a3 这三个特征下购买产品 A 和不购买产品 A 的概率，若购买产品 A 的概率大于不购买的，则预测该消费者会购买产品 A，反之则不购买产品 A，其中，a1 表示消费者在 NL 店忠诚度，a2 表示购买过 NL 店系列产品 B 的数量，a3 表示是否有 NL 店铺折扣优惠。在章节 1 中提到的数据中，不包含店铺是否有折扣，但比文献[1]中的数据多了店铺类别，因此，用户购买店铺的贝叶斯分类特征如表 2 所示。

Table 2. Bayesian classification characteristics of consumer U_i in store S_j

表 2. 用户 U_i 购买店铺 S_j 的贝叶斯分类特征

贝叶斯分类特征	离散值	说明
用户 ID	U_i	用户 ID
店铺 ID	S_j	店铺 ID
用户 U_i 对店铺 S_j 的忠诚度(C_1)	$C_1 = c(c \in \{0,1,2,3,4,5,6\})$	c 的具体值是 k-means 算法聚类后的结果，其中 0 表示用户 U_i 在店铺 S_j 未使用过支付宝。
购买店铺 S_j 之前用户购买店铺 S 的数量($\mathbf{C}_2$)	$\mathbf{S} = \{S_1, S_2, \cdots, S_k, \cdots, S_n\}$ $C_2 = \{c_1, c_2, \cdots, c_k, \cdots, c_n\}$	S 是利用关联规则的计算结果获取购买店铺 S_j 之前可能购买的 n 个店铺，c_k 对应第 k 个店铺的数量。
购买同类店铺的数量(C_3)	$C_3 = c(c \in N^*)$	在与店铺 S_j 相同类型的店铺支付次数，c 仅表示一个数值。

贝叶斯分类特征 $\mathbf{C}_2$ 的店铺关联性计算步骤如下[7] [8]：

1) 数据处理：考虑到店铺存在重复购买的情况，在获取用户同一天店铺支付信息时在购买的多个店铺 ID 后面增加小数进行区分，如某用户在 2016 年 5 月 10 日不同时间点在店铺 156 支付三次，那么该用户 2016 年 5 月 10 日的支付信息为(156, 156.1, 156.2)；

2) 初始化最小支持度 min_Sup 和最小置信度 min_Conf：考虑到用户多次一天仅使用一次支付宝导致推出的规则支持度过小的情况，min_Sup 的值设置为 0.001，min_Conf 设置为 0.5 [9] [10]；

3) 找出支持度大于 min_Sup 的 1 项集 L_1；

4) for $i = 2 : L_i \neq \text{null}$；

a) 根据 L_{i-1} 项集找支持度大于 min_Sup 的频繁 L_i 项集。

5) end；

6) 计算最大频繁项集 L 的所有非空子集的置信度，并找出置信度大于 min_Conf 的强关联规则；

7) 将店铺 ID 取整还原为原始店铺 ID，并去掉重复的规则，如关联规则 156 => (156.1, 156.2)和关联规则 156.1 => (156.2, 156)，对数据进行取整为 156 => (156, 156)和 156 => (156, 156)，两个规则表示的均为店铺 156 可能被购买第二次甚至第三次，去重后可表示为 156 => 156，表示店铺 156 有被重复购买的可能性。

其中，支持度(Support)计算公式如式(1)所示，置信度(Confidence)计算公式如式(2)所示：

$$\text{Support}_{S_1 => S_2} = P(S_1 \cup S_2) \tag{1}$$

$$\text{Confidence}_{S_1 => S_2} = P(S_2 \mid S_1) \tag{2}$$

根据以上信息对用户进行贝叶斯分类特征计算后，再进行用户贝叶斯店铺预测，用户贝叶斯店铺预测的步骤如下[11] [12] [13]：

1) 数据处理：根据表 2 进行贝叶斯分类特征提取，并将训练数据和测试数据分别存入到 **Bayes_train_data** 和 **Bayes_test_data** 中，将用户对店铺的行为存入到 **Bayes_train_class** 和 **Bayes_test_class** 中，其中，**Bayes_train_data** 和 **Bayes_test_data** 的第 j 行数据可表示为：$\{\text{user\_id}, \text{shop\_id}, C_1, \mathbf{C}_2, C_3\}$，**Bayes_train_class** 和 **Bayes_test_class** 的第 j 行数据表示为：$\{\text{user\_id}, \text{shop\_id}, \text{action\_type}\}$，action_type 表示用户在店铺的行为，行为包含两种购买和未购买分别用 1 和 0 表示；

2) 计算 **Bayes_train_data** 数据中的购买概率 $P(V_0)$ 和未购买概率 $P(V_1)$，其中，$P(V_0) = V_0$ 的个数/数据总据，$P(V_1) = V_1$ 的个数/数据总据；

3) 根据贝叶斯定理计算 **Bayes_train_data** 数据中每个类别条件下特征属性的划分概率 $P(C_m \mid V_n)$，其中，$m=1,2,3; n=0,1$；

4) 根据步骤 3)及步骤 4)的训练结果和公式(3)分别计算 Bayes_test_data 中用户 U_i 的店铺 S_j 属于类别 V_0 的概率 $P(V_0 \mid S_j)$ 和属于类别 V_1 的概率 $P(V_1 \mid S_j)$，公式(3)如下所示：

$$P(V_n \mid S_j) = \sum_{i=1}^{3} P(C_m \mid V_n) * P(V_n) \tag{3}$$

5) 比较 $P(V_0 \mid S_j)$ 和 $P(V_1 \mid S_j)$ 大小，若 $P(V_0 \mid S_j) > P(V_1 \mid S_j)$，则说明用户 U_i 不在店铺 S_j 购买商品，反之，则说明用户 U_i 在店铺 S_j 购买商品，并将计算结果保存到变量 prediction_results 中，其中，prediction_results 的 j 行数据可表示为：$\{user\_id, shop\_id, prediction\_class\}$。

4. 实验方案及结果分析

4.1. 实验数据预处理

本文的实验数据采用天池大数据提供的 IJCA117_data 数据集，该数据包含支付宝用户从 2015 年 7 月 1 日到 2016 年 10 月 31 日的支付日志和浏览日志以及 2000 个店铺信息，其数据格式如表 3、表 4 和表 5 所示。

Table 3. Customer payment log
表 3. 用户支付日志

user_id	shop_id	pay_time
12152353	1894	2016-09-18 09:00:00

Table 4. Customer browsing log
表 4. 用户浏览日志

user_id	shop_id	view_time
171893	1894	2016-08-26 19:00:00

Table 5. Store information
表 5. 店铺信息

shop_id	per_pay	score	comment_cnt	shop_level	cate_1_name	cate_2_name	cate_3_name
846	17	4	0	1	超市便利店	超市	
847	11	3	3	0	美食	休闲茶饮	奶茶

其中，user_id 表示用户 ID，shop_id 表示店铺 ID，pay_time 表示支付时间，view_time 表示浏览时间，per_pay 表示店铺平均支付金额，score 表示店铺评分，comment_cnt 表示店铺被评论的次数，shop_level 表示店铺档次，cate_1_name 一级类别名称、cate_2_name 二级类别名称和 cate_3_name 三级类别名称，其中，score、comment_cnt 和 shop_level 的数值越大代表评分或店铺等级越高、评论次数越多。

选取 2016 年 1 月 1 日到 2016 年 9 月 30 日的支付日志和浏览日志作为训练数据，2016 年 10 月 1 日到 2016 年 12 月 31 日的支付日志和浏览日志作为测试数据，其中包含 453,412 名用户的支付和浏览信息。店铺信息中不包含店铺类型 ID，仅包含三个级别的类别名称，因此需要对 cate_1_name、cate_2_name 和 cate_3_name 进行数字化处理，处理方法为：以 cate_1_name 和 cate_2_name 为主分类 cate1，cat_3_name 为次分类 cate2，则类别 IDcate_id=cate1*100+cate2，如店铺 847 的主分类为美食－休闲茶饮、次分类为

奶茶，设 cate1=10，cate2=4，则分类为美食－休闲茶饮－奶茶的类别 ID cate_id=1004。

4.2. 评测方法

实验选择贝叶斯预测结果的准确率作为实验的评价标准，计算公式如(4)所示：

$$\text{accuracy\_rate} = \frac{\textbf{prediction\_results} \odot \textbf{Bayes\_test\_class}}{\text{N}} \tag{4}$$

其中，N 表示 Bayes_test_class 的总行数，即测试数据总数量，prediction_results⊙Bayes_test_class 表示计算预测结果和实际结果相同的数量。

4.3. 推荐算法实验及结果分析

本文将实验结果与文献[1]的算法进行对比，并将本文提出的算法用 KBR 表示。本文主要考虑的参数有 k-means 算法中的用户忠诚度类别数量 k 和购买商品 A 前所有可能购买商品的数量 C2。计算结果如表 6 所示。

Table 6. Comparison of the results of the KBR and KBS algorithms
表 6. KBR 算法和 KBS 算法的计算结果对比

k-means 聚类数目	KBS 算法的准确率	KBR 算法的准确率
3	0.723058824378503	0.921133256108519
4	0.700246830715268	0.921129647472331
5	0.690220235066561	0.921138669062801
6	0.709820542522364	0.921140473380895

根据表 6 的计算结果可知，KBS 算法的准确率随着 k-means 聚类数目的增加而减少，而随着 k-means 聚类数目的增加 KBR 算法的准确率并无较大变化。当 k-means 聚类数目一致时，KBR 算法的准确率明显比 KBS 的准确率高，因此，KBR 算法是有效的和可行的。

5. 结论

本文提出了一种基于店铺忠诚度分类和店铺关联性的朴素贝叶斯推荐算法。该算法以天池大数据提供的 IJCA117_data 数据集为测试数据，通过 SQLServer 2008 搭建数据库，通过 MATLAB 进行数据计算及对比。该算法首先对用户店铺忠诚度进行聚类，再根据用户的历史支付信息进行店铺关联性计算，最后，根据用户店铺忠诚度聚类结果以及店铺关联性进行朴素贝叶斯训练和测试。实验结果表明，本文提出的算法对 KBS 算法的改进是有效的。在本文中，k-means 的店铺忠诚度聚类特征和贝叶斯分类特征仅选取了与店铺相关的信息，因此，在特征的选取上将是未来的研究重点。

基金项目

广东省科技厅项目(2016A070708002, 2015A070706001, 2014A07070 8005)；研究生教育创新计划项(2016SFKC_42, YJS-SFKC-14-05, YJS-PYJD-17-03)资助；教育部“云数融合、科教创新”基金项目(2017B02101)资助。

参考文献

[1] 黄益国. 基于数据挖掘技术的淘宝店铺客户及商品销售分析[D]: [硕士学位论文]. 泉州: 华侨大学, 2016.

[2] Titah, V., Lapian, S.L.H.V.J., Rumokoy, F.S. (2018) Analysing Factors That Drive Customers Purchase Intention of Licensed Team Merchandise Sports Station Manado. *Jurnal EMBA*: *Jurnal Riset Ekonomi*, *Manajemen*, *Bisnis dan Akuntansi*, **6**, 1418-1427.

[3] Ratanasawadwat, N. (2015) Loyalty in Department Store Online Shopping. *XIV International Business and Economy Conference* (*IBEC*), Bangkok, 5-8 January 2015, 1-7. https://doi.org/10.2139/ssrn.2593909

[4] Hwang, Y.J. (2014) Understanding the Different Influences of Online Trust on Loyalty by Risk Takers and Avoiders. *International Journal of Human-Computer Interaction*, **30**, 977-984. https://doi.org/10.1080/10447318.2014.925384

[5] Zhang, M. and Peng, H. (2010) An Empirical Study of Influence Factors in Customer Loyalty for Department Store Service. 2010 *International Conference on Management and Service Science*, Wuhan, 24-26 August 2010, 1-4. https://doi.org/10.1109/ICMSS.2010.5577406

[6] Bora, D.J. and Gupta, A.K. (2014) Effect of Different Distance Measures on the Performance of K-Means Algorithm: An Experimental Study in Matlab. *International Journal of Computer Science and Information Technologies* (*IJCSIT*), **5**, 2501-2506.

[7] 陈君山, 刘京菊, 李振汉. 基于关联规则挖掘的口令字典生成技术[J]. 微电子学与计算机, 2018, 35(12): 110-114.

[8] Brijs, T., Swinnen, G., Vanhoof, K. and Wets, G. (2004) Building an Association Rules Framework to Improve Product Assortment Decisions. *Data Mining and Knowledge Discovery*, **8**, 7-23. https://doi.org/10.1023/B:DAMI.0000005256.79013.69

[9] Suhyun, H.A., Noh, K., Shin, M., *et al*. (2015) Identifying Multi-Component Drug Candidates in Natural Products via Association Rule Mining. *Chinese Journal of Pharmacology and Toxicology*, **29**, 99-100.

[10] Kamsu-Foguem, B., Rigal, F. and Mauget, F. (2013) Mining Association Rules for the Quality Improvement of the Production Process. *Expert Systems with Applications*, **40**, 1034-1045. https://doi.org/10.1016/j.eswa.2012.08.039

[11] 许召召, 李京华, 陈同林, 等. 融合 SMOTE 与 Filter-Wrapper 的朴素贝叶斯决策树算法及其应用[J]. 计算机科学, 2018, 45(9): 65-69+74.

[12] Jafar, M.J. (2013) A Tools-Based Approach to Teaching Data Mining Methods. *Journal of Information Technology Education*: *Innovations in Practice*, **9**, 24 p.

[13] Muda, Z., Yassin, W., Sulaiman, M.N. and Udzir, N.I. (2011) Intrusion Detection Based on K-Means Clustering and OneR Classification. 2011 7*th International Conference on Information Assurance and Security* (*IAS*), Melacca, 5-8 December 2011, 192-197. https://doi.org/10.1109/ISIAS.2011.6122818