计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2020,29(5):136−143 [doi: 10.15888/j.cnki.csa.007382]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于全域市场数据感知的终端客户推荐①

何利力, 张　星

(浙江理工大学 信息学院, 杭州 310018)
通讯作者: 张　星, E-mail: starsvictor@163.com

摘　要: 终端客户推荐系统是大型制造商终端营销的一种有效工具. 如何在互联网+环境下通过采集全域市场数据, 设计一个寻找最佳目标客户的推荐方法成为了一项挑战. 为解决这一问题, 本文提出一种基于全域市场数据感知的终端客户推荐方法 (GMF). 即采用全域分析的思想对全国范围内的客户数据进行预处理, 建立全方位, 多角度的评估指标, 得到目标客户价值. 然后通过域子空间分解的方法, 在域子空间中对数据进行分解分析, 得到某一区域内的客户评价标准, 将二者分析结果进行有效融合, 通过计算耦合对象相似度, 并筛选出最相似的 TopN 个数据作为最佳目标客户结果集. 在大型制造商营销活动所生成的数据集上的实验结果表明: 本文提出的推荐算法其性能明显优于当前主流的协同过滤算法.

关键词: 全域市场; 客户价值; 矩阵分解; 耦合对象相似度; 推荐算法

引用格式: 何利力,张星.基于全域市场数据感知的终端客户推荐.计算机系统应用,2020,29(5):136–143. http://www.c-s-a.org.cn/1003-3254/7382.html

Terminal Customer Recommendation Based on Global Market Data Perception

HE Li-Li, ZHANG Xing

(School of Information Science and Technology, Zhejiang Sci-Tech University, Hangzhou 310018, China)

Abstract: The end-customer recommendation system is an effective tool for large-scale manufacturer terminal marketing. How to design a search method for finding the best target customer by collecting global market data in the Internet+ environment has become a challenge. To solve this problem, This study proposes a terminal customer recommendation method based on global market data perception (GMF). That is to use the idea of global analysis to preprocess the customer data nationwide, establish a comprehensive, multi-angle evaluation index, and obtain the target customer value. Then, through the method of domain subspace decomposition, the data is decomposed and analyzed in the domain subspace, and the customer evaluation criteria in a certain region are obtained. The analysis results of the two are effectively merged, and the similarity of the coupled objects is calculated, and the most similar TopN data is used as the best target customer result set. The experimental results on the data set generated by the large-scale manufacturer marketing activities show that the proposed algorithm is significantly better than the current mainstream collaborative filtering algorithm.

Key words: global market; value of customer; matrix decomposition; coupled object similarity; recommendation algorithm

随着“互联网+”的快速推进, 大型制造商在投入巨大资金建设大数据的同时也面临着数据因多源化而变得分散难以管理的难题. 如何有效融合 PC 端, 移动端及线下数据等多渠道采集的客户信息, 将大量的不统一的数据碎片通过数据挖掘汇聚成可视化的整体并从中发现其个性化需求变得更加困难. 为了走出因多源

① 基金项目: 浙江省科技厅 (重大) 项目 (2015C03001)
Foundation item: Major Program of Science and Technology Bureau, Zhejiang Province (2015C03001)
收稿时间: 2019-10-07; 修改时间: 2019-10-29; 采用时间: 2019-11-05; csa 在线出版时间: 2020-05-07

数据而产生的数据孤岛的困局, 大型制造商力图通过建设全域市场数据体系打通跨屏, 多源数据间的障碍, 实现以全域市场数据来驱动业务, 让数据发挥更大价值. 为了寻找最佳目标客户实现产品精准投放提高营销利润, 开发智能化的终端客户推荐系统成为解决这一难题的有效手段, 近年来受到学术界和工业界的广泛关注.

目前使用的推荐系统绝大部分是基于协同过滤技术的推荐. 它是一种基于用户偏好且对所有用户无差别的推荐, 这可能导致大型制造商在获利甚微的客户身上投入不适当的成本. 营销活动的持续开展使得新生产品日益丰富, 每个产品在进行选户投放时可能存在众多销量一样而客户属性相差巨大的不同客户候选集, 制造商不可能通过将产品均衡投放到所有客户的方式开展系列营销活动, 这就导致了用户-产品矩阵非常稀疏. 因此, 利用推荐系统预测评分矩阵中的缺失项为目标产品寻找机会点, 并将结果以个性化列表的形式推荐给客户经理成为了一项挑战.

针对目前客户推荐方法所造成的推荐效果差和营销成本大的问题, 本文提出了一种基于全域市场数据感知的终端客户推荐方法. 首先通过采集全国范围内的客户订单交易数据对客户进行全方位价值评估, 然后利用子空间分解的方法对各个区域内产品的购买情况进行分析, 通过构建区域特色系数对客户购买产品情况进行衡量, 最后结合客户自身价值与区域特色系数构建全域用户项目评分矩阵. 在此基础上, 利用基于耦合对象相似度的推荐算法计算各个客户之间的相似度, 深度挖掘全域市场下不同客户之间的隐式关联, 为大型制造商推荐最佳目标客户.

1 协同过滤推荐算法

在个性化推荐领域, 协同过滤 (Collaborative Filtering, CF)[1,2]虽然是使用最广泛的技术之一, 但由于其存在严重的项目稀疏性和冷启动问题, 使得大量的学者针对此进行不断研究. Lika 等[3]提出了一种利用已知分类算法创建用户组, 并结合语义相似性技术识别相似行为用户的方法缓解冷启动问题. Ji 等[4]提出了一种利用因子矩阵分解模型, 结合用户和项目的内容信息来缓解冷启动问题的推荐算法. Forsati 等[5]提出了一种动态微调正则化参数的矩阵分解算法被广泛关注并证明了矩阵分解推荐算法的有效性.

矩阵分解算法 (Matrix Factorization, MF)[6,7]是利用降维的思想将用户项目矩阵分解为用户隐特征向量矩阵和项目隐特征向量矩阵, 然后通过两个隐特征向量矩阵的点积计算预测评分矩阵的缺失项. 其中典型的矩阵分解算法包括: 概率矩阵分解 (Probabilistic Matrix Factorization, PMF)[8], 最大间隔矩阵分解 (Maximum Margin Matrix Factorization, MMMF)[9], 非负矩阵分解 (Nonnegative Matrix Factorization, NMF)[10], 正则化奇异值分解 (Regularization Singular Value Decomposition, RSVD)[11], 贝叶斯概率矩阵分解 (Bayesian Probabilistic Matrix Factorization, BPMF)[12], SVD++[13]等. 但用户项目评分矩阵的稀疏性造成的项目冷启动问题仍然是目前亟待解决的问题, 为了进一步提高推荐算法的有效性, 近年来已有不少学者提出利用不同类型的信息源来解决项目冷启动问题. Yang 等[14]提出了一种社交信任网络的矩阵分解模型, 利用额外的信任数据来解决这一问题. Gurini 等[15]提出了一种融合情感分析的在社交网络推荐算法, 在矩阵因子分解过程中利用在社交平台所生成的内容中提取到的用户情感信息, 为目标用户推荐要关注的偏好用户.

区别于基于简单相似匹配 (SMS)[16]衡量客户相似性的推荐算法, 本文在矩阵分解时利用耦合对象相似度 (Coupled Object Similarity, COS)[17,18]捕获客户属性信息来改善推荐效果. 通过目标约束条件, 利用客户属性信息约束矩阵分解的过程来学习客户间的隐特征关系, 使推荐结果更具可解释性.

1.1 传统的矩阵分解方法

矩阵分解模型将用户和项目映射到维度为 K 的低维联合隐特征空间, 而用户-项目交互信息可以被建模为该空间中的内积. 每个用户 u 对应于一个列向量 $P_u \in R^k$, 每个项目 i 对应于一个行向量 $Q_i \in R^k$. 对于一个给定的用户 u, P_u的元素度量了这个用户对相应的项目特征的偏好程度. 对于给定的项目 i, Q_i的元素度量了该项目拥有这些特征的程度. m 个用户和 n 个项目分别形成用户隐特征矩阵 $P \in R^k$ 和项目隐特征矩阵 $Q \in R^k$. 其内积 $P_u^{\mathrm{T}} Q_i$就是用户 u 对项目 i 交互的建模. 因此, 若给定特征向量维数 K, 用户-项目评分矩阵可分解为 P 和 Q 两部分:

$$R_{u,i} \approx \hat{R} = P^{\mathrm{T}} Q \tag{1}$$

通过用户对项目的评分数据, 最小化目标函数学习隐特征矩阵 P 和 Q, 目标函数定义如下:

$$L=\frac{1}{2}\sum_{u=1}^{m}\sum_{i=1}^{n}(R_{u,i}-\hat{R}_{u,i})^2+\frac{\lambda_1}{2}\|P_u\|_F^2+\frac{\lambda_2}{2}\|Q_i\|_F^2 \quad (2)$$

其中, $\frac{\lambda_1}{2}\|P_u\|_F^2+\frac{\lambda_2}{2}\|Q_i\|_F^2$用来避免过拟合. λ_1和λ_2为正则化超参数, 用来控制正则化项对隐特征向量的影响. 通常使用随机梯度下降法 (Stochastic Gradient Descent, SGD) 求解此目标函数的局部最优解.

1.2 耦合对象相似度

借鉴文献[17], 客户u和客户u'之间的相似度定义如下:

$$\cos(u,u')=\sum_{i=1}^{n}\delta_i^A(u_i,u_i') \quad (3)$$

其中, u_i, u_i'是客户u和u'在特征i上的属性值, δ_i^A为耦合属性值相似度 (Coupled Attribute Value Similarity, CAVS).

耦合属性值相似度 (CAVS) 形式化的由 IaAVS 和 IeAVS 两部分组成. 其中 IaAVS 表示特征内耦合属性值相似度, IeAVS 表示特征间耦合属性值相似度.

特征i上属性值u_i和u_i'之间的耦合属性值相似度定义如下:

$$\delta_i^A(u,u')=\delta_i^{\mathrm{Ia}}(u_i,u_i')*\delta_i^{\mathrm{Ie}}(u_i,u_i') \quad (4)$$

其中, δ_i^{Ia}表示特征内耦合属性值相似度 (IaAVS), δ_i^{Ie}表示特征间耦合属性值相似度 (IeAVS).

特征内耦合属性值相似度 (IaAVS) 度量属性值相似度从频率分布角度刻画属性值间的相似度, 其在计算相似度时仅考虑了同一特征a_i内的属性值u_i和u_i'之间的相互关系, 定义如下:

$$\delta_i^{\mathrm{Ie}}(u_i,u_i')=\sum_{k=1,k\neq i}^{n}a_k\delta_{i|k}(u_i,u_i') \quad (5)$$

其中, a_k是特征$a_k(k\neq i)$下的第k个属性的权重参数. $\sum_{k=1}^{n}a_k=1$, $a_k\in[0,1]$. $\delta_{i|k}(u_i,u_i')$是属性值u_i和u_i'在特征$a_k(k\neq i)$下的特征间耦合属性值相似度, 定义如下:

$$\delta_{i|k}(u_i,u_i')=\sum_{w\in\cap}min\left\{P_{k|i}(\{w\}|u_i),P_{k|i}(\{w\}|u_i')\right\} \quad (6)$$

其中, $\cap$表示特征a_i取属性值u_i条件下特征a_k的属性值的所有取值集合与特征a_i取属性值u_i'条件下特征a_k的属性值得所有取值集合的交集. $P_{k|i}(\{w\}|u_i)$和$P_{k|i}(\{w\}|u_i')$是信息条件概率, 其定义如下:

$$P_{k|i}(\{w\}|u_i)=\frac{|g_k(w)\cap g_j(u_i)|}{g_j(u_i)} \quad (7)$$

其中, $P_{k|i}(\{w\}|u_i)$描述了特征a_i取属性值u_i条件下, 特征a_k取值w的属性值分布特征.

2 基于全域市场数据感知的推荐算法

全域市场数据感知推荐方法是一种从数据源头进行全面的整体趋势和性能分析, 对全国范围内的客户评估其总体价值, 然后通过域子空间分解的方法, 对各个区域客户购买产品情况进行分析构建初始的用户-项目评分矩阵, 最后融合全域客户价值和各区域初始评分得到最终的全域用户项目评分矩阵.

2.1 全域用户项目评分矩阵

区别于产品推荐在协同过滤算法中的输入, 本文提出的全域用户项目评分矩阵由两部分构成: 1) 利用全国范围内的客户订单交易数据评估客户的总体价值; 2) 根据某一区域内的客户订单交易数据以及该区域的特色构建初始评分矩阵; 最后, 融合 1) 和 2) 计算的结果作为全域用户项目评分矩阵.

假设用v_u表示客户u在全域范围内其自身的价值. 借鉴文献[19], 利用 RFM 模型通过对客户最近消费时间 R, 消费频率 F 以及消费金额 M 记录来计算客户价值. 其中, R (Recency) 表示客户最近一次交易时间的间隔; F (Frequency) 表示客户在给定的时段内消费的次数; M (Monetary) 表示客户在给定的时段内总共消费的金额数.

在 RFM 模型中, 对于时间间隔 R 来说, 当客户最近一次交易时间的间隔越短时, 则 R 值越大, 客户在短时间内也最有可能产生新的消费行为. 随着 R 的不断增大, 客户的相关信息也越来越完善, 因为随着时间间隔的不断缩短, 客户再次购买产品的可能性会逐渐变大. 对于消费频率 F 来说, 消费频率与客户忠诚度成正比, 客户消费频率越高, 说明该客户的忠诚度也越高, 为制造商创造的价值也越大. 对于消费金额 M 来说, 它是客户对制造商贡献值大小的最直接体现方式, 消费金额越大, 说明客户为制造商带来的价值也越大. 通过以上分析, 使用 RFM 模型从时间间隔, 消费频率和消费金额 3 个维度描述客户的消费行为, 可以较好的体现出客户为制造商所创造的现实价值, 也就是客户自身价值.

根据客户在最近一年内的购买行为, 利用 RFM 模型计算客户价值的过程如下:

1) 获取客户最近一年内消费时间 R, 消费频率 F, 消费金额 M 这 3 个行为指标;

2) R, F, M 按照其对大型制造商收益的贡献值大小将数据区间从高到低分别用 5, 4, 3, 2, 1 进行赋值;

3) 采用 z-score 标准化 (zero-mean normalization) 对 RFM 模型的指标数据进行标准化处理;

4) 利用层次分析法 (Analytic Hierarchy Process, AHP) 对 RFM 模型的指标权重进行评估;

5) RFM 模型中在已知 R, F, M 3 个指标权重分别为 a, b, c 的情况下, 计算客户价值v_u:

$$v_u = a*R + b*F + c*M \tag{8}$$

与此同时, 通过域子空间分解的方法, 先将全国范围内的客户数据按省, 市进行归类, 然后利用市内 POI 数据分布的特点, 将客户划分到旅游区域, 商业区域, 办公区域等不同的区域内. 根据不同区域客户购买产品的数量指标不同, 本文提出用区域特色系数来衡量客户购买产品情况.

对于市内的两个不同区域 (如商业区域和旅游区域), 位于商业区的客户占有地理位置优势, 平时客流量大, 此区域内的客户订单量大且购买产品的频次高, 制造商对该区域的偏好程度也大. 而位于旅游区的客户受季节性因素影响, 平时客流量小, 只有在旅游旺季产品购买量才会有明显上升. 为了描述这种因区域因素造成的客户购买产品情况的差异, 本文使用片区域特色系数表示市内不同区域购买产品的整体差异. 片区域特色系数是将市内各区域的客户购买量与该市的所有客户总购买量做比值运算, 比值越高的区域, 客户整体购买量也越高, 制造商对该区域的偏好程度也越大. 与此类似, 为了描述省内因城市内部因素造成的客户购买产品情况的差异, 本文使用市区域特色系数表示省内不同城市购买产品的整体差异, 用各个城市的客户购买量与该省所有客户的购买量的比值表示制造商对该城市的偏好程度. 为了描述全国因各省份内部因素造成的客户购买产品情况的差异, 本文使用省区域特色系数表示全国各省份购买产品的整体差异, 用各个省份的客户购买量与全国所有客户的购买量的比值表示制造商对该省份的偏好程度. 在某一片区域内, 则由客户自身购买产品的数量与该区域的客户购买产品总量的比值表示客户在片区域内与其他客户购买产品的差异, 用客户购买量系数来表示. 基于以上讨论, 本文使用客户所在区域, 所处城市和所属省份计算得到片区域特色系数, 市区域特色系数和省区域特色系数, 加上客户购买量系数表示全国范围内客户购买产品情况, 并以此构建初始的用户-项目评分矩阵:

$$r_{u,i}=\begin{cases} \mu, & \text{客户已购买产品} \\ 0, & \text{客户未购买产品} \end{cases} \tag{9}$$

其中, $r_{u,i}$表示用户u对项目i的初始评分. μ表示区域特色系数, 计算过程如下:

1) 计算全国客户总的购买量N_g, 省内客户的购买量N_s, 市内客户的购买量N_c, 片区域内客户的购买量N_p, 客户u的购买量N_u;

2) 计算省区域特色系数$\mu_s=N_s/N_g$, 市区域特色系数$\mu_c=N_c/N_s$, 片区域特色系数$\mu_p=N_p/N_c$和客户u的购买量系数$\mu_u=N_u/N_p$;

3) 采用 z-score 标准化对μ_s, μ_c, μ_p和μ_u分别进行标准化处理, 得到$\hat{\mu}_s$, $\hat{\mu}_c$, $\hat{\mu}_p$和$\hat{\mu}_u$;

4) 计算区域特色系数$\mu=\hat{\mu}_s+\hat{\mu}_c+\hat{\mu}_p+\hat{\mu}_u$.

最后, 本文给出用户u对项目i的最终评分:

$$r'_{u,i} = (1-\alpha)*r_{u,i} + \alpha*v_u \tag{10}$$

其中, $\alpha=1/\log(1+N(i))$, 表示用户-项目评分的平衡因子, v_u表示客户u的自身价值, $r_{u,i}$表示用户-项目的原始评分, $N(i)$表示客户购买产品的数量, $N(i)$越小, 表示该客户购买的产品数量越少, 同时$\log(1+N(i))$的值越小, α的值就越大. 对于新客户来说, 其购买产品的数量很小, 平衡因子α的值就会越大, 此时用户-项目的评分基本上是由客户自身价值来决定, 这在一定程度上解决了推荐系统冷启动问题.

最后根据$r'_{u,i}$得到改进后的用户项目评分矩阵$R'_{u,i}$, 并将数据归一化处理后映射到区间[1, 5], 得到归一化后的评分矩阵$R''_{u,i}$:

$$R''_{u,i} = \frac{(R_{\max}-R_{\min})*(R'_{u,i}-R_{\min})}{(R_{\max}-R_{\min})+R_{\min}} \tag{11}$$

2.2 基于全域市场数据感知的推荐算法框阵

基于全域市场数据感知的推荐算法 (Global Market Data Perception Matrix Factorization, GMF) 是利用客户属性信息约束矩阵分解的过程, 缓解推荐系统冷启动问题. 本文利用客户属性信息构建客户关系正则化项, 并假设当客户u和u'的属性信息相似时, 他们隐特征向量p_u和$p_{u'}$也尽可能相似.

根据文献[20], 利用与已知评分的最小平方逼近误差, 定义损失函数为:

$$\frac{1}{2}\min_{P,Q}\sum_{u=1}^{n}\sum_{i=1}^{m}I_{u,i}(R_{u,i}-P_u^{\mathrm{T}}Q_i)^2 \tag{12}$$

其中, $I_{u,i}$是指示函数, 等于 1 是表示客户 u 购买过产品 i, 等于 0 时表示客户 u 未购买过产品 i,$R_{u,i}$是已知评分矩阵.

为解决过拟合问题, 本文在上述模型的基础上加入低秩分解因子的范数$\|P\|_F^2$和$\|Q\|_F^2$对 P 和 Q 的训练过程进行控制, 使模型分解保持稳定. 考虑到客户之间的不同, 在损失函数中加入正则化约束项和偏置项信息:

$$\frac{\beta}{2}\sum_{u=1}^{N}\sum_{u'=1}^{N}\cos(u,u')\|p_u-p_{u'}\|_F^2+\frac{\lambda_1}{2}\|P\|_F^2 +\frac{\lambda_2}{2}\|Q\|_F^2+\frac{\lambda_3}{2}\|b\|^2 \tag{13}$$

其中, $\cos(u,u')$表示基于属性信息的客户u和u'间的相似度, p_u和$p_{u'}$表示分别客户u和u'的特征向量, β表示先验参数, 用于衡量客户相似度信息对矩阵分解约束的强度, 该值越大, 说明客户相似度的增强对于客户潜在特征表示越重要, λ_1, λ_2, $\lambda_3>0$表示正则项的调节参数, 其作用是防止过拟合.

3 实验分析

为了验证所提出的推荐算法的准确性, 本文在大型制造商真实数据集上进行了实验.

3.1 实验数据

由于本文使用的数据来自多个渠道 (如国家统计局, 营销人员走访客户采集, 相关并行系统以及互联网等), 因此需要先将数据清洗操作, 然后将多数据源进行实体唯一, 属性唯一, 编码取值统一及数据全链路的全域一致性处理, 最后集成到数据中台. 最后从数据中台中收集 2018 年全国范围内的客户基本数据及订单交易数据, 进行实验研究. 其中, 数据库客户订单交易部分字段如表 1 所示.

3.2 评价指标

本文选择平均绝对误差 (Mean Absolute Error, *MAE*) 和均方根误差 (Root Mean Squared Error, *RMSE*) 两种评价指标来评估推荐算法的质量. 下面是两种误差的计算方法:

$$MAE=\frac{1}{|T|}\sum_{u,i\in T}\left|r_{u,i}-\hat{r}_{u,i}\right| \tag{14}$$

$$RMSE=\sqrt{\frac{1}{|T|}\sum_{u,i\in T}\left|r_{u,i}-\hat{r}_{u,i}\right|^2} \tag{15}$$

其中, $r_{u,i}$和$\hat{r}_{u,i}$分别表示实际的评分值和推荐预测的评分值, T 表示测试数据集大小. *MAE* 或 *RMSE* 的值越小, 推荐算法的推荐质量越高.

表 1 客户订单交易部分字段示例

字段	描述
ORDER_ID	交易编号
USER_ID	客户编号
USER_ADDRESS	客户地址
USER_TYPE	客户类型
COMMODITY_NAME	交易产品名称
COMMODITY_QUANTITY	交易产品数量
ORDER_TOTALAMOUNT	交易金额
COMMODITY_TYPE	交易产品类型
COMMODITY_START_TIME	订货时间
COMMODITY_END_TIME	收货时间

3.3 实验过程与结果分析

3.3.1 RFM 客户聚类实验

为了验证利用客户最近消费时间, 消费频率以及消费金额 3 个因素计算客户价值的合理性, 我们利用 K-Means 聚类算法进行验证.

通过层次分析法计算得到 R 参数权重a为–0.287, F 参数权重b为 0.548, M 参数权重c为 0.165. 将 RFM 各参数作为聚类变量, 利用 K-Means 聚类算法将客户分为 5 类后, 分别计算这 5 类客户的 R 参数, F 参数及 M 参数的平均值, 将其代入式 (8) 中, 可得用户分类结果如表 2 所示.

表 2 RFM 模型客户分类结果

类别	R 参数值	F 参数值	M 参数值	客户价值	客户占比	分类精确度 (%)
1	–3.86	3.54	32.75	8.45	0.014	98
2	–3.67	2.92	5.50	3.56	0.285	91
3	–2.32	3.44	1.92	2.85	0.058	87
4	–2.16	2.59	1.61	2.30	0.043	92
5	0.35	–0.23	–0.35	–0.28	0.600	99

通过指标分析可将客户分为监管户, 流失户, 价值户, 连锁户和核心户 5 种类别. 从分类结果呈现非线性聚集和客户分类的精准度可以看出, 利用 RFM 模型计算客户价值可有效区分各不同类别客户购买产品能力

的差异.

3.3.2 区域特色系数的影响实验

为了验证区域特色系数在用户对项目最终评分准确度的影响, 我们将未加入区域特色系数前和加入区域特色系数后的用户项目评分进行了对比.

由图 1 可知, 未加入区域特色系数前的用户项目评分大多集中在 2–3 分之间, 而加入区域特色系数以后的用户项目评分在 1–2, 2–3 和 3–4 间的比例均有增加, 消除了因客户价值导致的项目评分趋于一致性的问题, 使得用户-项目评分更加准确.

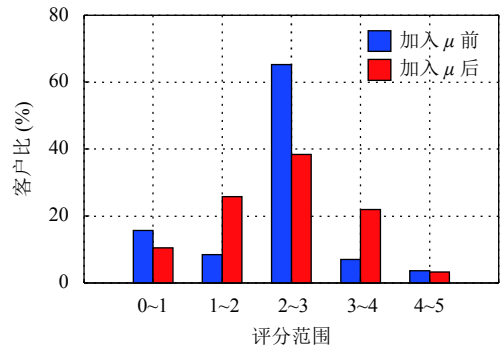

图 1 加入区域特色系数前后的客户比

3.3.3 与经典方法的对比实验

为了验证基于全域市场数据感知算法的有效性, 本文选择以下经典方法在 *MAE* 和 *RMSE* 两个指标上进行对比实验:

(1) PMF[8]: 该方法仅考虑用户对物品的评分信息进行概率矩阵分解预测缺失项.

(2) MMMF[9]: 该方法引入计算序数回归排序损失函数进行矩阵分解模型预测缺失项.

(3) NMF[10]: 该方法限定在训练学习过程中隐特征向量更新仅包含非负项进行矩阵分解预测缺失项.

(4) RSVD[11]: 该方法基于 SVD 模型中引入正则化项进行奇异值分解预测缺失项.

(5) BPMF[12]: 该方法使用马尔科夫链蒙特卡洛方法进行近似推理预测缺失项.

(6) SVD++[13]: 该方法同时考虑偏置信息以及用户隐式反馈信息进行矩阵分解预测缺失项.

为了公平比较, 我们根据各个对比算法的参考文献或者实验结果设置对比算法的参数. 在这些参数设置下, 各对比算法取得最佳性能. 我们设置 $\lambda_1=\lambda_2=\lambda_3=0.01$, 学习率 $\eta=0.005$, 同时, 我们将处理后的数据集每次随机抽取 80% 的数据作为训练数据, 剩下的 20% 的数据作为测试数据进行 5 折交叉验证. 最后, 取 5 次不同测试数据集上运行结果的平均值作为实验的 *MAE* 和 *RMSE* 的最终结果. 实验结果如表 3 所示.

表 3 GMF 与其他算法质量对比

K	Metric	PMF	MMMF	NMF	RSVD	BPMF	SVD++	GMF
5	*MAE*	0.7652	0.7796	0.7782	0.7695	0.7502	0.7683	**0.7401**
	RMSE	0.9584	0.9684	0.9697	0.9523	0.9442	0.9536	**0.9335**
10	*MAE*	0.7615	0.7785	0.7762	0.7565	0.7525	0.7651	**0.7416**
	RMSE	0.9541	0.9631	0.9656	0.9496	0.9464	0.9523	**0.9362**

(1) 实验参数 *K* 的影响

隐特征向量的维数 *K* 的取值对推荐算法的性能有很大影响, 在实验中, 本文将 *K* 初始值设置为 5, 同时设置步长为 5, 直至 *K* 值递增到 50.

实验结果如图 2 和图 3 所示.

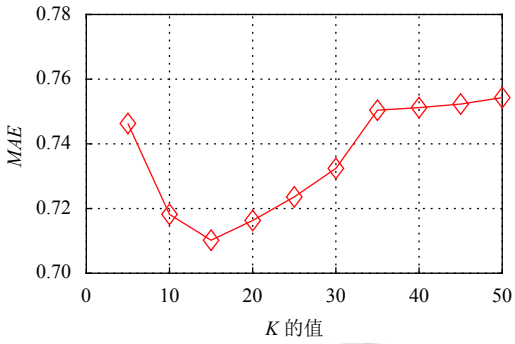

图 2 *K* 对 *MAE* 的影响

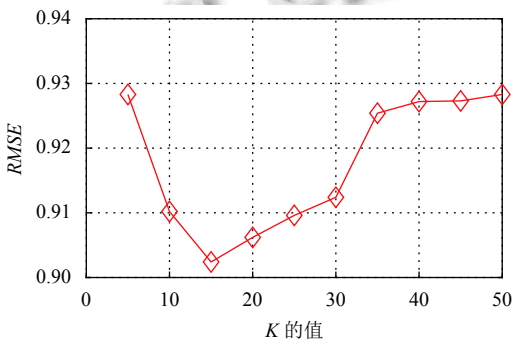

图 3 *K* 对 *RMSE* 的影响

当 *K* 小于 15 时, 推荐算法随着 *K* 的增加其质量不断提高, 但当 *K* 大于 15 以后继续增加 *K* 的值推荐算法的质量不再提高. 这说明隐特征数量的增加会在一定范围内提高推荐算法质量, 一旦超过某一阈值以后可能就不会再提高推荐算法的质量. 造成这一现象的原因可能是本文所选的数据集在 *K* 大于 15 以后用户和项目的隐特征向量已经能够很好的刻画其隐特征,

而继续增加 K 的值反而会因为噪音的影响降低了推荐算法的质量.

(2) 实验参数β的影响

β控制着 GMF 算法中的客户的属性信息对学习隐特征向量的影响. 若β=1时, 客户隐特征向量将直接与它邻居的特征向量相似, 忽略了评分数据的影响; 若β= 0时, 仅使用评分信息进行矩阵分解预测缺失评分. 本文在大型制造商数据集上, 设置隐特征向量维度 K 为 10, β的值从 0 到 1 并以步长 0.1 的间隔逐渐增加. 实验结果如图 4 和图 5 所示, 随着β值的增长, *MAE* 和 *RMSE* 的值先下降后递增. 这说明β的值一旦超过某一阈值后, 推荐算法的性能就会下降. 也就是说, 不依赖或完全依赖客户属性信息都会使得推荐系统性能下降, 推荐结果不可靠.

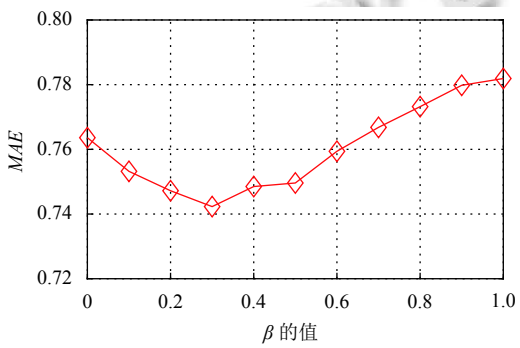

图 4 β对 *MAE* 的影响

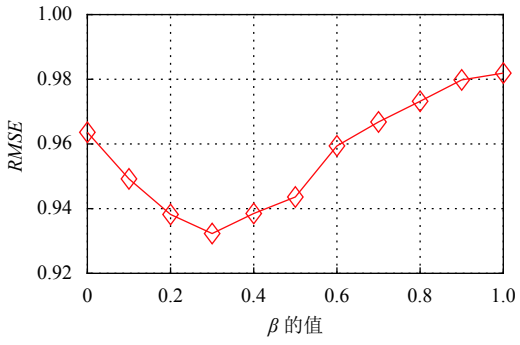

图 5 β对 *RMSE* 的影响

(3) 冷启动对推荐系统性能的影响

为了验证利用客户自身价值计算用户-项目的评分缓解推荐系统冷启动问题的有效性, 我们根据项目的评分数量对数据集进行分组后, 在每个组上与其他推荐算法做对比分析.

在所选取的大型制造商 35 万数据集上, 先根据项目的评分数量情况分成了 6 组, 分别是“0”, “1–20”, “21–40”, “41–80”, “81–160”, “>160”, 然后对分组后的数据集上进行对比实验. 实验结果如图 6 和图 7 所示, 从图中可以看出, 本文使用客户自身价值计算用户项目评分在 6 组实验中均有好的推荐效果, 特别是在评分少的项目上效果比较明显, 说明在一定程度上缓解了推荐系统冷启动问题. 为了进一步解决冷启动项目对推荐系统性能的影响, 对于新用户将通过以用户所在地理位置为圆心, 向外进行雷达扩散式寻找周边近距离的客户进行推荐. 随着评分数量的增多, GMF 方法相比其他推荐方法的仍然有一定的优势. 这是因为本文使用在矩阵分解推荐算法的过程中考虑了耦合对象相似来捕获客户间的属性特征, 从而产生更加可靠的推荐结果.

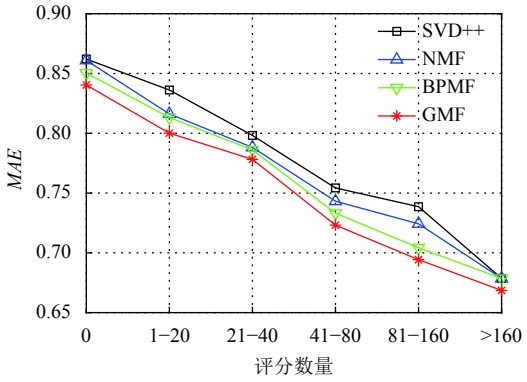

图 6 评分数量对 *MAE* 的影响

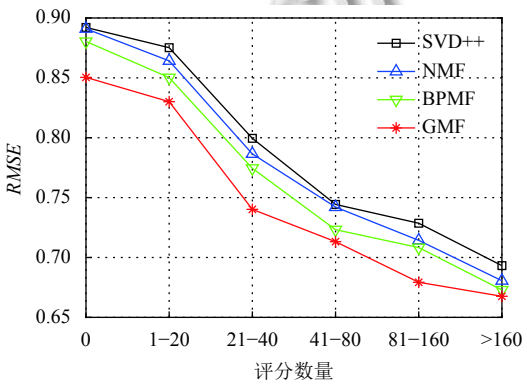

图 7 评分数量对 *RMSE* 的影响

4 结论与展望

本文提出一种基于全域市场数据感知的推荐方法 GMF 寻找最佳目标客户. 首先根据客户订单交易数据及客户属性信息获取原始用户-项目评分矩阵和客户自身价值, 然后在两者间引入平衡因子α通过归一化处理后得到最终的用户-项目评分矩阵. 再根据用户属性

通过耦合对象相似度计算客户间的相似度为产品寻找目标客户. 在大型制造商数据集上进行的实验表明, 本文提出的算法在准确性上优于当前流行的典型推荐算法. 同时, 在大型制造商精准营销实践中的结果表明: 利用本文提出的 GMF 方法效益提升了 26.8%.

在后续的研究中, 将针对 RFM 模型进行进一步研究, 因为 RFM 模型中 3 个指标描述的是客户的行为特征, 并不能代表客户的大多数行为, 为了更好的衡量客户价值, 可以考虑将客户的第一次交易至最近一次交易期间的间隔时长, 某一时间段内的最高消费金额和客户平均收入等因素考虑在内进行模型优化, 建立一个更加全面准确的客户价值体系.

参考文献

1 Wang Y, Deng JZ, Gao J, *et al*. A hybrid user similarity model for collaborative filtering. Information Sciences, 2017, 418–419: 102–118. [doi: 10.1016/j.ins.2017.08.008]

2 Lü LY, Medo M, Yeung CH, *et al*. Recommender systems. Physics Reports, 2012, 519(1): 1–49. [doi: 10.1016/j.physrep.2012.02.006]

3 Lika B, Kolomvatsos K, Hadjiefthymiades S. Facing the cold start problem in recommender systems. Expert Systems with Applications, 2014, 41(4): 2065–2073. [doi: 10.1016/j.eswa.2013.09.005]

4 Ji K, Shen H. Addressing cold-start: Scalable recommendation with tags and keywords. Knowledge-Based Systems, 2015, 83: 42–50. [doi: 10.1016/j.knosys.2015.03.008]

5 Forsati R, Mahdavi M, Shamsfard M, *et al*. Matrix factorization with explicit trust and distrust side information for improved social recommendation. ACM Transactions on Information Systems (TOIS), 2014, 32(4): 1–38.

6 Koren Y, Bell R, Volinsky C. Matrix factorization techniques for recommender systems. Computer, 2009, 42(8): 30–37. [doi: 10.1109/MC.2009.263]

7 Hernando A, Bobadilla J, Ortega F. A non negative matrix factorization for collaborative filtering recommender systems based on a Bayesian probabilistic model. Knowledge-Based Systems, 2016, 97: 188–202. [doi: 10.1016/j.knosys.2015.12.018]

8 Salakhutdinov R, Mnih A. Probabilistic matrix factorization. Proceedings of the 20th International Conference on Neural Information Processing Systems. Vancouver, BC, Canada. 2008. 1257–1264.

9 Weimer M, Karatzoglou A, Smola A. Improving maximum margin matrix factorization. Machine Learning, 2008, 72(3): 263–276. [doi: 10.1007/s10994-008-5073-7]

10 Huang KJ, Sidiropoulos ND, Swami A. Non-negative matrix factorization revisited: Uniqueness and algorithm for symmetric decomposition. IEEE Transactions on Signal Processing, 2014, 62(1): 211–224. [doi: 10.1109/TSP.2013.2285514]

11 Nguyen J, Zhu M. Content-boosted matrix factorization techniques for recommender systems. Statistical Analysis and Data Mining, 2013, 6(4): 286–301. [doi: 10.1002/sam.11184]

12 Salakhutdinov R, Mnih A. Bayesian probabilistic matrix factorization using Markov chain Monte Carlo. Proceedings of the 25th International Conference on Machine Learning. New York, NY, USA. 2008. 880–887.

13 Koren Y. Factorization meets the neighborhood: A multifaceted collaborative filtering model. Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA. 2008. 426–434.

14 Yang B, Lei Y, Liu JM, *et al*. Social collaborative filtering by trust. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(8): 1633–1647. [doi: 10.1109/TPAMI.2016.2605085]

15 Gurini DF, Gasparetti F, Micarelli A, *et al*. Temporal people-to-people recommendation on social networks with sentiment-based matrix factorization. Future Generation Computer Systems, 2018, 78: 430–439. [doi: 10.1016/j.future.2017.03.020]

16 Gan GQ, Ma CQ, Wu JH. Data Clustering: Theory, Algorithms, and Applications. Philadelphia: SIAM, 2007.

17 Yu YH, Wang C, Wang H, *et al*. Attributes coupling based matrix factorization for item recommendation. Applied Intelligence, 2017, 46(3): 521–533. [doi: 10.1007/s10489-016-0841-8]

18 Lian DF, Zheng K, Ge Y, *et al*. GeoMF++: Scalable location recommendation via joint geographical modeling and matrix factorization. ACM Transactions on Information Systems (TOIS), 2018, 36(3): 33.

19 Chiang WY. Identifying high-value airlines customers for strategies of online marketing systems: An empirical case in Taiwan. Kybernetes, 2018, 47(3): 525–538. [doi: 10.1108/K-12-2016-0348]

20 Pirasteh P, Hwang D, Jung JJ. Exploiting matrix factorization to asymmetric user similarities in recommendation systems. Knowledge-Based Systems, 2015, 83: 51–57. [doi: 10.1016/j.knosys.2015.03.006]