基于天猫交易数据的分析与挖掘*

王希梅
清华大学 软件学院
北京 100084
wxm17@mails.tsinghua.edu.cn

于千山
清华大学 软件学院
北京 100084
yqs17@mails.tsinghua.edu.cn

陈善宇
清华大学 软件学院
北京 100084
abc321094@gmail.com

摘要

我们利用 2014 年双 11（11 月 11 日）前六个月天猫的用户行为日志进行数据挖掘任务。在本次任务中，我们提出四个问题，并分别设计算法进行解决，同时给出了实验结果并进行讨论。这几个问题的解决，对于商家降低促销成本，提高投资回报率（ROI）非常重要。

关键词

数据挖掘，聚类分析，行为预测

1. 问题描述

我们提出以下四个问题：

1. 商品频繁模式挖掘：即分析哪些商品会被一起购买。
2. 同一商家用户的聚类分析: 商家会拥有很多用户，对于商家来说，如果能够把握用户的行为，清楚用户的类别，就能够据此来针对不同用户提升服务品质，而为客户提供更好的服务的同时，也能大大的提升投资回报率。
3. 查找相似商家：对商家来说，除了透过促销吸引客户注意之外，关注自身的竞争对手也是非常重要的，由此来分析相似商家的商品及价格，商家也能为自家的商品找到更好的价格定位。
4. 用户重复购买的预测：为了吸引大量的新买家，商家有时会在特定的日子进行大促销)。然而，很多买家都是一次性买家，这些促销活动可能对销售的影响不大。我们期望预测未来哪些特定商家的新买家将成为忠实客户。这些新买家将来再次购买同一批商品的可能性。

*具体数据集信息参见 https://tianchi.aliyun.com/datalab/dataSet.htm?id=5

2. 方法设计

这一节中，我们考虑对第 1节提出的问题进行方法设计。

2.1 商品频繁模式挖掘

我们使用 FPGrowth 算法进行商品关联规则挖掘，FPGrowth 算法使用分治的策略将频繁项集的数据放进一棵 FP-Tree 中，再对其进行挖掘，寻找频繁模式。其中算法主要可以分为: 建构 FP-Tree 和挖掘频繁模式两个部分。在本问题中，根据 action_type 可以筛选出用户在商家的购买记录，根据购买记录建立 FP-Tree, 再通过 FPGrowth 算法求解频繁项集，取 min_support 为 3%。与此同时，本文还对关联规则进行了探究，取最小置信度 min_prob 为 0.9.

2.2 同一商家用户的分类

对于每一商家而言，找出用户的客群 (Target Market) 有助于让商家针对其客群，调整自身销售的产品类型、种类以及价格，其中年龄 `age_range` 与性别 `gender` 最直接的影响其消费购买能力以及消费的商品种类，是以我们希望借由这样的分类，能让商家更了解更能因应客户调整商品，也能得到更好的投资回报率。因此我们选择年龄和性别作为特征，尝试对提取的特征进行 tsne 降维和 meanshift 聚类。

2.3 查找相似商家

我们基于商家的共同用户数量来判断相似商家。对于商家来说，可以找到其竞争对手，方便其促销和价格比较。考虑到问题的效率和数据规模，我们选用共同用户数量作为相似度度量特征。实现中，用户的输入为待查找的商家编号 mid，我们使用以下步骤来计算相似商家：

1. 找出 mid 对应的用户集合 `user_set`；
2. 对于 `user_set` 中的每个元素 u_i，分别找到对应的商家集合 M_i；
3. 对 M_i 多重集合 M（多重集合同时保存了元素及其计数值，其中计数值为共同用户数量）；
4. 返回 M 中计数值前五的元素作为该商家 mid 的相似商家。

2.4 用户重复购买的预测

我们提取了维特征，具体信息如表 1所示。并选用了 Random Forest 、xgboost、Logistic Regression 等三种模型，使用 train_format1.csv 中 80% 的数据训练网络，另外 20% 的数据进行预测。三种模型的具体内容如下：

Table 1: 属性列表

名称	描述
similar_merchant_id[5]	商家的最相似的 5 个商家
similar_merchant_num[5]	商家与其相似商家的共同用户数量
click_num	当前用户对此商家的点击数
add_to_favourite_num	当前用户对此商家的收藏数
add_to_cart_num	当前用户对此商家的添加购物车数
purchase_num	当前用户对此商家的购买数
action_ratio_in11	当前用户对此商家於双十一的四种行为比例
u_repeat_buy_ratio	用户重复购买的商家占用户所购买的所有商家之比例
u_repeat_buy_before...	用户在双十一之前重复购买比例
u_age_range	用户年龄
u_gender	用户性别
u_action_days	用户行为天计数
u_daily_action_factor	用户平均每天行为因子 (4 种操作加权和)
u_click_ratio_in11	双十一的点击行为占所有时间点击行为之比例
u_add_to_cart...	双十一的加购物车行为占所有时间加购物车行为之比例
u_purchase_ratio_in11	双十一的购买行为占所有时间购买行为之比例
u_fav_ratio_in11	双十一的收藏行为占所有时间收藏行为之比例
u_action_ratio_in11	双十一的所有行为占所有时间所有行为之比例
u_is_new_user	用户是否为新用户
m_repeat_purchased...	所有用户在当前商家重复购买的比例
m_clicked_num	所有用户对当前商家的点击数
m_faved_num	所有用户对当前商家的收藏数
m_added_to_cart_num	所有用户对当前商家的添加购物车数
m_purchased_num	所有用户对当前商家的购买数
m_repeat_purchased...	双十一之前重复购买的比例
m_purchased_11_ratio	双十一当天该商家购买数占所有购买数的比例
m_regular_user_ratio	该商家所拥有的老用户比例

1. Random Forest: 透过训练多个决策树，随机森林能够避免过拟和的状况发生，在预测时，每个树的都会预测一个结果，再将每个结果加权表决来决定最后的预测结果。
2. xgboost: 主要是用来解决有监督学习问题，利用包含多个特征的训练数据，来预测目标变量。
3. Logistic Regression: 逻辑回归通过历史数据的表现对未来结果发生的概率进行预测，回归分析用来描述自变量 x 和因变量 Y 之间的关系，或者说自变量 X 对因变量 Y 的影响程度，并对因变量 Y 进行预测。

我们的实验框架如图 1所示。

3. 结果与分析

我们所有的实验均在如下所示的主机上运行：

- **CPU:** 2.3GHz Xeon E5 CPU 72 Cores
- **RAM:** 250GB

在实际的模型中，本文首先测试了 xgboost、Random Forest、Logistic Regression 的 3 种不同模型的性能表现，如图 2， 3，

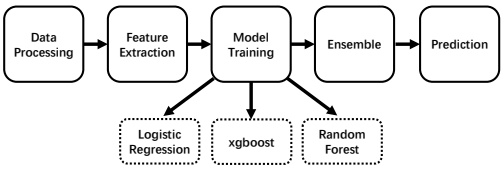

Figure 1: 多模型融合的预测分析框架

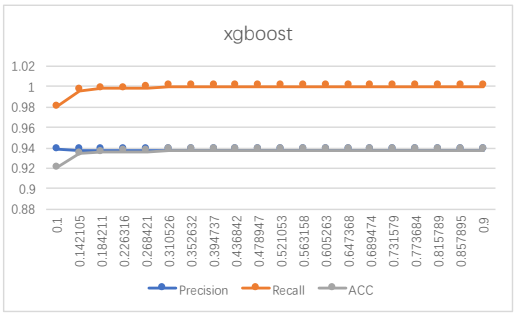

Figure 2: xgboost 模型的性能表现

4所示。每种模型的性能表现都展示了准确性、精确度、召回率随概率阈值的变化情况，可以看到，xgboost 模型的性能提升最快，Random Forest 次之，Logitstic Regression 最慢，但最终的准确性都相近，在图中 5，展示了 3 种模型融合后的效果，而图 6中展示了不同模型与融合模型的准确率的比较。

我们将得到的结果以图表的形式呈现。

4. 参考文献

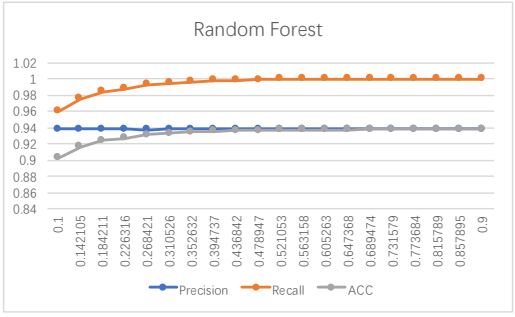

Figure 3: RF 模型的性能表现

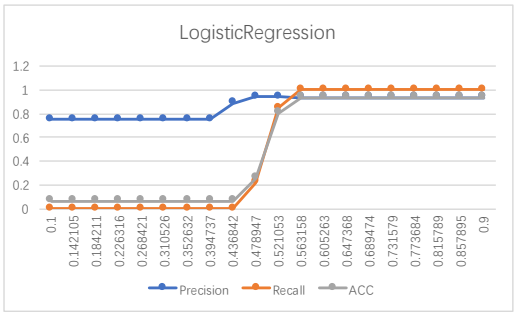

Figure 4: LR 模型的性能表现

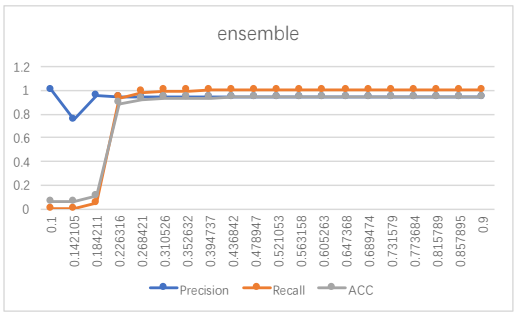

Figure 5: ensemble 模型的性能表现

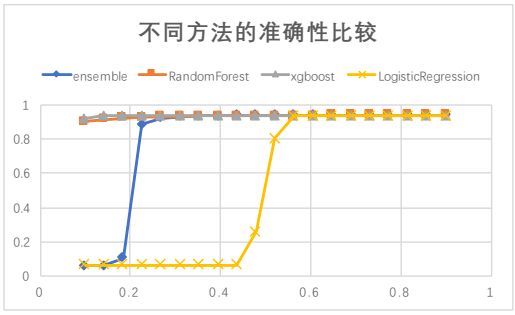

Figure 6: 不同模型的准确率对比