基于加权回归模型预测商家未来客流量

BRYAN 桑榆 云泛天音

摘要

IJCAI-17 口碑商家客流量预测(Passengers flow forecast for Koubei's merchants) 是一个根据商家历史支付流量数据来预测口碑网商家在未来一段时间内支付流量的问题。本次竞赛给出了 2000 个商家在历史 16 个月的客流量数据，我们的任务是对这 2000 个商家在未来两周内每天的客流量做出预测。显然这是一个回归问题。本文详细介绍了 CAT 战队所采用的两种回归方法，它们分别是时间序列加权回归模型（Time Series Weighted Regression, TSWR），和基于现有树模型的回归模型（Xgboost and RandomForest Regression,XRR）。在竞赛第一阶段我们队在 A 榜排名 11，在竞赛第二阶段我们队在 B 榜排名第一。

1 简介(introduction)

随着移动定位服务的流行，阿里巴巴和蚂蚁金服逐渐积累了来自用户和商家的海量线上线下交易数据。蚂蚁金服的 O2O 平台“口碑”用这些数据为商家提供了包括交易统计，销售分析和销售建议等定制的后端商业智能服务。举例来说，口碑致力于为每个商家提供销售预测。基于预测结果，商家可以优化运营，降低成本，并改善用户体验。

预测客户流量对商家的经营管理至关重要。在口碑平台上，我们将客户流量定义为“单位时间内在商家使用支付宝消费的用户人次”。在这个问题中，我们将使用口碑网提供的用户浏览和支付历史数据，以及商家相关信息，并在论坛获取了相应时间的天气、空气质量数据。以此预测所有商家在接下来 14 天内，每天的客户流量。

IJCAI-17 口碑商家客流量预测的目的在于，鼓励选手使用先进人工智能技术来解决上述问题。参赛选手可以获得国内最大 O2O 平台口碑网大量的商家用户日志数据。与过去其他数据挖掘竞赛不同的是，本次竞赛提供了大量商家历史流量数据，和海量用户行为数据，并且可以自由使用其他平台相关的数据。

本次竞赛中，我们可以使用从 2015.07.01 到 2016.10.31（除去 2015.12.12）

的 69,674,110 条大约 2.1G 用户支付行为数据，5,556,715 条大约 174M 用户浏览行为数据和 2000 个商家信息数据，参赛选手可以自由使用天气数据。需要选手预测未来 14 天每个商家每天的支付流量，其中 1000 个商家用于 A 榜测试，另外 1000 个商家用户 B 榜测试。比赛方式为线下训练模型，提交预测结果。开发的工具使用没有限制。

IJCAI-17 口碑商家客流量预测的评估公式如下：

$$\mathrm{L} = \frac{1}{nT}\sum_{i}^{n}\sum_{t}^{T}\left|\frac{c_{it} - c_{it}^{g}}{c_{it} + c_{it}^{g}}\right| \tag{1}$$

其中，L 为排行榜显示的选手测评结果；c_{it}为第 t 天，商家 i 的客户流量预测值(由参赛选手提供)；c_{it}^{g}为第 t 天，商家 i 的客户流量实际值。

在本次竞赛中，我们主要使用了两种回归方法，它们分别是时间序列加权回归模型（Time Series Weighted Regression, TSWR），和基于现有树模型的回归模型（Xgboost and RandomForest Regression）。在时间序列加权回归模型中，我们将目标拆分为最有常值回归，加权回归，三种周期性回归，以及对天气的优化。在基于树模型的回归中，基于历史信息，商家信息，日期信息，天气信息来对单个和整体商家进行建模。建模框架如图 1 所示：

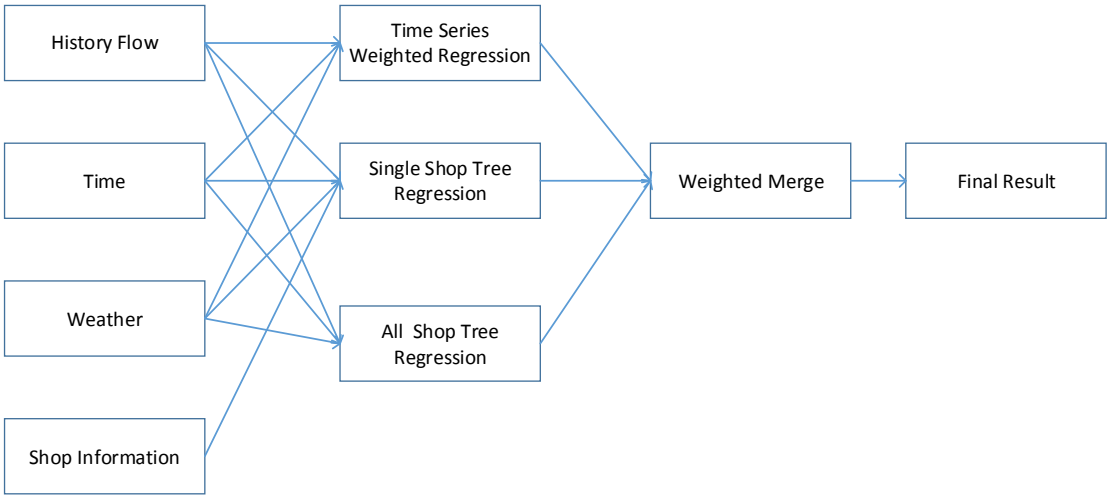

图 1 建模框架

我们 CAT 团队使用简单方案在初赛中排名 11，使用融合方案在决赛中排名第 1，证明了我们团队所使用的时序回归方法和树模型方法融合后在商家客流量预测中的有效性。

2 数据分析和预处理(data analysis and preprocessing)

2.1 数据分析(data analysis)

2.2 线下验证(offline evaluation)

2.3 数据转换(data transforming)

本次比赛额外数据来自于天池论坛其他选手公开的天气与雾霾信息数据，我们将这些数据 one-hot 数值化后作为样本特征加入到模型训练中。对于周均值的变化程度使用方差来衡量，找出大于方差阈值的商家。直接删除波动过大的时段的数据，例如：国庆节。

3 模型设计(model design)

我们采用了时间序列加权回归模型和树回归模型，将两种模型以 7:3 的比例进行融合，得到最终的预测结果。

3.1 时间序列加权模型(Time Series Weighted Regression)

对每个商家的历史客流量进行加权回归，并在测试时间段做出预测。

3.1.1 最优常值回归

我们将 2016.10.11-2016.10.31 及 2016.9.20-2016.9.26 共计 28 天作为训练时间段。选择这 28 天的原因是，这 28 天是距离测试时间段最近并且不受 2016 年的国庆节及中秋节影响的 28 天，而更久远的时间段与测试时间段之间的差异过大，没有被包含在其中。

一个最基本的思路是，对于每个商家，将测试时间段的 14 天的客流量预测为其训练时间段的客流量的均值，即

$$Y_A = \frac{\sum_{m \in M}(K_{A,m})}{|M|} \tag{2}$$

式中，Y_A 为商家 A 的预测值；M 为训练时间段日集；|M|表示训练时间段的日数；$K_{A,m}$ 表示商家 A 在日 m 的客流量值。我们把均值看做是对训练时间段的客流量的回归，那么显然我们可以找到一个比均值更优秀的回归值，这个回归值应该满足在训练时间段的 28 天上的损失值最小。我们将这个值作为这个商家在待测的 14 天上的预测。即

$$\mathrm{Y_A} = argmin_y \left(\sum_{m \in M} L(y, K_{A,m}) \right) \tag{3}$$

式中， L 为赛题给定的损失函数。显然 $\mathrm{Y_A}$ 满足

$$\mathrm{Y_A} \in \left[\min(K_{A,m}), \max(K_{A,m}) \right], m \in M \tag{4}$$

由于小于 1 的预测值差异对成绩的影响基本可以忽略，为简单起见，我们遍历(4)所确定的范围内的所有整数，选择其中最优秀的 $\mathrm{Y_A}$。

3.1.2 时间加权回归

通过对数据的观察和对问题的分析，我们发现商家的经营情况是不断变化的，越靠近测试时间段的样本重要性越大；此外，相对于老商家，一个刚刚开始营业的商家对上述的重要性应该更加敏感。因而我们给选取的每一天加上一个权重，使得越靠近测试时间段的样本的权重越大，同时对于一个商家，越远离该商家开始营业的时间的样本的权重越大。经过线下测试，我们设计了如下三个权重函数：

$$\mathrm{w}_{m,A,1} = m - \min(M_A) \tag{5}$$

$$\mathrm{w}_{m,A,2} = (m - \min(M))^3 \times \left(m - \min(M_A) \right) \tag{6}$$

$$\mathrm{w}_{m,A,3} = \frac{1}{m_0 - m} \tag{7}$$

式中 $\mathrm{M_A}$ 表示商家 A 的所有营业的日子；$\mathrm{m_0}$ 表示待测时间段的首日；二日相减表示二者的日数距离。我们将(5)、(6)、(7)三者融合，得到最终采用的权重函数

$$\mathrm{w}_{m,A} = \alpha \mathrm{w}_{m,A,1} + \beta \mathrm{w}_{m,A,2} + \gamma \mathrm{w}_{m,A,3} \tag{8}$$

式中 α，β，γ 均为 0，1 之间的实数且满足 α+β+γ=1。我们依据线下测试选择了合适的 α，β，γ，并将(2)改进为

$$\mathrm{Y_A} = argmin_y \left(\sum_{m \in M} \left(L(y, K_{A,m}) \times w_{m,A} \right) \right) \tag{9}$$

3.1.3 曜日(day of week)周期回归

通过对数据的观察和对问题的分析，我们发现商家的客流量具有明显的周期性，而且不同商家的周期性不同。例如一些美食商家周末的客流量更低，而一些超市受周末的影响不大，而不同的美食商家的情况也不相同。为此，我们设计了曜日（day of week）权重函数

$$v(\psi_1, \psi_2)$$

其中，ψ_1 为训练时间段中的一个曜日；ψ_2 为测试时间段的一个曜日。v 是曜日权重函数，且满足若当 ψ，χ 同为工作日或同为周末时，该函数得返回值较

大，否则返回值较小。凭借对数据的观察，并依据线下测试集，我们确定了 v 的形式。此时，(9)被改进为

$$Y_{A,\mu,1} = argmin_y \left(\sum_{m \in M} \left(L(y, K_{A,m}) \cdot w_{m,A} \cdot v(\psi(m), \psi(\mu)) \right) \right) \quad (10)$$

式中，μ 为测试时间段的一天；ψ(m)表示 m 日的曜日。

(10)式的问题是对样本的利用程度不高，预测某一天时，不能尽用每一个样本。而时间序列问题的特点是时间样本较少，因而样本利用程度非常重要。为此，我们又做了如下的改进。

我们选取排除掉节日等异常因素的近六个月的时间段作为第二训练时间段。在这个时间段上，我们为每个曜日计算出一个最优系数，即能够在乘以这个系数后最小化损失值的系数，即

$$p_{A,\psi} = argmin_p \left(\sum_{m \in M'_\psi} \left(L(pY_{(3),A,m}, K_{A,m}) \right) \right) \quad (11)$$

式中，$p_{A,\psi}$ 表示商家 A 在曜日 ψ 上的曜日占比；M'为第二训练时间段的日集，M'_ψ 表示第二时间段上属于曜日 ψ 的日集，$Y_{(3),A,m}$ 表示以 m 所在周作为训练时间段，利用(4)的形式对商家 A 在 m 日的客流量做出的预测。我们排除训练样本中的周期性成分，然后重新加在对测试时间段的预测中，即

$$Y_{A,\mu,2} = argmin_y \left(\sum_{m \in M} \left(L\left(y, \frac{K_{A,m}}{p_{A,\psi(m)}}\right) \cdot w_{m,A} \cdot v(\psi(m), \psi(\mu)) \right) \right) \cdot p_{A,\psi(\mu)} \quad (12)$$

(12)式的问题是曜日占比 $p_{A,\psi}$ 难以准确计算。(11)和(12)各有优缺点，我们将二者加权相加，即

$$Y_{A,\mu} = \delta Y_{A,\mu,1} + \varepsilon Y_{A,\mu,2} \quad (13)$$

式中 δ，ε 均为 0，1 之间的实数且满足 δ+ε=1。我们依据线下测试选择了合适的 δ，ε。

3.1.4 天气因素

我们在天池论坛上取得了赛题训练时间段的天气情况及测试时间段的天气预报情况。通过对数据的观察和对问题的分析，我们发现商家的客流量受天气情况的影响，例如大多数商家在阴雨天气的客流量明显比晴天的客流量低。我们按照暴雨、大雨、中雨、小雨、阴、多云、晴等天气将天气数值化，并依据线下测试将该数值线性映射为系数。我们将(10)，(11)分别改进为

$$Y_{A,\mu,3} = argmin_y \left(\sum_{m \in M} \left(L\left(y, \frac{K_{A,m}}{t(m,s(A))}\right) \cdot w_{m,A} \cdot v(\psi(m), \psi(\mu)) \right) \right) \cdot t(\mu, s(A)) \quad (14)$$

$$Y_{A,\mu,4} = argmin_y \left(\sum_{m \in M} \left(L\left(y, \frac{K_{A,m}}{p_{A,\psi(m)} \cdot t(m,s(A))}\right) \cdot w_{m,A} \cdot v(\psi(m), \psi(\mu)) \right) \right) \cdot p_{A,\psi(\mu)} \cdot t(\mu, s(A)) \quad (15)$$

式中 s(A)表示商家 A 所在地的市名，t(m,s)表示 m 日 s 市的天气系数。并将(13)改进为

$$Y_{A,\mu} = \delta Y_{A,\mu,3} + \varepsilon Y_{A,\mu,4} \quad (16)$$

(16)中的结果便是时间序列加权回归模型预测的最终结果。

3.2 回归树模型(Xgboost and RandomForest Regression,XRR)

在本次竞赛中，我们采用了两种基于商家信息，日期信息，天气信息的回归建模方法： (1)单独对每个商家训练一个模型。每个商店的时序都有各自的特点，这样可以减小商家与商家之间的影响；(2)所有的商家一起建模，我们在方法 1 的基础上增加了商家特征，从 shop_info 表里将每个商家的信息处理成数值特征。

我们使用了额外的数据，比如热心参赛者在论坛中共享的天气和雾霾数据。因为人们外出与天气和空气质量有关、如果当天天气好，并且空气质量较高，那么那一天用户出行消费的概率越大，天气数据和雾霾数据是有用的。在预处理阶段，我们对类别特征和天气特征都经过了 one-hot 编码处理。

在本次竞赛中，Xgboost 和 RandomForest 具有较好的表现，故选取它们为回归模型。Xgboost 具有较强的泛化能力，同时能自主的选取特征，也能够减小我们做特征选择的工作量。此外，我们还选取了 Random Forest 模型，有文献显示，随机森林能较好的控制过拟合现象，此外，在实际的比赛和项目的使用中，我们发现，在数据量较小等一些情况下，Random Forest 能够表现得比 GBDT 更好。

对于本赛题需要预测未来 14 天 2000 家商店每日的客流量。我们通过基本的分析发现商店类别为：快餐、超市、便利店、休闲茶饮、小吃、休闲食品、烘焙糕点中餐、其他美食、火锅、烧烤、汤/粥/煲/砂锅/炖菜、网吧网咖、药店、本地购物、个人护理、美容美发。他们全都是需要外出购物的商店，而不是外卖服务。对于消费者来说，会更倾向外卖服务。因此，对于需要自己外出消费的服务，消费者更愿意在天气状况更好的情况下外出，比如说：约会、聚会等，所以天气特征对是否消费有影响。同时，每个人的行为都有一定的规律性在里面，比如：有的人喜欢晚上外出吃饭，而有的人更偏好中午；对于时间来说，是否是周末或者节假日也是影响整体消费的重要因素；是否是寒暑假影响着学生的消费，比如:放假了，学校附近的商店的客流量急剧降低，学生更倾向于在城市中心的商圈消费；是否是法定节假日，距离放假还有几天也影响着上班族的消费；对于商家来说，商家的历史客流量，商家的地理位置，商家的规模也影响着该店铺在未来呈现一个什么样的趋势，这一类属于商家的历史统计特征。

依据线下测试，我们对各个模型得到的预测结果进行加权相加；这种融合方式简单有效而泛化能力强，具有一定的提升。

3.3 其他优化

此外，我们还做了以下的优化：(1)采用了重采样技术，选取不同于前述 28 天训练集的训练集，用同样的方法预测，并将所有得到的结果加权相加；(2)我们发现 2015 年的 11 月 11 日“双十一购物狂欢节”的商家客流量比平常要高。参照 2015 年，我们将 2016 年的 11 月 11 日的所有客流量乘以 1.1。

4 实验结果

In this section, we evaluate our solutions for IJCAI-17 Competition in second stage. In our experiments, we split two weeks of the training set to generate an offline test set and the others to generate an offline training set for modeling training and parameter tuning. The results of individual models at the second stage are listed in Table 1,while the top 5 teams at the second stage are listed in Table2. In particular, Table 1 shows the effectiveness of ensemble of individual models, Table 2 shows the effectiveness of the methods obtained through our depth of business understanding.

Table1: The results of individual models at the leaderboard B

Model	Time series	Tree regression	Mixed
Score	0.0771	0.0786	0.0762
Rank	3	16	1

Table2: Top 5 teams at the leaderboard B.

Rank	Team Name	Score
1	CAT	0.076263
2	Life is full of rhythm	0.076817
3	---------Baseline--------	0.077137
4	Flamingo	0.077188
5	VanillaTwilight	0.077201

5 Conclusions

In this paper, we present our solutions for IJCAI-17 Competition :Passengers flow forecast for Koubei's merchants.我们基于商家历史数据进行了深入分析，采用了时间序列加权回归和树回归两种方法对商家进行预测，并取得了良好的验证效果。在第一阶段的 A 榜中，我们使用简单的时间序列加权回归方法获得了第 11 名，在第二阶段的 B 榜中，我们使用时间序列加权回归和树回归两种方法融合最终获得第一名。