

文章编号: 2095-2163(2020)10-0054-03

中图分类号: TP391

文献标志码: A

可解释性旅游需求量预测模型综述

王馨玉

(西安石油大学 计算机学院, 西安 710065)

摘要: 本文针对传统的旅游需求量预测模型, 建立一种可解释性模型。此外, 用 LIME 方法对该模型的可解释性进行研究, 分析对旅游需求量数据影响权重最大的因素, 为相关单位提供决策依据和其他旅游景区旅游需求量预测提供参考。

关键词: 旅游需求量; 可解释性模型; 可解释性; LIME

Overview of interpretable tourism demand forecasting models

WANG Xinyu

(School of Computer Science, Xi'an Shiyou University, Xi'an 710065, China)

[Abstract] According to the traditional tourism demand forecasting model, an interpretable model is established. In addition, LIME method is used to study the interpretability of the model, and analyze the factors that have the greatest influence on tourism demand data, so as to provide decision basis for relevant units and reference for the prediction of tourism demand of other tourist attractions.

[Key words] Tourism demand; Interpretability model; Interpretability; LIME

0 引言

随着社会的发展, 越来越多的人选择出门旅游。对于城市、景区、酒店而言, 需要依据淡旺季、客流量等因素调整规划和安排。因此, 准确预测旅游需求量显得至关重要。

目前有很多的旅游需求量预测模型, 但可解释性高的旅游需求量预测模型更值得信赖, 也更有可能被有关部门所采纳。对于旅游规划者而言, 探索出最影响旅游规划统筹的因素是非常重要的。通过对模型应用的解释, 找出对旅游量影响的最主要的因素, 从而制定出更好的旅游规划。对于模型开发者而言, 模型的可解释性有助于在特定的情况(如过拟合等)下对模型进行修改。因此, 本文针对旅游需求量预测模型进行了研究。

1 预测模型

1.1 时间序列模型

时间序列模型解释了一个关于其自身过去的变量和一个随机干扰项, 探索所涉及的时间序列的历史趋势和相关因素(如季节性), 并根据模型中确定的趋势和因素预测该序列的未来。

在使用时间序列预测技术的研究中, 超过 2/3 的研究使用了不同版本的 ARIMA 模型。根据时间序列的频率, 可以是简单的 ARIMA, 也可以是季节性的 ARIMA。由于季节性是旅游业的一个主要特征, 因此决策者对旅游需求的季节性变化非常感兴趣, 所以在近几年中, 季节性的 ARIMA 来越受欢迎。

关于 ARIMA 和 SARIMA 模型的预测性能, 实证研究提供了相互矛盾的证据。例如, ARIMA 模型在所有情况下都优于 AR 模型和 MA 模型。SARIMA 模型优于 ARIMA、ARMA 等其他 8 种时间序列方法, 而非季节性 ARIMA 模型的性能高于所有考虑的预测模型的平均值。

1.2 其他预测模型

与时间序列模型相比, 计量经济学方法的主要优势之一在于能够分析旅游需求(因变量)及其影响因素(解释变量)之间的因果关系。计量经济学分析“除了作为一种产生预测的工具之外, 还发挥了许多有益的作用; 例如, 这些模型巩固了现有的关于经济如何运作的经验和理论知识, 为渐进的研究战略提供了框架, 并有助于解释其自身的失败”。就旅游需求而言, 计量经济学分析从经济学家的角度解释旅游需求变化、证明政策建议以及评价现有旅游政策的有效性等方面具有实证意义。相反, 在企业 and 政府主要关注的情况下, 时间序列模型对于旅游需求和其他相关因素之间相互依赖的关系是没有帮助的。

粗糙集方法通过结合经典集理论对不精确、不确定或不完整的知识(数据)进行分类分析, 是一种决策规则归纳法, 用于对一组混合的数值变量和非数值变量之间的关系进行建模。它在旅游需求分析

作者简介: 王馨玉(1995-), 女, 硕士研究生, 主要研究方向: 自然语言处理。

收稿日期: 2020-05-19

中的应用分别用于旅游购物、餐饮和观光支出分析。与经典的回归模型不同,粗糙集方法非常关注人口特征等分类变量,并根据这些变量之间的模型关系,预测与每个人口类别相关的旅游需求水平(而不是精确值),因此被认为是从微观角度分析旅游需求的计量经济模型的一个有用的补充工具。

模糊时间序列法在分析具有有限观测值的短时间序列时具有优势。灰色理论同样关注模型的不确定性和信息不足。研究者将模糊时间序列和灰色理论应用于旅游需求预测,利用模糊时间序列、灰色模型和马尔科夫修正灰色模型等三种模型,仅利用 12 项年度数据,对香港、德国和美国赴台旅游人数进行建模和预测。模糊时间序列模型预测性能的一致性有待进一步研究。

遗传算法(GAs)是基于自然选择和遗传进化思想的自适应、启发式搜索算法,通常被认为是一种优化方法。大量研究表明,GAs 适合解释旅游需求构成的变化。支持向量机(SVM)是另一种可用于解决分类、非线性回归估计和预测问题的人工智能技术^[1]。

2 旅游需求量预测模型

2.1 建立预测模型

为了实现对时间维度和数据特征的可解释预测,本文开发了一个两阶段 CNN 架构。第一阶段是一个卷积层和一个 $k \times 1$ 维卷积滤波器,学习过滤器识别出的,出现在不同情况下的重要的行为模式;第二阶段是一个 1×1 维卷积滤波器,可用于最先进的网络,如应用于最初的模型中,将第一阶段生成的特征映射的数量减少到 1,即可以在第二阶段使用一维卷积。一维卷积使用的滤波器大小为 $k \times n$,其中 n 是特征数。使用一维过滤器可以提取所有特征中出现的重要行为模式。

通过这两个阶段,在网络模型中保持了多元时间序列的时间和空间动态,基于梯度的方法来生成显著性图,也称为属性图,以提取网络的注意力,可以认为它与对时间间隔和特征的预测最相关。

使用 grad-CAM 方法生成属性图。分别对产生 $f_{maps} = [f_2d, f_1d]$ 特征图的两个阶段的最后一层应用 grad-CAM,对于每个通用特征映射 A 上的每个激活单元 u ,获得一个与特定类输出 c 相关联的重要性权重 w^c ,计算输出分数 y^c 相对于 A 的梯度,对 A 进行全局平均,公式(1):

$$w^c = \frac{1}{z} \sum_u \frac{\delta y^c}{\delta A_u}, \quad (1)$$

其中, Z 为 a 的单位总数,在二维情况下,激活单位 u 为二维坐标 $\{i, j\}$ 。

使用 w^c 计算 c 类所有特征映射之间的加权组合,使用 ReLU 来删除负面影响,公式(2):

$$L_{1/2D}^c = \text{ReLU}(\sum_{f_{maps}} w^c A). \quad (2)$$

$L_{1/2D}^c$ 用于查找输入数据中对 c 类网络的决策起主要作用的区域, L_{2D}^c 将在不同的时间间隔突出特征的贡献,而 L_{1D}^c 将突出所有特征的共同贡献^[2]。

2.2 模型的可解释性

一些模型本身是具有可解释性的,如线性模型、决策树等。而很多机器学习模型是不具备可解释性的,对不可解释性模型的解释需要通过可解释性模型,即用可解释性模型去模拟拟合不可解释性模型。

将 x^1, x^2, \dots, x^N 输入到一个“黑盒模型”中,输出结果为 y^1, y^2, \dots, y^N 。再将 x^1, x^2, \dots, x^N 输入到将要训练的、具有解释性的模型中,训练此模型的目的是要它的输出 $\tilde{y}^1, \tilde{y}^2, \dots, \tilde{y}^N$ 尽可能地接近 y^1, y^2, \dots, y^N ,使两模型在行为上尽可能地拟合。这样的局部解释性模型的方法叫做 LIME^[3]。

模型拟合过程如图 1 所示。

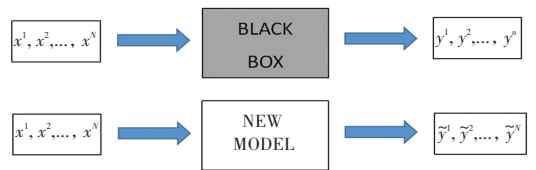


图 1 模型拟合

Fig. 1 Model fitting

无法用一个新模型去完全地模拟黑盒模型的行为,只能用新模型模仿一部分黑盒模型的行为。LIME 模型的原理如图 2 所示。

- (1) 找出一个待解释的数据点;
- (2) 在该数据点的附近进行数据点的取样;
- (3) 训练一个新模型使其可以模仿这个黑盒模型在样本点这个区域内的行为;

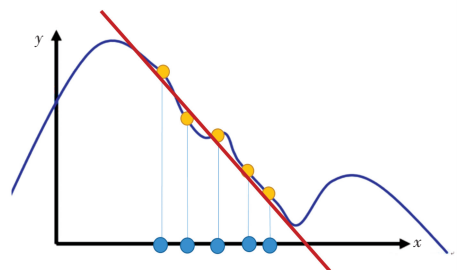


图 2 LIME 方法

Fig. 2 LIME Method