

多目标时间序列混合特征动态变权组合预测研究

董志学, 宫越, 胡勇, 甘孟壮

(北京汽车之家信息技术有限公司算法部, 北京 100190)

✉602932093@qq.com; gongyue.best@163.com; hyalone@126.com; 531609777@qq.com



摘要: 在通过LightGBM、Prophet、HotWinters等单一预测算法, 以及通过引入搜索指数等消费者行为数据作为预测变量等无法达到汽车品牌销量预测精度的情况下, 基于销量数据特征遴选出HotWinters、Prophet、LightGBM三个预测模型, 并自主构建了Musgrave方法, 以此四个算法构建了组合预测模型, 并结合熵值法作为权重动态变化的方法, 构建了“动态变权组合预测策略”。本策略使用“单地区多品牌维度、多品牌维度、多地区多品牌维度”三种方式进行六期预测并检验预测效果, 结果表明三种方式预测误差中位数分别为7.50%、6.11%、9.61%, 因此, 本策略能够满足对具有复杂多变特征的数据进行预测的需要。

关键词: 动态变权; 组合预测; Musgrave; 熵值法

中图分类号: TP311 **文献标识码:** A

Research on Dynamic Variable Weight Combination Prediction with Mixed Characteristics of Multi-objective Time Series

DONG Zhixue, GONG Yue, HU Yong, GAN Mengzhuang

(Algorithm Department, Beijing Autohome Information Technology Co., Ltd., Beijing 100190, China)

✉602932093@qq.com; gongyue.best@163.com; hyalone@126.com; 531609777@qq.com

Abstract: Prediction accuracy of automobile brand sales cannot be achieved through single prediction algorithms such as LightGBM, Prophet, HotWinters, and the introduction of consumer behavior data such as search indexes as predictors. Based on the characteristics of sales data, this paper proposes to select three predicting models, LightGBM, HotWinters and Prophet, and independently build Musgrave method. With these four algorithms, a combined forecasting model is constructed, and a "dynamic variable weight combination predicting strategy" is constructed by taking the entropy method as a method of dynamic weight change. This strategy uses three methods of "single-region multi-brand dimension, multi-brand dimension, and multi-regional and multi-brand dimensions" to conduct six-period predictions and test the prediction results. Results show that the median prediction errors of the three methods are 7.50%, 6.11%, 9.61%. Therefore, this strategy can meet the needs of predicting data with complex and changeable characteristics.

Keywords: dynamic variable weight; combined prediction; Musgrave; entropy method

1 引言(Introduction)

作为我国最重要的产业部门之一, 汽车产业无论横向还是纵向都跨越多个行业部门, 同时又与百姓生活关联紧密, 对汽车市场销量态势进行准确预测有助于从国家层面洞察各

个汽车品牌的经营态势, 也有利于从消费者方面了解不同价值等级消费者选择不同汽车品类过程中体现出来的消费结构和能力的变迁。然而, 汽车销量数据除了具有周期性(如车型换代)、季节性(如金九银十)、节假日(如春节)等特征, 还容易

受政府调控(如限行限号、车险费率改革)、车企不定期促销、优惠措施(如购置税减半)等行为的影响,同时又具有随机性特征,而现在新兴的直播购车等互联网形式的营销活动又进一步增加了预测的复杂性。

为提升预测精度,有关学者分别从应用并改进人工智能算法^[1]、引入搜索指数等消费者行为数据作为变量^[2]等方面进行预测研究。虽然此类方法能够在特点品牌^[3]、特定时段^[4]产生较高的预测精度,但更多实践表明,使用ARIMA、X-11、HotWinters等时间序列预测方法难以覆盖全部的数据特征,预测结果容易出现“有时准”但难以做到“实时准”,而当今流行的Prophet、LightGBM等机器学习算法具有难以将数据全部特征提取完全的弊端。所以,在单一静态预测无法满足需要的前提下,动态组合预测成为一种改善单项模型预测性能的有效策略。

本文基于相关学者的研究成果,在组合预测的基础上引入动态变权的策略,以广东地区30大品牌销量数据为基础,在数据特征识别、预测模型动态选择依据、各算法价值权重的自动调整方面进行预测研究。同时,为进一步验证动态变权组合预测策略的有效性,将预测对象从广东地区30大汽车品牌扩展到全国42个地区的30大汽车品牌以及全国30大汽车品牌进行扩展预测,以期达到能够真正指导实践的目的。

2 模型设计(Model design)

2.1 组合预测模型的原型

组合预测最早可追溯到1969年BATES和GRANGER^[5]提出的将单个预测组合成复合预测的原理,组合模型凭借可以有效去除滞后变量和非平稳性对预测的干扰^[6]、在单项预测结果存在有偏性的情况下通过组合能产生具有无偏性的预测结果^[7]等优势得到了研究学者的广泛应用,尤其是机器学习预测算法的引入,对通过识别长时间序列数据自身具有的规律性特征进而提升预测精度具有显著效果。组合预测模型如下:

假设选取 n 种预测模型 $f_{it}(i=1,2,\dots,n, t=1,2,\dots,m)$,每个模型在预测过程中均能够依据误差产生依据时间序列而变动的权重 w_{it} ,则变权组合预测模型 Y_t 可表示为:

$$Y_t = \sum_{i=1}^n f_{it} \times w_{it} + \mu_t \quad (1)$$

式中, μ_t 为随机噪音。

2.2 单项预测模型的选择

预测模型种类繁多且各有优劣势,从统计学层面来看,可分为回归问题和分类问题两大类。而在机器学习层面,

可分为线性模型、树型模型和神经网络模型,例如,谢如贤等^[8](1992)使用ARMA、季节变量回归模型、指数自回归模型组成变权组合模型对社会商品零售总额进行了预测,将平均误差(ME)降到了0.10;王永刚等^[9](2013)以灰色Verhulst模型、Brown指数平滑模型及非线性幂函数回归模型为单项模型,构建航空运输事故征候的最优变权组合预测模型,组合预测模型的平均绝对百分比误差(MAPE)为0.02323;朱周帆等^[10](2020)将SVM、RF、XGBoost与ARIMA模型相组合应用于汽车市场预测,平均绝对百分比误差(MAPE)为0.0297。通过对现有研究成果的归纳总结,本文选择的纳入组合模型的预测算法需要重点考虑能够处理季节性、趋势性、节假日等特征的模型,因此,初选模型包括ARIMA、HotWinters、Prophet、LightGBM、XGBoost、GM(1,1)。

2.3 权重设计与动态计算

预测效果的评估方法有绝对偏差、均方误差、均方根误差、平均绝对相对误差、平均绝对误差等,而权重的确定方法有等权重法、最小方差法、误差倒数法、优势矩阵法、权重收缩法等。在预测效果评价方面,本文选择绝对偏差法($\frac{|\text{预测值}-\text{真实值}|}{|\text{真实值}|}$)作为评价准则;在权重方法选择方面,考虑到预测数据收到季节性、节假日等因素的影响,各预测算法对预测对象的预测结果存在动态误差^[11],权重同样需要考虑实时、动态的特征,所以本文使用熵值法^[12]计算组合预测模型的参数 w_{it} 。熵值法的基本原理是:假设有 n 个评价对象(各预测算法),有 m 个评价指标(各品牌预测误差),视计算结果情况将其归一化后转变为标准化数据 $\{d_{ij}\}$,则其第 i 个指标的熵的计算公式为:

$$w_{ji} = \sum_{j=1}^m a_j \times f_{ij} \quad (i=1,2,\dots,n) \quad (2)$$

式中, a_j 为权数, f_{ij} 为第 j 项指标下第 i 年占该指标的比重。将式(2)代入式(1)即得到基于熵值法的组合预测模型。

3 实例研究(Case study)

3.1 数据特征认知

本文研究使用的销量数据以月为单位,数据对象是汽车30大品牌的新车销量数据,数据范围包括全国销售数据和42个主要地区的销售数据,选择2017年1月至2020年9月作为测试集,选择2020年10月至2021年3月数据作为验证集。因此,从数据的时间跨度来看,数据内涵既包括各品牌的销售行为等微观特征,又包括各地区经济发展、市场推广等宏观特征,这无疑都加大了预测难度。

(1) 平稳性检验

为进一步确定数据集的特征, 本文使用ADF进行平稳性检验。平稳性是选择预测模型的先决条件, 对于平稳型数据可使用AR、MA、ARMA等方式直接建模, 而对于非平稳型数据则需要通过取对数、差分等方法处理后使用ARIMA进行建模。对平稳性的检验分为依据时序图和自相关图进行判断的图检验法以及构造检验统计量的假设检验法, 前者主要依据研究者的经验以主观方式判断, 后者可通过单位根检验法进行客观判别, 因此, 本文采取ADF方法进行检验。从如表1所示的检测结果表可以看出, 30大品牌中不平稳对象占大多数, 仅有7个品牌(东风本田、长安福特、长安、凯迪拉克、荣威、吉利、领克)在数据期内平稳, 因此, 在模型选择中, ARMA、ARIMA应包含在选择对象之内。

表1 部分对象平稳性检测结果

Tab.1 Stability test results of some objects

品牌	ADF检验值	1%	5%	10%	P值	是否平稳
东风本田	-4.7126	-3.5685	-2.9214	-2.5987	0.0001	稳定
长安福特	-4.5237	-3.6104	-2.9391	-2.6081	0.0002	稳定
长安	-3.6607	-3.5685	-2.9214	-2.5987	0.0047	稳定
凯迪拉克	-5.2878	-3.5685	-2.9214	-2.5987	0.0000	稳定
荣威	-4.0623	-3.5685	-2.9214	-2.5987	0.0011	稳定
吉利	-5.7477	-3.5685	-2.9214	-2.5987	0.0000	稳定
领克	-4.1633	-3.6104	-2.9391	-2.6081	0.0008	稳定

(2) 季节性判定

长期趋势、循环波动、季节性变化、随机波动是数据序列波动的四大类因素, 因此, 对季节性因素进行识别是提升预测精度的一个有效手段。本文使用圆形分布法结合集中度进行是否存在季节性以及集中月份的识别, 该方法通过三角函数变换使得原始数据成为线性数据, 进而识别季节分布特征。圆形分布法先将12个月变换成360度, 1年按365.2563天计算, 1天相当于0.9856度(360/365.2563), 以1月1日零时为起点, 以每个月月中值作为组中值折算为度, 如: 1月份31天, 组中值为15.5天, 转换角度为15.2769度; 2月份28天, 组中值为45(31+28/2)天, 转换角度为44.3524度; 其他以此类推。通过公式计算圆形分布的 R 值、平均角 α 和角标准差 S 。由平均角 α 和角标准差 S 推算该对象的集中时间和增长高峰期, 由于圆形分布法对数据格式的要求必须是整年数据, 因此, 选取2017年、2018年、2019年、2020年作为研究数据, 并选取0.5作为标准差系数, 经计算可得广东地区

2017—2020年度30大品牌销售高峰的月度, 如表2所示。

表2 部分品牌销售高峰的月度分布示例

Tab.2 Example of monthly distribution of peak sales of some brands

品牌	2017年			2018年			2019年			2020年		
	高峰前期	高峰后期	高峰后期	高峰前期	高峰后期	高峰后期	高峰前期	高峰后期	高峰后期	高峰前期	高峰后期	
广汽本田	4.13	7.01	9.18	17.10.19	17.12.30	3.11	4.26	7.01	9.05	19.10.29	1.01	3.05
广汽	4.19	6.30	9.10	4.10	6.30	9.19	4.27	6.30	9.02	4.30	6.29	8.27
吉利	4.26	6.30	9.03	4.22	7.01	9.09	4.24	6.30	9.05	4.26	6.28	8.30
奔驰	4.17	7.01	9.13	4.12	7.01	9.20	4.23	7.01	9.08	4.23	6.29	9.03
红旗	16.11.02	16.12.31	3.01	17.11.05	17.12.30	2.23	18.10.14	18.12.31	3.18	19.11.02	1.01	2.29

通过以上计算结果可以看出, 各个品牌在研究期间内的季节性呈现出多样化的特征, 各品牌差异化程度较高, 例如广汽本田的销售高峰期分别是6月底7月初和12月底1月初, 红旗品牌的销售高峰期则是12月底1月初, 而奔驰品牌的销售高峰期则是6月底7月初。同时可以看出, 广汽、吉利品牌的销售高峰近4年均均在6月底7月初。由此可见, 各个品牌的季节性规律并不统一, 基于此种数据特征, 特将HotWinters、Prophet、LightGBM列入选择范围。

3.2 单项模型选择

通过对每个算法的预测进行实验, 根据预测误差进一步淘汰了中位数误差较大的ARIMA、GM(1,1)、XGBoost模型, 最终选取了HotWinters、Prophet、LightGBM三个模型, 考虑到能够满足数据特征的预测算法相对较少, 本文结合数据特征, 基于Musgrave和Henderson系数单独设计了算法进行预测(下文简称Musgrave方法)。

针对本文研究的数据特征, 本文使用Musgrave非对称滤子和Henderson滤子思想构建基于数据对称分布特征的预测^[13]。在X-11中, 使用Henderson移动平均可以从季节调整后序列中提取趋势这一成分, 在应用 $2p+1$ 项中心化移动平均时, 由于结构限制, 得不到序列最初 p 项和最后 p 项的平滑估计值, 这容易给预测带来困扰, 因为通常序列中最重要的点就是最末的那个点, 因此在实际应用中需要采用非中心化移动平均来估计这些值。Musgrave的思路就是构建这样的非对称移动平均, 能使得未来对估计值做出的修正达到最小。其基本原理是:

令 $x_1, x_2, \dots, x_t$ 为原始数据序列, $x_{t+1}, x_{t+2}, \dots, x_{t+n}$ 为待预测数据序列, 可使用X-12算法进行数据分解:

(1) 趋势序列为: $U_1, U_2, \dots, U_t$, 进一步扩展为