计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2019,28(5):1−9 [doi: 10.15888/j.cnki.csa.006905]

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

基于交叉验证网格寻优支持向量机的产品销售预测①

张文雅[1], 范雨强[1], 韩　华[1], 张　斌[2], 崔晓钰[1]

[1](上海理工大学 能源与动力工程学院, 上海 200093)
[2](上海交通大学 机械与动力工程学院, 上海 200240)
通讯作者: 崔晓钰, E-mail: xiaoyu_cui@usst.edu.cn

摘　要: 综合考虑影响汽车销售的多种因素, 运用交叉验证网格搜索优化支持向量机的惩罚系数和核函数参数, 建立了适合汽车销售的预测模型. 仿真实验结果表明, 改进支持向量机优化汽车销售预测模型的预测效果比某公司当前采用的模型更佳, 该模型具有较高的预测精度和较大的可信度, 可为企业决策层提供较为准确的销售预测参考.

关键词: 支持向量机; 销售预测; 汽车销售; 网格搜索; 交叉验证

引用格式: 张文雅,范雨强,韩华,张斌,崔晓钰.基于交叉验证网格寻优支持向量机的产品销售预测.计算机系统应用,2019,28(5):1–9. http://www.c-s-a.org.cn/1003-3254/6905.html

Product Sale Forecast Based on Support Vector Machine Optimized by Cross Validation and Grid Search

ZHANG Wen-Ya[1], FAN Yu-Qiang[1], HAN Hua[1], ZHANG Bin[2], CUI Xiao-Yu[1]

[1](School of Energy and Power Engineering, University of Shanghai for Science & Technology, Shanghai 200093, China)
[2](School of Mechanical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China)

Abstract: Considering various factors affecting automobile sales, the penalty coefficient and kernel function parameters of support vector machine are optimized by cross validation and grid search, and a prediction model suitable for automobile sales is established. The simulation results show that the forecasting effect of the improved support vector machine optimized automobile sales forecasting model is better than that of the current model adopted by a company. The model has higher forecasting accuracy and greater credibility, and can provide more accurate sales forecasting reference for enterprise decision-making level.

Key words: Support Vector Machine (SVM); sales forecast; car sales; grid search; cross validation

21 世纪的飞速发展, 人们生活水平得到了极大提高, 越来越多的家庭购买汽车作为代步工具, 我国汽车市场已进入品牌营销时代, 市场竞争也从传统的产品和价格竞争转移到品牌和渠道的竞争[1]. 汽车制造企业若能在生产、制造、销售等环节实现定量化预测, 为其决策提供必要依据, 则可在满足客户个性化需求的同时, 使其在日益激烈的市场竞争中占得先机.

前人已经进行了一些销售预测的尝试. 2011 年, Yu Y 等[2]提出了一种最大程度学习机制对人工神经网络进行了优化, 并通过对某品牌销量数据的学习训练, 在给定时间内精准的预测了服装产品的月销量. 张闯等[3]采用后向传播 (Back Propagation, BP) 神经网络预测方法, 通过新浪微博数据预测电影票房, 模型拟合效果较佳, 但因存在数据不完全和干扰数据的情况, 使预

① 基金项目: 国家自然科学基金 (51506125)
Foundation item: National Natural Science Foundation of China (51506125)
收稿时间: 2018-11-28; 修改时间: 2018-12-18; 采用时间: 2018-12-28; csa 在线出版时间: 2019-05-01

测精度不够. 严洪森等[4,5]分别采用了混沌 v-支持向量机和扩展的径向基函数核支持向量机建立了产品销售预测模型,在预测精度上具有一定的优势, 但也增加了模型的复杂度、需要优化的参数个数和最优参数组合的获取难度, 使模型难以推广. 本文拟采用支持向量机这种先进的机器学习方法, 在尽可能不增加参数及少量增加模型复杂度的情况下, 对模型进行优化,以期实现较为精准的汽车产品的销售预测.

支持向量机 (Support Vector Machine, SVM) 是近几年来发展起来的基于统计学习的机器学习方法[6]. 它以统计学习理论为基础, 直接从小样本出发, 放弃了传统的经验风险最小化 (Empirical Risk Minimization, ERM) 准则, 而采用结构风险最小化 (Structural Risk Minimization, SRM) 准则, 在最小化样本误差的同时, 考虑模型的结构因素, 从根本上提高了泛化能力. 支持向量机解决小样本、非线性及高维模式识别问题中表现出许多特有的优势, 它既能够有限的训练样本得到小的误差, 又能够保证对独立的测试集仍保持小的误差, 而且支持向量机算法是一个凸优化问题. 因此, 局部最优解一定是全局最优解. 在支持向量机的具体应用中, 惩罚系数 C 和核函数参数 g 的选取对预测性能具有关键性的影响[7]. 目前, 支持向量机的参数选择方法主要有: 网格搜索法、遗传算法和混沌优化等, 其思想主要是在初始化范围内进行寻优以获得模型最佳效果时的参数[8]. 王宁等[9]在训练过程中采用网格搜索法对支持向量机回归模型参数进行优化, 提出基于支持向量机回归组合模型的中长期降温负荷预测方法, 成功的把预测值与真实值的误差控制在 5% 以下, 且该模型得到了实际应用. Gao 和 Hou[10]为了提高 SVM 预测的精度和减少计算负荷, 采用了网格搜索 (GS) 算法优化 SVM 参数, 进而预测田纳西伊斯曼 (TE) 过程的状态, 发现 GS 方法比产生类似分类精度的遗传算法 (GA) 和粒子群优化算法 (PSO) 效率更高. Gencoglu MT 等[11]将混沌理论与 SVM 结合, 通过重构相空间的饱和嵌入维数确定 SVM 最佳输入变量的选取, 以混沌序列的最大 Lyapunov 指数确定 SVM 预测模型的最大有效预测步数, 但所处理的时间序列必须具有混沌性. 本文所处理的是小样本汽车销售数据, 时间序列的混沌性并不显著, 采用基本的 SVM 并采用 GS 算法进行参数优化进行销售预测是可行的.

为了增加模型的鲁棒性, 有效地避免过学习以及欠学习状态的发生, 使得到的结果更加可靠, 所以在优化过程中结合了 K-fold 交叉验证[12], 降低了支持向量机参数选择随机所带来的误差, 提高了模型的推广能力. 本文提出了基于交叉验证网格寻优的支持向量机方法,分别建立了采用每 3 个月、6 个月、9 个月、12 个月、18 个月和 24 个月的汽车销售额数据预测下一个月销售额的预测模型, 对预测结果进行详细的比较分析, 以期找到最佳的预测模型, 为汽车制造商及销售商提供可信度更高的销售预测数据, 作为决策参考.

1 基于网格搜索与交叉验证的 SVM 回归模型

1.1 SVM 回归基本理论

支持向量机 (Support Vector Machine, SVM) 由 Cortes 和 Vapnik 等于 1995 年提出, 此后, Vapnik 又提出引入 ε 不敏感损失函数[6]的 ε-SVR 算法,将支持向量机应用于回归领域. ε-SVR 通过事先确定 ε 来控制算法大致希望达到的精度. ε 不敏感损失函数的用途在于能够用稀疏数据点来表达如下要找的回归函数.

设样本向量为$\{(x_0,y_0),(x_1,y_1),\cdots(x_k,k_k)\}$ $(x_i\in R^n, y_i\in R, i=1,2,\cdots,k)$, 其中 k 为样本个数. 支持向量机回归的基本思想是通过一个非线性映射 Φ, 将数据x_i映射到高维空间 F 中, 并在这个高维空间中构造最优线性回归函数:

$$f(x)=\omega\phi(x)+b \tag{1}$$

式中, ω和$\phi(x)$为 m 维向量, b 为偏置量. 支持向量机采用结构风险最小化原则 (SRM)[13]确定参数ω和 b 的值, 即:

$$\min R_{\text{reg}}=\frac{1}{2}\|\omega\|^2+CR_{\text{emp}} \tag{2}$$

式中, R_{reg}为正则化风险; $\|\omega\|^2$为控制模型的复杂度; C 为惩罚系数, 用来调节模型复杂度和训练误差, C 越大, 对数据的拟合程度越高, 但过大时会使机器学习复杂度较大, 易造成过学习; $R_{\text{emp}}=\frac{1}{k}\sum_{i=1}^{k}L_g[x_i,y_i-f(x_i)]$为误差控制函数, 通常采用 ε-不敏感函数来度量, 定义如下:

$$L_g=\begin{cases}|y-f(x)|-\varepsilon, & |y-f(x)|\geq\varepsilon\\ 0, & |y-f(x)|\leq\varepsilon\end{cases} \tag{3}$$

根据结构风险最小化原则, 考虑在数据集上获得的回归模型的复杂度, 持向量机回归本质上就是求解一个优化问题[11]:

$$\min \ \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{k}\left(\xi_i + \xi_i^*\right)$$
$$\text{s.t.}\begin{cases} [\omega\phi(x_i)] + b - y_i \le \varepsilon + \xi_i \\ y_i - [\omega\phi(x_i)] - b \le \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0 \end{cases} \tag{4}$$

式中, ξ_i和ξ_i^*为松弛变量, 此问题称为支持向量机的原始问题. 由于ω维数很大, 为了便于求解, 根据强对偶定理引入 Lagrange 乘子α_i和α_i^*, 建立 Lagrange 函数, 将这一优化问题转化到对偶空间中得到原始问题的对偶问题[14]:

$$\min \frac{1}{2}\sum_{i,j=1}^{k}\left(\alpha_i^* - \alpha_i\right)\left(\alpha_j^* - \alpha_j\right)K\left(x_i, x_j\right) + \varepsilon\sum_{i=1}^{k}\left(\alpha_i^* - \alpha_i\right) - \sum_{i=1}^{k} y_i\left(\alpha_i^* - \alpha_i\right)$$
$$\text{s.t.}\begin{cases} \sum_{i=1}^{k}\left(\alpha_i^* - \alpha_i\right) = 0 \\ 0 \le \alpha_i^*\alpha_i \le C(i = 1, 2, \cdots, k) \end{cases} \tag{5}$$

式中, $K\left(x_i, x_j\right)$为核函数, 可将原问题通过非线性变换, 映射为某个高维特征空间上的线性问题, 进行求解. 本文采用的汽车销售数据属于非线性数据, 故需采用核函数. α_i^*, α_i是对偶问题的解, 由此可得回归函数为:

$$f(x) = \sum_{i=1}^{k}\left(\alpha_i^* - \alpha_i\right)K\left(x_i, x_j\right) + b \tag{6}$$

1.2 核函数的选择

在高维特征空间中, 线性问题中内积运算可以用核函数来代替, 即

$$K\left(x_i, x_j\right) = \phi(x_i)\cdot\phi\left(x_j\right) \tag{7}$$

支持向量机不同的内积核函数将形成不同的算法,回归支持向量机常用的核函数有三种, 即多项式核函数、径向基核函数和 Sigmoid 核函数.对于多项式核函数, 当特征空间位数很高时, 其计算量将大大增加,甚至对某些情况无法得到正确的结果, 而径向基函数不存在这个问题. 另外, 径向基函数的选取是隐含的, 每个支持向量机产生一个以其为中心的局部径向基函数, 使用结构风险最小化原则, 能找到全局的径向基函数参数[15]. 对某些参数, RBF 与 Sigmoid 核函数具有相似的性能, 在一般情况下, 首先考虑的是 RBF[16]. 因此本文选取径向基核函数 (RBF) 建立预测模型[17–19], 即:

$$K(x, x_i) = \exp\left(-\frac{\|x - x_i\|}{2\sigma^2}\right) = \exp\left(-g\|x - x_i\|\right) \tag{8}$$

式中, σ 为径向基函数的宽度, σ 越小, 径向基函数的宽度越小, 越有选择性. $g = \frac{1}{2\sigma^2}$是径向基核参数, g 越大, 径向基函数越有选择性. 把径向基核函数代入 (6) 式得到完整的回归函数, 即:

$$f(x) = \sum_{i=1}^{k}\left(\alpha_i^* - \alpha_i\right)exp\left(-g\|x - x_i\|\right) + b \tag{9}$$

研究发现, 支持向量机计算过程中涉及到的两个参数: 惩罚系数 C 和核函数参数 g, 是影响支持向量机回归模型的主要因素.

1.3 基于网格搜索的 SVM 参数优化

在本案例中支持向量机 (SVM) 的核函数采用的是径向基核函数 (RBF), 径向基函数中的参数 g 和惩罚系数 C 的选择对汽车销售量的预测值有着很大的影响, 为了寻找最佳的参数 C 和 g, 本文根据前文和样本特性选择的是网格搜索法 (grid search). 网格搜索法首先是要把所有的可能的参数值做统计然后进行分组,分组的依据是由步距决定的网络. 然后对逐个网络中可能的最优参数值进行计算, 并验证观察结果是否最优,即找到的最优参数[19].

交叉验证 (cross validation) 是一种消除取样随机性所带来的训练偏差的统计学方法. 常用的交叉验证方法有重复随机抽样法、K-fold 交叉验证法、留一法等.基于支持向量机原始预测模型, C 和 g 的初始值均为 1,预测精度较低, 当使用交叉验证网格寻优的方法以后, C 和 g 值在设定的范围内进行寻优, 对每个预测模型均进行 MSE 值得比较, 这样可以建立最佳的预测模型,保证 MSE 值为最小, 避免原始模型预测精度低的问题.所以本文将交叉验证与网格搜索相结合, 以 MSE 最小化为参数优选的目标, 提高了参数优选的效率和准确性, 同时极大规避了样本的抽样随机性对模型性能的影响[20,21]. 网格搜索法参数优化的基本流程如下, 流程图见图 1 所示.

(1) 先初始化网格搜索中惩罚系数 C 和核函数参数 g 的搜索范围和搜索步长, 本文在寻优时分为粗略选择和精细选择.

(2) 进行粗略选择, 粗略选择时 C 的取值范围是 $[2^{-8}, 2^8]$, 当输入变量 (C 的取值范围、g 的取值范围、交叉验证的折数等) 个数小于 8, 则指数的步长为 0.8, g 的取值范围是$[2^{-8}, 2^8]$, 指数的步长为 0.8. 得到粗略选择的 C 和 g.

(3) 根据粗略选择结果再进行精确选择, C 和 g 的取值范围是粗略选择后确定的范围, 指数步长为 0.5.

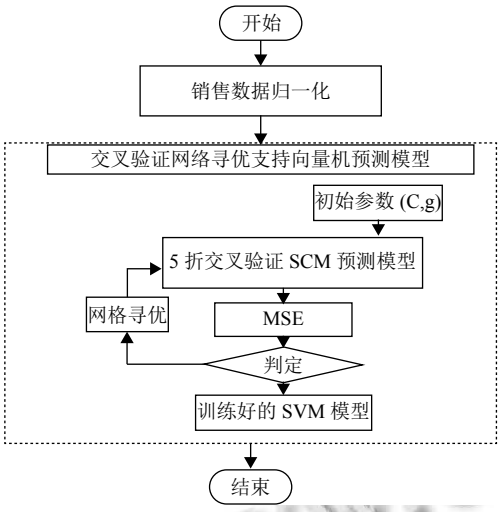


图 1 基于交叉验证网格寻优的支持向量机预测模型

图 2 给出 Model-12m 模型 (每 12 个月预测下一个月) 精细选择参数后的等高线图和 3D 视图. 以 $\log_2 C$ 为横坐标, $\log_2 g$ 为纵坐标, MSE(下文公式求得) 为 Z 轴 (见下文公式). 如图 1 所示, 图中红点就是精细选择时找到的最佳参数点, 此时 $\log_2 C = 1/2$, $\log_2 g = 1/2$, 所以得到最佳参数 $C = \sqrt{2}$, $g = \sqrt{2}$.

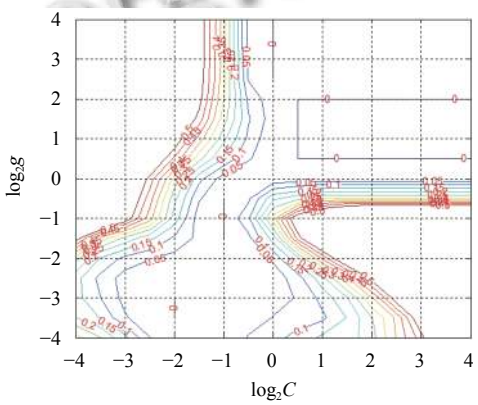
2 当前的预测模型

进行销售预测时, 由于预测对象可能是产品销售数量、销售价格、销售金额, 预测的地域、客户、时间长度等的不同, 可以有不同的预测方法分类. 本文案例主要研究某汽车公司的销售额预测, 适用于这种场合的常用预测方法分为两大类: 定性预测方法和定量预测方法, 见图 3.

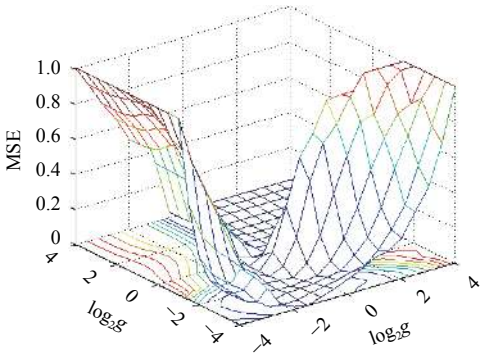
第一类是 (qualitative) 预测方法. 依据人们对过去及现在的经验、判断和直觉作预测. 一般常用的定性销售预测方法有四种: 高级经理意见法、销售人员意见法、购买者期望法和德尔菲法.

第二类是定量预测方法. 包括时间序列分析法和 (causal) 预测方法. 本文提出的基于支持向量机优化模型的预测方法属于时间序列分析方法.

图 4 亦可见, 销售预测及其预测方法的选择受到诸多内部因素及外部因素的交互影响, 相当复杂. 某汽车公司当前的销售预测模型是建立在汽车行业预测的基础上, 是定性与定量相结合的预测模型, 而汽车行业的经济预测是建立在宏观经济预测的基础上的.某汽车公司通过对中国车市的发展阶段、增长潜力、国家补助政策、各地区的引导政策、消费者收入与开支、出口、利率、企业车型开发进度、企业投资和与整车有关的其他重要因素和事件进行预测, 得到公司未来某一阶段的产出/销售预测. 定性预测模型预测流程图如图 3 所示, 可见该方法包含一些简单的定量分析, 但更多地依赖于上级决策者的决定及销售人员的建议, 偏重于定性方法.



(a) SVR 参数优化等高线图 [GridSearchMethod]
Best C=1.4142 g=1.4142 CVmse=0.14655



(b) SVR 参数优化 3D 视图 [GridSearchMethod]
Best C=1.4142 g=1.4142 CVmse=0.14655

图 2 Model-12m 模型参数精细选择等高线图和 3D 视图

按照该流程, 以 2009 年的实际销售额预测 2010 年到 2015 年共计 72 个月的销售额, 示于图 5 中. 图中, Sales 为实际销售额, PredSales 为某汽车公司当前预测模型预测的销售额. L 汽车公司的预测销售额与实际销售额的绝对误差相差幅度较大, 绝对误差最大值如图中红圈所示, 是 2013 年 12 月 (约 14 万元), 最小值是 2014 年 11 月 (约 8.4 万元), 最大值是最小值的 160 倍.

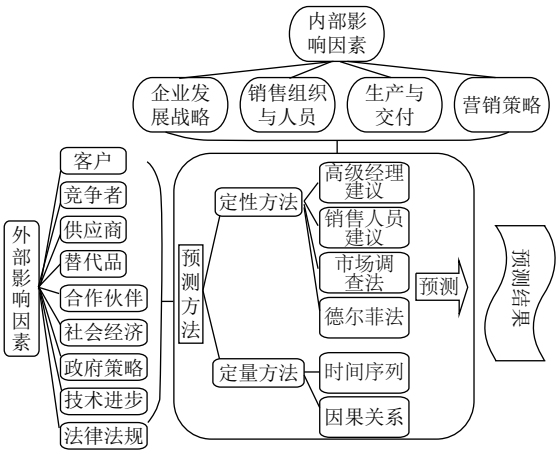


图 3 销售预测内在的逻辑关系图

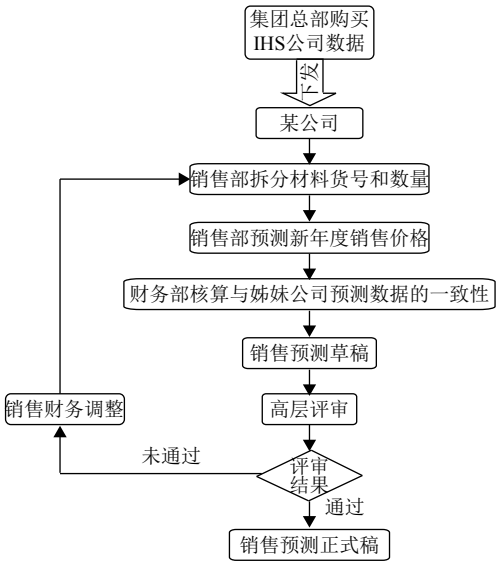


图 4 某公司当前的销售预测方法流程图

将该预测的相对误差示于图 6 中. 相对误差最大值是 2015 年 8 月, 达 61.4%; 最小值是 2014 年 11 月, 为 0.27%, 前者是后者的 200 倍. 在 72 个月的总样本中, 有 15 个月的相对误差在 30% 以上, 占样本总数的 21%; 8 个月的相对误差在 20% 和 30% 之间, 占样本总数的 11%; 其余 49 个月的相对误差小于 20%, 占样本总数的 68%. 可见, 某公司目前所采用模型的预测并不理想. 下文将采用基于网格搜索交叉验证的 SVM 优化算法对某公司的销售额进行预测, 以期获得更加准确的销售预测模型, 为某公司的生产及销售决策提供更为可靠的参考与指导.

3 基于 SVM 优化模型的汽车销售预测与分析

采用基于网格搜索和交叉验证的 SVM 回归模型对某公司 2009 年到 2015 年共计 7 年 (84 个月) 的销售额进行预测, 选取 2010 年到 2015 年共计 72 个月的预测数据与实际销售额进行比较分析. 多次尝试的结果表明, 采用一个季度 (3 个月) 或多个季度的销售数据进行预测较其他不以季度为周期的预测模型预测效果更佳. 假定每 3 个月数据预测下一个月销售额的模型为 Model-3m,其他各模型名称见表 1.

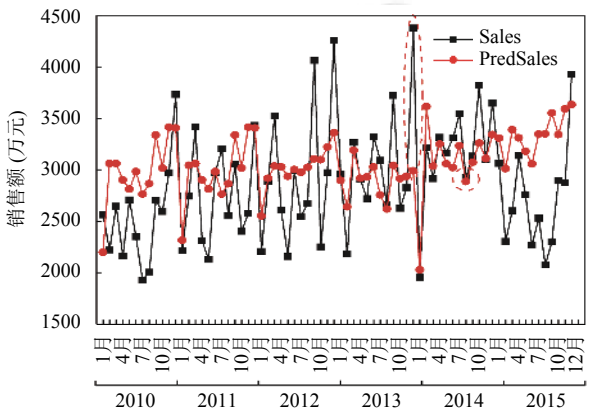


图 5 某公司预测销售额和实际销售额比较图

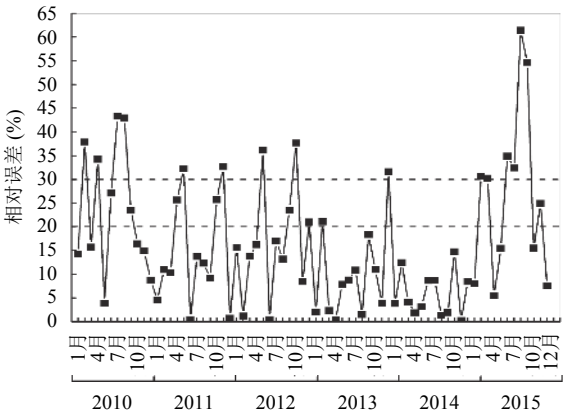


图 6 某公司预测销售额和实际销售额的相对误差值

为了说明所建预测模型的优劣, 将预测模型的预测值和真实值的均方误差 (Mean Squared Error, MSE)、绝对误差 (Absolute Error, AE) 和相对误差 (Relative Error, RE) 作为评价指标来评价模型, 其中均方误差主要评价预测模型的整体性能, 相对误差和绝对误差可用于评价预测模型的局部性能[22], 以季度为周期的预测模型的绝对误差相较别的预测模型更小, 对整体性能亦可作为参考.

表 1 各模型名称

模型名称	模型描述
Model-X	某公司当前预测模型
Model-3m	采用 3 个月数据预测
Model-6m	采用 6 个月数据预测
Model-9m	采用 9 个月数据预测
Model-12m	采用 12 个月数据预测
Model-18m	采用 18 个月数据预测
Model-24m	采用 24 个月数据预测

$$\begin{cases} \text{MSE} = \dfrac{1}{k}\sum_{i=1}^{k}(y_i - y_i')^2 \\ \text{AE} = \left|y_i - y_i'\right| \\ \text{RE} = \left|\dfrac{y_i - y_i'}{y_i}\right| \end{cases} \tag{10}$$

式中, y_i 为原始销售额, y_i'为预测销售额.

3.1 Model-3m 预测模型

参见图 1 交叉验证网格寻优的支持向量机预测模型流程图, 使用原始 C, g 建立原始预测模型 Model-3m-original, 此时 C=1, g=1. 模型的平均相对误差为 11.849%, 均方误差为 9.232×10^{-2}, 决定系数为 0.509 73. 其预测性能虽优于 Model-X, 但较下文所提的经过交叉验证与网格搜索的 Model-3m 预测性能仍显不足, 所以证明经过交叉验证与网格搜索的支持向量机预测模型得到改进, 提升了预测精度和可靠性.

经网格搜索与交叉验证寻优, 采用三个月数据预测下一个月销售额的 Model-3m 模型的最优 SVM 参数组合为 $C=2^{-3/2}$, g=2, 预测结果见图 7. 图中, 绝对误差最大值是 2015 年 12 月 (约 1,1.57 万元), 最小值是 2011 年 3 月 (7.8 万元), 最大值是最小值的 150 倍. 相较于某公司当前采用的 Model-X 模型, 实际销售额和预测销售额之间相差幅度有所减小. Model-3m 模型预测结果的相对误差示于图 8 中, 相对误差最大值为 38.02% (2015 年 8 月), 比 Model-X 的最大相对误差 61.4% 小 23.38%; 最小值为 0.23% (2011 年 3 月), 且 75% 的样本 (54 个月) 相对误差在 20% 以下, 比 Model-X 增加了 5 个月; 相对误差在 20% 和 30% 之间的有 14 个月, 占样本总数的 19.4%, 比 Model-X 增加了 6 个月; 相对误差 30% 的仅有 4 个月, 占样本总数的 5.6%, 比 Model-X 减少了 11 个月. 可见, Model-3m 模型对销售额的预测效果明显优于 Model-X 模型, 后文将对基于优化 SVM 的其他模型进行研究, 以期获得更优的预测效果.

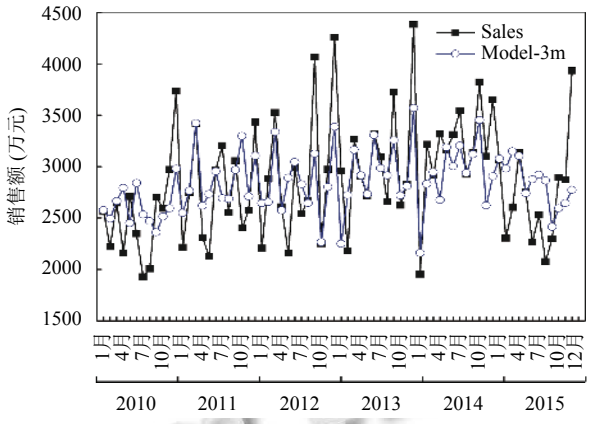


图 7 Model-3m 预测模型实际销售额和预测销售额对比图

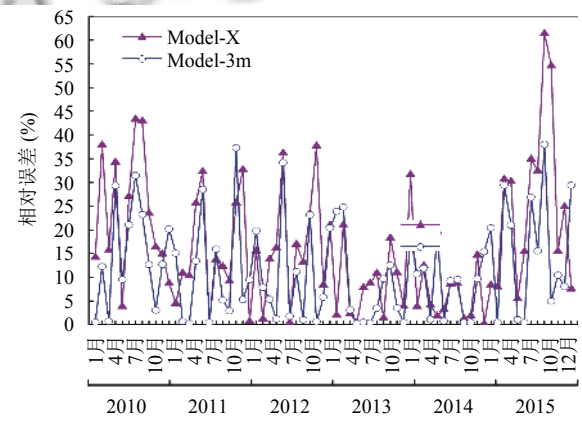


图 8 Model-X 模型和 Model-3m 模型相对误差对比图

3.2 Model-6m 预测模型

采用 6 个月数据预测下一个月销售额的 Model-6m 模型优化后的 SVM 参数组合为 C= 2, g=4. 该模型预测销售额与实际销售额的绝对误差较 Model-X 模型和 Model-3m 模型均有大幅下降, 绝对误差最大值约为 14 万元 (2010 年 2 月), 最小值是约为 4.6 万元 (2010 年 1 月), 前者仅为后者的 3 倍, 而非 200 倍 (Model-X) 或 150 倍 (Model-3m). Model-6m 的绝对误差主要集中在 13 万元到 14 万元之间, 幅度比较稳定.

由 Model-6m 预测模型的相对误差图 9 可见, 该模型的相对误差基本以 0.45% 为中心上下浮动, 落在 0.15% 到 0.75% 之间, 最大相对误差是 2014 年 1 月的 0.724%, 最小的相对误差值是 2010 年 1 月的 0.18%, 二者较为接近.

将 Model-6m 与 Model-3m 的相对误差示于图 10 中进行比较分析. 与 Model-3m 模型相比, Model-6m 模型的相对误差紧贴着横轴, 总体上明显较小, 除 3 个月

的相对误差略有上升外 (2011 年 3 月, 2013 年 4 月, 2014 年 8 月), 其余月份的相对误差均大幅下降, 降幅最大的是 2015 年 8 月, 达 37.39%; 降幅超过 10% 的有 33 个月, 占样本总数的 45%. 表明, Model-6m 模型的预测效果较 Model-3m 模型有显著提高.

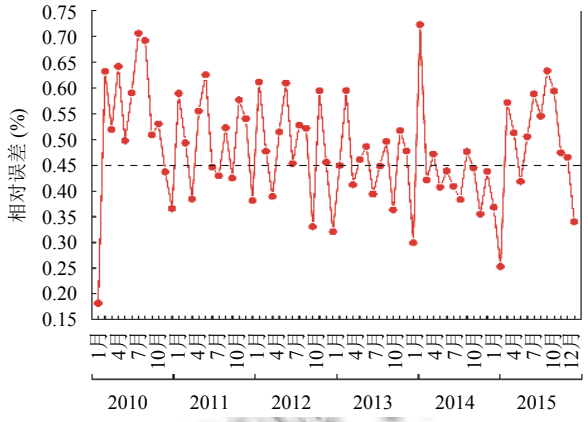


图 9 Model-6m 预测模型的相对误差图

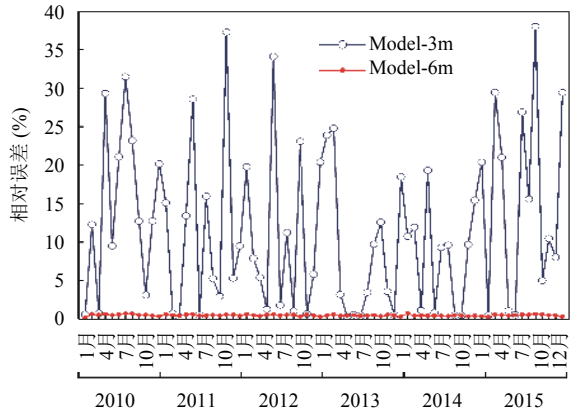


图 10 Model-3m 模型和 Model-6m 模型相对误差对比图

3.3 Model-9m 和 Model-12m 预测模型

采用 9 个月数据预测下一个月销售额的 Model-9m 模型, 其优化后的 SVM 的最优参数组合为 $C=\sqrt{2}$, $g=2$. 以一年 (12 个月) 数据作为预测基准的 Model-12m 模型, 优化后的 SVM 的参数组合为 $C=\sqrt{2}$, $g=\sqrt{2}$.

Model-9m 模型和 Model-12m 模型预测销售额与实际销售额绝对误差相差较小. Model-9m 模型绝对误差最大值是 2012 年 10 月的 14 万元, 最小值是 2011 年 6 月的 12 万元; Model-12m 模型绝对误差最大值是 2010 年 10 月的 12.8 万元, 最小值是 2013 年 4 月的 7.45 万元. 将 Model-9m 模型、Model-12m 模型与前述最佳模型 Model-6m 的相对误差共同示于图 11 中. 可见, Model-6m 的相对误差在 0.15%~0.75% 之间, Model-9m 模型的相对误差在 0.25%~0.75% 之间, Model-12m 模型的相对误差在 0.25%~0.65% 之间. 三个模型相对误差低于 0.45% 的月份数分别为 29 个月, 29 个月和 42 个月, 分别占样本总数的 40%, 40% 和 58%. 相对于 Model-9m 模型, Model-12m 模型每一个月的相对误差均有所下降; 相对于 Model-6m 模型, Model-12m 模型除了 2010 年 1 月、2015 年 1 月相对误差分别增大了 0.286% 和 0.152%, 其余的月份均有不同程度下降. 表明, 以 6 个月、9 个月、12 个月的数据进行销售额预测, 效果均较佳, 其中 Model-12m 模型的整体性能更好. 数据有限时, Model-6m 模型亦可实现较为准确的销售额预测.

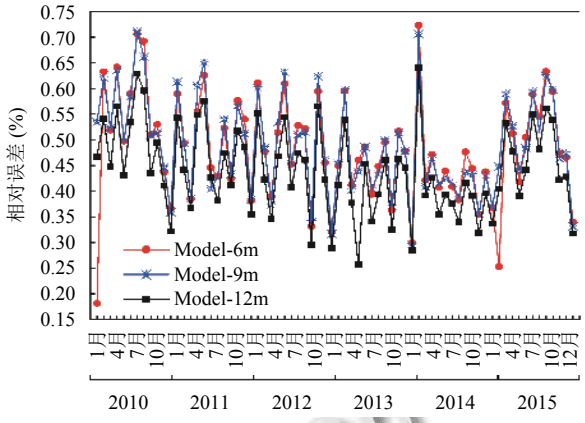


图 11 Model-6m,Model-9m 和 Model-12m 三种模型的相对误差对比图

总体上, 图 11 中的 3 种模型在中间段预测能力较好, 尤其是 2013 年和 2014 年, 相对误差多在 0.45% 附近, 而 2011 年和 2015 年的预测能力相对较差. 2011 年正处于全球金融危机, 大部分消费者有危机感, 更愿意把钱存在银行, 对投资买车可能处于观望状态, 不同月份销售的汽车数量波动较大, 因而影响预测.而 2013 年和 2014 年, 金融危机缓和, 国家大力提倡人民消费, 每个月的销售量都较为平稳, 有利于预测. 2015 年国家出台限制公车购买量的政策, 一定程度上影响了每个月的销售额, 增加了预测难度.

3.4 Model-18m 预测模型

Model-18m 模型优化后的 SVM 参数组合为 $C=1$, $g=0.707107$. 该模型在 2013 年 12 月的相对误差达到了 3.10%, 比 Model-6m、Model-9m 和 Model-12m 的最大相对误差均大三倍以上, 此处不加详细讨论.

3.5 Model-24m 预测模型

采用 24 个月数据预测下一个月销售额的 Model-24m 模型, 其优化后的 SVM 参数组合为 C=1, g=0.5. Model-24m 模型的相对误差在 0.25%~0.65% 之间, 相对于 Model-12m 模型, 有 38 个月的相对误差减小, 22 个月的相对误差增大. 与 Model-6m 和 Model-9m 模型一样, Model-24m 模型在中间月份,即 2013 年和 2014 年的预测性能较好, 而在起始和末端月份的预测性能较差.

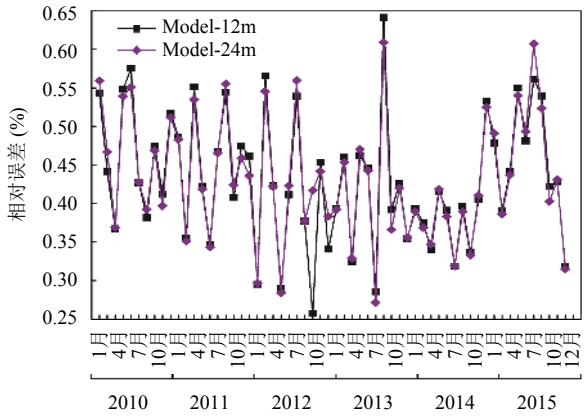


图 12 Model-12m 和 Model-24m 相对误差对比图

综合以上可知, 以年 (12 个月) 的数据为周期的模型预测效果最佳, 因 Model-24m 模型所需数据量较为庞大, 性能提高却并不明显, Model-12m 模型即可进行较为准确的销售预测.

3.6 各模型评价指标的比较与分析

决定系数 R^2 是预测值拟合程度的指标[23], R^2 的数值大小可以反映实际销售额与预测销售额之间的拟合程度, R^2 越大, 数据拟合程度越高, 预测销售额的可靠性就越高.

表 2 列出了各模型的决定系数, 平方相对误差和均方误差三项评价指标. Model-24m 模型的决定系数最大, 平均相对误差最小; Model-12m 模型均方误差最小, 决定系数和平均相对误差与 Model-24m 相近. 在基于优化 SVM 的模型中, 采用三个月数据进行预测的 Model-3m 模型, 决定系数是某公司当前采用的 Model-X 模型的近 3 倍, 平均相对误差小 4.67%, 而 Model-24m 和 Model-12m 的平均相对误差更是 Model-X 的 1/25(即 4%), 充分说明基于交叉验证网格搜索的 SVM 预测模型整体性能非常好, 最佳模型是 Model-12m 和 Model-24m, 当数据有限时, 亦可以采用 Model-6m 模型进行预测. 为了验证经过交叉验证和网格搜索后的支持向量机预测模型的优越性, 对最佳模型 Model-12m 随机选取销售数据进行测试, 并计算其各项评价指标, 列于表 2 中: 平均相对误差为 0.446%, 均方误差 1.012×10^{-4}, 决定系数为 0.99970, 各项指标与 Model-12m 模型性能相差不大, 可见经过交叉验证和网格搜索后的支持向量机预测模型预测精度高, 鲁棒性强.

表 2 各模型评价指标

模型名称	决定系数 (R^2)	平均相对误差 (%)	均方误差 (MSE)
Model-X	0.181 07	16.338	
Model-3m	0.525 42	11.673	9.232×10^{-2}
Model-6m	0.999 68	0.482	9.926×10^{-5}
Model-9m	0.999 63	0.489	1.010×10^{-4}
Model-12m	0.999 74	0.441	9.872×10^{-5}
Model-12m-test	0.999 70	0.446	1.012×10^{-4}
Model-18m	0.999 14	0.481	2.846×10^{-4}
Model-24m	0.999 75	0.432	1.013×10^{-4}

4 结论与展望

本文针对汽车销售预测问题的特点, 运用了交叉验证和网格搜索方法优化了支持向量机的惩罚系数和核函数参数的选择, 建立了改进支持向量机汽车销售预测模型, 提高了汽车销售的预测精度. 尽管预测效果可能受到国家政策、消费模式等的影响, 但本文提出的基于改进支持向量机优化的预测模型仍然可达到较小的预测误差, 预测数据可靠性高, 可给汽车企业在日常生产、销售管理中, 提供科学有效的预测方法,从而为决策者制定或调整相关计划提供可靠的理论依据, 具有一定现实意义及应用价值.

参考文献

1 贾鸣镝, 郑鑫, 叶明海, 等. 汽车经销商能力评价模型及其实证. 汽车工程, 2012, 34(1): 85–90. [doi: 10.3969/j.issn.1000-680X.2012.01.019]

2 Yu Y, Choi TM, Hui CL. An intelligent fast sales forecasting model for fashion products. Expert Systems with Applications, 2011, 38(6): 7373–7379. [doi: 10.1016/j.eswa.2010.12.089]

3 张闯, 姜杨, 吴铭, 等. 基于社会化媒体节点属性的信息预测. 北京邮电大学学报, 2012, 35(4): 24–27. [doi: 10.3969/j.issn.1007-5321.2012.04.006]

4 吴奇, 严洪森. 基于混沌 v-支持矢量机的产品销售预测模型.

机械工程学报, 2010, 46(7): 128–135.
5 徐歆, 严洪森. 基于扩展的径向基函数核支持向量机的产品销售预测模型. 计算机集成制造系统, 2013, 19(6): 1343–1350.
6 Vapnik VN. The Nature of Statistical Learning Theory. New York: Springer, 2000. 138–167.
7 史峰, 王小川, 郁磊, 等. MATLAB 神经网络 30 个案例分析. 北京: 北京航空航天大学出版社, 2010.
8 杨洪, 古世甫, 崔明东, 等. 基于遗传优化的最小二乘支持向量机风电场风速短期预测. 电力系统保护与控制, 2011, 39(11): 44–48, 61. [doi: 10.7667/j.issn.1674-3415.2011.11.008]
9 王宁, 谢敏, 邓佳梁, 等. 基于支持向量机回归组合模型的中长期降温负荷预测. 电力系统保护与控制, 2016, 44(3): 92–97.
10 Gao X, Hou J. An improved SVM integrated GS-PCA fault diagnosis approach of Tennessee Eastman process. Neurocomputing, 2016, 174: 906–911. [doi: 10.1016/j.neucom.2015.10.018]
11 Gencoglu MT, Uyar M. Prediction of flashover voltage of insulators using least squares support vector machines. Expert Systems with Applications, 2009, 36(7): 10789–10798. [doi: 10.1016/j.eswa.2009.02.021]
12 纪昌明, 周婷, 向腾飞, 等. 基于网格搜索和交叉验证的支持向量机在梯级水电系统隐随机调度中的应用. 电力自动化设备, 2014, 34(3): 125–131. [doi: 10.3969/j.issn.1006-6047.2014.03.021]
13 罗公亮. 从神经网络到支撑矢量机 (上、中、下). 冶金自动化, 2001, 25(5): 1–5, 2001, 25(6): 1–4, 2002, 26(1): 1–5.
14 王观玉, 郭勇. 支持向量机在电信客户流失预测中的应用研究. 计算机仿真, 2011, 28(4): 115–118, 312. [doi: 10.3969/j.issn.1006-9348.2011.04.028]
15 张金玉, 张炜. 装备智能故障诊断与预测. 北京: 国防工业出版社, 2013.
16 付元元, 任东. 支持向量机中核函数及其参数选择研究. 科技创新导报, 2010, (9): 6–7. [doi: 10.3969/j.issn.1674-098X.2010.09.004]
17 马向东, 卢占庆, 谭永彦, 等. 基于支持向量机的分类辨识方法及应用. 控制工程, 2016, 23(5): 768–772.
18 王霞, 王占岐, 金贵, 等. 基于核函数支持向量回归机的耕地面积预测. 农业工程学报, 2014, 30(4): 204–211. [doi: 10.3969/j.issn.1002-6819.2014.04.025]
19 Najafi G, Ghobadian B, Moosavian A, *et al*. SVM and ANFIS for prediction of performance and exhaust emissions of a SI engine with gasoline-ethanol blended fuels. Applied Thermal Engineering, 2016, 95: 186–203. [doi: 10.1016/j.applthermaleng.2015.11.009]
20 罗小燕, 陈慧明, 卢小江, 等. 基于网格搜索与交叉验证的 SVM 磨机负荷预测. 中国测试, 2017, 43(1): 132–135, 144.
21 郭李娜, 樊贵盛. 基于网格搜索和交叉验证支持向量机的地表土壤容重预测. 土壤通报, 2018, 49(3): 512–518.
22 亢生彩. 网格搜索法 SVM 参数优化在主扇风机故障诊断中的应用. 煤炭技术, 2015, 34(1): 295–297.
23 王琪, 孙玉坤, 倪福银, 等. 一种混合动力电动汽车电池荷电状态预测的新方法. 电工技术学报, 2016, 31(9): 189–196. [doi: 10.3969/j.issn.1000-6753.2016.09.024]