计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2020,29(4):18−23 [doi: 10.15888/j.cnki.csa.007254]

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

正则化和交叉验证在组合预测模型中的应用①

张欣怡, 袁宏俊

(安徽财经大学 统计与应用数学学院, 蚌埠 233030)
通讯作者: 张欣怡, E-mail: xybuty@126.com

摘　要: 组合预测模型的权重确定方式对于提高模型精度至关重要, 为研究正则化与交叉验证是否能改善组合预测模型的预测效果, 提出将正则化和交叉验证应用于基于最小二乘法的组合预测模型. 通过在组合模型的最优化求解中分别加入 L_1、L_2 范数正则化项, 并对数据集进行留一交叉验证后发现: L_1、L_2 范数正则化都对组合模型的预测精度具有改善效果, 且 L_1 范数正则化比 L_2 范数正则化对组合预测模型的改善效果更好, 并且参与组合预测的单项预测模型越多, 正则化的改善效果越好, 交叉验证对组合预测模型的改善效果则与给定实验数据量呈现正相关.

关键词: 组合预测模型; 正则化; 交叉验证; 最小二乘估计

引用格式: 张欣怡,袁宏俊.正则化和交叉验证在组合预测模型中的应用.计算机系统应用,2020,29(4):18–23. http://www.c-s-a.org.cn/1003-3254/7254.html

Application of Regularization and Cross-Validation in Combination Forecasting Model

ZHANG Xin-Yi, YUAN Hong-Jun

(Institute of Statistics and Applied Mathematics, Anhui University of Finance and Economics, Bengbu 233030, China)

Abstract: To determine the weight of the combined forecasting model is very important to improve the accuracy of the model. Applying the regularization and cross-validation to the combined forecasting model based on the least squares method is for studying whether the regularization and cross-validation can improve the prediction effect of the combined forecasting model. It is carried out by adding the L_1 and L_2 norm regularization terms to the optimization solution of the combined model and using leave-one-out-cross-validation in the data set. The result shows that both the L_1 and L_2 norm regularization can improve prediction accuracy of the combined model to a certain degree. Moreover, the L_1 norm regularization is better than the L_2 norm regularization for the combined forecasting model, and the more single forecasting models participating in the combined forecasting, the better the regularization improvement effect. In addition, there is a positive correlation between the cross-validation improvement effect and amount of experimental data given.

Key words: combined forecasting model; regularization; cross-validation; least square estimation

对于一个时间序列趋势分析问题, 一般不只限于一种时序预测模型可以使用, 特别在数据量较大的情况下, 数据量的增多意味着信息的增多, 只根据误差大小选择单一模型进行预测可能会导致部分有用信息的流失. 因此对于数据量较大的时间序列趋势分析, 进行组合预测很有必要. 组合预测模型是将多个候选预测模型的预测结果赋予合适的权值进行组合, 核心问题包括两个方面: 预测模型选择、权重确定方式.

① 基金项目: 国家社会科学基金 (13CTJ006); 安徽省教育厅高校人文社会科学重点研究项目 (SK2018A0431); 安徽财经大学研究生科研创新基金 (ACYC 2017236); 安徽财经大学重点科研基金 (ACKY1713ZDB)
Foundation item: National Social Science Foundation of China (13CTJ006); Major Project of Humanities and Social Science Research of Higher Education of Education Bureau, Anhui Province (SK2018A0431); Graduate Research and Innovation Fund of Anhui University of Finance and Economics (ACYC2017236); Major Scientific Research Fund of Anhui University of Finance and Economics (ACKY1713ZDB)
收稿时间: 2019-06-16; 修改时间: 2019-07-12; 采用时间: 2019-07-24; csa 在线出版时间: 2020-04-05

可供选择的单项预测模型包括灰色预测模型[1]、时间序列模型[2]以及机器学习模型, 前两者是传统的统计学模型, 机器/深度学习模型应用于组合预测模型则是比较新的探索[3–6].

组合预测模型的权重选择方式包括层次分析法[6]、熵权法[7,8]和最优加权法[9], 以上方法原理是基于预测数据本身的性质; 还有一类方法是基于单项模型预测误差来确定权重的, 如最小二乘法[10,11], 改进后的广义加权最小二乘法[12,13], 局部加权最小二乘法[14,15].

正则化和交叉验证是机器学习建立模型时使用的一种用于优化模型参数的方法, Hansen 和 Racine[16]提出了基于交叉验证的 Jackknife 模型平均 (JMA) 方法, 并证明了 JMA 方法可使模型权重选择得到优化; Zhao 等[17]在 JMA 方法的基础上开发了一个基于 K 折交叉验证的权重选择准则, 并证明了使用该准则可以产生更准确的预测. Sebastian Bayer[18]则在组合权重的选择中使用正则化降低单项预测间的多重共线性, 并证实这种方法可以通过稳定模型来提高模型预测能力.

通过阅读以上文献, 发现以往针对正则化和交叉验证的研究有两个特点: 一方面, 正则化和交叉验证对组合模型的影响都独立研究, 不能够发现这两者共同作用于组合模型的影响规律; 另一方面, 以上研究着重于比较正则化或交叉验证与其他权重选择方法的影响能力, 不注重讨论它们本身应用时的优化规律. 有鉴于此, 本文将以最小二乘加权组合预测模型为基础, 这种组合预测模型属于常规组合模型设计, 但其结构灵活, 使用广泛, 具有一定的改良和研究价值; 使用正则化和交叉验证对其进行优化, 正则化项选择 L_1 范数和 L_2 范数, 交叉验证选择 K 折交叉验证和留一交叉验证, 实验数据选择我国铁路客运量 2005~2018 年的月度数据, 由于该数据具有明显时间序列周期性, 组合预测的单项模型选择为适合分析周期性时间序列的 SARIMA 模型和 Holt-Winters 模型.

1 组合预测模型理论分析

1.1 单项模型理论

(1) SARIMA 建模原理和步骤

SARIMA 模型称为季节性差分自回归移动平均模型, 能够通过在 ARIMA 模型的基础上通过适当次数的季节性差分来消除序列的不稳定性, 适合处理具有趋势性和周期性的时间序列[19]. 也写作 SARIMA(p, d, q)(P, D, Q)[m]模型, 由 6 个部分组成: AR(自回归)、I(整合/差分)、MA(移动平均)、SAR(季节性自回归)、SI(季节性整合/差分)、SMA(季节性移动平均), 每个模型均是 SARIMA 的特殊情况, 各自对应的参数为 p, d, q, P, D, Q, 还有一个参数[m]表示季节性周期长度, 本文的实验数据 m 值为 12; 模型结构为:

$$\Phi(B)\Phi_m(B)\nabla^d\nabla_m^D Y_t = \Theta(B)\Theta_m(B)\varepsilon_t \tag{1}$$

$$\begin{cases} \Theta(B) = 1-\theta_1 B-\cdots-\theta_q B^q \\ \Phi(B) = 1-\varphi_1 B-\cdots-\varphi_p B^p \\ \Theta_m(B) = 1-\theta_1 B^m-\cdots-\theta_Q B^{Qm} \\ \Phi_m(B) = 1-\varphi_1 B^m-\cdots-\varphi_P B^{Pm} \end{cases} \tag{2}$$

其中, B 是滞后算子, d 是趋势差分次数, D 是以周期 m 为步长的季节差分次数, $\{\varepsilon_t\}$为白噪声序列.

使用 SARIMA 模型建模的过程为:

1) 检验原序列平稳性, 对不平稳的原序列通过 d 次差分使其满足平稳性条件;

2) 对平稳化后的时间序列数据绘制自相关图和偏自相关图, 以此初步识别模型的 p、q、P、Q 可以取值的范围, 然后使用 AIC 准则作为寻找最佳模型的准则, 在所有通过检验的模型选择 AIC 值最小的模型作为最适合的 SARIMA 模型;

3) 对已经选定的模型进行检验, 观察模型参数是否通过检验以及残差序列是否为白噪声序列, 如果模型参数未能通过检验或残差序列存在相关性, 则模型需要重新拟合;

4) 使用通过检验的 SARIMA 模型进行时间序列预测.

(2) Holt-Winters 模型

Holt-Winters 模型是一种可用于周期性和趋势性时间序列预测问题的模型, 也称作三次指数平滑法. 模型思想可认为是对一组时序数据长期趋势、周期变化和随机扰动的分解, 因为 Holt-Winters 模型是在保留了随机扰动信息的一次指数平滑模型和保留了趋势信息的二次指数平滑模型的基础上添加一个新参数, 使其对时间序列的周期性进行描述, 模型分为累加模型和累乘模型两种.

累加模型的数学表达如下:

$$\begin{cases} s_t = \alpha(X_t - p_{t-k}) + (1-\alpha)(s_{t-1} + t_{t-1}) \\ i_t = \beta(s_t - s_{t-1}) + (1-\beta) i_{t-1} \\ p_t = \gamma(x_t - s_t) + (1-\gamma) p_{t-k} \\ X_{t+h} = s_t + h \cdot i_t + p_{t-k+h} \end{cases} \tag{3}$$

累乘模型的数学表达为:

$$\begin{cases} s_t = \alpha(\dfrac{X_t}{p_{t-k}}) + (1-\alpha)(s_{t-1}+t_{t-1}) \\ i_t = \beta(s_t - s_{t-1}) + (1-\beta)i_{t-1} \\ p_t = \gamma(\dfrac{x_t}{s_t}) + (1-\gamma)p_{t-k} \\ X_{t+h} = (s_t + h\cdot i_t)p_{t-k+h} \end{cases} \tag{4}$$

在上述公式中, X_t 为 t 时刻的观测值, s_t、i_t、p_t 分别为 t 时刻的稳定成分、周期成分和趋势成分, h 为区间外预测期数, k 为周期长度, α、β、γ 为平滑参数且 α、β、$\gamma \in [0,1]$.

累加模型适合于周期趋势不随时间变化而发生变化的时间序列, 累乘模型适合于周期趋势随时间变化发生变化的时间序列, 在使用 Holt-Winters 模型进行单项模型预测时需要根据实验数据自身特点选择其中一个模型.

1.2 组合模型理论

对于一个 n 期时间序列预测问题, 观测数据为 $x = \{x_1, x_2, \cdots, x_t, \cdots, x_n\}$, 其中 x_t 表示变量在 t 时刻的观测值, 选取 J 个单项预测模型对其进行预测, 得到 J 个预测模型 $\{\varphi_1, \varphi_2, \cdots, \varphi_J\}$, 则在 t 时刻对序列的预测值为 $\{\varphi_1(t), \varphi_2(t), \cdots, \varphi_J(t)\}$, 组合预测的思想就是对每个模型的预测值通过适当的加权方法进行组合, 设单项预测模型的权重为 $\{\omega_1, \omega_2, \cdots, \omega_J\}$, 且 $\sum_{i=1}^{J}\omega_i = 1$, 则组合模型的数学表达式为:

$$\Phi_t = \omega_1\varphi_1(t) + \omega_2\varphi_2(t) + \cdots + \omega_J\varphi_J(t) = \sum_{i=1}^{J}\omega_i\varphi_i(t) \tag{5}$$

权重选择的常见思想是根据每个模型的预测精度, 即预测值与观测值之间的误差来进行权重选择, 可供使用的误差包括相对误差、绝对误差、对数误差和误差平方和等, 本文将在使用误差平方和作为加权依据的最小二乘加权法的基础上, 加入正则化项并使用交叉验证对加权法则进行优化, 进而研究正则化和交叉验证对组合预测模型的优化能力.

(1) 最小二乘加权法

在组合模型中经常使用预测值与观测值之间的误差平方和最小化来计算权重系数, 这种方法被称为最小二乘法[20], 对于一个线性组合预测模型:

在 t 时刻组合预测模型的预测值为 Φ_t, 第 i 种单项预测模型的预测值为 $\varphi_i(t)$, 观测值为 x_t; 那么在 t 时刻, 第 i 种单项预测模型的预测误差为 $e_i(t) = \varphi_i(t) - x_t$, 则组合预测模型在 t 时刻的预测误差为 $e_t = (\Phi_t - x_t) = \sum_{i=1}^{J}\omega_i[\varphi_i(t) - x_t] = \sum_{i=1}^{J}\omega_i e_i(t)$, 为了使组合模型的预测值最优, 将预测值最优问题转化为组合模型性误差最小化问题, 组合预测模型的误差平方和为 $Q = \sum_{t=1}^{n}e_t^2 = \sum_{t=1}^{n}\left(\sum_{i=1}^{J}\omega_i e_i(t)\right)^2$, 在实际运算时一般采用平均误差平方和 MSE 作为组合预测模型的误差衡量准则, 则基于上述理论, 最小二乘加权法的数学表达为:

$$\begin{cases} \min Q = \dfrac{1}{n}\sum_{t=1}^{n}\left(\sum_{i=1}^{J}\omega_i e_i(t)\right)^2 \\ \text{s.t.}\ \sum_{i=1}^{J}\omega_i = 1 \\ \Phi_t = \omega_1\varphi_1(t) + \omega_2\varphi_2(t) + \cdots + \omega_J\varphi_J(t) = \sum_{i=1}^{J}\omega_i\varphi_i(t) \end{cases} \tag{6}$$

(2) 正则化

正则化技术广泛应用于机器学习和深度学习算法中, 常用于防止模型过拟合或提高模型泛化能力, 其思想是在经验风险 (本文指误差平方和最小化函数) 中加上一个正则化项来控制权重系数矩阵的值使权重矩阵缩减或变得相对稀疏, 从而避免有些特征在模型中影响力过高导致的模型过拟合[21].

在组合预测模型中, 如果几个单项预测模型的预测效果表现的比较接近, 直接选取其中预测精度最高的模型进行组合可能会失去一些有用的信息, 但如果将所有单项预测模型一起进行组合, 在传统的加权方式下, 组合模型权重选择会很大程度上偏向在训练数据中预测误差最小的单项预测模型, 形成的组合模型会在训练数据上表现的比较好, 但是在预测数据上则可能会出现“过拟合”现象, 从而降低组合模型对新数据的预测能力. 基于这种思考, 本文将正则化方法引入组合预测模型的加权过程调整每个单项预测模型的权重, 在回归问题 (最小二乘估计) 中, 正则化一般具有以下形式:

$$\min \frac{1}{N}\sum_{i=1}^{N}L(\omega_i x_i - y_i) + \lambda\left(\sum_{i=1}^{m}|\omega_i|^p\right)^{\frac{1}{p}} \tag{7}$$

式 (7) 被称作代价函数, 其中, N 表示数据集的样本总量, m 表示需要被加权的特征总量, ω_i 表示权重系数矩阵, x_i 表示权重系数矩阵对应的样本特征的向量, 加号前是损失函数, 加号是正则化项, λ 为调整两者之间的关系的系数, 一般称之为正则项系数, 常用的正则项有两种: L_1 范数正则化和 L_2 范数正则化.

1) L_1 范数正则化

L_1 范数得名原因是因为式 (7) 中正则化项的 p 取值为 1, 即正则化项为$\lambda\sum_{i=1}^{n}|\omega_i|$, 其中的$\sum_{i=1}^{n}|\omega_i|$就是 L_1 范数, 一般写作$\|\omega_i\|$.

在一般的最小二乘加权模型中, 模型会通过最小化平方误差和来选择权重, 在这个过程中模型只会注重误差平方和最小化而忽略了对权重的控制, 将 L_1 范数加入最小化函数中后, 如果对单个模型赋予的权重过大, 最小化函数仍然无法实现最小化, 这就会使得模型在选择权重时有两个方面的考虑: 误差平方和最小化和权重系数矩阵最小化.

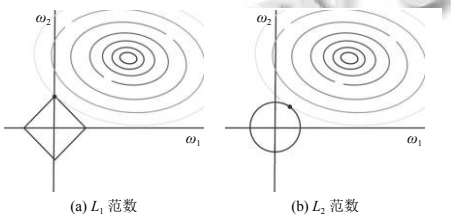
将 L_1 范数代入本文的组合预测模型中, 得到基于 L_1 范数的组合预测模型的代价函数:

$$\min f(\omega)=\frac{1}{n}\sum_{t=1}^{n}\left(\sum_{i=1}^{J}\omega_i e_i(t)\right)^2+\lambda\|\omega_i\| \tag{8}$$

2) L_2 范数正则化

与 L_1 范数类似, L_2 范数得名原因是由于式 (7) 中正则化项的 p 取值为 2, 即为$\left(\sum_{i=1}^{m}|\omega_i|^2\right)^{\frac{1}{2}}=\sqrt{\sum_{i=1}^{m}|\omega_i|^2}=\|\omega_i\|^2$, L_2 范数与 L_1 范数的工作原理类似, 都是在最小化损失函数的过程中降低权重系数矩阵的值, 但是达到的效果有所不同, 具体见图 1.

如图 1 所示, 图左是 L_1 范数, 图右是 L_2 范数, 旋涡部分代表损失函数最小化的过程, 在整体代价函数最小化的过程中, L_1 范数的图像与权重ω_i的图像在零点以外的地方只会无限接近不会重合, L_2 范数的图像与权重ω_i的图像则有可能在零点以外的地方重合; 这意味着 L_1 范数可能会使部分权重变为 0, 而 L_2 范数则只会让每个特征尽可能的变小而不会变为 0.



(a) L_1 范数 (b) L_2 范数

图 1 L_1 范数与 L_2 范数在代价函数最小化中的图像

将 L_2 范数代入本文的组合预测模型中, 得到基于 L_2 范数的组合预测模型的代价函数:

$$\min f(\omega)=\frac{1}{n}\sum_{t=1}^{n}\left(\sum_{i=1}^{J}\omega_i e_i(t)\right)^2+\lambda\|\omega_i\|^2 \tag{9}$$

这样就得到了带有正则化项的组合预测模型:

$$\begin{cases}\min f(\omega)=\dfrac{1}{n}\sum_{t=1}^{n}\left(\sum_{i=1}^{J}\omega_i e_i(t)\right)^2+\lambda\|\omega_i\|^p, p\in\{1,2\}\\ \text{s.t.}\sum_{i=1}^{J}\omega_i=1\end{cases}$$
$$\Phi_t=\omega_1\varphi_1(t)+\omega_2\varphi_2(t)+\cdots+\omega_J\varphi_J(t)=\sum_{i=1}^{J}\omega_i\varphi_i(t) \tag{10}$$

基于正则化的组合模型在为单项预测模型加权的过程中, 就不会出现因为某个单项预测模型在训练数据上表现过好或过差而为其赋予很大的权值, 避免了组合预测模型在预测数据上表现出“过拟合”的情况.

(3) 交叉验证

交叉验证是机器学习中另一种选择模型、优化模型的方法, 常用于数据量对于拟合模型不是很充足的时候, 其思想是重复的使用数据, 把给定的数据进行切分, 将切分的数据集分为训练集和测试集, 在这个基础上反复进行训练、测试以及模型选择. 交叉验证方法包括简单交叉验证、K 折交叉验证和留一交叉验证.

如图 2 所示, K 折交叉验证的思想是将已给数据切分为 K 个大小相同、互不相交的子集, 利用其中 K–1 个子集进行训练, 剩下的一个子集用来测试模型, 这种操作重复进行 K 次, 最后选择在 K 次评测中误差最小的模型; 留一交叉验证是是 K=N 时的 K 折交叉验证, 这种方法往往在数据量较小的情况下使用, 本文所使用的数据量相对于较小, 故选用留一交叉验证.

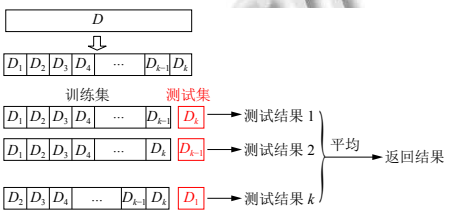


图 2 K 折交叉验证示意图

(4) 模型优劣判定

本文将把总数据集分为训练集和测试集, 总样本量设为 N, 训练集容量为 m, 选用平均绝对百分比误差 $MAPE$ 对测试集上的模型预测结果进行评价:

$$MAPE=\frac{1}{N-m}\sum_{t=m+1}^{N}\left|\frac{\Phi_t-x_t}{x_t}\right| \tag{11}$$

2 实证分析

2.1 数据说明

为了保证正则化和交叉验证有一定的学习率, 本文在一定程度上扩大了时间序列的数据量, 采用的时

间序列数据为我国铁路客运量 2005 年 1 月~2018 年 11 月的月度数据, 样本量为 168, 数据来源为国家统计局, 时间序列趋势图如图 3.

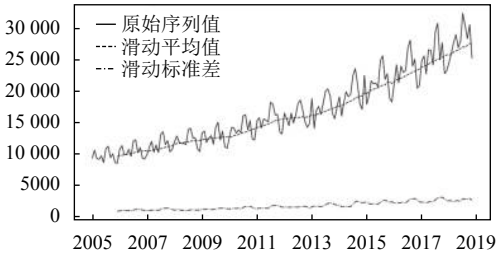


图 3 时间序列均值-标准差分析图

如图 3 所示, 我国铁路客运量有明显的升高趋势和周期性趋势, ADF 检验表明数据存在不稳定性, 是不平稳序列, 适合使用 SARIMA 模型和 Holt-Winters 模型进行时间序列预测.

2.2 单项预测模型选择

根据自相关和偏自相关图、AIC 准则、模型显著性以及实验数据序列周期性与时间的相关性, 本文使用的单项预测模型如下:

1) SARIMA(1,1,1)(1,1,1)[12]模型;

2) SARIMA(1,1,1)(0,1,1)[12]模型;

3) SARIMA(0,1,1)(0,1,1)[12]模型;

4) Holt-Winters 累加模型.

另外, 以上 3 个 SARIMA 模型的 AIC 值接近, 参数只有一个未通过检验但是非常接近给定的 P 值,

为了保留可能存在的潜在信息, 对这 3 个 SARIMA 模型的预测结果都予以保留.

2.3 组合预测模型

本文选用的组合预测模型如下:

1) 最小二乘法组合模型;

2) 基于 L_1 范数的组合模型($\lambda = 1$);

3) 基于 L_1 范数和交叉验证的组合模型;

4) 基于 L_2 范数的组合模型($\lambda = 1$);

5) 基于 L_2 范数和交叉验证的组合模型.

在铁路客运量的组合预测中, 将 2005 年至 2017 年的数据作为训练集, 2018 年的数据值作为预测集, 计算每个模型在预测集上的误差, 得到模型在测试集上的表现如表 1.

由表 1 所示的各模型在预测集上的预测误差可观察到, 简单最小二乘法组合预测模型的预测误差为 3.9576%, 比所有的单项预测模型的预测误差都要低; 引入 L_1 范数和 L_2 范数后的组合模型的预测误差分别为 3.7993%、3.8118%, 比最小二乘法组合预测模型的预测误差略低; 因此根据实验数据, 说明在组合预测模型进行权重选择时引入正则化项能够使组合预测模型的预测误差降低, 并且基于 L_1 范数正则的组合模型比基于 L_2 范数正则的组合模型对组合预测模型的优化效果更好; 此外, 交叉验证在一定程度上改善了 L_1 范数正则化模型的预测效果, 对 L_2 范数的模型起到的作用很小, 几乎可以忽略不计.

表 1 各模型在预测集上的预测误差 (单位: %)

预测模型	*MAPE*
SARIMA(1,1,1)(1,1,1)[12]模型	4.6521
SARIMA(1,1,1)(0,1,1)[12]模型	4.8742
SARIMA(0,1,1)(0,1,1)[12]模型	4.7462
Holt-Winters 累加模型	4.1757
最小二乘法组合模型	3.9576
基于 L_1 范数的组合模型	3.7993
基于 L_1 范数和交叉验证的组合模型	3.7622
基于 L_2 范数的组合模型	3.8118
基于 L_2 范数和交叉验证的组合模型	3.8118

结合各组合预测模型对单项预测模型的权重选择进行分析, 由表 2 所示的各组合模型中单项模型的权重, 可发现: 1) 所有的组合模型的权重分布都满足 $\sum_{i=1}^{J}\omega_i = 1$的条件; 2) 在最小二乘组合模型的权重选择的基础上, 加入 L_1 范数和 L_2 范数的单项预测都使其绝对值降低了, 但能够观察到, 加入 L_1 范数正则的组合模型在调整权重时, 权重绝对值的下降速度更快, 而加入 L_2 范数正则的组合模型对权重的调整则并不明显.

2.4 实验结论与理论分析

L_1 范数是绝对值函数, L_2 范数则是二次函数, 这种差异导致 L_1 范数正则在使模型权重衰减时始终保持同一个速度, 而 L_1 范数正则会在权重较大时提高衰减速度, 权重较低时降低衰减速度, 最终, L_1 范数正则在降低模型复杂度时保留了权重较大的单项模型, 并且尽量使模型变得稀疏, L_2 范数正则在降低模型复杂度时则使模型更加符合规则化.

相对于 L_2 范数正则经常被使用的机器学习特征分配的应用场景, 组合预测模型需要被分配权重的单项预测模型个数较少, 权重数值小, 这使得 L_2 范数正则在调整组合预测模型的权重时没有 L_1 范数正则的下降速度快, 改善效果明显.

基于以上分析, 认为 L_1 范数正则化更适合用于组合预测模型. 另外在实验过程中发现, 参与组合预测的单项预测模型越多, 正则化的改善效果越好.

表 2 各组合模型的权重选择

组合预测模型	单项预测模型			
	1	2	3	4
最小二乘法组合模型	–1.399 525	0.937 399	1.053 596	0.401 297
基于 L_1 范数的组合模型	–0.607 698	0.869 332	0.949 926	–0.211 560
基于 L_1 范数和交叉验证的组合模型	–0.607 690	0.869 337	0.949 950	–0.211 597
基于 L_2 范数的组合模型	–1.381 567	0.936 118	1.049 200	0.396 249
基于 L_2 范数和交叉验证的组合模型	–1.381 466	0.936 101	1.049 116	0.396 249

交叉验证则在一定程度上改善了 L_1 范数正则化模型的预测效果, 对 L_2 范数的模型起到的作用很小, L_1 范数的代价函数在组合预测模型中表现出的惩罚力度大, 使得交叉验证的改善效果强于 L_2 范数的惩罚力度, 并且在实验中发现, 训练数据分配的越多, 交叉验证的改善效果就越好.

3 结论

本文按照常规的组合模型设计思路, 在基于最小二乘法的组合预测模型中, 引入了正则化和交叉验证用以优化组合预测模型的权重选择, 通过实验发现: 首先, 组合预测模型优于所有单项预测模型的预测精度, 正则化和交叉验证都能在不同程度改善基于最小二乘法的组合预测模型, 且 L_1 范数正则化比 L_2 范数正则化的改善效果更好; 其次, 实验中发现候选的单项预测模型越多、数据量越大, 正则化和交叉验证表现的效果就越好, 因此这种组合预测方法可以推广到基于大数据的多模型组合预测中.

参考文献

1 邓聚龙. 灰色系统基本方法. 武汉: 华中科技大学出版社, 2005.
2 何书元. 应用时间序列分析. 北京: 北京大学出版社, 2003.
3 孙轶轩, 邵春福, 计寻, 等. 基于 ARIMA 与信息粒化 SVR 组合模型的交通事故时序预测. 清华大学学报 (自然科学版), 2014, 54(3): 348–353, 359.
4 宋国君, 国潇丹, 杨啸, 等. 沈阳市 $PM_{2.5}$ 浓度 ARIMA-SVM 组合预测研究. 中国环境科学, 2018, 38(11): 4031–4039. [doi: 10.3969/j.issn.1000-6923.2018.11.005]
5 王祥雪, 许伦辉. 基于深度学习的短时交通流预测研究. 交通运输系统工程与信息, 2018, 18(1): 81–88.
6 王锟, 王洁, 刁迎春. 基于 LS-SVM 组合预测的地空导弹发射车液压系统油液污染度预测. 传感技术学报, 2012, 25(5): 712–717. [doi: 10.3969/j.issn.1004-1699.2012.05.029]
7 郝少峰, 方源敏, 杨建文, 等. 基于熵权法的组合模型在滑坡变形预测中的应用. 测绘工程, 2014, 23(7): 62–64. [doi: 10.3969/j.issn.1006-7949.2014.07.015]
8 温廷新, 陈晓宇. 基于组合赋权的混合粒子群优化支持向量机的岩爆倾向性预测. 安全与环境学报, 2018, 18(2): 440–445.
9 李长锦, 谭满春. 基于最优加权法的改进交通流组合预测研究. 暨南大学学报 (自然科学与医学版), 2010, 31(5): 457–461.
10 董艳, 贺兴时. 一种组合预测模型及其应用. 西安工程大学学报, 2010, 24(1): 128–130. [doi: 10.3969/j.issn.1674-649X.2010.01.029]
11 李佩, 彭斯俊. 一种新的组合权重在组合预测模型中的应用. 河南科技大学学报 (自然科学版), 2018, 39(2): 87–93.
12 孙炯, 梁锦强, 刘凯. 一种基于最小二乘法的广义加权组合预测模型. 科技通报, 2013, 29(8): 10–12. [doi: 10.3969/j.issn.1001-7119.2013.08.004]
13 梁锦强, 孙炯, 刘凯. 广义加权最小二乘组合预测法在装备故障率预测中的应用. 计算机与数字工程, 2012, 40(9): 39–40, 50. [doi: 10.3969/j.issn.1672-9722.2012.09.014]
14 郭伟, 李京. 基于改进的优化组合方法的旅游需求预测. 统计与决策, 2011, (8): 75–77.
15 Hansen BE, Racine JS. Jackknife model averaging. Journal of Econometrics, 2012, 167(1): 38–46. [doi: 10.1016/j.jeconom.2011.06.019]
16 Berrar D. Cross-validation. Encyclopedia of Bioinformatics and Computational Biology, 2019, 1: 542–545.
17 Zhao SW, Zhou JH, Yang GR. Averaging estimators for discrete choice by *M*-fold cross-validation. Economics Letters, 2019, 174: 65–69. [doi: 10.1016/j.econlet.2018.10.014]
18 Bayer S. Combining value-at-risk forecasts using penalized quantile regressions. Econometrics and Statistics, 2018, 8: 56–77. [doi: 10.1016/j.ecosta.2017.08.001]
19 Xu SJ, Chan HK, Zhang TT. Forecasting the demand of the aviation industry using hybrid time series SARIMA-SVR approach. Transportation Research Part E: Logistics and Transportation Review, 2019, 122: 169–180. [doi: 10.1016/j.tre.2018.12.005]
20 李静. 变权重组合预测模型的局部加权最小二乘解法. 统计与信息论坛, 2007, 22(3): 44–47. [doi: 10.3969/j.issn.1007-3116.2007.03.009]
21 Hou RR, Xia Y, Bao YQ, *et al*. Selection of regularization parameter for l_1-regularized damage detection. Journal of Sound and Vibration, 2018, 423: 141–160. [doi: 10.1016/j.jsv.2018.02.064]