基于关注度的 CPI 指数构建

王艺楷
15300180076

摘要

居民消费价格指数，简称 CPI，是一个颇受各界关注的宏观经济指标。长期以来，官方的月度 CPI 公布日期在每月的中上旬（10 号左右）。作为反映上月度宏观经济情况的一项指标相对具有滞后性。在本文中，我们基于互联网用户对 CPI 指数相关商品、服务的关注度与 CPI 指数具有相关性的假设，通过实时的搜索引擎指数表达关注度，利用机器学习的相关模型预测当月的 CPI 指数，可以做到对这一指标实时性的预测，且与官方公布的 CPI 指数具有较好的近似度。本文利用关注度构建指数的框架与以往利用各种价格指数构建 CPI 指数的框架不同，对于 CPI 指数的构建的一种新思路进行了探索，且该框架具有良好的迁移性，可用于各种指数的构建，为经济学研究提供帮助。

1 引言

CPI 指数的预测问题是学术界研究的一个热点问题。作为一项反映宏观经济情况的重要指标，由于官方的 CPI 指数统计具有相对的滞后性，为了更快的发现宏观经济情况的变化以提前做出应对，很多学者对 CPI 指数的预测提出了一系列的预测模型，主要的预测方法有：时间序列预测方法，组合预测方法和灰色预测方法。这些方法在当期 CPI 指数预测上都有一定效果，但也存在着种种缺陷。传统模型一般基于各种价格指数数据进行预测，值得指出的是，这些预测模型所采用的多为抽样调查所得数据，一方面由于抽样可能存在的偏差使得其不如官方统计指标那么客观，另一方面由于统计数据所需时间也在一定程度上影响到了模型的时效性。此外，价格指数本身已经是二手数据，存在破坏原有信息的可能，通过价格指数来预测 CPI 指数一定程度上增大了预测的偏差。[4]

近些年来，在互联网的蓬勃发展下，基于互联网大数据的模型也开始逐渐进入学界的视线。自谷歌使用搜索引擎数据预测流感爆发 [6] 以

来，基于搜索引擎数据来进行预测逐渐的被用于各领域。本质而言，搜索引擎数据反映的是使用搜索引擎的群体对特定关键词的关注度，而以关注度为基础建立的预测所作出的假设在于：特定群体的关注度的度量及趋势变化与预测指标之间具有明显的相关性。在此基础上，基于搜索引擎数据构建预测模型的两个关键点在于关键词的选取和转换。关键词的选取可以粗略的分为两种方法，其一为根据研究者的主观经验判断及他人经验初步选定关键词，再根据模型效果筛选关键词；其二为先纳入大量关键词，再利用算法自动选取关键词。关键词的转换则是利用模型将网络上各种非结构化数据转化为结构化数据，常用的方法有用于语义特征提取的自然语言处理方法等。基于关注度的 CPI 指数预测模型是一种较为新颖的 CPI 指数构建思路，该类模型抛开了传统的利用价格指数来反映价格指数的思路，而依赖于关注度与价格指数的相关性，将相关关系运用在 CPI 指数构建模型中。

2 相关研究

在基于关注度的 CPI 指数构建模型研究中，中外学者根据不同的理论基础采用了多种多样的模型，其中较为启发性的模型有 MIDAS 模型、Elastic net 估计方法和 SARIMA-GMDH 模型。

2.1 MIDAS 模型

MIDAS 模型 [5][9] 选取了 CPI、物价、价格、涨价、降价、通货膨胀、通货紧缩七个关键词，通过使用含有超参数的滞后权重多项式函数构建模型，利用数值优化的方法估计混频数据模型中的最优参数。

$$\begin{aligned} log(CPI)_t^M &= c + \alpha log(CPI)_{t-1}^M + \epsilon_t^M \\ &\beta_0 \sum_{j=0}^{30-1} \gamma_{0,j} INDEX_{t-j/30}^D + \\ &\beta_1 \sum_{j=0}^{30-1} \gamma_{1,j} INDEX_{t-1-j/30}^D \end{aligned} \tag{1}$$

其中 M 代表月度低频数据，$\gamma_{0,j} = \frac{\theta_{k,1}j+\theta_{k,2}j^2}{\sum_{j=1}^{3} 0\theta_{k,1}j+\theta_{k,2}j^2}$，$\gamma_{1,j} = \frac{1}{30}$

MIDAS 模型可以实现对于 CPI 的实时预测，例如：每月中旬可以利用该模型对当月 CPI 做出初步预测，待 30 天数据收集完全后，再更新预测值，得到最终的预测。因此，MIDAS 模型可以做到对 CPI 价格指数任意度量尺度上的预测。

2.2 Elastic net 估计方法

Elastic net[7] 是一种有效的变量选择方法，该方法在选择变量时既能防止出现逐步回归的变量选择所产生的模型稳定性较低的问题，又能防止出现 Lasso 方法只能得到稀疏模型的情况。[11] 对于线性回归模型：

$$Y = X\beta + \epsilon \tag{2}$$

定义：

$$L(\theta,\gamma,\beta)=\sum_{i=1}^{n}(Y_i-X_i^T\beta)^2+\theta\sum_{j=1}^{p}\beta_j^2+\gamma\sum_{j=1}^{p}|\beta_j| \tag{3}$$

我们有：

$$\hat{\beta}_{Elasticnet}=(1+\theta)argmin_{\beta}L(\theta,\gamma,\beta) \tag{4}$$

在参数估计过程中，首先给出 θ 的参数空间，对空间中的每一个 θ 结合 LARS-EN 算法产生所有解决路径，再利用交叉验证选择参数 γ，最终利用均方误差最小的 θ 和 γ 确定 $\hat{\beta}$.

Elastic net 估计方法可以有效减少由于 CPI 价格指数基期五年一轮换所带来的模型稳定性问题，有效提高模型的预测精度。

2.3 SARIMA-GMDH 模型

SARIMA 模型，即乘积季节模型，来源于自回归求积滑动平均模型，其一般表达形式如下：

$$f_p(L)F_p(L^s)(1-L)^d(1-L^s)^D X_t=q_q(L)Q_q(L^s)v_t \tag{5}$$

其中 L 是滞后算子，$f_p(L)$ 是平稳 AR 算子，$q_q(L)$ 是可逆 MA 算子，$F_p(L^s)$ 是季节 AR 算子，$Q_q(L^s)$ 是季节 MA 算子，v_t 为白噪声过程。将 SARIMA 模型与 GMDH 模型结合起来构建的组合模型，使得 SARIMA 模型的效果更加显著。

SARIMA-GMDH 模型 [10] 将低频数据与高频数据结合起来进行组合预测，将组合预测的方法运用在 CPI 价格指数构建中，获得了良好的效果。

3 模型

本文构造的模型主要基于机器学习的回归模型。根据机器学习的相关知识与经验，结合本问题数据矩阵的特点，我们选取了神经网络模型、Kernel Ridge 回归模型、Bagging 回归模型和支持向量回归模型（SVR）四种回归模型来进行 CPI 指数的预测。

3.1 神经网络

神经网络是一种模仿动物神经网络行为特征1，进行分布式并行信息处理的算法数学模型，神经网络可以任意逼近非线性模型，同时可以快速提取特征，获得优异的效果。神经网络可以粗略的分为输入层，隐藏层，输出层三种，每一层含有特定数量的神经元，神经元之间可以进行非线性联系2。输入可以类比为神经元的树突，而输出可以类比为神经元的轴突，计算则可以类比为细胞核。

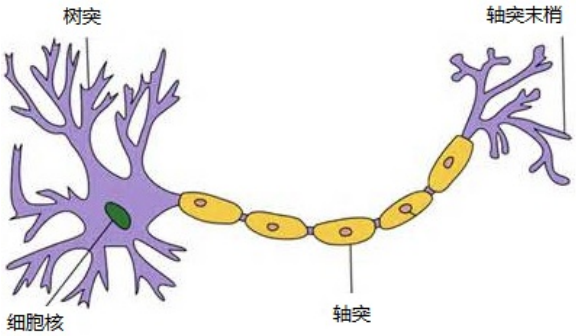

图 1: 人脑中的神经元示意图

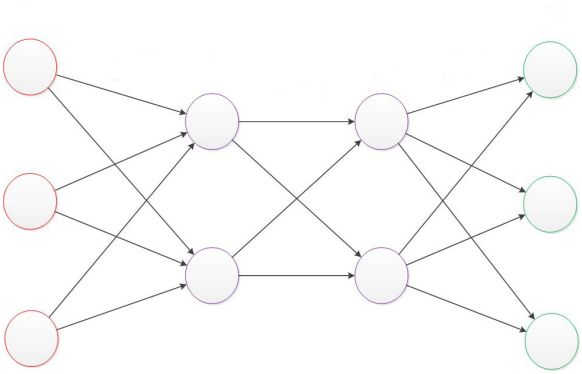

图 2: 一个含有两个隐藏层的简单神经网络

一般地，神经网络主要运用在分类任务中，对神经网络的诸多改进也多在于提高其分类性能。在本文中，我们的神经网络模型从分类模型修改为回归模型3。主要的改动在于，将输出层从 n 个不同的固定值改为 1 个任意值，然后利用最小均方误差拟合其与真实值的差距，将结果反馈给神经网络进行训练。[8]

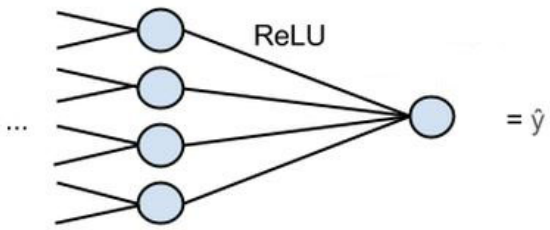

图 3: 回归神经网络

在本问题中，我们构建的神经网络模型具有一个输入层，两个隐藏层和一个输出层，其中输入层含有 1800 个神经元，两个隐藏层分别含有 500 个神经元与 200 个神经元，输出层含有一个神经元。每一层采用 sigmoid 函数作为激活函数，以最小二乘法作为损失函数，训练次数为 500 次。

3.2 Kernel Ridge 回归

我们的任务中的数据集合矩阵存在多重共线性，使用一般的线性回归模型 $y = w^T x$ 时，最小二乘法求得的参数 w 在数值上会非常的大，如果输入变量 x 有一个微小的变动，其反应在输出结果上也会变得非常大，即一般的线性模型对输入变量总的噪声非常敏感。如果能限制参数 w 的增长，使 w 不会变得特别大，那么模

型对输入 w 中噪声的敏感度就会降低。这就是 Ridge 回归的基本思想。为了限制模型参数 w 的数值大小，就在模型原来的目标函数上加上一个惩罚项，如果惩罚项是参数的 l2 范数，就是 Ridge 回归。而 Kernel Ridge 回归则是利用核函数将数据集合矩阵映射到高维空间后进行 Ridge 回归 [1]。在 Kernel Ridge 回归模型中，损失函数定义如下：

$$LOSS = ||k(w^T)k(x) - k(y)||^2 + ||k(w^T)\tau|| \quad (6)$$

3.3 Bagging 回归

在机器学习领域中，决策树回归算法是一个可读性很强、回归快，同时不需要对数据进行归一化、缩放的处理的算法。但是决策树算法的不足之处在于易过拟合。Bagging 是 bootstrap aggregatin 的缩写，属于机器学习中模型融合的一种方法 [3]。利用 Bagging 来融合决策树，可以有效的解决决策树过拟合的问题。Bagging 的基本思想在于利用重抽样在有限的样本上生成大量的决策树，在进行回归时对大量决策树的结果进行综合给出最终回归结果。Bagging 较单棵决策树来说，降低了方差，由于将多棵决策树的结果进行了平均，这损失了模型的可解释性，但可以利用每个特征使树的 RSS（或 Gini 指数）平均降低的量来度量特征的重要性。

3.4 SVR

SVR 回归，即为利用核函数找到一个回归平面，让一个集合的所有数据到该平面的距离最近，从而对数据进行回归 [2]。假设核函数为 ϕ,SVR 回归模型要解决的问题是在给定

$$-\epsilon - \xi_i^* \leq w^T\phi(x_i) + b - z_i \leq \epsilon + \xi_i, \xi_i, \xi_i^* \geq 0 \forall i \quad (7)$$

的条件下求解

$$min_{w,b,\xi,\xi^*} \frac{1}{2} w^T w + C\sum \xi_i + C\sum \xi_i^* \quad (8)$$

SVR 回归模型在小样本高维数据中具有优异的表现。

4 数据处理

2013 年版的居民消费支出分类将居民消费支出划分为食品烟酒，衣着，居住，生活用品及服务，交通和通信，教育、文化和娱乐，医疗保健，其他用品和服务等 8 个大类。根据 8 个大类的不同特点，考虑到搜索引擎所反应人群的特殊性，我们从 8 个大类的每个小类中依据与用户搜索习惯匹配度由高到低和反应市场价格波动相关性由强到弱的原则选取了以下 60 个关键词作为我们模型的特征。见表 1:

食品烟酒	外卖、白酒、肉制品、牛奶、蔬菜、啤酒、饮料、茶叶、食用油
衣着	长裤、短裤、外套、鞋、衬衫
居住	天然气、煤、电费、装修、水费、宾馆、房租、物业
生活用品及服务	洗面奶、面膜、床上用品、家政、洗衣机、香水、空调、冰箱、家具
交通和通信	手机、飞机票、汽油、快递、火车票、柴油、汽车、电动车、石油
教育、文化和娱乐	旅游、健身、游戏、电子书、笔记本、辅导、培训、玩具、电影、教材
医疗保健	保健品、门诊、医院、药品、挂号
其他用品和服务	美容、金融、保险、首饰、手表

表 1: 关键词

由于从 2016 年 1 月起我国开始使用 2015 年作为新一轮的对比基期，为了排除不必要的因素影响，我们从百度指数网站上爬取了从 2016 年 1 月 1 日至 2017 年 12 月 31 日每个关键词每日的百度指数作为关注度的度量。由于百度指数网站上的搜索指数分为 PC 端和移动端两种，我们两种指数都进行了爬取，所有的数据共计 60 * 731 * 2= 87720 （条）。

由于百度指数网站采取了一定的反爬虫措施，在分析所得到的数据时，我们发现存在一定数量（3165）的数据为异常数据（数值小于 10），这些数据呈离散的分布状态，我们尝试了几种缺失数据补全方法，在之后的实验中发现数据补全方法对结果并无明显影响，因此在最终的实验中，我们采用插值法填充缺失数据。

由于我们的算法要求各月间的特征长度需相等，我们以每个月 30 天为基准，对于不足 30 天的 2 月，我们将 3 月 1 日（2016 年）或 1 月 31 日与 3 月 1 日（2017 年）算进 2 月的特征中，对于含有 31 天的月份，我们尝试了去除每月的第一天与每月的最后一天两种方式，发现对结果没有明显影响，在最终的实验中采取了去除每月最后一天的方式选取特征。

由于我们关注的是特征的变化趋势而非特征值，故我们对每一关键词的数据进行了标准化处理，即对每一关键词减去该关键词的均值后除以该关键词的标准差。

我们从国家统计局网站上获得了 2016 年

1 月至 2017 年 12 月共 24 个月份的 CPI 指数月度数据，该数据以上年同月为 100 进行计算，其调查目录依据国家统计局发布的《居民消费支出分类（2013）》制定而成，与我们使用的关键词重合度较高，减少了我们预测的偏差。

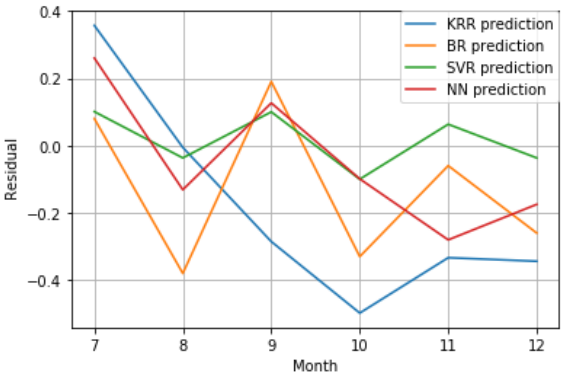

图 5: 模型预测结果与真值的残差

5 结果分析

我们使用 2016 年 1 月至 2017 年 6 月共 18 个月份的数据作为训练集，2017 年 7 月至 2017 年 12 月共 6 个月份的数据作为验证集。图 4 展示了四个模型的预测结果，图 5 展示了四个模型的残差。从图中我们可以看出，SVR 模型对数据的预测结果最好，神经网络和 Bagging 回归模型的结果次之，Kernel Ridge 回归模型对于该问题不能起到很好的作用。

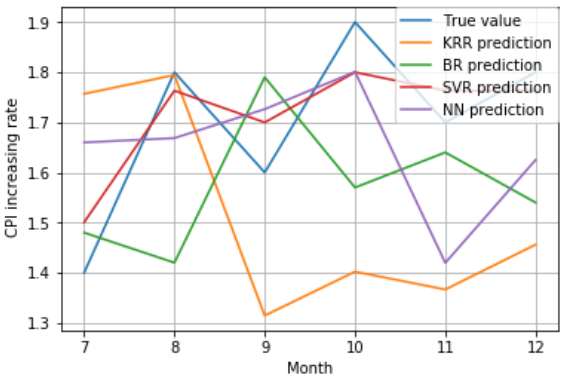

图 4: 模型预测结果与真值之间的比较

更细致的比较可以发现，Bagging 回归模型对于 CPI 指数的预测偏保守，神经网络模型对 CPI 指数的波动性预测偏保守，但可能会出现大的波动，SVR 则对 CPI 指数的数据和波动都有较好的预测。结果表明在合适的回归模型下基于关注度的搜索引擎指数和 CPI 指数之间具有良好的相关性，因此利用基于关注度的模型预测 CPI 指数的方法是可行的。

6 讨论

本文表明了基于关注度的 CPI 指数模型的构建是可行的，从实验结果来看还有一定的提升空间。从关键词的选取来看，我们选取关键词的原则为主观的经验判断，在一定程度上影响了关键词对 CPI 指数的反应程度，另一方面，我们选取的搜索引擎为百度搜索，在价格相关的关注度模型中，百度指数有时并不能较好的体现关注度与价格变化之间的关系，相较之下，

淘宝等电商的指数或许是一个更好的选择。但由于技术条件的限制，我们并不能获取淘宝指数，所以将其作为模型改进的可能选择之一。从回归模型的结果来看部分模型的回归结果并不尽如人意，其中一个可能的原因是由于我们用来做训练集的数据只有 18 个，一定程度上存在过拟合的问题。解决模型的过拟合问题一方面可以利用 Bootstrap 等方法增加训练集，另一方面也可以从选取更好的惩罚因子，降低训练次数等方面着手防止模型的过拟合。

本模型具有良好的可迁移性，虽然在本文中我们利用其预测的是 CPI 指数，但通过关键词的改变和指数的改变，可以非常容易的用来预测别的指数。同时，在本文中我们利用的是月度数据进行预测，但利用 back-off 的方法即可将其变为一个实时预测的模型，即将我们的特征由

$$Feature = DATA_{month} \tag{9}$$

变为

$$Feature = \sum_{i=1}^{n} \lambda_i DATA_i \tag{10}$$

其中，$DATA_i$ 为该月从第一天到第 i 天的数据。

相关性作为一种较弱的关系在如今越来越受到研究者的重视，本文的模型也可视作相关性关系的一个较为成功的应用，利用机器学习方法提取大数据中人为难以分别的特征，不仅可以用来进行预测，也可以重新基于大数据构建相关的指数来反映经济社会的运行情况。随着大数据的发展，机器学习的方法和模型也会与经济学的原理相结合，发挥更好的作用。

参考文献

[1] Senjian An, Wanquan Liu, and Svetha Venkatesh. Fast cross-validation algorithms for least squares support vector machine and kernel ridge regression. Pattern Recognition, 40(8):2154–2162, 2007.

[2] Debasish Basak, Srimanta Pal, and Dipak Chandra Patranabis. Support vector regression. Neural Information Processing-Letters and Reviews, 11(10):203–224, 2007.

[3] Leo Breiman. Bagging predictors. Machine learning, 24(2):123–140, 1996.

[4] W. Erwin Diewert. Index number issues in the consumer price index. Journal of Economic Perspectives, 12(1):47–58, March 1998.

[5] E. Ghysels, P. Santa-Clara, and R. Valkanov. The midas touch: Mixed data sampling regressions. Cirano Working Papers, 5(1):512–517, 2004.

[6] Jeremy Ginsberg, Matthew H Mohebbi, Rajan S Patel, Lynnette Brammer, Mark S Smolinski, and Larry Brilliant. Detecting influenza epidemics using search engine query data. Nature, 457(7232):1012, 2009.

[7] Zou Hui and Trevor Hastie. Addendum: Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society, 67(2):301–320, 2005.

[8] Donald F Specht. A general regression neural network. IEEE transactions on neural networks, 2(6):568–576, 1991.

[9] 徐映梅、高一铭. 基于互联网大数据的 cpi 舆情指数构建与应用——以百度指数为例. 数量经济技术经济研究, (1):94–112, 2017.

[10] 皮进修、赵清俊、彭建文. 基于 sarima-gmdh 的 cpi 组合预测模型. 统计与决策, (17): 22–25, 2016.

[11] 董莉、彭凯越、唐晓彬. 大数据背景下的 cpi 实时预测研究. 调研世界, (8):51–54, 2017.