计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2023,32(10):54−64 [doi: 10.15888/j.cnki.csa.009266]

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

基于文本生成语言模型的股指预测①

温灿红[1], 陈　思[1], 杨海生[2]

[1](中国科学技术大学 管理学院, 合肥 230026)
[2](中山大学 岭南学院, 广州 510275)
通信作者: 杨海生, E-mail: yhaish@mail.sysu.edu.cn

摘　要: 股指预测是金融领域中一个重要课题. 随着计算能力和技术的发展, 从在线新闻中识别和量化有价值的信息为提高股指预测表现创造了机会. 本文为将关于股票指数预测框架的计量经济学文献扩展到高维文本数据提出了一种基于生成语言模型的股票指数预测框架. 该预测框架可以分为两个步骤. 首先, 使用有监督生成语言模型快速过滤噪声词语, 并将剩余文本聚合成可以充分解释股指变动的新闻指数. 其次, 将该新闻指数和历史股指数据共同作为时变参数预测模型的自变量来预测股指未来价值. 该框架不仅丰富了股票指数预测的影响因素并且揭示了这些因素与股票指数价值之间的时变动态关系. 实证研究展示了该预测框架解释能力和样本外预测能力. 在预测的 6 个行业股指中, 本文提出的预测框架得到的均方误差普遍小于传统时间序列和机器学习方法. 与没有考虑新闻信息的时变参数预测模型和长短期记忆网络相比该预测框架也表现了更好的预测性能.

关键词: 深度学习; 分布式多项回归; 负二项回归; 股指预测; 文本分析; 时变参数模型

引用格式: 温灿红,陈思,杨海生.基于文本生成语言模型的股指预测.计算机系统应用,2023,32(10):54–64. http://www.c-s-a.org.cn/1003-3254/9266.html

Stock Index Prediction with Text Generative Language Model

WEN Can-Hong[1], CHEN Si[1], YANG Hai-Sheng[2]

[1](School of Management, University of Science and Technology of China, Hefei 230026, China)
[2](Lingnan College, Sun Yat-sen University, Guangzhou 510275, China)

Abstract: Stock index prediction is an important topic in the field of finance. With the development of computing power and technologies, there are opportunities to improve the performance of stock index prediction by identifying and quantifying valuable information from online news. In order to extend the econometric literature on stock index prediction frameworks to high-dimensional textual data, a stock index prediction framework based on generative language models is proposed. The prediction framework can be divided into two steps. First, a supervised generative language model is used to filter out noisy words quickly and aggregate the remaining text into a news index that can fully explain stock index changes. Second, the news index and historical stock index data are jointly used as independent variables of the time-varying parameter predictive model to predict future stock index values. The framework not only enriches the influencing factors of stock index prediction but also reveals the time-varying dynamic relationship between these factors and stock index values. Empirical research demonstrates the explanatory and out-of-sample predictive power of the proposed prediction framework. Among the six industrial stock indices predicted, the mean square error obtained by the proposed prediction framework is generally lower than that by traditional time series and machine learning methods. Compared with the time-varying parameter predictive model and long short-term memory model that do not consider news

① 基金项目: 国家自然科学基金面上项目 (72173141); 广东省自然科学基金面上项目 (2023A1515012434)
收稿时间: 2023-03-28; 修改时间: 2023-05-06; 采用时间: 2023-05-15; csa 在线出版时间: 2023-08-09
CNKI 网络首发时间: 2023-08-10

information, the proposed prediction framework also exhibits better predictive performance.

Key words: deep learning; distributed multinomial regression; negative binomial regression; stock index prediction; text analysis; time-varying parameter model

股票指数的多变性、非线性和非平稳性为股指预测带来了非常大的挑战. 探索股指预测工具, 提高股指预测精度一直以来都是金融, 经济统计等领域的研究重点和热点[1]. 传统基于公司财务数据和历史交易数据构建时间序列模型预测股指的方法存在局限性. 首先, 使用基本指标数据忽略了未来价格变动的重要驱动因素, 如在线新闻和政策文件. 其次, 传统时间序列回归模型难以准确地表示影响因子和股票指数之间复杂的时变关系. 因此, 研究如何从选择影响因子和改进预测模型两个方面来提高股指的预测精度是有意义和价值的.

为股指预测选择适当的驱动因素对提高预测准确性至关重要. 以往的研究主要依赖历史交易数据、宏观经济变量或资产价格预测股票指数, 但这些特征所解释的指数变化方差仍然不足. 在这种背景下, Gürkaynak 等人[2] 指出新闻文章在解释事件窗口内几乎所有收益率曲线变动方面发挥着重要作用. 这为股票指数背后真正的驱动因素提供了更全面的理解. 幸运的是, 计算能力的提高使得分析大量非结构化的数据如文本、音频和视频等数据具有可行性. 因此, 本文聚焦于挖掘新闻文章来预测股票指数.

在金融领域中, 文本分析的应用仍处于萌芽阶段. Foster 等人[3] 使用主成分回归基于文本数据开发了一个房地产定价模型. Kelly 等人[4] 和 Sert 等人[5] 指出使用命名实体提取、情感分析和潜狄利克雷分配 (latent Dirichlet allocation, LDA) 主题建模可以进行合理的股价预测. Bai 等人[6] 提出了一种基于语义辅助非负矩阵分解的短新闻标题主题指标. 近年来, 随着计算能力的极大提升, 自然语言技术开始被应用于金融市场预测. Ko 等人[7] 使用基于 Transformer[8] 的双向编码器表征 (bidirectional encoder representation from Transformers, BERT)[9] 识别新闻和论坛中的投资者情绪. 这些方法存在一些缺陷. 例如, 主题模型在构建文本模型时没有去除与经济变量不相关的噪声词语, BERT 模型不具有可解释性, 且训练需要消耗大量的计算空间和时间.

为了减轻噪声词语对经济变量预测的影响, 并保留模型的可解释性, Taddy[10] 提出在有监督生成语言模型——多项逻辑逆回归 (multinomial logistic inverse regression, MNIR) 上添加 Gamma-Lasso 惩罚[11]. 生成语言模型假定文档中的词语是通过感兴趣变量确定条件下的词语生成概率定义生成的. 根据感兴趣变量是否可观测, 生成语言模型又可分为有监督生成模型和无监督生成模型. 其中 LDA 主题模型是无监督生成模型的经典模型. 常见的有监督生成模型包括朴素贝叶斯分类器[12] 和 MNIR. 在 MNIR 中词语生成概率与感兴趣变量之间通过多项逻辑函数建立依赖关系. Taddy[13] 详细介绍了 MNIR 在属性预测、治疗效果估计和文档索引等各种应用中的使用. 但 MNIR 在面临协变量高维情形时同样会出现内存不足, 计算时间长等问题. Taddy[14] 将多项回归转化成多个可并行计算的泊松回归, 使得模型具有计算可伸缩性, 解决了 MNIR 模型中的计算问题. 该方法称为分布式多项回归 (distributed multinomial regression, DMR). 然而, DMR 中关于单词计数服从泊松分布的核心假设往往过于严格. 泊松分布具有期望和方差一致的性质, 而实际中文本词频的均值和方差往往并不相同. 例如, 如图 1 所示, 在《中国证券报》发表的房地产相关新闻报道中, “科技队伍”一词在一篇文章中出现次数的均值和方差分别为 5.32 和 15.46, 两个值显然并不一致. 因此, 本文提出使用条件独立的负二项 (negative binomial, NB) 回归代替泊松回归来逼近多项回归. 负二项分布中包含能够调整期望与方差之间的数量关系的尺度参数. 当尺度参数趋近于零时, 负二项分布趋近于泊松分布. 所以, 更通用的负二项分布比泊松分布更适合作为词频的假设条件分布, 使用多个条件独立的负二项回归代替泊松回归来逼近多项回归具有重要研究意义.

基于负二项分布的分布式多项回归 (DMR-NB) 模型可以通过添加惩罚项的方法保留与股票指数相关的词语, 过滤掉噪声词语. 但是, 保留的词语数量仍旧数以千计. 若将这些词语特征不加处理地直接放入预测模型中, 会带来了高维性的挑战. Gürkaynak 等人[2] 指

出可以从新闻中整合出一个解释事件窗口内所有收益率曲线方差的潜在指数. 与 Kelly 等人[4] 一致, 本文基于 DMR-NB 构建文本词频矩阵在经济变量 (即股指) 上的低维投影, 并将该投影作为新闻的综合潜在指数代替高维文本作为影响因子添加到预测模型中, 从而克服由文本高维性带来的挑战.

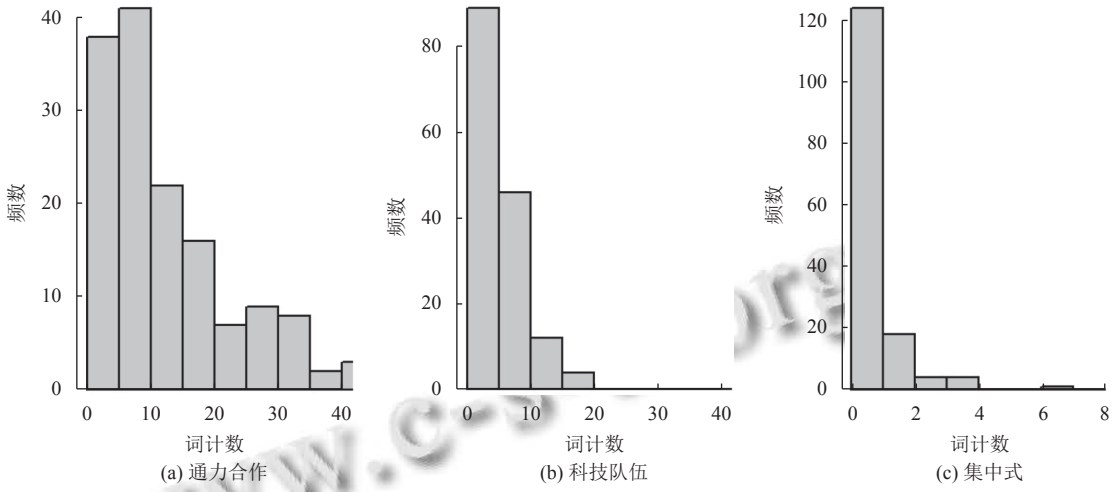

图 1 “通力合作”“科技队伍”和“集中式”词计数分布直方图

在股指预测问题中, 除影响因子的选择外, 预测模型的选择也是研究的重点和热点. 金融市场预测模型主要可以分为基于统计学和概率论的时间序列方法, 以及基于非统计原理的机器学习, 深度学习等方法. 传统时间序列方法包括差分自回归移动平均 (autoregressive integrated moving average model, ARIMA) 模型[15]、贝叶斯向量自回归 (Bayesian vector autoregressive, BVAR) 模型[16,17] 等. 机器学习方法包括长短期记忆网络 (long short term memory, LSTM), 支持向量机等[18–20]. 机器学习方法相比于传统时间序列模型虽然可以较大程度地模拟变量的特征, 但是具有无法解释特征的内在影响机制的缺陷. 股市的强烈不稳定性使得传统时间序列模型中参数不随时间变化的假定变得不可靠. 考虑到时变参数 (time-varying parameter, TVP) 模型[21] 可以有效捕捉渐进变动的关键优势, 本文将其作为本文预测框架中的预测模型. 该模型具有与传统时间序列模型类似的经济结构, 且不再假定模型参数是常数. 实证分析结果也验证了 TVP 模型对可观测股指的拟合几乎与原始股指价值重合.

在 TVP 模型中, 响应变量的预测依赖于参数状态的预测. Koop 等人[22] 阐述了两种最常用的状态预测方法: (1) 直接使用时刻 T 的系数状态作为时刻 T+1 的系数状态的预测值; (2) 允许系数在样本外演化, 通过模拟随机游走状态以生成 T+1 系数状态的预测值. 这两种方法的局限性在于仅考虑了滞后一阶的系数状态, 未考虑滞后多阶的系数状态和特征变量对未来系数状态的影响. 为此, 本文提出使用深度神经网络 (deep neural networks, DNNs) 作为系数状态预测工具[23,24]. 与深度学习在金融预测中常用作“黑匣子”预测器不同, 本文基于 DNNs 预测 TVP 模型的系数状态而非直接预测响应变量的方式保留了经济模型的可解释性, 具有经济学意义.

总之, 为了同时在影响因子和预测模型两个方面做出改进, 本文搭建了一个基于 DMR-NB 模型的股指预测框架. 该预测框架具有以下 3 个贡献.

(1) 本文提出的 DMR-NB 模型通过分布式计算和正则化方法从超高维和稀疏性的在线新闻数据快速识别和解释词语信息对股指价值波动的影响.

(2) 利用 TVP 模型对多变的、非线性、非平稳的股指数据进行预测可以揭示影响因素与股指之间的时变关系. 通过高度拟合时变系数状态可以有效提高股指拟合精确度.

(3) 本文提出基于 DNNs 将学习 TVP 模型时变系数状态和股指预测损失函数优化结合起来同时进行. 该方法确保深度学习过程遵循基础经济结构, 在利用 DNNs 卓越的学习能力的同时保留了经济模型的可解释性.

1 基于文本的股指预测框架

本文提出的股指预测框架如图 2 所示, 其涵盖了从收集在线新闻报道到最终评估预测效果的所有阶段. 该框架包含 5 个不同的步骤: 第 1 步是收集与股票指数相关的在线新闻文章; 第 2 步是使用生成语言模型 (如, 多项逻辑回归) 提取新闻文本中影响股指的所有信息, 去除不相关的信息; 在第 3 步, 将保留的词语词频矩阵通过投影聚合为一个低维新闻指数; 第 4 步将得到的聚合指数和历史交易数据应用于预测模型 (如, TVP 模型) 来预测股票指数; 最后使用适当的评估标准评估模型预测性能. 其中第 2–4 步应用的具体方法与算法将在第 1.1–1.5 节中详细阐述.

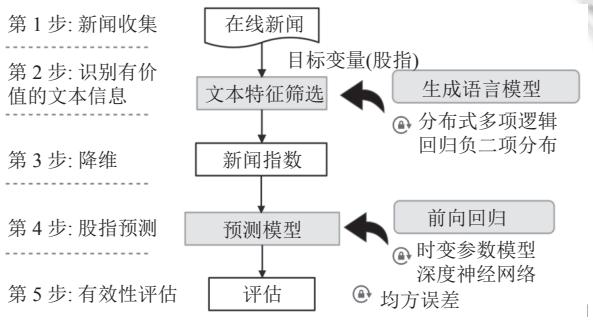

图 2 基于在线新闻的股指预测框架

1.1 有监督生成语言模型

假设拥有原始的文档数据集, 记为$\mathcal{D}$, 并将其映射到一个非负数据矩阵C上. $C\in\mathbb{N}^{n\times d}$中每一行向量$c_i$表示$d$个不同的词语在第$i$篇新闻文档中出现的次数, $c_i=(c_{i1},\cdots,c_{id})$中的每个元素分别对应一个词语. 原始文档集$\mathcal{D}$被划分为$n$个单独的文档集$\{\mathcal{D}_i\}$的标准根据$p$个行业股指$V$的划分时间所确定. 如果$V$是希望从某月的新闻文本中预测的该月最后一日股指价值, 那么按月划分新闻文本是有意义的. 股指v_i在训练集中是可观测的, 而在测试集中是不可观测且需要被预测的. 因此, 在训练集中变量v_i可以被用来指导文本生成.

$$\begin{cases} p(c_i\mid v_i,m_i)=MN(c_i;q_i,m_i), \quad i=1,\cdots,n \\ q_{ij}=\dfrac{\mathrm{e}^{\eta_{ij}}}{\Lambda_i}, \quad \Lambda_i=\displaystyle\sum_{l=1}^{d}\mathrm{e}^{\eta_{il}}, \quad j=1,\cdots,d \\ \eta_{ij}=\alpha_j+v_i^{\top}\varphi_j \\ m_i=\displaystyle\sum_{j=1}^{d}c_{ij} \end{cases} \tag{1}$$

式 (1) 即为有监督生成语言模型, 也称为多项逻辑回归模型.

当C具有大量响应类别时, 由于模型 (1) 需要对每个词语计算Λ_i来保证所有词语生成概率和为 1, 使得模型在参数估计上会消耗大量的时间和空间. 因此, 本文使用可以通过尺度参数调整期望和方差之间的数量关系的负二项分布代替 DMR 模型中的泊松分布, 将多项逻辑分布分解成一系列独立的负二项分布.

$$\begin{aligned} p(c_i\mid v_i,m_i)=MN(c_i;q_i,m_i)&\approx\prod_j NB\left(c_{ij};\xi_j,m_i\mathrm{e}^{\eta_{ij}}\right) \\ &=\prod_j\binom{\frac{1}{\xi_j}+c_{ij}-1}{c_{ij}-1}\left(\frac{1}{1+\xi_j\mu_{ij}}\right)^{\frac{1}{\xi_j}}\left(\frac{\xi_j\mu_{ij}}{1+\xi_j\mu_{ij}}\right)^{c_{ij}} \end{aligned} \tag{2}$$

其中, $\mu_{ij}=m_i\mathrm{e}^{\eta_{ij}}$, ξ_j表示第j个词语对应的负二项分布尺度参数. 式 (2) 即为本文提出的文本特征筛选模型——基于负二项分布的分布式多项回归 (DMR-NB) 模型. DMR-NB 模型假设每一个词语计数向量c_j服从一个m_i给定下的条件负二项分布, 从而将多项逻辑回归分解为d个并行的负二项回归. 因此可以并行地使用负对数似然估计每个词语类别 j 对应的负二项分布参数α_j和φ_j. 该分布式计算方式使得每个词向量可以独立地在不同计算机器上运行, 具有计算可伸缩性. 每个词语对应的负对数似然函数去除常数项后为:

$$\begin{aligned} l\left(\alpha_j,\varphi_j,\xi_j\mid c_j,V\right)=&\sum_{i=1}^{n}\left[\left(\frac{1}{\xi_j}+c_{ij}\right)\log\left(1+\xi_j m_i\mathrm{e}^{\alpha_j+v_i^{\top}\varphi_j}\right)\right] \\ &-\sum_{i=1}^{n}\left[c_{ij}\left(\alpha_j+v_i^{\top}\varphi_j\right)\right] \end{aligned} \tag{3}$$

1.2 文本特征筛选

当可观测变量 V 的维度较高时, 往往存在部分变量只与部分词语相关联. 将所有变量考虑在模型中增加了模型的复杂度. 正则化估计方法往往可以有效降低模型的复杂度. 这种方法通过对系数施加惩罚实现模型性能和复杂性之间的平衡, 从而防止模型过度拟合. 本文为每个词计数的负二项回归应用加权ℓ_1正则化来控制不同变量V前系数的压缩程度.

$$\begin{cases} F(\alpha_j,\varphi_j,\xi_j)=l\left(\alpha_j,\varphi_j,\xi_j\mid c_j,V\right)+n\lambda_j\displaystyle\sum_{k=1}^{p}\omega_{jk}\left|\varphi_{jk}\right| \\ \hat{\alpha}_j,\hat{\varphi}_j,\hat{\xi}_j=\underset{\alpha_j,\varphi_j,\xi_j}{\arg\min}\left\{F(\alpha_j,\varphi_j,\xi_j)\right\} \\ \lambda_j,\omega_{jk}\geqslant 0 \end{cases} \tag{4}$$

关于权重ω_{jk}的选取, 本文采用 Gamma-Lasso 算法来设定权重ω_{jk}的变化路径. 若权重用ω_{jk}表示, 则其每次迭代过程中与$|\hat{\varphi}_{jk}|$成比例下降.

$$\omega_{jk}^{t}=\left(1+\gamma_j\left|\hat{\varphi}_{jk}^{t-1}\right|\right)^{-1}\quad \text{for}\ \ \gamma_j\geqslant 0 \tag{5}$$

显然, 当$\gamma_j=0$时, Gamma-Lasso 算法等同于标准 Lasso 算法, 当$\gamma_j>0$时, Gamma-Lasso 提供了减小偏差正则化效应的方法, 使得强信号比弱信号更不易缩减至零. 最终, 若参数估计$\hat{\varphi}_{jk}=0$则表示词语 j 对于第 k 支行业股指的变动没有影响. 因此根据$\hat{\varphi}_{jk}\neq 0$规则可以筛选出与第 k 支行业股指变动相关的词语特征.

1.3 DMR-NB 模型参数估计

对于模型 (2), 我们需要估计的参数包括α_j、φ_j和ξ_j. 本文提出使用模块更新路径的方法交替更新α_j、φ_j和ξ_j. 估计的α_j、φ_j和ξ_j用$\hat{\alpha}_j$、$\hat{\varphi}_j$和$\hat{\xi}_j$表示. 对于$\hat{\alpha}_j$和$\hat{\varphi}_j$, 通过坐标下降 (coordinate descent, CD) 迭代算法[25]得到, $\hat{\xi}_j$可以使用 Newton-Raphson 算法进行求解. 上述两个迭代过程在交替中重复执行.

在给定$\hat{\xi}_j$的条件下, 基于当前参数值$\hat{\varphi}_j^{t-1}$, φ_j极大似然估计的 Newton-Raphson 更新方式为$\hat{\varphi}_j^{t}=\hat{\varphi}_j^{t-1}-g/H$, 其中 H 是基于$\hat{\varphi}_j^{t-1}$的信息矩阵.

$$-H=\sum_{i=1}^{n}\frac{\mu_{ij}(1+\hat{\xi}_j c_{ij})}{(1+\hat{\xi}_j\mu_{ij})^2}v_i v_i^{\top}=V^{\top}WV$$

而 g 是基于$\hat{\varphi}_j^{t-1}$的系数梯度向量:

$$g=\sum_{i=1}^{n}\frac{v_i(c_{ij}-\mu_{ij})}{1+\hat{\xi}_j\mu_{ij}}=V^{\top}W(\pi-\hat{\eta}_j)$$

其中, $\mu_{ij}=m_i e^{\hat{\eta}_{ij}}$, $\hat{\eta}_{ij}=v_i^{\top}\hat{\varphi}_j^{t-1}$, 以及$\pi=\hat{\eta}_j+\left(c_j-\mu_j\right)\partial\hat{\eta}_j/\partial\mu_j$. 将 g 和 H 代入$\hat{\varphi}_j$的更新方式中可以得到:

$$\hat{\varphi}_j^{t}=\left[V^{\top}WV\right]^{-1}V^{\top}W\pi \tag{6}$$

此时式 (6) 可以被视为如下加上ℓ_1惩罚项的带权最小二乘问题中参数的估计解:

$$\underset{\alpha_j,\varphi_j\in\mathbb{R}}{\arg\min}\sum_i\frac{w_i}{2}\left(\alpha_j+v_i^{\top}\varphi_j-\pi_i\right)^2+n\sum_k\left(1+\gamma_j|\hat{\varphi}_{jk}^{t-1}|\right)^{-1}\lambda_j|\varphi_{jk}|$$

本文提出的 CD 算法具体展示在算法 1 中. 算法 1 只有在第 1 次迭代中才进行所有参数的完整更新, 此后在每次迭代中不会更新非活动$\hat{\varphi}_{jk}$ (即那些值为零的$\hat{\varphi}_{jk}$). 当每次完整迭代的$\hat{\varphi}_{jk}$最大平方变化小于指定的容差阈值时迭代终止. 其中模型 (4) 中的超参数λ_j使用一步估计路径 (path of one-step estimators, POSE) 算法[11]来选择.

算法 1. CD 算法

输入: V, c_j, $\hat{\xi}_j$
输出: $\hat{\alpha}_j$, $\hat{\varphi}_j$

1) 记 $wh_k=\sum w_i(v_{ik}-\bar{v}_k)^2$和 $wv_k=\sum w_i v_{ik}$, 其中$k=1,\cdots,p$;
2) while $\max_{k=1,\cdots,p}wh_k\Delta_k^2>\epsilon$ do
3) for $k=1,\cdots,p$ do
4) 定义$wg_k=-\sum v_{ik}w_i(z_i-\hat{\eta}_i)$;
5) 定义$ghb=wg_k-wh_k\hat{\varphi}_{jk}$;
6) if $|ghb|<n\lambda_j\omega_{jk}$ then
7) $\Delta_k=-\hat{\varphi}_{jk}$;
8) end
9) else
10) $\Delta_k=-\left(wg_k-sign(ghb)n\lambda_j\omega_{jk}\right)/wh_k$;
11) end
12) 更新$\hat{\varphi}_{jk}=\hat{\varphi}_{jk}+\Delta_k$, $\hat{\alpha}_j=\hat{\alpha}_j-wv_k\Delta_k$, wh_k, wv_k, 以及$\hat{\eta}_j=\hat{\alpha}_j+v^{\top}\hat{\varphi}_j$;
13) end
14) end

为了参数$\hat{\alpha}_j$, $\hat{\varphi}_j$和尺度参数$\hat{\xi}_j$同时收敛到局部最优值, 本文提出了一种基于模块交替更新的方法. 算法 2 给出了具体实现. 当某次迭代中模型 (4) 的似然和$\hat{\tau}_j=1/\hat{\xi}_j$的变化都小于预定义的容差水平时停止迭代, 使得模型 (4) 收敛到局部最小值.

算法 2. DMR-NB 模型参数估计流程

1) 初始化$\hat{\tau}_j^1=1$, $\hat{\alpha}_j^1=\hat{\varphi}_j^1=0$;
2) 记负二项分布的负对数似然 (4) 为 l;
3) while $\left|\Delta l/\sqrt{2df}\right|+|\Delta\tau_j|\geqslant\epsilon$ do
4) $gt=\partial l(\tau_j|\hat{\alpha}_j,\hat{\varphi}_j,c_i)/\partial\tau_j$;
5) $ht=\partial^2 l(\tau_j|\hat{\alpha}_j,\hat{\varphi}_j,c_i)/\partial\tau_j^2$;
6) $\Delta\tau_j=-gt/ht$;
7) $\hat{\tau}_j^{m+1}=\hat{\tau}_j^{m}+\Delta\tau_j$;
8) 给定$\hat{\tau}_j$的情形下使用 CD 算法更新$\hat{\alpha}_j$, $\hat{\varphi}_j$;
9) end
10) 返回$\hat{\xi}_j=1/\hat{\tau}_j$, $\hat{\alpha}_j$和 $\hat{\varphi}_j$.

1.4 时变参数预测模型

词频在股指上的投影 ($Z=\hat{\Phi}^{\top}C$) 可以看作是解释事件窗口中行业股指 V 由新闻带来的所有变动方差的低维聚合新闻指数. 考虑到这一点, 本文将聚合新闻指数和历史股票指数作为未来股指预测的驱动因素纳入到 TVP 预测模型中. TVP 模型具体形式为:

$$\theta_{kt}=\theta_{k,t-1}+\zeta_{kt},\quad \zeta_{kt}\sim\mathcal{N}(0,\Omega_k) \tag{7}$$

$$v_{kt}=x_{kt}^{\top}\theta_{kt}+\epsilon_{kt},\quad \epsilon_{kt}\sim\mathcal{N}(0,\sigma_k) \tag{8}$$

其中, $x_{kt}=[1,v_{k,t-1},\log(z_{k,t-1})]^{\top}$是一个 J=3 维向量. $k=1,\cdots,p$. 而随时间变化的回归系数θ_{kt}是一个独立随机

游走, 系数状态更新方差$\Omega_k = \mathrm{diag}(\omega_{k1},\cdots,\omega_{kJ})$是一个$J\times J$大小的对角矩阵, 它控制着回归系数随时间变化的程度, 且系数间的状态变化是相互独立的.

本文使用状态空间模型的非中心参数化来估计时变回归系数θ_{kt}[26]. 将系数θ_{kt}分成两部分:

$$\theta_{kt} = \theta_{k0} + \sqrt{\Omega_k}\,\widetilde{\theta}_{kt} \tag{9}$$

其中, $\widetilde{\theta}_{kt}$满足标准正态随机游走, 即$\widetilde{\theta}_{kt} = \widetilde{\theta}_{k,t-1} + r_{kt}$, $r_{kt}\sim\mathcal{N}(0, I_J)$. 此时可以将时变参数模型 (7) 和 (8) 重写为状态空间模型$v_{kt} = x_{kt}^{\top}\theta_{k0} + x_{kt}^{\top}\sqrt{\Omega_k}\,\widetilde{\theta}_{kt} + \epsilon_{kt}$. 在选择初始条件$\theta_{l0}$的先验分布之后, 参数贝叶斯统计推断可以通过 MCMC 方法直接进行[21]. 它主要在Ω_k和σ_k给定的条件下使用卡尔曼滤波器从状态空间模型中抽样$t = 1,\cdots,T$时刻的$\widetilde{\theta}_{kt}$. 在给定时变参数的完整历史路径$\left\{\widetilde{\theta}_{kt}\right\}_{t=1}^{T}$下, 从高斯后验分布中以块的形式抽样状态更新方差$\Omega_k$、初始状态$\theta_{k0}$和$\sigma_k$. 最后, 通过式 (9) 即可估计$\theta_{kt}$.

1.5 时变参数模型参数预测

为了预测下一个时期的股指价值$v_{k,T+1}$, 首先需要基于历史时期参数预测下一时期 T+1 的参数 $(\theta_{k,T+1})$. 本文开创性地提出 DNNs 实现对系数状态$\theta_{k,T+1}$的预测. 随后, 通过式 (8) 预测下一时期的股指$v_{k,T+1}$. 本文将标准的全连接前馈网络——多层感知器 (multi-layer perceptions, MLPs) 和 TVP 模型合并在一起, 使用历史时变参数来预测未来时间点的时变参数.

图 3 为参数预测结构示意图. 它将 MLPs 用于倒数第 2 个“参数层”来学习系数$\theta_{k,t+1}^{\diamond} = (\theta_{k,t+1,2},\cdots,\theta_{k,t+1,J})^{\top}$, 截距项$\theta_{l,t+1,1}$在时间 t+1 保持恒定, $\hat{\theta}_{l,t+1,1} = \theta_{lt1}$. 学习到的系数在网络的最终“模型层”中放入 TVP模型计算$\hat{v}_{l,t+1}$的平方损失. 该 DNNs 模型通过优化$\hat{v}_{l,t+1}$的损失函数学习$\theta_{k,t+1}^{\diamond}$可以确保深度学过程遵循基础经济结构, 结合了深度学习技术和时间序列模型的双重优势, 在保证模型可解释性的前提下提高股指预测精度.

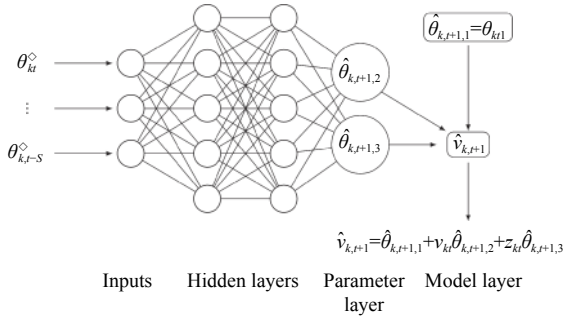

图 3 DNNs 预测时变参数结构示意图

另外, DNNs 输入层中时变参数滞后阶数 S 的选择会影响到预测结果. 本文利用实证数据将超参数 S 在 [1, 5] 范围内搜索选取出最优值. 结果显示当 S=4 时, DNNs 在实证行业股指$v_{k,t+1}$预测中产生了最小均方误差.

2 实验

2.1 数据选择

本研究主要包含两个数据集, 分别为在线新闻数据和行业股指数据. 其中在线新闻数据来源于在中国领先的证券报纸——《中国证券报》, 该报以其权威性和在金融行业中排名前二的发行量而闻名. 由于《中国证券报》覆盖了上海证券交易所和深圳证券交易所的信息, 因此可以提供较为全面的报道. 在行业股指数据方面, 本文选取农林行业, 制造行业, 水电行业, 批发零售行业, 运输仓储行业以及房地产行业 6 个代表行业的行业股指在月末的收盘价 (CP) 对数. 该 6 个行业的代码分别为 399231、399233、399234、399236、399237 和 399241.

本研究收集的数据时间跨度为 2005 年 1 月 1 日–2017 年 7 月 31 日. 共计收集到《中国证券报》在此期间发表的 259 039 篇文章. 同期 6 个行业股指数据可以从网站 http://cn.gtadata.com 处获得.

2.2 数据描述

图 4 展示了《中国证券报》平均每月在线新闻发表数量. 图 4 中显示在 1、3、7、8、11 和 12 月发表文章数量增加, 展现了跨年、假期以及季度收益效应对新闻发表数量的影响. 图 5 提供了《中国证券报》每年发表在线新闻报道的总量. 图中显示, 文章数量在 2004–2007 年间呈现出一个持续上升的趋势, 然后在 2007–2010 年期间经历了一个下降期. 这种波动可以归因于一些因素, 包括 2005 年实施的股票交易改革及其对金融结构市场导向创新的影响, 以及 2008 年的全球金融危机及其对股市的影响.

为了初步展示 6 个行业新闻报道的主要信息, 本文绘制了与 6 个行业相关的新闻报道的词云图, 结果如图 6 所示. 通过对词云图的视觉分析, 可以发现在 6 个行业中最普遍的术语是“经济”和“增长”. 这说明了挖掘包含与经济相关的有价值信息的新闻文章的重要性. 另外, 房地产行业的词云图也显示“开发”“贷款”和“需求”这些与房价直接相关的术语也具有相对较高的出现频数.

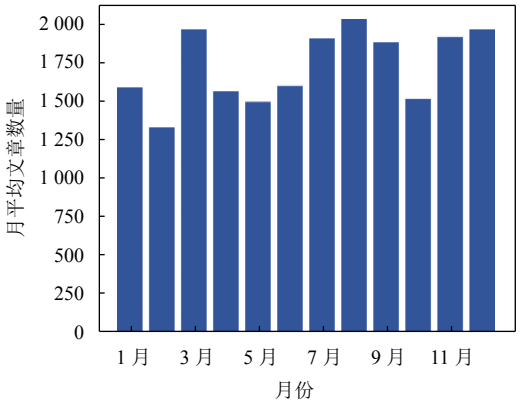

图 4 月平均发布新闻数量统计

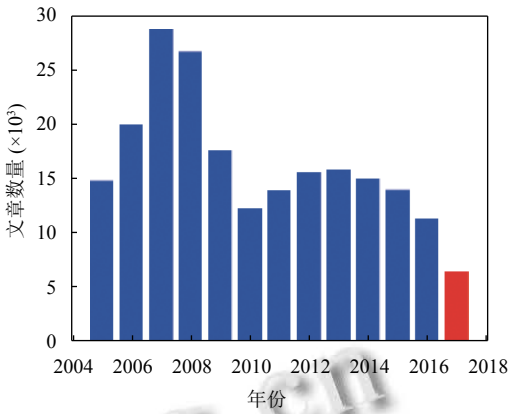

图 5 每年发布新闻总量统计

图 6 6 个行业的新闻词云图

2.3 数据集划分

本文以月为观测单元, 2005 年 1 月 1 日–2017 年 7 月 31 日期间共计 151 个观测单元. 代表着完整两年时间段的最后 24 个观测单元作为测试数据集 (X^{te}, V^{te}), 其余数据则作为训练数据集 (X^{tr}, V^{tr}).

3 实验结果与分析

3.1 实验环境

实验环境: Windows 10 操作系统 16 GB 内存; 开发工具: R 4.0.5 和 Python 3.8.3; CPU: AMD R7-4800H @2.90 GHz. 实验是在一台配备有 AMD R7-4800H CPU 处理器 (8 个内核, 2.90 GHz, 16 MB). 其中基于 R 语言实现 DMR-NB 模型文本特征筛选和 TVP 模型参数估计; Python 3.8.3 实现 DNNs 时变参数参数预测.

3.2 文本词语特征筛选结果分析

基于行业指数及在线新闻报道构建的 DMR-NB 模型可以通过$\hat{\varphi}_{jk} \neq 0$规则筛选股指变动相关词语特征, 并且可以根据$|\hat{\varphi}_{jk}|$的大小反映词语频数与股指变动的相关程度. $|\hat{\varphi}_{jk}|$越大, 说明文档中第 j 个词语出现的频数越大, 股指价格同向变动幅度越大. 表 1 展示了 6 个行业最终通过 DMR-NB 筛选的词语特征数目和$|\hat{\varphi}_{jk}|$最大的 10 个词语. 结果显示 DMR-NB 模型筛选出满足$\hat{\varphi}_{jk} \neq 0$条件的词语数量约占语料库词语数量的 20%, 有效地过滤了与股指变动无关的词语, 实现了高效降维. DMR-NB 模型的解释能力表现在它具有直观和明显的解释词语和股指之间关系的能力. 例如, 运输仓储行业的股票指数可能会受到最近宣布的“强制措施”或“信仰”的显著影响. 农业“销售”的减少可能会导致农业

衰退, 最终体现为股指的下跌.

为了对比 DMR-NB 和 DMR 模型在新闻有效信息提取方面的表现, 本文基于两模型得到的$\hat{\varphi}_k^{NB}$和$\hat{\varphi}_k^{Po}$分别计算了文本矩阵投影, 即新闻指数:

$$\begin{cases} z_{k,t-1}^{NB} = c_{k,t-1}\hat{\varphi}_k^{NB} \\ z_{k,t-1}^{Po} = c_{k,t-1}\hat{\varphi}_k^{Po} \end{cases}$$

其中, $k = 1,\cdots,p$; $p = 6$. 表 2 统计了 6 个行业的新闻指数对数 $\ln(z_k)$ 与股指收盘价对数 ln(CP) 之间的相关系数. 结果显示由 DMR-NB 得到的新闻指数对数对比于 DMR 与行业股票指数之间的相关系数显著更高. 说明 DMR-NB 更能够捕捉与股指变动相关的信息.

3.3 6 种对比预测框架介绍

表 3 展示了 6 种股指预测框架的组成部分. TVP.z.NB 为本文提出的基于 DMR-NB 模型构建的新闻指数和历史股指结构化数据使用 TVP 模型预测未来股指价值的股指预测框架. TVP.z.Po 为基于 DMR模型构建的新闻指数和历史股指结构化数据使用 TVP模型预测未来股指价值的股指预测框架. TVP.nonz 和 LSTM.nonz 表示只基于历史股指结构化数据使用 TVP 模型或者 LSTM 模型预测未来股指价值. TVP.LDA 表示基于 LDA 模型提取的相关主题和历史股指结构化数据使用 TVP 模型进行股指预测. BVAR.LDA 表示基于 LDA 模型提取的相关主题和历史股指结构化数据使用 BVAR 模型进行股指预测.

表 1 6 个行业指数词语特征筛选结果

行业	有效词语数量	$\|\varphi_{jk}\|$最大的前10个词语
农林行业	4 777	“总得”“对冲”“采矿业”“高收”“病毒”“趋淡”“使为”“低档车”“销业”“试图”
制造行业	18 503	“前置”“守则”“谈崩”“美食”“不加区分”“等待”“太”“自然生态”“新思路”“店面”
水电行业	11 828	“角落”“俭朴”“涌现出”“察觉”“国际局势”“采矿”“钩”“预降”“打上”“印花”
批发零售行业	7 633	“皮影”“套保者”“美好”“价年”“护盘”“混合区”“煤球”“近距离”“衡量标准”“掌握”
运输仓储行业	9 325	“强制措施”“葵阳”“煤矸石”“撮合”“新闻纸”“信条”“区人”“横轴”“违规者”“分辨率”
房地产行业	12 684	“赫然”“低价股”“链分析”“赌场”“空格”“紧量”“不败”“扭转”“通货”“旧款”

表 2 聚合新闻指数对数与行业指数对数之间的相关系数

模型	农林	制造	水电	批发零售	运输仓储	房地产
DNR	0.87	0.89	0.87	0.89	0.91	0.87
DMR-NB	0.90	0.90	0.90	0.92	0.92	0.89

表 3 6 种预测框架结构

预测框架	影响因素类型		识别和量化文本信息方法			预测模型
	结构化数据	文本	DMR-NB	DMR	LDA	
TVP.nonz	Π	—	—	—	—	TVP
LSTM.nonz	Π	—	—	—	—	LSTM
TVP.z.Po	Π	Π	—	Π	—	TVP
TVP.z.NB	Π	Π	Π	—	—	TVP
TVP.LDA	Π	Π	—	—	Π	TVP
BVAR.LDA	Π	Π	—	—	Π	BVAR

3.4 样本内拟合效果对比

图 7 展示了测试集中后 24 期基于 6 种不同的股指预测框架得到的$\hat{v}_k$和实际股指v_k的时序图. 结果显示所有由 TVP 模型作为预测模型的预测框架样本内拟合的$\hat{v}_k$与实际股指v_k几乎完全重合. 由 BVAR 和 LSTM 模型拟合的$\hat{v}_k$相比于实际v_k有滞后一阶的问题. 说明 TVP 模型相比于 BVAR 和 LSTM 模型在样本内拟合精度更高, 也表明 TVP 模型中的时变参数估计值$\hat{\Theta}$可以被视为真实Θ来预测未来一期参数.

3.5 样本外预测效果对比

在样本外预测中, 本文基于第 1.5 节所介绍的时变参数预测方法构建了一个包含 3 层隐藏层, 每个隐藏层有 16 个隐藏单元, 激活函数都为 ReLU 函数的 DNNs 来学习训练集中历史 TVP 模型时变参数和未来一期参数的之间的联系. 训练过程中的验证集来源于训练集中随机挑选的 20% 数据. 基于训练完成的 DNNs 可以使用滑动窗口法预测样本外也即测试集中的 24 期 TVP 模型参数. 滑动窗口法以 6 个时期为一个窗口, 每次滑动一个时期, 在每个窗口内分别以经 TVP 模型估计完成的前 5 个时期 $(\hat{\theta}_{k,t-4}^{\diamond},\hat{\theta}_{k,t-3}^{\diamond},\cdots,\hat{\theta}_{kt}^{\diamond})$ 的时变参数作为 DNNs 模型的输入来预测最后一个时期的参数$\theta_{k,t+1}^{\diamond}$, 并通过模型 (8) 预测最后一时期$v_{k,t+1}$.

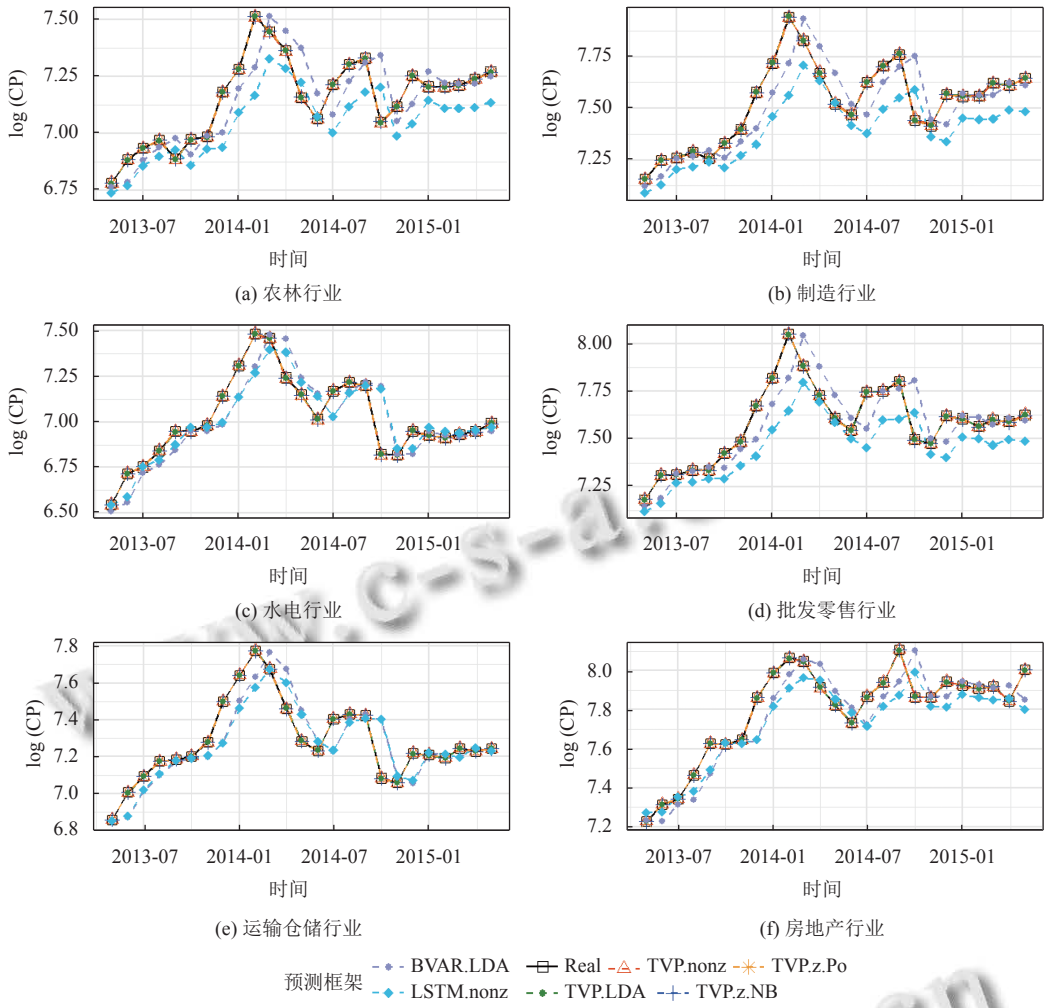

图 7 对数行业指数的样本内拟合值和实际值时序图

为了评估模型对股指的预测性能, 选择均方误差 (mean square error, *MSE*) 作为评价指标来量化模型的性能. *MSE* 计算公式为$MSE_k = \|\hat{v}_k - v_k\|_2^2/(T-4)$, T=24. 对 DNNs 分别进行 100 次实验, 取 100 次结果的平均值, 括号内为标准差, 统计的最终实验结果如表 4 所示. 从表中可以分析出如下 4 点结论.

(1) TVP.z vs. TVP.nonz: 相比于 TVP.nonz 预测框架, 通过 TVP.z.NB 和 TVP.z.Po 将在线新闻信息整合到股票指数预测模型中可以降低预测误差. 这一发现支持了考虑新闻驱动因素对提高股票指数预测准确性具有重要价值的观点.

(2) TVP.z.NB vs. TVP.z.Po: 在分析除运输仓储行业外的 5 个行业的股票指数预测误差时发现 TVP.z.NB 方法比 TVP.z.Po 方法具有更小的均方误差. 我们建议有文本数据访问权限的用户使用 TVP.z.NB 方法进行股票指数预测.

(3) TVP.z.NB vs. TVP.LDA: 在样本外预测方面, TVP.z.NB 通常优于 TVP.LDA. 这表明在一致的 TVP 预测模型条件下, 由生成语言模型筛选的新闻信息用于股票指数预测方面比 LDA 模型聚合的主题更有价值.

(4) TVP.nonz vs. LSTM.nonz: 在 4 个行业股指中, LSTM.nonz 得到的 *MSE* 小于 TVP.nonz. 说明在结构化时序数据中, TVP 模型具有不弱于 LSTM 模型的预测能力.

(5) 最优模型: 相比于其他 5 种预测框架 TVP.z.NB 在性能评估方面具有最好的结果. 本文建议若快速提取新闻文本信息遇到阻碍则可以使用 TVP.nonz 模型. 否则为了减少股指预测误差, 建议使用 TVP.z.NB 预测框架.

表 4 不同预测框架 *MSE* 比较

预测框架	农林行业	制造行业	水电行业	批发零售行业	运输仓储行业	房地产行业
TVP.z.NB	**0.0046**	**0.0053**	**0.0025**	**0.0049**	0.0020	**0.0065**
	(0.0017)	(0.0023)	(0.0012)	(0.0012)	(0.0003)	(0.0019)
TVP.z.Po	0.0060	0.0063	0.0048	0.0059	**0.0019**	0.0215
	(0.0060)	(0.0080)	(0.0056)	(0.0054)	(0.0004)	(0.0155)
TVP.nonz	0.0122	0.0078	0.0063	0.0060	0.0106	0.0218
	(0.0046)	(0.0019)	(0.0018)	(0.0017)	(0.0026)	(0.0071)
TVP.LDA	0.0051	0.0055	0.0098	0.0063	0.0071	0.0066
	(0.0006)	(0.0003)	(0.0029)	(0.0010)	(0.0020)	(0.0019)
BVAR.LDA	0.0100	0.0061	0.0157	0.0073	0.0078	0.0073
	(0.0004)	(0.0002)	(0.0003)	(0.0002)	(0.0006)	(0.0003)
LSTM.nonz	0.0149	0.0176	0.0090	0.0117	0.0090	0.0128
	(0.0059)	(0.0062)	(0.0009)	(0.0043)	(0.0008)	(0.0042)

4 总结

本文从影响股指预测的因子选取和预测模型两个方面入手, 提出了一个新的基于在线新闻的股指预测框架. 首先, 在影响因素选择方面将包含了内在驱动金融市场变动的新闻报道纳入考虑, 并基于 DMR-NB 模型过滤文本中的噪声词语降低文本特征维度, 实现了从在线新闻数据中捕捉与股指变动相关的, 有价值的信息. 相比于自然语言模型, 分布式计算和直观解释词语与股指依赖关系的能力是有监督生成模型的优势. 而实证结果也表明, DMR-NB 在捕捉文本特征方面优于 DMR 和 LDA 模型. 因此, 若收集文本信息便捷时, 使用 DMR-NB 模型来挖掘相关文本特征是可行且有效的. 其次, 在预测模型方面, 选择 TVP 模型作为股指预测模型, 并将深度学习领域的 DNNs 方法用于时变参数预测. 该方法在保留了 TVP 经济模型结构的同时利用 DNNs 的强学习能力. 相比与 LSTM 和 BVAR 模型, 本文提出的 TVP 预测方法在实际股指样本内拟合和样本外预测方面都表现出了它的优势.

本文提出的预测框架使用 DMR-NB 从新闻中筛选的相关特征, 丰富股指预测模型的驱动因素. 然而 DMR-NB 模型在文本挖掘中仍旧存在一些不足, 例如没有考虑词语的位置和语境. 相同的词语在不同的语境或者位置有时会表现不同的含义. 这些不足有可能导致 DMR-NB 模型遗漏或者错会了重要信息. 因此, 考虑将词语的位置信息是有意义的. 虽然自然语言处理是考虑文本位置关系方面的先驱, 但是它不具有生成语言模型的可解释性. 研究如何将文本位置信息和生成语言模型实现融合任重而道远.

参考文献

1 Huang JY, Liu JH. Using social media mining technology to improve stock price forecast accuracy. Journal of Forecasting, 2020, 39(1): 104–116. [doi: 10.1002/for.2616]

2 Gürkaynak RS, Kisacikoğlu B, Wright JH. Missing events in event studies: Identifying the effects of partially measured news surprises. American Economic Review, 2020, 110(12): 3871–3912. [doi: 10.1257/aer.20181470]

3 Foster DP, Liberman M, Stine RA. Featurizing text: Converting text into predictors for regression analysis. Philadelphia: The Wharton School of the University of Pennsylvania, 2013.

4 Kelly B, Manela A, Moreira A. Text selection. Journal of Business & Economic Statistics, 2021, 39(4): 859–879. [doi: 10.1080/07350015.2021.1947843]

5 Sert OC, Şahin SD, Özyer D, *et al*. Analysis and prediction in sparse and high dimensional text data: The case of Dow Jones stock market. Physica A: Statistical Mechanics and its Applications, 2020, 545: 123752. [doi: 10.1016/j.physa.2019.123752]

6 Bai Y, Li XX, Yu H, *et al*. Crude oil price forecasting incorporating news text. International Journal of Forecasting, 2022, 38(1): 367–383. [doi: 10.1016/j.ijforecast.2021.06.006]

7 Ko CR, Chang HT. LSTM-based sentiment analysis for stock price forecast. PeerJ Computer Science, 2021, 7: e408. [doi: 10.7717/peerj-cs.408]

8 Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2017. 6000–6010. [doi: 10.5555/3295222.3295349]

9 Devlin J, Chang MW, Lee K, *et al*. BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis: Association

for Computational Linguistics, 2019. 4171–4186 [doi: 10.18653/v1/N19-1423]

10 Taddy M. Multinomial inverse regression for text analysis. Journal of the American Statistical Association, 2013, 108(503): 755–770. [doi: 10.1080/01621459.2012.734168]

11 Taddy M. One-step estimator paths for concave regularization. Journal of Computational and Graphical Statistics, 2017, 26(3): 525–536. [doi: 10.1080/10618600.2016.1211532]

12 Murphy KP. Machine Learning: A Probabilistic Perspective. Cambridge: MIT Press, 2012. 82–95.

13 Taddy M. Document classification by inversion of distributed language representations. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Beijing: Association for Computational Linguistics, 2015. 45–49. [doi: 10.3115/v1/P15-2008]

14 Taddy M. Distributed multinomial regression. The Annals of Applied Statistics, 2015, 9(3): 1394–1414. [doi: 10.1214/15-AOAS831]

15 Mashadihasanli T. Stock market price forecasting using the Arima model: An application to Istanbul, Turkiye. Journal of Economic Policy Researches, 2022, 9(2): 439–454. [doi: 10.26650/JEPR1056771]

16 Bessler W, Lückoff P. Predicting stock returns with Bayesian vector autoregressive models. Data Analysis, Machine Learning and Applications. Berlin: Springer, 2008. 499–506. [doi: 10.1007/978-3-540-78246-9_59]

17 Larsen VH, Thorsrud LA. The value of news for economic developments. Journal of Econometrics, 2019, 210(1): 203–218. [doi: 10.1016/j.jeconom.2018.11.013]

18 Lei BL, Liu ZD, Song YP. On stock volatility forecasting based on text mining and deep learning under high-frequency data. Journal of Forecasting, 2021, 40(8): 1596–1610. [doi: 10.1002/for.2794]

19 李博, 孟志青, 朱爱花. 时态支持向量机模型在股票操纵模式发现上的研究. 系统科学与数学, 2023, 43(2): 356–378. [doi: 10.12341/jssms22213]

20 王德广, 马恒锐, 梁叶. 基于 ATLG 混合模型的股票价格预测. 计算机系统应用, 2023, 32(3): 171–179. [doi: 10.15888/j.cnki.csa.008964]

21 Hauzenberger N, Pfarrhofer M. Bayesian state-space modeling for analyzing heterogeneous network effects of US monetary policy. The Scandinavian Journal of Economics, 2021, 123(4): 1261–1291. [doi: 10.1111/sjoe.12436]

22 Koop G, Korobilis D. Large time-varying parameter VARs. Journal of Econometrics, 2013, 177(2): 185–198. [doi: 10.1016/j.jeconom.2013.04.007]

23 Farrell MH, Liang TY, Misra S. Deep neural networks for estimation and inference. Econometrica, 2021, 89(1): 181–213. [doi: 10.3982/ECTA16901]

24 Kropotov YA, Proskuryakov AY, Belov AA. Method for forecasting changes in time series parameters in digital information management systems. Computer Optics, 2018, 42(6): 1093–1100. [doi: 10.18287/2412-6179-2018-42-6-1093-1100]

25 Wright SJ. Coordinate descent algorithms. Mathematical Programming, 2015, 151(1): 3–34. [doi: 10.1007/s10107-015-0892-3]

26 Frühwirth-Schnatter S, Wagner H. Stochastic model specification search for Gaussian and partial non-Gaussian state space models. Journal of Econometrics, 2010, 154(4): 85–100. [doi: 10.1016/j.jeconom.2009.07.003]

(校对责编: 牛欣悦)