高维序列数据降维方法在证券市场的应用研究

陈　赛[1],刘文杰[1],黄国耀[1],卢凌峰[1],李华康[2],孙国梓[1*]
(1. 南京邮电大学 计算机学院,江苏 南京 210023;
2. 西交利物浦大学 人工智能与高级计算学院(太仓),江苏 苏州 215123)

摘　要:证券市场数据分析与预测,作为一个经典的大数据分析场景,很多数据挖掘方法已经在本领域得到实际应用。但是鉴于企业本身情况的变化以及证券市场的人为操作等情况,现有的各种大数据挖掘方法无法应对不可见或者未出现的情况,为此论文探索使用易经方法,将其应用在证券市场的数据挖掘和分析预测。利用数据挖掘进行特征筛选、数据降维,通过滑动时间窗、随机森林、三才映射等方法实现传统易经体系中的断卦步骤,将易经概念、规则抽象成算法并对卦辞分类,由解卦算法得出预测值。与先前的预测模型相比,该模型融合易经预测体系与机器学习,充分利用了证券市场的场景特征与历史数据,最终对证券市场平稳、上升、下跌三种发展趋势进行预测。使用10 年内股票证券交易公共数据集进行实验,准确率优于 SVM、XGBoost 等流行的机器学习算法,并在分行业建模中进一步提升了效果。

关键词:数据挖掘;易经;特征筛选;证券预测;机器学习
中图分类号:TP391　　**文献标识码**:A　　**文章编号**:1673-629X(2023)04-0190-08
doi:10.3969/j.issn.1673-629X.2023.04.028

Research on Application of Dimension Reduction Method of High Dimensional Sequence Data in Securities Market

CHEN Sai[1],LIU Wen-jie[1],HUANG Guo-yao[1],LU Ling-feng[1],LI Hua-kang[2],SUN Guo-zi[1*]
(1. School of Computer Science,Nanjing University of Posts and Telecommunications,Nanjing 210023,China;
2. School of Artificial Intelligence and Advanced Computing,XJTLU Entrepreneur College (Taicang),
Suzhou 215123,China)

Abstract:As a classic big data analysis scenario,many data mining methods have been practically applied in the field of securities market data analysis and predicting. However,various existing data mining methods cannot deal with invisible or non-existent situations caused by the company diversification and the human intervention. Therefore,IChing method is used in data mining,analysis and prediction of securities market. Data mining is used for feature selection and data dimension reduction,sliding time window,random forest,three-way mapping and other methods are adopted to realize the steps of hexagrams breaking in the traditional IChing system,and the concepts and rules of IChing are abstracted into an algorithm and the hexagrams are classified,and the predicted values are obtained from the hexagrams breaking algorithm. Compared with the previous prediction model,the proposed model integrates the IChing prediction system and machine learning,makes full use of the scene characteristics and historical data of the securities market,and finally predicts the stable,rising and falling development trends of the securities market. Using the public data set of stock exchange within 10 years for experiment,the accuracy is better than that of SVM,XGBoost and other popular machine learning algorithms,and the effect is further improved in the sub-industry modeling.

Key words:data mining;IChing;feature screening;securities forecasting;machine learning

0　引　言

易经是中国古代人类文明智慧的结晶,其包含了宇宙间万事万物的规律,能引导人类社会的正向发展,易经作为一种高维序列数据的降维方法,降维后能够帮助人们做出合理的预测,趋利避害。如今,有的周易研究者利用这种高维序列数据降维方法来预测股票未来的发展趋势。陈永伟[1]通过研究易经中的四个卦象来指导企业如何面对困境;郭仪等人[2]通过易经中的

收稿日期:2022-11-19　　**修回日期**:2023-02-28
基金项目:国家自然科学基金(61906099);自然资源部城市自然资源监测与仿真重点实验室开放基金(KF-2019-04-065)
作者简介:陈　赛(1996-),男,硕士研究生,研究方向为数据挖掘、自然语言处理;通讯作者:孙国梓(1972-),男,教授,博士后,CCF 会员(12099S),研究方向为计算机取证、区块链安全、数据挖掘。

象数模型来预测股市未来的行情变化。目前利用高维序列数据降维方法实现股市预测的研究存在以下问题:

(1)研究所使用的股票数据不够全面,只研究一个经济指标如收盘价,容易忽略其他重要的经济指标。

(2)由于采用的数据太少,不具有统计意义,实验结果缺少说服力。

(3)高维序列数据降维方法中的重要概念在预测过程中没有很好的进行抽象和应用。

针对以上问题,论文结合大数据和机器学习算法从《易经》中抽取出一套基于高维序列数据降维方法的宏观预测模型来预测股票的发展趋势。高维序列数据降维方法通过卦象和爻的变化,并通过相应的规则结合卦象和爻的解释对股票未来的发展进行一定范围内的预测,引导人们做出正确判断并指导公司进行相应的政策调整。特征选择是构建宏观预测模型的重要一步,特征选择主要是使用机器学习算法从大量的特征中选取出对标签重要的特征。过往的学者用高维序列数据降维方法预测股市行情时都是人为选择经济指标进行预测,这种方法经实验证明效果较差。而论文提出的特征选择方法可以从大量复杂的特征中找到重要性排名较高的一些特征。目前,特征选择方法有很多研究,大都是传统的机器学习算法。黄新等人[3]提出基于变量重要性的偏最小二乘特征筛选法来选择对红外光谱变量起重要作用的指标。刘云翔等人[4]使用随机森林算法筛选出造成肝癌原因的重要因子,实验证明随机森林算法的效果好于决策树。因此如何利用数学思想抽象易经,结合机器学习算法来构建易经宏观模型是论文的核心。

论文首先使用机器学习算法进行特征重要性排名,再借助高维序列数据降维方法中的天地人思想,将众多的特征进行重要性排序并选择最重要的 6 个特征。之后根据动态时间滑动窗口将特征对应的数值转化成四象值,下一步根据高维序列数据降维方法中的阴阳爻的变化规则将四象值进行转化形成本卦和变卦。最后根据高维序列数据降维方法中解卦原则综合本卦和变卦的卦辞得到最终的解卦结果,使用解卦结果和标签值进行对比得到最后的预测结果。

传统机器学习受限于数据集本身特性,普适性较差。而基于高维序列数据降维方法的模型能够根据天地人、四象等思想,并通过卦象和爻的变化来构建出不同场景下的宏观预测模型,并将股价涨跌的预测结果展示给公司。实验结果表明,论文提出的预测模型在股票未来发展的预测效果上要好于对比实验中的其他机器学习方法。因此,基于高维序列数据降维方法构建的宏观模型有实际的研究意义和应用价值。

1　相关工作

证券数据预测是近年来的研究热点,很多学者使用不同的算法对证券进行预测。例如,文献[5]使用不同机器学习方法预测股票未来趋势,实验表明 Adaboost 和贝叶斯网络效果相对较好。文献[6]提出一种基于财务指标和数据挖掘相结合的模型来对股票未来进行预测,实验结果显示,各行业的准确率在 60% 左右。Kannan K S 和 Sekar P S[7] 等提出使用五种方法来挖掘历史交易数据中隐藏的信息并对股票未来发展趋势进行预测,结果表明,该方法的预测准确率大于 50% 。Ou P 和 Wang H[8] 等使用十种不同的数据挖掘技术来预测香港股市恒生指数的价格走势,最终表明 SVM 和 LS-SVM 算法具有更好的预测性能。综上可见,很多算法在证券领域预测趋势任务中表现一般。因此,论文结合数据挖掘技术和高维序列数据降维方法构建一套预测模型。高维序列数据降维方法能够通过卦象实现对证券数据的宏观预测。该方法中的天地人思想结合数据挖掘技术能够筛选出重要的 6 个特征,并通过四象思想将 6 个特征转化成 6 位序列,形成本卦,接着通过变爻思想形成变卦,最后综合本卦和变卦得到最终的卦象结果,实现预测。

1.1　变量重要性评分-随机森林算法

随机森林(Random Forest,RF)是 Breiman 等人于 2001 年提出的[9],至今为止 RF 已经被普遍应用到数据挖掘等领域。RF 具有较高的预测准确率,对于离异值和噪声较多的数据有着非常强的容忍度,可以处理高维数据,能够在分析高维数据的同时,给出不同变量的重要性评分。这些优势让 RF 非常适用于高维数据的研究,在数据挖掘领域有着较高的使用价值。杨明悦和毛献忠[10]通过随机森林算法对水环境中的各个影响因子进行特征重要性评分,最终选取重要的 5 个水质指标用于水环境的评估。肖美丽、晏春丽等人[11]采用随机森林算法通过变量重要性评分对产后抑郁影响因素进行重要程度排序,最终获取排名前 10 的影响因子,并针对影响因子进行定量分析,有效进行产后预诊工作。

由于 RF 在处理高维特征数据时能够很好地对各变量进行重要程度排序,因此论文通过 RF 进行变量筛选工作。

1.2　线性回归

线性回归分析主要用于研究因变量与自变量间的线性关系,通过适当的数学模型将变量间的关系准确表达,进而通过自变量的取值来预测因变量的取值。很多研究中都把线性回归方法用于股票价格预测,通过线性回归方法建立一个预测新股上市第一天开盘价的模型,该模型能够较好地拟合股票价格曲线。苏

晴[12]通过线性回归并结合循环神经网络构建融合模型对股票价格进行预测,实验结果表明预测未来 10 天的股价准确率能达到 70%,预测未来 20 天准确率能达到 60%,但随着预测时间的增长,模型准确率越来越低。

论文通过线性回归拟合股票价格,从而得到股票未来的发展趋势并最终根据趋势构建标签数据。

1.3 高维序列数据降维方法与证券市场

高维序列数据降维方法的预测理论探析的文章中提到,在高维序列数据降维方法的六十四卦中,每个卦象的初爻、二爻对应着地位,三爻、四爻对应着人位、五爻、上爻对应着天位,从卦象中也可以看出人处在天地之中,受限于天地,作用于天地。在一个研究证券市场上的天、地、人的文章中,提出了如下观点:在证券市场的场景下,天时指的是一种经济周期性的波动规律以及国家政策的宏观因素,若该证券所属公司顺着国家政策走,则该证券符合天时;地利指的是当下证券所属公司的基本情况,若该证券的市场流动性很高,则该证券符合人和;从而得出结论,在证券市场的场景下天对应着国家政策调控(经济运行规律),地对应着公司基本面,人对应着市场流动性,且在证券市场的场景下,天地人的重要性等级如下:天>地>人[13]。文献[14]提到易经中有少阳、老阳、少阴和老阴,股票市场当中对应的则是小阳线、大阳线、小阴线和大阴线,大阴线到达一定程度股票价格就会上升,对应易经中的阴极必阳,与易经时刻变化相同,社会经济一直在不停的发展着,有时这种发展还十分显著和迅速,忽起忽落,一盛一衰,成为一种波浪起伏的动荡状态。郭仪等人使用高维序列数据降维方法中的象数模型对股市行情进行预测,其通过收盘价进行起卦,收盘价整数相加除八为上卦,一收盘价小数相加除八为下卦,以整数和小数相加除八为变爻。最终预测股票走势实验发现效果一般。

文献[15-16]等提出使用证券市场中的一些因子作为特征并使用机器学习算法进行证券预测,实验表明算法准确率在 60% 上下浮动。文献[17-18]等提出使用融合 Attention 机制的 LSTM 模型对证券历史数据进行建模和预测,实验结果表明该模型效果好于传统的机器学习。

高维序列数据降维方法能够通过卦象进行预测,并通过卦象的千变万化适应证券市场的变化,从而能够较好地预测证券市场的未来发展趋势。

2 基于高维序列数据降维方法的模型

2.1 模型整体框架

论文提出基于高维序列数据降维的方法,首先通过随机森林算法将大量经济指标进行重要性排名,之后根据天地人思想最终选取 6 个指标。之后通过动态时间窗口和四象映射算法将数据转化成四象值,最后通过四象值得到本卦和变卦,综合本卦和变卦得到最终的卦象结果,实现预测功能,并通过与标签进行对比统计得到预测准确率等相关评价指标的结果。该模型的整体架构如图 1 所示。

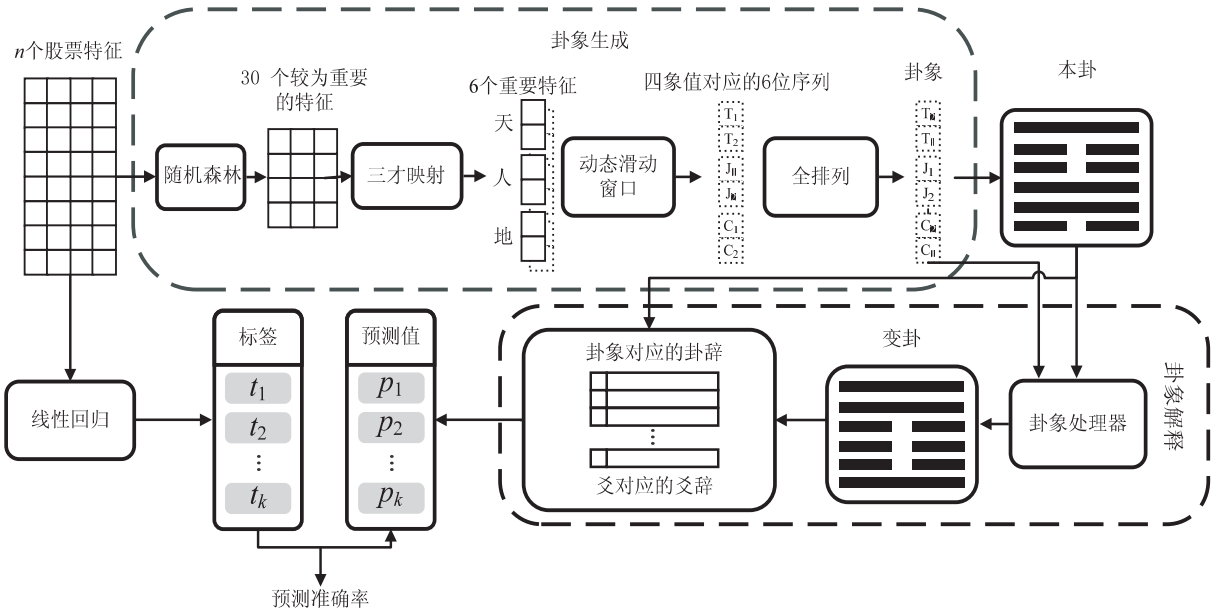

图 1 模型整体架构

如图 1 所示,本模型主要有三大模块,标签构造模块、卦象生成模块和卦象解释模块。标签构造模块主要是通过线性回归的方法将股票价格进行拟合,得到股票未来一段时间内的涨跌趋势;卦象生成模块主要

有以下几个步骤,通过随机森林算法和天地人思想选取特征,并通过滑动窗口将特征对应的值转化为四象值,最终得到本卦;卦象解释模块主要有两个步骤,首先由本卦生成变卦,之后综合本卦和变卦以及对应的爻辞和卦辞得到预测的结果。

2.2　标签构造

由于在股票数据中,存在较多的噪声数据,而 huber regression 具有很好的鲁棒性,对异常的 y 的鲁棒性较强,能够很好地解决数据中的噪声点。论文主要通过 huber regression 拟合收盘价得到收盘价的变化趋势。采用的是一元线性回归,表达式为:$y = ax + b$。y 表示因变量的预测值,x 表示单个自变量,a、b 是回归模型的待定参数,其中 a 又称为回归系数。

huber regression 的损失函数为 huber loss,其计算公式如下:

$$L_\delta(y, f(x)) = \begin{cases} \delta \mid y, f(x) \mid - \dfrac{1}{2}\delta^2 & \text{otherwise} \\ \dfrac{1}{2}(y - f(x))^2 & \text{for} \mid y - f(x) \mid \leqslant \delta \end{cases} \tag{1}$$

其中,δ 为需要调整的超参数。

通过不断的迭代,使得损失值达到最小得到最优拟合函数,获取函数中的回归系数 a 以及偏移量 b。

提出的标签构造算法的具体描述如算法 1 所示:

算法 1:标签构造算法。

输入:每只股票在所选时间段内的所有收盘价数据 price_list。每只股票一个月的收盘价列表 price_list$_i$。

输出:股票各个月的涨跌趋势对应的标签集 month_label_list。

1. for price_list$_i$ in price_list
2. 构建一元线性函数 $y = a_i * x + b_i$
3. a_i, b_i = huber regression(price_list$_i$)
4. a. append(a_i)
5. end for
6. 将 a 中的数据按正态分布划分,设均值 x,方差为 μ
7. for a_i in a
8. if $a_i \in (-\infty, x - \mu)$
9. $a_i = 0$ //表示股价这个月呈下降趋势
10. month_label_list. append(a_i)
11. else if $a_i \in (x - \mu, x + \mu)$
12. $a_i = 1$ //表示股价这个月呈平稳趋势
13. month_label_list. append(a_i)
14. else if $a_i \in (x + \mu, +\infty)$
15. $a_i = 2$ //表示股价这个月呈上涨趋势
16. month_label_list. append(a_i)
17. end if
18. end for
19. return month_label_list

算法 1 描述的是:首先根据线性回归算法拟合每个月的股票价格并计算出一元线性函数中的变量 a_i,循环计算出所选时间段内所有月份的 a_i,之后根据 a 的分布判断每个月涨跌趋势所对应的标签。

2.3　特征筛选

可以比较每个特征在随机森林中每棵分类树上的贡献大小并计算基尼指数,以此判断每个特征的重要程度。节点的基尼指数表示节点的不纯度。以指标的平均基尼减小值占所有指标平均基尼减小值总和的百分比来评估每个风险指标对总风险的贡献程度。

基尼指数计算公式如下:

$$\mathrm{GI}_n = \sum_{k=1}^{N} \hat{p}_{nk}(1 - \hat{p}_{nk}) \tag{2}$$

其中,N 为样本集的类别数,$\hat{p}_{nk}$ 为节点 n 属于第 K 类样本的概率,当二分类时($N = 2$),节点 n 的 Gini 指数为:

$$\mathrm{GI}_n = 2\hat{p}_n(1 - \hat{p}_n) \tag{3}$$

其中,$\hat{p}_n$ 为样本点在节点 n 属于任何类别的概率。

变量 X_j 在节点 n 的重要性,即节点 n 分枝前后 Gini 指数变化量的计算公式如下:

$$V_j^{\mathrm{Gini}} = \mathrm{GI}_n - \mathrm{GI}_l - \mathrm{GI}_r \tag{4}$$

其中,GI_l 和 GI_r 分别表示由节点 n 分裂而成的两个新节点的 Gini 指数。

如果变量 X_j 在第 i 棵数中出现 M 次,则变量 X_j 在第 i 棵树的重要性为:

$$V_{ij}^{\mathrm{Gini}} = \sum_{n=1}^{N} V_{jn}^{\mathrm{Gini}} \tag{5}$$

变量 X_j 在随机森林中的 Gini 重要性定义为:

$$V_j^{\mathrm{Gini}} = \frac{1}{m}\sum_{i=1}^{m} V_{ij}^{\mathrm{Gini}} \tag{6}$$

其中,m 为随机森林中分类树的数量。

通过将每个特征对标签的 Gini 指数的重要性进行排序,最终选取 Gini 指数重要性排名前 30 的特征。

将股票特征的物理含义与地人天思想进行映射和归类,将所有的特征分为地人天 3 类。通过随机森林算法筛选出 30 个特征后,根据特征与地人天类别的对应表,将 30 个特征映射到地人天 3 个类别中。选取每类中重要性排名靠前的 2 个特征,一共筛选出 6 个特征,将筛选后的特征按照地人天的顺序依次排列。

2.4　四象值的计算和卦象的生成

得到 6 个特征之后,按照高维序列数据降维方法的相关理论,需将 6 个特征对应的数值转化成四象对应的值。论文通过动态时间窗的方法实现四象值的计算。

论文提出的四象生成算法的具体描述如算法 2 所示:

算法 2:四象生成算法。

输入:窗口大小 k,四象比例值 old_rate = 1/8, young_rate =

3/8,时序数据列表 vec = $\{v_1, v_2, \cdots, v_n\}$,已排好序的列表 sortedvec,最早添加进列表的元素对应的下标 m。

输出:时序数据列表对应的四象值列 sixiang_list。

1. for i =0→ k −1 do
2. inivec[i] = vec[i]
3. end for
4. 定义四象对应数值,老阴:6,少阴:8,少阳:7,老阳:9
5. for j = k to n −1 do
6. 排序,sortedvec = sort(inivec)
7. maxv = max(sortedvec) ,minv = min(sortedvec)
8. cha = maxv−minv
9. A = old_rate * cha, B = young_rate * cha
10. laoyin = [minv, minv + A) , shaoyin = [minv + A, min + A + B) , shaoyang = [min+A+B, min+A+2B) , laoyang = [min+A+2B, maxv)
11. if vec[j] ∈ laoyin
12. sixiang_list[$j - k$] = 6
13. else if vec[j] ∈ shaoyin
14. sixiang_list[$j - k$] = 8
15. else if vec[j] ∈ shaoyang
16. sixiang_list[$j - k$] = 7
17. else
18. sixiang_list[$j - k$] = 9
19. del sortedvec[m], sortedvec[k] = vec[j]
20. end if
21. end for
22. returnsixiang_list

卦由爻组成,根据高维序列数据降维方法的思想,定义阳爻为 1,阴爻为 0。得到四象值后,将四象值与 1 和 0 进行映射,数值 6 和 8 代表阴爻,7 和 9 为阳爻,随着时间窗口的不断滑动,得到每只股票在每个季度点对应的 6 个特征组成的 6 位 0 和 1 的序列,将此 6 位 0 和 1 的序列按照地人天的排序得到本卦。

根据阳极必阴和阴极必阳的思想,此高维序列数据降维方法中存在变卦现象,当一个卦象的 6 个爻中存在老阳爻和老阴爻时会发生变化,即老阳爻转化而成的 1 会变成 0,老阴爻则相反。最终根据本卦中存在的变爻形成变卦。

2.5 高维序列数据降维方法的解卦算法

定义高维序列数据降维方法中 64 个卦象对应的类别以及每个卦象中每个爻对应的类别。

《易经》有六十四卦及每个卦对应的卦辞,三百八十四爻及每个爻对应的爻辞,每个卦辞都与该卦的卦象紧密关联;每一卦的爻辞也都与其对应的阴爻或阳爻在其卦中所处的位置有关系。在形成本卦和变卦之后需要对卦象进行解卦,论文解卦方法由此高维序列数据降维方法中解卦思想抽象而来。

提出的解卦算法的具体描述如算法 3 所示。

算法 3:解卦算法。

输入:本卦 $Orig_i$,变卦 $Change_i$,变量 $flag_i = 0$,爻变个数 $YbNum_i$,64 个卦象和卦辞映射字典 Gua_dict 以及 384 个爻和爻辞映射字典 Yao_dict,$Orig_i$中的变爻 StaY,$Orig_i$中处于高位的变爻 Hp_StaY,$Orig_i$中处于低位的不变爻 Lp_UstaY,$Orig_i$ 中的不变爻 UstaY。

输出:本卦和变卦对应的解卦结果列表 res_i。

1. for k =0→5 do
2. if $Orig_i$[k] ! = $Change_i$[k]
3. $flag_i$ = 1, $YbNum_i$ = $YbNum_i$ +1
4. end if
5. end for
6. if $flag_i$ = 0
7. res_i = Gua_dict_i[$Orig_i$]
8. end if
9. else
10. if $YbNum_i$ = 1
11. res_i = Yao_dict[StaY]
12. else if $YbNum_i$ = 2
13. res_i = Yao_dict[Hp_StaY]
14. else if $YbNum_i$ = 3
15. res_i = Gua_dict[$Orig_i$]
16. else if $YbNum_i$ = 4
17. res_i = Yao_dict[Lp_UstaY]
18. else if $YbNum_i$ = 5
19. res_i = Yao_dict[UstaY]
20. else
21. res_i = Gua_dict[$Change_i$]
22. end if
23. return res_i

地人天内部的爻位排序不确定,需要分别在地人天对应的两个爻进行排列,共 2 * 2 * 2 = 8 种排列组合。训练模型,每次选择 8 种排列中的 1 种六位序列,将对应的卦的解卦结果作为预测值,与真实标签值做比较,得到每种排列下的准确率和 F1 值,通过 8 种情况下的比对结果,找到 F1 值最高的那种排序作为最后排序进而构建出最后的宏观预测模型。

3 实验结果与分析

3.1 实验数据集

实验所使用的证券数据是网易财经网站上爬取的 3 000 只股票 2010 年 3 月至 2020 年 3 月这 10 年内的历史交易数据集。该数据集包含如下两部分:(1)财务数据集;(2)资金流向数据集。其中财务数据集是以季度为单位的数据集,记录了每个季度的公司净利润率、负债率等反映公司业绩的数据,资金流向数据集是以天为单位的数据,记录了每天的公司开盘价、收盘价、换手率等资金流向的数据。所有的股票分为 10 个行业。实验根据近一个季度的历史数据来预测未来一年内的发展趋势。

论文使用的主要参数有时间滑动窗口 win_len,由于论文所述的高维序列数据降维方法中四象的分布比例为 1∶3∶3∶1,所以此处将 win_len 设置为 8。

3.2　评价指标

论文研究的分类问题常用的评价指标包括结果的精确率(Precision)、召回率(Recall)和 F1 值。

精确率 P 指所有预测正确的数量占总量的比例,计算公式如下:

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{7}$$

召回率 R 指正确预测为正的占全部实际为正的比例,计算公式如下:

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{8}$$

F1 值是一个综合了 P 和 R 的指标,是基于 P 和 R 的加权调和平均,计算公式如下:

$$\mathrm{F1} = \frac{2PR}{P + R} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{9}$$

其中,TP 表示被模型预测为正类的正样本,FP 表示被模型预测为正类的负样本,FN 表示被模型预测为负类的正样本。使用 F1 值来评价分类器性能时,其值越高,说明分类器的性能越好。

3.3　实验结果

为验证模型的优越性,主要进行了两组对比实验,第一组是未分行业模型的各评价指标与其他算法进行对比验证本模型的优越性。以下介绍对比模型的输入和输出。

SVM:输入数据为易经模型选取的 6 个特征对应的数据。输出是分类的准确率、召回率和 F1 值。

XGBoost:输入数据为易经模型从原始特征中选取的 6 个特征对应的每个季度的数据。输出是分类的准确率、召回率和 F1 值。

RF:输入为易经模型从原始特征中选取的 6 个特征对应的每个季度的数据。输出是分类的准确率、召回率和 F1 值。

KNN:输入数据为易经模型从原始特征中选取的 6 个特征对应的每个季度的数据。输出是分类的准确率、召回率和 F1 值。

GRU:序列数据通常使用循环神经网络建模,由于 GRU 参数少、速度快,能在一定程度上避免出现过拟合,因此本实验采用其作为对比模型。输入数据与机器学习算法一致。

Lstm-Att:在 LSTM 基础上加入 Attention 机制,对于重要特征基于更高权重。输入数据与机器学习算法一致。

第二组是分行业建模后本模型与其他模型的对比以及与未分行业的结果对比。结果如表 1 所示。

表 1　各模型股票预测的性能对比

Model	P/%	R/%	F1/%
GRU	60.15	60.71	60.43
Lstm-Att	61.85	60.33	61.08
SVM	45.51	67.46	54.35
XGBoost	46.31	68.05	55.11
RF	46.46	68.16	55.26
KNN	52.00	56.77	54.05
Ours(IChing)	65.41	61.51	63.33

从表 1 中可以看出,在本数据集中,论文提出的方法要优于其他算法。在传统的机器学习算法中,XGBoost 和随机森林算法的 F1 要稍优于 SVM 和 KNN 算法。KNN 算法的准确率要优于其他三个机器学习算法。两个循环神经网络模型的效果要优于传统机器学习算法,表明在证券数据此种序列数据上循环神经网络效果较好。原因在于循环神经网络能够捕捉到不同时刻数据之间的依赖信息,能够提取到上一时刻的重要信息,因此 GRU 和 Lstm-Att 模型效果好于传统的机器学习,除此之外,Lstm-Att 模型的 F1 值稍高于 GRU,因为在加入 Attention 机制后,模型能够加权学习,对于重要信息给予更大权重,从而提高预测效果。

论文提出的基于易经的宏观预测方法在 3 000 只股票数据集上除 recall 外各评价指标都要高于其他方法,P 值比 SVM 高 20 百分点,比 RF 高 19 百分点,比 KNN 高 13.4 百分点,论文提出方法的 F1 值均高于其他模型且本模型各评价指标之间相差较小,相对 SVM、XGBoost 和 RF 模型各指标要更加稳定。

因此,基于高维序列数据降维方法和数据挖掘技术的证券预测模型的效果要好于其他算法。在宏观预测方面,本模型比传统的机器学习方法效果更好,主要在于高维序列数据降维方法能够根据每个公司不同的经营状况以及市场流动性进行降维并选取不同的指标作为基础来进行预测,能够很好地适应证券市场的变化,因此在预测未来发展趋势方面有很大优势。

由于不同行业的公司发展速度和行业规律不同,接下来将分行业建模,验证分行业后的模型和不分行业的模型的性能对比。结果如表 2 和表 3 所示。

从表 2 可以看出,在分行业建模后,大多数行业的 P 值都比总体建模的 P 值高,且每个行业 P 值的提升幅度不稳定,比如提升幅度较低的行业如化学制品行业的 P 值比总体的 P 值高 4.7 百分点,提升较高的行业如交通物流行业比总体的 P 值高 13.4 百分点,最高的如水电燃气行业 P 为 84.56,比总体建模 F1 高了将近 20 百分点。另外,从表中还可看出,9 个行业中有 7

表 2 本模型分行业和总体建模的实验结果

Industry	*P*/%	*R*/%	F1/%
总体(不分行业)	65.41	61.51	63.33
交通物流	78.81	68.65	73.20
信息技术	56.64	58.46	57.53
化学制品	70.14	63.36	66.43
建筑业	74.15	67.13	70.22
房地产业	80.49	68.87	73.99
水电燃气	84.56	70.84	76.94
通信设备	58.99	58.45	58.72
通用设备制造	71.80	65.09	68.17
采矿业	79.99	68.73	73.69

个行业的各评价标准都比总体建模高,只有两个行业:专用设备制造行业和通信设备行业的效果比总体建模效果差。从表 3 中可得出,提出的模型的 F1 值要优于其他对比模型。

分析表 2 和表 3 的数据可得出如下结论:每个行业的行业规律和特点是不同的,而高维序列数据降维方法能够根据每个行业不同的特点选取对该行业重要的指标作为构建易经卦象的爻,因此得到的卦象结果也就更加准确。而有两个行业效果差于总体建模的原因则是因为参数 win_len 设置的值不是最优值,因为信息技术行业发展速度较快,在短短几个月内该行业的发展情况就会有大变化,并且作为现在备受关注的行业,会受到国家政策和新闻导向等更多不确定的宏观因素影响,因此在预测该行业几个月后的发展趋势时不确定性高,导致最终的预测准确率较低。

表 3 各模型分行业建模的 F1 值对比结果

行业	GRU	Lstm-Att	SVM	XGBoost	RF	KNN	IChing
总体	60.43	61.08	54.35	55.11	55.26	54.05	63.33
交通物流	63.91	64.55	62.81	61.17	60.4	60.81	73.2
信息技术	44.31	43.26	36.93	30.5	36.1	37.45	57.53
化学制品	62.01	60.02	46.38	47.2	45.22	49.15	66.43
建筑业	58.44	59.31	52.93	55.24	57.53	50.22	70.22
房地产业	63.65	65.16	58.73	58.01	58.53	57.13	73.99
水电燃气	65.51	66.58	64.87	65.96	65.51	64.66	76.94
通信设备	50.03	47.11	34.51	34.88	36.06	41.34	58.72
通用设备制造	48.82	48.92	45.8	44.51	45.31	46.12	68.17
采矿业	66.76	64.78	61.34	60.13	60.28	59.62	73.69

图 2、图 4 直观地显示了表 1 和分行业后各模型对于不同行业的 F1 值的实验对比结果,图 3 显示了本模型在整体建模下模型预测时长与模型预测准确率的变化关系。

从图 2 中可看出,论文所提出模型的效果明显优于所列出的其他传统机器学习算法,精确率和 F1 值比其他算法均高。因此,论文提出的算法在证券市场宏观预测上表现出了较好的效果,除此之外,从图 3 中可看出,论文提出的模型在整体建模下,模型预测时长在一年内随着时长的增加准确率不断提升,当预测时长为 12 个月即一年时,准确率不再提升,模型收敛,此时预测准确率最高为 62% 。

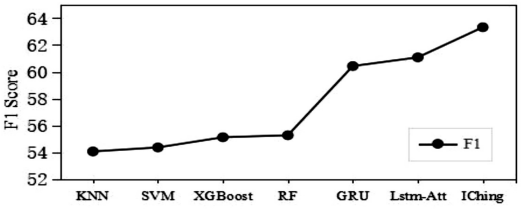

图 2 各模型预测结果的性能对比

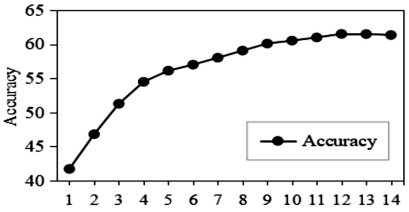

图 3 不同预测时长的准确率

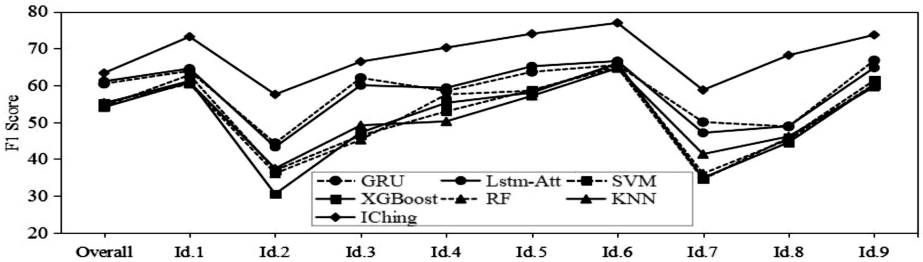

图 4 各行业实验结果 F1 值对比

从图 4 中可看出,论文提出的模型在分行业后大多数行业预测的 F1 值均大于未分行业时的性能,然而其他机器学习算法则不然,大部分行业与总体建模时的指标结果相比无明显提高。因此本模型对于不同行业的证券的宏观调控具有很好的指导性和实用性。

4　结束语

此高维序列数据降维方法中的六十四卦的理论在股市技术分析中虽已有不少的体现,但都没有结合大数据和数据挖掘技术对股票各方面的经济指标进行分析和建模。论文结合机器学习算法针对股票数据构建了一套基于高维序列数据降维方法的宏观预测模型,该模型将阴阳、四象、爻、卦象等概念进行提取和抽象,首先根据天地人思想并结合机器学习算法筛选得到 6 个特征,之后使用动态滑动窗口计算得到四象值,然后根据四象值得到本卦和变卦,最后结合本卦和变卦得到解卦值并与标签进行对比,从而得到最终的预测结果。实验表明,本模型比 SVM、XGBoost、RF、KNN、GRU 和 Lstm-Att 等模型效果更好。另外,由于不同的行业其发展周期和指标的重要性不同,因此论文还对股票数据进行分行业建模,实验证明,论文提出的模型在分行业建模后的效果比总体建模效果好。

但是,论文提出的模型中也存在一些问题:(1)论文在形成四象值时采用的动态时间窗口大小是固定的,然而不同行业的公司有不同的发展周期,因此在形成四象值时,应该采取不同的时间窗口大小;(2)论文没有使用证券新闻等文本数据,现实中,新闻信息能侧面看出证券公司的经营状况以及公司和国家出台的经济政策,而这些信息对于证券的未来发展趋势具有一定的影响;(3)论文只研究了证券未来涨跌趋势,而没有研究证券未来涨跌的原因。因此,下一步的研究方向在于:针对不同的行业采取的时间窗口大小应设置不同的值;使用文本数据作为预测的辅助工具;当预测到证券未来发展情况不利时,要能根据模型判断出导致证券发展不利的原因,并根据原因指导公司调整经营策略,避免危机,从而使本模型达到诊断和预警的作用。

参考文献:

[1] 陈永伟. 企业如何面对困境? ——来自《易经》"四大难卦"的启示[J]. 清华管理评论,2021(6):29-37.

[2] 郭　仪,鲁　珩.《易经》象数模型在股市行情预测中的应用探析[J]. 现代营销:下旬刊,2020(3):46-47.

[3] 黄　新,刘伟平. 基于变量重要性和偏最小二乘的近红外特征筛选方法研究[J]. 湖南城市学院学报:自然科学版,2021,30(6):50-54.

[4] 刘云翔,陈　斌,周子宜. 一种基于随机森林的改进特征筛选算法[J]. 现代电子技术,2019,42(12):117-121.

[5] GHAZANFAR M A. Using machine learning classifiers to predict stock exchange index[J]. International Journal of Machine Learning and Computing,2017,7(2):24-29.

[6] YONG H, FENG B, ZHANG X. A prediction model for stock market:a comparison of the world's top investors with data mining method[C]//The 12th Wuhan international conference on e-business. Wuhan:[s. n.],2013:107.

[7] KANNAN K S, SEKAR P S, MOHAMED S M. Financial stock market forecast using data mining techniques[J]. Lecture Notes in Engineering & Computer Science, 2010, 2180(1):4.

[8] OU P, WANG H. Prediction of stock market index movement by ten data mining techniques[J]. Modern Applied Science, 2009,3(12):28-42.

[9] BREIMAN L. Random forests[J]. Machine Learning,2001, 45(1):5-32.

[10] 杨明悦,毛献忠. 基于特征重要性评分-随机森林的溶解氧预测模型及其在深圳湾的应用[J]. 中国环境科学,2022, 42(8):3876-3881.

[11] 肖美丽,晏春丽,付　冰,等. 随机森林算法在产后抑郁风险预测中的应用[J]. 中南大学学报:医学版,2020,45(10):1215-1222.

[12] 苏　晴. 基于循环神经网络和回归分析的证券价格区间值预测与应用研究[D]. 郑州:郑州大学,2021.

[13] 刘海啸. 股票市场上的天时、地利、人和[J]. 燕山大学学报:哲学社会科学版,2007,8(2):111-117.

[14] 刘　铁.《易经》中的股票投资智慧[J]. 中外企业家,2017(6):268.

[15] ZHAN Y, YU X H. Research on macro influencing factors based on stock market stability[C]//2019 annual meeting on management engineering (AMME 2019). New York: ACM,2019:63-70.

[16] NTAKARIS A, MIRONE G, KANNIAINEN J. Feature engineering for mid-price prediction with deep learning[J]. arXiv:1904.05384,2019.

[17] ZHENG H Y, ZHOU Z Q, CHEN J Y. RLSTM: A new framework of stock prediction by using random noise for overfitting prevention[J]. Computational Intelligence and Neuroscience,2021,2021(3):1-14.

[18] LI Y, ZHU Z, KONG D, et al. Ea-lstm: evolutionary attention-based lstm for time series prediction[J]. Knowledge-Based Systems, 2019, 181: 104785. 1-104785. 8.