DOI:10.3969/j.issn.1003-5060.2024.10.006

基于情绪向量的隐半马尔可夫模型股市拐点预测方法

姚宏亮，　江永生，　杨　静，　俞　奎
(合肥工业大学 计算机与信息学院,安徽 合肥　230601)

摘　要:股市的情绪化倾向是股票市场具有高度不确定性的主要原因,直接利用历史数据的股票趋势预测方法难以适应市场情绪的多变性,在实际应用中效果不理想。文章针对市场情绪的不稳定性导致股市拐点难以预测的问题,提出一种基于情绪向量的隐半马尔可夫模型股市拐点预测方法(hidden semi-Markov model stock turning point prediction method based on sentiment vector,SV-HSMM)。针对市场情绪不可观察性,选取与市场情绪相关的主要特征,使用马尔可夫毯融合成市场情绪;利用隐半马尔可夫模型建模市场环境,构建市场情绪、市场状态和状态持续时间之间的结构关系;引入情绪向量平滑情绪的多变性,并利用 Kullback-Leibler(KL)距离量化情绪热度;利用隐半马尔可夫模型的动态推理实现股市拐点预测。结果表明情绪向量方法具有更好的预测效果。

关键词:市场情绪;情绪向量;隐半马尔可夫模型(HSMM);Kullback-Leibler(KL)距离
中图分类号:TP181;TP202　　**文献标志码**:A　　**文章编号**:1003-5060(2024)10-1335-06

A method for stock turning point prediction with hidden semi-Markov model based on sentiment vector

YAO Hongliang,　JIANG Yongsheng,　YANG Jing,　YU Kui
(School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230601, China)

Abstract: The sentimental stock market is the main reason for the high degree of uncertainty in the market trend. The stock trend prediction method using historical data directly is difficult to adapt to the variability of market sentiment, and the effect is not ideal in practical application. Aiming at the problem that it is difficult to predict the turning point of stock market due to the instability of market sentiment, a hidden semi-Markov model stock turning point prediction method based on sentiment vector(SV-HSMM) is proposed. Firstly, for market sentiment is not observable, the main features related to market sentiment are selected and fused into market sentiment by Markov blanket. Secondly, the HSMM is used to model the market environment, and the structural relations among market sentiment, market state and state duration are constructed. Furthermore, sentiment vector is introduced to smooth the variability of sentiment, and Kullback-Leibler(KL) distance is used to quantify sentiment heat. Finally, the dynamic inference of HSMM is used to predict the turning point of stock market. Experimental results show that the sentiment vector method has better prediction effect.

Key words: market sentiment; sentiment vector; hidden semi-Markov model(HSMM); Kullback-Leibler(KL) distance

股市是一个复杂的非线性动态系统,容易受到各种因素的影响,情绪化特征比较明显,使股市走势具有高度的不确定性。市场情绪是影响市场多种因素的一种综合表现,情绪本身存在非理性

收稿日期:2022-03-03
基金项目:国家重点研发计划资助项目(2020AAA0106100);国家自然科学基金资助项目(62187620;62176082)
作者简介:姚宏亮(1972—),男,安徽桐城人,博士,合肥工业大学副教授,硕士生导师;
俞　奎(1979—),男,安徽合肥人,博士,合肥工业大学教授,博士生导师.

的一面,市场情绪化会进一步提升股市波动的不确定性,导致基于历史交易数据的股市预测模型难以适应市场情绪变化的不确定性。因此,股市趋势预测一直是机器学习领域的一个挑战性难题。

拐点发现是股市趋势预测中的关键问题,学者们对其进行了研究。文献[1]将历史数据直接用分段线性表示方法分解为不同的段,结合分段线性表示与反向传播神经网络模型(piecewise linear representations & back propagation network,PLR-BPN)对股市的拐点进行预测;文献[2]针对PLR-BPN过度拟合、陷入局部最小值的缺点,提出融合加权支持向量机(weighted support vector machine,WSVM)的PLR-WSVM模型预测股市拐点;文献[3]提出一种适应度函数自动选择PLR阈值和过采样拐点改进的PLR-WSVM;文献[4]在缠论基础上进行拐点标注,构建基于深度学习的拐点预测模型。由于股市波动具有高度不确定性,传统的直接利用历史数据进行拐点预测的方法难以适应情绪化的市场环境。

股市中投资者具有情绪化倾向[5],特别是在中国股市,以散户为主的投资者结构,更易受到情绪影响[6]。因此学者们从情绪角度研究股市趋势预测,主要分析股市走势与市场情绪存在高度相关性,表明市场走势可以由市场情绪解释[7]。文献[8]提取股票论坛中的情绪因素和股市历史数据中的技术指标,运用卷积神经网络对隐藏的情绪进行分类,利用长短期记忆神经网络(long short term memory network,LSTM)进行收盘价预测;文献[9]构建宏观和微观混合投资者情绪,使用经验模式分解确定预测范围,但仅考虑了各个情绪指标,市场情绪的整体表示能力弱,算法的适应性不强;文献[10]讨论市场情绪对股市走势的影响程度,表明市场情绪与股市走势具有联动性;文献[11]在输入变量中增加代表投资者情绪的技术指标训练LSTM预测股市,但未考虑市场情绪在时间上的演化,导致模型预测效果一般。

隐半马尔可夫模型(hidden semi-Markov model,HSMM)是一种含隐变量的动态贝叶斯网络,已成为异常检测和自动推理领域中的一种有力工具。文献[12]利用HSMM挖掘网络流量特征中最大似然概率的分段模式,基于序列相的似然概率检测网络中的异常攻击行为;文献[13]采用HSMM对设备退化过程时间序列进行动态建模,根据训练的模型预测设备退化状态。研究表明HSMM对隐变量及持续时间建模,能更有效地表示复杂建模场景中的结构关系。在股市数据中,由于市场情绪不稳定且交易数据中噪声较多,将利用滑动窗口提出情绪向量对数据进行整合,以提升融噪能力。

针对市场情绪的动态多变,本文提出一种基于情绪向量的隐半马尔可夫模型股市拐点预测方法(hidden semi-Markov model stock turning point prediction method based on sentiment vector,SV-HSMM)。首先对主要情绪特征利用马尔可夫毯进行特征融合,提高市场情绪表达能力;构建市场情绪的HSMM,建模情绪作用下市场环境的动态变化;为了提高模型的融噪和稳定表示市场情绪的能力,引入情绪向量,并利用Kullback-Leibler(KL)距离进行情绪热度度量;最后根据SV-HSMM对股市拐点进行推理预测。

1 基于马尔可夫毯的情绪特征融合

1.1 情绪特征离散化

市场情绪不能直接观测,是通过相关的特征表现出来的。将与市场情绪显著相关特征称为情绪特征。选取成交量X_1、涨幅超5%个股数量X_2、跌幅超5%个股数量X_3、涨跌家数比X_4、龙虎榜上榜股票数量X_5、龙虎榜买入金额与卖出金额比X_6 6个指标作为市场情绪的主要表现特征。根据建模的需要,利用均值和方差法对特征的状态进行离散化,将每个特征的取值都离散化为{1,0,−1}。特征离散计算公式如下:

$$X_i=\begin{cases}1, & X_i>E(X_i)+\alpha H(X_i);\\ 0, & E(X_i)+\alpha H(X_i)>X_i>\\ & E(X_i)-\alpha H(X_i);\\ -1, & X_i<E(X_i)+\alpha H(X_i)\end{cases}\tag{1}$$

其中:$i=1,2,3,4,5,6$;$E(X_i)$为特征X_i的平均值;$H(X_i)$为X_i的标准差;α为调节$H(X_i)$倍数的参数,这里取1。

1.2 情绪特征融合

定义1(马尔可夫毯) 在一个贝叶斯网络中,一个结点X_i的马尔可夫毯是由X_i的父结点、子结点和子结点的父结点组成的结点集。结点X_i与贝叶斯网中所有非马尔可夫毯结点条件独立,结点X_i的马尔可夫毯记作$M(X_i)$,则有:

$$P(X_i\mid M(X_i))=P(X_i\mid Z,M(X_i))\tag{2}$$

其中,Z为贝叶斯网络中的所有非马尔可夫毯结点。

用 S 表示市场情绪，$M(S)$ 为市场情绪的马尔可夫毯，满足 $X_i \in M(S)$，利用条件概率 $P(S|X_1,X_2,\cdots,X_n)$ 对市场情绪进行计算。条件概率 $P(S|X_1,X_2,\cdots,X_n)$ 表示当前情绪特征组合为 $X_1,X_2,\cdots,X_n$ 时市场情绪 S 的概率，即

$$P(S=s^j \mid X_1,X_2,\cdots,X_n)=\frac{P(X_1,X_2,\cdots,X_n \mid S=s^j)P(S=s^j)}{P(X_1,X_2,\cdots,X_n)} \tag{3}$$

其中：$j=1,2,3$；s^1、s^2、s^3 分别为市场情绪热度的 3 种状态。

2　市场情绪结构建模

2.1　隐半马尔可夫模型

HSMM 是在隐马尔可夫模型(hidden Markov model，HMM)基础上进行扩展的特殊概率图模型，HSMM 为了更有效地描述隐状态转移关系，在 HMM 基础上增加了状态驻留时间，因此 HSMM 能够更客观地描述网络状态。

一个典型的 HSMM 模型[14]是一个四元组 $(\boldsymbol{\pi},\boldsymbol{A},\boldsymbol{B},P_i(d))$，其中：$\boldsymbol{\pi}=(\pi_1,\pi_2,\pi_3)$ 为初始概率分布矢量，$\pi_i=P(S_1=s^i)$，$1\leqslant i\leqslant 3$，S_1 为初始时刻的状态变量；$\boldsymbol{A}=(a_{ij})_{3\times 3}$ 为状态转移概率矩阵，$a_{ij}=P(S_{t+1}=s^j \mid S_t=s^i)$；$1\leqslant i,j\leqslant 3$；$\boldsymbol{B}=(b_{jk})_{3\times M}$ 为观测值概率矩阵，$b_{jk}=P(O_t=o^k \mid S_t=s^j)$，$1\leqslant j\leqslant 3$，$1\leqslant k\leqslant M$，$M$ 为观测值的状态数，O_t 为 t 时刻观测变量；$P_i(d)=P(D_t=d_{s,t} \mid S_t=s^i)$ 为状态持续时间分布，D_t 为 t 时刻的持续时间变量，$d_{s,t}$ 为 t 时刻状态 s^i 的持续时间。

为了更直观地表示状态转移过程，观测变量、隐变量和持续时间变量的序列如图 1 所示。

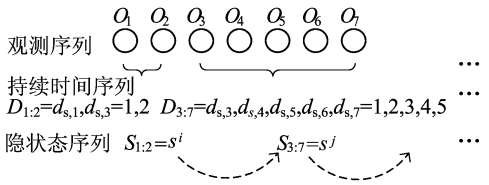

图 1　HSMM 模型

为了简化模型，通常在给定当前状态情况下，假设状态转移独立于状态持续时间，有

$$P(S_t=s^j,D_t=d_{s,t} \mid S_{t-d_{s,t}}=s^i,D_{t-d_{s,t}}=d_{t-d_{s,t}})=P(S_t=s^j,D_t=d_{s,t} \mid S_{t-d_{s,t}}=s^i)=P(S_t=s^j \mid S_{t-d_{s,t}}=s^i)P(D_t=d_{s,t} \mid S_t=s^j)=a_{ij}P_j(d_{s,t})。$$

其中：$S_t=s^j$ 为 t 时刻的隐变量状态 s^j；$d_{s,t}$ 为 t 时刻隐变量状态已经持续的时间，则上一个状态的结束时间为 $t-d_{s,t}$；a_{ij} 为每段状态的转移过程对应的状态转移概率，$a_{ij}=P(S_t=s^j \mid S_{t-d_{s,t}}=s^i)$；$P_j(d_{s,t})$ 为状态持续时间概率，$P_j(d_{s,t})=P(D_t=d_{s,t} \mid S_t=s^j)$。

2.2　隐半马尔可夫模型建模市场情绪

在股票市场中，市场情绪是一个无法直接观测的隐变量，可以通过相关的可观测变量感知市场情绪，使用 HSMM 模型建模市场情绪环境，其结构如图 2 所示。图 2 中：X_t 为 t 时刻情绪特征观测结点集；R_t 为 t 时刻收盘价结点；S_t 为 t 时刻情绪结点；D_t 为 t 时刻状态持续时间结点。

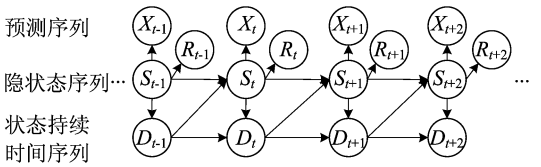

图 2　市场情绪隐半马尔可夫模型

根据市场环境的结构关系，观测序列、市场状态序列和持续时间序列的联合概率分布表示为：

$$P(O_{1:T},S_{1:T},D_{1:T})=P(X_{1:T},R_{1:T},S_{1:T},D_{1:T})=\prod_{t=1}^{T}P(X_t,R_t \mid S_t)P(S_t \mid S_{t-1},D_{t-1})P(D_t \mid D_{t-1},S_{t-1}) \tag{4}$$

其中：$P(X_t,R_t \mid S_t)$ 表示在情绪状态 S_t 情况下，观察变量的概率，有

$$P(X_t,R_t \mid S_t)=P(O_t \mid S_t)=P(O_t=o^k \mid S_t=s^i)=b_{ik} \tag{5}$$

其中，$O_t=\{X_t,R_t\}$ 为 t 时刻的观测结点集；o^k 为 $O_t=o^k$ 时观测结点 O 在 t 时刻的取值。

式(4)中的 $P(S_t|S_{t-1},D_{t-1})$ 表示情绪转移概率，可描述为：

$$P(S_t=s^j \mid S_{t-1}=s^i,D_{t-1}=d_{s,t-1})=\begin{cases}\delta(s^{t-1},s^t), & d_{s,t-1}>1;\\ a_{ij}, & d_{s,t-1}=1\end{cases} \tag{6}$$

其中：当 $s^{t-1}=s^t$ 时，$\delta(s^{t-1},s^t)$ 为 1，否则为 0；$a_{ij}=P(S_t=s^j|S_{t-1}=s^i)$。

式(4)中的 $P(D_t|D_{t-1},S_{t-1})$ 表示状态持续时间概率，可描述为：

$$P(D_t=d_{s,t} \mid S_t=s^i,D_{t-1}=d_{s,t-1})=\begin{cases}\delta(d_{s,t},d_{s,t-1}+1), & d_{s,t-1}>1;\\ P_i(d_{s,t}), & d_{s,t-1}=1\end{cases} \tag{7}$$

其中：$d_{s,t-1}>1$ 表示当上个时刻的趋势持续时间

变量 d_{t-1} 还在持续,下一时刻的状态持续时间 d_t 直接在 d_{t-1} 上加 1;$d_{s,t-1}=1$ 表示下一时刻将会开始新的状态,而新的状态持续时间 d_t 会根据新的状态持续时间概率分布重新生成。

2.3 市场情绪模型的参数估计

市场情绪的 HSMM 模型中需要学习的参数有 $\boldsymbol{\pi}$、$\boldsymbol{A}$、$\boldsymbol{B}$、$P_i(d)$。使用极大似然估计法对状态转移概率和初始概率进行估计。具体过程如下:从状态类型 s^i 转移到状态类型 s^j 的概率估计 $\hat{a}_{ij}$ 是用样本集的频数 $\text{count}(s^i \to s^j)$ 除以 s^i 转移到所有非自身状态类型的频数 $\text{count}(s^i \to \Omega_S)$,$\Omega_S$ 表示变量 S 的状态空间。$\hat{a}_{ij}$ 计算公式为:

$$\hat{a}_{ij}=\frac{\text{count}(s^i \to s^j)}{\text{count}(s^i \to \Omega_S)} \tag{8}$$

由式(6)、式(8)导出状态转移概率 $P(S_t \mid S_{t-1}, D_{t-1})$。而初始状态概率矢量估计 $\hat{\pi}$ 中的 $\hat{\pi}_i$ 则通过所有样本状态类型出现的频率进行估计,即

$$\hat{\pi}_i=\frac{\text{count}(s^i)}{\text{count}(\cdot)} \tag{9}$$

其中,$\text{count}(\cdot)$ 为样本总数。同理可得观测概率 $P(O|S)$ 为:

$$P(O=o^k \mid S=s^i)=b_{ik}=\frac{\text{count}(s^i,o^k)}{\text{count}(s^i)} \tag{10}$$

状态 s 对应持续时间 d 的概率为:

$$P_i(d)=\frac{\text{count}(d_{s,t},S_t=s^i)}{\text{count}(s^i)} \tag{11}$$

结合式(7)、式(11)得到持续时间的转移概率 $P(D_t=d_{s,t} \mid S_t=s^i, D_{t-1}=d_{s,t-1})$。由 HSMM 参数定义得到四元组$(\boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{B}, P_i(d))$。

3 情绪向量的热度度量

针对市场情绪易变性,利用情绪向量来平滑情绪波动,提升情绪的稳定性。

定义 2(情绪向量) 引入长度为 m 的滑动窗口,得到情绪向量 $\boldsymbol{W}_t=\{S_{t-m+1}, S_{t-m+2}, \cdots, S_t\}$,其中:$t-m+1$ 为窗口开始的时间;t 为窗口结束的时间。

$\boldsymbol{W}_t$ 蕴含了市场情绪在窗口内情绪冷热及其变化情况的高阶信息,将其概括为情绪热度。为了度量窗口的情绪热度,从历史数据中选择走势平缓的数据作为情绪平稳的基准分布,将情绪向量数据分布与选取的基准情绪数据分布的距离作为情绪热度度量结果,利用 KL 距离度量 2 个概率分布之差。KL 距离用来衡量相同事件空间里 2 个概率分布的差异程度,情绪向量与情绪热度基准数据的 KL 距离表示为 $\boldsymbol{D}_{\text{KL}}(\boldsymbol{F}_{\boldsymbol{W}_t}, \boldsymbol{F}_{H_t})$,计算公式如下:

$$\begin{aligned}\boldsymbol{D}_{\text{KL}}(\boldsymbol{F}_{\boldsymbol{W}_t}, \boldsymbol{F}_{H_t}) &= \sum \boldsymbol{F}_H \text{lb}\ \boldsymbol{F}_{\boldsymbol{W}_t}{}^{-1} - \\ \sum \boldsymbol{F}_H \text{lb}\ \boldsymbol{F}_H{}^{-1} &= \sum \boldsymbol{F}_H \text{lb}\ (\boldsymbol{F}_H \boldsymbol{F}^{-1}{}_{\boldsymbol{W}_t})\end{aligned} \tag{12}$$

其中:$\boldsymbol{F}_{\boldsymbol{W}_t}$ 为情绪向量的数据分布;$\boldsymbol{F}_H$ 为情绪热度基准数据的数据分布;$\sum \boldsymbol{F}_H \text{lb}\ \boldsymbol{F}_{\boldsymbol{W}_t}{}^{-1}$ 为用情绪向量的分布 $\boldsymbol{F}_{\boldsymbol{W}_t}$ 表示基准分布 $\boldsymbol{F}_H$ 所需信息量;$\sum \boldsymbol{F}_H \text{lb}\ \boldsymbol{F}_H^{-1}$ 为情绪热度基准数据分布信息熵。

定义 3(情绪热度) 表示时刻 t 情绪向量偏离基准情绪的程度,其数值由两者的相似度来计算,即

$$h_t=\exp[\gamma \boldsymbol{D}_{\text{KL}}(\boldsymbol{F}_{\boldsymbol{W}_t}, \boldsymbol{F}_H)] \tag{13}$$

其中:h_t 为 t 时刻情绪向量的情绪热度;γ 为调整 $\boldsymbol{F}_{\boldsymbol{W}_t}$ 和 $\boldsymbol{F}_H$ 的 KL 距离与相似度之间关系的系数,$\gamma>0$。

4 基于情绪向量的股市拐点预测

基于情绪向量的股市拐点预测是在给定观测序列 $\boldsymbol{O}_{1:t_1}$ 和观测序列 $\boldsymbol{O}_{t_1:t_2}$ 的条件下(t_1 为当前状态开始时间片,t_2 为当前时间片),首先选择窗口长度 m 得到情绪向量序列;然后计算情绪热度确定市场情绪状态;最后寻找概率最大的状态类型和对应持续时间。将求解的状态类型和对应持续时间分别记为 s^*、d^*,即

$$\begin{aligned}&(s^*, d^*)= \\ &\arg\max_{i,d_{s,t_2}} P(S_{1:t_1-1}, S_{t_1-1} \neq S_{t_1}, S_{t_1:t_2}= \\ &\qquad s^i, D_{t_2}=d_{s,t_2} \mid \boldsymbol{W}_{1:t_2})\end{aligned} \tag{14}$$

将情绪向量带入市场情绪隐半马尔可夫模型中,然后根据式(4),对式(14)中的条件概率进行因子分解,有

$$\begin{aligned}&P(S_{1:t_1-1}, D_{1:t_1-1} \mid \boldsymbol{W}_{1:t_1-1}) P(S_{t_1:t_2}, D_{t_1:t_2} \mid \\ &D_{t_1-1}) P(\boldsymbol{W}_{t_1:t_2} \mid S_{t_1:t_2}) = P(S_{1:t_1-1}, D_{1:t_1-1} \mid \\ &\boldsymbol{W}_{1:t_1-1}) P(S_{t_1} \mid S_{t_1-1}) P(D_{t_1} \geqslant \\ &t_2-t_1+1 \mid S_{t_1}) P(\boldsymbol{W}_{t_1:t_2} \mid S_{t_1:t_2})\end{aligned} \tag{15}$$

由式(14)可知,$P(S_{1:t_1-1}, D_{1:t_1-1} \mid \boldsymbol{W}_{1:t_1-1})$ 取得最大值的状态,并不能保证式(15)整体取得最大值。为了寻找使得条件概率最大的状态序列,预测拐点包含状态识别和持续时间概率最大化 2 个部分。首先根据模型推导 t_1-1 时刻最优状态序列,然后根据最优状态序列推导 t_1-t_2 时刻目标状态并计算持续时间的最大后验概率。

为简化模型表示,引入中间变量 $\varphi_t(s,d)$ 和 $\rho_{s^i}(S_{t_1}=s^j, \Delta t, \boldsymbol{W}_{t_1:t_2})$。

$\varphi_t(s,d)$表示 t 时刻以状态 s^i 和对应持续时间 $d_{s,t}$结尾的最优状态序列的概率，即

$$\varphi_t(s,d)=\max_{S_{1:t-d}} P(S_{1:t-d},S_{t-d}\neq S_{t-d+1},S_{t-d_{s,t}+1},S_t=s^i,S_t\neq S_{t+1},\boldsymbol{W}_{1:t})。$$

$\rho_s^i(S_{t_1}=s^j,\Delta t,\boldsymbol{W}_{t_1:t_2})$表示当 t_1 时刻前最后一个已完成状态为 s^i 时，尚未结束的当前的状态类型为 s^j 到目前 t_2 为止的时间 $\Delta t=t_2-t_1+1$ 和到目前为止的情绪向量序列为 $\boldsymbol{W}_{t_1:t_2}$ 的概率，即

$$\rho_s^i(S_{t_1}=s^j,\Delta t,\boldsymbol{W}_{t_1:t_2})=P(S_{t_1}=s^j\mid S_{t_1-1}=s^i)P(D_{t_1}\geqslant\Delta t\mid S_{t_1}=s^j)P(\boldsymbol{W}_{t_1:t_2}\mid S_{t_1:t_2})。$$

拐点推理过程的算法步骤如下：

1）输入一段时间情绪特征观测序列和对应的上证指数收盘价序列。

2）按照 2.3 节参数估计方法学习 HSMM 模型参数$(\boldsymbol{\pi},\boldsymbol{A},\boldsymbol{B},P_i(d))$。

3）选取窗口长度 m，划分情绪向量序列 $\boldsymbol{W}$ 计算情绪热度，然后根据收盘价确定尚未结束的当前趋势开始时间 t_1。

4）当 $t=1$ 时，由变量的初始值计算 $\varphi_t(s,d)$ 的边界概率，即 $\varphi_1(s,d)\leftarrow\pi_sP_s(d)$。

5）当 $1<t\leqslant t_1-1$ 时，由 $\varphi_t(s,d)$定义递推计算每个时间片中间变量 $\varphi_t(s,d)$。① 如果 $t-d_{s,t}>0$，那么 $\varphi_t(s,d)=\max\limits_{s_{t-d_{s,t}},d_{s,t-d_{s,t}}}\{\varphi_{t-d_{s,t}}(s,d)\times a_{s_{t-d_{s,t}}s^i}\}b_i(\boldsymbol{W}_{t-d_{s,t}+1:t})P_i(d_{s,t})$；② 如果 $t-d_{s,t}=0$，那么 $\varphi_t(s=s^i,d)=\pi_ib_i(\boldsymbol{W}_{t-d_{s,t}+1:t})p_i(d)$。根据计算结果，可知 t_1 时间片前最后一个状态 s^i 的概率 $\varphi_{t_1-1}(s=s^i,d)$。

6）状态 s^i 的持续时间为 d_{s,t_1-1}，求解 $t_1\sim t_2$ 时刻状态类型为 s^j 的概率 $\varphi_{t_1-1}(s=s^i,d)\times\rho_s^i(s^j,\Delta t,\boldsymbol{W}_{t_1:t_2})$，遍历情绪状态得到使之最大的状态类型 s^*，即

$$s^*\leftarrow\arg\max_{i,d_{s,t},j}\varphi_{t_1}(s=s^i,d)\rho_s^i(s^j,\Delta t,\boldsymbol{W}_{t_1:t_2})。$$

7）由状态持续时间分布得到状态类型 s^* 最大可能持续时间 d^*，$d^*=\arg\max\limits_d p_{s^*}(d)$。

8）用最大可能持续时间 d^* 减去经过的时间 t_2-t_1，则估计拐点将于 $d^*-(t_2-t_1)$ 个时间片后发生。

5　实验分析与结论

5.1　数据集

实验使用通达信软件收集 2018-01-02—2020-12-31 共 730 个交易日的股市数据、通过爬虫程序爬取新浪财经对应时间龙虎榜数据作为原始数据，从中提取 6 个市场情绪指标，划分 500 个训练样本和 230 个测试样本。考虑到情绪状态过多会造成收敛速度慢且不符合真实市场情绪状态，将市场情绪状态数设为 3，表示市场情绪的冷、温、热。

选择上证指数作为市场表现情况的指数并使用 PLR 生成拐点标签数据，并在训练数据上生成，为更直观地表示，对上证指数进行了标准化处理，结果如图 3 所示。

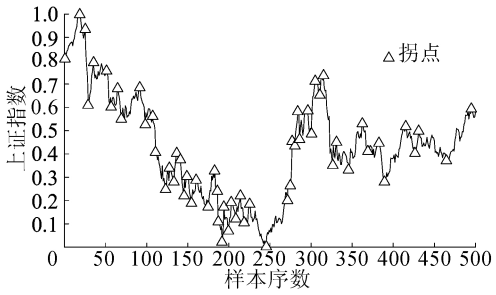

图 3　训练集和拐点

5.2　评价标准

为了验证算法有效性，采用精确度 P、召回率 R、综合指标 F_1 作为分类预测指标，即

$$P=\frac{n_{\mathrm{tp}}}{n_{\mathrm{tp}}+n_{\mathrm{fp}}},R=\frac{n_{\mathrm{tp}}}{n_{\mathrm{tp}}+n_{\mathrm{fn}}},F_1=\frac{2PR}{P+R}。$$

其中：n_{tp}为被模型预测为正的正样本数；n_{fp}为被模型预测为正的负样本数；n_{fn}为被模型预测为负的正样本数。

5.3　实验结果

对比 $m=5$ d 和 $m=10$ d 下不同方法的拐点预测结果，可知 m 的选取决定情绪向量的长度，见表 1、表 2 所列。当 $m=5$ d 时，SV-HSMM 方法的 P、R、F_1 值均优于现有方法；当 $m=10$ d 时，4 种方法预测效果均有不同程度下降，这是由于窗口长度变化导致输入的情绪向量变化，进而影响方法预测结果。

表 1　$m=5$ d 的预测结果　　%

方法	P	R	F_1
PLR-BPN	57.89	47.82	52.38
LSTM	59.09	56.52	57.78
PLR-WSVM	63.64	60.87	62.22
SV-HSMM	65.21	65.21	65.21

为了研究 m 对 SV-HSMM 预测精确率的影响，选取不同 m 进行实验，m 与 SV-HSMM 方法的 P 值见表 3 所列。当 $m<5$ 时，P 随 m 增大呈

显著上升趋势;当 $m>5$ 时,P 随窗口增大呈下降趋势。表明 m 决定了划分的情绪向量的长度,进而决定情绪热度所能稳定表达市场情绪的程度。因此,当 m 设置较小时,情绪的多变性导致预测结果不佳;当 m 选择过大时,情绪表达趋势的能力降低,不能有效捕捉情绪变化产生的拐点。

表2 $m=10$ d **的预测结果** %

方法	P	R	F_1
PLR-BPN	42.86	39.13	40.91
LSTM	50.0	52.17	51.06
PLR-WSVM	57.14	52.17	54.54
SV-HSMM	54.16	56.52	55.32

表3 m **与** SV-HSMM **方法** P **值**

m/d	P/%	m/d	P/%
1	45.45	8	59.09
2	51.72	9	56.52
3	57.69	10	54.16
4	62.50	11	52.00
5	65.21	12	50.00
6	65.00	13	48.15
7	61.90	14	46.43

综上所述,当 $m=5$ d 时,SV-HSMM 预测精确率取得最大为 65.21%。同时,结合表1可知,此时 P、R、F_1 值均优于现有方法。虽然相较于主流的 PLR-WSVM 方法,SV-HSMM 预测精确率提升不明显,但其 R、F_1 值也高于该方法,说明引入情绪向量并建模其动态变化能够提高模型预测的精度和稳定性。

6 结 论

本文针对目前历史数据学习模型的方法忽略股市受市场情绪多变性的影响,导致股市拐点难以预测的问题,提出一种融合市场情绪特征的基于情绪向量的市场隐半马尔可夫拐点预测方法(SV-HSMM)。首先在历史数据中提取情绪特征,使用马尔可夫毯融合情绪隐变量;然后在构建市场情绪隐半马尔可夫模型的基础上,进一步结合窗口提取情绪向量并使用 KL 距离度量情绪热度,提高稳定表达市场情绪变化的能力;最后根据模型推理预测拐点的发生时机。实验结果表明,该方法具有更高的预测精度和稳定性。优化预测模型,使之能在市场情绪基础上进一步筛选并预测市场热点股的走势是重点研究的方向。

[参 考 文 献]

[1] CHANG P C, LIAOTW, LIN J J, et al. A dynamic threshold decision system for stock trading signal detection[J]. Applied Soft Computing, 2011, 11(5): 3998-4010.

[2] LUO L K, CHEN X. Integrating piecewise linear representation and weighted support vector machine for stock trading signal prediction[J]. Applied Soft Computing Journal, 2013, 13(2): 806-816.

[3] TANG H M, DONG P W, SHI Y. A new approach of integrating piecewise linear representation and weighted support vector machine for forecasting stock turning points[J]. Applied Soft Computing Journal, 2019, 78: 685-696.

[4] 田红丽,杨莹莹,闫会强. 结合缠论和深度学习的股价拐点预测研究[J]. 计算机工程与应用,2022,58(16):319-325.

[5] PILAR C, ELENA F, RAFAEL S. Investor sentiment effect in stock markets: stock characteristics or country-specific factors[J]. International Review of Economics and Finance, 2013, 27: 572-591.

[6] 朱菲菲,李惠璇,徐建国,等. 短期羊群行为的影响因素与价格效应:基于高频数据的实证检验[J]. 金融研究,2019(7):191-206.

[7] WANG Z, HO S B, LIN Z. Stock market prediction analysis by incorporating social and news opinion and sentiment [C]//2018 IEEE International Conference on Data Mining Workshops(ICDMW). [S. l.]: IEEE, 2018: 1375-1380.

[8] JING N, WU Z, WANG H. A hybrid model integrating deep learning with investor sentiment analysis for stock price prediction[J]. Expert Systems with Applications, 2021, 178: 115019.

[9] LU S, LIU C, CHEN Z. Predicting stock market crisis via market indicators and mixed frequency investor sentiments [J]. Expert Systems with Applications, 2021, 186: 115844.

[10] 詹冰清,屈波怡. 市场情绪对股票走势的影响分析及预测[J]. 科技和产业,2021,21(4):51-57.

[11] JIN Z G, YANG Y, LIU Y H. Stock closing price prediction based on sentiment analysis and LSTM[J]. Neural Computing and Applications, 2019, 32(13): 9713-9729.

[12] 吴志军,李红军,刘亮,等. 基于小波能谱熵和隐半马尔可夫模型的 LDoS 攻击检测[J]. 软件学报,2020,31(5):1549-1562.

[13] 苏春,李乐. 基于隐半马尔可夫退化模型的非等周期预防性维修优化[J]. 东南大学学报(自然科学版),2021,51(2):342-349.

[14] YU S Z. Hidden semi-Markov models[J]. Artificial Intelligence, 2010, 174(2): 215-243.

(责任编辑 张 锴)