第 57 卷 第 3 期 厦门大学学报(自然科学版) Vol. 57 No. 3
2018 年 5 月 Journal of Xiamen University (Natural Science) May 2018
doi:10.6043/j.issn.0438-0479.201710011

基于隐马尔科夫模型的市场指数量化择时研究

傅中杰,吴清强*
(厦门大学软件学院,福建 厦门 361005)

摘要:量化择时是量化投资领域的重要组成,主要负责评判何时进行交易.为了验证隐马尔科夫模型(hidden Markov model,HMM)应用到量化择时的可行性,基于股票市场原始数据计算得到候选特征集,并利用 HMM 对各个单特征进行特征筛选,最后使用选出的特征集训练得到综合模型,预测交易日的市场状态.实验结果表明,基于 HMM 的交易策略比双均线策略和基于 k-均值(k-means)聚类的策略都有更好的表现,且具有较强的识别市场状态、规避系统性风险以及获取超额收益的能力.

关键词:隐马尔科夫模型;市场择时;交易策略
中图分类号:TP 391 **文献标志码**:A **文章编号**:0438-0479(2018)03-0404-09

精准预测金融市场是一件非常困难的任务,但是预测市场在未来一段时间的趋势或状态依旧是可行的.国内外许多文献已经指出,在金融市场尤其是股票市场和债券市场中,存在崩溃、缓慢成长、熊市和恢复阶段[1].De Angelis 等[2]提出了检测市场稳定和混乱状态的框架模型,并且预测了两种状态之间的转换.Salhi 等[3]使用隐马尔科夫模型(hidden Markov model,HMM)完成了金融危机和稳定时期的分类任务.与此同时,多种机器学习算法被应用于股票市场预测,如支持向量机(support vector machine,SVM)[4-6],神经网络[7],集成学习[8],深度学习[9]等.Galeshchuk[7]发现了拥有最佳预测能力的神经网络,并用于交易所数据预测.Bebarta 等[8]应用集成学习方法,组合人工神经网络、HMM 和遗传算法构建了一个股票预测系统.并且随着深度学习的兴起与成熟,越来越多的研究开始利用深度神经网络进行股票涨跌预测[9].然而由于为股票价格序列做标注难度较大,难以进行有效的监督学习,故 HMM 等非监督方法也被大量研究.Hassan 等[10]利用 HMM 找到与预测当天最相似的历史数据,用以预测下一天的股价,为股票预测提供了一种新范式.之后,Park 等[11]使用连续 HMM 预测下一天收盘价的变化方向.Seethalakshmi 等[12]利用 HMM 识别股价的危机期和稳定期.但这些研究多数直接使用携带噪音更多的日内开盘价、最高价、最低价和收盘价作为模型的输入特征进行训练,没有在特征的选择上进行探究,并且没有充分发挥 HMM 对隐状态的刻画能力,也缺乏策略层面的充分验证.

为了在环境多变、难以预测的股票市场中研发可靠的量化交易策略,取代主观性较强的人工交易方式,有效地保障资产组合保值增值,本研究利用 HMM 自身特点来识别和预测市场状态,将可观察的特征作为观测值,将金融市场状态作为预测目标的隐状态.在此基础上,生成指数基金的量化交易策略,对策略的有效性通过各个层面予以评估,且通过与当前业内较为常用的基于技术指标以及基于无监督聚类的择时方法对比说明了该方法的有效性,同时也分析了这种方法的局限性和缺陷.

1 股票市场的 HMM 算法

股票市场周期可以粗略区分为牛市、熊市和震荡市,还能分成更多细粒度的市场状态,这些市场状态难以观察,且转换过程往往并非一蹴而就,存在一定的过渡期.图 1 为 HMM 模型量化金融示例图,HMM 假设当前交易日 t 的市场状态 s_t 仅依赖于前一交易

收稿日期:2017-10-15 录用日期:2018-03-21
*通信作者:wuqq@xmu.edu.cn
引文格式:傅中杰,吴清强.基于隐马尔科夫模型的市场指数量化择时研究[J].厦门大学学报(自然科学版),2018,57(3):404-412.
Citation:FU Z J,WU Q Q.Research of market index quantitative timing based on hidden markov model[J].J Xiamen Univ Nat Sci,2018,57(3):404-412.(in Chinese)

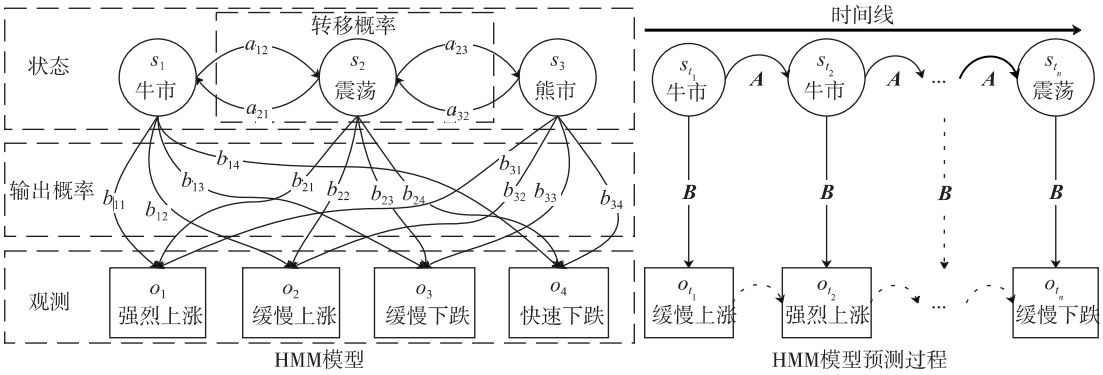


图 1　HMM 量化金融示例

Fig. 1　Example of HMM on quantitative finance

日状态 s_{t-1},故市场状态之间的转移概率分布为

$$P(s_t \mid s_{t-1}, s_{t-2}, \cdots, s_1, o_1, o_2, \cdots, o_{t-1}) = P(s_t \mid s_{t-1}), \tag{1}$$

其中 o_t 为观测状态,这里可观测变量很多,一般表现为股价和成交量的波动,也可表现为一些技术指标的数值变化.进而,转移概率矩阵 $\boldsymbol{A}$ 可表示为

$$\boldsymbol{A} = (a_{ij}) = (P(s_t = j \mid s_{t-1} = i)). \tag{2}$$

市场状态 s_t 映射到观测状态 o_t 的输出概率分布为

$$P(o_t \mid s_t, s_{t-1}, \cdots, s_1, o_1, o_2, \cdots, o_{t-1}) = P(o_t \mid s_t), \tag{3}$$

$$\boldsymbol{B} = (b_{jk}) = (P(o_t = k \mid s_t = j)). \tag{4}$$

其中 $\boldsymbol{B}$ 为输出概率矩阵.根据 HMM,在上述转移概率和输出概率的基础上,还需要有作为市场状态序列开端的初始状态概率分布 π.

$$\boldsymbol{\pi} = (\pi_i) = (P(s_1 = i)). \tag{5}$$

因此,股票市场可由一个完整的 HMM 模型 $\boldsymbol{\theta} = (\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi})$的三元组描述.HMM 的训练过程目前尚无最优解方法,一般采用鲍姆-韦尔奇(Baum-Welch)算法,根据期望-最大化(EM)原理确定局部最优的 $\boldsymbol{\theta}$ 三元组.

预测过程则使用维特比(Viterbi)算法,定义状态空间为 $\boldsymbol{S} = \{s_1, s_2, \cdots s_t\}$,令 V_{t,s_k} 为在已有 t 个观测值下以状态 s_k 为结尾的最可能状态序列的概率如式(6)和(7)所示,则当前状态 s_t 可由式(8)得到.

$$V_{1,s_k} = P(o_1 \mid k) \cdot \pi_k, \tag{6}$$

$$V_{t,s_k} = \max_{s \in \boldsymbol{S}}(P(o_t \mid s_k) \cdot a_{s,s_k} \cdot V_{t-1,s_{t-1}}), \tag{7}$$

$$s_t = \operatorname{argmax}_{s \in \boldsymbol{S}}(V_{t,s}). \tag{8}$$

HMM 保留了与当前交易日关联最强的前一交易日的信息,简化了建模过程,降低了模型复杂度;且 HMM 相较于神经网络等其他模型来说具有更好的可解释性;另外 HMM 也假设每个观测值仅依赖于当前的隐状态,该假设符合人们对于市场的理解认知.在牛市和熊市期间,市场的收益率和波动率分布具有显著差异[13],因此有理由相信在不同市场状态下观测值拥有不同的分布.

总的来说,HMM 拥有结构简单、鲁棒性强和可解释性好的优势,它在手写体识别[14-15]、体态识别[16]、自然语言处理[17-18]和语音识别[19]等时序相关领域的应用已经相当成熟;其不足主要在于部分远期历史信息的丢失以及缺乏对每个状态的直接解释.

2　基于 HMM 的量化择时模型

模型整体流程如图 2 所示.首先,通过开源数据接口获取股票市场指数的原始日频历史行情数据,并从中计算得到相关候选特征.其次,为了能够找到拥有卓越盈利能力的特征,使用 HMM 对单个候选特征建模以进行回测检验,并根据评估指标选取有效特征.随后,利用有效特征集合对给定状态数的 HMM 模型进行训练.最后在回测阶段,模型被用于预测后一交易日的市场状态,并发出相应的交易信号.通过上述步骤,得到模型的最佳参数,并用于实际的量化交易中.

2.1　特征准备

文中使用经过计算的特征取代前文所述的原始特征.在股票市场中,候选特征主要分为技术面特征和基本面特征两类.技术面特征主要是一些常用的技术指标,如相对强弱指数(relative strength index,RSI)、平均真实波动率(average true range,ATR)等,在这些指标的计算中往往对股价信息做了不同程度的平滑处理,在一定程度上减少了噪声,且这些指标在市

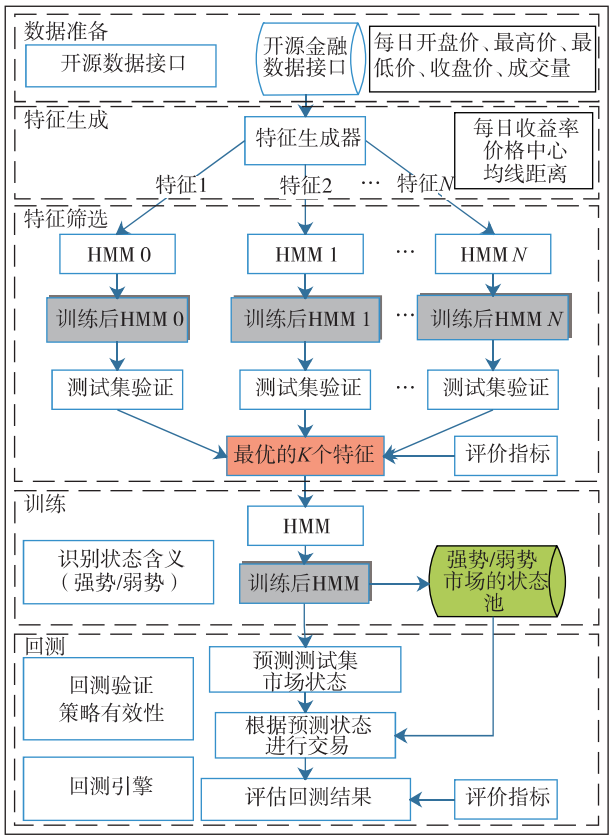


图 2 模型流程图

Fig. 2 Flow diagram of the model

场上被广泛运用多年,具有很强的可靠性.基本面特征主要是一些能够反映国家宏观经济状态的指标.上述特征都从各自的特定角度揭示了当前金融经济市场的状态.为了使模型更具通用性,本文中的主要研究对象为技术面特征.

2.2 特征筛选

特征筛选的主要任务是从所有候选特征中挑选高质量的特征.不同特征从不同角度反映了市场状态,本文中将这些特征组合在一起以期能够起到互补效果,提高模型表现.

根据图 2,在生成各个候选特征后,为每个特征构建单独的 HMM 模型并执行单特征检验.训练集上的单特征时间序列作为 HMM 模型的输入,可得到一个经过单特征训练的 HMM 模型,随后观察该模型在回测验证集上的预测表现.由于训练过程还涉及到状态数目的确定,本文中为每个单特征分别检验了状态数 $n\in[2,13]$ 的回测情况,并对各个状态数下得到的回测结果的评估指标计算平均值获得该特征的综合回测性能表现.通过观察每个单特征模型在数据集上的回测表现,能够比较这些特征,评估它们的效用.在量化金融领域,常用年化收益率、最大回撤、夏普比率、交易胜率和交易频率等作为评估指标.年化收益率是衡量盈利能力的直接指标,而最大回撤是衡量模型可能遇到的最坏情况的有效指标,夏普比率衡量了单位风险水平下可获得的收益,交易胜率代表了盈利交易的比例,交易频率则衡量了交易成本的数量.在由专家给出指标的相对重要性后,评估指标 RSI、平滑异同移动平均线(MACD)、ATR、资金流指标(MFI)、顺势指标(CCI)、乖离率指标(BIAS)的权重可由层次分析法获得,层次分析结构如图 3 所示.

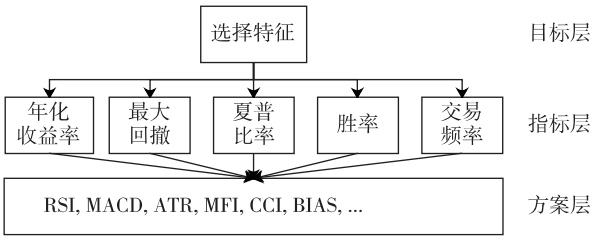


图 3 层次分析图

Fig. 3 Analytic hierarchy process diagram

评估出指标权重 w_i 后,可计算每个特征的综合得分,可以得到

$$\text{Score}=\sum_{i=1}^{n}w_i e_i, \tag{9}$$

其中,n 为评价指标数量,e_i 为评价指标值.据此筛选得到表现靠前的 K 个特征,组合成训练特征集用于最终模型的训练.

2.3 状态识别

HMM 模型的输出包含多个状态区间如图 4 所示,为了识别出盈利和非盈利状态,需要在训练集上为每种状态执行多头策略(即在预测为该状态的交易日买入持仓,在非该状态的交易日卖出),并且统计它们各自的累积收益率曲线.

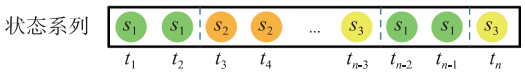


图 4 状态区间

Fig. 4 Interval of states

一种简单的方法是根据在训练区间上统计得到的状态累积收益情况将状态分为 2 类,累积收益为正,代表可盈利状态,累积收益为负,代表不可盈利状态.根据状态所属类型,相应的执行多头或空头策略.若在考察期间共有 T 个交易日,第 t 个交易日的累积收益为 r_t,本文中识别可盈利状态的判定条件为:

$$r_T>15\%\ \vee\ (r_T>5\%\ \wedge\ (\forall t\leqslant T,$$

$\exists r_t > -3\%)$). (10)

式(10)定义的判定条件实际上代表了两类可盈利状态:第一类是具有明显盈利能力的状态,其收益率显著为正;第二类是微盈利、小波动的状态,多数出现在震荡上行的行情.从长期来看,这两类状态都能够带来一定的利润.

一些盈利状态在某个特定时期可能会变得失效,一些状态也可能会从非盈利变为可盈利,这一方面是由于状态的含义不精准,导致状态识别存在一定的误差,另一方面也是由于市场本身的不稳定性造成的.尤其是当所选择状态数量多、比例大的时候,更有可能引入"不准确"状态.为了解决这个问题,采用一种直观的方法来满足需求——动态状态池.对每个状态在每个交易日进行监控,并根据式(10)条件进行检验,一旦某个状态满足条件,则将其加入到候选状态池中,相应地在该状态执行多头策略.相反,如果状态不能满足条件,则移出池子.

在状态识别的基础上,综合模型还需要确定最佳的状态参数.该步骤与单特征筛选的过程较为类似,主要通过分析一定范围状态数的综合模型在数据集上的回测验证结果,综合国家经济运行背景,选取其中状态区分度大,且盈利状态所获累积收益加和最高的状态数作为模型最终的状态参数.

2.4 回　测

训练后的模型能够预测后一天的市场状态,根据预测得到的市场状态在相对应的指数交易所交易基金(exchange traded funds,ETF)上进行交易.交易规则如式(11)所示,p_{t+1}为下一个交易日仓位,s_{t+1}表示预测的下一个交易日的市场状态类型,盈利状态为 1,其他状态为 0.

$$p_{t+1}=\begin{cases}1, s_{t+1}=1,\\ 0, s_{t+1}=0.\end{cases} \tag{11}$$

卓越的投资表现主要由以下 3 个因素构成:

1) 回报率衡量投资组合获取绝对收益和超额收益的能力,也是一个交易策略的基本要求.通常用年化收益率和超额收益率来衡量回报率,它分别衡量投资策略获取盈利的速度以及战胜市场的能力.

2) 风险也是一个重要的指标.一个稳定的策略能够规避糟糕的市场环境带来的巨大损失,低风险的策略能够做到低回撤.

3) 交易成本也是需要考虑的因素之一.在大多数情况下,过量的交易会带来大量的无效交易.在中国,证券公司会对每笔交易收取成交额的 0.15%作为佣金手续费,从长期来看,这是一笔巨大的开支并且会侵蚀一部分的利润.因此,保持相对高的胜率和控制合适的交易频率是必要的.

日频数据的使用也意味着本文中的策略是日频策略,策略模型监控每天的市场状态并且在下一个交易日做出相应的交易决策.添加仓位控制、止盈止损措施可以在很大程度上进一步将风险控制在更低的水平.

3 实　验

3.1 数据介绍

为了简化实验,本文中选择市场指数作为实验对象,国内市场选取沪深 300(CSI 300)指数,在国际市场选取标准普尔 500(S&P 500)指数作为建模标的,其中选取跟踪 CSI 300 指数的 300ETF(510300.OF)来模拟真实的交易场景.原始数据主要包含了交易日期、开盘价、最高价、最低价、收盘价、成交量、成交额信息.本研究选择 CSI 300 作为跟踪标的主要原因为:CSI 300 是我国股票市场最为重要的交易指数之一,其跟踪标的流动性好,不易受到操纵,且波动性适中;同时希望通过指数择时,能够为指数增强型基金产品提供一些新思路,以期获得超过纯被动型指数基金的超额收益.

3.2 特征筛选

如 2.2 节所述,本研究根据 5 个评价指标对 18 个候选特征进行综合打分并筛选特征.在保证一定盈利水平、可控制风险、维持一定的稳定性的目标下,由专家指导给出每个指标的重要程度分值,可得如式(12)所示的评估指标判断矩阵 $\boldsymbol{C}$.

$$\boldsymbol{C}=\begin{bmatrix}1 & 6/5 & 3/2 & 2 & 3\\ 5/6 & 1 & 5/4 & 5/3 & 5/2\\ 2/3 & 4/5 & 1 & 4/3 & 2\\ 1/2 & 3/5 & 3/4 & 1 & 3/2\\ 1/3 & 2/5 & 1/2 & 2/3 & 1\end{bmatrix}. \tag{12}$$

矩阵元素 c_{ij}表示指标 i 相对于指标 j 的相对重要程度,矩阵自左向右和自上向下分别表示年化收益率、最大回撤、夏普比率、交易胜率、交易频率.通过对式(12)中的 $\boldsymbol{C}$ 进行特征向量分析,取最大特征值对应的特征向量进行归一化,即可得到每个指标具体对应的权重,本研究中上述指标最终权重向量 $\boldsymbol{W}=(w_i)=[0.30, 0.25, 0.20, 0.15, 0.10]$.其中最大回撤和交易频率指标是反向指标,因此在计算最终分值时需要给予负值处理.另外由于评价指标存

在量纲不一致问题,需要在计算前进行标准化,主要依据各个评估指标的含义,以及其在量化投资领域常见的取值范围进行最大最小标准化(max-min normalization),夏普比率主要取值范围在(0,10),其余指标取值范围主要在(0,100),映射后可使所有指标不会带有额外的权重影响.

经过上述步骤并结合式(9),便可为每个候选特征打分,选取其中得分高于 5 分、最大回撤小于 30% 且夏普比率高于 1 的特征作为筛选后特征集.表 1 列出了各个候选属性的得分,最终选出了 6 个特征:14 日 ATR、20 日价格效率、每日收益率、MACD、意愿指标和 MFI.表 2 给出了这些特征作为单特征模型在 2014—2016 年间的表现.表中的结果是状态数分别为 2～13 时的测试结果统计平均值.一些特征在特定的状态数上表现突出.以这些特征为输入的模型的回测结果均获得了正的年化收益率,胜率大多超过 50%,且交易次数适中.

3.3 状态识别

如图 5 所示,将表 2 中选出的特征组合为每个交易日维度 $d=6$ 的特征向量,作为综合 HMM 模型的输入.对于不同的市场,有不同的理想状态数.实验结果表明,一些状态可能从非盈利状态转换为盈利状态,意味着动态状态池是必要的.实际上,在模型中隐状态的含义是不确定的,隐状态可能代表了某段时期的赚钱效应,也可能代表了市场的波动率.在一定程度上,可以通过输入不同的特征集来控制隐状态的内在含义,这解释了为什么在表 2 中"每日收益率"的回测结果能够取得最高的累积收益率.

表 3 第 1 行为配合动态状态池并取得最佳回测表现的 9 状态 HMM 模型得到的结果,在回测过程中,动态状态池由 3 状态增加到 4 状态,说明其检测到了新的盈利状态.为了探索通过动态状态池能够获得的利润空间上限,引入未来函数对比实验结果.训练集被定义在 2005—2013 年,回测验证集定义为 2014—

表 1 候选特征得分

Tab. 1 Scores of candidate features

指标	得分	指标	得分	指标	得分
14 日 ATR	26.3	人气指标	6.9	BIAS	−2.0
20 日价格效率	16.6	MA5/MA20 距离	6.2	close/MA5 距离	−2.1
每日收益率	15.8	每日振幅	5.6	RSI	−2.9
MACD	13.8	ROC	2.6	PVT	−6.1
意愿指标	12.9	CCI	0.7	5 日价格效率	−7.0
MFI	7.3	close/MA20 距离	−0.9	OBV	−7.3

表 2 所选特征的平均测试结果

Tab. 2 The average testing results of selected features

特征	胜率/%	频率/次		累积收益率/%	年化收益率/%	最大回撤率/%	夏普比率/%	交易成本/元	R
		开仓	平仓						
14 日 ATR	71.63	4.3	3.4	47.03	15.04	12.76	1.40	1 396	3.82
3*	(50.00)	(5)	(4)	(73.45)	(23.48)	(15.26)	(1.38)	(1 853)	(4.81)
20 日价格效率	55.03	68.5	67.5	59.79	19.12	22.95	1.34	26 163	2.81
4*	(59.65)	(58)	(57)	(92.34)	(29.52)	(24.05)	(1.32)	(25 651)	(3.84)
每日收益率	51.05	18.2	17.5	47.97	15.34	26.86	1.10	7 072	2.31
6*	(53.33)	(16)	(15)	(97.20)	(31.07)	(19.53)	(1.38)	(7 139)	(4.98)
MACD	51.74	28.7	28.7	35.19	11.25	16.95	0.83	8 679	2.34
5*	(55.56)	(18)	(18)	(67.83)	(21.68)	(21.92)	(1.41)	(6 539)	(3.09)
意愿指标	51.58	31.0	31.0	37.01	11.90	22.68	1.24	10 529	1.64
9*	(48.48)	(33)	(33)	(49.10)	(15.70)	(23.14)	(1.29)	(11 789)	(2.12)
MFI	44.54	32.3	32.3	18.96	6.09	21.39	0.95	9 814	0.86
4*	(56.52)	(23)	(23)	(42.96)	(13.73)	(22.29)	(1.24)	(7 904)	(1.93)

注:* 表示单特征下表现最好的状态数,() 表示拥有最高累计收益率的状态数,R 表示累积收益率与最大回撤率比值,下同.

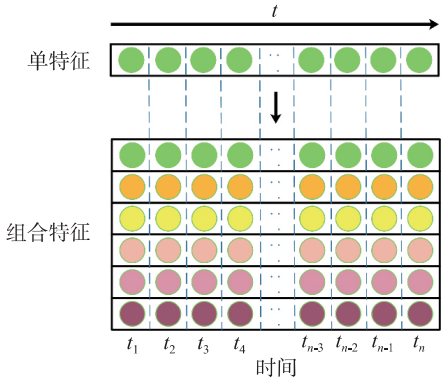


图 5　组合特征

Fig. 5　Combined features

2016 年.通过直接观察 2005—2016 年间满足式(10)定义的条件的状态,得到一个包含未来信息的“未来状态集”.

表 3 中第 2 和 3 行分别展示了 9 状态模型在静态状态集中的预测能力和获取潜在更高累积收益的能力.对比静态状态集,对于大部分状态数模型,随着时间推移,在回测区间上至少出现一个能够获取收益的新状态,部分状态数模型甚至出现 2～3 个新增的候选状态,且这些新状态所带来的收益是非常显著的.从表 3 中可以看出,9 状态动态状态池模型结果已经非常接近 9 状态下所能达到的最大潜在收益率 105.14%,这说明动态状态池是有效且必要的.图 6 展示了 CSI 300 指数 2014—2016 年期间的市场运行情况,CSI 300 指数基准收益率为 43.95%,9 状态模型获得了超过 60%的超额收益.

为了检验本文中模型状态识别能力的通用性,也对模型在国际市场的表现进行了测试.由于本文中的股票特征具有明确的市场意义,其有效性具有一定的跨市场通用性,因此在国际市场中,直接使用 3.2 节中在国内市场上筛选得到的特征集.选取美国金融市场作为测试目标,实验中使用 S&P 500 指数 2000—2006 年期间的数据作为训练数据,2007—2016 年作为回测集.表 4 给出了 4 状态 HMM 在 S&P 500 上的回测结果.状态数为 3 时,得到了 236.63%的最大潜在累积收益,而指数的基准收益为 140.27%.图 7 分别展示了 HMM 预测结果在回测区间上的状态转移情况和各个状态根据转移情况得到的净值曲线.从图 7 可知,长期而言,状态 0,1 代表了上升趋势,状态 2 代表市场反弹,同样可以带来利润.另外,状态 1 在开始阶段经历了损失,直到 2009 年才成为一个候选状态,而后为整个资产组合带来了利润.状态 3 代表了剧烈的下跌趋势.可以看到训练后的 HMM 模型避免了 2008 年金融危机给股市带来的剧烈下跌.受益于更加

表 3　9 状态数 HMM 在 CSI 300 指数上的回测结果(2014—2016 年)

Tab. 3　Backtest results on CSI 300 index (2014—2016) of 9 states HMM

状态集	选择状态数	胜率/%	频率/次		累积收益率/%	年化收益率/%	最大回撤/%	夏普比率/%	交易成本/元	R
			开仓	平仓						
动态	3→4	66.67	13	12	101.87	32.57	12.29	1.32	5 596	8.29
静态	3	55.56	10	9	65.53	20.95	12.27	1.42	3 906	5.34
静态	4	69.23	14	13	105.14	33.61	12.30	1.32	6 063	8.55

注:灰色表示使用了未来状态集的结果,下同.

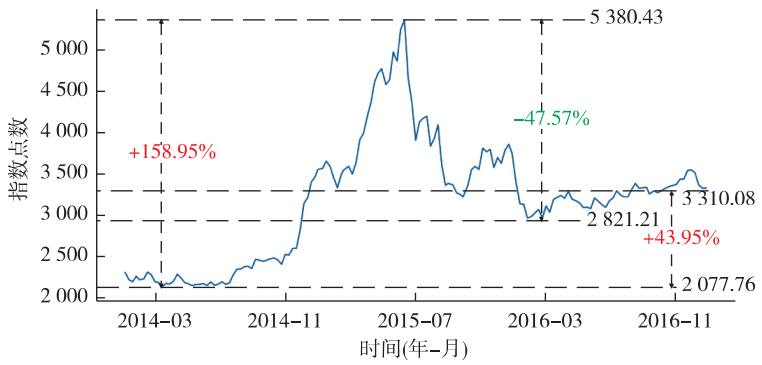


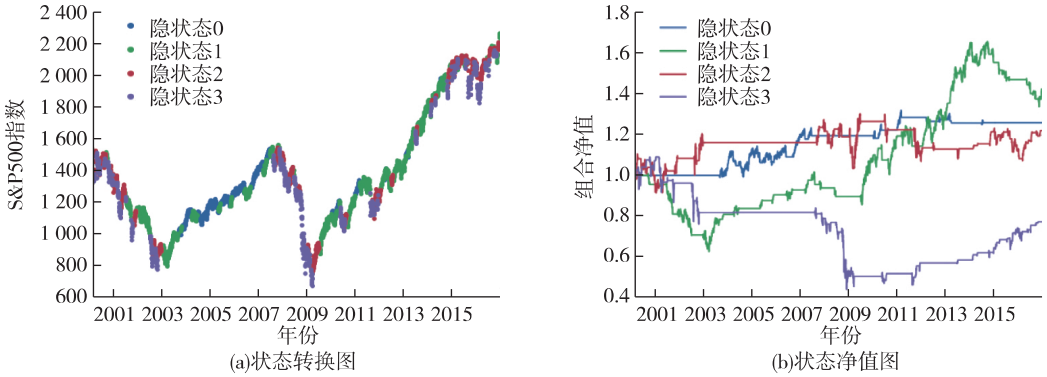
图 6　CSI 300 指数走势图(2014—2016 年)

Fig. 6　Trend chart for CSI 300 index (2014—2016)

表 4 S&P 500 指数部分结果(2007—2016 年)

Tab. 4 Part of results on S&P 500 index (2007—2016)

选择状态数	胜率/%	频率		累积收益率/%	年化收益率/%	最大回撤/%	夏普比率	交易成本/元	R
		开仓	平仓						
2	58.00	50	50	103.42	10.27	12.23	1.81	22 847	8.46
3	54.29	36	35	236.63	23.49	12.23	1.41	24 043	19.35



(a)状态转换图 (b)状态净值图

图 7 S&P 500 指数 4 状态 HMM 模型

Fig. 7 4-states HMM model for S&P 500 index

成熟的市场环境,在美国市场只需要更少的状态就能够充分描述市场变化,获取稳定收益.

3.4 结果分析

文中将基于 HMM 的量化择时模型与基于双均线技术指标、基于 k-均值聚类算法的量化择时模型进行了性能对比.

1) 基于双均线技术指标的量化择时模型

移动平均线是股票和期货市场上历史最为悠久、应用最为广泛的技术指标之一,基于均线的择时策略也因其简单易懂而广泛运用.双均线策略主要利用短期移动平均线和长期移动平均线之间的相对位置关系发出交易信号[20-21],短期均线运行在长期均线上方则视为持仓信号,反之则视为空仓信号.均线模型主要涉及的参数为长短期均线的周期.

2) 基于 k-均值聚类算法的量化择时模型

近年来,已有研究将机器学习方法运用在量化投资领域[22].本文中基于 k-均值聚类算法将每天的特征向量作为对每个交易日的描述,并将欧式距离作为样本间相似度的度量标准,将拥有相似特征的相似交易日进行聚类并预测为同一种市场状态,根据预测得到的状态发出相应交易信号.

3 种方法的比较结果如表 5 所示.通过组合长短期均线周期(5/10/20/30/60),对双均线择时模型共 10 个参数组合进行了测试,MA10/MA60 参数组合获得了最佳表现.在 2014—2016 年期间,该模型共发生 6 笔交易,获得 23.73%的年化收益率,此时最大回撤 25.68%.从结果上看,双均线策略的平均最大回撤率分别比 HMM 高出 16.5 个百分点,比 k-均值聚类算法高出 13.56 个百分点,均线策略本身的延迟性造成了幅度更大的回撤区间.利用表 2 中的特征对基于 k-均值聚类算法的量化择时模型进行测试.在没有使用动态状态选择的情况下,5 状态 k-均值聚类算法在 2014—2016 年期间取得了 15.19%的年化收益率和对应 13.29%的最大回撤.通过使用动态状态选择,8 状态 k-均值聚类算法能够带来 84%的潜在收益,对应最大回撤仅 12.74%.不过在大多数情况下,其胜率都低于 50%,且交易频率偏高.

综合对比结果可知,HMM 拥有参数少、鲁棒性强和可解释性强的优势.从最大回撤率的角度来看,基于 HMM 的策略的表现皆优于 k-均值聚类算法和双均线策略,表明相较于其他常见策略,HMM 在规避风险的敏感性和在控制回撤的有效性上表现更佳.从期望收益率的角度看,基于 HMM 的策略的表现优于 k-均值聚类算法,与均线策略相当.这主要源于 HMM 合理的马尔科夫性质假设.另外,HMM 策略带来了长期高于 50%的交易胜率,以及更低的交易频率.从平均意义上来说,它产生更少的交易费用.

表 5　3 种方法统计结果

Tab. 5　Statistical results of 3 methods

策略模型	统计	胜率/%	频率/次		累积收益率/%	年化收益率/%	最大回撤率/%	夏普比率	交易成本/元	R
			开仓	平仓						
HMM	Avg	56.82	13.3	12.8	44.88	14.35	14.95	1.37	4 558	3.37
	Max	100.00	26.0	26.0	83.69	26.76	24.37	1.44	8973	5.48
	Min	34.62	2.0	1.0	3.24	1.04	6.98	1.26	483	0.32
	SD	17.70	7.2	7.5	20.33	6.50	5.79	0.06	2462	1.71
移动平均线	Avg	47.53	15.6	15.5	45.33	15.13	31.52	0.49	5428	1.65
	Max	61.54	44.0	44.0	71.12	23.73	46.60	0.78	1 3902	3.32
	Min	33.33	6.0	5.0	10.85	3.62	20.26	0.13	1 990	0.27
	SD	8.37	11.2	11.3	18.94	6.32	8.07	0.20	3 452	0.95
k-均值聚类算法	Avg	40.11	34.1	34.0	28.39	9.47	17.96	0.60	11 062	1.82
	Max	53.06	80.0	80.0	45.54	15.19	35.99	1.08	2 3056	3.43
	Min	31.03	20.0	20.0	−1.32	−0.44	13.29	−0.02	7065	−0.04
	SD	7.15	16.2	16.2	11.49	3.83	5.90	0.27	4451	0.90

注:Avg 表示平均值,Max 表示最大值,Min 表示最小值,SD 表示标准差.

均线策略的最大缺陷来源于均线的延迟性,这会在趋势不明显的震荡市带来长期的亏损.只有在趋势明显的牛市或熊市期间均线策略才能够带来大幅利润或避免大幅损失.然而通常来说,在经济周期的影响下,市场每经过 7～8 年才会有一轮趋势明显的牛市或熊市,因此均线策略所需回报周期更长.另外,找到合适的均线参数也是相当困难的,很容易造成过拟合.而 k-均值聚类算法则完全没有考虑相邻交易日之间的关系,导致了频繁的状态转换和高昂的交易成本.

4　结　论

本研究围绕量化投资领域的量化择时问题,研究了 HMM 在该领域的应用,给出了如何基于 HMM 构建量化交易策略的完整流程.同时将基于移动平均线和 k-均值聚类算法的量化择时模型作为对比,对实验结果进行了分析,验证了 HMM 具有识别市场中长期状态的能力.HMM 的主要原理在于马尔科夫性质的假设,相邻时序样本之间的关联信息能够被有效利用,隐状态的转换存在一定的概率分布,因而它能够选择合适的交易时机,并在市场迎来暴跌时有效保护资产组合.另外相较于其他两种常见策略,它在敏感性和稳定性上有更好的表现.

尽管 HMM 在量化择时上有较为优越的表现,但是仍然存在一些不足,在后续的研究中将进一步改进.

1) 当前模型的状态选择依然属于静态规则,不能有效监控各个状态的动态变化.当某个状态当前不能满足特定条件时,很容易错过巨额的利润空间,而满足条件时,又很容易遭遇巨幅回撤,也即在动态状态监控上存在滞后.因此需要对动态状态选择机制进行进一步的智能优化.

2) 择时往往只是一个量化交易系统的一部分,通过合理利用策略融合技术,将隐马尔科夫择时模型与其他的策略组合使用,充分利用择时模型对市场环境的预判能力,能够更加有效地挖掘个股股价的利润空间,规避市场风险.

3) 本文中提出的策略流程,尚未考虑任何仓位控制和止盈止损措施,如何合理通过添加风险控制措施来提高策略的可控性也是重要的研究课题.

4) 当前针对的主要跟踪交易对象为市场指数,在后续研究中,将进一步针对预测难度更大的个股以及期货品种进行模型研究.

参考文献:

[1] GUIDOLIN M, TIMMERMANN A. Asset allocation under multivariate regime switching[J]. Journal of Economic Dynamics & Control, 2007, 31(11): 3503-3544.

[2] DE ANGELIS L, PAAS L J. A dynamic analysis of stock markets using a hidden Markov model[J]. Journal of Applied Statistics, 2013, 40(8): 1682-1700.

[3] SALHI K, DEACONU M, LEJAY A, et al. Regime switching model for financial data: empirical risk analysis[J]. Physica a — Statistical Mechanics and Its Applications, 2016, 461: 148-157.

[4] 杨新斌, 黄晓娟. 基于支持向量机的股票价格预测研究[J]. 计算机仿真, 2010, 27(9): 302-305.

[5] 谢国强. 基于支持向量回归机的股票价格预测[J]. 计算机仿真, 2012, 29(4): 379-382.

[6] CAO L, TAY F E H. Financial forecasting using support vector machines[J]. Neural Computing & Applications, 2010, 10(2): 184-192.

[7] GALESHCHUK S. Neural networks performance in exchange rate prediction[J]. Neurocomputing, 2016, 172(C): 446-452.

[8] BEBARTA D K, SUDHA T E, BISOYI R. An intelligent stock forecasting system using a unify model of CEFLANN, HMM and GA for stock time series phenomena [C] // Emerging Ict for Bridging the Future. Berlin: Springer-Verlag Berlin, 2015: 485-496.

[9] 李文鹏, 高宇菲, 钱佳佳, 等. 深度学习在量化投资中的应用[J]. 统计与管理, 2017, 9(8): 104-106.

[10] HASSAN R, NATH B. Stock market forecasting using hidden Markov model: a new approach[C] // 5th International Conference on Intelligent Systems Design and Applications. Los Alamitos: IEEE, 2005: 192-196.

[11] PARK S H, LEE J H, SONG J W, et al. Forecasting change directions for financial time series using hidden Markov model[C] // Rough Sets and Knowledge Technology. Berlin: Springer, 2009: 184-191.

[12] SEETHALAKSHMI R, KRISHNAKUMARI B, SAAVITHRI V. Gaussian kernel based HMM for time series data analysis[C] // Conference Proceedings of 2012 International Conference on Management Issues in Emerging Economies. Thanjavur: IEEE, 2012: 105-109.

[13] 李腊生, 翟淑萍, 关敏芳. 证券市场收益率分布时变性的经济学分析及其我国的经验证据[J]. 统计研究, 2011, 28(11): 66-78.

[14] FIERREZ J, ORTEGA-GARCIA J, RAMOS D, et al. HMM — based on-line signature verification: feature extraction and signature modeling[J]. Pattern Recognition Letters, 2007, 28(16): 2325-2334.

[15] 肖明, 贾振红. 基于轮廓特征的 HMM 手写数字识别[J]. 计算机工程与应用, 2010, 46(33): 172-174, 211.

[16] CHEN F S, FU C M, HUANG C L. Hand gesture recognition using a real — time tracking method and hidden Markov models[J]. Image and Vision Computing, 2003, 21(8): 745-758.

[17] KHAN W, DAUD A, NASIR J A, et al. A survey on the state—of-the-art machine learning models in the context of NLP[J]. Kuwait Journal of Science, 2016, 43(4): 95-113.

[18] 韩普, 姜杰. HMM 在自然语言处理领域中的应用研究[J]. 计算机技术与发展, 2010(2): 245-248, 252.

[19] PARAMONOV P, SUTULA N. Simplified scoring methods for HMM — based speech recognition[J]. Soft Computing, 2016, 20(9): 3455-3460.

[20] 谭磊. 趋势跟踪类策略的内在逻辑[J]. 当代经济, 2017, 27(6): 144-145.

[21] 景泰然. 量化投资在期货交易中的应用[J]. 现代商业, 2015, 12(18): 152-153.

[22] 张文俊, 张永进. 4 种数据挖掘典型分类方法在股票预测中的性能分析[J]. 安徽工业大学学报(自然科学版), 2017, 34(1): 97-102.

Research of Market Index Quantitative Timing Based on Hidden Markov Model

FU Zhongjie, WU Qingqiang*

(Software School of Xiamen University, Xiamen 361005, China)

Abstract: Quantitative market timing constitutes an important part of quantitative investment to choose the best trading opportunity. To verify the feasibility of applying hidden markov model (HMM) to quantitative market timing, we creatively calculate candidate features set based on raw data, use HMM to test performance on each single feature, and train a comprehensive model using selected features to predict the market state of the next trading day. Experimental results show that HMM—based strategy enjoys better stability and profitability compared with strategies based on moving average or k-means. Finally, HMM can skillfully identify market states, avoid systematic risk and obtain excess return.

Key words: hidden Markov model(HMM); market timing; trading strategy