探寻西蒙斯投资之道：基于 HMM 模型的周择时策略研究

——模式识别方法应用系列报告一

罗军 研究员 电话：020-87555888-655 eMail：lj33@gf.com.cn
史庆盛 研究助理
胡海涛 研究员 电话：020-87555888-406 eMail：hht@gf.com.cn
李明 研究助理 电话：020-87555888-687 eMail:lm8@gf.com.cn
蓝昭钦 研究助理 电话：020-87555888-667 eMail：lzq3@gf.com.cn

隐马尔科夫模型（HMM）是西蒙斯大奖章基金的主要工具之一

成立以来到2008年，大奖章基金的平均年度净回报是35.6%。2000年科技股灾，标普500下跌10.1%，大奖章净回报98.5%；2008年，全球金融危机，大奖章净回报80%。无论牛熊市，基金表现非常不错。

复兴技术成立初期有三位著名的科学家对公司的长期发展产生重要影响，率先提出隐马尔科夫模型的鲍姆就是其中之一。1993 年加盟复兴技术的剑桥大学数学博士尼可.帕特森也是全球 HMM 领域公认的专家。复兴技术使用 HMM 模型的可能性非常大。

HMM 模式识别模型的优势：动态刻画价量推动的过程

股价的变化来自于未知力量的价量推动，推动的完成其需要一个动态的过程，而HMM在描述该过程的动态变化方面相对其他模式识别工具具备明显的优势，这也正是我们采用HMM尝试对股价预测的原因。

本文将HMM模型应用到我国股市的预测中，通过对股票数据序列的模式识别来对股市每周趋势进行预测。我们将股票的未来走势分别划分为两种（涨、跌）和三种（涨、跌、平）状态，把股市的波动预测转化为分类问题，并通过 HMM 模型进行识别。

量化模型输入指标的选取：基于高频的资金流指标为主

在具体构建指数预测模型时，特征提取即输入变量的选择我们考虑了若干与收盘价及成交资金量相关的指标，并最终选取了标的指数的日收益率数据、日净资金占比、日总资金环比以及标准化的日总资金作为量化择时模型的基础指标。

模型的预测准确率：大幅超过随机预测概率

实证中我们选取了沪深300指数作为量化择时策略的标的指数，对两类波动模式的HMM模型我们做了 115 期的样本外预测，准确率达到了 60.87%，对三类波动模式的 HMM 模型我们做了 96 期的样本外预测，准确率则达到了 47.37%，均高于两类模型的随机预测概率 50%以及 33.4%。

HMM 预测择时策略的收益远超基准

在两类波动模式的 HMM 模型中，量化择时策略于 2007 年 12 月 19 至 2010 年 4 月 29 日共 115 周间，在不考虑交易费用的情况下资产收益率分别为+67.84%，而同期内沪深 300 指数收益率为-33.3%；在三类波动类型的 HMM 模型中，2008 年 5 月 12 至 2010 年 4 月 29 日共 96 周间，在不考虑交易费用情况下资产收益率分别为+13.08%，而同期沪深 300 指数收益率为-14.2%。量化择时策略在绝大多数的时期都明显地超越了沪深 300 指数的收益表现。

.

目录索引

图表索引

HMM与量化投资：探寻西蒙斯投资之道

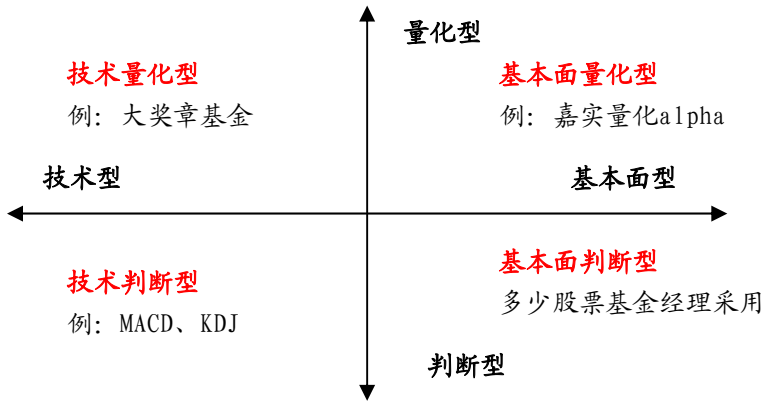
投资流派的分类：判断型 V.S 量化型、技术型 V.S 基本面型

投资方法通常根据其决策的方式分为两类：判断型和量化型。判断型投资者根据各种信息以及个人过去的经验确定买卖什么、买卖多少、什么价位执行等。索罗斯与巴菲特应该都是判断型的投资者。判断型投资的中心枢纽是人的大脑。各类信息汇入大脑，不同知识结构以及性格的人会得出不同的结论。但是，人类对自己大脑的认知是非常有限的，大脑是典型的“黑箱”。然而，由于量化工具的理论深奥、应用局限性较大等原因，通常把量化投资称为黑箱投资。量化投资是把最新的各类信息输入各自不同的模型，由模型产生投资指令。量化型投资者最大的不同在于其不需人脑判断，完全根据模型决策。

投资方法根据信息源可以分为基本面型以及技术型两类。基本面投资者根据宏观经济或者公司基本面等信息进行投资决策，而技术型的投资者通常根据二级市场交易产生的价量进行分析。根据2007年的统计，全球70%的资金都是凭借基本面型的投资方法操作。技术型以及量化型的投资最近30年开始逐渐壮大，目前使用量化分析进行投资的资金额占比20%左右，在全球很多大型的股票交易所中，接近50%的交易量来自各类量化投资的方式。

综合上面两种分类方法，投资方法可分为基本面判断法、基本面量化法、技术判断法和技术量化法。包括巴菲特在内的众多明星基金经理大多应用基本面判断法，民间个人投资者应用较多的各类技术分析指标均为技术判断法。而西蒙斯掌管的大奖章基金主要为技术量化法。目前，国内共同基金逐渐出现了基本面信息分析为主的量化基金，基本面量化法目前是国内金融工程领域各机构最普遍的应用方法。

图表 1 投资方法分类



数据来源：广发证券发展研究中心

西蒙斯的主要应用工具之一——隐马尔科夫模型（HMM）

1988年，著名的数学家西蒙斯成立了大奖章基金，该基金采用纯粹的技术量化方法进行投资。成立以来到2008年，大奖章基金的平均年度净回报是35.6%。并且，在期间的几次有名的股灾中均表现优异。2000年科技股灾，标普500下跌10.1%，大奖章净回报98.5%；2008年，全球金融危机，各类资产价格下滑，大奖章净回报80%。这些收益均为剔除了管理费以及收益提出的净回报。大奖章基金各类费用非常高，每年固定管理费2%，外加20%收益提出，2002年剔除比例提高到36%，2008年提高到44%。这些优异的收益数据令众多投资者对其神秘的投资方法感到好奇。

1988年3月，复兴技术成立初期有三位著名的科学家对公司的长期发展产生重要影响，鲍姆就是其中之一。统计学中著名的鲍姆-威尔士算法就是以他的名字命名，该算法是确定某种不可确知的变量出现的概率，被广泛应用于语音识别等领域。同时，鲍姆也是率先提出隐马尔科夫模型的科学家之一，该方法在语音识别中得到非常成功的应用。在金融领域中，HMM可以被用来推测目前的市场状态究竟是趋势还是震荡，究竟是高波动还是低波动。1993年加盟复兴技术的剑桥大学数学博士尼可.帕特森就是全球HMM领域公认的专家。复兴技术使用HMM模型的可能性非常大。

数据挖掘的技术工具很多，包括SVM、神经网络、HMM等，这些工具在众多领域应用都取得非常大的成功。我们认为HMM在股价预测应用上更合适，主要考虑股价的变化来自于未知力量的价量推动，推动的完成其需要一个动态的过程，而HMM在描述该过程的动态变化方面相对其他工具具备明显的优势，这也正是我们采用HMM尝试对股价预测的原因。

预测的好坏不仅来自于量化工具的选择，更重要的是输入特征的提取，本报告在解析HMM预测方法的同时，也在尝试输入特征提取上的实证，为采用其他工具进行预测的投资者提供新的特征提取思路。

HMM模型简介

HMM的基本理论

HMM 是在 Markov 链的基础上发展起来的，与 Markov 链模型不同的是，HMM 模型更加复杂，其观测值与状态不是一一对应的，而是通过一组概率分布相联系。HMM 是一个双内嵌式随机过程，即 HMM 是由两个随机过程组成，一个是隐含的状态转移序列，它对应一个单纯的 Markov 过程；另一个则是与隐状态有关的观测序列。并且在这两个随机过程中，有一个随机过程（状态转移序列）是不可观测的，只能通过另一个随机过程的输出观测序列进行推断，所以称之为隐马尔可夫模型，HMM 模型的基本要素包括:

(1) 模型的状态数 N。如果 S 是状态集合，则 $S=\{S_1,S_2,...S_N\}$。模型在时间 t 的状态记为 $q_t \in S$，$1 \le t \le T$，T 是观察序列的长度。模型经历的状态序列记为 $Q=\{q_1,q_2,...q_T\}$。

(2) 观察符号数 M。设 V 是所有观察符号的集合，则 $V=\{v_1,v_2,...v_M\}$

(3) 状态转移的概率分布 A。状态转移的概率分布可表示为 $A=\{a_{ij}\}$，其中，$a_{ij}=P\{q_{t+1}=S_j \mid q_t=S_i\}$，$1\le i,j\le N$，且满足 $a_{ij}\ge 0,\sum_{i=1}^{N}a_{ij}=1$，表示时刻 t 从状态 S_i 转移到时刻 t+1 状态 S_j 的转移概率。

(4) 状态 S_i 条件下输出的观测变量概率分布 B。假设观测变量的样本空间为 V，在状态 S_i 时输出观测变量的概率分布可表 $B=\{b_i(v),1\le i\le N,v\in V\}$，其中，$b_i(v)=P\{Q_t=v \mid q_t=S_i\}$，$Q_t$ 为时刻 t 的观测随机变量，可以是一个数值或向量，观测序列记为 $O=\{O_1,O_2,...O_t\}$。值得注意的是，此处观测变量的样本空间和概率分布可以为离散型，也可为连续型。

(5) 系统初始状态概率分布 π。系统初始状态概率分布可表示为 $\pi=\{\pi_i,1\le i\le N\}$，其中，$\pi_i=P\{q_1=S_i\}$。

综上可知，要描述一个完整的 HMM，需要确定模型参数 $\{N,M,A,B,\pi\}$。为了简化，常用下面的形式来表示，即 $\lambda=\{A,B,\pi\}$。此外，对于一个标准 HMM 模型，需要解决模型训练、隐状态估计和似然计算三个基本问题。

HMM模式识别模型的关键算法

将隐马尔可夫模型（HMM）应用到实际中，需解决以下三个基本问题:

问题 1： 给定模型参数 $\lambda=\{M,N,A,B,\pi\}$ 和观测序列 $O=(o_1o_2...o_T)$，如何快速求出在该模型下，观测序列发生的概率 $P(O\mid\lambda)$？

问题 2: 给定模型参数 $\lambda=\{M,N,A,B,\pi\}$ 和观测序列 $O=(o_1o_2...o_T)$，如何找出对应的最佳状态序列 $S=\{S_1,S_2,...S_N\}$？

问题 3: 如何确定模型的参数 λ，使得条件概率 $P(O\mid\lambda)$ 最大化?

对应于 HMM 三个问题的求解，产生了三个关键算法：向前－向后算法、Viterbi 算法以及 Baum-Welch 算法。

（1） 向前—向后算法

给定模型 λ 产生某一状态序列 $Q=\{q_1,q_2,...q_t\}$ 的概率为 $P(O\mid\lambda)$，使用向前—向后算法对 $P(O\mid\lambda)$ 进行求解如下:

首先，定义前向变量：$\alpha_t(i)=P(o_1o_2...o_t,q_t=s_i\mid\lambda)$，$\alpha_t(i)$ 可按如下步骤进行迭代计算:

1) 初始化:

$$\alpha_1(i)=\pi_i b_i(o_1)\qquad(1\le i\le N);$$

2) 递归计算:

$$\alpha_{t+1}(j)=[\sum_{i=1}^{N}\alpha_t(i)\alpha_{ij}]b_j(o_{t+1}),1\le t\le T-1;\quad 1\le j\le N;$$

3) 终止:

$$P(O|\lambda)=\sum_{i=1}^{N}\alpha_T(i)。$$

其次，类似地定义后向变量: $\beta_t(i)=P(o_{t+1}o_{t+2}...o_T\mid q_t=i,\lambda)$，$(o_{t+1}o_{t+2}...o_T)$表示从终止时刻 T 到时刻 t+1 的观测事件序列，则 $\beta_t(i)$ 可按如下步骤进行迭代计算:

1) 初始化:

$$\beta_T(i)=1,(1\le i\le N)$$

2) 递归计算:

$$\beta_t(i)=\sum_{j=1}^{N}a_{ij}b_j(o_{t+1})\beta_{t+1}(j),t=T-1,T-2,...,1;\quad 1\le i\le N$$

3) 终止:

$$P(O|\lambda)=\sum_{i=1}^{N}\pi\beta_1(i)$$

则给定模型 λ 下，产生状态序列 $Q=\{q_1,q_2,...q_t\}$ 的概率为

$$P(O\mid\lambda)=\sum_{i=1}^{N}\alpha_t(i)\beta_t(i)\qquad 1\le t\le T \tag{1}$$

（2） Viterbi 算法

给定观测序列 O 以及模型 λ，如何选择对应的状态序列 $Q=\{q_1,q_2,...q_t\}$，使得 Q 能够最为合理的解释观测序列 O?

首先定义: $\delta_t(i)=\max_{q_1,q_2...q_{t-1}}P[q_1q_2...q_{t-1},q_t=i,o_1o_2...o_t\mid\lambda]$，我们所要找的就是 T 时刻最大的 $\delta_T(i)$ 所代表的那个状态序列。可使用运用 Viterbi 算法求解 Q，步骤如下:

1) 初始化:

$$\delta_1(i)=\pi_i b_i(o_1);$$

2) 递归计算:

$$\delta_t(j)=\max_{1\le i\le N}[\delta_{t-1}(i)a_{ij}]b_j(o_t)\quad 2\le t\le T\quad 1\le j\le N$$

$$\Psi_t(j)=\arg\max_{1\le i\le N}[\delta_{t-1}(i)a_{ij}]\quad 2\le t\le T\quad 1\le j\le N$$

3) 终结:

$$P^* = \max_{1\le i\le N}[\delta_T(i)]\,,\quad q_T^* = \arg\max_{1\le i\le N}[\delta_T(i)]$$

4) 求序列 Q:

$$q_t^* = \Psi_{t+1}(q_{t+1}^*) \qquad t = T-1, T-2, ..., 1$$

（3） Baum-Welch 算法

Baum-Welch 算法是曾经在复兴技术工作过的统计学家鲍姆提出来的，实际上就是为了解决 HMM 的训练问题，即 HMM 模型的参数估计问题，也就是给定一个观测值序列 O，调整模型参数 λ，使其产生观测值序列的概率 $P(O|\lambda)$ 最大。根据前向变量和后向变量的定义，可以推出前向变量和后向变量的混合的 $P(O|\lambda)$ 概率公式:

$$P(O|\lambda)=\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_t(i)a_{ij}b_j(o_{t+1})\beta_{t+1}(j),\quad 1\le t\le T\text{-}1$$

实际上，当给定的训练序列有限时，不存在一个最佳的方法来估计 λ，只可能找到某些方法，其在特定的几个性能上有比较强的优势。在这种情况下，Baum-Welch 算法利用递推的思想，使 $P(O|\lambda)$ 局部最大，最后得到模型参数。

首先定义 $\xi_t(i,j) = P(q_t = S_i, q_{t+1} = j \mid O, \lambda)$ 表示 t 时刻状态为 S_i，t+1 时刻状态为 S_j 的概率。根据向前向后变量可以推出:

$$\xi_t(i,j) = \frac{P(q_t = S_i, q_{t+1} = S_j, O \mid \lambda)}{P(O \mid \lambda)} = \frac{\alpha_t(i)a_{ij}b_j(o_{t+1})\beta_{t+1}(j)}{P(O \mid \lambda)}$$

$$= \frac{\alpha_t(i)a_{ij}b_j(o_{t+1})\beta_{t+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_t(i)a_{ij}b_j(o_{t+1})\beta_{t+1}(j)}$$

再定义 $\gamma_t(i) = \sum_{j=1}^{N}\xi_t(i,j)$ 表示 t 时状态为 S_i 的概率，则可得到 Baum-Welch 算法的重估公式为:

$$\tilde{\pi}_i = \gamma_1(i)\,,\quad \tilde{a}_{ij} = \frac{\sum_{t=1}^{T-1}\xi_t(i,j)}{\sum_{t=1}^{T-1}\gamma_t(i)}\,,\quad \tilde{b}_j(k) = \frac{\sum_{\substack{t=1\\ o_t=v_k}}^{T}\gamma_t(j)}{\sum_{t=1}^{T}\gamma_t(j)}\,。$$

这里，$b_i(v) = P\{Q_t = v \mid q_t = S_i\}$ 服从离散分布，而实际应用中，我们常常需要假设观测矢量服从连续分布，也称为连续隐马尔可夫模型，其中使用得最多的是高斯分布，其密度函数为:

$$P(o_t \mid S_t = i)=\sum_{i=1}^{N}\omega_{im}N(o_t,\mu_{im},\Sigma_{im})\,,\quad \omega_{im}\ge 0, \sum_{i=1}^{N}\omega_{im}=1$$

则可得到 Baum-Welch 算法的重估公式为:

$$\tilde{\mu}_{im}=\frac{\sum_{i=1}^{T}\delta_{im}o_t}{\sum_{i=1}^{T}\delta_{im}(t)},\quad \tilde{\sigma}_{im}=\frac{\sum_{i=1}^{T}\delta_{im}(o_t-\tilde{\mu}_{im})(o_t-\tilde{\mu}_{im})'}{\sum_{i=1}^{T}\delta_{im}(t)},$$

$$\tilde{\omega}_{im}=\frac{\sum_{i=1}^{T}\delta_{im}(t)}{\sum_{i=1}^{T}\delta_{i}(t)},\quad \tilde{\alpha}_{ij}=\frac{\sum_{i=1}^{T}\alpha_{ij}(i)a_{ij}b_j(o_{t+1})\beta_{t+1}(j)}{\sum_{t=1}^{T}\alpha_t(i)\beta_t(i)}。$$

在前面介绍的三个问题种，HMM 的训练，也就是参数估计问题，是 HMM 在模式识别应用的关键问题，同其他两个问题相比，这也是最困难的一个问题。

HMM模型的应用

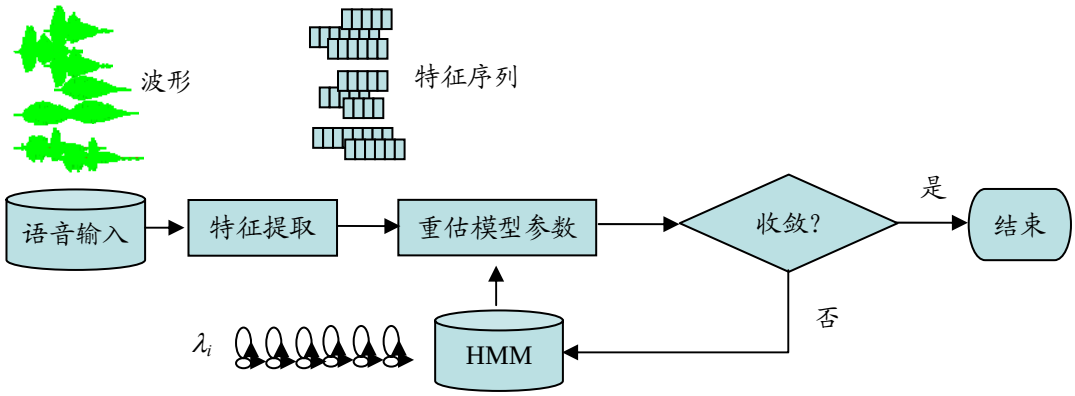
近年来，国内外对 HMM 模型的研究不断增加，其中，在语音识别方面的应用便取得了巨大的见效。随着其应用日趋广泛，许多相关领域都试图引入该模型。本文尝试将 HMM 模式识别模型引入到股票价格波动预测问题中，通过解决 HMM 模型中的学习问题和识别问题，构建两个股票波动模式识别模型，并相应地运用该模型对股票指数波动情况进行预测。

语音识别

HMM 是序列数据处理和统计学习的一种重要概率模型，近年来已经被成功应用到许多语言处理的任务中，且取得了很好的成果。经典 HMM 语音识别过程如下：

（1）首先，从输入的语音中提取相应的数字特征序列，并对模型进行训练，得到局部最优参数估计。HMM 语音识别模型训练过程如下图：

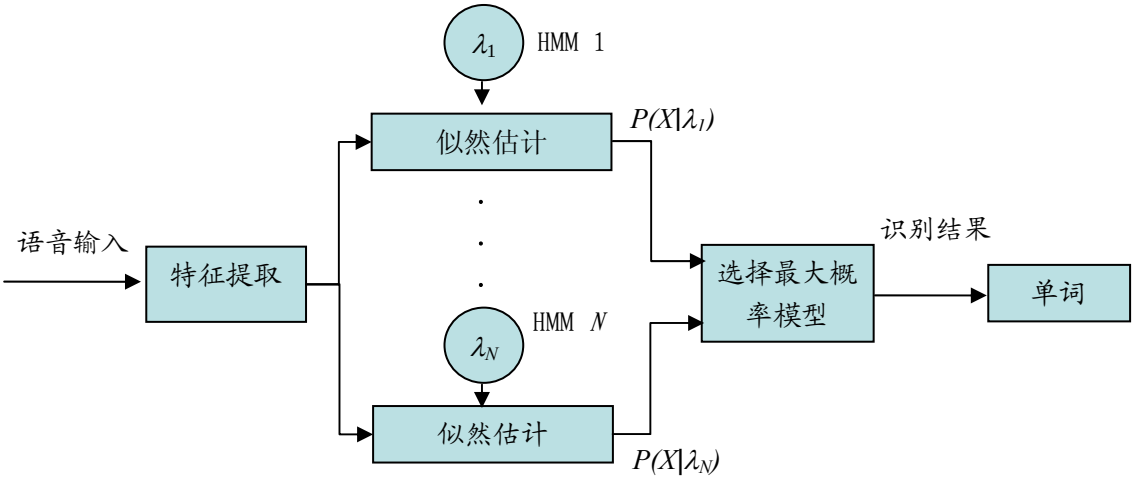
图表 2 经典 HMM 语音识别训练过程



数据来源：广发证券发展研究中心

（2）其次，输入需要进行识别的语音，通过提取相应的数字特征序列，再运用向前-向后算法对各类模型进行似然估计，得到最大概率的模型输出，从而实现识别功能。HMM 模型语音识别过程如下图：

图表 3 经典 HMM 语音识别过程



数据来源：广发证券发展研究中心

股市预测

股市预测，是金融工程的一个重要分支，也是一个重大难题。通常的股票预测模型都基于以下三个基本假设：

1) 有效市场假设：当证券价格能够充分地反映投资者可以获得的信息时，我们称证券市场就是有效的，即在有效市场中，无论随机选择何种证券，投资者都只能获得与投资风险相当的正常收益率；

2) 供求决定假设：即认为股票的价格是由供需关系决定的，企业的盈利和股息的影响是很小；

3) 历史相似假设：假定根据过去资料建立的趋势外推模型能适合未来，能代表未来趋势变化的情况，即未来和过去的规律一样，或者是说，历史是相似甚至历史会重演。

启发于 HMM 模型在语音识别中的应用，我们尝试将其思想引入到股票价格预测中。并补充如下两个模型假设：

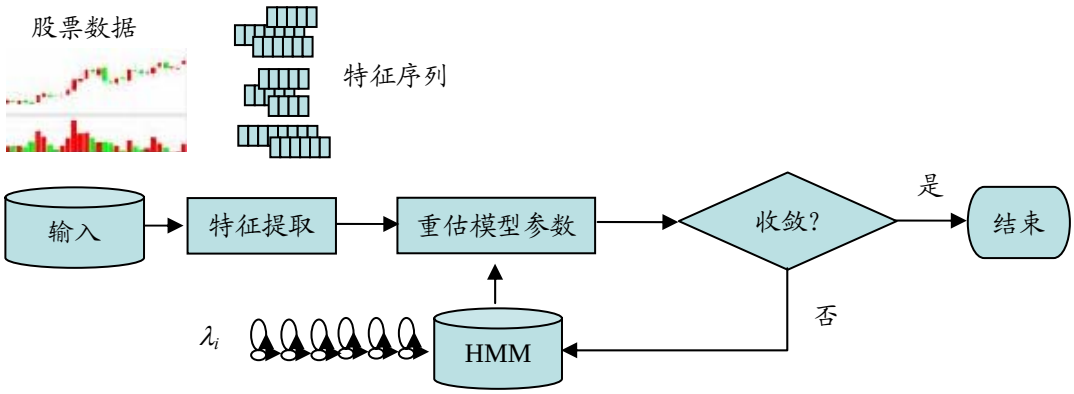
4) 首先，对股票价格未来走势进行分类，假设股票指数的每一种涨跌情况都存在一种明确的模式与其对应，可由一个 HMM 模型来表示；

5) 假设股市的外在数据表现由有限个服从马尔可夫过程的隐状态所决定，即任何一个时段的股市表现由一个隐藏的马尔可夫链所决定。

基于以上五个假设，我们提出基于 HMM 模式识别模型的股市走势预测思想如下：

（1）首先，按照事先分类，选取历史上属于同类走势的日期以及该日期之前若干个星期的股票数据，提取股票数据中某些特征指标（成交价格，成交量，等等）形成相应的序列作为模型的输入，并应用 Baum-Welch 算法对各类模型进行训练，训练过程如下图：

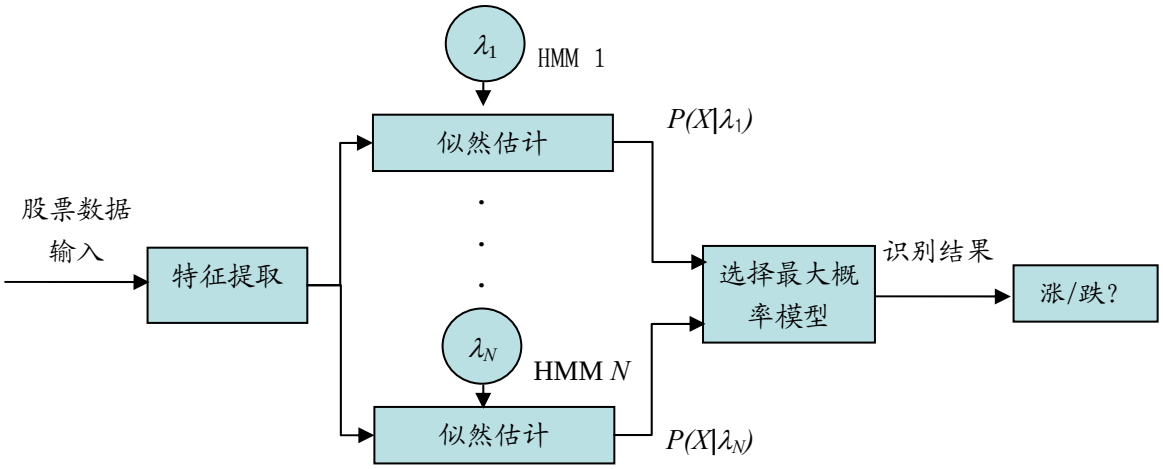
图表 4　HMM 股票走势预测模型训练过程



数据来源：广发证券发展研究中心

（2）其次， 根据训练好的 HMM 模型，选取若干个星期的股票特征指标（成交价格，成交量，等等）序列作为输入，应用向前－向后算法计算各个模型发生的概率，选取最大概率对应的模型，从而得到下一阶段股票走势的识别结果。识别过程如下图：

图表 5　HMM 股票走势预测模型识别过程



数据来源：广发证券发展研究中心

模型参数选择

变量选择

在应用 HMM 模型对股票走势进行预测时，首先需要解决的关键问题是输入变量的选择。目前大多数择时模型所选用的输入变量可分为公司基本面、技术指标以及市场资金信息三大类。在前期的报告《基于高频数据的市场情绪择时研究》中，我们对基于资金面的择时模型进行了研究，发现资金面的波动能够有效地衡量股票市场的波动，且对于择时具有一定的指导意义。因此，本报告同样基于资金面对股票波动的考虑出发，首先从股票市场的高频数据中提取出以下三个基础指标，他们是：每日收盘价、

每日资金流入和每日资金流出，关于资金流入和资金流出的详细定义参见前期的报告《基于高频数据的市场情绪择时研究》。由以上三个指标，再构造出以下我们认为对股价具有较强相关性的四个模型输入指标：

X_1：股票日收益率；

X_2：资金日净流入占当日所有流动资金的比例；

X_3：日总流动资金环比；

X_4（日总流动资金－过去一年平均流动资金)/过去一年流动资金波动率。

数据及参数选择

在 HMM 股票波动模式识别模型中，我们选取了沪深 300 指数 2005 年 4 月 8 日上市至 2010 年 4 月 30 日的数据进行实证分析，并假设观测概率服从连续高斯分布，隐状态数量为 Q=3。经实验，我们发现以(X_3, X_4)，(X_2, X_3, X_4)和(X_1, X_2, X_3, X_4)作为输入向量，且每类训练样本数量大约为 30～40 时得到的预测结果比较好。本报告中，我们分别将股票的波动模式分为两类（涨、跌）和三类（涨、跌、平），并对其进行分析和比较。

下面，我们给出针对不同训练样本数量、样本长度以及不同输入向量对应的预测结果。

图表 6 分为“涨”、“跌”两类模式的预测准确率

样本数量 / 样本长度	X_3 X_4			$X_2X_3X_4$			$X_1X_2X_3X_4$		
	300	400	500	300	400	500	300	400	500
5	0.5111	0.5000	0.5474	0.5556	0.5487	0.5898	0.5704	**0.6078**	0.5579
10	0.5481	0.5043	0.5053	0.5259	0.5043	0.484	0.5333	0.5000	0.4842
15	0.5111	0.5478	0.5263	0.5333	0.5565	0.5579	0.5407	0.5739	0.5579
20	0.5407	0.5739	0.5579	0.563	0.5652	0.5474	0.5481	0.5739	0.5684

数据来源：广发证券发展研究中心

图表 7 分为“涨”、“跌”、“平”三类模式的预测准确率

样本数量 / 样本长度	X_3 X_4			$X_2X_3X_4$			$X_1X_2X_3X_4$		
	300	400	500	300	400	500	300	400	500
5	0.3704	0.4348	0.4211	0.3481	0.4174	0.4421	0.3704	0.4000	**0.4737**
10	0.3778	0.3739	0.4316	0.3178	0.4300	0.4211	0.3852	0.3650	0.4210
15	0.4074	0.4000	0.4105	0.3302	0.4087	0.4211	0.3402	0.4000	0.4100
20	0.3235	0.4348	0.3895	0.3462	0.4174	0.4316	0.3128	0.3910	0.4526

数据来源：广发证券发展研究中心

由实验结果我们发现：在两类模型中，当训练样本长度为 T=1 周即五个工作日，且选取输入向量为(X_1, X_2, X_3, X_4)时，预测准确率较高；分别设置两个概率阈值 0.57 和 0.43，并相应地标记出两类模型中准确率较高的参数组合。可以看出，在两类股票波动模式模型中，当训练样本数量为 400 周，预测准确率最高为 P_1=0.6087>0.5，在三类股票波动模式模型中，

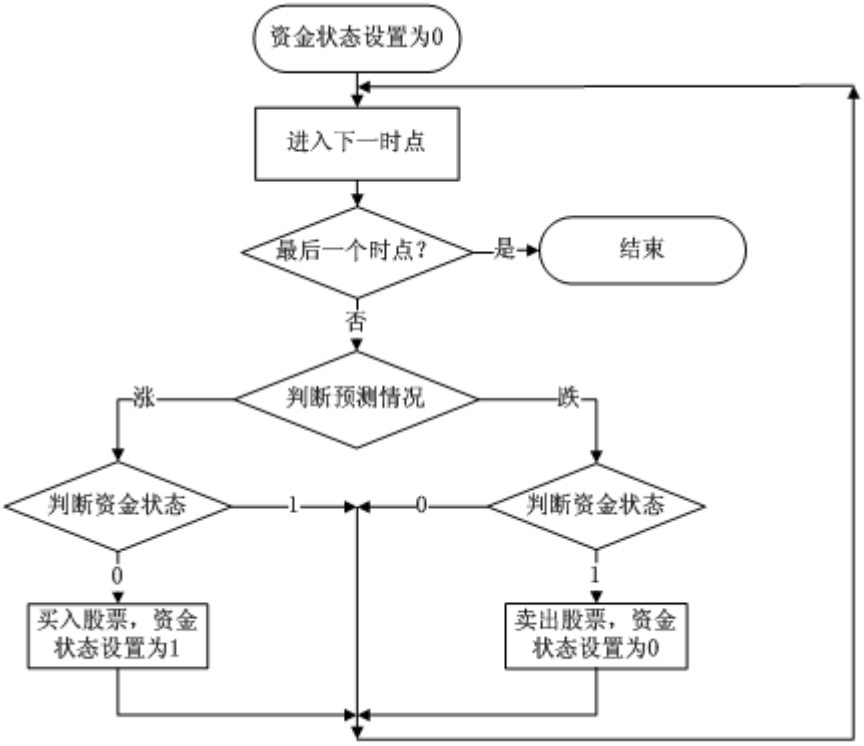
当训练样本数量为 500 周时，预测准确率最高为 P_2=0.4737>0.33。

算法和结果分析

两类波动模式下的择时策略及交易结果

基于上一节所确定的各个模型参数，下面我们首先给出两类股票波动模式情况下对应的股票交易择时策略，具体如下图所示。

图表 8　两类波动模式的 HMM 预测模型交易择时策略

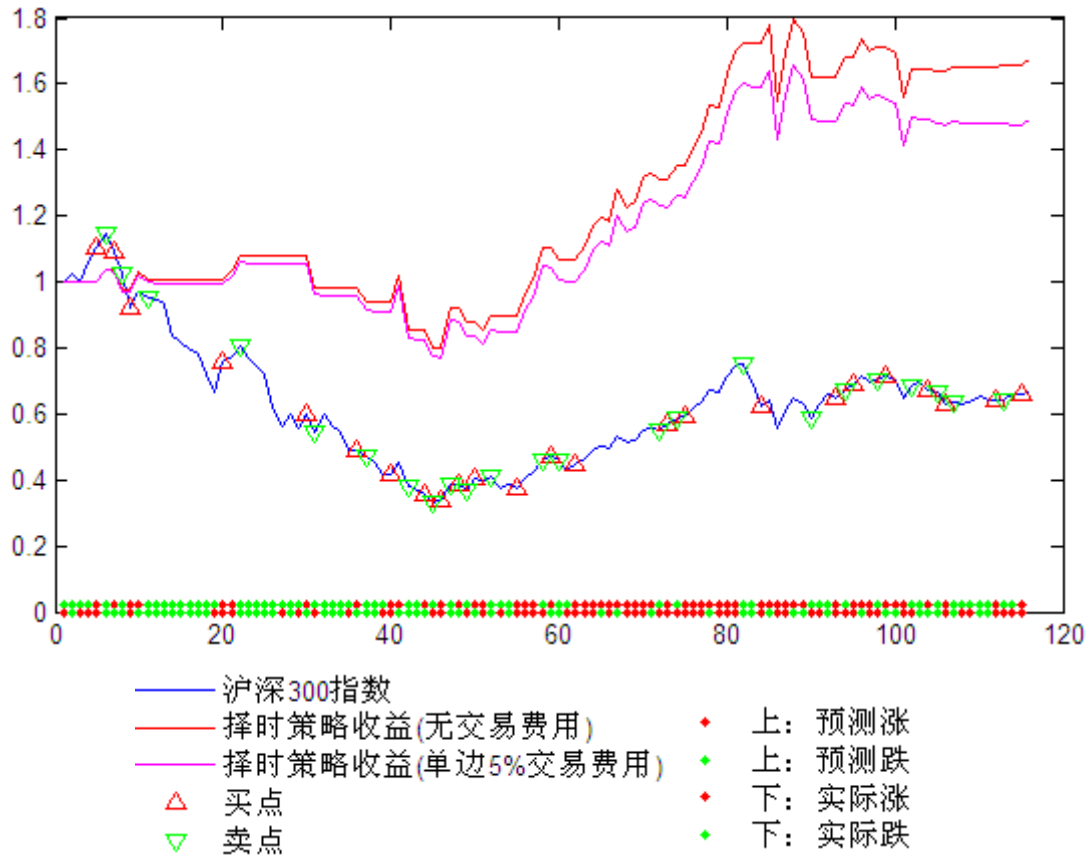


数据来源：广发证券发展研究中心

应用以上择时策略进行交易，我们得到交易的情况如下：

在 2007 年 12 月 19 至 2010 年 4 月 29 日共 115 周间，共发出了 24 次买入信号和 23 次卖出信号，平均每 2.45 周交易一次。其中，预测结果准确为 70 周，准确率为 60.87%高于随机预测概率 50%，若根据预测情况进行模拟交易，则期末的资产收益率为 67.84%，而同期沪深 300 指数的收益率为-33.3%，两者资产比值为 2.5162。若考虑 0.5%的单边交易费率，则模拟交易的期末收益率降为 49.56%。

图表 9　两类波动模式的 HMM 预测模型交易结果



数据来源：广发证券发展研究中心

通过对表 9 的判读，我们还可以得到如下结果:

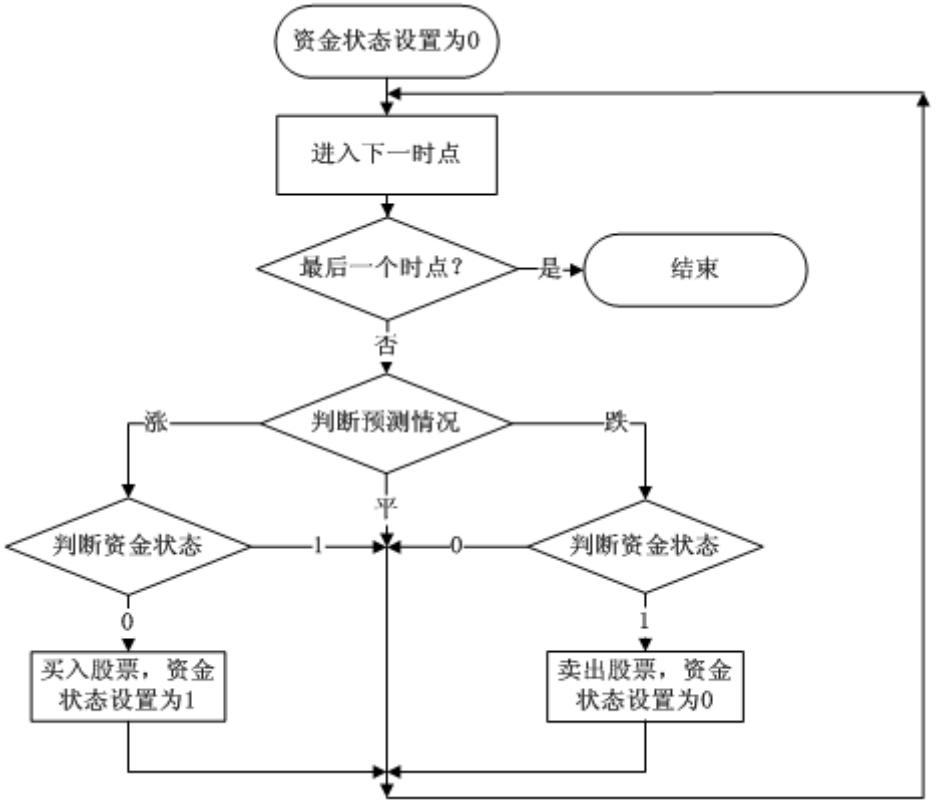
（1） 在 2007 年 12 月 19 至 2010 年 4 月 29 日共 44 周的单边下跌行情以及 2009 年 1 月 20 日至 2009 年 8 月 3 日共 26 周的单边上涨行情中，模型有较好的预测效果，预测准确率分别为 65.9%和 69.2%。这就保证了该预测模型能够有效地避开单边下跌行情，同时比较准确地判断了单边上涨行情，从而保证了较高的收益;

（2） 由于本模型中仅取“涨”、“跌”两种模式，根据我们设计的交易择时策略，则会出现较为频繁的交易操作，尤其在考虑交易费用的情况下，收益将明显下降，这将给投资者带来一定的困难。

三类波动模式下的择时策略及交易结果

下面，我们换一个角度来进行思考，由于股市每时每刻都处于波动状态，波动的方向以及幅度都会对我们的投资结果产生巨大的影响，由于考虑到了交易费用，因此，我们有必要避免部分无谓的买卖操作，这就要求我们对股市的波动进行重新分类。下面，假设当预测股票指数的周收益率位于区间[-2%，2%]时，我们认为没有必要对股票进行买卖操作。相应地，可将股票的波动模式分为“涨”、“跌”、“平”三类，并给出对应的交易择时策略如下图所示。

图表 10　三类波动模式的 HMM 预测模型交易择时策略

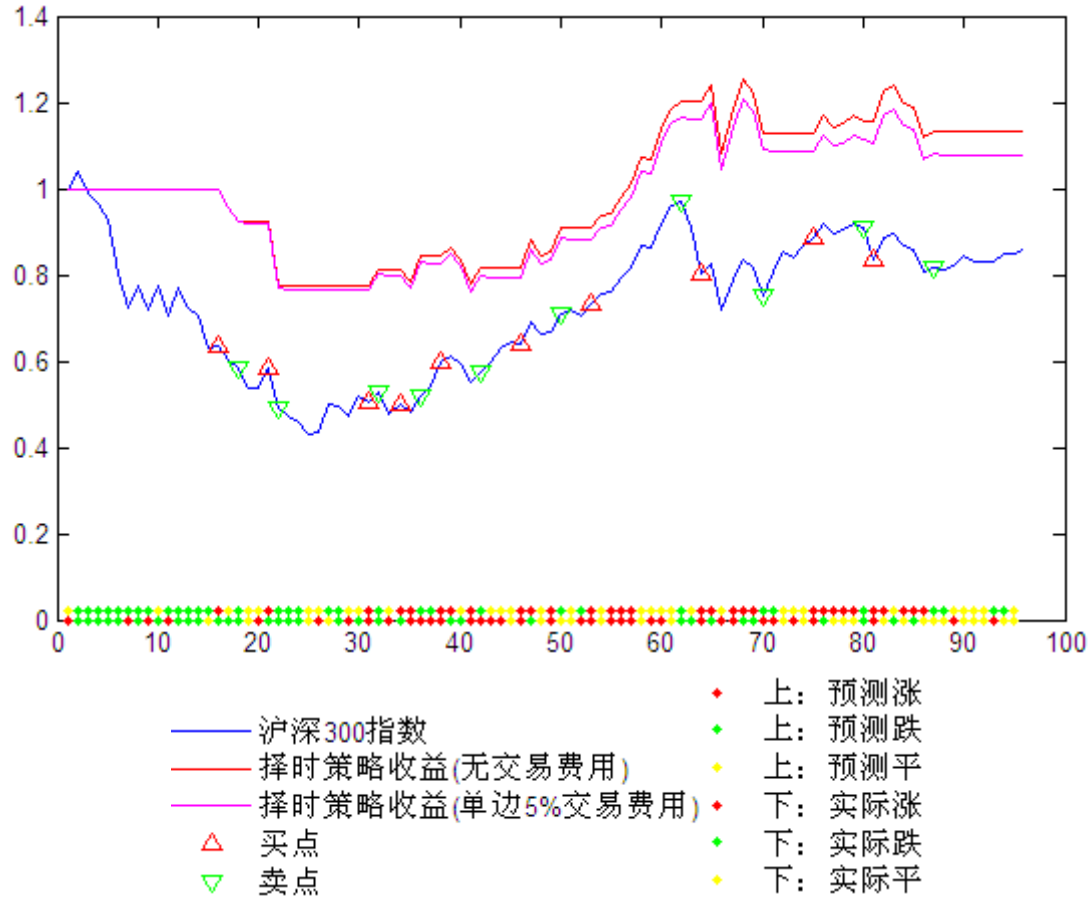


数据来源：广发证券发展研究中心

应用以上择时策略进行交易，我们得到交易的情况如下：

在 2008 年 5 月 12 至 2010 年 4 月 29 日共 96 周间，共发出了 10 次买入信号和 10 次卖出信号，平均每 4.8 周交易一次。其中，预测结果准确为 45 周，准确率为 47.37%高于平均概率 33.3%。若根据预测情况进行模拟交易，则期末的资产收益率为 13.08%，而同期沪深 300 指数的收益率为-14.2%，两者资产比值为 1.3152。若考虑 0.5%的单边交易费率，则模拟交易的期末收益率降为 7.55%。

图表 11　三类波动模式的 HMM 预测模型交易结果



数据来源：广发证券发展研究中心

通过对表 11 的判读，我们还可以得到如下结果：

（1） 类似地，在 2008 年 5 月 12 至 2008 年 10 月 31 日共 24 周的单边下跌行情以及 2009 年 1 月 20 日至 2009 年 8 月 3 日共 26 周的单边上涨行情中，模型有较好的预测效果，预测准确率分别为 58.33%和 53.84%，远高于平均概率 33.3%。这就保证了该预测模型能够有效地避开单边下跌行情，同时准确判断单边上涨行情，从而保证了较高的收益；

（2） 由于本模型中将股票波动分为“涨”、“跌”、“平”三种模式，因此，在股票指数波动较为频繁的时候，根据我们设计的交易择时策略，则能够避免由于频繁交易操作而耗费大量的交易费用的问题。由图 10 可以看出，在考虑交易费用的情况下，该择时策略的期末累计收益下降并不是特别明显，从而为投资者提供了一个更为稳健的操作策略。

总　结

研究意义和创新点

本报告首次提出将 HMM 模式识别模型引入到股票价格波动预测问题中，通过解决 HMM 模型中的学习问题和识别问题，从而建立了一个基

于股票日收益率以及日资金流等变量的对股票指数择时模型，经实证检验，无论是预测准确率和择时策略收益，该模型都取得了比较不错的效果，具有相当的理论和现实意义。

由于 HMM 模型的相关算法相当成熟，且具有效率高，效果好以及易于通过已有的数据进行模型训练等特点，因此选用 HMM 模型进行股票波动模式识别不仅是一个较大的创新，更是一个值得探讨的选择。

模型的不足

本报告提出的模型虽然取得了较好的预测结果，但是仍然存在以下不足之处：

（1）预测准确率尚不够高，在两类波动模式和三类波动模式中，模型的预测准确率分别为 60.87%和 47.37%，均有待进一步提高；

（2）输入向量所选的时间段过短，由于本文的实证数据选取了沪深 300 指数，数据始于 2005 年 4 月 8 日，而在提取相应的特征向量时，则截取了前 250 个数据用于计算后续数据的波动率；在剩下的数据中，又截取了部分数据作为模型训练样本，这就导致了用来测试的样本数据有限，难以有效地代表我国股市的历史特征；

（3）输入向量的选择比较局限，由于本研究主要是基于资金流对股市波动的影响，因此所选的输入变量在某种程度上均与资金流有关，这就可能导致我们根据输入向量所训练出来的模型，所能提取的股市波动规律受到局限，从而导致模型的预测准确率有待提高。

后续研究方向

由于本报告只是对 HMM 模型在股市择时研究中的一个初步尝试，在随后的系列报告中，我们将继续对量化交易策略进行研究，并将对该模型作进一步的修正和改进，初步有如下几点改进设想：

（1）选用更多的指数数据，通过大量的实证，寻找应用 HMM 模型的最有效标的指数；

（2）本篇报告对指数周行情即中短期趋势进行预测，未来我们会对日内的短期趋势进行实证研究。

（3）鉴于市场对股指期货的关注度日趋加大，部分投资者开始尝试通过股指的杠杆效应进行套利操作，因此接下来我们将可能尝试对与股指关系较为紧密的一些指标进行检验，以期能够找到新时期的股市波动规律。

广发证券—公司投资评级说明

买入（Buy）	预期未来 12 个月内，股价表现强于大盘 10％以上。
持有（Hold）	预期未来 12 个月内，股价相对大盘的变动幅度介于-10%～+10%。
卖出（Sell）	预期未来 12 个月内，股价表现弱于大盘 10％以上。

广发证券—行业投资评级说明

买入（Buy）	预期未来 12 个月内，行业指数优于大盘 10％以上。
持有（Hold）	预期未来 12 个月内，行业指数相对大盘的变动幅度介于-10%～+10%。
卖出（Sell）	预期未来 12 个月内，行业指数弱于大盘 10％以上。

相关研究报告

	广州	深圳	北京	上海
地址	广州市天河北路 183 号大都会广场 36 楼	深圳市民田路华融大厦 2501 室	北京市月坛北街 2 号月坛大厦 18 层 1808 室	上海市浦东南路 528 号证券大厦北塔 17 楼
邮政编码	510075	518026	100045	200120
客服邮箱	gfyf@gf.com.cn			
服务热线	020-87555888-612			

注：本报告只发送给广发证券重点客户，不对外公开发布。

免责声明