文章编号：1002-1566(2016)02-0215-10
DOI: 10.13860/j.cnki.sltj.20160322-014

财经新闻与股市预测 — 基于数据挖掘技术的实证分析

孔翔宇 毕秀春 张曙光

(中国科学技术大学 统计与金融系，安徽 合肥 230026)

摘要：本研究深度挖掘了财经新闻主题内容与股市市场的相关性，并提出了一种基于理解当日新闻主题分布来分析中国股市涨跌的预测模型。具体来说，我们使用自动文本分析技术与机器学习技术，首先通过概率主题模型对财经新闻文档进行聚类得到其中的主题分布，再结合实际股票市场的交易数据分析其与市场之间的关联程度，最后引入支持向量机算法对股市走势进行预测。实验部分我们抽取了近三个月的新闻数据与沪深股市数据进行分析，结果表明：新闻中国际贸易以及城市化相关主题与股市变动关系密切，通过本文提出的算法能较准确得预测当日股市涨跌，而建立在其上的股指期货策略也取得了很好的效果。

关键词：数据挖掘；潜在狄利克雷分配；支持向量机

中图分类号：F832, O212 **文献标识码**：A

Financial News and Prediction for Stock Market: An Empirical Analysis Based on Data Mining Techniques

KONG Xiang-yu BI Xiu-chun ZHANG Shu-guang

(Department of Statistics and Finance, University of Science and Technology of China, Anhui Hefei 230026, China)

Abstract: In this paper, we attempt to explore the correlation between the topic distribution of financial news and the movement of stock market, and therefore deliver a prediction for stock market by analyzing the distribution. Specifically, we adopt probabilistic topic model with data mining techniques, and maintain the topic distribution by clustering financial news documents. The distribution is thereafter analyzed in conjunction with actual market data to understand its impact on the market, and is used for the prediction with aid of support vector machines algorithm. As shown in the experimental result, the topics about international commerce and utilization are strongly related with the movement of stock market, besides, an accurate prediction system is proposed and the trading strategy based on it delivers good performance.

Key words: data mining, latent Dirichlet allocation, support vector machines

收稿日期: 2014 年 10 月 14 日
基金项目: 国家重点基础研究发展计划项目 (973:2007CB814901) 资助。

0 引言

近年来，量化投资在国内渐渐引起重视，市场上量化型基金产品层出不穷。量化投资简单来说就是利用数学、统计学、信息技术来管理投资组合，与定性投资相比，投资广度得到扩大，客观性纪律性也得到加强。随着量化投资分析的蓬勃发展，量化分析大致可以分为两类：传统的使用数理统计方法比如分形理论、随机过程、小波分析等；新兴的使用数据挖掘技术比如机器学习、神经网络、支持向量机等。

传统的金融数据分析处理的是大量反映经济表现的结构化数据，然而大数据时代带来的是巨大的数据体量和繁多的数据类型，结构化数据已经逐渐不能完全满足量化投资分析的需求。最近海外有对冲基金尝试将文本挖掘 (Textual Mining)[1] 技术与机器学习 (Machine Learning) 技术 [2] 应用在对单个公司股价预测上，整合从各种渠道提取的信息以建立操作策略，而这些技术可以从相关新闻报道中提取与公司有关的信息，借此解释其股价的实时变化。因此本文将利用概率主题模型 (Probabilistic topic model)[3] 对财经新闻进行语义分析，再与实际市场数据结合分析，采用支持向量机 (Support vector machine)[4] 算法实现对市场进行预测的目的。

本文构建了一个分析财经新闻的文本挖掘系统以考察其与实际沪深股票市场的行为的相关度，目标是识别出对股票市场有冲击的主要事件，分析这些事件的出现特征并利用其预测市场变化趋势。使用的数据包括沪深指数从 2014 年 1 月 1 日到 2014 年 3 月 14 日间的时间序列数据，以及同时间段从 ChinaDaily 新闻网站取得的带有时间戳的财经新闻英文文本。而用来衡量沪深股市的指标有：每日收盘价，每日波动率与日内变动百分比。

其中每日收盘价指的是沪深 300 指数的每日收盘价格，波动率指的是其每日最高价格与最低价格间的差，而日内变动百分比 (percentage change)，定义为：. .

$$\text{日内变动百分比} = \frac{\text{股指开盘价}}{100 \times \text{股指收盘价股指开盘价}}. \tag{1}$$

构建的系统主要包含以下三部分：

1. 通过对财经新闻库的主题做聚类分析，识别主要事件。
2. 分析这些事件与股票市场表现的相关联系。
3. 建立利用当前新闻预测股指走势的模型。

本文剩余内容将组织如下：第二部分将概述国内外的相关研究工作；第三部分中 LDA (Latent direclet allocation)[5] 算法将被应用与财经新闻的文本分析中，分析后每篇新闻是一组主题上的概率分布；第四部分将每日的新闻主题分布与当日股市参数进行关联，分析不同主题对股市的影响并建立股市走势的预测模型；最后是本研究的结论及分析。

1 文献综述

随着互联网的蓬勃发展，特别是 Web2.0 与大数据时代的到来，通过数据挖掘技术来分析股市受到越来越多的学者与工业界的关注。Wutrich et al (1998)[6] 认为财经新闻报纸中的文章内容除了市场表现结果外还包含造成其的潜在原因，因此利用 k-NN 聚类算法和神经网络模型等挖掘这些文本中的信息以对全球主要股市指数的当天收盘价进行预测。Lavrenko 等 (2000)[7] 提出了一个预测股价趋势的系统，通过分段线性拟合将股价时间序列中的趋势找出，使用自然语言模型描述新闻文字中的特点以寻找最可能影响未来趋势的新闻内容。Kloptchenko 等 (2002)[8] 指出公司财务报表中文字部分包含远比财务指标多的信息，而对前者的分析手段

寥寥无几，因此将数据挖掘方法结合到财务报表的分析中以找到决定未来公司表现的隐含信息。Mittermayer (2004)[9] 提出实现了一个对新闻发行后股价的即时趋势的预测系统，流程分为三步：通过文本处理技术应用从新闻中提取相关信息，将这些新闻分入预设的分类，通过对分类中内容分析构建交易策略。Seo 等 (2004)[10] 认为公司财务展望方面的新闻报道与其对投资者的吸引力有正相关的关系。然而鉴于人们无法追踪阅读每一条新闻报道，利用新的文本分类方法及抽样方法，实现了自动分析新闻报道反应的公司未来表现倾向的系统。Ingvaldsen 等 (2006)[11] 搭建了结合信息检索，信息抽取和自然语言处理技术的数据挖掘框架，并利用此框架从信息流中抽取关键元素，并分析其与市场的相关度。

而金融市场预测的研究工作除了传统的时间序列方法之外，也与支持向量机算法结合应用。Trafalis 等 (2000)[12] 验证了支持向量机算法在金融预测应用上的表现优于其他技术，但是其训练过程会导致二次规划问题。Tay 等 (2001)[13] 将支持向量机算法应用在金融时间序列预报领域，实验结果证实其表现在各种衡量标准下都优于神经网络算法。Yang 等 (2002)[14] 将支持向量回归技术应用在金融预测任务中，通过调整间隔大小反应出金融数据中波动性的变化，最后给出了恒生指数的良好预测。Huang 等 (2004)[15] 将支持向量机在公司信用评级分析的表现与神经网络算法相比较，并建立了关于美国与台湾市场间区别因素的市场比较分析。Cao 等 (2005)[16] 基于支持向量机技术，通过利用每日汇率数据训练模型，成功建立了外汇汇率预测模型。Bao 等 (2005) 提出了模糊支持向量机回归算法，并将其应用在股票成分指数预测方面，在并数据处理过程中显示了其算法的优良效果 [17]。

国内关于此领域的研究，赵丽丽等 (2012)[18] 采用文本挖掘技术将财经新闻内容量化为影响因子，并利用多元回归分析其对中国股市的影响。本文采用潜在狄利克雷分配 (LDA) 方法对财经新闻进行深度挖掘，并实现了对股市未来走势的预测。

2 语义分析财经新闻

对于理解文本研究内容最为关键的一步是确定其中隐含的核心主题。我们采用潜在狄利克雷分配 (Latent Dirichlet Allocation, LDA) 算法从财经新闻中聚类分析其主题。

LDA 是一种统计模型 [5]，确切的说是一个概率主题模型，其中每个文档被认为由一些主题组合形成，而文档中每个词语都与这些主题之一相关联。给定一个文档的语料库，LDA 可以完成以下步骤：确定一个主题集合；对每一个主题关联一个词语集合；对每个文档定义一个不同主题构成的比例。

我们通过词语，文档，语料库等概念出发建立 LDA 模型。

定义 3.1 词语 (word) 是离散数据中的基本单位，定义为一个词汇表上索引标记为 $\{1,\cdots,V\}$ 的条目。我们使用只有一个元素为 1 其余为 0 的单元基础向量来代表词语，于是用上标表示内容，词汇表中的第 v 个词语用 V- 向量 w 代表，因此当 $u \neq v$ 时 $w^v = 1, w^u = 0$。而文档 (document) 定义为 N 个词语的序列，记为 $W = (w_1, w_2, \cdots, w_N)$，其中 w_n 是序列中第 n 个词语。语料库 (corpora) 则是 M 个文档的集合，记为 $D = \{W_1, W_2, \cdots, W_M\}$。

LDA 模型是一个从语料库 D 生成的概率主题模型，其基本思想是文档由不同的主题的随机混合而成，而这些主题通过在词汇表上的离散分布来刻画。

因此对全集 D 中的每个文档 W，LDA 模型假定以下的生成过程：

(1) 随机选择文档中词语的数量 N：$N \sim Poisson(\xi)$。

(2) 从狄利克雷先验分布随机确定主题的参数 θ：$\theta \sim Dir(\alpha)$。

其中 $\theta=(\theta_1,\cdots,\theta_k),\alpha=(\alpha_1,\cdots,\alpha_k)$，$\theta_i$ 代表第 i 个主题被选中的概率，而 k 维狄利克雷分布 (Dirichlet distribution) 定义在 $(k-1)$ 维单形体上：

$$\Delta^{k-1}=\Big\{(\theta_1,\cdots,\theta_k)\in R^k\Big|\sum_{i=1}^{k}\theta_i=1,\ \text{且}\ \forall i,\theta_i>0\Big\},$$

$$\text{且}\ Dir(\theta_1,\cdots,\theta_{k-1};\alpha_1,\cdots,\alpha_k)=\frac{1}{B(\alpha)}\prod_{i=1}^{k}\theta_i^{\alpha_i-1},\quad \text{其中}\ B(\alpha)=\frac{\prod_{i=1}^{k}\Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{k}\alpha_i)}.$$

(3) 利用参数 θ 和多项分布随机生成一个主题：$z_n\sim Multi(\theta)$。

(4) 根据已有的主题 z_n，从其条件多项分布选择一个词语：$w_n\sim p(w_n|z_n,\beta)$。

其中 $k\times V$ 维矩阵 β 是将词语出现频率的参数化，$\beta_{ij}=p(w^j=1|z^i=1)$。

我们从 ChinaDaily 取得 2013 年 11 月 29 日至 2014 年 3 月 14 日间共 73 日的关于中国经济的 557 个英文财经新闻文档，组成供 LDA 分析的语料库。将语料库中所有词语编为词汇表，并通过统计每个文档中词语出现的次数把语料库转化为 557 个词语频数的向量，表明了文档中出现的词语在词汇表中对应位置。原始词汇表中有 10786 个词语，我们将冠词，介词，连词和代词等对文档内容无意义且出现频率过高的 54 个词语删去，使用剩下的 10732 个词语组成的词汇表对 557 个数组实施 LDA 算法，从处理过的语料库识别聚类出 25 个主题，表 1 是每个主题中出现频率前 5 的关键词。

表 1 主题对应频率最高的词语

序号	第 1 个词语	第 2 个词语	第 3 个词语	第 4 个词语	第 5 个词语
Topic1	shanghai	financial	zone	ftz	foreign
Topic2	hungary	france	business	greece	gao
Topic3	2014	gdp	bank	economist	target
Topic4	migrant	workers	cases	home	million
Topic5	province	guangdong	city	regions	area
Topic6	cooperation	countries	visit	asia	asian
Topic7	world	global	emerging	economies	countries
Topic8	kong	hong	mainland	students	innovation
Topic9	grain	food	agricultural	gm	prices
Topic10	projects	project	construction	infrastructure	city
Topic11	reform	premier	national	session	central
Topic12	pension	party	population	committee	social
Topic13	negotiations	fta	air	talks	round
Topic14	not	years	been	can	such
Topic15	workers	people	my	who	job
Topic16	income	rural	region	people	poverty
Topic17	australia	britain	zealand	uk	past
Topic18	companies	private	state	owned	capital
Topic19	cities	urbanization	estate	real	housing
Topic20	companies	industry	manufacturing	industries	products
Topic21	eu	us	africa	says	fdi
Topic22	overseas	brands	countries	global	water
Topic23	index	month	december	january	manufacturing
Topic24	debt	local	governments	financing	trillion
Topic25	trillion	total	3	1	8

聚类与分类不同，是把相似的对象分为不同的子集的过程，因此聚类出的主题未必都有实质的意义，比如第 14 个和第 25 个主题。但是可以看出，通过 LDA 算法聚类出的其他主题都有各自的核心意义。

我们从这 25 个主题中提炼归纳出以下 10 个主题：自由贸易区谈判 (如第 1, 13)；国际商贸合作 (如第 2, 6, 10, 17, 21)；宏观经济指标 (如第 3, 23)；农村劳动力转移 (如第 4, 15, 20)；区域经济发展计划 (如第 5, 16, 24)；新兴经济体发展 (如第 7)；海内外留学交流 (如第 8, 22)；食品安全与农业经济 (如第 9)；私企发展与国企改革 (如第 11, 18)；养老金并轨等城市化问题 (如第 12, 19)。

进一步的，我们通过聚类分析得到主题的结果后，可以计算得到每个文档中各个主题所占比例情况，将之汇总得到 25 个主题在整个语料库中的占比如图 1。可以看出没有实质意义的第 14 个和第 25 个主题总体占比最高。

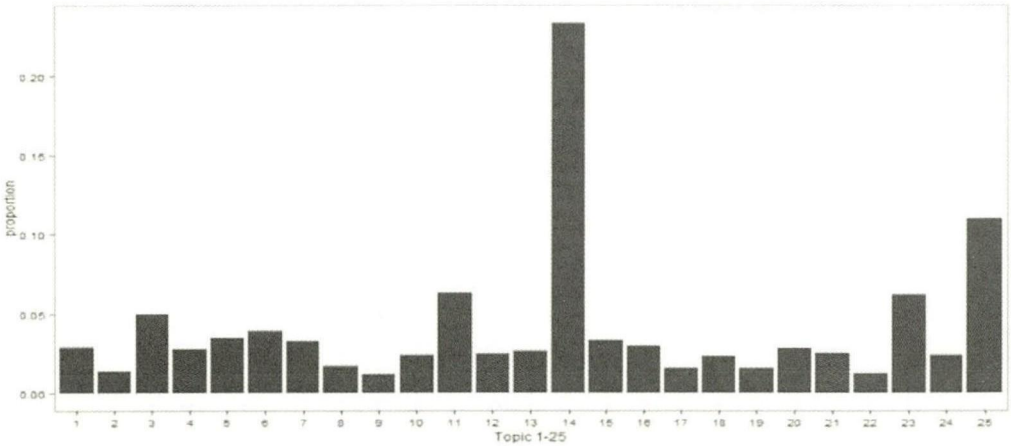

图 1 25 个主题在语料库中总占比

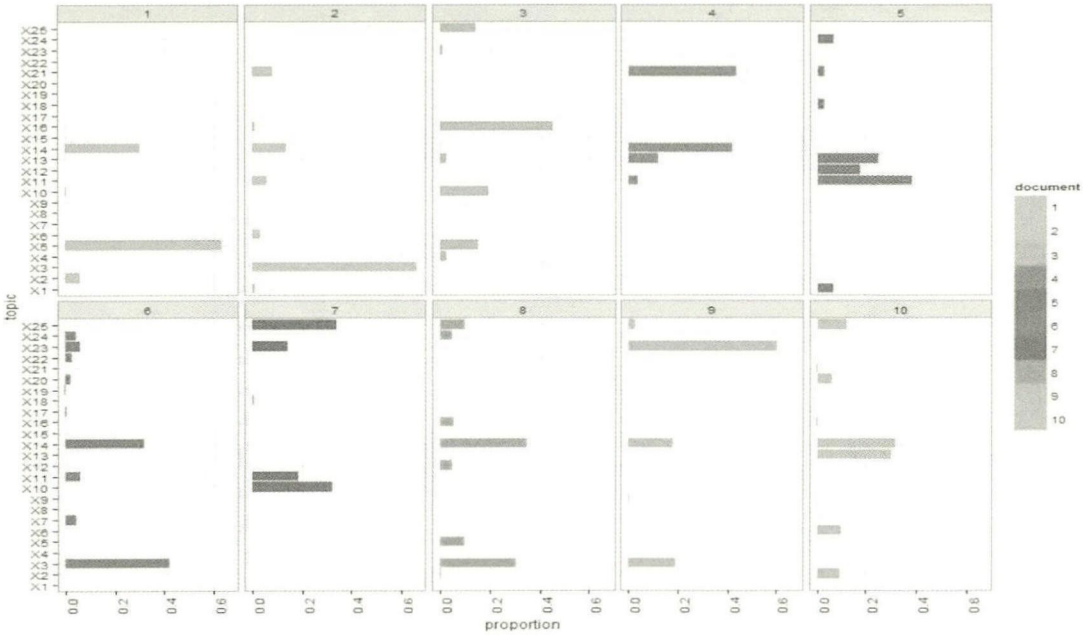

图 2 前 10 个文档中的主题构成

更加具体的，我们将语料库的 557 个文档中前 10 个文档取出，图 2 显示了这 10 个文档中 25 个主题的占比情况。

可以看出每个文档除了由无实质意义的主题外，还包含有意义的主题，例如第一个文档中占比最多的是第 5 个关于国内区域经济发展的主题，第二个文档中占比最多的是第 2 个关于宏观经济调控的主题。但是由于分布较发散，因此有意义的主题总体占比并不高。

接着，从国泰安金融数据库读取 2014 年 1 月 1 日至 2014 年 3 月 14 日间共 47 个交易日的沪深股指的每日收盘价，每日波动率与日内价格变动百分比。然后将 557 个文档按照各自的时间戳计算出 47 个交易日对应的 25 个主题所占比例的数值，并通过计算 Pearson 相关系数考察每个主题与股市趋势之间的关联程度。其中，Pearson 相关系数的计算公式如下：

$$r=\frac{\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\overline{X})^2\times\sum_{i=1}^{n}(Y_i-\overline{Y})^2}}. \tag{2}$$

得到 25 个主题与三个指标间的相关系数如表 2 所示。

表 2 25 个主题与股市指标间 Pearson 相关系数

主题序号	每日收盘价	每日波动率	日内变动百分比
Topic1	−0.0252	0.1856	−0.2336
Topic2	0.1785	−0.1134	0.2884
Topic3	−0.1142	−0.0454	0.1836
Topic4	0.0374	0.0544	−0.1146
Topic5	−0.1233	0.0829	−0.0705
Topic6	0.1583	−0.1740	0.2253
Topic7	−0.0436	−0.2376	−0.0268
Topic8	0.0039	0.1411	0.1852
Topic9	0.2348	−0.0656	0.0591
Topic10	0.0373	0.0496	−0.0418
Topic11	−0.2242	0.0789	0.1619
Topic12	0.0910	−0.0654	−0.1387
Topic13	0.1035	−0.2069	0.1000
Topic14	−0.1497	0.0308	0.0708
Topic15	−0.0405	0.0449	−0.2679
Topic16	−0.1917	−0.1504	0.0505
Topic17	−0.0266	0.0669	0.2044
Topic18	−0.0502	0.0822	−0.0665
Topic19	0.3051	0.0004	0.1657
Topic20	−0.0315	0.2610	−0.1252
Topic21	0.1584	0.1767	0.0507
Topic22	−0.1654	0.1190	−0.1823
Topic23	0.0282	0.2575	−0.1893
Topic24	0.1198	−0.0587	0.2082
Topic25	0.0080	−0.2405	−0.0754

表 2 中显示，与每日收盘价格关联度最高的是关于养老金并轨等城市化问题的第 19 个主题，随着我国城市化率稳步突破 50%，股市反应出对城市化进程带来经济增长的期待；而其次的是关于食品安全与农业经济的第 9 个主题，我国人均粮食消费量已然逼近 400 公斤，而粮

食自给率却跌破 90%，因此粮食安全成为了一大经济焦点。与每日波动率关联度最高的是关于农村劳动力转移的第 20 个主题，随着工业化进程加快，解放农村劳动力的趋势无法逆转，但是人们对于农业剩余劳动力就业能否得到妥善解决态度还是不乐观；而其次的是关于宏观经济指标的第 23 个主题，该主题的第 6 个关键词是 pmi (采购经理指数)，今年一二月份公布的 pmi 跌破 50% 的消息使股市陷入制造业收缩导致经济衰退的恐慌情绪之中，波动率也随之提升。而与日内变动百分比关联度最高的两个主题是都关于国际商贸合作的第 2 个和第 6 个主题，随着经济发展更多转向内需驱动，现如今我国正经历从世界工厂到世界市场的转变，这一转变的实现能够带来一个对中国和世界的双赢局面：对内有助于提高国内居民生活水平，对外能提振世界经济形势，因此相关方面新闻对于维持股市信心很有作用；而与日内变动百分比负相关程度最高的是关于农村劳动力转移的第 15 个主题，再一次说明农业剩余劳动力的就业问题是近期中国经济面临问题的重中之重。

3 预测沪深指数走势

金融预测一直都是学界研究的热点，近年在文献书籍中也涌现大量时间序列预测理论与算法。通常来说，时间序列预测的目标是在给定过去以及当前样本的条件下，对某个未来的价值做出估计，而估计方式一般可以分为两类：线性与非线性的。在过去的几十年间，人们在利用过去及当前数据的线性组合估计未来值方面做出了众多努力。而现实的金融市场受到政策和投资人心理等复杂因素影响，因此擅长处理分类问题的支持向量机算法被引入这一研究领域。

支持向量机 (Support Vector Machines) 是建立在统计学习理论基础上的如今广泛应用的数据挖掘技术，本质上是一个二类分类模型，可以定义为使特征空间上的间隔最大化的线性分类器 [4]。而线性分类的起源可以追溯至 Fisher 在 1936 年提出的如何区分来自两个不同正态总体的样本的问题，而 Fisher 线性判别方法的基本思路是将 m 维向量 $\vec{x}=(x_1,x_2,\cdots,x_m)$ 通过线性函数投影到实数轴上：

$$y(\vec{x})=\vec{w}\cdot\vec{x}=\sum_{j=1}^{m}w_jx_j, \tag{3}$$

其中 $\vec{w}$ 是一个 m 维权重向量，根据投影的结果推断原向量之间的差异大小。于是针对训练集：

$$(\vec{x}_1,y_1),\cdots,(\vec{x}_n,y_n)\qquad \vec{x}\in R^m,\quad y_i\in\{-1,1\},$$

一个线性分类器的学习目标就是要找到一个 m 维空间中的分类超平面：

$$\vec{x}\cdot\vec{x}+b=0. \tag{4}$$

以利用其将训练集分入两个类中，即：

$$y(\vec{x})=\mathrm{sign}(\vec{w}\cdot\vec{x}+b). \tag{5}$$

显然，分类超平面并不唯一，如图 3 所示，可以定义当前超平面下的间隔：

$$\rho(\vec{x},b)=\min_{\{\vec{x}_i|y_i=1\}}\frac{\vec{x}_i\cdot\vec{w}}{|\vec{w}|}-\max_{\{\vec{x}_i|y_i=-1\}}\frac{\vec{x}_i\cdot\vec{w}}{|\vec{w}|}. \tag{6}$$

而支持向量机算法就是通过最大化上述间隔，找到最优分类超平面对应的 $\vec{w}_0$，而后提出的引进核函数使其可以类似得处理非线性问题。

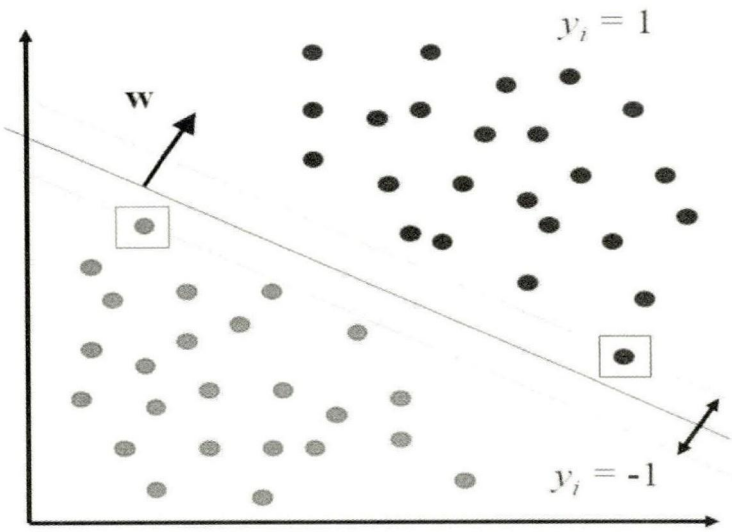

图 3 分类超平面

我们将在上一小节中得到的结果分为两部分：训练集以及预测集. 其中 2014 年 1 月 1 日至 2 月 18 日间共 29 个交易日被划作训练集，而 2014 年 2 月 19 日至 3 月 14 日间共 18 个交易日被划作预测集. 在 $(\vec{x}_1, y_1), \cdots, (\vec{x}_n, y_n)$ $\vec{x}_i \in R^m$, $y_i \in \{-1, 1\}$ 训练集中 m 取 25, n 取 29, $\vec{x}_i = (x_{i,1}, \cdots, x_{i,m})$ 即为之前得到的每一个交易日对应的主题分布向量，而 y_i 根据当天的沪深 300 指数开盘价和收盘价之间涨跌关系记为 $Rise(1)$ 和 $Fall(-1)$.

表 3 对沪深股指涨跌的预测结果

时间	预测值	实际值	是否正确	累计收益	平均每日收益	回撤率
2014-02-19	Rise	Rise	正确	8318.7 元	8318.7 元/天	0%
2014-02-20	Fall	Fall	正确	16314.3 元	8157.2 元/天	0%
2014-02-21	Rise	Fall	错误	13093.5 元	4364.5 元/天	-38.7%
2014-02-24	Rise	Fall	错误	7061.7 元	1765.4 元/天	-37.0%
2014-02-25	Fall	Fall	正确	24643.5 元	4928.7 元/天	0%
2014-02-26	Rise	Rise	正确	29343.9 元	4890.7 元/天	0%
2014-02-27	Fall	Fall	正确	34354.5 元	4907.8 元/天	0%
2014-02-28	Rise	Rise	正确	43212.0 元	5401.5 元/天	0%
2014-03-03	Fall	Rise	错误	42773.1 元	4752.6 元/天	-12.8%
2014-03-04	Rise	Rise	正确	42897.6 元	4289.8 元/天	0%
2014-03-05	Fall	Fall	正确	49910.4 元	4537.3 元/天	0%
2014-03-06	Fall	Rise	错误	45639.3 元	3803.3 元/天	-10.0%
2014-03-07	Rise	Rise	正确	47233.8 元	3633.4 元/天	0%
2014-03-10	Rise	Fall	错误	39129.0 元	2794.9 元/天	-17.8%
2014-03-11	Fall	Rise	错误	34313.1 元	2287.5 元/天	-10.2%
2014-03-12	Rise	Rise	正确	37713.0 元	2357.1 元/天	0%
2014-03-13	Rise	Rise	正确	43948.8 元	2585.2 元/天	0%
2014-03-14	Fall	Fall	正确	46059.0 元	2558.8 元/天	0%

于是我们利用 R 中的支持向量机算法对训练集的数据找到最优分类超平面，其中核函数选用的是应用最广泛的高斯核函数，通过训练集得出的最优分类超平面，对预测集的每日主题占比预测其对应的日内沪深股指走向趋势. 进一步，按照预测结果建立股指期货策略：若预

测当天股指上涨则开盘开多仓收盘平仓；若预测当天股指下跌则开盘开空仓收盘平仓。按照下式计算累计收益，平均每日收益以及回撤率：

$$\begin{aligned}
&\text{每日收益}_k = \text{每日开仓方向}_k \times (\text{每日收盘价}_k - \text{每日开盘价}_k),\\
&\text{累计收益}_i = \sum_{k=1}^{i} \text{每日收益}_k,\\
&\text{平均每日收益}_i = \frac{\text{累计收益}_i}{\text{累计天数}_i},\\
&\text{回撤率}_i = \min\left(0, \frac{\text{每日收益}_i}{\text{累计收益}_{i-1}}\right).
\end{aligned} \tag{7}$$

结果如表 3 所示：

我们提出的方法对日内股指走势预测的正确率达到了 66.7%，这个结果对于信奉大数定律的量化交易领域是一个很有意义的结果。从按照预测结果建立的股指期货策略结果可以看出：较高的预测正确率可以带来相当可观的收益，而预测错误虽然会带来较大的回撤率，但总体而言的平均每日收益还是非常稳定，该方法对于短期时间效果显著。另外，常规的量化分析研究通常建立在财务指标以及基于价量关系的技术指标的基础上，但是不同的行业在会计准则上会有较大的差异，目前也缺乏关于如何使用技术指标准确预测股指趋势的严谨理论。而我们通过对以往难以量化的财经新闻核心主题内容分析推断出未来的股市走势，这有进一步深入的价值。

4 结论分析

随着改革开放带来的经济发展和人们转变的投资意识，股票已经成为中国人投资理财的重要一环，因此对股票市场的分析研究有着重要的理论意义与应用价值，随着时代的进步，核心本质为数学、统计学以及计算机科学在金融市场应用的量化分析研究在金融投资领域中的地位也日渐重要。自从量化交易在欧美市场出现，追寻减少市场系统性风险和获取超额收益率机会的统计套利就一直是量化分析研究的重点。但伴随着计算机技术发展，量化分析研究不再局限于统计套利领域，数据挖掘技术已经被应用于金融市场，并取得了不菲的成绩。

本文提出了一个通过分析财经新闻的主题内容识别市场中主要事件的股市预测系统。此系统通过对财经新闻文档做语义分析得到其主题构成，分析得到对整个股市有影响力的主要事件。进一步的，结合支持向量机理论，本文提出了一种基于新闻主题分布分析股市的未来走势的预测模型。支持向量机作为基于统计学习理论的一种模式识别方法，已经在生物信息学和文本手写识别等大量领域成功应用，本文说明它在金融市场预测领域也能发挥关键性作用。通过支持向量机对获取到的每日新闻主题进行分析，能以较高的概率 (66.7%) 预测当日股市的涨跌，这对量化交易将产生巨大的价值。

[参考文献]

[1] Hearst M A. Text data mining: Issues, techniques, and the relationship to information access [R]. Presentation notes for UW/MS workshop on data mining, 1997.

[2] Goldberg D E, Holland J H. Genetic algorithms and machine learning [J]. Machine Learning, 1988, 3:95-99.

[3] Landauer T K, McNamara D S, Dennis S, et al. Handbook of latent semantic analysis [B]. Lawrence Erlbaum, 2007.

[4] Cortes C, Vapnik V. Support-vector networks [J]. Machine Learning, 1995, 20: 273–297.

[5] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet allocation [J]. The Journal of Machine Learning Research, 2003, 3: 993–1022.

[6] Wuthrich B, Permunetilleke D, Leung S, et al. Daily prediction of major stock indices from textual www data [J]. HKIE Transactions, 1998, 5: 151–156.

[7] Lavrenko V, Schmill M, Lawrie D, et al. Mining of concurrent text and time series [C]. In KDD-2000 Workshop on Text Mining, 2000, 2000: 37–44.

[8] Kloptchenko A, Eklund T, Karlsson J, et al. Combining data and text mining techniques for analysing financial reports [J]. Intelligent systems in accounting, finance and management, 2004, 12:29–41.

[9] Mittermayer M A. Forecasting intraday stock price trends with text mining techniques [C]. Proceedings of the 37th Annual Hawaii International Conference on System Sciences, 2004.

[10] Seo Y W, Giampapa J A, Sycara K. Financial news analysis for intelligent portfolio management [R]. Robotics Institute, 2004..

[11] Ingvaldsen J E, Gulla J A, Laegreid T, et al. Financial news mining: Monitoring continuous streams of text [C]. Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence, 2006: 321–324.

[12] Trafalis T B, Ince H. Support vector machine for regression and applications to financial forecasting [C]. Proceedings of IEEE-INNS-ENNS International Joint Conference on Neural Networks, 2000, 6: 6348–6348

[13] Tay F E, Cao L. Application of support vector machines in financial time series forecasting [J]. Omega, 2001, 29: 309–317.

[14] Yang H, Chan L, King I. Support vector machine regression for volatile stock market prediction [C]. In Intelligent Data Engineering and Automated Learning, 2002: 391–396.

[15] Huang Z, Chen H, Hsu C J, et al. Credit rating analysis with support vector machines and neural networks: a market comparative study [J]. Decision Support Systems, 2004, 37: 543–558.

[16] Cao D Z, Pang S L, Bai Y H. Forecasting exchange rate using support vector machines [C]. Proceedings of 2005 International Conference on Machine Learning and Cybernetics, 2005, 6:3448–3452.

[17] Bao Y K, Liu Z T, Guo L, et al. Forecasting stock composite index by fuzzy support vector machines regression [C]. Proceedings of 2005 International Conference on Machine Learning and Cybernetics, 2005, 6: 3535–3540.

[18] 赵丽丽, 赵茜倩, 杨娟等. 财经新闻对中国股市影响的定量分析 [J]. 山东大学学报, 2012, 47: 70–75.