DOI:10.13502/j.cnki.issn1000-7636.2007.06.006

基于数据挖掘技术的股票收益预测研究

蒋艳霞　柯大钢　解青芳

[内容提要] 近年来，数据挖掘技术在各领域的应用越来越广。数据挖掘工具能够从大量数据中获取有用的知识，并对将来的趋势和行为做出预测，为使用者提供决策依据。本文主要分析了数据挖掘技术在股票收益预测方面的应用前景并阐述了相关的应用程序。

[关 键 词] 数据挖掘　应用　股票收益

中图分类号：F 832　　文献标识码：A　　文章编号：1000-7636（2007）06-0036-05

一、数据挖掘技术的引入

股票收益预测是财务研究的一个重要方面，其蕴含的假设是过去的公开信息对未来的股票收益具有一定程度的预测能力。有效市场假设认为未来股票收益是不可预测的，因为当前股价已经反映了所有可以影响股票价值的信息（Jensen，1978），是自身价值的最佳估计。但大量实证研究表明，市场往往并不是有效的（发展中国家市场的有效性程度更低），市场中存在着可以预测的短期趋势。Balvers，Cosimano 和 McDonald（1990），Breen，Glosten 和 Jagannathan（1989），Fama 和 Schwert（1977），Fama 和 French（1988，1989），Ferson（1989），Keim 和 Stambaugh（1986），Schwert（1990）等都发现股票收益是可以通过公开信息来预测的。可以用来预测股票收益的公开信息，主要包括宏观经济变量以及特定公司信息（比如资产负债表，股利等）。财务报表（资产负债表，损益表，现金流量表）信息是投资者进行投资决策的主要依据。投资者从财务报表中提取出与公司价值相关的信息，预测出公司的未来收益变化，从而指导自己的投资决策，买进价值被低估的股票，卖出价值被高估的股票，以获得较高收益。

随着信息技术的飞速发展，越来越多的公司开始以电子形式来提供会计信息。在大型商业数据库和互联网上存在着大量与公司财务业绩有关的信息，这些信息对公司的各利益相关者（股东，债权人，审计人员，管理当局等）是非常有用的，他们可以随时从中提取相关信息以帮助自己决策。但面对如此海量的数据，如何准确有效地发现其中的潜在规律是大多数投资者面临的任务，数据挖掘技术为此提供了便利。从数据库中发现知识（knowledge discovery in database，KDD）是20世纪80年代末开始的。KDD一词是在1989年8月在美国底特律

收稿日期：2007-04-25

作者简介：蒋艳霞　西安交通大学管理学院博士研究生，西安市，710049；
柯大钢　西安交通大学管理学院教授，博士生导师；
解青芳　山东建筑大学商学院讲师。

召开的第一届 KDD 国际学术会议上正式形成的。1995 年在加拿大召开了第一届知识发现和数据挖掘（data mining，DM）国际学术会议，由于把数据库中的“数据”形象地比喻成矿床，“数据挖掘”一词很快流传开来。简单地说，数据挖掘就是一个处理过程，它利用一种或多种计算机学习技术，从数据库的数据中自动分析并提取知识，目的是确定数据的趋势和模式（pattern）。数据挖掘应用的领域很广，包括卫生保健，商业和金融，欺诈侦测等。近来数据挖掘技术也开始作为财务分析工具，用来预测股票收益。

二、数据挖掘技术研究

1. 国外研究。财务报表一直是股票价值估计的主要信息来源。由于市场的非完全有效性，股票价格往往会围绕股票价值上下浮动，但又不会偏离太远，最后还会逐渐回归到自身价值（Rendleman，Jones 和 Latane，1982；Foster，Olsen 和 Shevlin，1984）。基本面分析就是运用财务报表中的会计信息来估计股票价值的过程，以找出当前价格偏离价值的股票，从而采取相应的投资战略，获得超额收益。最早分析财务报表基础信息的学者是 Graham，Dodd 和 Cottle（1962），他们合作出版的《证券分析》一书经过多次修订，一直被称为证券投资领域的圣经。Graham 是证券投资中价值学派理论的鼻祖，他认为，任何一家公司公布的资料都是经过修饰、掩盖的，如何分析、整理它们，使之露出本来的面目，需要借助一些方法，只有通过真实的信息，投资者才能决定取舍（20 世纪 90 年代中期，美国证券界掀起了一股热潮，其主题就是对 Graham 机制投资观念进行重新挖掘。之所以会这样，主要原因是以 Buffett 为代表的价值投资学派与主流的效率市场理论相互竞争，而 Buffett 则取得了显著的成功）。

研究财务报表信息与股票证券价格关系的会计文献起始于上世纪 60 年代。Ball 和 Brown1968 年 发表在《Journal of Accounting Research》上的文章是当时最有影响的研究。他们认为，财务报表信息和股票的市场价值是相关的。他们证明，股票价格的变化可以由未预期的年净收益变化的符号来预测。未预期收益定义为根据市场模型预计出的收益与实际实现收益之间的差额。如果未预期收益用 $\hat{u}_{jt}$ 表示，则其定义可用符号表示为：

$$\hat{u}_{jt} = VI_{jt} - V\hat{I}_{jt}$$

其中 VI_{jt} 表示 j 公司第 t 期净收益的实际变化，$V\hat{I}_{jt}$ 是预期净收益的变化。预测模型中净收益的变化是通过如下方法得到的：首先，用市场平均收益的变化 $VM_{j,t-\tau}$ 对单个股票的收益变化 $VI_{j,t-\tau}$ 做回归：

$$VI_{j,t-\tau} = \hat{a}_{1jt} + \hat{a}_{1jt} VMmj,\ t-\tau + \hat{u}_{j,t-\tau}$$

然后用得到的系数来预测单个公司净收益的变化：

$$VI_{jt} = \hat{a}_{1jt} + \hat{a}_{1jt} VM_{jt}$$

Ball 和 Brown 用 1946～1966 共 21 年的公司收入数据和月股票价格进行分析，并把样本限制在至少有 100 个月股价资料和在每年 12 月 31 日出具财务报表的公司。研究表明，未预期的会计收益能够预测股票价格变化，会计收益是价值相关的。Ball 和 Brown（1968）的推论是如果会计收益能够预计，则股票收益也可以预计。

McKibben（1972）仿效 Ball 和 Brown（1968）的研究，用工业股票的数据来预测股票收益，自变量为每股收益、股利和前期股价资料。根据他的模型，表现最好的股票（占总数的 10%）8 年平均收益可以达到 29.5%，而平均收益水平仅为 16.5%。Basu（1983）的方法稍微不同，他把报表数据和市场技术资料结合起来进行分析，并根据价格收益比率（PE）的值把样本分成五部分，计算 12 个月的股票收益，结果发现，PE 比率和风险调整收益之间存在着显著关系（低的 PE 会导致后期产生高股票收益）。DeBondt 和 Thaler（1987）也发现，当前业绩表现很好的股票，前期的表现往往很差。

Ou 和 Penman（1989）的研究是会计研究中采用无假设归纳方法的一个很好的例子，他们对会计变量的显

著性没有做任何假设，而是采用逐步回归程序来选择模型将要包含的变量。虽然在研究的开始限制了将要分析的变量的数目，他们仍考察了 68 个财务报表变量对财务收益和股价的影响。他们运用 LOGIT 模型来预计未来收益增加的概率，自变量是 t 年的财务报表变量，因变量是调整后的 t + 1 年的每股收益变化。根据预测出的收益增加的概率，来决定股票是买进还是卖出。结果表明，财务报表中的价值相关变量可以稳定地用来预测未来股票收益。Ou 和 Penman 的套利组合样本内的收益为 14.53%，表现最好的组合收益为 21.91%。

近年来，随着数据挖掘工具的发展，模型设定技术也取得了巨大飞跃。神经网络技术是用的比较多的一种，常用来研究财务报表变量和股价之间的非线性关系以及财务危机预警。Longo（1995）发现用神经网络模型来选择股票比较有效，优质股票投资组合的年收益可达到 31.2%，而市场收益仅为 18.36%。Hill，O' Connor 和 Remus（1996）比较了神经网络和传统统计预测方法，发现神经网络方法的预测效果明显优于传统方法。他们认为神经网络特别适用于不连续的时间序列。Serrano-Cinca（1998），Back 等（2001），以及 Kloptchenko 等（2004），都采用了自组织图对财务比率进行分析，以预测公司未来的财务表现。在输入变量的选择上，Enke 和 Thawornwong（2005）提出用数据挖掘中的信息增益技术，计算每个变量的信息增益值，来决定是否将其引入预测模型。通过比较几种神经网络模型，他们发现分类模型可以产生较高的风险调整收益。关于神经网络的应用研究还有 Lam（2004）、Thawornwong 和 Enke（2004）等。利用神经网络来研究财务预警的文章也比较多，如 Kiviluoto（1998）、McKee 和 Greenstein（2000），Tung、Quek 和 Cheng（2004）等。

遗传算法也是数据挖掘技术的一种，1975 年由美国 Michigan 大学的 Holland 提出，近年来也开始逐渐应用在财务预测领域（Grupe 和 Jooste，2004）。相对于神经网络，遗传算法要求的训练集比较小。由于一般的时间序列都相对较短，所以遗传算法是时间序列模型的理想工具。另外，遗传算法适合用于变化的环境条件，当数据集存在结构变化时，它能直接从以前的优化结果中形成新的结论，而不用重新对数据进行训练。Mahfoud 和 Mani（1996）以 1600 多支股票为试验对象，比较神经网络和遗传算法在财务预测中的表现，发现遗传算法要优于神经网络，但二者相结合的方法又优于遗传算法。Dvhar、Chou 和 Provost（2000）；Telbany（2004）；Oh，Kim 和 Min（2005）分别用遗传算法模型预测了 S&P500，埃及和韩国股市的股票收益，模拟结果都比较好。

2. 国内研究。国内学者的相关研究比较零散，不成系统，而且样本量普遍偏少，采用的方法也较单一。赵宇龙（1998）采用 Ball 和 Brown 的方法对取自上海证券交易所的样本研究后发现，1996 年度的盈余披露具有比较明显的信息含量和市场效应。黄志忠（2002）以 1994 年 12 月 31 日之前在上海证券交易所上市的公司为样本，分析了利润率和未预期利润对超额股票收益的影响。分析结果表明，在中国股票市场上，不仅未预期盈余具有信息含量，利润率本身也具有信息含量。而且我国证券市场具有初级市场所具有的不成熟的特征，即价值观念的易变性。陆静、孟卫东和廖刚（2002）采用皮尔逊相关系数和扩展的 Ohlson-Feltham 股票计价模型比较了会计盈余和现金流量对股票价格的影响程度，结果表明每股收益比现金流量能较好地解释股票价格。另外，他们发现我国资本市场存在功能锁定的现象，投资者对上市公司价值的评估往往局限于每股收益，忽视了现金流量等其他财务指标。王喜刚、丛海涛和欧阳令男（2003）也发现会计指标在解释上市公司市场价值方面具有很高的信息价值。其他研究还有赵春光（2004），兰永、李仕明和李金（2005）等。

受国外研究的影响，国内学者用数据挖掘技术研究财务预警的相对比较多。吴德胜、梁樑、殷尹（2004），刘洪、何光军（2004），杨淑娥、黄礼（2005），张玲、陈收和张昕（2005）比较了不同模型在财务预警中的应用，发现神经网络效果比较好。研究同时表明在现有会计制度和会计准则下，财务报表能提供预测财务困境的大量有用信息。吴超鹏、吴世农（2005）用人工神经网络分析了影响公司财务状态变化的因素。姚宏善、沈轶（2005）利用遗传算法的全局寻优能力，构造了一个预测财务困境的遗传神经网络模型，实证结果表明，该模型比神经网络模型具有更好的预测财务困境的能力。在股价行为方面，甘霖敏、杨炘（2004）采用 BP 神经

网络方法研究了公司规模、交易量、β因子和年收益价格比等因素与股票收益率之间的关系。陈兴、孟卫东、严太华（2001），武振、郑丕谔（2004）则从技术分析的角度用神经网络与遗传算法相结合的方法对股价进行预测。杨成、程晓玲、殷旅江（2005）以钢铁行业为研究对象，应用BP神经网络模型分析了公司各个层面的财务指标对股价的预测能力。

由以上讨论可知，现有的关于股票收益预测的研究只是期望证明财务报表信息是否确实对股票收益产生了影响，而具体的影响方式、程度、途经是什么一直没有得到深入研究。从研究的对象上看，现有研究的取样范围和区间都很不完整，使研究结论缺乏一般性而难以推广。从研究方法上看，以上研究大都采用演绎法，这就容易导致不能充分利用财务报表信息，少数研究即使利用了属于归纳法的数据挖掘技术，但也都是采用比较单一的神经网络方法。因此，如何应用先进的数据挖掘方法来深入系统地研究财务报表信息与股票收益的关系，是学者们今后应该努力的方向。

三、数据挖掘方法在股票收益预测中的应用程序

我们以财务报表基本面分析为例，来说明数据挖掘在股票收益预测中的应用。这里，我们的目标是研究财务报表信息与股票收益之间的关系。目标确定后，我们应用数据挖掘方法进行分析，其步骤如图1所示：

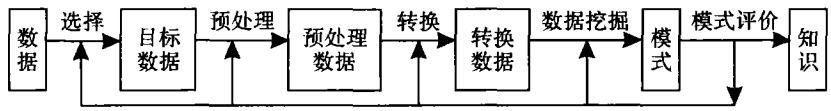

图1 数据挖掘方法

1. 数据收集。研究所需的数据经常存在于企业的数据仓库或各种大型商业数据库中。在此阶段，我们的任务就是从数据库中获取基本分析所需的数据：财务报表变量和股票收益信息。首先根据财务报表变量预测出每股收益的变化，每股收益预期变化又会影响到股票价格。国泰安数据库CSMAR中的中国股票市场交易数据库，中国上市公司财务数据库以及中国上市公司财务指标分析数据库可以为我们提供关于上市公司股票收益和财务变量的数据。关于财务指标数量的选择，一般而言，指标越多，归纳研究越容易发现数据中存在的潜在规律。但如果选择的指标过多，符合条件的样本就会减少，从而影响预测效果。因此，需要采用一个合理的标准来决定财务变量的数目。

2. 数据预处理。数据预处理一般包括消除噪声、推导计算缺值数据、消除重复记录等。数据预处理可以通过SAS软件的频率分析来实现。如果同一个变量的缺失值很多，为了保存样本数，可以考虑丢掉这个变量。对异常数值的处理有两种方法：一是直接去掉这些异常数值，二是寻找不受异常数值点影响的健壮性建模方法。由于我国股票市场形成时间较短，而且经济政治环境变化比较剧烈，股市数据存在不少异常点，应慎重处理。

3. 数据转换。数据转换的主要目的是削减数据维数或降维，即从初始特征中找出真正有用的特征以减少数据挖掘时要考虑的特征或变量个数。维归约（dimension reduction）技术主要有零维特征法（逐步向前选择法）和全维特征法（逐步向后删除法）。另外还可以应用数据编码或变换，得到原数据的归约或压缩表示，普遍使用的方法有小波变换法和主成分分析法。在本研究中，很多原始财务报表数据都要转换成比率和百分比。因变量［调整后的每股收益变化 =（$t+1$）期每股收益 - t期每股收益 -（$t+1$）期之前四年每股收益变化的平均值］则根据其值的符号，取0或1。

4. 数据挖掘。数据挖掘阶段首先要确定挖掘的任务或目的，如数据分类、聚类、关联规则发现或序列模式发现等。确定了挖掘任务后，就要决定使用什么样的挖掘算法。算法的选择有两个考虑因素：一是不同的数据有不同的特点，因此需要用与之相关的算法来挖掘；二是要根据用户或实际运行系统的要求，有的用户可能希望获取描述型的、容易理解的知识（采用规则表示的挖掘方法显然好于神经网络之类的方法），而有的用户希

望获取预测精度尽可能高的预测型知识。选择了挖掘算法后，就可以实施数据挖掘操作，获取有用的模式。我们的目的是获得每股收益的预测模型，可以选择神经网络算法、遗传算法、或者二者相结合的模型。

5. 结果的解释和评估。数据挖掘阶段发现出来的模式，经过评估可能存在冗余或无关的模式，这就需要将其剔除；也有可能模式不满足用户要求，则需要退回到发现过程的前面阶段，如重新选取数据，采用新的数据变换方法，设定新的参数值，甚至换一种挖掘算法等等。另外，还可能需要对发现的模式进行可视化，或者把结果转换为用户易懂的另一种模式表示，如把分类决策数转换为“IF…THEN…”规则。在我们的研究中，此阶段的任务就是把数据挖掘阶段预测的股票收益与前人研究相比较。如果根据数据挖掘方法获得的股票收益比采用其他方法得到的高，则说明数据挖掘方法是一种有效的研究方法。

综上所述可以发现，数据挖掘方法是非常适合做股票收益预测的。鉴于国内学者的相关研究比较零散，不成系统，而且样本量普遍偏少，采用的方法也较单一，因此，未来的研究应在系统性和研究深度上有所突破。

参考文献：

[1] Back, B., Torvonen, J., Vanharanta, H., and Visa, A. Comparing numerical data and text information from annual reports using self – organizing maps [J]. International journal of accounting information systems, 2001(2).

[2] Ball, R., and Brown, P. An empirical evaluation of accounting income numbers [J]. Journal of accounting research, 1968(6).

[3] Basu, S. The relationship between earnings' yield, market value and return for NYSE common stocks: further evidence [J]. Journal of financial economics, 1983(12).

[4] DeBondt, W., and Thaler, R. Further evidence on investor overreaction and stock market seasonality [J]. Journal of finance, 1987(42).

[5] Enke, D., and Thawornwong, S. The use of data mining and neural networks for forecasting stock market returns [J]. Expert systems with applications, 2005(29).

[6] Graham, G., Dodd, D. L., and Cottle, S. Security analysis: principles and techniques [M], 4thed. New York: McGraw – Hill, 1962.

[7] Grupe, F. H., and Jooste, S. Genetic algorithms – a business perspective [J]. Information management & computer security, 2004(12).

[8] Hill, T., OConnor M. and Remus, W. Neural network models for time series forecasts [J]. Management science, 1996(42).

[9] Kloptchenko, A., Eklund, T., Karlsson, J., Back, B., Vanharanta, H., and Visa, A. Combining data and text mining techniques for analyzing financial reports [J]. Intelligent systems in accounting, finance and management, 2004(12).

[10] Lam, M. Neural network techniques for financial performance prediction: integrating fundamental and technical analysis [J]. Decision support systems, 2004(37).

[11] Longo, J. M. Selecting superior securities: using discriminant analysis and neural networks to differentiate between winner' and loser' stocks [D]. Ph. D. dissertation, Rutgers University, 1995.

[12] Mahfoud, S., and Mani, G. Financial forecasting using genetic algorithms [J]. Applied artificial intelligence, 1996(10).

[13] McKibben, W. Econometric forecasting of common stock investment returns: a new methodology using fundamental operating data [J]. Journal of finance, 1972(27).

[14] 甘霖敏,杨炘. 用人工神经网络方法对股票收益率影响因素的实证分析[J]. 清华大学学报(哲学社会科学版),2004(2).

[15] 黄志忠. 论上市公司盈余的信息含量[J]. 经济评论,2002(3).

[16] 刘洪,何光军. 基于人工神经网络方法的上市公司经营失败预警研究[J]. 会计研究, 2004(2).

[17] 陆静,孟卫东,廖刚. 上市公司会计盈利、现金流量与股票价格的实证研究[J]. 经济科学,2002(5).

[18] 吴超鹏,吴世农. 基于价值创造和公司治理的财务状态分析与预测模型研究[J]. 经济研究,2005(11).

[19] 吴德胜,梁樑,殷尹. 不同模型在财务预警实证中的比较研究[J]. 管理工程学报, 2004(2).

[20] 杨淑娥,黄礼. 基于 BP 神经网络的上市公司财务预警模型[J]. 系统工程理论与实践,2005(1).

责任编辑：范子奇