Advances in Applied Mathematics 应用数学进展, 2021, 10(3), 654-665
Published Online March 2021 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2021.103071

Hans汉斯

基于多因子模型和CART分类回归树的证券市场的预测及应用

王喜月，刘海军，马璐瑶，樊　宇，苏　晴

郑州大学数学与统计学院，河南　郑州
Email: 1933108512@qq.com

收稿日期：2021年2月8日；录用日期：2021年3月4日；发布日期：2021年3月11日

摘　要

本文基于多因子模型和CART分类回归树算法，利用因子分析法，研究了证券市场的预测问题。首先依据Fama和French因子分析方法选取了8大类公司财务指标建立了多因子分析模型，其次利用统计学中的因子分析法对所选因子进行了特征提取，得到贡献最大的五个特征，并分析了所得市场特征的市场意义，然后根据提取的市场特征及其市场技术指标建立了预测市场季收益率的CART分类回归树模型，并随机选取了了上交所20家公司进行了实证分析，结果表明预测模型具有较好的预测精度，其次，依据预测结果构造了顺势投资策略，获得了优于市场平均收益的良好回报。

关键词

多因子模型，CART分类回归树，因子分析，预测，投资策略

Forecasting and Application for Securities Market Based on Multi-Factor Model and CART Classification Regression Tree Algorithm

Xiyue Wang, Haijun Liu, Luyao Ma, Yu Fan, Qing Su

School of Mathematics and Statistics, Zhengzhou University, Zhengzhou Henan
Email: 1933108512@qq.com

Received: Feb. 8th, 2021; accepted: Mar. 4th, 2021; published: Mar. 11th, 2021

文章引用: 王喜月, 刘海军, 马璐瑶, 樊宇, 苏晴. 基于多因子模型和 CART 分类回归树的证券市场的预测及应用[J]. 应用数学进展, 2021, 10(3): 654-665. DOI: 10.12677/aam.2021.103071

Abstract

Based on Famma and French multi-factor model and CART classification regression tree algorithm, this paper forecasts the securities market. First, according to the Fama and French fmulti-factor analysis method, eight major types of company financial indicators were selected to make a multi-factor analysis. Second, the factor analysis method in statistics was adopted to extract the features of the selected factors, and five features that contribute the most weight were obtained. Then these five features and market risk indicator were taken to train a CART classification regression tree model for predicting market quarterly returns. Meanwhile 20 companies from the Shanghai Stock Exchange were randomly chosen for empirical analysis. The results show that the forecasting model has good forecasting accuracy. Especially, a simulation investment has made with history data and obtained a better performance than the average market.

Keywords

Multi-Factor Model, CART Classification Regression Tree, Factor Analysis, Forecast, Investment Strategy

1. 引言

证券市场在现代社会中发挥着重要作用，并且影响着国家的经济发展[1]。一直以来，对证券市场收益预测的研究是金融统计研究领域的热点问题，寻找影响证券投资收益的因素并分析这些因素的作用对提高证券投资效率与防范市场风险具有十分重要的理论与实际意义。特别是，近十年来，相关研究得到金融工程和统计领域的广泛关注[2]。

分析和预测证券市场行为的两种常用方法是基础分析和技术分析。这两种方法的主要目标都是依据已有的证券市场信息预测其未来走向及数值。前一种分析是依据影响市场走势的经济因素，如公司经营数据等，它较适合于长期的预测。而技术分析主要依据证券市场的交易数据去分析预测市场未来的变化[3]，一般依据多个技术指标，对某只股票或市场指数的历史数据进行整理和分析，从而预测该金融资产的未来价值。

近来，人工智能(AI)在各个领域获得了极大成功[4]，其被认为是解决许多金融问题的有力工具，如利用人工神经网络[5] [6] [7]、决策树(DT) [8] [9] [10] [11]、深层神经网络[12]、梯度树(GBDT)、随机森林(RF) [13]和 CART 树[14]等，对股票进行分类或预测，取得了不错的预测效果。尽管中国股市已经发展了几十年，但与股票回报相关的特征研究仍显不足。因此，借助于机器学习来说明证券收益及其潜在影响因素之间的关系是很有价值的。为了对股票进行预测，本文通过使用 CART 分类回归树和多因子模型，通过基本分析面和技术分析相结合研究证券市场的预测及其应用。首先依据 Fama 和 French 因子分析方法选取了 8 大类指标给出了因子得分，其次利用因子分析法提取了五个重要因子，并分析其市场意义，最后根据提取的市场特征及其市场技术指标建立了预测市场季收益率的 CART 分类回归树模型，并进行了仿真模拟投资。

论文组织如下：第二部分是数据及其处理方法，第三部分是数据建模，第四部分是预测结果及其分析，第五部分给出了模拟投资结果，最后是结论。

2. 数据预处理

2.1. 数据来源

本文选用的数据来自锐思数据库，选择上海交易所从2002年12月31号到2019年12月31号的200家公司季度财务数据和交易数据。文章选取了32个变量，其中31个候选因子，选取股票的市场投资的季收益率作为目标变量，即预测变量。所选的候选因子包括8类指标，有每股指标，盈利能力，偿债能力，成长能力指标，营运能力指标，现金流量指标，资本结构指标和价值指标。具体候选因子见表1：

Table 1. Candidate factors
表 1. 候选因子

类型	因子名称
每股指标	每股净资产、每股营业收入、每股收益、每股营业利润
盈利能力指标	资产报酬率、资产净利率、净利润/营业收入、营业利润率、营业利润/营业总收入、营业总成本/营业总收入
偿债能力指标	流动比率、速动比率、产权比率、现金流动负债比
成长能力指标	每股经营活动现金流量增长率、净资产收益率、营业收入增长率、营业利润增长率、净利润增长率
营运能力指标	存货周转率、应收账款周转率、非流动资产周转率、总资产周转率
现金流量指标	总资产现金回收率
资本结构指标	资产负债率、流动资产/总资产、权益乘数
价值指标	成交量、市盈率、市净率、市销率

2.2. 数据处理

在获取数据中存在空缺值，数据量纲不统一等问题，为此对其进行预处理，以便模型的预测。数据预处理的方法包括缺失值处理和归一化处理。缺失值处理是用前后两个值的平均值进行填充。归一化处理的公式如下：

$$\overline{x_{ij}} = \frac{x_{ij} - x_{i\min}}{x_{i\max} - x_{i\min}}$$

其中 x_{ij} 是第 i 行第 j 列的真实值，$x_{i\min}, x_{i\max}$ 分别是这一列的最小值和最大值。

2.3. 因子得分

由Fama和French的多因子模型，因子得分是根据因子收益率沿着一个或者多个因子的维度，将股票分成不同的部分，根据这些维度计算出整个市场不同部分的收益率，其差值即为因子得分。

2.3.1. 五因子得分

Fama和French五因子的因子得分计算过程如下(以市值因子 SMB_t 为例)：

根据Fama-French [2]中所使用的2 × 3方法构建因子，沿着两个因子维度，将股票分为2 × 3 = 6个组合。其中一个因子固定为规模，按照中位数分为两层，另外一个为BP (账面市值比)、OP (营运利润率)

或者 INV (投资风格)，按照 30%、70%分位点分为 3 层。具体构建步骤如下：

1) 按股票市值的中位数把全体股票分成小市值(S)和大市值(B)两组。

2) 按账面市值比的 30%和 70%分位点把样本分成高(H)、中(N)、低(L)三组，将两个指标交叉，可把全体分成 SH，SN，SL，BH，BN，BL 共 6 个组合。

3) 用同样的方法，以营运利润率和投资风格代替账面市值比，用稳健(R)、集中(N)、较弱(W)来划分盈利能力、用保守(C)、居中(N)、激进(A)来划分投资风格，可把全体分为 12 个组合。SR，SN，SW，BR，BN，BW 和 SC，SN，SA，BC，BN，BA。

4) 计算上述各组合每一期的市值加权平均季收益率，接着计算因子得分，具体公式如表 2。

估值因子 HML_t、盈利因子 RMW_t 和投资因子 CMA_t 构建方法类似市值因子，公式如表 2。

Table 2. Five-factor calculation method
表 2. 五因子计算方法

因子名称	计算方法
市值因子 SMB_t	$SMB_{BM}=\frac{SH+SN+SL}{3}-\frac{BH+BN+BL}{3}$ $SMB_{OP}=\frac{SR+SN+SW}{3}-\frac{BR+BN+BW}{3}$ $SMB_{INV}=\frac{SC+SN+SA}{3}-\frac{BC+BN+BA}{3}$ $SMB_t=\frac{SMB_{BM}+SMB_{OP}+SMB_{INV}}{3}$
估值因子 HML_t	$HML_t=\frac{BH+SH}{2}-\frac{BL+SL}{2}$
盈利因子 RMW_t	$RMW_t=\frac{BR+SR}{2}-\frac{BW+SW}{2}$
投资因子 CMA_t	$CMA_t=\frac{BC+SC}{2}-\frac{BA+SA}{2}$

2.3.2. 本文选取的 31 个因子得分

与 Fama-French 五因子得分计算原理相似，不同之处是五因子是从两个或两个以上维度来计算因子得分，而这 31 个因子的得分是从一个维度来计算的，下面以净资产收益率(ROE)为例说明因子得分的具体计算步骤：

1) 按净资产收益率的 30%和 70%分位点把样本分成高(H)、中(N)、低(L)三组，可把全体公司分成 H, N, L 共 3 个组合。

2) 计算上述各组合每一期的市值加权平均季收益率，高组的市值加权平均季收益率减去低组的市值加权平均季收益率，具体公式如下：

$$ROE_t=H-L$$

同样的步骤和方法计算每一季度的其他 30 个因子的得分。计算公式如下：

$$ELSE_t=H-L$$

2.4. 筛选因子

为减轻模型的复杂度，根据 Fama-French 多因子分析方法[15] [16]，按照规模－估值，规模－盈利，

规模－投资三种维度将所有股票分成 3 个 5 × 5 宫格的资产组合，每一个宫格内股票的季收益率都会形成一个时间序列，把它作为回归的因变量，由 2.3.2 计算出来的 31 个因子得分也是时间序列，以它们分别作为回归的自变量，做一元回归，利用回归的 t，p 值筛选因子。

1) 按照规模－估值筛选

我们构造了规模－估值的 5 × 5 宫格的资产组合，其时间序列作为因变量，构造的 31 个因子的得分时间序列作为自变量，对 25 个宫格的时间序列分别对每一个因子进行单因子回归，会得到 p 值和 t 值，如果 25 个宫格的时间序列和某个因子回归得到的 p 值大于 0.05 的宫格个数大于 90%，则保留该因子，否则剔除该因子，按照规模－估值分组回归得到的 t 值和 p 值如表 3：

Table 3. The t and p values of factor regression
表 3. 因子回归的 t，p 值

第 1 个因子的 t 值(左)和 p 值(右)											
	Small	2	3	4	Big		Small	2	3	4	Big
Low	−8.371	−10.378	−8.544	−7.888	−10.044	Low	0.000	0.000	0.000	0.000	0.000
2	−6.528	−5.725	−7.177	−8.566	−8.303	2	0.000	0.000	0.000	0.000	0.000
3	−6.063	−8.168	−7.050	−5.338	−7.292	3	0.000	0.000	0.000	0.000	0.000
4	−5.377	−4.533	−6.799	−5.440	−6.994	4	0.000	0.000	0.000	0.000	0.000
High	−0.997	−2.407	−3.231	−1.953	−3.773	High	0.322	0.018	0.002	0.055	0.000
第 2 个因子的 t 值(左)和 p 值(右)											
	Small	2	3	4	Big		Small	2	3	4	Big
Low	−3.896	−4.431	−4.424	−4.799	−4.241	Low	0.000	0.000	0.000	0.000	0.000
2	−4.735	−4.989	−3.991	−5.584	−4.645	2	0.000	0.000	0.000	0.000	0.000
3	−3.071	−4.104	−3.468	−3.195	−4.622	3	0.003	0.000	0.000	0.002	0.000
4	−2.408	−2.279	−4.298	−3.001	−4.082	4	0.018	0.025	0.000	0.003	0.000
High	−0.810	−1.165	−1.632	−0.975	−1.771	High	0.420	0.247	0.107	0.333	0.080
		……						……			
第 31 个因子的 t 值(左)和 p 值(右)											
	Small	2	3	4	Big		Small	2	3	4	Big
Low	0.333	0.816	0.778	1.501	−0.604	Low	0.739	0.417	0.438	0.138	0.547
2	0.664	0.917	1.122	0.194	−0.469	2	0.508	0.362	0.265	0.846	0.640
3	1.276	−0.092	3.907	2.252	0.375	3	0.206	0.926	0.000	0.027	0.708
4	2.437	2.251	2.141	1.702	1.672	4	0.017	0.027	0.036	0.093	0.099
High	3.227	2.968	4.753	4.374	4.186	High	0.001	0.001	0.000	0.000	0.000

由于因子比较多，表 2 仅列举了三个因子，根据因子筛选规则，第一个因子给与保留，第二个因子舍去，第三十一个因子也舍去。按照规模－估值筛选最终保留了每股净资产、资产报酬率、资产净利率等 12 个因子。

2) 按照规模－盈利筛选

同 1) 按照规模－估值的筛选方法一样，按照规模－盈利筛选最终保留了营业利润/营业总收入、每股净资产、资产报酬率等 13 个因子。

3) 按照规模－投资筛选

同 1) 按照规模－估值的筛选方法一样，按照规模－投资筛选最终保留了每股营业收入、营业利润/营业总收入、每股净资产、资产报酬率等 14 个因子。

综上，选取 1) 2) 3) 中筛选之后的共同因子作为预测因变量。见表 4：

Table 4. Selected factors
表 4. 选取的因子

类型	因子名称
每股指标	每股净资产
盈利能力指标	资产报酬率、资产净利率、净利润/营业收入、营业利润/营业总收入、营业总成本/营业总收入、营业利润率、
偿债能力指标	产权比率
成长能力指标	营业收入增长率
营运能力指标	总资产周转率
价值指标	成交量、市盈率

2.5. 因子分析

为了进一步降低模型复杂度，这里选择因子分析继续降维.因子分析是通过减少因子之间的相关性，将高维变量变成低维变量，因子分析的具体步骤如表 5：

Table 5. Factor analysis steps
表 5. 因子分析步骤

因子分析步骤
1) 输入原变量 x 的观测数据阵并对其标准化，记之为 X；
2) 求 x 的子样相关阵 $R=\frac{1}{N}X'X$ ；
3) 求 R 的特征值 $\lambda_1 \geq \cdots \geq \lambda_m$ 及相应的单位特征向量 $u_1,\cdots,u_m$；
4) 确定公共因子的个数 k 值；
5) 计算初始因子载荷阵 $A=(a_1,\cdots,a_k)$；
6) 计算共同度 $H=(h_1^2,\cdots,h_m^2)$， $h_i^2=\sum a_{ij}^2$， $i=1,\cdots,m$；
7) 求正交(旋转)因子解—因子的正交旋转；
8) 计算因子得分，如可求 R^{-1} 从而得因子得分矩阵。

根据筛选后的 12 个因子和 Fama-French 的五因子，对 17 个因子进行因子分析，得到五个因子，本文选取因子分析后的五个因子，再添加市场因子，一共六个因子对本篇文章进行分析说明。通过因子分析方法得到的五因子旋转矩阵如表 6：

Table 6. Factor rotation matrix
表 6. 因子旋转矩阵

	第一因子	第二因子	第三因子	第四因子	第五因子
每股净资产	0.8546	−0.1661	−0.1327	−0.0378	0.2638
资产报酬率	0.9397	−0.0602	−0.0964	−0.1717	0.1511
资产净利率	0.9275	−0.0764	−0.0082	−0.2999	0.0474
净利润/营业收入	0.9175	−0.0085	−0.1336	−0.1377	0.2643
营业利润/营业总收入	0.9194	−0.1155	−0.1628	−0.1373	0.2260
营业总成本/营业总收入	−0.8903	0.2071	0.1316	0.1059	−0.2955
营业利润率	0.9193	−0.1158	−0.1628	−0.1366	0.2266
产权比率	−0.2723	0.3215	0.0112	0.8035	−0.0221
营业收入增长率	0.7169	−0.3767	0.0865	0.1865	−0.0972
总资产周转率	−0.2014	0.0668	0.1746	0.1886	−0.8693
成交量	−0.1262	0.9099	−0.1377	0.1748	−0.1281
市盈率	−0.5514	0.0906	0.7429	0.1749	−0.0432
市场因子	−0.2784	0.6276	0.3230	0.4385	−0.3455
市值因子	−0.8721	0.0313	0.1926	−0.0184	0.2257
估值因子	0.0089	0.1075	−0.9438	0.0305	0.0561
投资因子	0.1883	−0.3221	0.1678	0.4058	0.6643
盈利因子	0.8588	−0.1425	0.0326	−0.0629	0.0866

表 3 中做标记的是权重相对于较大的因子，绝对值都超过了 0.8。对于第一个因子，每股净资产、资产报酬率、资产净利率、净利润/营业收入、营业利润/营业总收入、营业总成本/营业总收入、营业利润率、等对应的因子权重较大，由于这些因素跟盈利能力指标关联较大，因此可以称之为盈利能力因子，同样的道理可以称第二因子为价值指标因子，第三因子就称为估值因子，第四因子称为偿债能力因子，第五个因子就称为营运能力因子。由此得到五个新的市场特征(因子)，它们和市场因子(指数)一起将作为预测模型的选择特征。

2.6. 模型评估

为了对模型的预测结果进行评估，本篇文章选用了 MAPE 和 RMSE 来评价模型的预测能力：

$$\text{MAPE} = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{\text{real}-\text{predict}}{\text{real}}\right|$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\text{real}-\text{predict})^2}$$

其中 real 是真实值，predict 是预测值。

3. 模型

3.1. 多 CART 分类回归树

这里借鉴集成学习思想，提出多 CART 分类回归树模型，其基本思想是对不同的特征训练数据重复应用回归树算法，得到多个回归树，然后将这多个回归树预测值进行平均得到一个最终的预测值。结果表明此多 CART 分类回归树可以明显的提高回归树的预测准确性。

首先来回顾一下回归树的基本概念。回归树是数据挖掘和机器学习领域应用最广的算法之一，它在拟合数据时，先将预测变量 X 的联合空间划分成互不重叠的 J 个小区域 R_j，称作树的终端节点(或叶子)；然后为每一个小区域拟合一个常数 γ_j 作为这个小区域内响应变量 Y 的预测值:

$$x \in R_j \Rightarrow f(x) = \gamma_j$$

因此一个回归树可以表示为：

$$T(x;\Theta) = \sum_{j=1}^{J}\gamma_j I(x \in R_j), \Theta = \{R_j, \gamma_j\}_1^J$$

回归树的两组基本参数是小区域 R_j 以及小区域上相应的常数 γ_j，将其统一记做 Θ，参数估计的标准是：

$$\hat{\Theta} = \arg\min_{\Theta}\sum_{j=1}^{J}\sum_{x_j \in R_j} L(y_i, \gamma_j)$$

其中 $L(.)$ 是损失函数。在回归树中，最常用的损失函数是平方损失函数，$L(y, f(x)) = (y - f(x))^2$，回归树的参数是使训练样本残差平方和最小的那一组。

为了防止数据的过拟合，回归树应该进行剪枝处理，通过剪枝处理可以提高模型的泛化能力。CART 分类回归树的剪枝算法就是从完全决策树的底端减去一些子树，从而使模型变的简单。

本文将建立 6 棵 CART 分类回归树，每棵树的分类特征是前面得到的五个因子和市场因子，比如第一棵树是根据第一个因子前两季度的因子得分来预测当季度的收益率，其值也就是 T_1，同样的计算方法用其他五个因子预测当季的 收益率，即 T_2, T_3, T_4, T_5, T_6，于是得到最终的预测值 T，T 的计算方法如下：

$$T = \frac{T_1 + T_2 + T_3 + T_4 + T_5 + T_6}{6}$$

3.2. 模型的实现

本文使用因子分析和多 CART 分类回归树算法来建立预测模。首先利用一元回归对多个财务因子进行分析筛选，得到筛选后的因子之后，并和 Famma 五因子合并一起进行因子分析，得到新的五因子，加上市场因子共六个因子，由它们作为分类特征，接着根据 CART 分类回归树算法对季收益率进行预测，然后给出相应的投资策略，对公司进行模拟投资。本文的算法基本思路如图 1 所示：

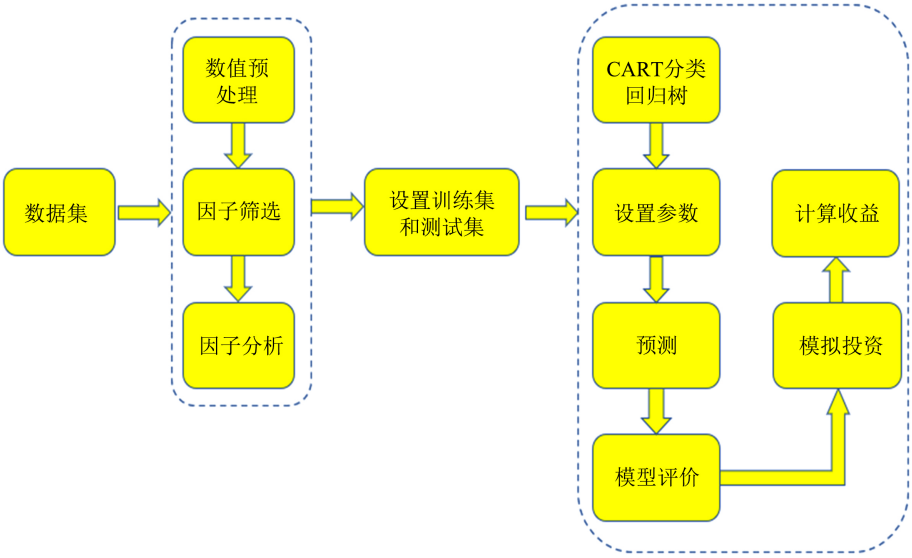

Figure 1. Flow chart
图 1. 流程图

4. 实证分析

4.1. 模型评价

根据 CART 分类回归树模型和因子的构造，最后得到市场及公司的每个季度数据，分别用六个因子中每个因子前两天的数据对季收益率进行预测，接着求平均，得到此因子的预测结果。根据 2002 年 12 月 31 号到 2019 年 12 月 31 号的数据，是随机选取 20 家公司，分别对每家公司 2002 年 12 月 31 号到 2015 年 12 月 31 号的数据进行训练，剩下的数据进行测试，模型的训练结果由 MAPE 和 RMSE 进行评估，得到表 7 如下结果：

Table 7. Model evaluation form
表 7. 模型评价表

公司	MAPE	RMSE	公司	MAPE	RMSE
1	0.0681	0.0784	11	0.0787	0.1066
2	0.0484	0.0904	12	0.0830	0.1323
3	0.0547	0.0757	13	0.0767	0.0889
4	0.0516	0.0654	14	0.0421	0.0545
5	0.0315	0.0436	15	0.0727	0.1009
6	0.0667	0.1041	16	0.0446	0.0824
7	0.0435	0.0603	17	0.0411	0.0465
8	0.0595	0.0907	18	0.0622	0.0826
9	0.0503	0.0645	19	0.0533	0.0803
10	0.0454	0.0522	20	0.0777	0.1088

由表 2 可知，通过基于因子分析与回归树的模型预测得出的结果，RMSE 最大值是 0.1323，最小值是 0.0436；MAPE 最大值是 0.0830，最小值是 0.0411，与以往结果比较，预测模型表现良好。

4.2. 模拟投资

为了更好地说明本文因子构造的合理性，对数据进行预测之后再进行模拟投资。投资过程是：如果投资者有一元人民币，当预测结果(未来投资收益率)大于 0，就进行买入；否则，就卖出股票。以下是对 20 家公司进行的模拟投资，用累计收益和夏普比率对收益和风险进行评价，其中累积收益计算过程如下：

$$CR = \prod_{i=1}^{n}(1+R_t)$$

其中 R_t 是 t 时间的季收益率。

夏普比率计算公式如下：

$$\text{sharperatio} = \frac{E(R_p)-R_f}{\sigma_p}$$

其中 $E(R_p)$ 是投资组合的期望收益率，R_f 是无风险利率，σ_p 是投资组合的标准差。

20 家公司四年的累计收益和夏普比率如表 8：

Table 8. Simulation investment evaluation table
表 8. 模拟投资评价表

公司	累计收益	夏普比率	公司	累计收益	夏普比率
1	5.5405	0.8618	11	5.4260	0.9006
2	6.0080	0.7621	12	7.7510	0.7670
3	6.4573	1.0585	13	5.8540	0.8984
4	5.7730	1.1779	14	7.4018	1.6251
5	3.1418	0.9745	15	6.4601	0.8199
6	4.5841	0.6780	16	5.1390	0.8393
7	2.9357	0.6788	17	4.6403	1.3952
8	3.3744	0.5524	18	3.8670	0.6265
9	3.3541	0.6932	19	6.9189	1.0417
10	4.8363	1.5570	20	4.3789	0.5992

从表 5 可以看出来，随机挑选的 20 家股票都能获得相应的收益，四年收益最高达到了 7.7510 倍，最低也达到了 2.9357 倍。正如下面几家公司的累积收益和市场收益图展示，如图 2，相比于市场收益，本文提出的方法获得的收益要比市场收益高很多，说明本文的方法是可行的。再来看夏普比率，最大值是 1.6251，最小值是 0.5524，说明相同的风险可以获得较高的收益。

Figure 2. Investment income comparison chart
图 2. 投资收益对比图

5. 结论

本文提出了基于多因子得分的 CART 分类回归树预测模型。首先利用财务数据对市场的影响效果对财务因子进行筛选，得到五个综合反映公司运营状况的五大特征，此五大特征揭示出证券市场不同的波动状态，接着根据市场特征和证券交易的前期收益率建立了多 CART 分类回归树预测模型，并依据预测结果构造出趋势投资策略，对 20 家公司进行模拟投资，预测和模拟投资结果都说明本文选用的算法对市场投资有良好的指导作用。

参考文献

[1] Huang, C., Huang, L.-L. and Han, T.-T. (2012) Financial Time Series Forecasting Based on Wavelet Kernel Support Vector Machine. 8*th International Conference on Natural Computation*, Chongqing, 29-31 May 2012, 79-83.

[2] Yoo, P.D., Kim, M.H. and Jan, T. (2005) Machine Learning Techniques and Use of Event Information for Stock Market Prediction: A Survey and Evaluation. *Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce*, Vienna, 28-30 November 2005, 835-841. https://doi.org/10.1109/CIMCA.2005.1631572

[3] Murphy, J.J. (1999) Technical Analysis of the Financial Markets. Institute of Finance, New York.

[4] Lee, M.-C. (2009) Using Support Vector Machine with a Hybrid Feature Selection Method to the Stock Trend Prediction. *Expert Systems with Applications*, **36**, 10896-10904. https://doi.org/10.1016/j.eswa.2009.02.038

[5] Khashei, M. and Bijari, M. (2010) An Artificial Neural Network (p, d, q) Model for Time Series Forecasting. *Expert Systems with Applications*, **37**, 479-489. https://doi.org/10.1016/j.eswa.2009.05.044

[6] Alberg, J. and Lipton, Z.C. (2017) Improving Factor-Based Quantitative Investing by Forecasting Company Fundamentals.

[7] Belciug, S. and Sandita, A. (2017) Business Intelligence: Statistics in Predicting Stock Market. *Annals of the University of Craiova, Mathematics and Computer Science Series*, **44**, 292-298. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85038626695&partnerID=40&md5=35b8b79a01a6b41338d3857b3e09c36c

[8] Sorensen, E.H., Miller, K.L. and Ooi, C.K. (2000) The Decision Tree Approach to Stock Selection. *The Journal of Portfolio Management*, **27**, 42-52. https://doi.org/10.3905/jpm.2000.319781

[9] Andriyashin, A., HHrdle, W.K. and Timofeev, R.V. (2008) Recursive Portfolio Selection with Decision Trees. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2894287

[10] Zhu, M., Philpotts, D. and Stevenson, M.J. (2012) The Benefits of Tree-Based Models for Stock Selection. *Journal of Asset Management*, **13**, 437-448. https://doi.org/10.1057/jam.2012.17

[11] Zhu, M., Philpotts, D., Sparks, R. and Stevenson, M.J. (2011) A Hybrid Approach to Combining CART and Logistic Regression for Stock Ranking. *The Journal of Portfolio Management*, **38**, 100-109. https://doi.org/10.3905/jpm.2011.38.1.100

[12] Chong, E., Han, C. and Park, F.C. (2017) Deep Learning Networks for Stock Market Analysis and Prediction: Methodology, Data Representations, and Case Studies. *Expert Systems with Applications*, **83**, 187-205. https://doi.org/10.1016/j.eswa.2017.04.030

[13] Krauss, C., Do, X.A. and Huck, N. (2017) Deep Neural Networks, Gradient-Boosted Trees, Random Forests: Statistical Arbitrage on the S&P 500. *European Journal of Operational Research*, **259**, 689-702. https://doi.org/10.1016/j.ejor.2016.10.031

[14] 刘建中, 殷其威. 基于 CART 算法的股票价格走势预测算法研究[J]. 计算机科学与应用, 2017, 7(6): 603-614.

[15] Fama, E.F. and French, K.R. (2015) A Five-Factor Asset Pricing Model. *Journal of Financial Economics*, **116**, 1-22. https://doi.org/10.1016/j.jfineco.2014.10.010

[16] Fama, E.F. and French, K.R. (1993) Common Risk Factors in the Returns on Stocks and Bonds. *Journal of Financial Economics*, **33**, 3-56. https://doi.org/10.1016/0304-405X(93)90023-5