Advances in Applied Mathematics 应用数学进展, 2018, 7(9), 1166-1173
Published Online September 2018 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2018.79136

Hans汉斯

A New Decision Tree Model for Securities Selection

Shuangshuang Li, Haijun Liu, Panpan Guo

School of Mathematics and Statistics, Zhengzhou University, Zhengzhou Henan
Email: 18937313871@163.com

Received: Aug. 22nd, 2018; accepted: Sep. 6th, 2018; published: Sep. 13th, 2018

Abstract

A new decision tree model for securities selection is established in this paper. First, the factors affecting the financial data of the stock are analyzed, and the factors are reduced. Secondly, the factors are discretized by K-means clustering method. Then, the decision tree algorithm is used to establish the stock forecasting model. Finally, the model is verified by the simulation investment.

Keywords

Securities Investment, Principal Component Analysis, Cluster Analysis, Decision Tree

一种新的证券选择的决策树模型

李双双，刘海军，郭盼盼

郑州大学，数学与统计学院，河南 郑州
Email: 18937313871@163.com

收稿日期：2018年8月22日；录用日期：2018年9月6日；发布日期：2018年9月13日

摘　要

本文建立了一种新的证券选择的决策树模型。首先，对影响股票的财务数据的因子做主成分分析，进行降维处理和因素选择；其次，利用K-means聚类法对因素进行离散化处理；然后，使用决策树算法建立股票预测模型；最后，利用模拟投资验证模型的有效性。

关键词

证券投资，主成分分析，聚类分析，决策树

文章引用: 李双双, 刘海军, 郭盼盼. 一种新的证券选择的决策树模型[J]. 应用数学进展, 2018, 7(9): 1166-1173.
DOI: 10.12677/aam.2018.79136

1. 引言

近年来，金融市场中的证券选择问题得到了大量的关注。证券选择就是利用投资标的的客观与微观信息对各种资产进行分类评估，并根据信息给出资源分配的方案。简单说，证券选择就是信息处理以及资源分配的方法。其中，人工智能与统计方法是信息处理的主要工具。人工神经网络[1] (ANN)、贝叶斯网络[2]以及 Logistic 回归[3] (LR)等这些方法已逐步受到市场人士的关注和认可。

决策树是人工智能中一种重要的方法，也正被应用于证券选择中。决策树是利用属性的信息增益对属性进行分类，建立树模型，给出决策方案。1986 年 Quinlan [4]提出了著名的 ID3 算法。在 ID3 算法的基础上，1993 年 Quinlan [5]又提出了 C4.5 算法。为了适应处理大规模数据集的需要，后来又提出了若干改进的算法[6]。近年来，很多专家把决策树用在股票选择上，Wu 和 Lin [7]在 2005 年提出了利用决策树技术的股票交易方法优于过滤规则；Zhou [8]在 2008 年使用了遗传算法和决策树相结合的方法；Jankowski [9]在 2016 年通过筛选特征和决策树结合的算法来预测股票的投资效率，都得到良好的结果。

本文首先用主成分分析法对财务数据进行降维处理，并用 K-means 聚类法对数据进行有效分类，最后用决策树得到了股票选择模型，并进一步利用模拟投资验证了该模型对股票选择有一定的价值。

2. 预备知识

2.1. 主成分分析

主成分分析是一种将多个变量降为少数几个综合变量的方法[10]。

设样本有 p 个变量，分别用 $X_1, X_2, \cdots, X_p$ 表示，构成 p 维随机变量 $X = \left(X_1, X_2, \cdots, X_p\right)'$，均值为 μ，协方差矩阵为 Σ。对 X 进行线性变换，形成新的综合变量 Y，如下所示[10]：

$$\begin{cases} Y_1 = a_1'X = a_{11}X_1 + a_{21}X_2 + \cdots + a_{p1}X_p \\ Y_2 = a_2'X = a_{12}X_1 + a_{22}X_2 + \cdots + a_{p2}X_p \\ \vdots \\ Y_p = a_p'X = a_{1p}X_1 + a_{2p}X_2 + \cdots + a_{pp}X_p \end{cases} \tag{1}$$

且有：

$$Var\left(Y_i\right) = Var\left(a_i'X\right) = a_i'\Sigma a_i \tag{2}$$

$$Cov\left(Y_i, Y_j\right) = Cov\left(a_i'X, a_j'X\right) = a_i'\Sigma a_j'. \tag{3}$$

定义 1 [10]设主成分 $Y_i\left(i = 1, 2, \cdots, p\right)$ 的特征值为 λ_i，则称 $\lambda_k \Big/ \sum_{i=1}^{p} \lambda_i$ 为主成分 Y_i 的贡献率；又称 $\sum_{i=1}^{m} \lambda_k \Big/ \sum_{i=1}^{p} \lambda_i$ 为主成分 $Y_1, \cdots, Y_m\left(m < p\right)$ 的累计贡献率。

2.2. K-means 聚类算法

K-means 算法是基于距离的聚类算法，它采用距离作为相似性的评价指标，即认为两个对象的距离越近，其相似度就越大。

首先，设 K 为聚类的个数，N 为数据样本，过程如下：

1) 从 N 个数据样本随机选取 K 个数据作为质心；

2) 对剩余的每个数据测量其到每个质心的距离(距离在本文中用欧氏距离)，并把它归到最近的质心的类；

3) 重新计算已经得到的各个类的质心；

4) 迭代 2~3 步直至新的质心与原质心相等或小于指定阈值，算法结束。

2.3. 决策树学习算法

决策树是一种典型的分类方法，是以信息增益来度量属性的选择，选择分类后信息增益最大的属性再次进行分类，直至分类结束。

首先定义一个刻画任意样例集纯度的度量标准，称为熵。如果某个目标属性具有 c 个不同的值，那么样例集 S 相对于 c 个状态的分类的熵定义为：

$$Entropy(S) \triangleq \sum_{i=1}^{c} -p_i \log_2 p_i \tag{4}$$

其中，p_i 是 S 中属于类别 i 的比例[11]。

然后定义属性分类训练数据的能力的度量标准，称为“信息增益”。简单的说，使用一个属性来分割样例导致的期望熵降低值就是这个属性的信息增益。一个属性 A 相对样例集合 S 的信息增益 $Gain(S,A)$ 被定义为：

$$Gain(S,A) \triangleq Entropy(S) - \sum_{v \in Values(A)} |S_v|/|S| \, Entropy(S_v) \tag{5}$$

其中，$Values(A)$ 是属性 A 所有可能值集合，S_v 是 S 中属性 A 的值为 v 的子集 $S_v = \{s \in S \mid A(s) = v\}$ [11]。

具体的计算步骤如下：

1) 计算总的熵；

2) 分别计算各个属性的信息增益；

3) 比较各个信息增益数值的大小，选择最大的作为根节点；

4) 在根节点的每个属性下，分别再求总的熵；

5) 再分别计算各个属性的信息增益并比较大小，选择下一个属性，…，直到每个属性被分配完，算法结束。

3. 因子与模型

3.1. 因子的选取与数据处理

不同于指标分析，基本分析面的分析侧重于长期分析，它的资料主要来自上市公司公开发布的财务报表、公司派息公告等，以季度为最短计量周期发布财务数据信息。本文选取了 11 个财务数据指标作为研究对象，如表 1 所示。

Table 1. Financial Index
表 1. 财务指标

反映偿债能力的指标	资产负债率、产权比率、权益乘数
反映盈利能力的指标	净利润、净资产收益率、营业利润率
反映股票每股平均能力的指标	每股净资产、每股营业收入、每股营业利润、每股资本公积金、每股未分配利润

本文所用数据均来自锐思数据库。选取上海证券交易所上市的2002年3月至2016年6月的54只股票。财务数据最短以季度为计量周期发布，周期较长，所以在分析时，共58组财务数据信息，我们需要预测的是下个季度的收益率。由于财务数据不存在数据缺失，因此不需要做缺失处理，只需对数据信息进行标准化即可。

3.2. 模型的结果与分析

3.2.1. 主成分分析

采用2.1的主成分分析法得到特征值。用MATLAB计算得出主成分结果，如图1所示。

从图中可以看出，从第四个主成分开始，特征值的变化趋势趋于缓慢，所以我们取三个主成分。经计算，我们可以算出主成分结果，如表2所示。

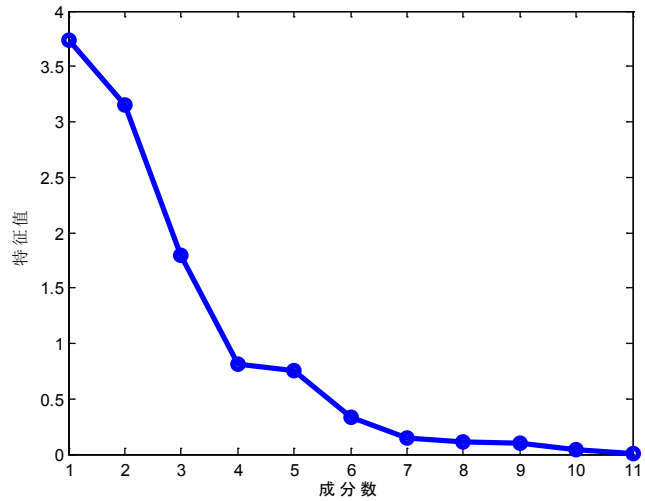

Figure 1. Eigenvalue
图 1. 特征值

Table 2. Principal Component Result
表 2. 主成分结果

属性	第一个主成分	第二个主成分	第三个主成分
Z_1：每股净资产	0.239	0.013	0.624
Z_2：每股营业收入	0.200	0.332	0.155
Z_3：每股营业利润	0.431	0.239	−0.105
Z_4：每股资本公积金	0.158	−0.179	0.500
Z_5：每股未分配利润	0.223	0.198	0.361
Z_6：净资产收益率	0.404	0.216	−0.277
Z_7：净利润	0.408	0.243	−0.106
Z_8：营业利润率	0.330	−0.027	−0.318
Z_9：产权比率	−0.263	0.466	0.044
Z_{10}：资产负债率	−0.255	0.469	0.044
Z_{11}：权益乘数	−0.266	0.466	0.041
特征值	3.838	3.355	1.996
贡献率	0.360	0.307	0.183
累计贡献率	0.360	0.667	0.850

从表中我们可以看出来，前三个主成分贡献率已达到 85%，所以我们可以用前三个主成分来代替原有的 11 个变量的数据信息。用 Y_1、Y_2、Y_3 表示第一、二、三主成分，则前三个主成分分别为：

$$\begin{aligned} Y_1 &= 0.239Z_1 + 0.200Z_2 + \cdots - 0.266Z_{11} \\ Y_2 &= 0.013Z_1 + 0.332Z_2 + \cdots + 0.466Z_{11} \\ Y_3 &= 0.624Z_1 + 0.155Z_2 + \cdots + 0.041Z_{11} \end{aligned} \tag{6}$$

由于得到的主成分不能很好的表示其实际意义，所以可以计算出载荷，然后对载荷阵做因子旋转，使得有因子分析得到的载荷阵可以实现载荷阵中所有元素或接近 0，或接近 ±1，旋转后的载荷阵如表 3 所示。

从表中可以看出，第一个主成分主要由净资产收益率、每股营业利润、净利润、营业利润率、每股营业收入来表示，第二个主成分主要由权益乘数、产权比率、资产负债率来表示，第三个主成分主要由每股净资产、每股资本公积金、每股未分配利润来表示。

3.2.2. K-means 聚类法

把 Y_1、Y_2、Y_3 作为新的数据，利用 K-means 聚类法对数据进行离散化处理。

结果如表 4。

Table 3. Rotating Load Array
表 3. 旋转后的载荷阵

指标	第一主成分	第二主成分	第三主成分
Z_6：净资产收益率	0.9439	−0.0520	−0.0314
Z_3：每股营业利润	0.9234	−0.0143	0.2042
Z_7：净利润	0.8907	0.0133	0.1874
Z_8：营业利润率	0.6495	−0.3751	−0.1655
Z_2：每股营业收入	0.5108	0.3714	0.3756
Z_{11}：权益乘数	−0.0602	0.9714	−0.0786
Z_9：产权比率	−0.0567	0.9684	−0.0722
Z_{10}：资产负债率	−0.0412	0.9670	−0.0660
Z_1：每股净资产	0.0884	−0.0954	0.9466
Z_4：每股资本公积金	−0.1360	−0.3453	0.7110
Z_5：每股未分配利润	0.3397	0.1720	0.6315

Table 4. Classification Situation
表 4. 分类情况

类别	分类	范围	样本数目
第一主成分	1	[−12.06, −0.99]	889
	2	[−0.99, 1.43]	1611
	3	[1.43, 9.80]	1285
第二主成分	1	[1.04, 8.15]	911
	2	[−5.84, −1.06]	895
	3	[−1.06, 1.04]	1326
第三主成分	1	[−0.72, 0.87]	1328
	2	[0.87, 5.24]	819
	3	[−4.83, −0.72]	985
收益	1	[−0.56, −0.1]	968
	2	[−0.1, 0.1]	1258
	3	[0.1, 2.37]	906

上表分别把每个主成分分成三类，所对应的收益也分成三类。类别分别用 1、2、3 来表示。将收益根据收益的数目，分为三类，分别为买入、不动、卖出，在表中也分别对应 1、2、3。

3.2.3. 决策树

进行分类之后，建立决策树。使用数据的前 2/3 作为训练数据，后 1/3 的数据作为测试数据。我们先求了主成分一、二、三的信息增益。为了方便在这里分别将主成分一、二、三记为 A、B、C。通过计算主成分 A、B、C。以及之下分类的信息增益，可以得到表 5。

通过上述表格中，计算信息增益，其中根据第一行得到主成分 A 为根节点，主成分 A 下收益为“1”的是主成分 B；主成分 A 下收益为“2”的是主成分 C；主成分 A 下收益为“3”的是主成分 B，结果得到一个决策树。在这么多的数据中，每个数据对应一个收益的分类，把主成分 A、B、C 分类全部相同的数据放在一起，看收益的分类，然后取概率最高的分类。

决策树图 2 所示。

其中，最下面一行是收益的分类。从图中可以得到，在所有的分类中，收益为“2”的占大多数，收益为“1”和“3”的占小部分。本文得到的这个决策树图可以预测股票的收益，给投资者规避风险。

例如，预测某一个样例的收益，查看上海浦东发展银行股份有限公司的某数据(表 6)。

经过对数据进行处理后，根据公式(7)，计算 Y_1、Y_2、Y_3 的值为−3.808、−1.401 和 0.222，根据 K-means 聚类法进行分类，结果分别为“1”、“1”和“3”。根据分类结果，对应图 2 决策树，找出所对应的收益分类为“2”至此可以大致预测收益的情况。

3.3. 模拟投资

通过对上述模型的构建，可以进行模拟投资。其中选取了几十家公司近 10 年的数据，按照上述的模型进行投资，可以得到累计收益率。

模拟投资步骤如下：

1) 根据建立的决策树模型和图 2，计算出每个数据的收益分类；

2) 收益分类为“1”，意为买入一支股票；收益分类为“2”，意为既不买入也不卖出；收益分类为“3”，意为卖出一支股票；

3) 设某个公司在一定时间的收益率为 $\omega_i\left(i=1,2,\cdots,n\right)$，$\lambda_i\left(i=1,2,\cdots,n\right)$ (其中 λ_i = 1，0 或−1)为在这段时间的收益分类，根据累计收益率的计算公式

$$\omega=\sum_{i=1}^{n}\lambda_i\omega_i-1 \tag{7}$$

即得到模拟投资的效果(表 7)。

通过上述步骤的计算，得到累计收益率表 7，从表中可以得到累计收益率在−0.6741 到 5.6254 之间，说明建立的决策树模型对股票的投资有着一定的价值。

Table 5. Information Gain
表 5. 信息增益

$Gain(S,A)=0.0293$	$Gain(S,B)=0.0022$	$Gain(S,C)=5.4544e-04$
$Gain(A1,B)=0.0383$	$Gain(A2,B)=0.0101$	$Gain(A3,B)=0.0021$
$Gain(A1,C)=0.0063$	$Gain(A2,C)=0.0192$	$Gain(A3,C)=0.0013$

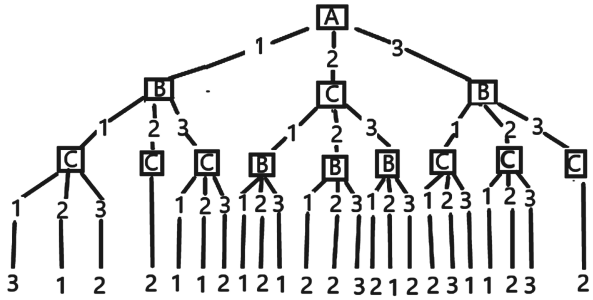

Figure 2. Decision Tree
图 2. 决策树

Table 6. Shanghai Pudong Development Bank Co., Ltd. March 31, 2002
表 6. 上海浦东发展银行股份有限公司 2002 年 3 月 31 日数据

每股净资产 (元/股)	每股营业收入 (元/股)	每股营业利润 (元/股)	每股资本公积金 (元/股)	每股未分配利润 (元/股)	净资产收益率 (%)
3.0424	0.7562	0.147	1.6141	0.1225	3.6912
净利润(元)	营业利润率(%)	产权比率(%)	资产负债率(%)	权益乘数	季累积收益率(%)
265,746,000	19.4377	2316.210378	95.861288	24.162104	0.0453

Table 7. Situating the Return of Individual Companies
表 7. 模拟投资各个公司收益情况

序号	公司全称	上市公司代码	时间	累计收益率
1	中国北方稀土(集团)高科技股份有限公司	C600111	2002.3.31~2016.6.30	5.6254
2	中船钢构工程股份有限公司	C600072	2002.3.31~2016.6.30	4.4689
…	…	…	…	…
53	甘肃亚盛实业(集团)股份有限公司	C600108	2002.3.31~2016.6.30	−0.289
54	广州发展集团股份有限公司	C600098	2002.3.31~2016.6.30	−0.6741

4. 结论

本文首先提出了对财务数据用主成分方法进行降维，用 K-means 聚类法进行分类，最后用决策树建立模型，来预测证券投资。经过在中国股票市场的实证检验，累计收益率在−0.6741 到 5.6254 之间，结果表明该方法具有一定的可行性，可作为经济分析预测工作的一种手段。

参考文献

[1] Dutta, A., Bandopadhyay, G. and Sengupta, S. (2012) Prediction of Stock Performance in the Indian Stock Market Using Logistic Regression. *International Journal of Business & Information*, **7**, 105-136.

[2] Moghaddam, A.H., Moghaddam, M.H. and Esfandyari, M. (2016) Stock Market Index Prediction Using Artificial Neural Network. *Journal of Economics Finance & Administrative Science*, **21**, 89-93. https://doi.org/10.1016/j.jefas.2016.07.002

[3] Yi, Z. and Kita, E. (2012) Stock Price Forecast Using Bayesian Network. Pergamon Press, Inc., Oxford.

[4] Quinlan, J.R. (1986) Induction of Decision Trees. *Machine Learning*, **1**, 81-106. https://doi.org/10.1007/BF00116251

[5] Quinlan, J.R. (1992) C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers Inc.

[6] Feng, Y. (2012) Study of Fault Diagnosis Method for Wind Turbine with Decision Classification Algorithms and Expert System. *Telkomnika Indonesian Journal of Electrical Engineering*, 10, No. 5.

[7] Wu, M.C., Lin, S.Y. and Lin, C.H. (2006) An Effective Application of Decision Tree to Stock Trading. *Expert Systems with Applications*, **31**, 270-274. https://doi.org/10.1016/j.eswa.2005.09.026

[8] Zhou, L., Si, Y.W. and Fujita, H. (2017) Predicting the Listing Statuses of Chinese-Listed Companies Using Decision Trees Combined with an Improved Filter Feature Selection Method. *Knowledge-Based Systems*, **128**, 93-101. https://doi.org/10.1016/j.knosys.2017.05.003

[9] Jankowski, D., Jackowski, K. and Cyganek, B. (2016) Learning Decision Trees from Data Streams with Concept Drift. *Procedia Computer Science*, **80**, 1682-1691. https://doi.org/10.1016/j.procs.2016.05.508

[10] 高慧璇. 应用多元统计分析[M]. 北京: 北京大学出版社, 2005.

[11] 曾华军, 张银奎. 机器学习[M]. 北京: 机器工业出版社, 2003.