浙江理工大学学报(自然科学版),第 33 卷,第 3 期,2015 年 5 月
Journal of Zhejiang Sci-Tech University (Natural Sciences)
Vol. 33, No. 3, May 2015

文章编号: 1673-3851 (2015) 03-0418-05

基于贝叶斯分类法的股票选择模型的研究

骆　桦, 张喜梅

(浙江理工大学理学院, 杭州 310018)

摘　要: 股票选择在证券投资中是一个重要问题,提出了一种基于朴素贝叶斯分类方法的股票选择模型。首先根据美股特斯拉的表现对沪深证券市场的能源股进行聚类分析,选取对股票投资价值影响显著的财务指标构造样本特征集,其次通过合理地选取贝叶斯分类器的参数对股票进行分类。经研究所选取股票的等权投资组合产生了44.6%的累计回报率,优于32.4%的基准回报率,表明朴素贝叶斯分类方法在股票选取方面具有较好的效果,值得投资者借鉴。

关键词: 股票选择; 投资决策; 聚类分析; 朴素贝叶斯分类方法

中图分类号: F830.91　　**文献标志码**: A

0　引　言

虽然股票市场是一个复杂的非线性系统,但未来的股票收益在一定程度上是可预测的。股票市场的各种预测方法一直是人们关注的焦点,许多文献已对此进行了探讨,如刘海玥等[1]用时间序列模型和神经网络模型对股票进行预测,钱颖能等[2]、左辉等[3]用贝叶斯分类法进行选股,获得的回报超出市场的平均回报。

本文提出了一种基于贝叶斯分类法的选股模型。首先根据美股特斯拉的表现对沪深证券市场的能源股进行聚类分析,再构造股票投资价值分类模型,通过模糊聚类对样本的财务指标进行简化;然后用朴素贝叶斯分类方法构造了两模式分类器,将股票分成两类,经研究所选取股票的等权投资组合的累计回报率优于基准回报率。

1　聚类分析

聚类分析是依据研究对象的个体特征,对其进行分类的方法[4],以某种度量为标准使得同一聚类中样本之间的相关性比其他聚类中样本之间的相关性更加显著。

设有 n 个样本,每个样本都有 p 个指标的观察值。设 x_{ik} 表示第 i 只股票的第 k 个指标的值,则这 n 只股票可以看成 p 维空间中的 n 个点。两只股票之间的相似度可以用 p 维空间中两点之间的距离来度量。d_{ij} 表示第 i 只股票与第 j 只股票的距离,距离越小表示这两只股票的性质越接近,可以归为一类。聚类分析中常用的距离定义为欧氏距离,可用下式表示:

$$d_{ij}=\left[\sum_{k=1}^{p}|x_{ik}-x_{jk}|^2\right]^{\frac{1}{2}}。$$

2　朴素贝叶斯分类方法简介

贝叶斯分类是一种基于贝叶斯定理的统计学分类方法,可以预测给定样本属于一个特定类的概率。由分类算法的比较研究可以发现,在实际应用中朴素贝叶斯分类算法在分类性能上可以与决策树和神经网络分类算法相媲美[5-6]。

2.1　贝叶斯定理

$P(C_i|X)=\dfrac{P(X|C_i)P(C_i)}{P(X)}$ 是在条件 X 下 C_i

收稿日期: 2014－06－12

基金项目: 国家自然科学基金项目(11401532)

作者简介: 骆　桦(1962－),男,浙江诸暨人,副教授,硕士生导师,主要从事金融数学和数理统计方面的研究。

发生的概率,称为 C_i 在条件 X 下的后验概率。$P(C_i)$是 C_i 的先验概率。类似的,$P(X|H)$是 X 在条件 H 下的后验概率,$P(X)$是 X 的先验概率。

贝叶斯定理提供了一种根据 $P(X)$、$P(H)$和 $P(X|H)$,计算后验概率 $P(H|X)$的方法,其公式可表示为:

$$P(H|X)=\frac{P(X|H)P(H)}{P(X)},$$

其中 $P(X)$、$P(H)$和 $P(X|H)$可以由给定的数据集合计算得到。

2.2　朴素贝叶斯分类方法

2.2.1　朴素贝叶斯假设

朴素贝叶斯分类假定一个属性值对给定类的影响独立于其他属性的值,即$P(x_1,x_2,\cdots x_n|C)=\prod_{i=1}^{n}P(x_i|C)$,作此假设可有效简化计算,在此意义下被称为"朴素的"。

2.2.2　朴素贝叶斯分类器参数估计

数据样本是由属性值组成的特征向量。每个数据样本用一个 n 维特征向量 $X=\{x_1,x_2,\cdots,x_n\}$表示,分别描述其 n 个属性的观察值。假定数据样本可以划分为 m 个不同类别,$C_1,C_2,\cdots,C_m$。各类别的先验概率可以由公式 $P(C_i)=\frac{S_i}{S}$进行估算,其中 S 为样本总数,S_i 为类 C_i 中的样本数。如果各类别的先验概率是未知的,通常可假定各类别是等概率的,即:

$$P(C_1)=P(C_2)=\cdots=P(C_m)。$$

2.2.3　朴素贝叶斯分类方法

a) X 为给定的未知类别的数据样本,朴素贝叶斯分类器将预测 X 属于具有最大后验概率(条件 X 下)的类。即,将 X 分类到类 C_i 当且仅当

$$P(C_i|X)>P(C_j|X)\quad 1\leqslant j\leqslant m,j\neq i,$$

其中 $P(C_i|X)$最大的类 C_i 被称为最大后验假定。

b) 由贝叶斯定理可知:

$$P(C_i|X)=\frac{P(X|C_i)P(C_i)}{P(X)},$$

假设 $P(X)$对于任一类别都是相同的,要 $P(C_i|X)$取最大,只需要 $P(X|C_i)P(C_i)$最大。其中 $P(C_i)$、$P(X|C_i)$的值可以根据朴素贝叶斯分类器参数估计来确定。

c) 为将未知样本 X 分类,对每个类别 C_i 相应的 $P(X|C_i)P(C_i)$进行估算,把样本 X 指派到类 C_i,当且仅当

$$P(X|C_i)P(C_i)>P(X|C_j)P(C_j)\quad 1\leqslant j\leqslant m,j\neq i。$$

3　样本的选取

3.1　特斯拉股票的波动率计算

股票价格波动率是定价模型中的重要因素之一,在其他参数不变的情况下,股票价格波动率越大,投资的风险也越大,但获利的机会也越大。

令 $S(t)$代表某股票在 t 时刻的价格,$S(t)$满足方程(1)[7]:

$$\mathrm{d}S=\mu S\mathrm{d}t+\sigma S\mathrm{d}B \qquad (1)$$

其中:μ 是股价漂移率,σ 是股价波动率,B 服从布朗运动。则方程(1)中的波动率可以用 $\sigma=\frac{S}{\sqrt{\Delta t}}$来估算。

下面对特斯拉股票的波动率进行估算,选取从 2012 年 3 月到 2014 年 2 月美股特斯拉汽车股票(TSLA)每月的股票收盘价格为股价样本,以月为计算单位,则 $\Delta t=\frac{1}{12}$,由此可以求出:

年波动率为:$\sigma=\frac{S}{\sqrt{\Delta t}}=\frac{0.383\,3}{\sqrt{\frac{1}{12}}}=1.327\,8=132.78\%$,

漂移率为:$\mu=\frac{\overline{U}+\frac{S^2}{2}}{\Delta t}=\frac{0.146\,5+\frac{0.383\,3^2}{2}}{\frac{1}{12}}=2.639\,6=263.96\%$。

3.2　应用聚类分析选取样本

从上面的结果来看,特斯拉股价的年平均收益率已高达 263.96%,年波动率高达 132.78%。大的波动率意味着投资风险大但获利机会也大。特斯拉高波动率伴随高收益率的表现对投资者有启示作用。笔者在沪深股市中随机挑选出了 100 家企业,统计 2013 年度各企业经营业务,用一个二维向量(x,y)表示。其中,当企业经营业务与能源有关时令 $x_i=1$,否则 $x_i=0$;当企业经营业务不仅与能源有关而且涉及到新能源时令 $y_i=1$,否则 $y_i=0$。定义特斯拉为(1,1),根据欧氏距离公式计算各企业之间的距离得到欧氏距离矩阵,由系统聚类法中的最短距离法,最终把与特斯拉性质相近的 56 家企业归为一类,剩余的为另一类。这 56 家企业经营业务涉及锂离子电池、镍氢电池、镍镉电池、燃料电池、太阳能电池等与特斯拉经营业务相近,分别为:德赛电池、京能电力、南都电源、新

宙邦、长信科技、江苏国泰、南洋科技、赣锋锂业、横店东磁、申能股份、亿纬锂能、复星医药、比亚迪、金瑞科技、杉杉股份、天齐锂业、佛山照明、特变电工、川投能源、金山股份、卧龙电气、风帆股份、航空动力、万向钱潮、粤水电、中材科技、包钢稀土、安泰科技、科力远、长城电工、中炬高新、航天机电、金风科技、隆基股份、新大洲 a、泰豪科技、风华高科、光电股份、拓邦股份、中信国安、国光电器、岷江水电、佛塑科技、方大集团、华芳纺织、湘电股份、曙光股份、中国宝安、天成控股、中科英华、江苏阳光、云天化、* ST 天威、动力源、乐山电力、青鸟华光。

4 股票投资价值分类模型的建立

4.1 股票分类模型的建立

第一步：选取数据样本，以股票的财务指标作为样本的属性值。

第二步：通过聚类、筛选对数据样本的财务指标进行简化，实现样本属性的选择与提取[8]。尽可能满足给定类变量时，属性变量之间条件独立，提高朴素贝叶斯分类方法的准确率和效率。

第三步：采用朴素贝叶斯分类方法构造股票分类器。根据其表现将样本分成两类，一类是高回报股票(C_1)，另一类是普通回报股票(C_2)。将第 i 个上市公司的财务指标用一个 n 维特征向量 $X=\{x_1,x_2,\cdots,x_n\}$ 表示，其中 x_k 代表属性 A_k 的取值，这样股票样本可以形成一个 n 维欧氏空间。股票的预期未来回报是一个二元相关变量 $y_i=\pm1$，其中 $+1$ 代表高回报股票，而 -1 代表普通回报股票。这样一个由 m 个上市公司财务数据组成的训练集可以表示为：

$T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m)\}\in(R^n\times Y)^m$，其中 $x_i\in R^n$，$y_i\in Y=\{+1,-1\}$，$i=1,2,\cdots m$。分类器可以看成是一系列将预测变量映射到分类值±1的函数。构造股票分类器的目的就是通过调整其参数以最小的出错率和高的效率将股票样本分类。

第四步：检验股票分类器的分类效果。本文根据各企业 2013 年和 2014 第一季度的财务数据，每一季度进行一次实验，基于所收集的数据，共进行了五次实验。

4.2 股票投资价值特征集的选取

根据公司价值的一般理论，公司价值主要体现在成长能力、盈利能力、运营能力、偿债能力等方面。股票投资价值的影响因素主要包括：每股收益 x_1，每股净资产 x_2，每股资本公积金 x_3，主营业务利润率 x_4，营业利润率 x_5，资产报酬率 x_6，净资产收益率 x_7，净资产增长率 x_8，总资产周转率 x_9，资产负债率 x_{10}。如果直接采用这 10 项指标作为股票投资价值的特征集，由于指标之间的相关性比较强，容易造成模型效率低下，也影响模型的准确性。为确保股票投资价值特征集的完备性和可操作性，本文对影响股票投资价值的因素进行聚类、筛选，过程如下。

第一步：选取 56 家上市公司 2013 年第一季度的财务数据作为训练集，每个样本用 10 个财务指标表示其性状，构造数据矩阵。

第二步：数据标准化，采用相关系数法建立模糊相似矩阵 R，如式(2)所示[9]。

$$\begin{pmatrix}
1.000 & 0.077 & 0.114 & 0.238 & 0.181 & 0.784 & 0.875 & 0.418 & 0.367 & 0.234\\
0.077 & 1.000 & 0.840 & 0.103 & 0.278 & 0.092 & 0.025 & 0.133 & 0.067 & 0.207\\
0.114 & 0.840 & 1.000 & 0.086 & 0.056 & 0.110 & 0.217 & 0.159 & 0.247 & 0.224\\
0.238 & 0.103 & 0.086 & 1.000 & 0.429 & 0.149 & 0.006 & 0.158 & 0.368 & 0.192\\
0.181 & 0.278 & 0.056 & 0.429 & 1.000 & 0.373 & 0.466 & 0.024 & 0.131 & 0.063\\
0.784 & 0.092 & 0.110 & 0.149 & 0.373 & 1.000 & 0.842 & 0.283 & 0.479 & 0.430\\
0.875 & 0.025 & 0.217 & 0.006 & 0.466 & 0.842 & 1.000 & 0.246 & 0.508 & 0.117\\
0.418 & 0.133 & 0.159 & 0.158 & 0.024 & 0.283 & 0.246 & 1.000 & 0.059 & 0.147\\
0.367 & 0.067 & 0.247 & 0.368 & 0.131 & 0.479 & 0.508 & 0.059 & 1.000 & 0.016\\
0.234 & 0.207 & 0.224 & 0.192 & 0.063 & 0.430 & 0.117 & 0.147 & 0.016 & 1.000
\end{pmatrix} \tag{2}$$

第三步：利用二次平方法求模糊相似矩阵 R 的传递闭包：$R\to R^2\to R^4$，$R^4\circ R^4=R^4$，即模糊等价矩阵，如式(3)所示。

$$\begin{pmatrix}
1.000 & 0.278 & 0.278 & 0.429 & 0.466 & 0.842 & 0.875 & 0.418 & 0.508 & 0.430 \\
0.278 & 1.000 & 0.840 & 0.278 & 0.278 & 0.278 & 0.278 & 0.278 & 0.278 & 0.278 \\
0.278 & 0.840 & 1.000 & 0.278 & 0.278 & 0.278 & 0.278 & 0.278 & 0.278 & 0.278 \\
0.429 & 0.278 & 0.278 & 1.000 & 0.429 & 0.429 & 0.429 & 0.418 & 0.429 & 0.429 \\
0.466 & 0.278 & 0.278 & 0.429 & 1.000 & 0.466 & 0.466 & 0.418 & 0.466 & 0.430 \\
0.842 & 0.278 & 0.278 & 0.429 & 0.466 & 1.000 & 0.842 & 0.418 & 0.508 & 0.430 \\
0.875 & 0.278 & 0.278 & 0.429 & 0.466 & 0.842 & 1.000 & 0.418 & 0.508 & 0.430 \\
0.418 & 0.278 & 0.278 & 0.418 & 0.418 & 0.418 & 0.418 & 1.000 & 0.418 & 0.418 \\
0.508 & 0.278 & 0.278 & 0.429 & 0.466 & 0.508 & 0.508 & 0.418 & 1.000 & 0.430 \\
0.430 & 0.278 & 0.278 & 0.429 & 0.430 & 0.430 & 0.430 & 0.418 & 0.430 & 1.000
\end{pmatrix} \tag{3}$$

第四步:聚类。求出模糊等价矩阵之后,通过计算 λ-截距阵获得不同的分类。并通过 F 统计量确定 λ 的最佳值为 0.430,将上述 10 项财务指标分成 4 类,具体分类为:$\{x_2, x_3\}$、$\{x_4\}$、$\{x_8\}$、$\{x_1, x_5, x_6, x_7, x_9, x_{10}\}$。针对第四类中的多个指标采用相关指数法筛选,选出 x_7 纳入特征集。通过聚类、筛选将原先的 10 项指标精简为 4 项:每股净资产、主营业务利润率、净资产收益率、净资产增长率。

5　分类实验

根据财务报表统计 56 家上市公司 2013 年第一个交易日的股票开盘价和 2013 年最后一个交易日的股票收盘价格,计算其涨幅,前 25%的股票被定义为高回报股票($y_i=+1$),其余的被定义为普通回报股票($y_i=-1$)。每个数据样本用一个 4 维特征向量 $X=\{x_1, x_2, x_3, x_4\}$ 表示,分别描述每股净资产、主营业务利润率、净资产收益率、净资产增长率的实际值,样本数据见表 1。

表 1　训练样本数据

	每股净资产/元	主营业务利润率/%	净资产收益率/%	净资产增长率/%	股票类别
复星医药	6.220 0	43.571 6	2.530 0	34.570 3	+1
拓邦股份	2.304 7	20.166 7	1.200 0	−1.175 2	+1
新宙邦	6.970 6	30.280 5	1.990 0	7.680 6	−1
南都电源	9.160 0	13.896 5	1.030 0	5.438 0	−1

数据说明:在训练样本集中,选取了 56 只股票,限于文章的篇幅,上表仅给出了其中部分股票的数据,此数据是以各公司 2013 年第一季度的财务报表为基础。表中复星医药公司控股子公司主要科研产品包括燃料电池轿车发动机、燃料电池大巴发动机。

评价一个股票分类器是否有用,仅注重预测精度是远远不够的,关键是看应用这个分类器所选出的股票的回报率如何,通常的做法是以基准回报率为参照物,将所选择的股票产生的回报率与其对比。基准回报率一般指由全部被选中进行分类的股票组成的等权投资组合所产生的回报。表 2 显示了每个季度对测试数据进行的分类结果。

表 2　朴素贝叶斯法则测试结果

项目	2013 年第 1 季度	2013 年第 2 季度	2013 年第 3 季度	2013 年第 4 季度	2014 年第 1 季度
总股票数	56	56	56	56	56
选出的高收益股票数	3	5	4	7	5
选出的低收益股票数	53	51	52	49	51
高收益股票精确度/%	100.00	60.00	75.00	42.86	100.00
低收益股票精确度/%	79.24	70.59	65.38	63.27	60.78
总精确度/%	89.62	65.30	70.19	63.27	80.39
回报/%	2.329	17.338	5.979	17.117	1.836
基准回报/%	0.625	14.417	3.109	13.539	0.742

从表 2 可以看出,朴素贝叶斯分类方法的预测精度虽然不是特别理想,但是由朴素贝叶斯分类器选出的高收益股票所组成的等权投资组合,在所测试的每个季度产生的回报都要超过基准回报。高收益股票在 2013 年第一季度至 2014 年一季度中共获得 44.6%的收益,明显优于基准回报 32.4%的收益。

朴素贝叶斯分类假定属性值相互条件独立。考虑到这样的独立性条件比较强,加之缺乏可用的概率数据,使得贝叶斯分类器的预测准确率受到影响。尽管如此,上述研究表明,贝叶斯分类器在股票选择问题上具有较好的分类效果,应该受到投资者的重视。

6 结 论

本文针对新能源股票使用朴素贝叶斯分类方法选股具有较好的效果,与市场基准回报率相比,运用该方法所选取的高收益股票的等权投资组合获取的收益具有明显的优势。由此可见使用贝叶斯分类方法在股票选取方面具有一定的实用价值。但是该方法的预测精度不是太理想,有待进一步提高。

参考文献:

[1] 刘海玥,白艳萍. 时间序列模型和神经网络模型在股票预测中的分析[J]. 数学的实践与认识,2011,41(4):14-19.

[2] 钱颖能,胡运发. 用朴素贝叶斯分类法选股[J]. 计算机应用与软件,2007,24(6):90-92.

[3] 左 辉,楼新远. 基于贝叶斯分类的选股方法[J]. 电脑知识与技术,2008(10):173-176.

[4] 李庆东. 聚类分析在股票分析中的应用[J]. 辽宁石油化工大学学报,2005,25(3):94-96.

[5] Sun L, Shenoy P P. Using Bayesian networks for bankruptcy prediction: some methodological issues[J]. European Journal of Operational Research, 2007, 180(2): 738-753.

[6] Perzyk M, Biernacki R, Kochański A. Modeling of manufacturing processes by learning systems: the naïve Bayesian classifier versus artificial neural networks[J]. Journal of Materials Processing Technology, 2005, 164: 1430-1435.

[7] Joseph S, Victor G. The Mathematics of Finance: Modeling and Hedging[M]. 北京:机械工业出版社,2003:86-89.

[8] Yu H, Chen R, Zhang G. A SVM stock selection model within PCA[J]. Procedia Computer Science, 2014, 31: 406-412.

[9] 谢季坚,刘承平. 模糊数学方法及其应用[M]. 2版. 武汉:华中科技大学出版社,2004:106-107.

Research on Stock Selection Model Based on Bayesian Classifier

LUO Hua, *ZHANG Xi-mei*

(School of Science, Zhejiang Sci-Tech University, Hangzhou 310018, China)

Abstract: Stock selection is an important problem in the securities investment. This paper proposes a stock selection model based on naive Bayesian classification method. Firstly, according to the performance of the US stocks Tesla, cluster analysis is made for energy stock on Shanghai and Shenzhen securities markets. Financial indexes which impose significant influenced on stock investment value are selected to construct the sample set. Secondly, we classify stocks by choosing the appropriate parameters of naive Bayesian classifier. After study, the equally weighted portfolio of stocks yields accumulative return rate of 44.6%, which is better than the benchmark return rate of 32.4%. The results show that naive Bayesian classification method has good effects on stock selection and investors should refer to it.

Key words: stock selection; investment decision; cluster analysis; naive Bayesian classifier

(责任编辑:康 锋)