聚类分析和支持向量机在股票研究中的应用

狄明明，孙德山
（辽宁师范大学 数学学院，辽宁 大连 116029）

摘 要：随着我国股市的逐步规范和完善，投资者愈加重视投资对象的选择。针对股票分类的特点，选取对上市公司股票走势有重要影响的7项主要财务指标，运用聚类分析和支持向量机相结合的方法对80家上市公司的股票进行分类。为了降低分类实验的复杂程度，在分类实验中采用因子分析法将原来的7项财务指标用3个综合指标概括。实验结果表明，这种方法大大降低了股票数据的维数，有很高的分类正确率，分类达到了让人满意的效果，证明这种分类方法是可行的。

关键词：因子分析；聚类分析；支持向量机；股票研究；分类
中图分类号：O235；TP181　　**文献标识码**：A　　**文章编号**：1673－629X(2009)06－0229－03

Applications of Cluster Analysis and Support Vector Machines to Stock Research

DI Ming-ming, SUN De-shan
(Institute of Mathematics, Liaoning Normal University, Dalian 116029, China)

Abstract: As China's stock market gradually standardized and improved, investors paid more attention to the selection of investment targets. Based on the characteristics of the stock classification, selected the data that greatly influenced the stock development trend of listed companies. 7 financial ratios from 80 stocks of listed companies had been studied by means of cluster analysis and support vector machines. Aiming at reducing the data complexity of the classification experiment, 7 primary financial ratios were summarized by three comprehensive indexes in classification experiment in the way of factor analysis. According to the experimental results, the method introduced in this paper reduced the dimensions of stock data greatly, and it had high classification accuracy. The satisfactory results of classification proved its feasibility.

Key words: factor analysis; cluster analysis; support vector machines(SVM); stock research ; classification

0 引 言

中国股市在不到二十年的时间里迅速发展壮大。随着股市的逐步规范和完善，价格向其内在价值回归是未来股市发展的重要方向。股票档次将不断拉开。投资者将会更加重视上市公司的经营业绩，重视股票自身的品质，即重视投资对象的选择。但是，随着股市发展、投资手法和证券监管方法的成熟，以及上市公司数量的不断增多，如何科学合理地进行股票的分析和选择是每一个投资者所要解决的首要问题。

笔者选取了80家上市公司的股票，根据2008年第一季度各股票的信息及数据，选取了每股收益、每股净资产、净资产收益率、投资收益、利润总额、净利润、流通A股这7项重要的财务指标进行分析，从而对股票进行分类。首先，应用因子分析将上市公司股票的数据指标(高维数据)变换到低维空间上。也就是采用因子分析进行特征提取，将原来每家股票的7项经济指标用少数几个不相关的综合因子来表示；其次用聚类分析对股票进行分类；最后用支持向量机进行训练并检测分类的正确率。

收稿日期：2008－09－25；**修回日期**：2008－12－09
基金项目：辽宁省高等学校科研项目(2008343)
作者简介：狄明明(1982－)，女，辽宁本溪人，硕士研究生，研究方向为概率统计；孙德山，博士，硕士生导师，研究方向为概率统计。

1 因子分析

在多元统计中，因子分析[1]是一种很有效的降维和信息浓缩技术，是用最少个数不可观测、互不相关的若干公共因子与一个特殊因子的线性组合，来描述原来一组可观测的相互有关的每个变量，其目的是尽可能合理地解释存在于原始变量之间的相关性，并简化变量的维数与结构。

多元因子分析法基于如下的多元因子数学模型：假设原有变量有 p 个，分别用 $x_1, x_2, x_3, \cdots, x_p$ 表示；因子变量有 m 个，分别用 $F_1, F_2, F_3, \cdots, F_m$ 表示，运

用多元因子分析法可建立如下数学模型[2]：

$$\begin{cases} x_1 = a_{11}F_1 + a_{12}F_2 + \cdots + a_{1m}F_m + a_1\varepsilon_1 \\ x_2 = a_{21}F_1 + a_{22}F_2 + \cdots + a_{2m}F_m + a_2\varepsilon_2 \\ \cdots\cdots \\ x_p = a_{p1}F_1 + a_{p2}F_2 + \cdots + a_{pm}F_m + a_p\varepsilon_p \end{cases}$$

该数学模型的矩阵形式为：

$$X = AF + a\varepsilon$$

其中 $F(F_1,\cdots,F_m)'$ 为公共因子，$A = \begin{pmatrix} a_{11}\cdots a_{1m} \\ a_{p1}\cdots a_{pm} \end{pmatrix}$ 为因子载荷矩阵，这里的 $a_{ij}(i = 1,2,\cdots,p, j = 1,2,\cdots, m)$ 称为因子载荷；$\varepsilon = (\varepsilon_1,\cdots,\varepsilon_p)'$ 为特殊因子，在实际中通常忽略不计。多元因子分析法就是运用数学方法找出因子载荷矩阵 A，对所得到的各因子，首先观察它们在哪些变量上的载荷较大，在哪些变量上的载荷较小，再根据载荷大的变量本身的内容说明因子的含义。

针对上述 7 项财务指标，采用 SPSS 软件运算，得到 3 个主成分，其累积贡献率达到 85.78%。说明用这 3 个综合因子可以概括出原来 7 项财务指标的绝大部分信息。因此可以用这 3 个综合因子来代表原来的 7 项财务指标。

2 聚类分析

聚类分析[3]又称群分析，是研究对样品或指标进行分类的多元统计方法。所谓“类”，通俗地说就是相似元素的集合。文中采用的是系统聚类方法，其基本思想是先将所有样品各看成一类，然后选择性质最接近的两类合并成一个新类，计算新类与其它类的距离，再合并，直到所有的样品合并为一类。文中，聚类分析采用离差平方和法聚类，相似性统计量采用欧氏距离系数。

其具体的计算过程如下[4,5]：

a. 原始数据标准化，构成标准化数据矩阵。

b. 计算欧氏距离矩阵，选出最小距离样本组。计算两两样本间欧氏距离构成距离矩阵，从中选出具有最小距离的样本组。将具有最小距离的样本组归并为一类，当计算新类与其他样本类之间的距离时，采用离差平方和法计算类之间的距离，计算完毕后再从中选择具有最小距离的两类。

c. 用新的样本类代替原来的一对样本类。

d. 对新形成的样本数据与其余样本数据重新计算欧氏矩阵，以代替原矩阵，再找出新矩阵中最小距离的对应样本类，如此重复 c 到 d 的步骤，直到把所有样本都归为一类为止。

e. 最后按下列原则连接成谱系图：

(1)若两个样本在已经归并成类的类中未出现过，则它们归类一个新类；

(2)若两个样本中有一个在某类中出现过，则另一个就加入该类；

(3)若两个样本都在同一类中，则这对样本不再归类；

(4)若两个都已在不同类中出现过，则把两类归并在一起。

上述计算过程用 SPSS 统计分析软件进行，并通过对聚类图的分析，将上面的 80 家股票划分为 3 类较合适。根据上市公司以往的表现和业绩，可以将这 3 类股票分为蓝筹股、一般股和垃圾股。

3 支持向量机(SVM)原理

SVM[6~8]是一种基于结构风险最小化原理的机器学习方法。它最初于 20 世纪 90 年代由 Vapnik 提出，由于其出色的学习性能，该技术已成为机器学习界的研究热点，并且在很多领域都得到了成功的应用。

文中是一个三类分类问题，属于多类分类问题。由于构造多类分类的一般方法具有很高的计算复杂性，因此以一类分类算法为基础建立一种多类分类算法。该方法是在高维特征空间中对每一类样本求出一个超球体中心，然后计算待测试样本到每类中心的距离，最后根据最小距离来判断该点所属的类。具体步骤如下所述：

设训练样本为 $\{(x_1,y_1),\cdots,(x_l,y_l)\} \subset R^n \times Y$，$Y = \{1,2,\cdots,M\}$，其中，$n$ 为输入向量维数，M 为类别数。将样本分成 M 类，各类分开写成，$\{(x_1^{(s)},y_1^{(s)}),\cdots,(x_{l_s}^{(s)},y_{l_s}^{(s)}), s = 1,\cdots,M\}$，其中，$\{(x_i^{(s)},y_i^{(s)}), i = 1,\cdots,l_s\}$ 代表第 s 类训练样本，$l_1 + \cdots + l_M = l$。首先给出原空间中的优化算法，为了求包含每类样本的最小超球体，同时允许一定的误差存在，构造下面的二次优化：

$$\min \sum_{s=1}^{M} R_s^2 + C\sum_{s=1}^{M}\sum_{i=1}^{l_s}\xi_{si} \tag{1}$$

约束为：

$$(x_i^{(s)} - a_s)^{\mathrm{T}}(x_i^{(s)} - a_s) \leqslant R_s^2 + \xi_{si}$$

$$s = 1,\cdots,M, i = 1,\cdots,l_s \tag{2}$$

$$\xi_{si} \geqslant 0, s = 1,\cdots,M, i = 1,\cdots,l_s \tag{3}$$

该优化问题的对偶形式为：

$$\max \sum_{s=1}^{M}\sum_{i=1}^{l_s}\alpha_i^{(s)} < x_i^{(s)}, x_i^{(s)} > - \sum_{s=1}^{M}\sum_{i=1}^{l_s}\sum_{j=1}^{l_s}\alpha_i^{(s)}\alpha_j^{(s)} < x_i^{(s)}, x_j^{(s)} > \tag{4}$$

约束为:

$$0 \leqslant \alpha_i^{(s)} \leqslant C, s = 1, \cdots, M, i = 1, \cdots, l_s \quad (5)$$

$$\sum_{i=1}^{l_s} \alpha_i^{(s)} = 1, s = 1, \cdots, M \quad (6)$$

借助核映射思想,首先通过映射 ϕ 将原空间影射到高维空间,然后在高维特征空间中进行上面的优化,并通过引入核函数 $K(x,y)$ 代替高维特征空间中的内积运算,于是可以得到核方法下的优化方程为:

$$\max \sum_{s=1}^{M} \sum_{i=1}^{l_s} \alpha_i^{(s)} K(x_i^{(s)}, x_i^{(s)}) - \sum_{s=1}^{M} \sum_{i=1}^{l_s} \sum_{j=1}^{l_s} \alpha_i^{(s)} \alpha_j^{(s)} K(x_i^{(s)}, x_j^{(s)}) \quad (7)$$

约束为:

$$0 \leqslant \alpha_i^{(s)} \leqslant C, s = 1, \cdots, M, i = 1, \cdots, l_s \quad (8)$$

$$\sum_{i=1}^{l_s} \alpha_i^{(s)} = 1, s = 1, \cdots, M \quad (9)$$

上面优化式是多类分类问题最终的优化方程,待优化的参数个数是样本总数 l。因此,该优化方程的计算复杂性主要与总的样本数量有关,而样本的分类数对算法复杂性的影响很小。由此可知,该算法在处理多类分类问题时比用 SVM 构造一系列二分类要简单得多。

根据KKT条件,对应于 $0 < \alpha_i^{(s)} < C$ 的样本满足:

$$R_s^2 - (K(x_i^{(s)}, x_i^{(s)}) - 2\sum_{j=1}^{l_s} \alpha_j^{(s)} K(x_j^{(s)}, x_j^{(s)}) + a_s^2) = 0 \quad (10)$$

利用式(10)分别计算出 R_s 的值,$s = 1, \cdots, M$。

给定待识别样本 z,计算它到各个中心点的距离:

$$f_s(z) = K(z,z) - 2\sum_{i=1}^{l_s} \alpha_i^{(s)} K(z, x_i^{(s)}) + \sum_{i=1}^{l_s} \sum_{j=1}^{l_s} \alpha_i^{(s)} \alpha_j^{(s)} K(x_i^{(s)}, x_j^{(s)}), s = 1, \cdots, M \quad (11)$$

比较大小,找出最小的 $f_k(z)$,则 z 属于第 k 类。同时可定义该分类结果的信任度如下:

$$B_k = \begin{cases} 1, & \text{当 } R_k \geqslant f_k(z) \\ \dfrac{R_k}{f_k(z)}, & \text{否则} \end{cases} \quad (12)$$

式(12)表明当所得的 $f_k(z)$ 值位于超球体内部时,此时的信任度为 1,否则,信任度小于 1,并且距离超球体中心越远,信任度越小。该算法的关键是找到各类的中心点,因此还可以通过适当调整参数 C 的取值来抑制噪声的影响。

从 80 个上市公司的股票中随机选取 50 个股票作为训练样本,剩下的 30 个股票作为测试样本。文中所构造的SVM模型的内积核函数,通过线性函数与多项式函数及径向基函数相互比较,最终确定采用最常用的径向基函数:$K(x, x_i) = \exp(- \| x - x_i \|^2 / \delta^2)$。

δ,C 等参数采用交叉验证方法确定为 $\delta = 0.1$,$C = 0.5$。具体程序略。

从实验结果可以看出,支持向量机在测试样本集中的测试准确率达到 92.86%。说明这种分类方法是可行的,对于股票的分类具有很好的效果。

4　结束语

采用聚类分析和支持向量机结合的方法对股票进行分类,并用因子分析对股票数据进行简化,这在股票研究中是一个新的尝试。其实质是先利用因子分析对传统的指标进行特征提取,把多维评价指标综合成少数几维评价指标。然后将聚类分析和支持向量机相结合,先用聚类分析将样本分类,再用支持向量机训练,找出每类训练样本的类中心,再算出测试样本点到每类类中心的距离,从而看出其归为哪一类。实验表明这一方法大大降低了特征空间的维数,并有很好的分类正确率。但由于该方向发展时间较短,因此还有很多问题值得去深入研究。

参考文献:

[1] Flad M, Jung R C. A common factor analysis for the US and the German stock markets during overlapping trading hours [J]. Journal of International Financial Markets, Institutions and Money, 2008, 18(5): 498 - 512.

[2] 童　帆. 因子分析模型在学生多元化评价中的应用[J]. 统计教育, 2007(11): 40 - 42.

[3] Tola V, Lillo F, Gallegati M, et al. Cluster analysis for portfolio optimization [J]. Journal of Economic Dynamics and Control, 2008, 32(1): 235 - 258.

[4] 于　华. 上市公司综合评估的聚类与主成分分析[J]. 证券经纬, 2007(9): 49 - 50.

[5] 柯　冰, 钱省三. 聚类分析和因子分析在股票研究中的应用[J]. 上海理工大学学报, 2002, 24: 372 - 374.

[6] 孙德山. 支持向量机分类与回归方法研究[D]. 长沙: 中南大学, 2004.

[7] 张晨希, 张燕平, 张迎春, 等. 基于支持向量机的股票预测[J]. 计算机技术与发展, 2006, 16(6): 34 - 36.

[8] Huang Wei, Nakamori Y, Wang Shou - Yang. Forecasting stock market movement direction with support vector machine [J]. Computers & Operations Research, 2005, 32(10): 2513 - 2522.