基于改进K-Means聚类算法的股票聚类研究

李开伟
中南民族大学，中国·湖北　武汉　430072

【摘　要】在受到多重因素复合影响的股票市场中，不同行业股票投资价值也不尽相同。研究投资者如何在不同行业合理配置资产是一个受众面广泛且具有实际应用意义的课题。本文以2020年第三季度A股市场中具有代表性的150支股票为例，利用皮尔逊相关系数对一系列股票市场行情指标进行分析，筛选出每股收益、营业收入同比增长等七个具有代表性的指标。本文利用标准化处理后的指标数据进行了对股票的聚类分析，根据聚类分析结果，本文最终得出A股市场在第三季度的表现中，制造业更适合稳健型投资者投资，农林牧渔业更适合风险性投资者投资，可用于分析其他季度的股票市场行情并给投资者应如何在股票市场优化投资策略带来一些参考性建议。

【关键词】聚类分析算法；A股市场配置优化策略；皮尔逊相关系数分析

Research on Stock Clustering Based on Improved K-Means Clustering Algorithm

Li Kaiwei
South-Central University for Nationalities, Wuhan, Hubei, China 430072

[Abstract]In the stock market affected by multiple factors, the value of stock investment in different industries is not the same. Studying how investors rationally allocate assets in different industries is a topic with a wide audience and practical application significance. Taking 150 representative stocks in the A-share market in the third quarter of 2020 as an example, this paper uses the Pearson correlation coefficient to analyze a series of stock market indicators, and selects seven key indicators such as earnings per share and year-on-year growth in operating income. representative indicator. This paper uses the standardized index data to conduct cluster analysis on stocks. According to the results of the cluster analysis, this paper finally concludes that in the performance of the A-share market in the third quarter, the manufacturing industry is more suitable for investment by stable investors, and agriculture, forestry and animal husbandry are more suitable for investment. Fishing is more suitable for investment by risky investors and can be used to analyze the stock market conditions in other quarters and bring some reference suggestions to investors on how to optimize their investment strategies in the stock market.

[Key words]Cluster analysis algorithm; A-share market allocation optimization strategy; Pearson correlation coefficient analysis

引言

随着我国市场经济的发展以及各项制度不断深化改革，我国资本市场也在逐步发展完善，股票市场已经成为我国资本市场的重要组成部分，人们的理财意识和投资意识日益增强，股票投资已经成为人们进行资产配置的重要手段。因此，如何根据股票行情指标选取股票，做好风险收益的平衡，合理优化股票配置具有十分重大的理论意义和非常普世的应用价值与指导意义 。

股票市场是受到各类信息与因素共同影响，难以进行准确预测。且不同股票市场在应对风险的稳健性不同。因此对于普通股民，如何选择合视的股票进行投资，确定股票的交易时机是关键所在。本文以2020年中国A股股票市场为例，利用聚类分析等方法研究不同行业股票市场行情的波动情况，给投资者进行股票投资提供参考。通过聚类结果，投资者可以直观的看到不同行业受到疫情冲击的程度，据此根据自身风险承受能力和投资风格最终进行一只或一组股票的选取搭配以达到投资风险与收益的合理预期。

1　基于改进K-Means聚类算法的股票聚类模型建立

聚类分析根据数据内在几何结构和数据间的相似性，发掘出数据中隐藏的结构特征，并可通过可视化的形式进行展现。在证券市场上，对数据的获取、利用和分析程度，直接关系到证券投资者是否可以获得满意的收益。但是证券数据具有种类繁多、结构复杂等特性，因此如何高效进行数据的表示和分析是一个具有挑战性的课题。

1.1 K-Means聚类算法在股票市场中的应用研究

国内外专家学者针对K-Means聚类算法及其在股票市场中的应用进行了一系列研究。刘骏，喻青[1]对K-Means聚类算法的优缺点进行了总结并提出了性能优化方法；唐绍聪[2]利用聚类算法生成了可视化类群图进行分析，但并未对聚类初始中心的生成进行优化；王子龙等[3]指出K-means 算法随机初始化聚类中心易导致算法陷入局部最优，因此提出一种基于距离和样本权重的算法以实现改进；此外，基于K-Means及相关改进算法，在股票市场行情规律变化方面，也有一系列研究开展。陈金林，杨林[4]基于股票价格时间序列数据进行相似性度量，对股票价格的状态及变化规律进行了探究；项睿等[5]基于聚类分析，将数据挖掘与股票分析结合实现了股票价格的预测；上述研究虽然较为充分，但完备性仍然较差且并且不具备较强的代表性。

因此本文利用针对股票的改进K-Means聚类算法对股票进行聚类研究，基于皮尔逊相关系数分析，筛选出若干具有代表意义且良好可解释性的股票市场行情指标，进行股票市场行情分析与投资策略建议。

1.1.1 算法优化

K-means算法由于运算方法相对简单，原理与其他聚类方法相比较容易，运算速度快等优点，因此这种聚类算法较为常用。然而K-means算法在拥有这些优点的同时，也存在一些缺点和不足。该方法下，初始中心的选择具有很大的随机性且K-means对初始中心的选择高度依赖。

在基于股票聚类模型的研究中，本文采用改进的K-means算法来提高结果的准确性，一方面，使用交叉验证，根据损失函数来确定最优K值，另一方面，初始聚类中心，及簇中心点之间的间距应该较大。因此，本文确定其优化策略为：

（1）计算所有样本点之间的距离，选择距离最大的一个点对（两个样本C1， C2）作为2个初始中心点，从样本点集中去掉这两个点；

（2）如果初始中心点个数达到k个，则终止。若否，则在剩余的样本点中，选一个点C3，这个点优化的目标如式（1）所示：

$$\begin{cases}\min\limits_{C3}\left\{\max\left\{|C_3-C_1|,|C_3-C_2|\right\}\right\}\\ \max\limits_{C_3}\left\{|C_3+C_1|+|C_3+C_2|\right\}\end{cases}\quad 式（1）$$

1.2 基于改进K-Means聚类算法的股票聚类模型

由于股票指标存在多重共线性，如果把全部维度指标拿来作为聚类指标会造成维度灾难问题，因此，本文通过皮尔逊相关系数分析，选取合适的指标。相关系数的绝对值越大，相关性越强，相关系数越接近于1或-1，相关度越强，相关系数越接近于0，相关度越弱。其中，本文所采用皮尔逊相关系数的计算如下式（2）所示：

$$r=\frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}}\quad 式（2）$$

其中，X_i，Y_i代表每一个样本的观测值，X，Y代表样本的平均值。因此，由以上推导本文利用SPSS，得出股票评价指标的皮尔逊相关系数分析如下图1所示：

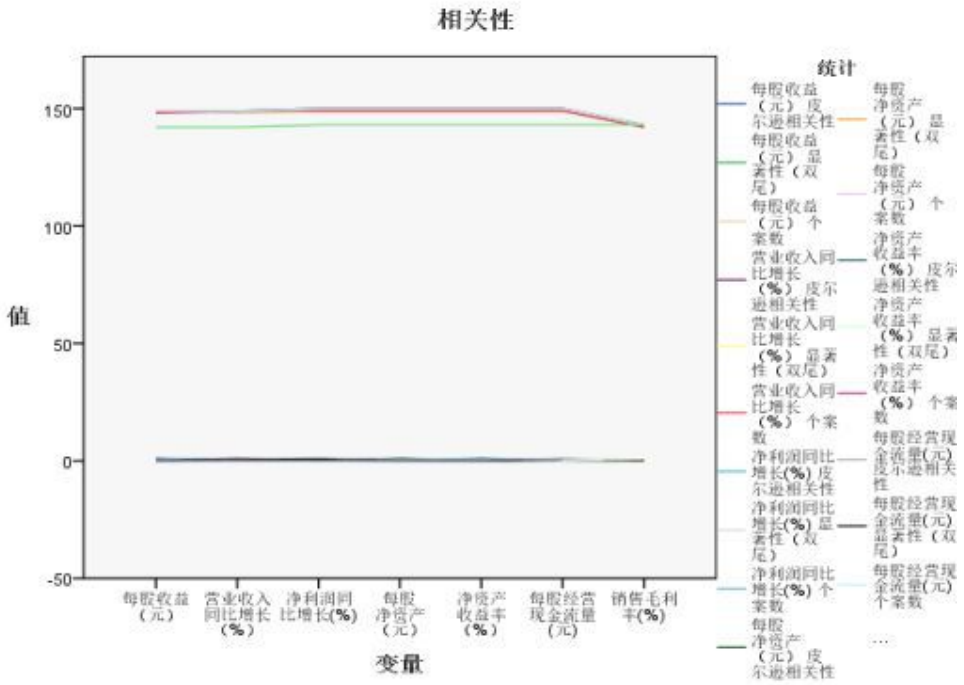

图1. 股票评价指标的皮尔逊相关系数比较示意图

2 算例分析

2.1构造算例

本文选取了沪市A股的三季报数据，以行业分类为代表，根据皮尔逊相关系数结果，挑选了每股收益、营业收入同比增长、净利润同比增长、每股净资产、净资产收益率、每股经营现金流量、销售毛利率作为聚类的指标，由于以上指标量纲不同，以下采用标准化方式为 Z-score 标准化进行数据预处理，结果如表1所示：

表1.标准化处理后的各类指标具体数据

股票序号	每股收益/元	营业收入同比增长/%	净利润同比增长/%	每股净资产/元	净资产收益率/%	每股经营现金流量/元	销售毛利率/%	所处行业
1	0.538	0.512	0.347	-0.6	1.39	0.273	-0.3	造纸
2	-0.06	-0.09	0.113	-0	-0.1	-0.324	-0.6	汽车
……	……	……	……	……	……	……	……	……
149	-0.43	-1.23	-0.02	-0.5	-0.4	-0.135	0.03	交运
150	3.378	0.449	0.197	3.07	0.72	3.477	-0.1	汽车

考虑到K-Means类聚类算法需要对所采用的各维度指标计算均值，要求各维度指标取值均为数值型，因此，本文对表1中的所处行业的数据进行语义分析并聚类，得到其数值化表达方法定义取值，并作出如下数值化规定如表2所示：

表2.股票所处行业数值化表达方法对应表

所属行业	农林牧渔	制造业	运输业	食宿服务	房地产
数值化取值	-9	-3	-1	-1/3	-1/9
所属行业	金融和保险	文体娱乐	其他行业	信息传输与技术、软件	
数值化取值	1/3	1/9	1	3	

2.2 结果分析

本文在AMD四核的计算机CPU，内存情况下做的测试，在Matlab12.0平台上编码实现。根据软件聚类数据，我们得到了每只股票的聚类结果，为保证结果可靠性，每个行业的样本总数均为16，结合每只股票的所属行业，我们绘制了行业聚类结果示意表如表3所示：

图中数据表明：从簇一到簇七的行业中每个维度的均值呈总体下降趋势，与预期的聚类结果相吻合，其中，结合表4，2020沪市A股的三季中表现比较稳健的有制造业、食宿服务业、文化体育和娱乐业，这可能与在国内新冠肺炎疫情得到整体控制下，第三季度各行业各企业顺利复工复产，各地各政府推出的刺激消费、拉动经济增长政策以及十一黄金周等节假日人们对出行旅游的意愿比较强烈有关。而像运输业等其他表现相对来说不是很好的行业，可能是因为受到全球疫情的影响，对外贸易数量减少的原因。

2.2.1簇间的平均距离

在Matlab12.0平台上进行编程后，我们得到了每个簇对应每个属性的坐标矩阵，据此可以得到簇间的平均距离，结果如下：

由簇间平均距离，可以看出簇一到簇七的之间的距离总体趋势在逐渐变大，符合聚类分析的预期结果，对于异常值，例如簇一到簇四的平均距离，结合簇内各维度的均值和簇内各行业样本数，我们可以分析可能是农、林、牧渔业在净利润同比增长这一指标下的发展情况较好，拉近了簇四和簇一这两个簇聚类中心的距离，但由于其他指标如净资产收益率、每股现金经营流量等，与簇一内对应行业的各指标仍存在一定的差异，因此两簇聚类中心还是存在一定距离。

此方法仍可以用于其他异常值的分析，结合结合簇内各维度的均值和簇内各行业样本数，我们可以得到两簇之内各行业造成异常值的指标，以帮助我们更好地整体把握相同簇内各行业股票各指标的稳健性，实现在行业类别上的投资合理配置。

表 3.行业聚类结果示意表

簇 / 样本数 / 行业	一	二	三	四	五	六	七
农、林、牧渔业	2	1	3	5	4	1	0
制造业	7	2	3	2	1	1	0
运输业	2	3	0	1	3	5	2
食宿服务	6	3	1	3	2	1	0
信息传输、软件和信息技术	4	6	0	1	3	0	2
房地产	2	4	3	3	1	2	1
文化体育和娱乐	7	1	3	2	1	1	1
金融和保险	2	3	5	1	1	2	2
其他行业	5	2	5	5	1	1	3

3 结论

在受到多重因素影响的股票市场，投资者如何根据不同市场的行情优化自己的投资策略是一个庞大的话题。本文以2020年A股市场第三季度不同行业的股票各类指标为基础，通过分析不同行业股票市场的稳健性，来对全行业的股票市场进行宏观分析，帮助投资者在不同行业中优化自己的投资策略给出建议我们在利用了皮尔逊相关系数对股票进行相关性分析后选取了七个数值型股票指标，并用数值化表达后的行业类型作为聚类分析的指标，根据分析结果我们得到2020年A股市场第三季度中制造业、食宿服务业、文化体育和娱乐业的市场稳健性比较好，这些行业可以作为保守型进行投资；而运输业、房地产业和金融保险业等行业的情况相对不佳，即便如此，在一些稳健性较差的行业中其某项指标也可能出现较好的趋势，如农、林、牧渔业的净利润同比增长等，综合考虑行业内股票的综合表现，这些行业可以作为风险型进行投资。保守型+风险型的投资策略可以使投资者根据自己的实际情况如最大风险承受能力等，在不同行业中获取满意的投资回报。

本文的模型在宏观上对各行业进行了分析，并给出了在不同行业中的投资建议，但是如果要确切到对行业中哪只股票投资，还要结合具体股票的具体指标数据等来进行综合分析，每个行业参考的侧重指标可能也有所不同。本模型也可以参考以往资料，对每个行业的指标进行加权分析以优化模型。

表 4. 簇间的平均距离

簇号	一	二	三	四	五	六	七
一	0.0000	1.1116	1.1980	0.6093	1.2313	1.3604	1.7267
二	1.1116	0.0000	0.4155	1.2795	0.7905	0.7205	0.9656
三	1.198	0.4155	0.0000	1.4649	0.6599	0.5907	0.8908
四	0.6093	1.2795	1.4649	0.0000	1.3518	1.6698	1.9197
五	1.2313	0.7905	0.6599	1.3518	0.0000	0.8316	0.8835
六	1.3604	0.7205	0.5907	1.6698	0.8316	0.0000	0.6281
七	1.7267	0.9656	0.8908	1.9197	0.8835	0.6281	0.0000

参考文献：

[1]刘骏,喻青.K-Means聚类算法及其性能优化研究[J].电子工业专用设备,20 20,49(05):46-49.

[2]唐绍聪.基于K-means算法的英语成绩聚类分析[J].信息技术与信息化, 2020 (07) :63-65.

[3]王子龙,李进,宋亚飞.基于距离和权重改进的K-means算法[J/OL].计算机工程与应用:1-11[2020-11-18].http://kns.cnki.net/kcms/detail/11.2127.TP.20201019.1 345.002.html.

[4]陈金林,杨林.基于聚类分析的股票价格状态和变化探究[J].时代金融,2018 (35):349+345.

[5]项睿,吴华玲,李琳,张立.基于K-Means聚类算法的股票技术指标分析[J].电脑编程技巧与维护,2019(12):4-7.