DOI:10.19473/j.cnki.1008-4940.2018.01.003

基于沪深 300 成份股的多因子量化选股策略研究

苏靖宇，方宏彬

（安徽大学　经济学院，安徽　合肥，230601）

[摘　要]　选取沪深 300 成份股作为样本股，截取 2007-2016 年财务数据和行情数据，通过模糊 C 均值聚类算法进行有效但冗余因子的剔除，构建多因子选股模型，将投资组合收益作为检验模型有效性的依据开展研究。结果表明，市盈率、市净率、每股收益等因子对股价波动规律有较强的关联性，多因子选股模型应用于投资实践能够取得稳定的相对于沪深 300 指数收益的超额收益，模糊 C 均值聚类算法与多因子选股模型的结合获得了较好的验证，为量化投资研究提供了新思路。

[关键词]　多因子选股模型；模糊 C 均值聚类；组合收益；超额收益

[中图分类号]　F224.0　　[文献标识码]　A　　[文章编号]　2096-3300（2018）01-0021-08

随着经济全球化的不断深入，资本在各国之间的流动日益频繁，世界经济愈来愈呈现出国际化态势，各国金融市场更加趋向融合性、紧密型发展。特别是中国，自改革开放以来，社会主义市场经济逐步确立，经济贸易逐步与世界接轨，金融市场与国际的联系更加紧密，短短 40 年，中国就发展成为世界第二大经济体，取得了辉煌的成就。理财、股票、信用卡消费等金融产品日益成为人们日常生活中不可或缺的必备产品，普通消费者从投资理财中获得了丰厚的回报，越来越多的投资者出没于证券行业以期获取更多的收益，促进了金融市场的蓬勃发展。

当然，机遇与挑战并存，在快速发展的过程中，我国金融市场也同样面临着巨大的挑战。首先，我国的资本市场是在借鉴国外先进的思想与方法的过程中不断发展与完善的，由于起步较晚，经济体制还不完善，必然存在着这样或那样的缺陷，这就使得信息不对称、投机行为、内幕交易等不利于金融市场发展的情况时有发生，严重损害了广大投资者的切身利益，使得大量的投资者望而却步，也令许多投资者因损失惨重而失去参与资本市场运作的信心，严重影响我国资本市场的良性发展。其次，传统的投资方法主要依赖于投资者对历史数据的经验判断以及投资者情绪等基本面分析和技术分析，尽管有数据的支持，但更多地依赖于投资者的经验累积，以定性分析为主。

随着计算机技术发展及人工智能时代的到来，计算机技术因其运算速度快、操作方便快捷、代

收稿日期：2017-09-23

作者简介：苏靖宇（1995-），男，安徽亳州人，硕士研究生，研究方向：量化投资；
方宏彬（1972-），男，安徽池州人，副教授，博士，研究方向：数据挖掘与量化投资。

替人工进行算法的高效处理等优势，越来越多地被应用于经济领域。量化投资依托于计算机技术的发展，采取合理的运算方法为投资决策提供有意义的指导方向，不断冲击着传统的投资理念，也渐渐被更多的投资者所接纳，所以，自量化投资出现以来，越来越多的投资者投入到量化投资的研究与应用中，并在此过程中取得了显著成果。基于以上考虑，采用量化投资的方法和理论能更好地适应当前金融市场的高速发展，成为金融投资分析的主要手段之一。

一、文献综述

国外学者都曾以上市公司基本面为研究对象，从财务状况、现金流、盈利能力等方面对股票的潜在价值进行深入的研究。Fama 等[1]的研究表明，股价的波动通常由多个因素的影响共同决定，单个因素的影响不能准确地描述上市公司内在价值；Asness[2]，Chen、Zhang[3]，Partha S. Mohanram[4]的研究表明，上市公司的基本面数据变化的同时伴随着股价的波动，良好的基本面使股价更易于偏向上升趋势，而基本面相对较差的则价格会出现不同程度的下跌，并且运用一定的数学逻辑方法从大量的基本面因子中选出最具代表性的几个因子，以这些有效因子为选股依据构建投资组合，依此更易获得稳定的超额收益。

当前，量化投资的研究已更加深入人心，多种量化选股策略在我国的 A 股市场上得到了很好的验证。范龙振[5]通过对 1995 年到 2000 年所有 A 股股票月收益率的研究，得出股票市场具有显著的市值效应、账面市值比效应、市盈率效应和价格效应，进行单因素对股价波动影响的 Fama-Macbeth 回归分析，并构造三因子影响模型，验证了三因子模型能很好地解释中国股票市场众多指数的差异；汪洋[6]是比较早开始研究股票收益与影响指标的学者，选取四个指标进行因子分析，结合中国的股市特点和实际国情，选取 19 家上市公司的股票作为样本股进行研究，得出股票的收益与股票的评估指标体系存在一定的相关性，通过对股票的评估指标体系研究可辅助投资者增加做出正确的投资决策的概率，尽管对影响股票收益率的因子进行研究有十分合理的依据，但是在因子选择等方面还尚不完善，不存在普遍性；王艳萍[7]基于多因子选股模型构建新的金融投资模型——多因子结构下的静态 MV 模型，通过详细的研究与分析得出在不允许卖空的条件下的解析最优解，该理论克服了现有文献对投资权重选择缺乏科学性和可操作性不强的缺点；丁鹏[8]在《量化投资——策略与技术》一书中首次将量化方面的研究进行系统化的阐述，为后来者对量化投资的研究提供了极好的参考价值，书中多因子选股模型将 30 个影响股价波动的因子作为候选因子，通过对因子下样本股票收益的对比，找出有效因子并对有效因子进行去冗处理，最终形成基于有效因子选取的多因子选股模型，通过其筛选的投资组合在 A 股中取得了较好的超额收益；孙守坤[9]以多因子选股模型与行业轮动模型相结合对沪深 300 股指期货进行检验并获取了超过指数的稳定收益率；王瑞[10]从众多影响股票波动的因子中选取 17 个简单易查的候选因子，通过有效性和冗余检验后使用基于等权重打分法的 Z 评分模型对投资组合进行实证分析，并通过累计收益率、超额年化复合收益率、夏普比率以及取得正收益的概率等指标对模型进行了评价。

本文在前人对多因子选股模型的研究基础上，以有效因子的检验、有效但冗余因子的剔除、投资组合的构建等三个步骤详细描写多因子选股模型的理论方法，其中对有效因子的冗余剔除方法作出了改进，创新之处在于将模糊 C 均值聚类算法应用到多因子选股模型，充分利用聚类算法对于具有相同或相似属性的影响股价波动因素的归类的原理，改进了有效但冗余因子的剔除，使模型更加合理、高

效，进一步为多因子选股模型的研究提供了一个可供参考的方向。

二、模型分析

（一）数据处理与指标选取

以 2016 年 7 月份调整后的沪深 300 指数成分股的全部样本作为样本依据，以 2007—2016 年共 10 年作为样本期，其中 2007—2014 年作为因子筛选期，2015 —2016 年作为选股策略的样本检验期，在整个模型期内包括了上涨、下跌以及震荡趋势，其目的为了检验不同趋势下模型的表现能否达到预期效果。所选股票样本为所有正常交易且上市时间超过一个季度的 A 股股票，业绩基准参考沪深 300 指数。候选因子主要包括财务数据和行情数据，数据均来自国泰安数据库（为了数据保持一致性，候选因子样本股股价均采用季度数据），本文从价值因子、成长因子、品质因子、技术因子中选取 16 个因子作为候选因子。价值因子反映股票的内在价值，能对股票价格高低进行初步的判断，能给投资者发现与买入被低估的股票提供参考，以期在价格上升时获取收益；成长因子是对公司成长性的度量指标，成长性的好坏关乎着公司未来的发展前景，具有高成长性的公司往往会在未来一段时间内股价上涨，从而受到众多投资者的关注；品质因子反映一段时间内公司的运用状况，能直观地反映公司的资金周转能力，也很大程度上决定了公司的经营效率；技术因子是投资者在日常投资中由技术面分析得到的，它是由一定的数理公式推导出来的，具有一定的解释能力。具体如表 1 所示。

表 1　候选因子

Tab. 1　Candidate factors

因子类型	价值因子	成长因子	品质因子	技术因子
候选因子	账面市值比	总资产收益率	资产负债率	换手率
	盈利收益率	净资产收益率	固定资产比率	
	现金收益率	营业收入净利率	股东权益周转率	
	市盈率	每股收益	固定资产周转率	
	市净率		流通市值	
	股息率			

（资料来源：国泰安数据库）

在进行数据处理时，剔除了数据缺失严重的样本，对于 ST 股在这里不作特殊处理，根据数据挖掘理论对部分数据处理的方法，本文对个别缺失数据使用前后加权平均的方法补充，由于缺失的股票数据较少，相对于大样本数据来说不会对实证结果造成显著影响。特别地，由于是用历史数据来推断未来股票价格波动规律即因子与样本之间存在一定的滞后性，因此在进行多因子选股模型前需要作滞后性处理以解决此类问题。本文以上一期的因子指标数据来表示本期股价波动的依据，消除滞后性的影响，使模型更加合理。

（二）多因子选股模型概述

在数据处理与候选因子选取之后，需要建立多因子选股模型对数据进行处理，多因子选股模型主要有候选因子有效性检验、有效但冗余因子的剔除、综合评分模型的构建和模型检验 4 个步骤。

1. 候选因子有效性检验

主要采用排序的方法，具体而言，对于每一个候选因子，在模型形成期的第一个持有期初开始计算市场中每只正常交易股票的该因子的大小，按从

大到小的顺序对样本股票进行排序，并平均分为 n 个组合，一直持有到期末，在下个持有期初再按相同的方法重新构建 n 个组合持有到期末，依次构建每个持有期一直到模型形成末期。组合构建完毕后，计算这 n 个组合的年化复合收益率、相对于基准业绩的超额收益、在不同市场情况下高收益组合跑赢基准和低收益组合跑输基准的概率。

2. 有效但冗余因子的剔除

对有效因子进行模糊 C 均值聚类分析，以此剔除具有相近或相似属性的有效因子，将有效因子按照一定的逻辑算法，将具有相近或相似性质的因子聚为一类，再从每一类中选择一个对股价波动影响最显著的因子。

模糊 C 均值聚类算法理论简单，易于理解，适用于解决很多学科范围的问题，方便计算机语言的实现，是一种非常有效的模糊聚类算法，计算每个样本隶属于各类的隶属度矩阵，该算法对于分类对象之间没有明确的界限，往往具有比较满意的聚类效果。模糊 C 均值聚类算法基本思想为给定样本观测矩阵 $X=\{x_1, x_2, x_3, \cdots, x_n\}$（$x_n \subset R^d$），其中元素 x_i 为 d 维向量，也就是说 X 是由 d 个样品的 n 个变量的观测值构成的矩阵。模糊 C 均值聚类就是将 n 个样品划分为 c 类（$2 \le c \le n$），记 $V=\{v_1, v_2, v_3, \cdots, v_c\}$ 为 c 个类的聚类中心，使得:

$$J(U, V)=\min \sum_{k=1}^{n} \sum_{i=1}^{c} d_{ik}^2 u_{ik}^2 ,$$

其中，$U=(u_{ik})_{c\times n}$ 为隶属度矩阵; $d_{ik}=\|x_k-v_i\|$，即样本 k 到聚类中心的距离; $0 \le u_{ik} \le 1$，$\sum_{i=1}^{c} u_{ik}=1$，u_{ik} 表示第 k 个样品 x_k 属于第 i 类的隶属度。依据最后的隶属度矩阵 U 中的元素的取值，可以确定所有样品的归属，当 $u_{jk}=\max\limits_{1\le j\le c}\{u_{ik}\}$ 时，可将样品 x_k 归于 c 类。筛选出各类中盈利能力表现最好的因子，以此达到冗余因子剔除的目的。

3. 综合评分模型的构建和选股

在选取去除冗余后的有效因子后，在模型检验期计算市场上正常交易的个股 i 的每个因子的最新得分，共 m 个持有期，按照等权重的方法获取 k 个因子的平均分，记为: $Z_i=\frac{1}{k}\sum_{k=1}^{K} Z(i, k)$ 。根据总的评分 Z_i，再将选取的所有股票进行排序，选取分数最高的总股票数的 $\frac{1}{5}$ 个股票进入投资组合。计算该组合在整个样本检验期的超额年化复合平均收益率，与 i 同时期的沪深 300 指数的年化复合平均收益率对比。

4. 模型的评价

模型评价的标准是检验模型有效性的重要指标，应从收益和风险角度考虑，实现较高的收益率、较低的风险水平，且无论在何种股价波动趋势下都能获得稳定的超额收益。可以作为评价模型收益率的指标有市场基准利率、投资组合收益率等。市场基准利率是衡量所构建投资组合业绩能力的一个重要参考，这个市场基准能够充分说明整个市场上股票价格波动的大体趋势，且具有投资性，一般被广大投资者作为衡量投资收益业绩的标准，主要包括深证指数、上证指数、沪深 300 指数等，本文选取沪深 300 指数，其选取的样本股覆盖了沪深两市六成左右的市值，具有良好的市场代表性。投资组合收益率是投资者构建的投资组合在投资期限内的投资收益体现，主要包括总收益率、年化收益率、相对收益率。总收益率是指整个投资时间内的收益水平，年化收益率是将收益细化到一年的收益水平，在此基础上与同期的无风险利率或市场基准利率作比较就能求得相对收益。

（三）模型估计与检验

1. 候选因子有效性检验

根据上述对候选因子有效性检验方法的论述，对所选取的 16 个因子进行检验，得到的结果如表 2 所示。

表 2　候选因子有效性检验结果

Tab. 2　Validity test results of candidate factors

因　子	年化复合收益/%	超额收益/%	分值与收益相关性	跑赢概率/%
账面市值比	2. 68	−4. 42	0. 46	59. 36
盈利收益率	10. 66	3. 56	0. 84	65. 63
市盈率	13. 01	5. 91	−0. 98	68. 75
总资产收益率	7. 17	0. 07	−0. 49	59. 36
市净率	11. 71	4. 61	−0. 83	59. 38
股息率	10. 48	3. 38	0. 92	62. 50
现金收益率	2. 81	−4. 29	0. 37	43. 75
净资产收益率	7. 87	0. 77	−0. 59	56. 25
营业收入净利率	6. 16	−0. 94	−0. 46	56. 26
资产负债率	5. 79	−1. 31	−0. 53	65. 63
固定资产比率	2. 99	−4. 11	−0. 64	53. 13
流通市值	25. 50	18. 40	−0. 99	65. 63
换手率	4. 02	−3. 08	−0. 88	62. 50
每股收益	15. 83	8. 73	−0. 81	62. 50
固定资产周转率	3. 48	−3. 62	0. 69	46. 86
股东权益周转率	1. 43	−5. 67	0. 15	53. 12

解释指标的相关含义：收益是评价投资股票回报的最好依据，是投资者最希望看到的结果，年化复合收益率就是年复利计算的收益率，它是将各因子下表现最好的股票组合在整个模型期的收益标准化，易于比较；超额收益是指超出同期沪深 300 指数的收益的那部分收益；分值相关性是指通过将股票按照因子大小排序后所分组合的收益率与分值的相关性，相关性在（−1，1）之间，小于零代表负相关，大于零表示正相关；跑赢概率是指计算每个持有期收益与当期沪深 300 指数收益的大小，计算大于沪深 300 指数收益的概率。表现最好的是流通市值这一指标，年化复合收益率为 25. 5%，而同期的沪深 300 指数的年化复合平均收益率为 7. 1%，且收益与分值的相关性达到 99%；表现最差的是股东权益周转率，分值与收益的相关性只有 15%，故其变化对股票价格波动的影响很小，不具有显著的代表性。尽管总资产收益率、营业收入净利率和资产负债率的年化复合收益率高于换手率等指标，然而其分值与收益的相关性都在 50%左右，因此这些指标的变动不能很好地说明收益的变动方向，故不能选为有效因子。将复合收益、分值与收益相关性、超额收益作为指标选取依据后，得到通过检验的有效因子有：盈利收益率、市盈率、市净率、股息率、净资产收益率、固定资产周转率、流通市值、换手率、每股收益，共 9 个有效因子。

2. 模糊 C 均值聚类剔除冗余

本文将 9 个有效因子聚类成 5 类，模糊 C 均值聚类算法的聚类原理是通过隶属度矩阵中每个样品在某类中的隶属度最大时归于某类，由于本文是由 16 个样品观测期组成的数据，因此所求隶属度矩阵是 16 个观测期隶属度矩阵的算术平均值，可能会造成部分隶属度区分度不明显的因子聚类出现一定的误差，因此处理方法为根据各时期隶属度矩阵所计算的聚类情况作为参考，当某个因子在不同类别中的隶属度相差不大时，可根据情况进行相应的分类处理以达到更好的分类效果。所求得隶属度矩阵如表 3 所示。

表 3 有效因子的隶属度矩阵

Tab. 3 Membership matrix of effective factors

有效因子	分类				
	第 1 类	第 2 类	第 3 类	第 4 类	第 5 类
盈利收益率	0. 246	0. 248	0. 182	0. 086	0. 238
市盈率	0. 125	0. 063	0. 375	0. 188	0. 250
市净率	0. 063	0. 188	0. 250	0. 312	0. 188
股息率	0. 246	0. 249	0. 181	0. 088	0. 236
净资产收益率	0. 247	0. 249	0. 183	0. 078	0. 242
流通市值	0. 212	0. 270	0. 185	0. 213	0. 119
换手率	0. 167	0. 222	0. 153	0. 249	0. 210
每股收益	0. 236	0. 244	0. 169	0. 136	0. 216
固定资产周转率	0. 428	0. 249	0. 007	0. 128	0. 188

根据上述隶属度矩阵，区分度最好的是市盈率、市净率、固定资产周转率，这三个因子在第 3 类、第 4 类、第 1 类中的隶属度都远远高于其他类别中的隶属度，且各个时期的聚类结果都相同，因此可先将这 3 个因子分为 3 类；而盈利收益率、股息率、净资产收益率、流通市值、换手率、每股收益这 6 个因子的区分度较小，在 2 类或 3 类中隶属度相差不大，故参考各时期的隶属度矩阵及其聚类结果，可将每股收益分到第 1 类，流通市值分到第 4 类，其他 4 个因子归为第 5 类。最终聚类结果为：第一类：每股收益、固定资产周转率；第二类：市盈率；第三类：市净率；第四类：流通市值；第五类：盈利收益率、净资产收益率、换手率、股息率。在保留盈利能力表现最好的因子后，剔除其余冗余因子，得到 5 个相似性较低的有效因子：市盈率、市净率、流通市值、净资产收益率以及每股收益。由于盈利收益率是由市盈率推导出来的，故在选择市盈率的条件下，对于盈利收益率将不予考虑，而每股收益与股息率分别在第一类和第五类中对股价影响最为显著。

3. 构建投资组合与选股

构建过程如下：

(1) 对该时期内每只正常交易的股票根据打分法进行打分，记因子 k 对股票 i 的打分为 Z_i^k；

(2) 按公式 $Z_i = \frac{1}{k}\sum_{k=1}^{n} Z_i^k$ (其中 n 为符合样本股票的个数) 得到股票 i 的综合评分 Z_i；

(3) 根据得到的股票 i 的综合评分对所有样本股票按照降序排列重新组合，提取最高得分的 20% 股票选入投资组合；

(4) 根据所选出的投资组合，在每季度等本金投资于该组合，计算其在持有期的收益率，同时计算同时期沪深 300 指数的收益率。

(5) 通过模拟历史数据得到了 5 个相似性较小的有效因子，为了检验模型的准确度，使用 2014 年第四季度—2016 年第四季度的数据作为模型检验期。检验期同样采用持有期为一季度计算构建投资组合。

根据上述方法计算出的收益率与沪深 300 指数的收益率如表 4 所示。

表 4　投资组合收益率与 HS300 指数收益率对比

Tab. 4　Comparison of returns between portfolio and HS300 index

组合	2014 年第四季度	2015 年第一季度	2015 年第二季度	2015 年第三季度	2015 年第四季度	2016 年第一季度	2016 年第二季度	2016 年第三季度
沪深 300	3. 32	0. 73	0. 49	−0. 74	0. 84	−0. 45	−0. 08	0. 13
投资组合	4. 18	1. 24	0. 88	1. 09	1. 60	1. 91	1. 37	1. 02

（四）模型结果分析

沪深 300 指数能很好地反应 A 股市场运行的整体状况，其成分股通常有良好的基本面，能够很好地解释在良好的经营条件下各因子与投资收益之间的影响关系，对于以基本面分析为主的多因子选股模型具有很强的适用性。同时沪深 300 指数成份股是不断更新变动的，因此为模型的持续改进提供了强有力的数据基础，并得到了广大投资者的认可。

在本文所选取的 16 个候选因子中，盈利收益率、市盈率、市净率、股息率等 9 个因子都有很好的市场表现。在剔除冗余后剩余 5 个相似性较低的因子：市盈率、市净率、流通市值、净资产收益率以及每股收益。市盈率是市场估值中最重要的参考因素，是对上市公司估值的参考性指标，对股价波动有良好的表现。从模型结果可以看出，市盈率与股价波动成负相关，即在一定范围内市盈率越低，股价表现越好，越容易引起股价上升，低市盈率也表明该只股票是被低估的，为投资者获利提供了可能；市净率是每股股价与每股净资产的比率，是资产的现在价值与账面价值的对比，一般来说企业经营业绩越好，净资产越高，在股价不高时相对的市净率越低，其投资价值越高，从结果来看市净率与投资收益成反比，得到了很好的验证；净资产收益率能在一定程度上反映公司的经营业绩，是投资者对公司发展评价的参考指标，模型结果也表明与收益具有很强的相关性，净资产收益率与收益成反比说明股票在该时期被低估，当市场发生波动时会推动股价的上升，从而为投资者获利提供可能；流通市值作为反映上市公司规模大小的最直观表现，是评价上市公司价值的依据，流通市值与收益率成负相关，流通市值越小，收益率越高，这从另一方面表明在沪深 300 指数成分股中，流通市值低的股票可能在一定时间内被低估，在未来的一段时间内会有所上升，因此投资者可以买入被低估的股票以获得投资收益；同样地，股息率是上市公司一段时期派息额与市价的比值，是衡量企业是否具有投资价值的重要尺度，股息率越高，该上市公司越被看好，越具有投资价值。

本文所建立的模型筛选出了 5 个有效因子，使用打分法，运用等权重赋权方法，构建综合评分选取的投资组合获得了良好的收益，跑赢同时期的沪深 300 指数收益，且获得了相对较多的超额收益，在熊市期间也能获取正的超额收益，因此验证了该模型的有效性。

三、结论与建议

多因子选股模型是量化投资选股模型中基本面分析最常见、应用最广泛的模型之一，投资者在投资实践过程中通过一系列数学公式变换处理并分析海量的上市公司历史数据，从中找寻被低估的股票。从以上模型检验可以得出：第一，部分基本面和财务数据因子与股价波动规律有较强的关联，为投资者获利提供了可能；第二，多因子选股模型应用于投资选股实践，能够取得稳定的相对于沪深 300 指数收益的超额收益，投资风险更小；第三，将模糊 C 均值聚类算法作为冗余因子的剔除的思路得到了较好的验证，拓展了多因子选股模型的理论方法。通过历史数据的模拟，可以确定多因子选股模型应用于股市是可行的，也为以后验证更多因子的有效性提供了参考，在经历无数次的验证与改进，更加符合市场行情波动。

本文在多因子选股中还存在不足之处。首先，因子的选取不够全面，可能导致具有更好市场表现的因子没有体现，对缺失数据的处理可能删除一些表现良好的股票；其次，在进行冗余因子剔除时，应不断地改进与验证新的方法，使之更加合理；最后，要想在市场中运用还需要不断地进行后期研究，比如投资策略选股数量化研究、股票择时研究等。

参考文献:

[1] FAMA E. Efficient capital markets: a review of theory and empirical work [J]. Journal of Finance, 1970, 25 (2): 383-417.

[2] ASNESS C S. The interaction of value and momentum strategies [J]. Financial Analysis Journal, 1997 (3): 29-36.

[3] CHEN N F, ZHANG F. Risk and return of value stocks [J]. Journal of Business, 1998, 71 (10): 501-535.

[4] PARTHA S, MOHANRAM P S. Separating winners from losers among low book-to-market stocks using financial statement analysis [J]. Review of Accounting Studies, 2005 (10): 133-170.

[5] 范龙振，余世典. 中国股票市场的三因子模型 [J]. 系统工程学报，2002 (6): 537-546.

[6] 汪洋. 基于估值与业绩的选股策略有效性研究 [D]. 成都: 电子科技大学，2010.

[7] 王艳萍，陈志平，陈玉娜. 多因子投资组合选择模型研究 [J]. 工程数学学报，2012，29 (6): 807-814.

[8] 丁鹏. 量化投资——策略与技术 [M]. 北京: 电子工业出版社，2012.

[9] 孙守坤. 基于沪深 300 的量化选股模型实证分析 [D]. 上海: 复旦大学，2013.

[10] 王瑞. A 股市场多因子量化选股研究 [D]. 济南: 山东财经大学，2016.

Research on Multiple-Factor Quantitative Stock Selection Strategy Based on CSI 300 Stocks

SU Jingyu, FANG Hongbin

(School of Economics, Anhui University, Hefei 230601, China)

Abstract: The CSI 300 constituent stocks are selected as the sample stocks and interception of 2007-2016 financial data and market data are used to build multiple-factor stock selection model through fuzzy C-means clustering algorithm eliminating the effective but redundancy factors. This paper makes a research based on the portfolio returns to test the effectiveness of the model. The results show that the p/e ratio, price-to-book ratio, earnings per share and other factors strongly correlate with the stock price fluctuation rule; the multiple-factor stock selection model is applied to the investment practice to obtain stable excess returns relative to the CSI 300 index; the combination of fuzzy C-means clustering algorithm and multiple-factor stock selection model is well verified, which provides a new idea for quantitative investment research.

Key words: multiple-factor stock selection model; fuzzy C-means clustering; portfolio returns; excess returns

(责任编辑: 杨成平)