改进k-means聚类在股价波动趋势上的应用

岑晓雪，秦江涛

（上海理工大学 管理学院，上海 200093）

摘要：股票价格在一段时间内的波动趋势对研究预测这支股票有着非常重要的作用。将K-means聚类利用层次聚类算法思想进行改进，对股价波动趋势进行聚类，并在一定范围内选择出最优聚类数。此方法在此问题上聚类效果优于原始k-means聚类算法，为未来对股价波动趋势的预测研究提供帮助。

关键词：股价波动趋势；K均值聚类；层次聚类

中图分类号：C931.1 **文献标志码**：A **文章编号**：1671－1807(2016)01－0144－05

2015年股市波动巨大，前所未有的大量投资者参与到股市中。投资者们必须要能选择可以长期投资的优良股票并且要在适当时机加入，适当时机退出，才可真正获利。由于精确预测股价不可能，所以股价在一段时间之内的波动趋势才是投资者们关注的焦点。本文中提到的股票波动趋势，指的是股票一段固定时间内波动情况，它有先上升后下降、先下降后上升等等所有可能的波动趋势，并不是股价在某一时点相较上一时点是上升还是下降。

道氏理论认为[1]，一旦股价变动形成一种趋势，便会持续相当长的时间，投资者可以顺应趋势找准自己的投资地位，当市场发出趋势转变的信号时，可立即作出合适的匹配动作避免损失。

现实生活中，大部分投资者由于时间与金钱的限制，只会进行中短期的投资操作。中短期内，股价波动趋势多变，却又有一定规律，可分出大致类别。对这种有规律性的股票波动状态进行研究或预测，对想要进行股票投资的投资者来说，是有意义的。

想要对一段时间内股价波动趋势进行下一步的研究，首先需要将股价波动走势进行聚类划分。文献[2]、[3]指出，在统计意义下，如果股市的基本波动趋势具有可识别性，个股不同时段波动趋势之间有相似性，我们就可以用聚类的方法对其进行聚类。对股价市场的聚类目前主要运用在划分股票板块，根据行业或指数等特征的不同而将股票分为不同的板块类别[4]。聚类分析方法能够快速有效的帮助股民准确客观的挑选出各板块中的绩优股、蓝筹股。

聚类分析是目前非常重要的数据挖掘方法之一，目前并没有一种通用的聚类方法能够对所有的实际问题进行最有效的聚类，主要有以下四种方法：

1)划分方法。首先确定分类个数，根据数据间的相似度将不同数据划分到不同的簇，循环迭代减少误差，直至每一个数据都恰在一个簇中。k-means聚类属于常用的划分聚类方法。

2)层次聚类方法。层次聚类方法则是将数据通过层次分解而进行聚类的一种方法。第一种为凝聚层次聚类。首先将所有数据都被各自视作一类，接下来根据计算出样本之间的距离或其他相似关系进而进行合并，直至所有数据被归于同一类中或直至满足某一条件。另一种为分裂层次聚类，先将所有样本都视做一类，用类似的相似关系将数据层层分裂，直到样本数据各自为一个簇或者满足某条件。

3)基于密度方法。此方法利用密度作为分割标准，密度相同则归于一类。

4)基于网格方法。该方法根据属性分割出若干区间创建网格。首先根据样本属性等的不同，分割出一个网格结构，在此基础之上进行聚类操作，数据集各数据根据属性的不同而归于不同的网格单元。

不同的实际问题需要选择不同的聚类方法，以下几个方面是选择聚类方法的主要标准：能否处理大数目或高维数据、适合的样本数据类型、算法速度与效率、对参数的依赖性、聚类可识别的形状、能否识别异常数据、样本顺序的影响程度等[5]。

本文根据实际股价波动趋势的数据特点与各算

收稿日期：2015－08－27

作者简介：岑晓雪(1992—)，女(苗族)，贵州都匀人，上海理工大学管理学院，管理科学与工程专业硕士研究生，研究方向：信息管理与决策支持系统；秦江涛，男，贵州毕节人，上海理工大学管理学院，副教授，研究方向：信息管理、先进制造系统等。

法的优缺点提出一种利用分裂层次聚类思想进行改进的 k-means 聚类方法[6]，利用自下而上的分裂层次聚类原理来获取 k-means 算法的初始聚类中心，使得不同类簇之间拥有层次关系。针对不同股票数据多个 10 日内波动情况聚类并得出聚类结果，并选择针对此实际问题最合适的聚类数。

1 改进 K-means 聚类

1.1 K-means 聚类算法

原始 K-means 聚类算法描述如下：

1)从所有样本数据中随机选择 K 个样本作为初始聚类中心；

2)计算每个样本到聚类中心的距离，将样本根据距离分别分配给最近的某个中心，得到不同的簇；

3)用每个类簇的均值点作为每个簇新的聚类中心；

4)重复执行 2)和 3)直至各个聚类中心不发生变化。

k-means 聚类算法拥有原理简单，效率高，伸缩性强的优势，对于例如股市指数等大规模数据来说非常适用。该算法的缺点在于，一开始就要主观判断最终聚类数，初始质心也只能随机选择，大大影响了聚类效果，只能得到局部最优。例如文献[7]利用 K-means 聚类就成功将大量时间间隔相同的股价波动趋势样本聚为 4 类。然而此文章实际数据类别的划分数量的主观判断与初始质心的随机选择直接影响了聚类效果。

1.2 层次聚类算法

层次聚类方法的优点在于，样本之间距离规则定义简单，不需要主观决定聚类数目，类与类之间的层次关系清晰；聚类形状不固定。而缺点在于，算法所需时间较多，复杂度高，不太能适应大数据集，并且每一次的合并或分裂都不能撤销，已经归入某一类的样本不能再归入另外的类中。

为解决层次聚类时间复杂度过高且合并或分裂不能撤销的缺陷，可利用 k-means 聚类先对样本进行分割，再对分割成的各个类簇进行下一步的分裂或者合并。

层次聚类算法最大优点在于算法思想简单易懂且合理，利用样本之间的距离或者其他数据间的相似性选择下一次进行聚类的样本。分裂层次聚类的原理正好可以为 k-means 聚类方法获得初始的质心，层层迭代从而得出质心间的层次关系。

1.3 改进 k-means 聚类算法

已有大量研究对 k-means 进行改进。比如胡伟[8]在研究中提出一种改进的层次 K-mean 聚类算法，此算法结合层次结构思想迭代执行聚类，利用聚类测度函数作为阈值判断是否结束聚类。

对于股价波动趋势聚类来说，胡伟的方法可以解决聚类数无法确定的问题，也可以在迭代中选择较为合适的聚类结果，然而，此方法并不适用于高维数据的分类，并且其利用特定阈值来作为迭代的结束条件，得到聚类数的数量之大并不适合股价波动趋势的进一步研究。

股价波动趋势需要利用多天的数据共同构成聚类的单个样本，因此，股价波动趋势的聚类事实上是高维数据的聚类，若要对股价波动趋势进行下一步的分析或预测，聚类数也应是在一定范围之内。在此基础上，本文将胡伟的方法再进行了改进，将 k-means 聚类与分裂层次聚类原理相结合，利用其思想为 k-means 聚类获取初始质心[9]，之后不断进行更细层次的 k-means 聚类[10]，不以样本之间距离或其他测度值作为聚类结束的标准，以股价波动趋势有可能的聚类数作为结束条件。

再次改进的 K-means 聚类算法计算步骤为：

1)首先选择包含 n 个数据对象的样本集 $A=\{A_1,A_2,A_3,\cdots,A_n\}$，每个样本有 a 个维度。设定初始聚类个数 K_1（初始聚类个数设为 2），聚类迭代次数 $times$ 初始化为 1，进行 k-means 聚类，聚类中心为 $cid=\{cid_1,cid_2,cid_3,\cdots cid_{k_1}\}$。

2)从 k-means 聚类结果中根据公式(1)分别计算每个簇的类半径，根据公式(2)选择类半径最大的一类。

$$dist(A_i^j,cid_i)=\sqrt{(A_i^{j1}-cid_i^1)^2+(A_i^{j2}-cid_i^2)^2+\cdots+(A_i^{ja}-cid_i^a)^2} \quad (1)$$

$$r_i=\max(dist(A_i^j,cid_i)) \quad (2)$$

其中，A_i^{ja} 指第 i 类中的第 j 个样本的第 a 维数，cid_i^a 指第 i 类类心的第 a 维数。

3)在此类中根据公式(1)选择距离类心最远的一个样本点 x，再得出距离 x 点最远的一点 y。

4)将此类中 x 和 y 这两个样本点与其他的聚类中心作为新的聚类中心，再次进行 K-means 聚类。

5)当聚类数目达到指定数目时，停止聚类。否则将继续执行步骤 2 与步骤 3，直至计算结束。

2 聚类过程与结果分析

为对聚类结果进行对比评价，本文选取万科 A、国中水务、世纪星源、国风塑业 4 支股票从 2015 年 7 月 3 日开始，再往前推 250 天每 10 天内的收盘价的波动趋势进行聚类。实验数据从大智慧平台获取，利

用 matlab7.0 进行计算。每支股票计算过程一致，因此下面只详细介绍万科 A 股票聚类情况。

为提高准确性，利用移动选取样本的方式，将 250 天内数据每隔 5 天选取 10 天的数据，因此，最后拥有 49 个样本，每个样本维数为 10。

由于每个样本数据大小不同，若直接进行聚类，结果会忽略股价波动的趋势而以数据大小进行分类。因此应该将每个样本集每个数据归一化之后再进行聚类计算。归一化公式为(3)：

$$X_i^j = (A_i^j - \min A_i)/(\max A_i - \min A_i) \qquad (3)$$

其中，X_i^j 指第 i 个样本第 j 维数的归一化值。A_i^j 指第 i 个样本第 j 维数。

实际计算时，将初始聚类数 K_1 设置为 2，由于股价波动趋势的研究需要在有限个聚类结果上进行，因此分别指定最终聚类数为 3 到 10 进行聚类。选择每个类簇的平均质心距离的加权平均值和类间距离的最小值来评价不同聚类数的聚类效果。平均质心距离加权平均值计算公式为：

$$\frac{\sum_{i=1}^{k}(||A_{ij} - cid_i|| \cdot n_i)}{n} \qquad (4)$$

其中，j=1,2,3,… n_i，A_{ij} 指第 i 类第 j 个样本，n_i 指第 i 类样本数目，n 指总样本数目。

选择每类质心与质心的距离作为类间距离，利用如同公式(1)的计算方法进行计算。

聚类数为 3 到 10 的平均质心距离的加权平均值变化曲线如图 1，类间距离最小值变化曲线为图 2。

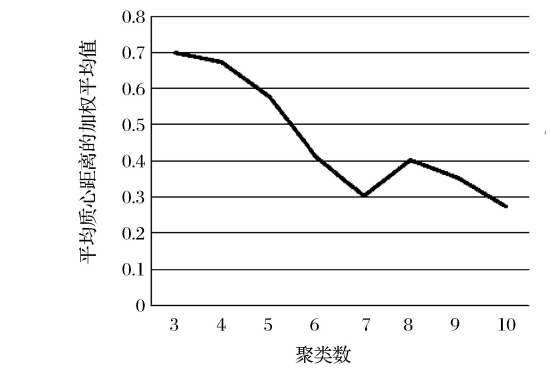

图 1　平均质心距离加权平均值变化曲线

从图 1 和图 2 上看，聚类数从 5 类到 6 类时，平均质心距离的加权平均值的下降速度是最快的，类间距离的最小值的上升速度也是最快的[11]，因此实验内最佳聚类数是 6(结果见图 3)。

从聚类结果上看，聚类成功将 10 天内股价波动走势分为大致如下六类：第一类为：先上升后下降再上升后又下降的反复震荡；第二类为先上升后下降；第三类为总体不断上升；第四类为先下降后上升再下降再小幅上升的反复震荡；第五类为先下降后上升；第六类为总体不断下降。

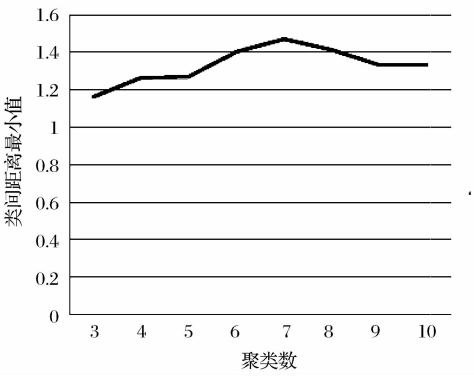

图 2　类间距离最小值变化曲线

用同样的方法对国中水务、世纪星源、国风塑业的股价波动趋势样本进行聚类，选择较佳聚类数。最后利用每个类簇的平均质心距离的加权平均值(P)和类间距离最小值(D)这两个评价指标评价原始算法与改进后算法的效果，结果见表 1。

表 1　原始算法与改进后算法聚类效果比较

样本集	实验内最佳聚类数	改进后聚类算法		原始 k-means 聚类算法	
		P	D	P	D
万科 A	6	0.412 0	1.400 3	0.493 1	1.373 4
国中水务	7	0.286 6	1.549 3	0.366 0	1.231 2
世纪星源	6	0.303 9	1.371 2	0.406 7	1.133 4
国风塑业	8	0.279 3	1.376 2	0.420 5	1.270 9

从表 1 上看，在同样的聚类数基础上，改进后算法的平均质心距离的加权平均值比原始算法要低，而类间距离最小值相较原始方法要更高。由此可见改进后算法聚类效果较好。

3　结论与展望

想要对股价波动趋势进行研究或预测，首先就要对其进行聚类分析。而本文提出的结合分裂层次聚类的改进后的 k-means 算法对股价波动趋势的聚类效果可能并不是最优的，但是却明显好于原始 k-means 算法。并且，本文提出能够针对本次研究的股价波动趋势而得出较优的聚类个数的方法。此方法未能摆脱一开始需要设定一个初始的聚类数的缺陷，但是将初始聚类数设置为 2 已经将聚类误差降到最小。此算法也未必适用于其他的研究问题。今后还要继续研究更简便且效率更高的聚类方法针对股价波动趋势进行聚类，以便于今后对股价波动趋势的预测研究。

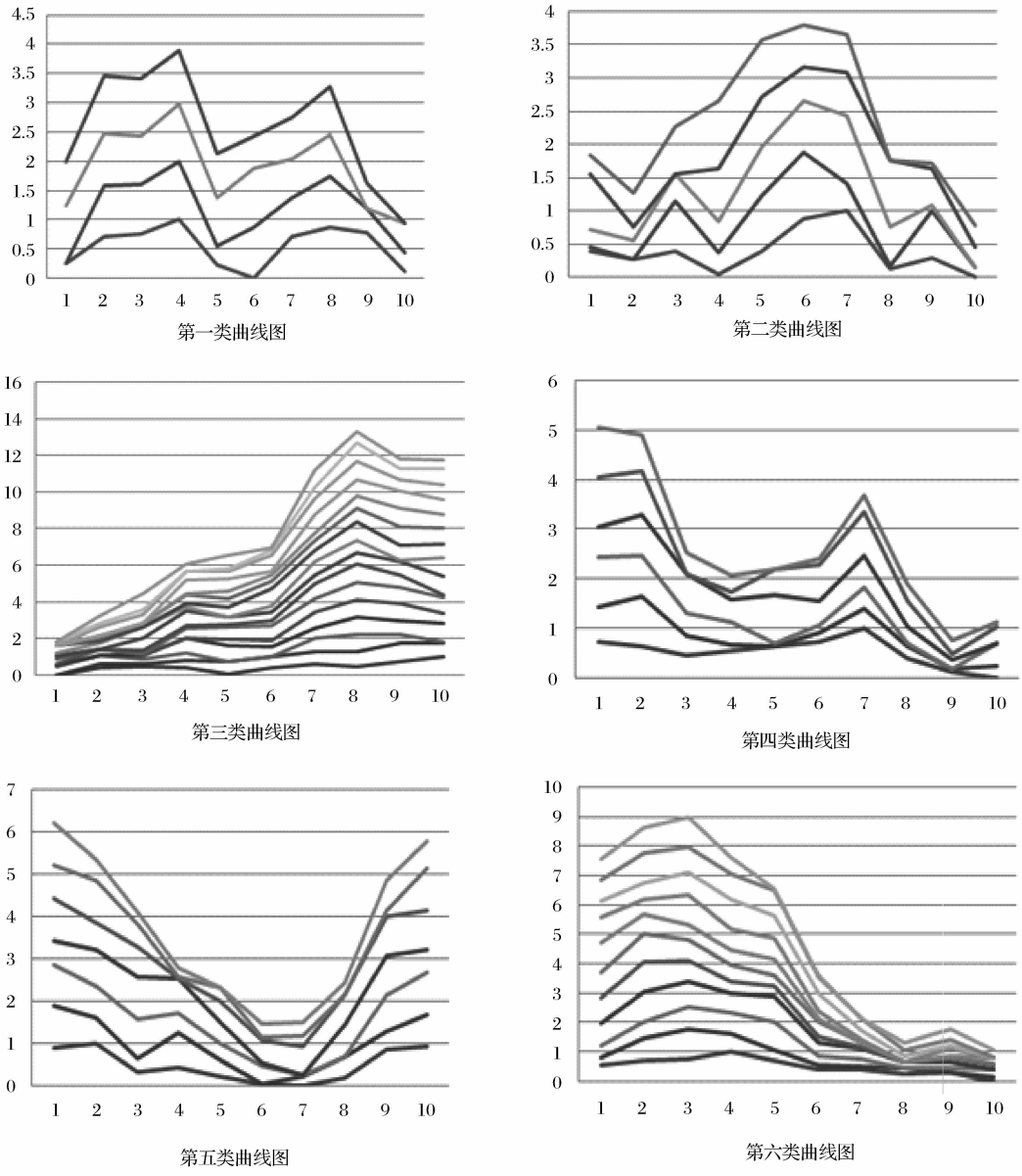

图 3　分类曲线图

参考文献

[1] 陈峰. 论股票指数的趋势分析[D]. 成都：西南财经大学，2007.

[2] 方国斌. 中国股市波动性聚类特征参数与非参数分析[J]. 技术经济，2007，26(10)：84－88.

[3] 张杨，宋恒. 基于聚类技术的股市基本趋势规律挖掘[J]. 世界经济情况，2006(10)：18－21.

[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(3)：94－96.

[5] 刘泉凤，陆蓓. 数据挖掘中聚类算法的比较研究[J]. 浙江水利水电专科学校学报，2005(2)：55－58.

[6] 罗晖霞，曲晓玲. 基于网络舆情的 K-Means 算法的改进研究[J]. 电脑开发与应用，2010(8)：4－6，15.

[7] 卢瑞瑞. 基于 K-means 聚类的马尔可夫过程在股价趋势预测中的应用[D]. 武汉：华中科技大学，2009.

[8] 胡伟. 改进的层次 K 均值聚类算法[J]. 计算机工程与应用，2013(2)：157－159.

[9] 何飞，蒋冬初. 基于向量空间模型的文档聚类算法研究[J]. 湖南城市学院学报，2003(3)：114－116.

[10] 郝洪星，朱玉全，陈耿，李米娜. 基于划分和层次的混合动态聚类算法[J]. 计算机应用研究，2011(1)：51－53.

[11] 周世兵，徐振源，唐旭清. 基于近邻传播算法的最佳聚类数确定方法比较研究[J]. 计算机科学，2011(2)：225－228.

Application of Improved K-means Clustering in Stock Price Fluctuation Trend

CEN Xiao-xue, QIN Jiang-tao

(College of Management, University of Shanghai for Science and Technology, Shanghai 200093, China)

Abstract: The fluctuation of stock price in a period of time is very significant to study and forecast the stock. In this paper, we use the Hierarchical clustering thought to improve K-means clustering, and to custer stock price fluctuation trend. Then select the optimal cluster number within a certain range. The clustering effect is better than that of the original K-means clustering algorithm, which will help to forecast the trend of stock price in the future.

Key words: stock price fluctuation trend; K-means; hierarchical clustering

(上接第 125 页)

The Relationship between the Aged Tendency of Population and the National Economic Development in New Normal

HE Min, ZHOU Zhao-xiong

(The University of Shanghai for Science and Technology, Shanghai 200093, China)

Abstract: The aged tendency of population has been global trends from 21 century. The aged tendency is the result of all kinds of development. Someone thinks that the result has negative effects, but as the policy of new normal, the aged tendency of population is a challenge, also ,it can be a chance. Taking advantage of the aged tendency of population, it can be well developed with economy.

Key words: new normal; the aged tendency of population; the development of economic

(上接第 130 页)

[9] 韩芳. 旅行社拓展在线市场的策略[J]. 中国电子商务, 2014 (15): 176－178.

[10] 伏六明. 论旅行社相关多元化经营策略[J]. 商业时代, 2005(6): 70－71.

The Choice of Business Strategy of Travel Agency Based on Probability Analysis Frame

FENG Zhao-sheng, GAO Ya-fang

(Tourism College, Northwest Normal Univetsity, Lanzhou 730070, China)

Abstract: With the continuous development of China market economy, the per capita net income of rural residents and the disposable income of urban residents achieve a substantial increase. More and more Chinese people prefer to spending on travel, however, the travel agency industry profits does not increase with the travel income increase, or even the profits begins to slow down. In order to make the travel industry profits of China to achieve a better development, this paper adopts the quantitative method to analysis the degree of interaction between the travel agency business strategy to adjust and modify the travel agency business strategy combination.

Key words: travel agency; business strategy; adjustment; SCP model