Hans Journal of Data Mining 数据挖掘, 2020, 10(3), 209-220
Published Online July 2020 in Hans. http://www.hanspub.org/journal/hjdm
https://doi.org/10.12677/hjdm.2020.103022

Hans汉斯

An Analysis of Plate Linkage Based on Chinese A-Share Market

Hao Yin

University of International Business and Economics, Beijing
Email: 646312620@qq.com

Received: Jun. 3rd, 2020; accepted: Jun. 18th, 2020; published: Jun. 28th, 2020

Abstract

The stock market is a very important part of current Chinese economic development. It plays a very crucial role not only in the real economy market, but also in the fictitious economy market. We use K-means to divide stock into different blocks and try to find the relevance of the block by using association rules.

Keywords

Stock, K-means, Association Rules, Relevance of the Block

我国股票市场的实证性板块联动分析

殷　浩

对外经济贸易大学，北京
Email: 646312620@qq.com

收稿日期：2020年6月3日；录用日期：2020年6月18日；发布日期：2020年6月28日

摘　要

股票市场在目前中国社会的经济发展中，是一个相当重要且不可分割的组成部分。不仅是在实体经济市场当中，而且在虚拟经济市场的方面，股票都发挥及充当了相当重要的作用和角色。本文使用K-means等聚类分析算法将股票分为不同的板块，然后使用关联分析算法，对不同的板块作关联分析，发现板块间蕴含的联动关系。

文章引用: 殷浩. 我国股票市场的实证性板块联动分析[J]. 数据挖掘, 2020, 10(3): 209-220.
DOI: 10.12677/hjdm.2020.103022

关键词

股票，K-means聚类，关联分析，板块联动

1. 问题背景

随着中国市场经济的快速发展，以及人们愈来愈强的金融意识以及投资意识，股票市场，作为市场经济组成的重要成分，也正渐渐地表现出它的成熟性和规范性。越来越多的投资者将目光和精力放在股票市场上。时间和经验已经表明，股票不仅在过去把可观的长期利益提供给投资者，而且在将来也会把良好的机遇提供给投资媒体。由于股市的变化不可测度，投资者以在股市投资中博取优厚的回报为核心目标，就需要以严肃的投资态度对待股票：研究上市公司的历史、业绩和发展前景，仔细分析上市公司的财务状况，秉持以基本分析、技术分析方法结合的投资理念，发掘真正存在投资价值的股票，对其进行合理的投资。

股票投资有宏观分析，中观分析，微观分析这三大种基本分析方式。宏观分析的对象是整个国家，分析我国的国民经济，实时政策。中观分析的对象是各个行业或者不同的地区。微观分析的对象是公司，分析公司的运营状况，财务数据等等。本文认为板块分析属于中观分析和微观分析的结合。

中国的股市白手起家，发展到如今已经具备了一定的规模，正在经历着从无到有，从有到优，从优到精的发展状态。在中国股市刚开始时，市场的规模比较小，上市公司的数量也不多，而且中国当时股民的投资思维和操作方法太过稚嫩，因此，投机的性质非常得强，板块分析的方法在那时候很少采取。但是，随着上市公司数量的不断增多，股市的发展，以及投资和操作手法的成熟，以前的投资理念和手法都渐渐失去了效用。面对整个 A 股市场上的 3000 多支股票，应该怎样选择？有些投资者不分青红皂白乱买一气，甚至被不同的股评、舆论、谣言所影响，不能理性地去投资，以致难以取得投资的成功。因此，在现在成熟的股市中，投资者如果想要成功，就要学会理性操作，学会树立板块投资的理念，学会板块分析。

板块是根据股票在某些指标上的共性来划分的股票集合。在股票市场中，可以根据所选指标的不同来划分各式各样的板块，可以从产业、行业、概念、地域、特殊题材等各个方面来划分。每一个划分的板块中都包含了几十甚至上百种股票。很多国内外的文献都是从理论方面介绍了板块联动的成因及其影响，对板块轮动的规律做了一定程度上的总结[1]，但并没有对整个股票市场上的数据做具体分析，没有用实际的结果去验证板块轮动的规律特点，及其结论的准确性。

市场上有大量的公司的股票，每个公司又有大量的财务数据，如何才能获取并且处理这些数据，将其变为所需要的数据格式，怎样去处理数据的不规整，去除数据的干扰内容，从而进行良好的分析？怎样才能发掘不同板块之间的联动关系呢？笔者采用了数据挖掘中的聚类分析和关联分析的方法，对上述问题作了一些研究及探讨。

2. 相关文献的研究现状

杜伟锦，何桃富(2005) [2]，文章的研究对象是上海证券市场上的各个板块，分析了 A 股和 B 股市场

上的 5 类传统指数，探究这 5 类指数之间的关联性。在验证了 B 股市场是独立的条件下，在 B 股市场中随机抽样了 24 个分类板块指数进行研究，采用方法为模糊聚类方法，最后的实证分析结果可以对投资操作提出具有参考意义的建议。

张建林，周超良(2013) [3]，使用关联规则挖掘股票市场中板块联动的关系，在使用过程中对关联规则算法进行了一些改进，在改进后的算法中，数据格式使用垂直格式，并且对产生候选项的连接方法也进行了改进。最后的实验结果表明，改进后的关联规则算法能够提升发现板块之间联系的处理速度。

冯甜(2014) [4]，作者认为股票市场中，各个板块之间的收益率在波动上具有显著的联动效应，一个或者数个大的板块带动的。文章同样利用关联规则的算法，分析内容涵盖了中国股票市场上的 30 个行业板块。

梁烨(2014) [5]，作者对股票指数之间的相关性进行距离测度，采用多维标量的方法，分析了各个行业板块之间的股指日收益率，分析行业数为 32，研究发现我国同类行业股指的收益率近年来相互之间的共同趋势越来越明显，而且随着时间加剧了某些行业股指间的整合，不同类行业股指的收益率的差异越来越明显。通过多维度的方法探讨我国 A 股市场近期 5 年的发展特点，发现了各个板块之间的联动规律。

3. 股票的板块聚类

3.1. 股票板块的定义

股票的板块指的是股票的一种集合，集合的元素是股票，在同一集合中的股票由于在某些维度上有较高的相似性，从而被人们归类在一起，而这些维度往往会被股民们所说的庄家用来进行炒作。股票板块的特色各式各样，有按概念划分的板块，如“锂电池板块”；有按地域划分的板块，如“山东板块”、“江苏板块”；有按行业分类的板块，如“电力板块”、“军工板块”、“房地产板块”、“银行板块”，有按上市公司经营状况划分的板块，如“购并板块”，总的来说划分板块的条件各式各样，只要这些板块能够成为股市炒作的题材。

3.2. 聚类

3.2.1. 聚类的定义

聚类是一个过程，这个过程将数据对象的集合分成相似的对象的类的集合[6] [7]。使得在某些维度上，同一个簇(或类)中的对象之间存在较大的共性，而不同类中的对象存在较大的差异性。例如，若有聚类集合{狗，鸡，猫，苹果，葡萄}，则根据动物和植物的概念，可以将其分为{狗，鸡，猫}和{苹果，葡萄}两个聚类。分类和聚类是完全不同的两个概念，它们具有显著的差异。对于分类来说，操作的对象的类别是已知的，笔者所要做的是如何对这些不同的类别进行处理。而在聚类中，操作的对象的类别是未知的，需要在某些维度上找寻它们的相似性，从而划分类别，因此聚类的难度要高于分类。

3.2.2. 相似性测度

对象之间的相似性是聚类分析的核心。对于不同的聚类应用，其相似度的定义方式是不同的。通常，各个对象之间的距离越小表示它们特征越相似。密度相似性度量：密度是指在单位区域内的对象个数。除了上面提到的相似性度量外，还有其他相似性度量，如相似系数。

3.2.3. 距离度量方法

若每个对象用 m 个属性来描述，即对象使用欧几里得距离表示为：

$$dist\left(O_i,O_j\right)=\left\|O_i,O_j\right\|=\sqrt{\sum_{k=1}^{m}\left(O_{ik}-O_{jk}\right)^2}$$

图 1 的具体含义是：距离的计算方法在二维图像上(即 $m=2$ 时)是如何逼近聚类中心的，图以一环套一环的同心圆的方式来逼近聚类中心。在本文的算法中，采用上述距离计算的方式，以同心圆的方式来逼近聚类中心。

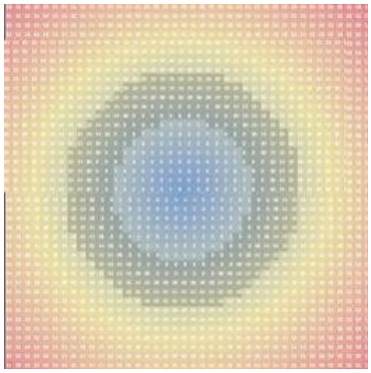

Figure 1. Euclidean distance
图 1. 欧几里得距离二维图像

3.2.4. 聚类过程

典型的聚类过程如图 2 所示。其中各部分的说明如下。

数据准备：为聚类分析准备数据，包括数据的预处理。

属性选择：从最初的属性中选择最有效的属性用于聚类分析。

属性提取：通过对所选属性进行转换形成更有代表性的属性。

聚类：采用某种聚类算法对数据进行聚类或分组。

结果评估：对聚类算法生成的结果进行评估。

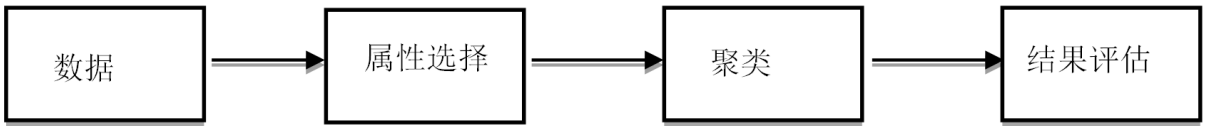

Figure 2. Clustering process
图 2. 聚类过程

4. K-means 算法

4.1. 算法流程

1) k 值作为聚类模型的输入参数，主要用来人为指定最终聚类模型的最终类数，即模型预设的希望分类的个数。

2) 根据输入参数的 k 值，随机在样本空间 S 中指定 k 个样本，并将这 k 个样本作为簇的质心，并设第 j 个质心为 C_{pj}。接下来每个质心周围的点将与此质心之间作计算。每个簇的质心代表了一个簇。这样得到的簇的质心集合为：

$$\text{Centroid}=\left\{C_{p1},C_{p2},\cdots,C_{pk}\right\}$$

3) 遍历样本空间 S 中的所有样本，并设第 i 个样本为 o_i，依次计算每个样本到各个质心中的距离，其中距离的计算方法采用欧几里得距离计算方法。然后依据距离最小原则，将样本进行簇归类，即把每一个样本归类到距离此样本最近的质心所在的簇。

4) 由过程 3 可得到样本初步归类完成的簇，现再对每一个簇内的所有样本进行计算，将簇内样本均值作为新的簇质心。设$|W_x|$为第 x 个簇W_x的样本数，设 n_x 为每一个簇对应的质心，即：

$$n_x = \frac{\sum_{o \in W_x} o}{|W_x|}$$

5) 如果这样划分的结果满足目标函数的要求，就说明了聚类已经达到事先要求的结果，终止算法。如果不满足要求，则需要迭代 3~5 这三个步骤，直到达到准则函数的要求。

4.2. 准则函数

准则函数主要以簇内样本与簇中心的总误差平方和为考量依据，以总误差平方和最小作为原则，通过不断的进行簇中心的更新和样本空间内所有的样本的簇再归类，以实现准则函数的收敛，并将这种准则函数的收敛作为模型终止条件。其中准则函数，即 SSE (sum of the squared error)，其定义如下：

$$\mathrm{SSE} = \sum_{i=1}^{k} \sum_{p \in C_i} \|p - m_i\|^2$$

基于此准则函数，最终训练得到的模型会在很大程度上保证各簇内的样本之间的紧凑性，以及各簇间样本的独立性。判断函数收敛的依据：1) 样本进行再归类时，样本所属簇不再变化。2) 簇中心更新前后的距离小于某一阈值，其中阈值的值是预先设定的。

5. 关联规则算法

5.1. 引论

关联规则作为一种常用的数据关联关系的挖掘方法，在很多领域均有较为明显的应用优势，比如说歌曲推荐，社交关联等。作为一种比较经典的推荐算法，关联规则主要能通过历史数据从概率的角度发掘出样本与特征之间的潜在关系，从而实现诸如顾客购买偏好等特征的量化[8]。

值得注意的是，关联规则的学习器是无监督学习，因此不需要对训练的数据打标签，这对数据预处理环节提出了较为宽松的要求，但因此也有一些对应的不足之处，即训练得到的模型缺乏科学有效的评估方式，普遍需要通过对模型结果进行人工观测以判断模型的合理性。

5.1.1. 例子——源数据

本例子所分析的数据为阅读网站的点击流数据，主要思想就是通过关联规则对不同类型的文章类型之间通过点击流数据的分析，以得到各文章类型间的关联关系。不同的访问者访问的版块，如表 1 所示：

Table 1. The sample clickstream data for the site
表 1. 阅读网站的点击流数据样例

访问 ID	List of media categories accessed
1	{新闻，金融}
2	{新闻，金融}
3	{体育，金融，新闻}
4	{艺术}
5	{体育，新闻，金融}
6	{新闻，艺术，环境}

5.1.2. 数据格式

关联规则是实现需要一定数据预处理，在本例中主要就是将原始数据进行稀疏矩阵化处理，具体规则如下：行名为文章类型，列名为序号，当序号为 i 的访问者访问过文章类型为 j 时，则设置稀疏矩阵的 (i,j)的值为 1，否则为 0。如表 2 所示：

Table 2. Source data after sparse matrix processing
表 2. 稀疏矩阵化处理后的源数据

访问 ID	新闻	金融	环境	体育
1	1	1	0	0
2	1	1	0	0
3	1	1	0	1
4	0	0	0	0
5	1	1	0	1
6	1	0	1	0

5.2. 术语和度量

5.2.1. 项集板块集

项集用来描述以文章类型作为元素的集合，例如{新闻，金融}=>{体育}，其中{新闻，金融}是一个项集实例，{体育}也是一个项集实例。

=>符号作为规则描述了两个项集之间的关联关系，即说明同时看过新闻类和金融类文章的访问者很有可能会看体育类文章。

=>符号左侧的项集被称为左项集 LHS，即 Left-hand-side，右侧的项集被称为右项集 RHS，即 Right-hand-side。

在基于关联规则的算法中，各项集间关联强度的量化主要依靠以下四个指标。

5.2.2. 支持度 Support

项集 i 的支持度的计算方式：

$$\text{项集的支持度} = \frac{\text{项集的出现次数}}{\text{总的记录数(交易数)}}$$

$$\text{Support}\left(\{\text{新闻}\}\right) = 5/6 = 0.83$$

$$\text{Support}\left(\{\text{新闻}, \text{金融}\}\right) = 4/6 = 0.67$$

$$\text{Support}\left(\{\text{体育}\}\right) = 2/6 = 0.33$$

支持度的概率意义在于对项集 i 的出现频次的度量，在进行关联规则的挖掘时，算法更倾向于优先对高频次项集的关联关系进行探索。

5.2.3. 置信度 Confidence

关联规则 $X \to Y$ 的置信度计算公式：

$$\text{confidence}\left(X \to Y\right) = \frac{\text{Support}\left(X \cup Y\right)}{\text{Support}\left(X\right)}$$

置信度的意义在于基于历史数据的量化了两个项集之间的关联强度，即通过概率学的方式将项集 X 和项集 Y 的关联强度用项集 X 出现时项集 Y 发生的概率来表示。

$$\text{Confidence}\left(\{\text{News},\text{Finance}\}\to\{\text{Sports}\}\right)=\frac{\text{Supports}\left(\{\text{新闻},\text{金融},\text{体育}\}\right)}{\text{Supports}\left(\{\text{新闻},\text{金融}\}\right)}=\frac{2/6}{4/6}=0.5$$

表示 50%的人访问过{新闻，金融}，同时也会访问{体育}。

5.2.4. 提升度 Lift

单纯的依靠支持度和置信度来进行关联强度的量化时，会有一定的缺陷。原因在于支持度高的项集往往在置信度上也表现优异，即对于高频次的项集 i，在任意项集出现的前提下，项集 i 发生的概率都会相对较高，这将破坏算法的平衡性和可靠性。

所以要引进 Lift 这个概念：

$$\text{Lift}\left(X\to Y\right)=\frac{\text{Support}\left(X\cup Y\right)}{\text{Support}\left(X\right)*\text{Support}\left(Y\right)}$$

例：

$$\text{Lift}\left(\{\text{新闻},\text{金融}\}\to\{\text{体育}\}\right)=\frac{\text{Support}\left(X\cup Y\right)}{\text{Support}\left(X\right)*\text{Support}\left(Y\right)}=\frac{0.333}{0.667*0.33}=1.5$$

对关联规则失效的情况进行概率上的统计，进一步完善对关联算法的评价指标体系。项集中含有 X 而不含有 Y 的概率为：

$$\text{Conviction}\left(X\to Y\right)=\frac{1-\text{Support}\left(Y\right)}{1-\text{Confidence}\left(X\to Y\right)}$$

5.3. 生成规则

一般两步：

第一步，高支持度项集的提取。遍历所有项集并分别计算其支持度，然后通过人为设置的阈值，对高支持度项集进行筛选。n 个板块，可以产生 2^n-1 个项集(板块集)。

第二步，高置信度规则的提取，针对步骤一筛选所得的高支持度项集，遍历的进行对应规则的置信度的计算，然后通过人为设置的阈值，对高置信度规则进行筛选。n 个板块，总共可以产生 $3^n-2^{n+1}+1$ 条规则。

从计算复杂度的角度考虑，步骤 2 的计算复杂度更高，步骤 1 相比于步骤 2，然后实际进行算法运算时，由于步骤 1 的阈值筛选掉了大量的低支持度项集，因此实际计算量步骤 1 更大。

5.4. Apriori 算法

Apriori Principle：1) 项集的频繁性是可遗传的。当项集 X 是高支持度项集时，此项集的子集也会是高支持度的。2) 项集的不频繁性是可追溯的。当项集 Y 是低支持度项集时，此项集的父集也会是低支持度的。例子：$\{X,Y\}$ 是频繁的，那么 $\{X\}$，$\{Y\}$ 也是频繁的。如果 $\{Z\}$ 是不频繁的，那么 $\{Y,Z\}$，$\{X,Y,Z\}$，$\{X,Z\}$ 都是不频繁的。

生成频繁项集的算法流程：

1) 通过预设的支持度阈值 sup，计算得到所有的高支持度项集。

2) 通过预设的支持度阈值 sup 过滤掉所有支持度低于 sup 的单一元素项集。

3) 基于步骤 1 过滤得到的单一元素高支持度项集，通过合并得到双元素项集，并计算支持度进行低支持度项集的过滤。

4) 基于步骤 1 和步骤 2 过滤得到的高支持度项集，通过合并得到三元素项集，并再次计算支持度进行低支持度项集的过滤。如表 3：

Table 3. Support data between different sections
表 3. 不同版块之间的支持度数据

One-Item Sets	Support Count	Support
{新闻}	5	0.83
{金融}	4	0.67
{环境}	1	0.17
{体育}	2	0.33
Two-Item Sets	**Support Count**	**Support**
{新闻，金融}	4	0.67
{新闻，体育}	2	0.33
{金融，体育}	2	0.33
Three-Item Sets	**Support Count**	**Support**
{新闻，金融，体育}	2	0.33

规则生成：1) 基于已得的高支持度项集，分别进行规则的置信度计算，并通过预设的置信度阈值过滤低置信度的规则，由于置信度的计算是基于已知的支持度，因此此过程计算量不高。2) 基于已得的高支持度项集和高置信度规则，进行规则的提升度计算，并通过预设的提升度阈值过滤低提升度的规则，以实现对规则的进一步完善。由于置信度的计算是基于已知的支持度和置信度，因此此过程计算量也不高。

6. 资料收集及基本思想

由于本文所研究的股票之间存在不同程度的相似性，衡量相似性的指标是股票之间的距离衡量。通过计算寻找统计量，用以衡量股票之间的相似度，作为划分类别的依据。不同的股票在不同的时间维度上存在不同的相似性，把一些相似程度大的维度的股票聚为一类，把另一些维度相似较大的股票聚为一类，直到把所有的股票聚合完毕[9]。本文选取了股票的交易日时间上的收益率维度，将两年中每只股票共 488 天的交易日的收益率作为划分类别的维度，每一个交易日的收益率都是一个维度，用 K-means 算法计算沪深市场上的股票在这 488 个不同维度上的相似性，将存在共同相似维度的股票划分为一类，将其作为一个板块。

获取数据的方式是从预测者网站下载了 2015 年到 2016 年的股票数据，包括了 488 天的股票数据，字段有股票代码，日期，开盘价，收盘价，当日最低价，当日最高价，涨跌幅等。本文只选用了涨跌幅这一字段，提取了每只股票的股票代码和涨跌幅，这样每只股票会拥有 488 个日期的涨跌幅，使用 R 语言算出每一只股票与其他股票的距离，即任意两只股票的 488 维欧几里得距离，根据不同的距离值选取聚类中心，对所有股票进行聚类，把具有近似距离值的股票归为一类，这样就能将涨跌幅近似的股票分为一类，即股票波动趋势近似的股票归为同一类。这样就实现了将波动作为特征而分类的目的，接下来将分成的类别作为一个板块，聚类的个数即为板块的个数。

读取本地沪深 300 的数据。提取沪深 300 的日期数据，代表了开盘的日期，结果显示两年内共有 488 天为开盘日期。并且读取本地所有股票的代码文件。处理股票代码，将.csv 后缀清洗掉。建立一个矩阵用来存放每只股票的开盘日的 change 值，若当日没开盘则记为缺失值 NA，矩阵的行表示所有 A 股市场上的股票，列则表示每个开盘日的 change 值。最后将建立的矩阵的部分值展示出来。对建立的矩阵 a 作处理，改变数据类型为数据框，并重命名其行列名，以便观察和研究。展示其重命名后的部分值。对数据框进行 K-means 聚类，选取聚类中心为 50，展示聚类结果。整理聚类结果，使同一类的结果排列在一起，并且将结果存入文本文件中。

整个 A 股市场的股票在 3000 支左右，在这 3000 支股票的数据中显示：在 2015 至 2016 年这两年的开盘时间中，有半数左右的股票停盘时间在 50 天以上。为了保证聚类结果集的准确性，在聚类的过程中，只选取了其中停盘时间不超过 50 天的股票作研究，这样能提高聚类结果的准去性及说服力。又由于从预测者网站下载的数据的股票名称都是用股票代码代替的，如浦发银行用 sh60000 来代替，虽然这种情况在作聚类时并不影响聚类的结果，但是使得聚类后的结果不够直观，因为每一类里都是股票代码。所以在处理聚类结果时笔者根据东方财富网的股票代码查询一览表将代码替换为了股票名称。处理的方式是首先使用了 Python 爬虫爬取东方财富网的股票代码查询一览表的网页，将其按照“股票代码，股票名称”的格式保存为 csv 文件。然后使用 R 语言读取此 csv 文件，一一对照聚类结果与 csv 文件，将股票代码替换为股票名称。

以聚类分析的结果划分板块，每一个板块说明了这类股票在波动上的相似性。接着再对分类的板块做收益率上的关联分析，分析这些板块之间在收益率上的联动性，得出板块之间的联动性规律及特性。从划分的 50 个聚类得到的每个板块中选取市场份额最大的股票作为这个板块的龙头股，从下载的数据中提取这 50 个龙头股的收益率信息，在一定程度上每个板块的龙头股代表了这个板块的整体信息，将其放在一个新的数据框中，作为接下来研究的对象。本部分采取了数据挖掘中关联分析这一方法，目的在于发现在同一交易日内涨跌幅状况同时为涨的板块之间的内在联系。关联分析是指关联规则挖掘，属于数据挖掘中一个重要且高度活跃的分支，其目标就是发现事务数据库中不同的项(如顾客购买的商品项)之间的联系，这些联系可以帮助用户找出某些行为特征(如顾客购买行为模式)，以便进行企业决策。那么，在股票板块研究中，就可以转化这个事例，股票的板块就相当于购物篮分析中的商品，某个板块的涨或跌就可以看作有或没有购买某个商品，每一个交易日可以看作顾客的每一次购买商品。研究的目的就变成了发现股票市场中不同的板块之间的联系，比如通过关联分析发现在大多数的交易日中收益率经常同时为涨的板块。可能是两个板块之间存在着关联关系，也可能是多个板块之间，它们之间互有关联，在某段交易时间，存在关联关系的板块集中大多数板块有上涨的趋势那么也预示着剩下的板块也很有上涨的可能，这个可能性要用关联分析中的指标如：支持度，置信度等去衡量。

关联规则算法能做到上述的分析与计算，并生成板块之间关联关系的规则。在接下来第七部分的内容中将展示部分分析与计算的结果，以及生成的关联规则。

7. 算法评价

7.1. 算法优点

本文的聚类结果分为 50 类，为了便于解释说明，笔者将 50 类板块分别命名为 T1~T50。从分类的结果上看，如 T1 板块含有的股票：浦发银行华夏银行民生银行上港集团中国石化招商银行上汽集团金地集团贵州茅台大秦铁路南京银行中国神华兴业银行北京银行农业银行中国平安交通银行工商银行中国太保中国人寿光大银行中国石油建设银行中国银行中信银行河钢股份本钢板材华东医药宁波银行。T1 结果展示出 T1 板块中的元素基本为银行股票，其中还含有一些大型国家企业，由此看来聚类的过程分析到了行

业因素。又如 T7 板块含有的股票：包钢股份北方稀土厦门钨业北矿科技金钼股份中色股份*ST 五稀中科三环章源钨业江粉磁材。T7 结果展示出 T7 板块中的股票基本为化学物质元素，说明聚类的过程分析到了物质类别元素。还有聚类算法分析到的很多方面这里就不一一列举了。总的来说聚类的结果比较符合期望的结果。

聚类算法不仅能够挖掘出这些股票之间为大家所知的类别关系，还能发掘股票时间隐藏的关系，这种隐藏的关系通常不能由人为的去发现，但算法能从数据上寻求到这种关系，这是聚类算法上的优点。

在寻求板块联动关系的过程中，关联规则起到了关键性的作用，由于采用了支持度，置信度等概率的衡量方法及展示方法，使得各个板块的关联性一目了然。而且关联规则能够快速准确的发现各个板块之间的规则，为研究提供了有力的支持。

7.2. 算法改进

本文使用的算法是 k 均值算法，“均值”二字体现了算法计算中心点的方法，即用求平均值的方法计算新的中心点，但这样做会存在一个问题，如果在计算中心店时有一个离群点，这会导致计算出的中心点远离大多数数据点，对聚类的准确性会造成不良的影响。解决此问题的方法为：使用 k-medoids 方法。k-medoids 采用的不是均值方法计算新的中心点，而是在数据集中寻求一个存在的对象，且这个对象必须满足在其所属的这一类里，是距其他对象距离和最小的对象，把这个对象作为新的中心点。以上做法便会改进 k 均值聚类离群点对聚类影响过大的问题。

本文在处理板块联动问题时，是将各个板块的龙头股的数据选出作为每个板块的代表以供接下来关联分析的数据。这样会简化运算的过程，提高处理计算的效率，但是如果将每个版块里所有股票的收益率按市场价值的权重求取平均值，作为一个新的指标，然后将这个新指标传递给关联分析，作为分析的数据会较好的提高模型的精确性，因为这考虑到了整个板块里所有股票的数据，而用龙头股代替就会失去这种面面俱到的效果。

8. 代码说明及结果

8.1. 关联规则的包

本文的关联分析方法主要基于 R 语言所带的 arules 模块实现的，其中模块的加载通过使用 R 语言命令 library(arules)。

8.2. 加载数据集

源数据为 StockType 数据集，每一行代表一天的不同板块的涨跌幅。数据转换：创建稀疏矩阵，每个板块一列，每一行代表一个交易日期。1 表示该交易日期中板块涨跌情况为涨的板块，0 表示没有涨。当然，在 R 语言里数据框是比较直观的一种数据结构，但是一旦板块比较多的时候，这个数据框的大多数单元格的值为 0，会占用大量内存。因为 R 语言处理数据时会将数据存入内存中，内存占用过高会引起计算机运行变慢，所以，R 引入稀疏矩阵，只去储存 1，这样会节省内存。函数 read.transactions 可以将源数据转换为稀疏矩阵。命令为：groceries=read.transactions("StockType.csv",format="basket",sep=",")。

接下来，使用 summary(groceries)查看数据集相关的摘要信息，以及数据集本身。summary 的信息包括三部分，第一部分：总共有 486 条每日的板块涨跌情况，50 个不同的板块。density = 0.4949794 表示在稀疏矩阵中 1 的百分比。第二部分：出现次数最高的板块名称及其次数统计信息。第三部分：涨跌情况为涨时包含的板块数目，及其对应的几个常见基本统计量，包括了最大最小值，三种常见的分位数及平均值的统计信息。如：仅 6 天包含了 T2 板块涨跌幅为涨的情况，仅 9 天包含了 T3 板块涨跌幅为涨的

情况。那段统计信息的含义是：第一分位数是 14，意味着 25%的日期包含不超过 14 个板块。中位数是 23 表面 50%的日期发生上涨的板块不超过 23 个。均值 24.75 表示所有的日期中平均发生上涨的板块为 24.75 个板块。第四段：如果数据集包含除了交易日期和板块之外的其他的列(如，用户 ID 等等)，会显示在这里。

8.3. 进行规则挖掘

为了进行关联规则的挖掘，通常预先设定一个合理的支持度作为阈值，这个阈值可以由具体研究的对象确定。本模型使用了 arules 包里的 apriori 函数。这里需要介绍 apriori 函数的用法，以及所用函数参数的说明：support，即支持度阈值参数，confidence，即置信度阈值参数。当 minlen = 1，说明{}=>{beer}这种关联规则是被允许的。但是本文的研究中不需要这种规则，所以本模型将 minlen 的参数值设置为 2。R 语言命令及产生结果如图 3 所示：

```
> groceryrules <-apriori(groceries, parameter = list(support = 0.28, confidence = 0.85, minlen = 2))
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport maxtime support minlen maxlen target   ext
       0.85    0.1    1 none FALSE            TRUE       5    0.28      2     10  rules FALSE

Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Absolute minimum support count: 136

set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[50 item(s), 486 transaction(s)] done [0.00s].
sorting and recoding items ... [50 item(s)] done [0.00s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 2 3 4 done [0.01s].
writing ... [147 rule(s)] done [0.00s].
creating S4 object  ... done [0.00s].
```

Figure 3. The result of Apriori function
图 3. Apriori 函数的使用及结果

在 R 语言命令行中使用 inspect 命令查看 apriori 函数产生的前六条规则，结果如表 4 所示：

Table 4. Produces the first six results of the association rule
表 4. 产生关联规则前 6 条的结果

lhs	rhs	support	confidence	lift
{T10}	=>{T16}	0.3559671	0.8160377	1.573787
{T8}	=>{T35}	0.3971193	0.8109244	1.608609
{T10,T30}	=>{T7}	0.2510288	0.8840580	1.691544
{T30,T35}	=>{T8}	0.2510288	0.8840580	1.805261
{T30,T8}	=>{T35}	0.2510288	0.8472222	1.680612
{T30,T8}	=>{T25}	0.2592593	0.8750000	1.707831

在 R 语言中也可以用可视化的展示方式来展示所得到规则的效果，如图 4 所示，LHS 代表了左侧项集，RHS 代表了右侧项集，在格点处绘制圆点，左右侧互相对应的项集便是圆点的位置。总共选取了按照提升度排序的前 20 条规则进行绘图，图中圆点的大小代表了支持度的大小，支持度越大，点的半径越

大。而提升度则用圆点颜色的深浅来表示，深度与提升度的大小呈正相关。

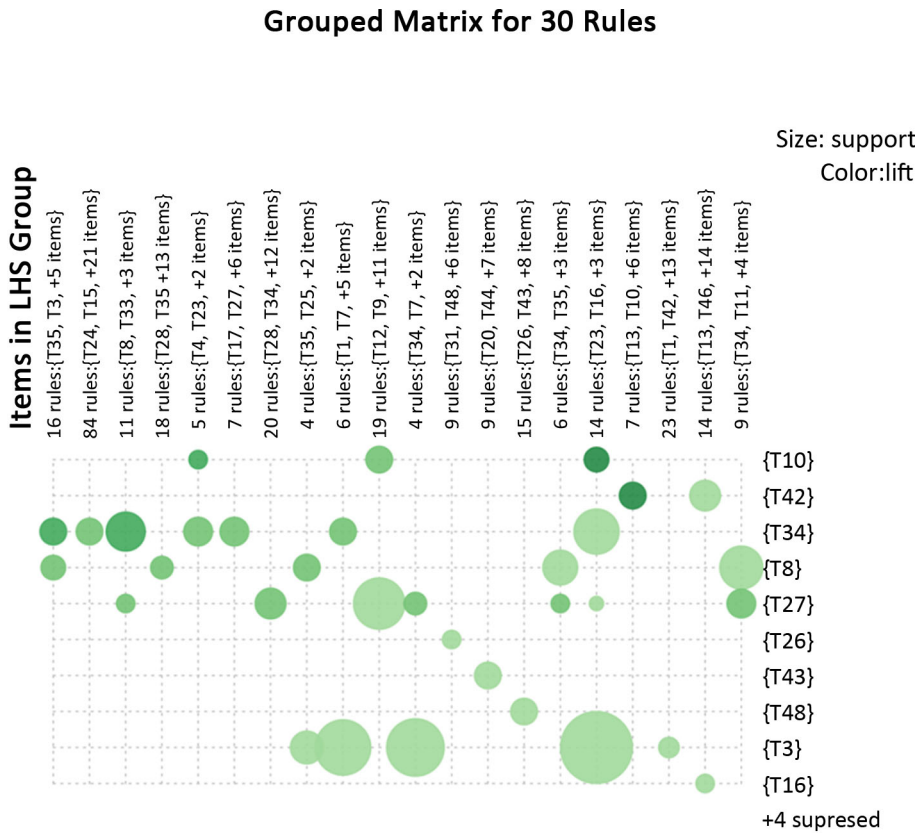

Figure 4. Visual presentation
图 4. 可视化的展示

参考文献

[1] 柳燕燕, 李兴平. 我国股票市场行业板块波动实证分析[J]. 科教文汇, 2013(3): 199-200.

[2] 杜伟锦, 何桃富. 我国证券市场的板块联动效应及模糊聚类分习[J]. 商业研究, 2005(22): 41-45.

[3] 张建林. 关联规则在股票板块联动分析中的应用[J]. 计算机工程与应用, 2013(2): 242-245.

[4] 冯甜. 我国股票市场行业板块联动效应的实证分析[J]. 时代金融, 2014(4Z): 158-159.

[5] 梁烨. 我国 A 股市场板块联动性效应分析[J]. 现代经济信息, 2014(12): 358-360.

[6] 杜巍, 赵春荣, 黄伟建. 改进的 k-means 聚类算法在客户细分中的应用研究[J]. 河北经贸大学学报, 2014, 35(1): 118-121.

[7] 张馨予. 基于 K-means 算法的北京市食品冷链农产品市场聚类分析[J]. 中国物流与采购, 2013(23): 68-69.

[8] 赵超. 数据挖掘关联规则的研究[J]. 网友世界, 2012(3): 43-45.

[9] 海沫, 牛怡晗, 张悦今. 面向大数据的并行聚类算法在股票板块划分中的应用[J]. 大数据, 2015(4): 9-17.