

基于影响力计算模型的股票网络社团划分方法

王 浩 李国欢 姚宏亮 李俊照

(合肥工业大学计算机与信息学院 合肥 230009)

(jsjxwangh@hfut.edu.cn)

Stock Network Community Detection Method Based on Influence Calculating Model

Wang Hao, Li Guohuan, Yao Hongliang, and Li Junzhao

(School of Computer and Information, Hefei University of Technology, Hefei 230009)

Abstract Taking advantage of the energy characteristics of complex system, a concept of influence is introduced to research community detection method, so that community structure could be discovered effectively. With regard to the stock closing price, by introducing the definition of influence and node centrality, a stock network is constructed with influence which is regarded as the edge weight. This paper proposes an algorithm named stock network hierarchical clustering based on the influence calculating model, which is referred to as BCNHC algorithm. Firstly, BCNHC algorithm introduces the definition of nodes' activity and influence, and puts forward the influence calculating model of node in networks in addition. Then, on the basis of measure criterion of the node centrality, the nodes with large node centrality value as the center nodes are selected, and the nodes' Intimacy and influence model are utilized to ensure the influence of association between neighbor nodes. Furthermore, the node with minimum degree is gathering toward to center nodes, so as to reduce the error clustering caused by the uncertainty of which community neighbor nodes belong to. On the basis, the neighbor communities are clustered with the average influence of association of communities. It guarantees that influence of association reach to maximization for all the nodes in the community, until the entire networks' modularity come to maximum. At last, comparison and analysis of experimental on stock network prove the feasibility of BCNHC algorithm.

Key words partial correlation; activity; stock network model; influence calculating model; influence of association; based on the center node hierarchical clustering algorithm about the influence calculation model of stock network (BCNHC)

摘 要 利用复杂系统的能量特性,引入影响力概念,研究动态复杂网络的社团划分方法,以有效地发现股票网络的社团结构.利用股票收盘价,通过引入影响力和结点中心性定义,构建以影响力为权值的股票网络,并提出一种基于影响力计算模型的股票网络中心结点层次聚类算法(based on the center node hierarchical clustering algorithm about the influence calculation model of stock network, BCNHC).BCNHC 算法首先引入结点活跃性和影响力的定义,并给出网络中结点的影响力计算模型;然后,基于所引入的结点中心性的度量准则,选取结点中心性大的结点为中心结点,并利用结点间的亲密性和影响力模型确定相邻结点之间影响力关联度;进而,通过优先选择度值最小的结点向中心结点聚集,以降低因相邻结点所属社团不确定而导致的错误聚类;在此基础上,利用社团平均影响力关联度对相邻社团进行聚类,保证社团内所有结点的影响力关联度最大化,直至整个网络模块度最大.最后,在构建的股票网络上的实验比较和分析,验证 BCNHC 算法的可行性.

关键词 偏相关性;活跃性;股票网络模型;影响力计算模型;影响力关联度;影响力计算模型的股票网络中心结点层次聚类算法

中图法分类号 TP181

复杂网络是一种要素、结构、功能和能量的有机聚合系统,而社团划分是研究复杂系统结构^[1]和功能特性的一种重要方法.与之相对应,当前的社团划分算法主要分两大类:1)基于结构特征的社团划分方法,如根据最短路径^[2]、结点聚集性^[3]等的社团划分方法;2)基于功能的社团划分方法,如根据结点相似性^[4]、加权等的社团划分方法.然而,这2类社团划分方法都没考虑到系统固有的能量特性,特别是动态复杂系统,其能量是系统演化的推动者.因而,在结构和功能划分基础之上引入能量要素,能更有效地实现对现实中动态复杂系统的划分,满足人们对于复杂系统特性分析的需要.

能量在不同的系统中有不同的存在和表现形式.在金融网络和社交网络中,能量表现为影响力.股市中每支股票都有各自的波动特性,波动性大表明其具有的能量大,对应股价涨跌幅度越大表明活跃性越高.另外,个体股票的波动对于其他股票所产生的影响力也是能量的体现.因而,可以将个体的活跃性和个体对于其他个体的影响力这2种能量的合力作为个体的总能量,通过提出影响力计算模型实现对能量大小的计算.

金融市场是一种动态复杂系统,构建金融市场的复杂网络是研究金融市场社团划分的基础.对于金融市场的复杂网络构建研究者进行了相关的研究,且当前的研究者主要利用股票价格波动序列^[5-6]的相关系数构建股票网络,如Wang选择不同行业股票的日收盘价数据,利用相关系数构建复杂网络,且网络结点的边权的绝对值大小呈现无标度特性^[7];Yang等人选取沪深300成分股、13个行业的284支股票日收益数据进行实证分析,用相关系数建立股票关联网络,得到网络的小世界性与无标度特性^[8].

然而,相关系数只能表明2个变量的共变联系,在实际应用中还要考虑这种相关性会受到其他因素的影响.为了更合理地表示股票之间的条件相关关系,将利用偏相关系数^[9]消除大盘走势对于股票间关系所产生的影响,并构建股票网络.偏相关系数又称为条件相关系数,先建立2个变量之间的相关关系,然后消除其他变量对于这2个变量之间关系的影响.因而,利用偏相关系数^[10]建立股票之间联系,能更准确地表示一个股票所具有的影响力.

复杂网络社团划分是利用聚类分析的思想,根据模块度来度量社团划分的效果.股票网络是一种加权网络,而加权网络的社团划分方法结合了复杂网络的结构特征和功能特征,是考虑双重因素的社会划分方法.CNM算法^[11]一种有权的社团划分算法,通过社团间模块度增量标准,实现结点的聚类;BGLL算法^[12]是一种基于树结构的分层加权社团划分方法,根据结点间的层次联系来实现划分.然而,当前的社会划分方法都是从复杂网络的结构或功能出发,没有考虑复杂系统的能量特性,存在社团结构、功能和能量优势的不一致性.股票网络中心结点类似于社交网络的意见领袖,是能量聚集的中心,因而研究了基于中心结点影响力计算的复杂网络分层社团划分方法.将基于边的强度和结点的影响范围,提出结点中心性的定义,改进已有的基于中心结点社团划分方法仅考虑结点度的不足.

本文基于复杂系统的能量特性,引入针对股票网络的活跃性和影响力定义,并给出网络中结点的影响力计算模型;进而通过结点中心性的定义和利用社团的层次划分思想,提出一种基于影响力计算模型的股票网络中心结点层次聚类算法(based on the center node hierarchical clustering algorithm about the influence calculation model of stock network, BCNHC).BCNHC算法基于层次聚类策略,利用结合影响力与结点的度给出结点中心性的度量标准,选取中心结点作为聚类的初始社团,将非中心结点以度的升序顺序选择关联度最大的中心结点进行聚类,第1层聚类主要是相邻结点的聚类;然后再根据已经得到的社团结构,依据社团平均关联度最大原则进行社团的聚类.以结点中心性大的结点为中心进行聚类,得到的群体结构内部影响力大,而社团间的影响力较小,由此可以根据中心结点进行股票的投资组合推荐^[13-14].最后从行业板块和模块度来评价股票是否受社团结构的影响.

1 影响力计算模型

1.1 相关性和偏相关性

首先建立无标度的加权网络模型.设股票 $i(i=1,2,\cdots,N)$ 在第 t 个交易日的收盘价为 $Y_i(t)$,交易日为 n 天,对应的大盘指数 s 在第 t 个交易日的收

盘指数为 $P_s(t)$; 其中, $S_i(t) = \ln Y_i(t) - \ln Y_i(t-1)$ 和 $Z_s(t) = \ln P_s(t) - \ln P_s(t-1)$, 分别表示第 i 支股票和大盘指数 s 的对数收益率。

设股票 i 和 j 之间的相关系数为 R_{ij} , 大盘指数的变化会对 R_{ij} 产生影响, R_{ij} 可表示如下:

$$R_{ij} = \frac{\langle S_i S_j \rangle - \langle S_i \rangle \langle S_j \rangle}{\sqrt{(\langle S_i^2 \rangle - \langle S_i \rangle^2)(\langle S_j^2 \rangle - \langle S_j \rangle^2)}}. \quad (1)$$

大盘指数与股票 i 之间的相关系数 R_{is} 可表示如下:

$$R_{is} = \frac{\langle S_i Z_s \rangle - \langle S_i \rangle \langle Z_s \rangle}{\sqrt{(\langle S_i^2 \rangle - \langle S_i \rangle^2)(\langle Z_s^2 \rangle - \langle Z_s \rangle^2)}}, \quad (2)$$

其中, $\langle S_i \rangle$ 为股票 i 在 n 天内的平均收益率, $\langle S_i \rangle = \frac{1}{n} \sum_{t=1}^n S_i(t)$; $\langle S_i S_j \rangle$ 为股票 i, j 同时时间收益率成绩的均值; $\langle S_i^2 \rangle$ 为股票 i 收益率平方的均值; $\langle S_i \rangle^2$ 为股票 i 收益率均值的平方。

消除大盘指数 s 对于 R_{ij} 所产生的影响, 得到股票 i 和 j 之间的偏相关系数 $R_{ij,s}$, 可表示如下:

$$R_{ij,s} = \frac{R_{ij} - R_{is} R_{js}}{\sqrt{(1 - R_{is}^2)(1 - R_{js}^2)}}, \quad (3)$$

其中, R_{js} 表示 j 和大盘指数 s 之间的相关系数, 股票间的偏相关系数 $R_{ij,s}$ 取值范围为 $-1 \sim 1$ 。

1.2 活跃性

活跃性大的股票具有更大的投资价值, 同时具有更大的风险性。在股市中, 股票的活跃性可通过换手率和成交量来体现。由于股票的成交量、换手率都难以消除大盘指数对其产生的影响, 而股票价格波动也能直观地体现股票的活跃性, 也就是价格波动越大股票越活跃; 因此, 利用股票价格的波动性描述股票的活跃性。

设个股涨跌幅为 $W_i(t) = [Y_i(t) - Y_i(t-1)] / Y_i(t-1)$, 大盘涨跌幅为 $U_s(t) = [P_s(t) - P_s(t-1)] / P_s(t-1)$ 。

用股票的收盘价相对于大盘指数的对数收益率变化之差的平方表示股票的活跃性, 则个股 i 相对大盘指数的活跃性 H_{is} 可表示如下:

$$H_{is} = \frac{1}{n} \sum_{t=1}^n [W_i(t) - U_s(t)]^2. \quad (4)$$

1.3 影响力计算模型

个股能量表现为活跃性和股价波动对于其他股票所产生的影响力, 这 2 个方面的合力构成个股的影响力。下面考虑股票 i 对于股票 j 的影响力, 设股票 i 和股票 j 之间的偏相关性为 $R_{ij,s}$, 若 $H_{is} > H_{js}$, 则股票 i 对股票 j 的影响力大于股票 j 对股票 i 的

影响力^[15], 表明股票 i 对股票 j 的影响力不具有对等性。股票 i 对股票 j 的影响力 $F(i \rightarrow j)$ 可以描述为

$$F(i \rightarrow j) = R_{ij,s} \times \frac{H_{is}}{H_{is} + H_{js}}. \quad (5)$$

给定偏相关系数 $R_{ij,s}$, $\frac{H_{is}}{H_{is} + H_{js}}$ 越大, 表明 i 对 j 的影响力 $F(i \rightarrow j)$ 越大。由于每支股票的活跃度不一致, 因而 F 是一个不对称矩阵, $F(i \rightarrow j)$ 简化为 F_{ij} 表示矩阵中的元素。

股票与股票间的影响力具有不对等性, 这也使得股票间的影响关系和股票间影响力的计算结果难以用于指导社团的划分。为了简化股票间的影响关系, 需要给出一个能度量二者间相互影响力的标准。如股票 i 和股票 j 之间的影响力 F_{ij} 和 F_{ji} 是不对等的, 可由股票自身的活跃性出发, 给出股票 i 和股票 j 影响力计算模型 W_{ij} , 可以描述如下:

$$W_{ij} = F(i \rightarrow j) \times H_{js} = F(j \rightarrow i) \times H_{is} = R_{ij,s} \times \frac{H_{is} H_{js}}{H_{is} + H_{js}}, \quad (6)$$

其中, W_{ij} 表示的 2 个结点间影响关系具有对等性, 当 H_{is} 和 H_{js} 越大则 W_{ij} 越大; W_{ij} 与 $F(i \rightarrow j)$ 变化趋势同步。

1.4 结点中心性

设 W_{ij} 表示结点 i 对结点 j 的影响力, $d(i)$ 表示结点 i 的度, 结点 i 的强度 $dn(i)$ 可描述如下:

$$dn(i) = \left[\sum_k count(W_{ik} \geq W_{ij}) \right], \quad W_{ij} \leq dn(i) / d(i), \quad (7)$$

其中, k 与 j 均是与 i 相连的邻居结点, $count()$ 函数统计在满足条件 $W_{ik} \geq W_{ij}$ 下的个数。

结点强度是衡量股票重要性的一个重要标准, 它是用度从影响范围^[16]的角度衡量结点在股票网络中的中心地位。同时, 若与该股票相连的股票较多, 但影响力强度都很小, 则该结点在网络中的影响力也弱, 其重要性也小。因此, 结合影响范围和影响强度定义股票结点 i 的中心性 $ION(i)$, 可描述如下:

$$ION(i) = \sum_{W_{ik} \geq W_{ij}} \frac{W_{ik}}{dn(i)} \times d(i) \times Infl(i) = \frac{1}{dn(i)} \sum_{W_{ik} \geq W_{ij}} W_{ik} \times d(i) \times Infl(i), \quad W_{ij} \leq dn(i) / d(i), \quad (8)$$

其中, $Infl(i) = \sum_j W_{ij}$ 表示与结点 i 相邻所有结点 j 的权值之和。当结点 i 的影响力大且影响的范围广时, 结点 i 在网络中的重要性越大中心性也就越强。

2 基于影响力计算的社团结构划分

2.1 股票结点的距离

股票网络是以股票为结点,偏向关系数为权边, i 结点与 j 结点之间的亲密性 d_{ij} ^[17] 描述如下:

$$d_{ij} = \sqrt{2(1-W_{ij})}, \quad (9)$$

其中, d_{ij} 越小表示 i 结点与 j 结点间的影响力越大,也就越亲密.

2.2 股票结点的关联度

Fortunato 等人提出了一种基于信息中心度的社团探测算法^[18],假设网络中 2 个结点之间的信息总是沿着最短路径传播的,认为结点之间信息传输的效率为它们之间最短路径长度的倒数.定义 2 个相邻结点 i 和 j 的关联度为

$$nodeLink(i, j) = 1 - C_{e_{ij}}, \quad (10)$$

其中,整个网络信息传输的有效性为各结点对信息传输有效率的平均值;信息中心度 $C_{e_{ij}}$ 为移除该边后整个网络的信息有效率减少的相对量.结点间的关联度定义表明:边的信息中心度越大,该边是社团之间边的概率越大,结点间的关联度越小.

由于关联度是用于无权网络,为了使其适用于有权的股票网络.设结点自身的关联度 $nodeLink(i, j) = 1$,网络中结点之间的关联度矩阵为 L ,主对角线上元素表示各结点自身的关联度 L_{ii} ,因而值均为 1; $L_{ij} (i \neq j)$ 表示网络中任意 2 个结点之间的关联度,则关联度矩阵 L 定义如下:

$$L_{ij} = \begin{cases} 1, & i = j, \\ nodeLink(i, j), & i \neq j, \end{cases} \quad (11)$$

其中,非相邻结点之间关联度为 2 结点之间最短路径上所有边的关联度之积.

2.3 社团评价标准

设股票网络 $G(V, E, W)$, 其中 $V = \{v_1, v_2, \dots, v_i, \dots\}$ 是股票网络 G 的结点集, E 是 G 的边集, W 是 E 上的权重,即为影响力,则相邻结点 v_i 和 v_j 之间边的影响力值为 w_{ij} . 则模块度 Q 定义如下:

$$Q = \frac{1}{2s} \sum_i \sum_j \left[w_{ij} - \frac{s_i s_j}{2s} \right] \delta(G_i, G_j), \quad (12)$$

其中, $s_i = \sum_j w_{ij}$ 表示与结点 v_i 相邻所有结点 v_j 边的影响力之和; $s = \sum_i \sum_j w_{ij}$ 表示股票网络中所有的边的权值之和; G_i 表示结点 v_i 所在的社团,当 G_i 与 G_j 相同时, $\delta(G_i, G_j) = 1$, 否则 $\delta(G_i, G_j) = 0$.

2.4 BCNHC 算法描述

基于影响力计算模型的中心结点层次聚类算法的基本思想:利用中心结点的层次聚类 BCNHC 算法思想,结合影响力与结点的度给出结点中心性的度量标准,选取中心结点作为聚类的初始社团,将非中心结点以度的升序顺序选择关联度最大的中心结点进行聚类,第 1 层聚类主要是相邻结点的聚类;然后再根据已经得到的社团结构,依据社团平均关联度最大原则进行社团的聚类.

算法 1. BCNHC 算法.

输入: 股票网络的影响力矩阵 F 及活跃性矩阵 H ;

输出: 网络的社团结构.

Step1. 第 1 层聚类: 设网络结点集为 V , 由式(8)计算网络中各结点的中心性 ION ; 将所结点的初始标志设 0, 即对于结点 $v \in V$, 令 $flags(v) = 0$, 表示结点未被访问.

Step1. 1. 根据股票结点中心性 ION 的大小, 选择结点中心性满足一定阈值条件的结点作为第 1 次聚类的中心结点, 得到一个中心结点集 V_1 ; 每个中心结点 $c \in V_1$ 可以看作是一个社团(该社团只有一个结点), 并且令 $flags(c) = 1$. 非中心结点集 $V_2 = V - V_1$ 中的结点将被聚集到不同的中心结点所在的社团中.

Step1. 2. 对非中心结点的度进行升序排序, 并根据结点间的关联度实现对非中心结点的聚类. 具体过程如下:

Step1. 2. 1. 选取度值最小的结点 v ;

Step1. 2. 2. 当 $flags(v) = 0$ 时, 计算 v 与所有社团的关联度;

Step1. 2. 3. 若 v 与已聚类的社团中的任意结点相邻, 则选择关联度最大进行的聚类, 并令 $flags(v) = 1$; 执行 Step1. 2. 1. 若待聚类结点 v 与各个社团中的结点均不相邻, 则不进行聚类操作, 并选择当前结点 v 的下一个结点, 进行 Step1. 2. 2 操作.

Step1. 2. 4. 循环 Step1. 2. 1~Step1. 2. 3, 直至所有的结点均被访问结束.

Step2. 在第 1 层聚类所得到的社团进行第 2 次划分.

先计算当前网络社团的模块度, 然后根据各社团平均关联度大小的比较, 选择关联度最大的 2 个社团 C_i 和 C_j 进行聚类; 然后, 再次计算聚类后的网络社团的模块度, 并计算新的各社团之间的平均关

联度,选择关联度最大的 2 个社团;如此操作,直到所有的社团被聚为整个网络为止。

Step3. 根据模块度最大化原理,从 Step2 所获得的模块度序列中选择当模块度最大时所对应的社团结构,为最佳划分结果。

2.5 BCNHC 算法复杂性分析

复杂网络 $G(V,E)$ 有 n 个结点、 m 条边,聚类得到的社团个数为 k 。BCNHC 算法主要分为第 1 层的基于中心结点的划分和第 2 层的层次聚类的思想,第 1 层中对已以结点度升序排序后,按照关联度最大原则选择与中心结点相邻的结点聚类,其时间复杂度最坏的情况为 $O(n^2)$;第 2 层在计算各个社团间的关联度,选择关联度最大的 2 个社团进行聚类,

其时间复杂度为 $O(k^2)$;则 BCNHC 算法的时间复杂度为 $O(n^2+k^2)$ 。由于,社团的个数远远小于网络中的结点数 $k \ll n$,则 BCNHC 算法复杂性近似为 $O(n^2)$ 。

3 实 验

3.1 个体股票网络

3.1.1 股票网络的偏相关系数分布

针对上证 180 支股票在 2009-03-03—2009-08-05 的日线收盘价数据,这个阶段是上证指数的上升段数据。通过去掉数据不完整的部分股票,最终选择其中符合要求的 94 支股票,如表 1 所示:

Table 1 Stock Name and Serial Number

表 1 股票名称和编号

Serial Number	Name	Serial Number	Name	Serial Number	Name	Serial Number	Name	Serial Number	Name
1	浦发银行	20	同方股份	39	南山铝业	58	西南证券	77	海螺水泥
2	武钢股份	21	上汽集团	40	海南航空	59	首开股份	78	中孚实业
3	上海机场	22	国金证券	41	云南城投	60	金地集团	79	北大荒
4	包钢股份	23	包钢稀土	42	中恒集团	61	盘江股份	80	青岛啤酒
5	华夏银行	24	东方航空	43	广汇能源	62	国电南瑞	81	鼎立股份
6	民生银行	25	中国卫星	44	广晟有色	63	小商品城	82	嘉宝集团
7	宝钢股份	26	兰花科创	45	北京城建	64	江淮汽车	83	万业企业
8	中国石化	27	铁龙物流	46	航天信息	65	吉恩镍业	84	申能股份
9	南方航空	28	重庆啤酒	47	恒瑞医药	66	中金黄金	85	爱建股份
10	中信证券	29	中国船舶	48	烟台万华	67	XD 驰宏锌	86	城投控股
11	三一重工	30	航天机电	49	荣华实业	68	烽火通信	87	豫园商城
12	招商银行	31	中体产业	50	洪都航空	69	天富热电	88	信达地产
13	保利地产	32	巨化股份	51	天房发展	70	方大炭素	89	福耀玻璃
14	中国联通	33	福田汽车	52	华发股份	71	康美药业	90	陆家嘴
15	五矿发展	34	太原重工	53	宏达股份	72	贵州茅台	91	哈药股份
16	南京高科	35	雅戈尔	54	阳泉煤业	73	山东黄金	92	川投能源
17	葛洲坝	36	兖州煤业	55	浙江龙盛	74	厦门钨业	93	中华企业
18	同仁堂	37	复星医药	56	江西铜业	75	天威保变	94	青岛海尔
19	特变电工	38	浙江医药	57	宁波韵升	76	海油工程		

根据相关系数和偏相关系数的式(1)~(3)计算 94 支股票的相关系数和偏相关系数,得出的概率分布如图 1 所示。

相关系数研究的是个体之间简单的相关关系,偏相关系数通过消除某些外界干扰因素,研究个体间比较纯粹的相关关系。由图 1 可知相同数据集获得的偏相关系数和相关系数的差异较大,偏相关系

数的概率分布更接近于正态分布,符合股市涨、跌平衡的原理,因而采用偏相关系数的方法研究股票间的相关关系。

3.1.2 股票的活跃性

股票活跃性定义为股票在股市中价格的波动性,它是股票的波动性能量特性之一,也是选股的一个重要因素。利用式(4)计算个股相对于大盘价格指

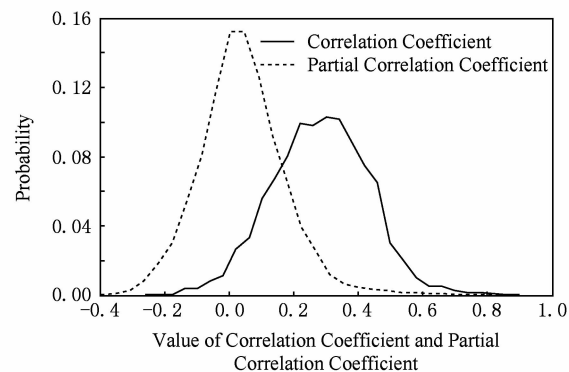


Fig. 1 Probability distribution of correlation coefficient and partial correlation coefficient.

图1 相关系数和偏相关系数的概率分布

数涨、跌幅度的方差,得到每支股票的活跃性如图2所示:

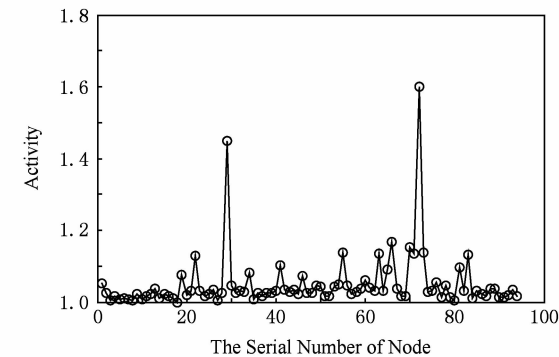


Fig. 2 Activity of each stock.

图2 个股的活跃性

图2的横坐标代表股票结点编号,纵坐标表示股票活跃性值;其中72号结点是贵州茅台活跃性最大,活跃值为1.6.

3.1.3 股票的影响力模型

影响力不仅考虑了股票间的偏相关关系,同时还结合了个股的活跃性;因而,相对于偏相关关系而

言,影响力能很好地体现股票的能量特性. 根据偏相关系数计算式(3)和影响力计算式(6),可得出如下的偏相关系数与影响力的概率分布,如图3所示:

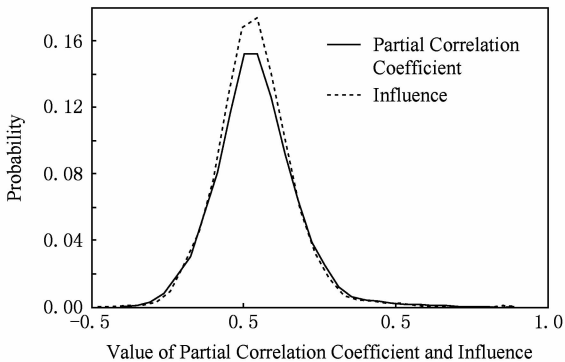


Fig. 3 Probability distribution of influence and partial correlation coefficient.

图3 影响力和偏相关系数的概率分布

偏相关系数是从股票对外的影响力角度分析股票结点间的相关关系,而影响力结合了偏相关性和活跃性2方面的概念. 从图3可以看出,偏相关系数与影响力的概率分布相近,其大部分结点影响力和偏相关系数集中在-0.2~0.2之间. 偏相关系数与影响力值相近,表明所引入的影响力定义的合理性,同时表明影响力与偏相关系数的差异性.

根据图3中偏相关系数与影响力的概率分布,设定各阈值均为0.2. 分别以偏相关系数和影响力的值作为网络的边,当 $|R_{ij,s}|$ 和 $|W_{ij}|$ 都大于阈值时,结点 i 和 j 之间有边;当 $|R_{ij,s}|$ 和 $|W_{ij}|$ 都小于阈值时,结点 i 和 j 之间无边. 则以偏相关系数为权值所构建的网络如图4(a)所示,以影响力为权值所构建的网络如图4(b)所示. 由此构建的偏相关系数股票网络结点的度分布如图5(a)所示,影响力股票网络结点的度分布如图5(b)所示.

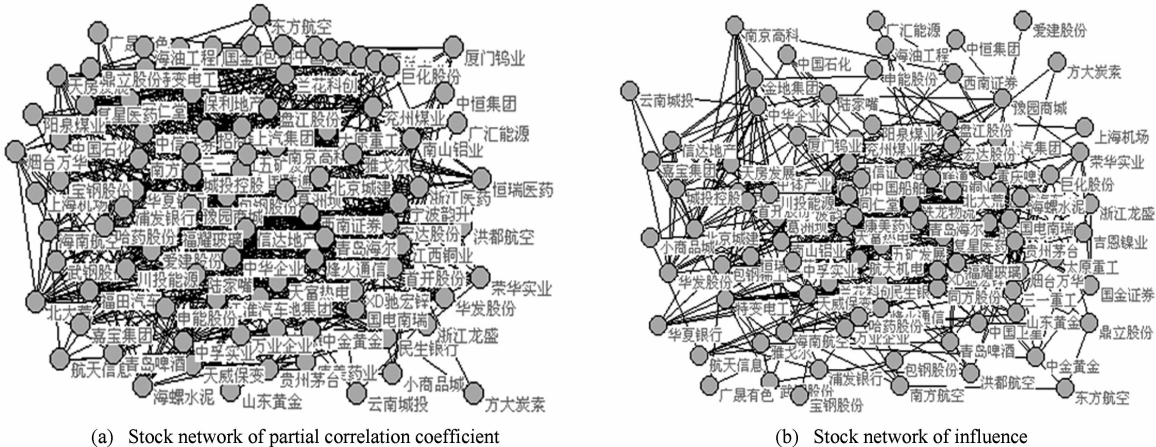


Fig. 4 Stock network of partial correlation coefficient and influence.

图4 偏相关系数股票网络和影响力股票网络

由图 5(a)和图 5(b)可知,在构建的股票网络中,股票结点度的分布具有无标度性;图 4(a)中的网络和图 4(b)中的网络都具有同样的特点:度较大的结点的个数少,度较小的结点的个数也少,处于中

间值 10 左右的结点数偏多. 度值大的结点在网络中占较少的部分,但对股市中的多数股票都有影响. 因此,根据式(8)分别计算股票网络中股票的结点中心性,如图 6(a)和图 6(b)所示.

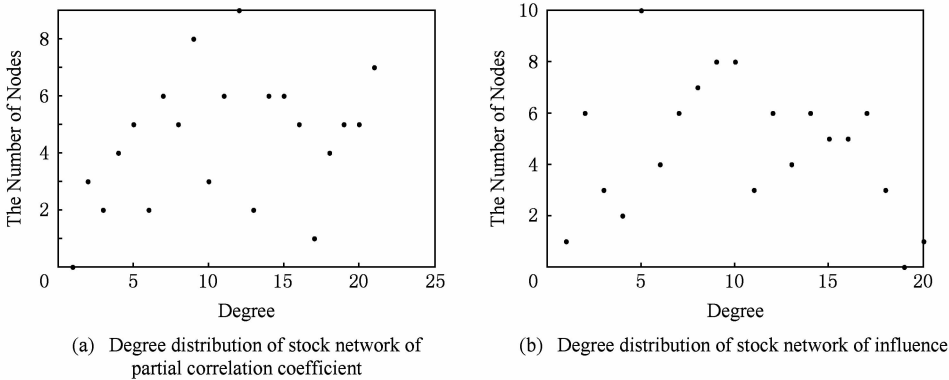


Fig. 5 Degree distribution of stock network of partial correlation coefficient and influence.

图 5 偏相关系数股票网络和影响力股票网络的度分布

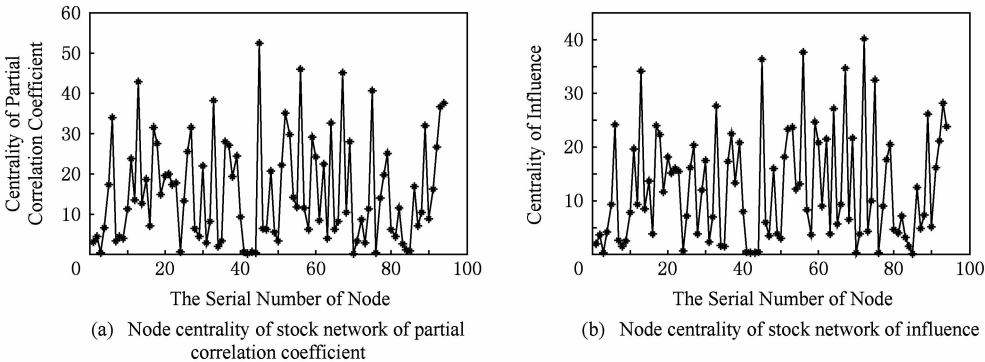


Fig. 6 Node centrality of stock network of partial correlation coefficient and influence.

图 6 偏相关系数股票网络和影响力股票网络的结点中心性

图 6(a)为偏相关系数股票网络的结点中心性,以偏相关系数为权值;图 6(b)为影响力股票网络的中心性,以影响力为权值. 根据式(8),偏相关系数股票网络的结点中心性计算就是将 W_{ij} 替换为 $R_{ij,s}$. 股票结点的中心性越大则股票在网络中核心地位越高,更有可能成为聚类的中心结点.

3.1.4 社团结构划分

根据所确定的核心结点作为聚类的中心结点,分别对偏相关系数股票网络和影响力股票网络用 BCNHC 算法进行社团划分.

算法 1 的过程:设定不同结点中心性的阈值,确定聚类的中心结点,进行第 1 层向中心结点的聚类. 再以第 1 层聚类的社团为聚类初始的结构,选择社团间平均关联度最大的社团两两聚类,直至为一个整体网络为止.

根据不同结点中心性的阈值,得到的中心结点的个数及相应的模块度值如表 2 所示:

Table 2 Different Modularity with Different Number of Central Nodes in Different Threshold Value

表 2 不同阈值选择的中心结点数和对应的模块度

Threshold Value of Node Centrality	Number of Central Node	Maximum Value of Modularity	
		Stock Network of Influence	Stock Network of Partial Correlation Coefficient
0.15	16	0.4113	0.2411
0.2	19	0.3609	0.3614
0.25	24	0.3615	0.3740
0.3	29	0.3587	0.3559
0.35	33	0.3833	0.3513

由表 2 可知,影响力股票网络结点中心性的最佳阈值为 0.15,偏相关系数股票网络的结点中心性的最佳阈值为 0.25. 以影响力股票网络为例,当选择阈值为 0.15 时,中心结点数为 16,社团的模块度达到最大值. 选择的 16 支股票如表 3 所示:

Table 3 Stocks with Larger Node Degree and Their Industry in Two Stock Networks

表 3 结点度数较大的股票及所属行业				
Serial Number	Stock Name	Degree of Nodes in Stock Network of Influence	Degree of Nodes in Stock Network of Partial Correlation Coefficient	Industry
72	贵州茅台	20	12	食品饮料
89	福耀玻璃	18	20	金属材料
13	保利地产	18	21	地产
17	葛洲坝	18	21	建筑业
33	福田汽车	17	20	机械仪表
45	北京城建	17	21	房地产
64	江淮汽车	17	19	机械仪表
94	青岛海尔	17	21	机械仪表
6	民生银行	16	20	金融
56	江西铜业	16	18	金属材料
75	天威保变	15	19	机械仪表
93	中华企业	15	21	房地产
52	华发股份	14	18	房地产
59	首开股份	14	17	房地产
67	XD 驰宏锌	14	16	采掘业
53	宏达股份	12	14	金属材料

由表 3 可知,构建的影响力股票网络中,结点度最大为 20.表 3 所列的股票结点的度均大于 10,处于股票网络的中心位置,是网络的核心结点.这 16 支股票所属的行业是我国经济的核心行业,金融、房地产、金属材料等都是中国经济的命脉.另外,在偏相关系数股票网络中结点的度最大为 21,这 16 支股票的度均大于 14,同样是偏相关系数股票网络的核心结点.再根据 2009 年各行业板块的龙头股列表,已选择的股票大部分都属于行业板块的龙头股.因此,用影响力选择影响力大的股票具有一定的可行性.

对于偏相关系数网络,选取结点中心性阈值为 0.25,得到的社团模块度序列如图 7(a)所示,选择影响力股票网络结点中心性阈值为 0.15,得到的社团模块度序列如图 7(b)所示.

由图 7(a)可以看出,当社团数为 6 时,偏相关系数股票网络对应的模块度为 0.374 0、为最大值;由图 7(b)可以看出,当社团数为 4 时,影响力股票网络对应的模块度为 0.411 3、为最大值.

用 BCNHC 算法分别对偏相关系数股票网络划分得到的社团如图 8(a)所示,用 BCNHC 算法分别对影响力股票网络划分得到的社团如图 8(b)所示.

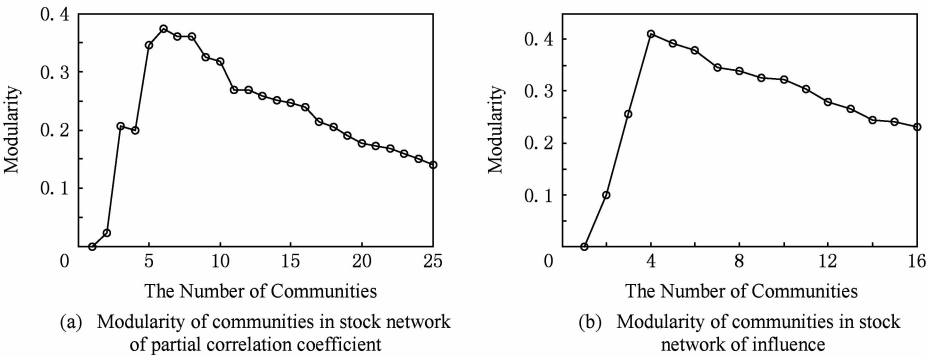


Fig. 7 Modularity of communities in stock network.

图 7 股票网络社团数对应的模块度

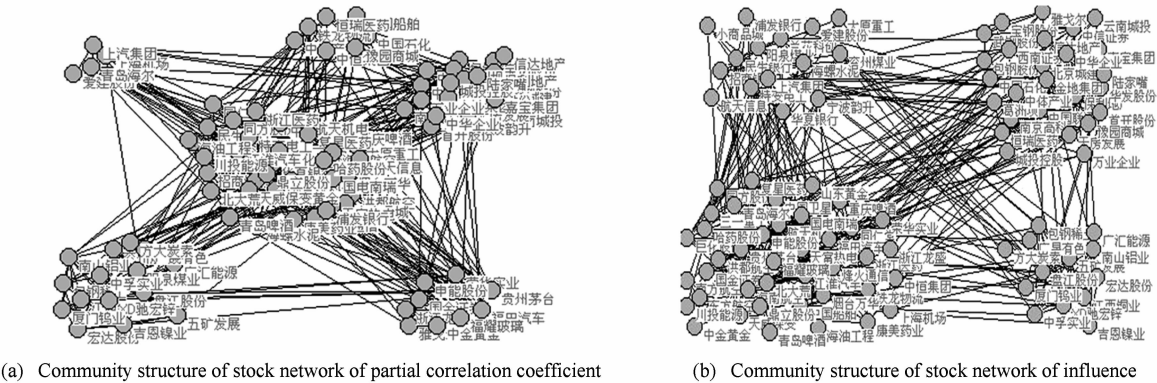


Fig. 8 Community structure of stock network of partial correlation coefficient and influence.

图 8 构建的 2 种股票网络的社团结构

3.1.5 实验结果的分析

经过社团结构的划分学习,偏相关系数股票网络被划分为 6 个社团结构,各社团的规模分别为 3, 7,10,15,22,37;影响力股票网络被划分为 4 个社团结构,各个社团结构的规模分别为 40,26,15,13. 构成网络的 94 支股票可大体分为银行金融类、金属材料、交通仓储、采掘业、机械仪表、房地产、生物医药、信息技术几个板块. 首先,从板块的角度分析,银行金融板块包括浦发银行、华夏银行、民生银行、招商银行、中信证券、西南证券被划入一个社团,房地产板块包括保利地产、中体产业、云南城投、北京城建、天房发展、华发股份、首开股份、金地集团、万业企业、信达地产、陆家嘴、中华企业均属于一个社团,属

于比较合理的划分. 其次,银行金融类、房地产是主导国家经济的主导因素,金属材料也与部分采掘业密不可分,机械仪表、信息技术、食品饮料、石油化工、生物医药板块与人们的生活息息相关. 从整个网络看,各行业、各板块都相互关联、相互影响,划分的社团结构是根据各自的关联度强强联合,所有具有强关联性的股票、强关联性的行业板块被划到同一个社团中.

利用 GN 算法、CNM 算法、BGLL 算法分别对已构建的相关系数、偏相关系数股票网络和影响力股票网络进行社团划分,且用模块度评价划分的结果,并与 BCNHC 算法进行比较,其结果如表 4 所示:

Table 4 The Result of Modularity with Different Community Partition Algorithm
表 4 不同划分算法的模块度

Algorithm	Stock Network of Influence		Stock Network of Partial Correlation Coefficient		Stock Network of Correlation Coefficient	
	Communities	Modularity	Modularity	Modularity	Modularity	Modularity
GN	3	0.3927	5	0.3523	4	0.1376
CNM	4	0.4488	4	0.4002	3	0.1245
BGLL	4	0.3870	6	0.3708	5	0.1452
BCNHC	4	0.4113	6	0.3740	5	0.1384

由表 4 可知,对于 GN,CNM,BGLL,BCNHC 这 4 种算法有同样的特点,利用相关系数的划分结果显著差于利用偏相关系数的划分结果,同时利用偏相关系数的划分结果要明显差于利用影响力的划分结果,表明引入的偏相关系数和影响力计算模型会使股票网络的社团划分结果得到显著提升,进而表明基于能量的社团划分思想的有效性.

另外,由表 4 可以看出得到 3 个社团,而 CNM,BGLL,BCNHC 算法均得到 4 个社团,且 GN 算法的模块度值最小,由此可以看出,所构建的影响力股票网络较适合的社团数为 4;BCNHC 算法在影响力股票网络和偏相关系数股票网络的模块度较 GN,BGLL 算法的大,反映 BCNHC 算法具有可行性;BCNHC 算法得到的社团模块度略小于 CNM 算法

得到的模块度,是由于 BCNHC 算法以股票网络结点间关联度为聚类标准,而 CNM 算法是以整个网络的关联度为聚类标准,CNM 算法的聚类要求更加严格,导致 BCNHC 算法的模块度值略小于 CNM 算法. 可见,BCNHC 算法对于股票网络社团划分的有效性.

3.2 行业板块网络

本文选取行业板块构建股票的行业板块网络,验证 BCNHC 算法的有效性. 该样本涵盖了金融、工业、建材等 32 个行业领域,每个板块的价格指数波动序列为 2008 年到 2009 年 183 个交易日收盘指数. 分别构建相关系数、偏相关系数板块网络和影响力板块网络,并对所构建的 3 个网络分别进行社团划分,用模块度进行结果的评价,其结果如表 5 所示:

Table 5 Result of Modularity in Plate Network with BCNHC Algorithm
表 5 BCNHC 算法对板块网络的划分的模块度结果

Algorithm	Stock Network of Influence		Stock Network of Partial Correlation Coefficient		Stock Network of Correlation Coefficient	
	Communities	Modularity	Modularity	Modularity	Modularity	Modularity
BCNHC	4	0.0446	5	0.0415	2	0.0013

由表 5 可知,BCNHC 算法对于行业板块网络划分的结果显示:利用相关系数的划分结果显著差于利用偏相关系数的划分结果,同时利用偏相关系数的划分结果要明显差于利用影响力的划分结果.这进一步表明 BCNHC 算法对于具有能量特征的股票网络进行社团划分的有效性.

4 结束语

本文利用复杂系统的能量特性,引入针对股票网络的活跃性和影响力定义,并给出网络中结点的影响力计算模型;进而通过结点中心性的定义和利用社团的层次划分思想,提出一种基于影响力计算模型的股票网络中心结点层次聚类算法 BCNHC.影响力计算模型不仅适用于个股构建的大型股票网络,而且也适合行业板块指数构建的复杂网络.基于能量特性的系统结构和功能分析将是下一步要研究的工作.

参 考 文 献

[1] Bonanno G, Lillo F, Mantegna R N. Levels of complexity in financial market [J]. Statistical Mechanics and Its Applications, 2001, 299(1/2): 16-27

[2] Newman M E J, Girvan M. Finding and evaluating community structure in networks [J]. Physical Review E, 2004, 69(2): 113-127

[3] Bolanos M, Bernat E M, He B, et al. A weighted small world network measure for assessing functional connectivity [J]. Journal of Neuroscience Methods, 2013, 212(1): 133-142

[4] Xie Fuding, Ji Min, Zhang Yong, et al. The detection of community structure in network via an improved spectral method [J]. Physica A: Statistical Mechanics and Its Applications, 2009, 388(15): 3268-3272

[5] Tse C K, Liu Jing, Lau F. A network perspective of the stock market [J]. Journal of Empirical Finance, 2010, 17(4): 659-667

[6] Roy R B, Sarkar U K. A social network approach to examine the role of influential stocks in shaping interdependence structure in global stock markets [C] //Proc of 2011 Int Conf on Advances in Social Networks Analysis and Mining. Los Alamitos, CA: IEEE Computer Society, 2011: 567-569

[7] Wang Peng. Analysis of the topology structure of network of Chinese stock market [C] //Proc of 2010 Int Conf on Management of e-Commerce and e-Government. Los Alamitos, CA: IEEE Computer Society, 2012: 403-407

[8] Yang Zhihui, Jia Hanmei. Network structure of the correlation between stock returns [C] //Proc of the 30th Chinese Control Conference (CCC 2011). Piscataway, NJ: IEEE, 2011: 5732-5736

[9] Zhang Jie, Chen Yejun, Zhai Dongsheng. Network analysis of Shanghai sector in Chinese stock market based on partial correlation [C] //Proc of 2011 Int Conf on Information Management and Engineering. Piscataway, NJ: IEEE, 2010: 321-324

[10] Kim J M, Jung Y S, Choi T, et al. Partial correlation with copula modeling [J]. Computational Statistics & Data Analysis, 2011, 55(3): 1357-1366

[11] Han Hua, Wang Juan, Wang Hui. A new algorithm to detect community structures on weighted network [C] //Proc of 2010 Int Conf on Computational Intelligence and Software Engineering. Piscataway, NJ: IEEE, 2010: 1-4

[12] Blondel V D, Guillaume J L, Lambiotte R, et al. Fast unfolding of community hierarchies in large networks [J]. Journal of Statistical Mechanics: Theory and Experiment, 2008, 10: 10008. [2014-04-01]. <http://works.bepress.com/lambiotte/>

[13] Qin Sen, Chen Xufeng, Sun Weigang. Optimal investment portfolio model with degree risk in complex stock network [C] //Proc of 2011 the 4th Int Workshop on Chaos-Fractals Theories and Applications. Los Alamitos, CA: IEEE Computer Society, 2011: 161-165

[14] Liao Shuhsien, Chou Shanyuan. Data mining investigation of co-movements on the Taiwan and China stock markets for future investment portfolio [J]. Expert Systems with Applications, 2013, 40(5): 1542-1554

[15] Arora N, Narayanan B, Paul S. Financial Influences and Scale-free Networks [M]. Berlin: Springer, 2006: 16-23

[16] Ge Xin, Zhao Hai, Zhang Jun. Degree correlation and its features of spreading on networks [J]. Journal of Computer Research and Development, 2013, 50(4): 741-749 (in Chinese)

(葛新, 赵海, 张君. 网路度相关及其传播特征研究[J]. 计算机研究与发展, 2013, 50(4): 741-749)

[17] Hang Weiqiang, Zhang Xintian, Yao Shuang. Topology property and clustering structure analysis of China's stock correlation network [J]. Journal of Management Sciences, 2009, 21(3): 94-103 (in Chinese)

(黄玮强, 庄新田, 姚爽. 中国股票关联网络拓扑性质与聚类结构分析[J]. 管理科学, 2009, 21(3): 94-103)

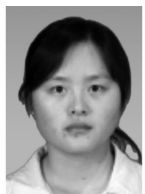
[18] Fortunato S, Latora V, Marchor IM. A method to find community structures based on information centrality [J]. Physical Review E, 2004, 70(5): 56-104



Wang Hao, born in 1962. PhD. Professor Senior member of China Computer Federation. His main research interests include artificial intelligence.



Yao Hongliang, born in 1972. PhD and assistant professor. Member of China Computer Federation. His main research interests include artificial intelligence and knowledge engineer.



Li Guohuan, born in 1988. Master candidate. Her main research interests include artificial intelligence and knowledge engineer.



Li Junzhao, born in 1975. Received his MSc degree in the University of Chinese Academy of Sciences. Currently PhD candidate in computer science. Lecturer of Hefei University of Technology. His research interests include machine learning and artificial intelligence.