申万一级行业指数长短期波动性 APRIORI 算法实证分析

摘要

本文主要选取申万一级行业 28 个行业股票指数，对其波动性进行了关联规则分析。主要采用 Apriori 算法发现满足最小支持度和置信度阈值的频繁项集和关联规则，并对最终的结果进行解释分析。本文选取 2018 年 1 月 1 日至 2020 年 3 月 27 日作为长期；2020 年 1 月 26 日到 2020 年 3 月 27 日（新冠肺炎疫情期间）作为短期分别进行分析，研究发现在长期中，随着机械设备指数下降，建筑装饰指数有 90.98%的概率下降；而当建筑指数上升，机械设备指数有 90.28%的概率上升。而短期行业指数之间的轮动以机械设备、传媒指数、化工、计算机、建筑装饰、电气设备、轻工制造和汽车指数为核心，传媒指数、机械设备指数表现出最强的联动性。

关键词：关联规则分析 Apriori 频繁项集

一、研究内容

股票数据的分析与预测具有重大的理论意义与应用价值；而关联规则作为数据挖掘领域一个非常重要的研究课题，目前已在零售业、电信、股票分析等领域得到了广泛应用。本文在关联规则算法的基础上，将其应用于股票数据分析。

本文主要对申万一级行业 28 个行业股票指数的波动性进行关联分析，使用 Apriori 算法发现满足最小支持度和置信度阈值的频繁项集和关联规则，分别选取 2018 年 1 月 1 日至 2020 年 3 月 27 日作为长期；2020 年 1 月 26 日到 2020 年 3 月 27 日（新冠肺炎疫情期间）作为短期，并对最终的结果进行解释分析。

二、方法介绍

（一）关联规则

定义 1 设$I = \{i_1, i_2, \ldots i_m\}$为不同项的一个集合，$i_j(1 \le j \le m)$表示项集中的第 j 个项。设任何相关的数据集 D 是数据库事务的集合，其中每一个事务 T 是一组项的集合，使得$T \subseteq I$。每个事务都与唯一的标识符 TID 相联。设 A 是一个项集，事务 T 包含 A 当且仅当$A \subseteq T$。一条关联规则就是一个形如$A \Rightarrow B$的蕴含式，其中$A \subseteq I, B \subseteq I, A \cap B = \emptyset$.

关联规则$A \Rightarrow B$成立的条件为：

（1）它具有支持度 S%，S%是事务集 D 中包含$A \cup B$的百分比，S%=support（$A \Rightarrow B$）= $P(A \cup B)$，即数据库 D 中至少有 S%的记录包含$A \cup B$。

（2）它具有可信度 C%，C%是事务集 D 中包含事务 A 同时包含事务 B 的百分比，C% = P(B|A) = $P(A \cup B)/P(A)$，即在数据库 D 中包含了 A 记录的同时至少有 C%也包含 B。

支持度定义了项在整个数据库中所占的比例，置信度定义了该规则的强度，满足最小支持度 min_sup 的项目集合称为频繁项集，同时满足最小支持度 min_sup 和可信度 min_coef 阈值的规则称为强关联规则，习惯上将强关联规则表示为$A \Rightarrow B(S\%, C\%)$.

（二）Apriori 算法

Apriori 算法是一种最有影响力的挖掘布尔关联规则频繁项集的算法。算法的名字基于这样的事实：算法使用频繁项集性质的先验知识。Apriori 使用一种称为逐层搜索的迭代方法。K-项集用于搜索(k+1)-项集。首先，找出频繁 1-项集的集合，该集合记为l_1，l_1用于找频繁 2-项集的集合l_2，而l_2用于找l_3.如此下去，直到不能找到频繁 k-项集。

三、分析过程

（一）数据选取及处理

从 Wind 数据库中下载申万一级行业指数日涨跌数据，具体包括['农林牧渔 1', '采掘 2', '化工 3', '钢铁 4', '有色金属 5', '电子 6', '家用电器 7', '食品饮料 8', '纺织服装 9', '轻工制造 10', '医药生物 11', '公用事业 12', '交通运输 13', '房地产 14', '商业贸易 15', '休闲服务 16', '综合 17', '建筑材料 18', '建筑装饰 19', '电气设备 20', '国防军工 21', '计算机 22', '传媒 23', '通信 24', '银行 25', '非银金融 26', '汽车 27', '机械设备 28']这 28 个行业指数。而后对涨跌状态进行预处理，用一个整数代表该指数涨跌状态。若该指数当日上涨末位设为 1，当日下跌末位设为 2，第一到第二位代表对应行业指数，如 231 代表传媒行业指数当日上涨。

（二）长期股票指数关联性分析

为了分析股票指数波动的长期关联性，本文选取了 2018 年 1 月 1 日到 2020 年 3 月 27 日间的长期数据，借助 Python 软件进行操作，为了更加全面把握股票指数的波动规律，保证关联规则的支持度，我们提取置信度大于 90%及支持度大于 40%的长期数据的关联规则，结果如表 1 所示。

表 1 长期股票指数的关联性分析结果

前项	后项	支持度	置信度
机械设备指数下降	建筑装饰指数下降	44.57%	90.98%
建筑装饰指数上升	机械设备指数上升	41.07%	90.28%

机械设备指数上升	化工指数上升	46.04%	90.25%

（三）短期股票指数的关联性分析

为了分析股票指数波动的短期效应，我们选取 2020 年 1 月 26 日到 2020 年 3 月 27 日（新冠肺炎疫情期间）作为短期，对期间的数据进行分析，取置信度大于 90%及支持度大于 50%的短期数据的关联规则，结果如表 2 所示。

表 2 短期的指数关联性分析结果

前项	后项	支持度	置信度
传媒指数上升	机械设备指数上升	53.33%	96.97%
公用事业指数上升	建筑装饰指数上升	50.00%	96.77%
通信指数上升	计算机指数上升	55.00%	94.29%
电气设备指数上升	计算机指数上升	53.33%	94.12%
化工指数上升	机械设备指数上升	53.33%	94.12%
传媒指数上升	化工指数上升	51.67%	93.94%
综合指数上升	化工指数上升	50.00%	93.75%
机械设备指数上升	计算机指数上升	53.33%	91.43%
机械设备指数上升	传媒指数上升	53.33%	91.43%
机械设备指数上升	化工指数上升	53.33%	91.43%
轻工制造指数上升	机械设备指数上升	51.67%	91.18%
化工指数上升	传媒指数上升	51.67%	91.18%
建筑装饰指数上升	机械设备指数上升	51.67%	91.18%
电气设备指数上升	机械设备指数上升	51.67%	91.18%
建筑装饰指数上升	轻工制造指数上升	51.67%	91.18%
轻工制造指数上升	建筑装饰指数上升	51.67%	91.18%
传媒指数上升	电气设备指数上升	50.00%	90.91%
传媒指数上升	机械设备指数上升	50.00%	90.91%
商业贸易指数上升	机械设备指数上升	50.00%	90.91%
汽车指数上升	通信指数上升	50.00%	90.91%
汽车指数上升	计算机指数上升	50.00%	90.91%
汽车指数上升	机械设备指数上升	50.00%	90.91%
传媒指数上升	计算机指数上升	50.00%	90.91%
传媒指数上升	电气设备指数上升	50.00%	90.91%
传媒指数上升	机械设备指数上升	50.00%	90.91%

三、实证结果

（一）长期股票指数关联分析结果

在2018年1月到2020年3月这一时间段中，机械设备指数下降时，有90.98%的把握建筑装饰指数会下降。而建筑装饰指数上升时，机械设备指数有 90.28%的把握会上升。从产业链的角度，机械设备是建筑装饰的上游产业，当建筑装饰处于上涨行情时，会对机械设备产生更大的需求，以此带动机械设备行情走高。而机械设备指数下跌时，反映出建筑装饰业对机械设备的需求下降，由此建筑装饰也同时进入下跌行情。同样也可以得出机械设备行业的繁荣景象可以促进化工产业的发展。

（二）短期股票指数联动性分析结果

短期行业指数之间的轮动以机械设备、传媒指数、化工、计算机、建筑装饰、电气设备、轻工制造和汽车指数为核心，特别是机械设备和传媒指数在表 2 中分别出现了 12 次和 9 次，占到所有满足最小置信度和支持度的所有项集的 18%和24%，说明这8个板块特别是机械设备以及传媒板块和其余板块有着密切的联系。通过它们之间的联动性进行板块组合分析，通常可以在某一板块出现上涨或下跌情况下，预测出与之相关联的板块的涨跌趋势，从而进行风险管理，来达到躲避风险或谋求利润的目的。对满足最小阈值条件的项集中的各指数出现次数及次数占比统计如表 3 所示。

表 3 短期中满足最小阈值条件的指数占比统计表

指数名称	次数	占比
机械设备	12	24%
传媒	9	18%
化工	5	10%
计算机	5	10%
建筑装饰	4	8%
电气设备	4	8%
轻工制造	3	6%
汽车	3	6%
通信	2	4%
公用事业	1	2%
商业贸易	1	2%
综合	1	2%
合计	50	100%

在短期内表现出传媒指数、机械设备指数之间联动性关系最强，在传媒指数上涨时，有 96.97%的把握机械设备指数会上涨。其次，在公用事业指数上升时有 96.77%的把握建筑装饰指数会上涨。也就是说，在短期板块联动过程中，它们的联动效应是显著的。

在短期分析中，所有满足阈值条件的指数状态都是当日上涨，这说明在上升行情下股票指数的联动性显著高于下跌行情下的联动性。