Computer Science and Application 计算机科学与应用, 2018, 8(6), 894-901
Published Online June 2018 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2018.86099

Hans汉斯

The Behavior Analysis of Stock Analysts Based on K-Prototypes Clustering Algorithm

Xiaomei Zhang, Diankai Hu

The School of Information Control, Qingdao University of Technology, Qingdao Shandong
Email: 772653711@qq.com, 843132974@qq.com

Received: Jun. 5th, 2018; accepted: Jun. 20th, 2018; published: Jun. 27th, 2018

Abstract

As the information intermediary, stock analysts provide information about the inner investment value of the stock by publishing research report and their behavior is more and more concerned by the investors. Because of the large number of stock analysts, the different style and quality of research report, the investors lacking the relevant knowledge experience have difficulty in choosing the research reports which are suitable for their own preference. This paper uses the K-prototypes clustering algorithm to analyze the behavior of stock analysts with mixed attributes, which solves the large and the dispersive characteristics of the stock analysts' group. By depicting the characteristics of different stock analysts' group, investors can know the analysts' group better to obtain more valuable data information and make investment rationally to reduce the risk of investment. The results can also provide the data basis for the follow-up multivariate analysis.

Keywords

Clustering, K-Prototypes Algorithm, Stock Analyst, Research Report

基于K-Prototypes聚类算法的股票分析师行为划分

张晓妹，胡殿凯

青岛理工大学信息与控制工程学院，山东 青岛
Email: 772653711@qq.com, 843132974@qq.com

收稿日期：2018年6月5日；录用日期：2018年6月20日；发布日期：2018年6月27日

文章引用: 张晓妹, 胡殿凯. 基于 K-Prototypes 聚类算法的股票分析师行为划分[J]. 计算机科学与应用, 2018, 8(6): 894-901. DOI: 10.12677/csa.2018.86099

摘　要

股票分析师作为信息中介，通过发布研报的形式提供股票内在投资价值的信息，其行为越发受到广大投资者的关注。由于股票分析师数量众多、研报风格迥异、质量良莠不齐，投资者缺乏相关知识经验难以去选择适合自身偏好的分析师研报。本文利用K-prototypes聚类算法分析具有混合属性的股票分析师行为数据，解决了股票分析师群体数据量大且分散的特性。通过刻画不同股票分析师群体的特征，帮助投资者了解分析师群体获取更多有价值的数据信息，进行理性投资降低投资风险，同时其结果为后续的多元分析提供数据基础。

关键词

聚类，K-prototypes算法，股票分析师，研究报告

1. 引言

金融行业发展至今，已从先前的传统金融机构模式，发展成为传统及各类新兴金融形态共存的态势。大数据技术在金融领域的广泛应用，为金融行业提升服务效率、规避投资风险、降低交易成本做出巨大贡献[1]。证券作为金融行业的重要组成部分，其金融产品在国民经济中也占有越来越高的比重，然而，经济活动瞬息万变，股票投资的风险日益增强[2]。股票分析师行业应运而生，在把握当前经济大环境的基础之上，将通过各种渠道收集到的上市公司信息整合成研究报告，内容包括股票相对走势、基本数据、投资优劣、风险提示等，并同时对该股票未来发展趋势作出投资评级、阶段性价位等投资性建议[3]。截至 2017 年 1 月 20 日，沪深两个主板市场 3092 只股票在交易，而股票分析师平均每年发布的研究报告超过 2 万份，其降低了信息的不对称性，成为主要机构投资者和部分散户进行投资决策的重要依据，在一定程度上左右着投资者的行为[4]。

然而，股票分析师行业入职门槛低，研究报告数量众多、风格迥异，没有受过专业训练的投资者，难以从众多研究报告中找到适合自身投资偏好的研报。更甚至有些股票分析师与上市公司存在内部联系，通过大力捧吹提升购买力度，从而获取额外收益，使得过于依赖研报的投资者产生盲目购买行为。因此，本文依托金融大数据的优势，利用数据挖掘技术从海量数据中提取能够提供投资价值的数据信息。通过聚类分析算法对股票分析师行为数据进行划分，充分刻画各股票分析师群体特征，得到具有指导投资意义的股票分析师分类，便于投资者根据需求从合适的分类中选择股票分析师，解决信息的不对称性，满足投资者的迫切需要。

2. 股票分析师行为分析的现状与发展

现阶段，随着资本主义市场的不断成熟，我国股票市场逐渐朝着规范化和成熟化的方向发展。但是影响股票市场稳定性的因素众多，股市涨跌无常，投资行为风险系数较大。投资者要想通过投资行为在股市中得到可观的收益，需要对上市公司的发展前景、历史业绩等相关财务情况有充分的了解，并在此

基础上合理的判断其股票价值。

国内许多学者围绕获取股票投资价值的目标提出许多研究方案。李毅以石油股票为例，利用模糊聚类算法对股票进行分类，评估上市公司的优劣[5]；谢桂标以金融业上市公司为研究对象，选取部分重要的财务指标进行聚类，结合因子得分客观评价每类上市公司的综合情况[6]；唐朝红采用社区发现算法根据股票波动相关性对股票进行聚类分析，找出涨跌幅度趋同性强的股票集合[7]；海沫将财务指标作为股票投资价值的衡量指标，采用谷本距离度量方式下的 K-means 算法对股票进行聚类分析，划分股票板块[8]；宋宗香利用模糊 C-均值聚类算法完成股票的选取分类，应用于股票投资分析中，为投资者提供可以区分并选择适合自己投资风格的股票分类参考[9]。从以上的研究中可以看出，目前国内大多数学者的研究分析仅针对于上市公司以及股票本身数据去做聚类，通过划分股票板块或行业公司分类来分析股票的购买价值。其中并没有考虑到，股票分类数据对于缺乏相关知识经验的投资者来说仅仅是一堆分类数据，各分类数据的股票特性并不足以支持投资者做出投资行为。

本文的数据基础为股票分析师行为，即股票分析师所发布研究报告情况。股票分析师通过各种渠道对上市公司的信息收集，针对某支股票的发展前景、历史业绩、投资风险等相关情况做出基本介绍，对未来长短期发展趋势作出投资性建议，所以研究报告所蕴含的数据价值更加可观。由于股票研究报告数据量大，且数据属性为数值型、符号型以及序数型等多数据类型的混合属性，选取用于处理混合数据类型的 K-prototypes 算法对股票分析师行为数据进行聚类分析。帮助投资者更加直观地了解分析师群体，根据自身需求选择并查阅股票分析师所发布的研究报告，获取更多的股票投资建议，做出理性的投资行为。同时为后续按照投资者自身持仓与投资偏好推荐分析师、寻求股票分析师行为与股票走势之间关联等多元分析提供数据基础。

3. K-Prototypes 算法在股票分析师行为分析中的运用

3.1. K-Prototypes 算法

K-prototypes 算法是由 K-Modes 算法和 K-Means 算法融合形成的用于对混合型数据集进行聚类分析的算法。在医学、社会学、生物学以及经济学等各现实领域中都存在着大量的混合数据集，这使得的 K-prototypes 算法的广泛应用成为必然。

K-prototypes 算法计算数据对象之间相异度的定义如下[10]：

定义 3.1：设 C_i 是聚类算法过程中的一个簇，其中 Z_i 是 C_i 的簇中心，对象 $X_j \in X$ 是由数值型与非数值型两部分组成，通过参数 β 来控制数值属性与分类属性的权重，则对象 X_j 与簇中心 Z_i 的相异度计算定义如下：

$$d\left(X_j, Z_i\right) = d_m\left(X_j, Z_i\right) + \beta d_n\left(X_j, Z_i\right) \tag{1}$$

d_m 和 d_n 分别表示对象与簇中心在数值型和非数值型属性描述下的相异度。其中 d_m 表示欧氏距离的平方，如下所示：

$$d_m\left(X_j, Z_i\right) = \sum_{l=1}^{P}\left(X_{jl} - Z_{il}\right)^2 \tag{2}$$

d_n 表示简单匹配相异度，定义为：

$$d_n\left(X_j, Z_i\right) = \sum_{l=p+1}^{q} \delta\left(X_{jl}, Z_{il}\right) \tag{3}$$

其中

$$\delta\left(X_{jl},Z_{il}\right)=\begin{cases}1, X_{jl}\neq Z_{il}\\ 0, X_{jl}=Z_{il}\end{cases} \tag{4}$$

K-prototypes 算法在计算数值型与非数值型属性的各簇中心时，分别使用均值和 Modes 法，具体的簇中心计算定义如下：

定义 1.2：设 C_i 是聚类算法过程中的一个簇，其中 Z_i 是 C_i 的簇中心，数值型与非数值型属性的簇中心 $Z_i^m=\left\{z_i^{m1},z_i^{m2},\cdots,z_i^{mp}\right\}$ 和 $Z_i^n=\left\{z_i^{n(p+1)},z_i^{n(p+2)},\cdots,z_i^{n(q)}\right\}$ 分别表示为[10]：

$$Z_i^{mj}=\frac{\sum_{w=1}^{|C_i|}C_i^{jw}}{|C_i|},\quad \left(1\le j\le p\right) \tag{5}$$

其中分母 $|C_i|$ 表示 C_i 类中的对象的个数，分子表示 C_i 类中的所有的对象在第 i 个数值属性上的属性值的和。$Z_i^{nj}\left(p+1\le j\le q\right)$ 表示类 C_i 中第 j 个属性中出现频率最高的属性值。

以下是 K-prototypes 算法的具体流程：

输入：数据集 $X=\{X_1,X_2,\cdots,X_n\}$ 类别数 k 个；

输出：k 个互不相交的簇；

1) 从数据集 X 中随机选取 k 个对象作为初始簇中心；

2) 根据定义 1.1 给出的距离度量公式分别计算对象与各簇中心之间的相异度，并将其分配至距离最近的簇中；

3) 根据定义 1.2 更新聚类的各簇中心点，其中数值属性部分通过计算同簇中属性的均值获得，非数值型属性部分则选取统计所得的各属性值的频率高值；

4) 重复 2)、3)步直到各簇中心不再发生变化为止。

3.2. 数据采集及预处理

本文收集的数据包括股票分析师所发布的研报、各股日收盘价、公司行业等级划分等，数据来源的广泛性造成数据冗余和结构不一性。对收集到的相关数据进行清理、变换、集成等预处理操作，摈弃数据中与本文挖掘目标关联度较低的数据，为后续聚类分析算法提供干净、准确、更有针对性的数据。

股票分析师的行为主要在其所发布的研报中体现，在数据采集、预处理的基础上建立股票分析师研报数据库，收集研究报告约 80 万份，利用股票分析师对分析股的评级情况、分析股所属的行业、分析股在各个阶段的收益情况、评级与股票后期走势是否相符等指标对分析师的单次行为进行描述，见表 1。

基于分析师的单次行为评价指标，汇总建立各股票分析师行为指标数据库，见表 2。

选取分析师行为数据库中部分主要指标，包括：公司级别、擅长领域、研报份数、收益率以及准确率等指标，汇总分析师数据约 1600 余条，作为本次聚类分析的实验数据，见图 1。

3.3. 实验与结果分析

采用聚类客观性评价指标紧密性和间隔性来分析聚类结果。紧密性(*CP*)代表每一个类内各点到聚类中心的平均距离，值越低越好，越低意味着类内聚类距离越近。间隔性(*SP*)代表各聚类中心两两之间的平均距离，值越高越好，越高意味类间聚类距离越远。

其中，紧密性与间隔性计算公式如下：

$$\overline{CP_n}=\frac{1}{|C_n|}\sum_{X_i\in C_n}\left\|X_i-Z_i\right\|$$

id	author	company	level	com	time	count	rate	d_avepri	m_avepri	y_avepri
1	蔡宇杰	安信证券	2	采掘业	2017-09-07	6	16.67	22.4	90.1	0.0
7	宋红欣	川财证券	3	房地产	2017-12-31	16	12.5	26.4	70.3	0.0
14	李雯婧	红塔证券	2	公用事业	2017-07-10	7	85.71	7.6	55.6	51.4
16	杨柳	中信建投	1	金融、保险业	2017-09-05	23	13.04	5.2	52.6	96.8
22	敖翀	中信证券	1	采掘业	2017-12-14	31	6.45	12.7	45.2	0.0
30	曾忠云	中航证券	3	采掘业	2017-10-27	13	23.08	7.7	34.7	0.0
35	蒋一樊	招商证券	1	制造业	2017-10-30	10	30	11.8	29.3	12.3
38	何方	汇丰前海证券	4	制造业	2017-02-09	6	16.67	8.0	26.9	0.0
39	李振	长江证券	2	信息技术业	2016-06-06	6	100	25.7	26.0	26.8

Figure 1. Example of stock analyst data
图 1. 股票分析师数据示例

Table 1. Data indicators of stock analysts' research and report
表 1. 股票分析师研究报告数据指标

指标名称	备注
股票名称	研究股名称
作者	研究报告的作者
评级	研究股的推荐程度，表明股票分析师对该分析股所持有购买的态度，统一划分为五个类别：买入|增持|中性|减持|卖出
发布日期	研究股发布日期
股票行业	研究股所属的行业范畴，参照证监会行业划分标准
短期收益	三个月内收益率，结合股票日收盘价，按照收益率公式计算
中期收益	六个月内收益率，结合股票日收盘价，按照收益率公式计算
长期收益	十二个月内收益率，结合股票日收盘价，按照收益率公式计算
是否准确	研究股六个月内，五个指标买入|增持|中性|减持|卖出分别相对沪深 300 指数涨幅在 15%+|5%~15%|−5%~5%|−15%~−5%|−15%−之间来判断，分为是|否两类

Table 2. Data indicators of stock analysts
表 2. 股票分析师数据指标

指标名称	备注
分析师	股票分析师姓名
公司级别	证监会划定的公司级别，公司级别与股票分析师获取的内部数据资源成正比，间接地通过证券公司在行业内风险管理能力及合规管理水平来反映在职分析师的水平
主占行业	股票分析师评级所擅长行业
同行业排名	股票分析师在同行业内的排名情况
工作年限	股票分析师的工作时长
明星分析师	新财富公布的明星分析师
研报数	发布评级报告的数量，量化评级报告情况反映其活跃程度
个股数	股票分析师所关注的股票个数
短期收益	发布评级报告的三个月内平均收益率
中期收益	发布评级报告的六个月内平均收益率
长期收益	发布评级报告的十二个月内平均收益率，通过三个阶段收益对比反映其所擅长的收益时长
准确率	反映发布的评级报告与后期走势是否一致的准确率指标，直接反应其可信程度

$$\overline{CP}=\frac{1}{k}\sum_{n=1}^{k}\overline{CP_n}$$

$$\overline{SP}=\frac{2}{k^2-k}\sum_{i=1}^{k}\sum_{j=i+1}^{k}\left\|w_i-w_j\right\|_2$$

选定聚类个数 $k = 5, 10, 15, 20$ 对分析师数据进行聚类分析，每个 k 值进行 50 次聚类，*CP* 值与 *SP* 值为 50 次聚类结果所求得的平均值，见图 2，图 3。由 *CP* 曲线图可知，随着 k 值的增大，各类内点到聚类中心的距离逐渐减小；由 *SP* 曲线图可知，随着 k 值的增大，各类中心点之间的距离逐渐增大。结合两个聚类评价指标可知，随着聚类中心数的增加，对股票分析师行为数据进行聚类效果的也就越好。

选定聚类中心个数值 $k = 10$ 时进行主观性聚类分析，各聚类中心点见表 3。

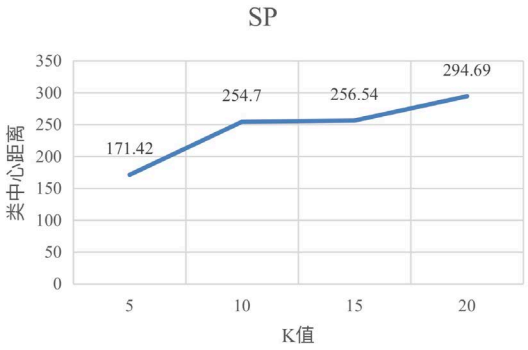

Figure 2. Line chart: The compactness of the result of clustering
图 2. 聚类效果的紧密性折线

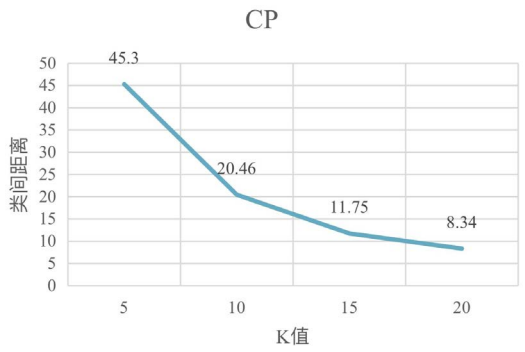

Figure 3. Line chart: The separation of the result of clustering
图 3. 聚类效果的间隔性折线

Table 3. Data of center points of clustering
表 3. 聚类中心点数据

标记	个数	等级	行业	数量	准确率	短期收益	中期收益	长期收益
1	51	2	制造业	12	0.79%	9.29	0.58	−0.66
2	98	1	金融保险业	144	33.66%	3.32	5.32	7.74
3	77	1	采掘业	256	41.67%	5.55	11.45	19.49
4	392	1	制造业	308	38.6%	3.89	6.77	13.00
5	181	2	制造业	31	49.35%	7.73	15.65	24.26
6	132	1	房地产	308	25.76%	2.52	3.00	5.94
7	265	2	信息技术业	211	31.86%	4.00	5.38	8.64
8	269	3	制造业	180	38.66%	4.27	8.08	15.28
9	87	1	信息技术业	39	49.12%	3.27	7.82	11.88
10	63	2	制造业	24	19.67%	−4.43	−9.23	−15.05

第一种类型的股票分析师，所在供职单位虽为 AA 级别，但是发布研报数量极少，活跃程度不高，其准确率指标更是低迷，接近于每次评级的结果与股票实际走势均不相符，唯一值得肯定的是其发布研报在三个月内的收益率指标。所以针对此类型的分析师，对制造业领域股感兴趣的投资者可以参考其所发布研报做出短期投资。

第二种类型的股票分析师，供职单位为 AAA 级别的金融保险业领域股，股票分析师活跃程度中等。针对此类型的分析师，对金融保险业领域股感兴趣的投资者可以对分析股做出理性投资，其各阶段的收益回报程度差别不大。

第三种类型的股票分析师，供职单位为 AAA 级别的采掘业领域股，股票分析师非常活跃，发布大批量的研报，研报准确率也近乎为 50%，其中期收益与长期收益水平要高于短期收益水平。针对此类型的分析师，投资者可以参考其研报购买意向对分析股进行六个月及以上的投资。

第四种类型与第五种类型的股票分析师相比，在同为制造业领域中，供职单位为较高级别，股票分析师活跃程度也非常高，但是研报准确率以及各阶段的收益水平均低于第五类别分析师，存在的原因可能是该类别的股票分析师重视数量但是没有考虑其质量，也或许是因为该类别中的数据量较大，对实验结果产生了误差，后期可以考虑单独对该簇数据进行聚类。第四类与第五类分析师相比，投资者可以重点关注第五类的分析师做出投资行为，各阶段的收益情况都相对比较理想。

第六种类型的股票分析师，供职单位为 AAA 级别的房地产领域股，股票分析师非常活跃，发布研报数量众多，其研报准确率仅为 25%，与研报发布数量相比，其收益情况不是很理想。针对此类型的分析师，投资者要谨慎参考研报，必要的话可以进行长期投资。

第七种类型与第九种类型的股票分析师相比，在同为信息技术业领域中，其供职单位为级别略低，股票分析师活跃程度虽然高，但是从准确率以及各阶段的收益水平均来看，其研报质量低于第九类别分析师，第九类别分析师虽然发布的研报数量不多，但是正确率近 50%，中长期收益程度都很可观。第七类与第九类分析师相比，投资者可以重点关注第七类分析师的短期股以及第九类分析师的中长期股做出投资行为。

第八种类型与第十种类型的股票分析师相比，在同为制造业领域的中小型公供职机构中，股票分析师活跃程度不高，但是第八类的小型供职单位的分析师所发布的研报准确程度较好，各阶段的收益情况相比较算理想，尤其是长期收益，但是第十类股票分析师的各阶段的收益情况均为负值。投资者可以关注第八类分析师的中长期股，但是对于第十类分析师的推荐股的购买要十分慎重。

综上所述，通过聚类分析的主观与客观评价可知，采用 K-prototypes 算法对股票分析师行为数据进行聚类分析，能够解决股票分析师群体数据分散的特性，通过数据来刻画各股票分析师群体特性，帮助投资者更加直观的了解各群体的特性，按照自身持仓以及投资喜好挑选适合的分析师，并查看其单次发表研报的行为，对分析股有更深层次的了解，从而进行理性投资行为。

4. 结束语

近年来随着投资者投资热情的不断高涨，对于缺乏投资经验的投资者来说仅依靠投机性进行非理性投资，严重加剧了股票投资风险。本文充分利用大数据的优势，将数据挖掘技术运用到金融科技中，降低金融投资风险。通过收集股票分析师研报数据，采取 K-prototypes 聚类算法对股票分析师行为进行分析，用数据的形式刻画股票分析师行为，有利于投资者辨别股票分析师的可信度与匹配度。后续可以在本文数据分析的基础上，进行投资者个性化推荐分析师、寻求股票分析师行为与股票走势之间关联程度等多元数据分析操作，充分体现数据价值，降低投资风险。

参考文献

[1] 刘钰. 大数据在金融风险管理中的应用研究[J]. 信息系统工程, 2018(4): 81-82.
[2] 刘苇. 分析师评价对股票预测价值的差异性研究[D]. 大连理工大学, 2016.
[3] 燕麟. 分析师跟踪与机构投资者持股的相互关系[J]. 浙江金融, 2016(9): 36-44.
[4] 汤泓. 中国股票分析师报告有效性研究[D]. 南京大学, 2017.
[5] 李毅. 模糊聚类在股票投资中的应用[J]. 现代计算机(专业版), 2014(28): 3-8.
[6] 谢桂标, 许姣丽. 因子分析和聚类分析在金融业股票投资中的应用[J]. 沿海企业与科技, 2016(4): 11-16.
[7] 唐朝红. 面向金融知识服务的股票聚类分析[D]. 哈尔滨工业大学, 2017.
[8] 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, 1(4): 9-17.
[9] 宋宗香. 模糊 C-均值聚类在股票投资中的应用[D]. 东北石油大学, 2017.
[10] 余文利, 余建军, 方建文. 混合属性数据 k-prototypes 聚类算法[J]. 计算机系统应用, 2015, 24(6): 168-172.