Computer Science and Application 计算机科学与应用, 2020, 10(12), 2447-2456
Published Online December 2020 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2020.1012260

Hans汉斯

基于因子分析和K-means聚类算法的行业聚类研究

曹　钰[1,2]，何国辉[1,2]，谭钜源[1,2]

[1]五邑大学智能制造学部，广东　江门
[2]江门市智能数据分析与应用工程技术研究中心，广东　江门
Email: 1986517165@qq.com

收稿日期：2020年11月29日；录用日期：2020年12月24日；发布日期：2020年12月31日

摘　要

工商登记信息中的企业经营范围记录了企业主要从事的生产经营活动，是反映企业所属行业类别的重要标准。对企业进行行业聚类，不仅方便国家管理企业，且有利于企业自身定位，顺应国家趋势发展经济。本文采用基于因子分析和K-means聚类算法，以国家发布的《国民经济行业分类》为标准文本，对企业经营字段样本进行行业聚类分析。首先通过因子分析算法得到K-means聚类的最佳聚类个数，然后通过K-means算法，对企业经营范围进行聚类分析，得到每个企业的所属行业类别，最终通过人工评价和戴维森堡丁指数(DBI)评价聚类结果，证明方法的有效性。

关键词

企业经营范围，行业聚类，因子分析，K-means聚类

Research on Industry Clustering Based on Factor Analysis and K-Means Clustering Algorithm

Yu Cao[1,2], Guohui He[1,2], Juyuan Tan[1,2]

[1]Department of Intelligent Manufacturing, Wuyi University, Jiangmen Guangdong
[2]Jiangmen Intelligent Data Analysis and Application Engineering Technology Research Center, Jiangmen Guangdong
Email: 1986517165@qq.com

Received: Nov. 29th, 2020; accepted: Dec. 24th, 2020; published: Dec. 31st, 2020

Abstract

The business scope of the enterprise in the industrial and commercial registration information records the main production and operation activities of the enterprise, which is an important

文章引用: 曹钰, 何国辉, 谭钜源. 基于因子分析和 K-means 聚类算法的行业聚类研究[J]. 计算机科学与应用, 2020, 10(12): 2447-2456. DOI: 10.12677/csa.2020.1012260

standard to reflect the industry category of the enterprise. Industry clustering is not only convenient for the state to manage enterprises, but also conducive to the positioning of enterprises and the development of economy in line with the national trend. In this paper, based on factor analysis and K-means clustering algorithm, and taking the national economic industry classification as the standard text, this paper conducts industry cluster analysis on enterprise business field samples. Firstly, the optimal number of K-means clustering is obtained by factor analysis algorithm, and then the business scope of enterprises is clustered by K-means algorithm, and the industry category of each enterprise is obtained. Finally, the clustering results are evaluated by artificial evaluation and Davies Bouldin index (DBI) to prove the effectiveness of the method.

Keywords

Business Scope, Industry Clustering, Factor Analysis, K-Means Clustering

1. 引言

在企业进行工商信息登记时通常要记录企业名称、企业经营方式和经营范围等信息，通过经营范围可大致了解企业所属的行业类别。但是对于有的企业来说，由于经营范围很广、经营种类比较繁杂，填写时不够严谨，导致仅仅从经营范围的文字中很难分辨出企业所属的行业及在每个行业中所占的比重。

为了较准确的获得企业的行业类别信息，方便对企业进行画像，为后期深层次挖掘应用提供数据基础，本文结合国家发布的《国民经济行业分类》，在众多学者研究成果的基础上，围绕企业经营范围内容研究运用基于因子分析和 K-means 聚类算法对企业行业进行聚类分析，最终得到企业所属行业的画像。

具体研究内容包括以下四部分：1) 预处理待聚类文本和标准文本《国民经济行业分类》；2) 通过因子分析计算出标准文本对待聚类文本的贡献率，进而得到最佳聚类数量 K；3) 运用 K-means 聚类算法实现行业聚类；4) 对聚类结果进行评价。

2. 相关工作

文本预处理是文本聚类的重要基础，其主要目的是完善和规范文本表述。本文的预处理过程如下：1) 处理待聚类文本的异常值；2) 对待聚类文本和标准文本同时进行 Jieba 分词，达到“归一化”；3) 构建待聚类文本和标准文本词袋，并对待聚类文本建立 Tfidf 模型；4) 计算待聚类文本和标准文本的相似度，以此构建聚类数据集合。

2.1. 国民经济行业分类

《国民经济行业分类》是指由国家统一制定的，按照生产的同一性，对于一个国家的国民经济的所有生产活动进行生产性质归属性分组而形成的规范标准[1]。具有国家强制推行的标准化特征[1]。具体内容如表 1 所示。

《国民经济行业分类》规定了全社会经济活动的分类与代码，能够满足国家在统计、计划、税收、工商等宏观管理中对经济活动的分类，也可用于信息处理和信息交换。

该标准是一个树形结构，总共分为四层，分别是：

第一层：门类(一级类别)，总共分为 20 个门类，编号为 A~T；

第二层：大类(二级类别)，总共分为 96 个门类，编号为 01~96；

第三层：中类(三级类别)，总共分为 960 个门类，编号为 011~960；

第四层：小类(四级类别)，总共分为 9600 个门类，编号为 0111~9600。

在对《国民经济行业分类》进行停用词处理后，将与待聚类文本共同构建聚类数据集。

Table 1. Classification and code table of national economic sectors
表 1. 国民经济行业分类和代码表

国民经济行业分类					
代码				类别名称	说明
门类	大类	中类	小类		
A				农、林、牧、渔业	本门类包括 01~05 大类
	01			农业	指对各种农作物的种植
		011		谷物种植	指以收获籽实为主，供人类食用的农作物的种植，如稻谷、小麦、玉米等农作物的种植
			0111	稻谷种植	
			0112	小麦种植	

2.2. 待聚类文本预处理

2.2.1. 文本来源及特点

待聚类文本是企业在相关系统中录入的经营范围内容。经营范围是企业从事经营活动的业务范围，同时也是企业所属行业类别的重要参考指标。从某系统中获得的部分企业录入的经营范围文本表示为例，其内容如表 2 所示。

Table 2. Business scope of some enterprises
表 2. 部分企业经营范围

企业注册码	经营范围
44070400000XXXX	销售：食品(凭有效的《食品经营许可证》经营)、日用杂品。
44070040002XXXX	生产经营牛仔服装系列产品。
440783NA000XXXX	城市建设管理咨询、代理；城市环境卫生维护、道路清扫。
440781301XXXX	加工、修理：拉链；五金。
44078200004XXXX	加工：五金。
44070230300XXXX	加工：五金.
194129XXXX	碎石，角石。

由表 2 可以看出，文本存在以下特征：

1) 内容不多：每条信息代表一个企业的经营范围，字数大概在几个字至几百字间。

2) 标点不规范：主要体现为中英文状态下标点符号的不准确使用。

3) 书写不规范：主要体现为“规则组合(如：加工、销售：机制沙)”和“非规则组合(如：城市建设管理咨询、代理；城市环境卫生维护、道路清扫)”。

4) 内容不直接：主要体现为“凭有效的《食品生产许可证》经营”等信息。

2.2.2. 文本异常值处理

聚类算法对异常值异常敏感，所以处理好潜在的异常值会进一步提高数据质量。针对企业经营范围中的文本特点，即异常值，本文预处理流程如图 1 所示。

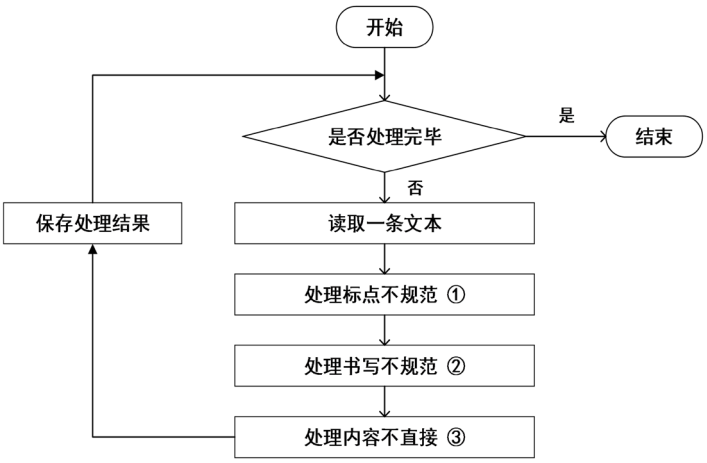

Figure 1. Pretreatment process
图 1. 预处理过程

针对图 1 中① ②的详细处理过程如下：

待聚类文本可根据冒号分为规则组合数据及非规则组合数据：

如果为规则数据，则根据分号进一步划分；

如果没有分号但有冒号(只有冒号)，则对冒号前后的动名词进行组合；

如果既有分号又有冒号，则首先按分号切分并分别保存，然后再对分号处理之后的分段数据按冒号前后的动名词进行组合；

如果为非规则数据，则对无冒号和分号及有分号无冒号两种情况，保留原始数据。

针对图 1 中③的详细处理过程如下：

对于类似于“(凭有效的《食品经营许可证》经营)”这样的信息，放入经营范围证书库中，因篇幅原因本文不再赘述。

经过上述处理后，得到结果如表 3 所示。

Table 3. Pretreatment results
表 3. 预处理结果

索引	原始内容	预处理结果
1	加工、销售：木材、木制品	[‘加工木材’，‘加工木制品’，‘销售木材’，‘销售木制品’]
2	加工：灯饰。	[‘加工灯饰’]
3	生产经营牛仔服装系列产品。	生产经营牛仔服装系列产品
…	…	…

2.3. 中文分词的“归一化”处理

虽然异常值处理后的聚类文本增加了分词效果，但与《国民经济行业分类》还存在差别，因此需要“归一化”处理。

如果只对待聚类文本进行分词，无法完全得到与标准文本相匹配的内容，如“生产经营牛仔服装系列产品”，在《国民经济行业分类》中没有与之相符的内容，在后期聚类中可能无法匹配到对应的行业，所以将对《国民经济行业分类》和待聚类文本同时使用 Jieba 分词，待聚类文本会得到“生产、经营、牛仔、服装、系列”，且《国民经济行业分类》经过分词后也会出现“生产、服装”等字眼，进而可以进行匹配。

另外，可以在一定程度消除自定义词典的语义歧义，即待聚类文本和《国民经济行业分类》对每个词组采用的切分规则是一致的，由此可以避免两个文本语义不同的情况。

2.4. 建立聚类数据集

本文以 1000 条待聚类样本为例，首先对“归一化”后的每个待聚类样本构建词袋模型，如[生产 2，销售，五金 4，服务 0，…]，并建立 Tfidf 模型；然后对《国民经济行业分类》构建词袋模型；最后计算两个文本间的相似度以此构建聚类数据集，其结果如表 4 所示。

Table 4. Clustering data set
表 4. 聚类数据集

索引	企业注册码	01	02	03	…	20	21	…	94	95	96
0	193942XXXX	0	0	0	…	0	0	…	0	0	0
1	193944XXXX	0	0	0	…	0.0112687	0.	…	0.0271559	0	0
2	193953XXXX	0	0	0	…	0.0143782	0.0147046	…	0	0	0
…	…	…	…	…	…	…	…	…	…	…	…

3. 因子分析抽取特征

因子分析是一种从变量群中提取共性因子的统计技术[2]，其本质上是一个降维的过程。因子分析法以变量为研究对象，通过分析变量间的相关性[3]，分析每个变量对信息的贡献度，进而在众多变量中找出最具有代表性的因子，使其具有较强的可解释性[4]。

因子分析模型如式(1)：

$$\begin{aligned} X_1 &= a_{11}F_1 + a_{12}F_2 + \cdots + a_{1n}F_n + e_1 \\ X_2 &= a_{21}F_1 + a_{22}F_2 + \cdots + a_{2n}F_n + e_2 \\ &\vdots \\ X_m &= a_{m1}F_1 + a_{m2}F_2 + \cdots + a_{mn}F_n + e_m \end{aligned} \tag{1}$$

其中，$X=(X_1,X_2,\cdots,X_m)$ 代表变量；

$a=(a_{11},a_{12},\cdots,a_{1n},\cdots,a_{m1},a_{m2},\cdots,a_{mn})$ 代表参数，即变量之间的相关系数，值越大，相关性越大；

$F=(F_1,F_2,F_3,\cdots,F_n)$ 代表公共(共性)因子，简称因子；

$e=(e_1,e_2,e_3,\cdots,e_m)$ 代表特殊因子，是不可直接观测的数据，在分析中一般省略[5]。

3.1. 主成分法提取因子

因子提取的方法有多种，使用最多的是主成分法，此外还有最小二乘法(least squares)、极大似然法(maximum likelihood)等[6]。本文将采用主成分法提取因子。

设观测数据为 m 个 n 维数据，其矩阵格式如式(2)：

$$X=\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{pmatrix} \tag{2}$$

对上述矩阵进行标准化处理：

$$x_{ij}^{*}=\frac{x_{ij}-\overline{x}_j}{\sqrt{\mathrm{var}\left(x_j\right)}}\ \left(i=1,2,3,\cdots,m;j=1,2,\cdots,n\right) \tag{3}$$

其中，对 $\overline{x}_j$ 和 $\mathrm{var}\left(x_j\right)$ 如式(4)和式(5)：

$$\overline{x}_j=\frac{1}{m}\sum_{i=1}^{m}x_{ij}\ \left(i=1,2,3,\cdots,m;j=1,2,\cdots,n\right) \tag{4}$$

$$\mathrm{var}\left(x_j\right)=\frac{1}{m-1}\sum_{i=1}^{m}\left(x_{ij}-\overline{x}_j\right)^2\ \left(j=1,2,3,\cdots,n\right) \tag{5}$$

然后计算相关系数矩阵，其矩阵如式(6)：

$$A=\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix} \tag{6}$$

为表示方便，设原始数据标准化处理后仍用 X 表示，则标准化后的相关系数如式(7)：

$$a_{ij}=\frac{1}{m-1}\sum_{l=1}^{m}x_{li}x_{lj}\ \left(i,j=1,2,3,\cdots,n\right) \tag{7}$$

再用雅克比方法求解相关系数矩阵的特征值和特征向量[7]，最后得到主成分，这里贡献率就是指某个主成分的方差占全部方差的比重[8]，其贡献率 C 表达式如式(8)：

$$C=\frac{g_i}{\sum_{i=1}^{n}g_i} \tag{8}$$

其中，g_i 表示单个主成分。贡献率 C 越大，表明该主成分所包含原始变量的信息越强[9]。主成分个数的选取，通常可以参考累积贡献率和主成分方差两个标准[10]，即一般要求累计贡献率达到 80%以上，且主成分方差尽可能大于 1。

将表 4 中 01 至 96 列的所有数值进行上述计算，且本文预设主成分个数 $K=5$ (可随机设置)，然后根据贡献率和主成分方差两个标准，不断调整，得到 $K=18$ 时，结果比较符合期望。数据如表 5 所示。

Table 5. Variance, proportion and cumulative rate of eigenvalues of correlation coefficient ($P=18$)
表 5. 相关系数特征值的方差、占比及累积率($P=18$)

成分	方差	方差占比	累积贡献度	成分	方差	方差占比	累积贡献度
0	22.32	0.234	0.234	9	1.76	0.019	0.706
1	16.10	0.169	0.403	10	1.60	0.017	0.723
2	7.93	0.083	0.486	11	1.39	0.015	0.738
3	5.93	0.062	0.548	12	1.25	0.013	0.751

Continued

4	3.20	0.034	0.582	13	1.19	0.013	0.764
5	3.12	0.033	0.615	14	1.09	0.011	0.775
6	2.72	0.029	0.644	15	1.07	0.011	0.786
7	2.28	0.024	0.668	16	1.03	0.011	0.797
8	1.81	0.019	0.687	17	1.01	0.011	0.808

3.2. 因子旋转

因子旋转是为了让因子载荷两极分化[11]，要么接近 0，要么接近 1，这样有助于分析因子的属性，便于后期因子命名。常用的方法有：方差最大正交旋转法和斜交旋转法[12]。本文采用方差最大正交旋转法进行因子旋转，其计算公式如式(9)：

$$g_j^2 = \sum_{i=1}^{m} a_{ij}^2 \tag{9}$$

每个公共因子的载荷系数 a_{ij} 平方和就是该公共因子的方差，本文运用 Spider 得到因子的旋转结果如图 2 所示：

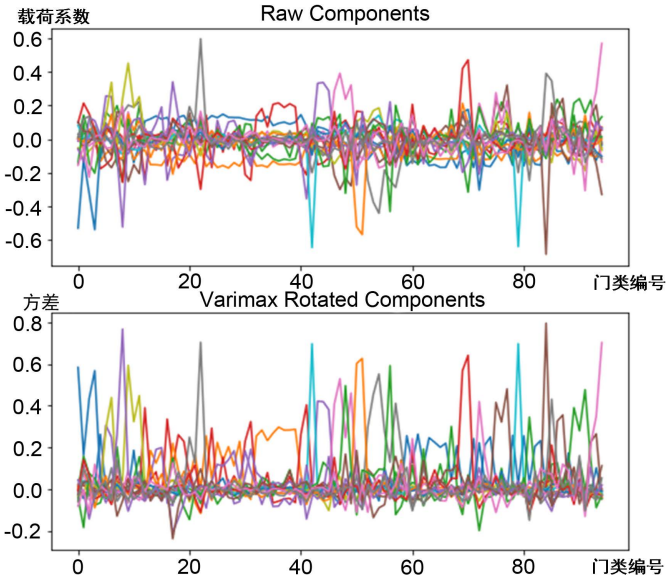

Figure 2. Factor analysis results
图 2. 因子分析结果

图 2 包括 Raw Components 和 Varimax Rotated Components 两部分，各部分含义如下：

① Raw Components (主成分分析图)：该图峰值围绕 0.0 的中轴线上下浮动，每个峰值表示每个主成分在每个变量上的权重值大小，但峰值比较不够明显。

② Varimax Rotated Components (因子旋转图)：该图峰值集中围绕在 0.0 中轴线以上，且每个峰值大的越大，小的越小，可较清楚地看出峰值间的差别。

4. 聚类实现过程与结果

K-means 聚类算法是一个迭代的过程[12]，其具体步骤如下：

1) 在样本中选取 K 个点作为初始质心，即每个 K 代表一个聚类中心；

2) 对每个样本点，本文通过欧式距离计算方式按照距离最近的原则将每个数据点划分到离它最近的聚类中心 K 所对应的类别中[13]；

3) 经过步骤 2)后，形成了 K 个集合，即 K 个类别，然后重新计算每个类别的质心，更新聚类中心的位置；

4) 在 3)中，如果新质心和旧质心间的距离小于某一阈值，则判断达到预期效果，算法终止，否则迭代 2)~3)步骤[14]。

通过聚类得到的数据类别如图 3 所示。

df - DataFrame

索引	企业注册码	原始数据	类别编号	行业类别
0	44060000002XXXX	生产经营牛仔服装系列产品。	7	零售业+纺织服装、服饰业
1	44070000000XXXX	设计、加工、生产、销售：模具、机械设备、金属制品。	7	零售业+金属制品业
2	44070000000XXXX	生产、加工、销售、维修：机械...	7	零售业+金属制品业+金属制品、机械和设备修理业+专用设备制造业
3	44070000001XXXX	生产、销售：麦克风配件，五金配件。	7	零售业+金属制品业
4	44070040004XXXX	生产、销售：五金卫浴、浴室洁具、浴室木柜。	7	零售业+金属制品业
5	44070400000XXXX	生产销售、来料加工：金属家具、铁线工艺品、五金制品。	7	零售业+金属制品业
6	44070040000XXXX	生产、加工、销售：金属制品、游艇高级装饰材料。	7	零售业+金属制品业
7	44070000002XXXX	生产、加工、销售：五金制品、冲压小五金、五金配件。	7	零售业+金属制品业
8	44078500000XXXX	设计、生产、销售：食品机械、...	7	零售业+金属制品业+通用设备制造业+专用设备制造业

格式设置 大小调整 背景颜色 列宽最小/最大化 保存并关闭 关闭

Figure 3. Data category display
图 3. 数据类别展示

图 3 中，每个索引对应一个企业，“原始数据”表示样本原始文本内容，“类别编号”是聚类得到的结果，“行业类别”是通过人工评价得到的企业具体所属的行业类别。

5. 评论与分析

每个企业经营范围包含多个行业类别，因此聚类结果中的每个类别是多个行业类别的融合。故本文对类别结果用 0~17 个数字表示，结果如图 4 所示。

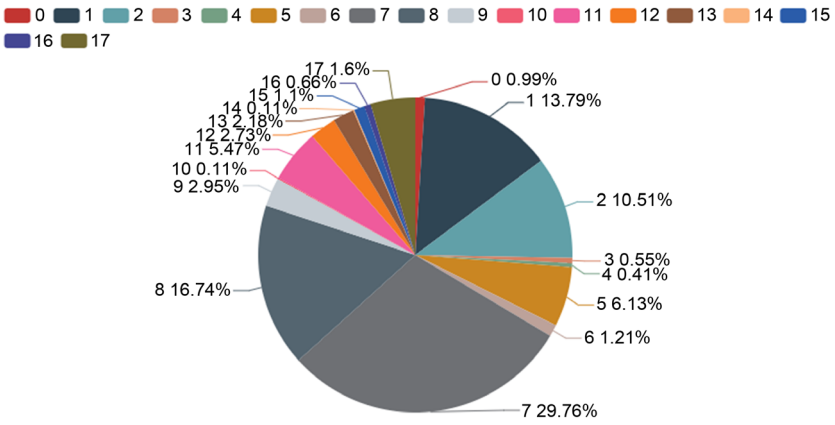

Figure 4. Cluster analysis results
图 4. 聚类分析结果

其中属于类别 7 (销售业和零售业)的类别占比最大，为 29.76%，其次是第 8 类(销售业，金属制品业，橡胶和塑料制品业及电气机械和器材制造业)，占比为 16.74%，

本文对聚类结果的评价方式是内部评价和人工评价[15]。

内部评价是通过某些模型生成聚类的参数，来统一判别聚类效果。本文采用戴维森堡丁指数(DBI)判定聚类结果。戴维森堡丁指数(Davies Bouldin index, DBI)是由大卫 L·戴维斯和唐纳德·堡丁提出的一种评估聚类算法优劣的指标[16]，其公式如式(10)：

$$DBI = \frac{1}{k}\sum_{i=1}^{k}\max_{i\neq j}\left(\frac{avg(S_i)+avg(S_j)}{dist(\omega_i,\omega_j)}\right) \tag{10}$$

其中，$S=\{S_1,S_2,\cdots,S_k\}$ 表示簇。首先计算两个簇 S_i、S_j 各自样本间的平均距离 avg(S)之和及两个簇中心点 ω 之间的距离，然后统计所有簇的相似度的最大值，对其求均值即可得到 DBI 指数。DBI 最小值为 0，DBI 指数越小，相同类别内元素之间距离越小，不同类别间距离越大，代表聚类效果越好[16]。本文的 DBI 指数为 0.21，证明聚类效果较好。

人工评价是指聚类分析的结果与人工评价越接近越好[16]。本文通过小范围的调查验证，对类别结果正确率进行判定。具体数据如表 6 所示。

Table 6. Artificial evaluation by factor analysis
表 6. 因子分析法人工评价

类别	文本数量	正确率	类别	文本数量	正确率
0	9	0.78	9	27	0.85
1	126	0.78	10	1	0
2	96	0.70	11	50	0.74
3	5	0.60	12	25	0.81
4	2	1	13	20	0.70
5	56	0.66	14	1	0
6	11	0.73	15	10	0.90
7	272	0.73	16	6	1
8	151	0.70	17	42	0.83

6. 结束语

本文打破了传统聚类的思路，以《国民经济行业分类》作为标准文本，通过因子分析计算出标准文本对企业经营范围的贡献率，进而得到最佳聚类个数，最后进行聚类分析。

通过评价结果证明本文的聚类结果较好，但由于企业经营范围数据是企业实际经营中填写的数据，在一定程度上会与《国民经济行业分类》有一定的区别，因此后续可以根据本次聚类结果适当的构建词库，进一步提高聚类的准确性。

基金项目

国家自然科学基金项目(No. 61771347)；广东省基础与应用基础研究基金(No. 2019A1515010716)；广东省普通高校基础研究与应用基础研究重点项目(No. 2018KZDXM073)。

参考文献

[1] 陈正伟. 国民经济行业分类及应用[Z]. 重庆: 重庆工商大学, 2014.

[2] 吴娇. 四川省各市州经济综合发展水平比较研究——基于因子分析和 K-means 聚类分析[J]. 知行铜仁, 2019(3): 35-39.

[3] 彭凯, 秦永彬, 许道云. 应用因子分析和 K-MEANS 聚类的客户分群建模[J]. 计算机科学, 2011, 38(5): 154-158, 198.

[4] 黎明, 熊伟. 基于因子分析与聚类分析的化妆品上市公司绩效评价[J]. 财会通讯, 2020(14): 96-99.

[5] 任恒妮. 大数据 K-means 聚类算法的研究与应用[J]. 信息技术, 2019, 43(11): 20-23.

[6] 王春枝. 因子分析中公因子提取方法的比较与选择[J]. 内蒙古财经学院学报(综合版), 2014, 12(1): 90-94.

[7] Martinez-Martin, P., Rojo-Abuín, J.M., Weintraub, D., Chaudhuri, K.R., Rodriguez-Blázquez, C., Rizos, A. and Schrag, A. (2020) Factor Analysis and Clustering of the Movement Disorder Society-Non-Motor Rating Scale. *Movement Disorders*, **35**, No. 6. https://doi.org/10.1002/mds.28002

[8] 韩雪, 张业, 朱聪慧. 企业经营范围文本自动分类方法探究[J]. 标准科学, 2012(1): 93-96.

[9] Martinez-Martin, P., Rojo-Abuín, J.M., Weintraub, D., Chaudhuri, K.R., Rodriguez-Blázquez, C., Rizos, A. and Schrag, A. (2020) Factor Analysis and Clustering of the Movement Disorder Society-Non-Motor Rating Scale. *Movement Disorders*, **35**, 969-975.

[10] Subramaniyam, B.A., Muliyala, K.P., Suchandra, H.H. and Reddi, V.S.K. (2020) Diagnosing Catatonia and Its Dimensions: Cluster Analysis and Factor Solution Using the Bush Francis Catatonia Rating Scale (BFCRS). *Asian Journal of Psychiatry*, **52**, 102002. https://doi.org/10.1016/j.ajp.2020.102002

[11] Wen, F., Du, H., Ding, L., Hu, J., Huang, Z., Huang, H., *et al.* (2020) Clinical Efficacy and Safety of Drug Interventions for Primary and Secondary Prevention of Osteoporotic Fractures in Postmenopausal Women: Network Meta-Analysis Followed by Factor and Cluster Analysis. *PLoS ONE*, **15**, e0234123. https://doi.org/10.1371/journal.pone.0234123

[12] 秦志勇. 安徽省医疗卫生机构服务水平综合评价——基于因子分析和聚类分析方法[J]. 合肥学院学报(综合版), 2020, 37(2): 63-68.

[13] Zhang, Q.H. (2019) Customers Segmentation Based on Factor Analysis and Cluster. *E-Commerce Letters*, **8**, 53-62.

[14] Wang, W. (2017) Stock Evaluation Based on Factor Analysis and Clustering. Chongqing Technology and Business University. In: *Proceedings of* 2017 2*nd International Seminar on Education Innovation and Economic Management* (*SEIEM* 2017), Atlantis Press, 473-476. https://doi.org/10.2991/seiem-17.2018.118

[15] 金涛, 戴玉刚. 浅析文本聚类有效性评价的方法[J]. 中文信息, 2018(5): 3.

[16] 黄越辉, 曲凯, 李驰, 司刚全. 基于 K-means MCMC 算法的中长期风电时间序列建模方法研究[J]. 电网技术, 2019, 43(7): 2469-2476.