Hans Journal of Data Mining 数据挖掘, 2019, 9(4), 117-130
Published Online October 2019 in Hans. http://www.hanspub.org/journal/hjdm
https://doi.org/10.12677/hjdm.2019.94015

Hans汉斯

Research on Classification of Urban Comprehensive Competitiveness Based on Data Mining

Liangliang Zhuang, Huilin Huang

College of Mathematical, Wenzhou University, Wenzhou Zhejiang
Email: a641292753@qq.com

Received: Sep. 30th, 2019; accepted: Oct. 14th, 2019; published: Oct. 21st, 2019

Abstract

In recent years, with the continuous development of China's economic strength, cities pay more and more attention to their comprehensive urban competitiveness. The establishment of the evaluation system and classification system of urban competitiveness can enable each city to grasp its own future development direction. In order to establish an evaluation system and grade classification, this paper extracts three main factors (Fa1: comprehensive economic and informatization degree, Fa2: urban environment and medical service level factor, and Fa3: economic growth benefit) through factor analysis, and set up city index system. On this basis, we learn and adopt K-center clustering, decision tree, neural network, KNN and the weighted KNN, and start from the three main factor scores to classify each city, and the comprehensive competitiveness of the official city in 2016. The rankings are compared, the classification accuracy of each method is judged, and the optimal method of city classification and the main factors affecting the comprehensive competitiveness of the city are compared. Based on the analysis of R language software, we get the following research conclusions: in the research of city classification, it is found that the classification accuracy of decision tree and neural network algorithm is the best, followed by weighted KNN, KNN algorithm and k-center clustering. Besides, the main factors affecting the comprehensive competitiveness of the city are the revenue within the budget, total retail sales of consumer goods, telephone penetration rate, Internet users, year-end deposit balance of financial institutions and per capita green space index.

Keywords

Urban Comprehensive Competitiveness, Factor Analysis, K-Center Clustering, Decision Tree, BP Neural Network, KNN

基于数据挖掘的各城市综合竞争力等级分类的研究

庄亮亮，黄辉林

文章引用: 庄亮亮, 黄辉林. 基于数据挖掘的各城市综合竞争力等级分类的研究[J]. 数据挖掘, 2019, 9(4): 117-130.
DOI: 10.12677/hjdm.2019.94015

温州大学数理学院，浙江 温州
Email: a641292753@qq.com

收稿日期：2019年9月30日；录用日期：2019年10月14日；发布日期：2019年10月21日

摘　要

近几年随着中国经济实力的不断发展，各个城市越来越注重自身的综合城市竞争力。城市竞争力评价体系、等级分类体系的建立能使各个城市有针对性的把握自身未来发展方向。为了建立评价体系与进行等级分类，本文通过因子分析提取出三个主因子(Fa1：综合经济和信息化程度、Fa2：城市环境与医疗服务水平因子、Fa3：经济增长效益)，并建立了城市指标体系。在此基础上，学习并采用K-中心聚类、决策树、神经网络、KNN与加权KNN等方法，从三个主因子得分入手，对各城市进行等级分类，与2016年官方城市综合竞争力排名进行比对，判断各方法的分类准确率，比较得出城市等级分类的最优方法以及影响城市综合竞争力的主要因素。基于R语言软件分析，我们得到以下研究结论：在对城市等级进行分类的研究中，发现决策树、神经网络算法分类准确率最优，其次分别是加权KNN、KNN算法和K-中心聚类。并且得到影响城市综合竞争力的主要因素分别是财政预算内收入、社会消费品零售总额、电话普及率、互联网用户数、金融机构年末存款余额与人均公园绿地面积指标。

关键词

城市综合竞争力，因子分析，K-中心聚类，决策树，BP神经网络，KNN

1. 引言

1.1. 研究背景与意义

近几年，随着经济全球化和贸易自由化，我国的经济实力正在飞速的发展，中国的国际地位也在不断提高，城市综合竞争力在中国社会经济中的地位正变得越来越重要，各个城市也开始关注起自身城市综合竞争力的发展[1]。根据城市的经济、文化、社会、科技指标，全面评估自身城市的发展水平，这一过程可以通过每年各城市各指标的汇总数据进行收集，再通过较为合理的方式建立指标体系，通过指标体系来给各个城市打分与排名。按照国家城市综合竞争力的得分和排名划分一线城市、新一线城市、二线、三线等级，有利于各城市政府更加深刻的了解自身城市发展的情况，通过与其他城市的比对，再结合自身城市实际情况来改善城市的资源分配；有利于各城市通过评价体系的得分，量化自身下一阶段应当提升的因素，使得城市能够良性发展；有利于我国经济的高质量发展，社会更加安定和谐，人民的幸福感不断提升。

因此，研究影响城市综合竞争力的因素以及指定合理的评价体系是国内外的一个重要课题。

1.2. 相关研究综述

近几年随着中国经济实力的飞速发展，城市综合竞争力在在中国社会经济中的地位越发重要，同时城市综合竞争力对经济现代化也有着很重要的意义。而城市综合竞争力是综合评价一个城市经济、社会、科技、环境的重要指标，对了解和改善城市资源合理分配和城市间良性竞争有着深远的意义。由于我国

具体国情的原因，城市综合竞争力不仅对资源分配有重要影响，而且与城市等级分类也有着密切的联系，因此将对城市综合竞争力影响因素的研究，转化为对影响城市等级分类原因的研究。近年来，城市综合实力等级分类问题引起了学者们越来越多的关注。

1.2.1. 国内研究情况

我国对于各城市综合实力进行分析的研究成果比较丰富，并且近年来我国政府也开始越来越重视城市排名，国内的很多学者利用各类不同的方法建立评价指标如：潘春彩，吴国玺(2012) [2]等人以河南省38个城市为研究对象，从经济发展与收益、社会与科教发展情况等四个方面运用主成分分析的方法，构建城市竞争力评价指标体系，评价河南省城市综合城市竞争力，并对其各个城市进行比较分析；曹清峰，倪鹏飞(2018)等人[3]，基于引力模型对 566 个城市的全球联系度进行测算，并且在此基础上建立了城市竞争力的评价体系；孙霞(2013) [4]则是基于自身提出的指标体系，运用因子分析评价和实证探索浙江省11个城市的城市竞争力，同时又利用聚类分析将11个城市分别分为3类和7类，研究并比较分类效果。

1.2.2. 国外研究情况

国外在城市综合竞争力等方面的研究的学者也比较多，其主要的成果有：Chun Feng Liu、Bao Min Hu、Zi Biao Li 等人(2010) [5]通过对《中国城市竞争力报告》中近 300 个地级市的综合竞争力比较，运用计量经济学模型对 50 多个主要城市的 12 个竞争力进行了评价，形成了 2009 年中国最具竞争力的 10 个城市，并提出了一种应用模糊层次分析法对城市综合竞争力进行评价的 Mathematica 程序；Chun Dong、Chunhua Wu、Xiaoli Sun 等人(2008) [6]通过对物体运动原理的分析，建立了一个包含 4 个子系统、12 个要素和 58 个指标的更加科学的评价指标体系，运用 TOPSIS 方法，对 2009 年中国 28 个城市群的 141 个城市进行了城市竞争力测算。

1.3. 本文主要研究内容

尽管学者们在城市的理论、实证研究以及城市分类上取得了很大的进展，但这些研究一般集中于利用单一分析或者主观分析来构建评价指标体系。例如，以上所提到的研究主要利用模糊层次分析法、因子分析、引力模型、基于自身提出的指标体系等。这些方法存在一定的问题，各有优缺点。而本课题主要通过因子分析建立综合竞争力指标体系，然后采用不同的分类算法进行比较，从五类分类算法中选择一种更为科学、合理的分类方法，并找出影响城市竞争力排名的主要因素。

本课题的主要研究内容如下：

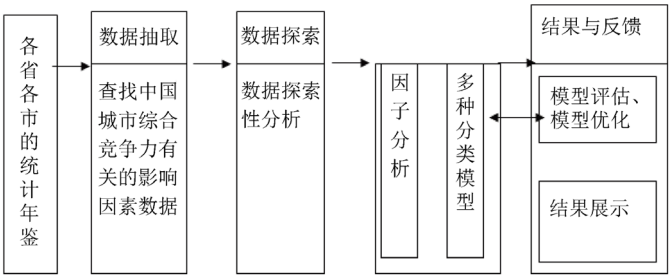
1) 利用各城市统计年鉴搜集 49 个城市的相关指标的实际数据，利用因子分析选取出主要的因子，并建立城市竞争力的评价体系，对各城市竞争力等相关研究给与补充和支持；

2) 利用因子得分数据来研究其他分类方法(K-中心聚类、决策树、BP 神经网络、KNN 算法以及加权 KNN 算法)对城市等级分类的准确性，丰富城市等级分类的方法，并通过准确率的高低来选择最优分类方法。以此作为下一年城市竞争力排名的依据之一；

3) 根据分类结果，并结合近些年我国的国家政策以及国际大环境对国内各个城市的发展，给出合理的建议与意见。

1.4. 论文组织结构

本文通过各省市的统计年鉴查找 59 个城市的相应数据，并对其进行了探索性分析。利用因子分析建立城市竞争力评价体系，在此基础上，通过各种分类模型对 59 个城市进行分类，并对模型进行比较与评估。



2. 变量指标选择

2.1. 数据选择

本文主要研究的是选择影响城市分类的影响因素，以及选择最优分类方法。根据 2016 年城市竞争力官方排名，以及前期结合前辈研究，总结整理出我们所需要的各因素指标，并通过统计年鉴搜集了相应数据[7]，各因素指标类别如表 1 所示。

Table 1. Each factor indicator category
表 1. 各个因素指标类别表

原始指标名称	简单命名	原始指标名称	简单命名
人均 GDP (万元/人)	X1	人口密度(人/平方公里)	X9
GDP 增长率(%)	X2	电话普及率(部/100 人)	X10
财政预算内收入(亿元)	X3	互联网用户数(万户)	X11
社会消费品零售总额(亿元)	X4	人均城市道路面积(平方米)	X12
第二、第三产业占 GDP 比重(%)	X5	每万人拥有公交车	X13
城乡居民人均储蓄年末余额	X6	万人拥有医生数	X14
金融机构年末存款余额(亿元)	X7	万人拥有病床数	X15
居民人均生活用水量(立方米/人)	X8	实际使用外资金额(亿美元)	X16

*所有数据来自各省统计信息网——统计年鉴。

2.2. 数据预处理和探索性分析

由于三线城市中内蒙古，齐齐哈尔等城市相关数据缺失，我们将三线城市进行剔除，主要以一线、新一线以及二线城市为研究对象。

下面，我们利用 R 语言中函数 summary( )计算出了各个指标的最小值、最大值、平均值和标准差，见表 2。

Table 2. Descriptive statistics of each indicator
表 2. 各个指标的描述性统计量

变量名	Min	Max	Mean	SD	变量名	Min	Max	Mean	SD
X1	2016	167,411	88,860	31,963.19	X9	181.2	542.3	787.3	470.11
X2	−5.6	10.7	7.877	2.18	X10	46.88	731.62	212.67	163.32
X3	115.5	6406.1	967.5	1178.48	X11	63	849	270.8	164.75
X4	653.9	11,005.1	3499.2	2287.78	X12	2.95	69.95	11.3	11.87
X5	0.68	1	0.9428	0.06	X13	1.16	86.67	10.15	12.43
X6	29,674	270,237	86,813	50,950.25	X14	16.6	83.1	38.4	15.13
X7	3125	32,792	19,000	23,554.35	X15	28.6	136.9	70.1	20.7
X8	2.56	78.14	27.48	17.7	X16	3781	44,524	28,274	31,699.53

从表 2 我们可知人均 GDP (X1)、城乡居民人均储蓄年末余额(X6)、实际使用外资金额(X16)、人均公园绿地面积(X17)标准差很大，而第三产业与第二产业产值比(X5)标准差、均值特别的小。我们有理由得知我国一线、新一线、二线城市的发展存在较大的差距，特别是经济、环境方面，这也是国内贫富差距的不断增大的一个缩影。

Table 3. Pearson correlation coefficient matrix table
表 3. Pearson 相关系数矩阵表

	X1	X2	X3	X4	X5	X6	X7	X8	X9
X1	1	0.39	0.4	0.4	0.55	0.48	0.39	0.27	0.24
X2	0.39	1	0	0.06	0.71	−0.05	−0.04	0.01	0.04
X3	0.4	0	1	0.87	0.27	0.54	0.94	0.17	0.54
X4	0.4	−0.06	0.87	1	0.17	0.45	0.88	0.15	0.42
X5	0.55	0.71	0.27	0.17	1	0.32	0.25	0.22	0.37
X6	0.48	−0.05	0.54	0.45	0.32	1	0.6	0.6	0.46
X7	0.39	−0.04	0.94	0.88	0.25	0.6	1	0.25	0.51
X8	0.27	0.01	0.17	0.15	0.22	0.6	0.25	1	0.32
X9	0.24	0.04	0.54	0.42	0.37	0.46	0.51	0.32	1
X10	0.37	0.01	0.87	0.89	0.23	0.62	0.88	0.27	0.52
X11	0.36	0.11	0.7	0.83	0.25	0.36	0.66	0.11	0.36
X12	0.33	0.05	−0.02	0.05	0.26	0.64	0.03	0.57	0.11
X13	0.46	0.06	0.34	0.21	0.21	0.76	0.36	0.52	0.38
X14	0.41	−0.04	0.38	0.29	0.28	0.83	0.47	0.61	0.23
X15	0.18	−0.13	0.25	0.27	0.12	0.65	0.3	0.47	0.05
X16	0.41	0	0.69	0.75	0.25	0.66	0.72	0.55	0.48
X17	0.41	0	0.69	0.75	0.25	0.66	0.72	0.55	0.48
X18	0.35	0.04	0.96	0.9	0.25	0.46	0.92	0.1	0.45

	X10	X11	X12	X13	X14	X15	X16	X17	X18
X1	0.37	0.36	0.33	0.46	0.41	0.18s	0.41	0.41	0.35
X2	0.01	0.11	0.05	0.06	−0.04	−0.13	0	0	0.04
X3	0.87	0.7	0.02	0.34	0.38	0.25	0.69	0.69	0.96
X4	0.89	0.83	−0.05	0.21	0.29	0.27	0.75	0.75	0.9
X5	0.23	0.25	0.26	0.21	0.28	0.12	0.25	0.25	0.25
X6	0.62	0.36	0.64	0.76	0.83	0.65	0.66	0.66	0.46
X7	0.88	0.66	0.03	0.36	0.47	0.3	0.72	0.72	0.92
X8	0.27	0.11	0.57	0.52	0.61	0.47	0.55	0.55	0.1
X9	0.52	0.36	0.11	0.38	0.23	0.05	0.48	0.48	0.45
X10	1	0.86	0.1	0.46	0.41	0.34	0.76	0.76	0.88
X11	0.86	1	−0.04	0.31	0.19	0.21	0.62	0.62	0.75
X12	0.1	−0.04	1	0.52	0.65	0.54	0.34	0.34	−0.02
X13	0.46	0.31	0.52	1	0.69	0.48	0.49	0.49	0.25
X14	0.41	0.19	0.65	0.69	1	0.77	0.52	0.52	0.32
X15	0.34	0.21	0.54	0.48	0.77	1	0.44	0.44	0.16
X16	0.76	0.62	0.34	0.49	0.52	0.44	1	1	0.68
X17	0.76	0.62	0.34	0.49	0.52	0.44	1	1	0.68
X18	0.88	0.75	−0.02	0.25	0.32	0.16	0.68	0.68	1

除此之外，我们分析下各个影响因素指标之间的相关性，下面我们使用变量 Pearson 相关系数矩阵来描述各个变量之间的关系(阴影数据表明两变量之间的相关系数大于 0.7)，由表 3 可以看出各个因素之间多重共线性较为严重，其主要是存在正相关关系，只有少部分几个变量之间存在负相关性，但负相关系数都不是很大在−0.13 之内。值得注意的是，X10 与 X3、X4、X7、X11、X16、X17、X18 有密切强正相关性；X18 与 X3、X4、X7、X10、X11 有密切强正相关性。而 X1、X8、X9、X12 与其他变量之间没有强相关性。

3. 模型建立

为了消除原始数据数量级和量纲的差异，以及部分算法的条件所限，我们先前先对数据进行了标准化处理，利用 R 语言中的 scale()函数进行标准化。

3.1. 因子分析

因子分析(Factor Analysis)是对相关系数矩阵内部结构的依赖性的研究，它将多个变量浓缩为几个不互相干扰的主要因素，从而来表现出原始数据与因子之间的相关关系。

模型的形式为：

$$X_i = a_1 Fa_1 + a_2 Fa_2 + \cdots + a_p Fa_p + U_i$$

其中，是第 i 个可观测变量($i = 1 \cdots k$)，是公共因子($j = 1 \cdots p$)，并且当 $p < k$ 时。是变量独自拥有的部分(无法被公共因子解释)。可以被看作是每个因子对复合可观测变量的贡献。

以下我们使用 R 语言对预处理数据执行因子分析，利用 psych 包对 18 个自变量进行因子分析，观察各因子的累计贡献率，提取出适当的因子数，再通过正交旋转变换得到相应的旋转成份矩阵并结合各因子所含主要自变量进行命名，最后就是建议相应的评价体系，并给 49 个城市打分、给出排名，根据结果给出相应的评价与建议。

3.1.1. 判断需提取的公共因子数

首先利用 psych 包中的 fa.parallel()来判断需提取的公共因子数，从表 4 所示，从中可知前 3 个因子已经提取出了原始数据的 76.48%的信息，并且特征值差异大，解释能力强。综合以上提取前三个因子作为主要因素，然后进行因子旋转。

Table 4. The cumulative contribution rate of each factor
表 4. 各因子的累积贡献率

因子	特征值	方差贡献率%	累计贡献率%
1	6.925	38.472	38.472
2	4.661	25.892	64.364
3	2.181	12.115	76.48

3.1.2. 因子旋转

三个主要因子与原始数据之间的相关性由因子载荷矩阵反映，由于旋转前各因子的信息结构不太明确，各个因子的解释能力不够强。于是我们进行了方差极大的正交旋转变换，利用 R 语言中的 fa( )函数，并利用正交旋转来旋转三个因子，以获得该相应的旋转成份矩阵，具体见表 5。

从表 5 中可以得知，含阴影的数据是各个因子中影响因素较大的指标系数。我们可以看出：第一个公因子对所有初始变量累积方差的贡献率达到 38.472%，其中 X3 (财政预算内收入)、X4 (社会消费品零

售总额)、X7 (金融机构年末存款余额)、X10 (电话普及率)、X11 (互联网用户数)和 X18 (人均财政教育费用支出)这 6 个指标上具有很大的负荷值，所以我们也可以得知这 6 个指标对因子一有很大的影响，结合实际情况以及变量的观察，我们发现这 6 个指标主要与社会经济、信息化普及程度有关，于是我们将因子一定义为：**综合经济和信息化程度因子**。

Table 5. The cumulative contribution rate of each factor
表 5. 各因子的累积贡献率

指标	Fa1	Fa2	Fa3	指标	Fa1	Fa2	Fa3
X1	0.289	0.321	0.649	X9	0.504	0.206	0.26
X2	-	−0.132	0.893	X10	0.916	0.252	-
X3	0.925	0.152	0.125	X11	0.826	-	0.176
X4	0.957	-	-	X12	−0.153	0.827	0.178
X5	0.133	0.152	0.89	X13	0.221	0.737	0.203
X6	0.411	0.823	0.145	X14	0.213	0.875	-
X7	0.909	0.224		X15	0.149	0.778	−0.113
X8	-	0.767	0.112	X16	0.726	0.506	0.103

第二个公共因子对所有初始变量的累计方差贡献率达到了 25.892%，其中在 X6 (城乡居民人均储蓄年末余额)、X8 (居民人均生活用水量)、X12 (人均城市道路面积)和 X14 (万人拥有医生数)四个指标中具有较大的负荷值，这 4 个指标主要与城市环境、基础设施以及医疗服务有相应的关系，于是我们将因子二定义为：**城市环境与医疗服务水平因子**。

第三个公共因子对所有初始变量的累计方差贡献率达到了 12.115%，其中在 X2 (GDP 增长率)与 X5 (第二、第三产业占 GDP 比重)两个指标中具有较大的负荷值，这 2 个指标主要与城市经济有相应的关系，于是我们将因子三定义为：**经济增长效益因子**。

3.1.3. 因子综合得分以及评价

利用 fa.diagram ( )获取正交图并通过 fa.varimax$weights 得到相应的因子得分，我们可以得到每个因子的表达式：

$$\begin{aligned} Fa1 = {} & 0.289X1 + 0.925X3 + 0.957X4 + 0.133X5 + 0.411X6 + 0.909X7 + 0.504X9 \\ & + 0.916X10 + 0.826X11 - 0.153X12 + 0.221X13 + 0.213X14 + 0.149X15 \\ & + 0.726X16 + 0.726X17 + 0.945X18 \end{aligned}$$

$$\begin{aligned} Fa2 = {} & 0.321X1 - 0.132X2 + 0.152X3 + 0.152X5 + 0.823X6 + 0.224X7 + 0.767X8 \\ & + 0.206X9 + 0.252X10 + 0.827X12 + 0.737X13 + 0.875X14 + 0.778X15 \\ & + 0.506X16 + 0.506X17 \end{aligned}$$

$$\begin{aligned} Fa3 = {} & 0.649X1 + 0.893X2 + 0.125X3 + 0.89X5 + 0.145X6 + 0.112X8 + 0.26X9 \\ & + 0.176X11 + 0.178X12 + 0.203X13 - 0.113X15 + 0.103X16 + 0.103X17 \\ & + 0.126X18 \end{aligned}$$

在此基础上，我们使用每个因子的方差贡献率与三个主要因子的总方差贡献率的比例作为每个因子的系数，并进行加权求和，我们可以可知因子综合得分的计算公式为：

$$Fa = \frac{0.385Fa1 + 0.259Fa2 + 0.121Fa3}{0.765}$$

各城市的主因子得分以及因子综合得分计算结果如表 6 所示。

Table 6. Factor score and comprehensive ranking of comprehensive competitiveness of each city
表 6. 各城市综合竞争力的因子得分及综合排名

城市	Fa1	Fa2	Fa3	因子综合得分	综合排名	城市	Fa1	Fa2	Fa3	因子综合得分	综合排名
北京市	3.32	0.48	−0.14	1.386	3	东莞	−0.89	3.75	0.11	0.642	5
上海市	3.92	0	0.05	1.515	1	昆明	−0.61	0.23	−0.26	−0.207	27
广州市	1.5	1.49	0.42	1.014	4	太原	−0.86	0.7	−0.13	−0.166	24
深圳市	1.41	2.94	1.3	1.462	2	南昌	−0.58	−0.53	0.4	−0.312	35
成都	0.94	−0.18	−0.19	0.292	9	南宁	−0.41	−0.17	−1.12	−0.337	41
杭州	0.38	0.24	0.58	0.279	10	贵阳	−0.83	0.04	0.34	−0.268	32
武汉	0.38	0.06	0.24	0.191	12	海口	−1.11	0.85	−0.33	−0.247	29
天津	0.8	−0.28	0.79	0.331	8	长春	−0.5	−0.41	−0.2	−0.323	37
南京	0.49	0.61	0.55	0.413	6	泉州	−0.2	−1.12	0.25	−0.337	40
重庆	1.95	−1.53	0.14	0.371	7	洛阳	−0.54	−0.92	−0.1	−0.458	49
西安	−0.01	−0.18	0.12	−0.036	20	常州	−0.6	−0.03	0.72	−0.152	23
长沙	−0.28	−0.04	0.55	−0.052	22	珠海	−1.51	1.83	0.95	0.008	15
青岛	0.16	−0.39	0.31	−0.002	16	金华	−0.49	−0.66	−0.03	−0.363	42
沈阳	0.36	0.71	−5.69	−0.366	44	烟台	−0.37	−0.72	0.05	−0.323	36
大连	−0.12	0	−1.86	−0.271	33	惠州	−0.69	−0.44	0.04	−0.375	46
厦门	−0.76	0.93	0.61	0.022	14	徐州	−0.15	−1.09	−0.2	−0.364	43
苏州	0.49	0.04	0.19	0.222	11	嘉兴	−0.49	−0.5	−0.05	−0.324	38
宁波	0.07	−0.34	0.48	−0.003	17	潍坊	−0.25	−1	−0.39	−0.402	48
无锡	−0.21	0.03	0.46	−0.017	18	南通	−0.24	−0.85	0.5	−0.252	31
福州	−0.13	−0.77	0.5	−0.189	26	扬州	−0.6	−0.87	0.59	−0.385	47
合肥	−0.42	−0.54	0.75	−0.211	28	汕头	−0.34	−1.09	0.36	−0.370	45
郑州	−0.02	−0.16	0.19	−0.026	19	哈尔滨	−0.2	−0.56	−0.87	−0.327	39
温州	−0.15	−0.93	0.41	−0.249	30	乌鲁木齐	−0.8	1.09	−1.18	−0.168	25
佛山	−0.43	0.94	0.04	0.083	13	石家庄	−0.09	−0.82	−0.48	−0.305	34
济南	−0.28	0.14	0.21	−0.046	21						

由表 6 可知，上海综合排名是 49 个城市中的第一名，其次是深圳市、广州市与北京市。广州是的 3 个因子得分较为均匀，比较适合任命生活与发展，而深圳市则是在经济方面得分较高，是我国经济不断增长的驱动力与助推器。相比而言，成都的综合经济和信息化程度因子得分较大，而其余两因子得分不高，而杭州市则是在因子三(经济增长效益因子)得分出众，作为浙江省的省会，确实是有自身城市的优越性和实力。其他城市在此就不做逐一分析，总之，通过因子分析我们可以建立相关的城市评价体系。接下来，在此基础上，我们利用各主因子得分作为变量数据，通过四种分类算法对此城市评价指标进行评估，来寻找最优分类方法以及对城市竞争力影响较大的因素。

3.2. 分类方法

通过因子分析，我们已经得出了相应的评价体系，并且得出了各城市的总得分与总排名，也对部分城市进行了分析。在此基础上，利用因子得分数据为我国的 49 个城市进行分类。

在进行每种分类算法的研究之前，我们首先将数据划分为训练集与测试集。其中，训练集占比 3/4，测试集占比 1/4。主要目的是为了防止模型出现过拟合情况，使得模型泛化性更强，利用训练集进行模型训练，利用测试集进行验证。

3.2.1. K-中心聚类

1) **基本原理**

聚类分析：根据事物的某个特征把数据对象划分成多个子集的过程，每个子集都是一个簇，并且使簇的对象彼此相近，但与其他簇中的对象不相近[8]。

相比 K-均值算法，K-中心点算法对异常值不敏感，它并不采用若干类中对象的平均值作为簇中心，而选用实际对象来代表所在的簇，以此为簇中心。

我们利用个城市进行聚类，通过训练集中 36 个城市进行聚类，并用测试集中的 13 个城市进行验证，与 2016 年官方排名进行比较，计算其准确率。

2) **实际应用**

用 R 软件对各城市二级指标数据进行 K-中心聚类。我们得到以下结果表 7 和表 8。

Table 7. Clustering results
表 7. 聚类结果

	聚类结果
1	北京市 上海市 广州市 深圳市 成都 杭州 天津 南京 西安 长沙 大连 苏州 宁波 无锡
2	武汉 重庆 青岛 沈阳 厦门 东莞 昆明 太原 海口 珠海 乌鲁木齐
3	福州 合肥 郑州 温州 佛山 济南 南昌 南宁 贵阳 长春 泉州 洛阳 常州 金华 烟台 惠州 徐州 嘉兴 潍坊 南通 扬州 汕头 哈尔滨 石家庄

Table 8. Performance results of K-center clustering model
表 8. K-中心聚类模型性能结果

测试集	官网数据	测试集中分类			
		A	B	C	合计
官网分类	A	1	0	0	1
	B	0	1	3	4
	C	0	1	7	8
	合计	1	2	10	13

由上两个表我们可以看出，用 K-中心聚类分出来的城市类别，导致了部分城市与已有的城市分类数据相差较大，准确率不是很高，有 16 个城市分类错误，其中分类后属于一线城市的数量竟然多达 14 个，有 10 个城市分错了类。其中，在测试集中的 13 个城市内，有 4 个城市分类错误，该算法准确率仅有 69.23%。

3.2.2. 决策树分类

1) 基本原理

决策树(Decision Tree)是一种预测模型，它可以用来做回归也可以做分类，它是一种类似于流程图的树结构，我们可以将它分为三个部分：决策节点，分支和叶节点。决策节点表示对属性上的测试，属性上的不同测试结果代表为分支；分支表示为某个决策节点的不同取值，每个叶节点代表一种可能的分类结果[9]。

本文主要利用利用 C4.5 算法(从候选划分属性中找出信息增益高于平均水平的属性，再从中选择增益率最高的属性)，对训练集中 36 个城市进行训练，并画出其决策树，然后再用测试集进行验证。

2) 实际应用

利用 R 软件中 RWeka 包的 J48( )函数，建立决策树模型。再通过加载 partykit、grid 包，我们得到以下的决策树模型，见图 1。

由图 1 可以看出，此时的分类规则是：如果 Fa1 (综合经济和信息化程度)小于等于−0.02 时，则属于二线城市，其中有 24 个城市属于此范畴；当 Fa1 大于−0.02 且 Fa1 小于等于 0.94 时，属于新一线城市，有 8 个城市在此范畴内；当 Fa1 大于 0.94 时，此范畴属于一线城市，并有 4 个城市落入其中。

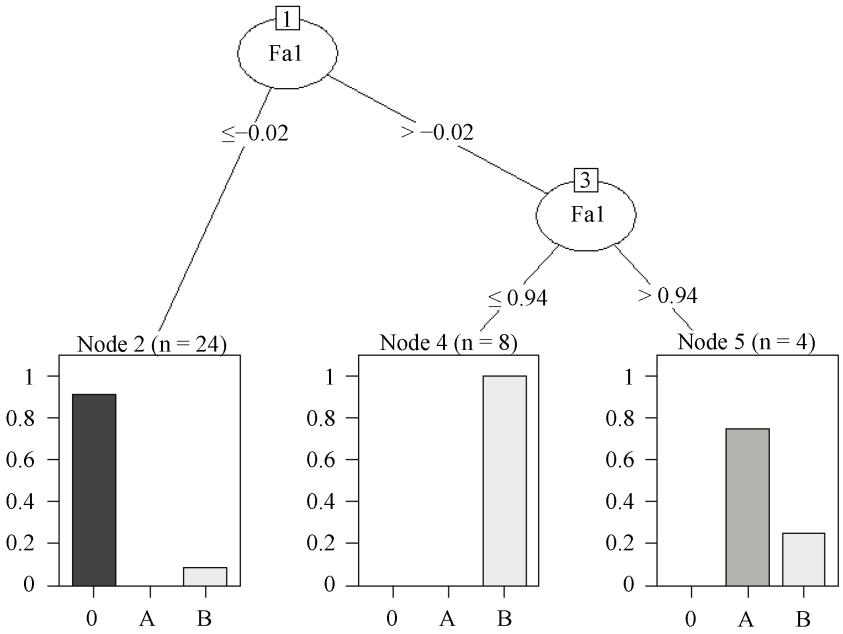


Figure 1. Training decision tree based on C4.5 algorithm
图 1. 基于 C4.5 算法的训练决策树

Table 9. Decision tree model performance results
表 9. 决策树模型性能结果

测试集	官网数据	测试集中分类			
		A	B	C	合计
官网分类	A	1	0	0	1
	B	0	3	1	4
	C	0	1	7	8
	合计	1	4	8	13

我们可以从表 9 看出 Fa1 对城市等级分类有重大影响，即影响城市等级分类的影响因素是综合经济和信息化程度，而 Fa2、Fa3 与城市等级分类并没有太大的影响。利用测试集数据，发现 13 个城市中方只有 2 个城市分类错误，我们可以得知分类准确率为 84.62%。可知：该模型的一线城市、二线城市以及三线城市的分类准确率较高，并且影响城市等级分类的主要因素为 Fa1 中系数较大的指标。说明用决策树对城市进行分类是可行的，并且分类错误率较低。

3.2.3. BP 神经网络

1) 基本原理

BP 神经网络是一种根据误差的反向传播，对多层前馈网络进行训练，该算法称为 BP 算法，主要思想是梯度下降法，利用梯度搜寻技术，为了最小化网络的实际输出量和预期输出之间的误差均方误差[10]。

本文主要利用 BP 神经网络进行分类处理，并对测试集中的 13 个城市进行分类，计算准确率。

计算过程为：① 网络状态初始化；② 前向计算过程。

2) 实际应用

用 R 软件对数据建立一个包括 4 个隐藏层节点的神经网络模型，本次共进行了 4634 次迭代，迭代结束时损失函数为 0.513，权值的最大调整量为 0.009，再对神经网络进行可视化，可接着得到图 2。

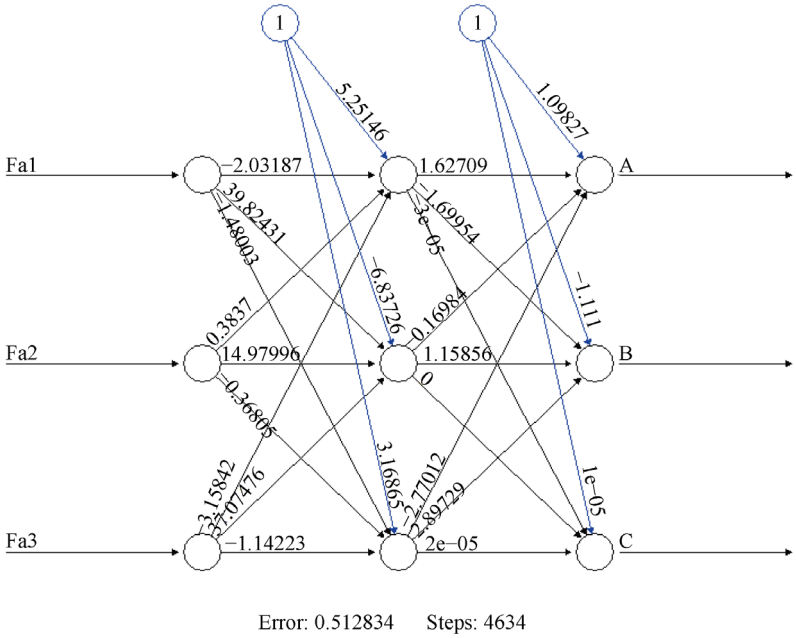


Figure 2. BP neural network
图 2. BP 神经网络

接着对模型进行性能预测评估，通过 R 软件，发现只有两个城市分类错误，见表 10，得到混淆矩阵精度为 84.62%，Kappa 为 0.726，总的看来该模型性能还是比较高的；再从城市分类的灵敏度看，除了一线城市和新一线城市的灵敏度较低外，二线城市的灵敏度还是比较高的，而一线城市和新一线城市的灵敏度较低，可能与原始数据中，这两类的城市数据较少有关，基于此认为总体模型性能还是挺好的。

Table 10. Performance results of BP neural network model
表 10. BP 神经网络模型性能结果

测试集	官网数据	测试集中分类			
		A	B	C	合计
官网分类	A	1	0	0	1
	B	0	3	1	4
	C	0	1	7	8
	合计	1	4	8	13

3.2.4. KNN 算法

1) 基本原理

K 最近邻(KNN)分类算法，是最简单的机器学习算法之一，虽然它的想法很简单，但是该方法功能及其强大。K 最近邻方法主要思想是：将未分类的数据归入到与他们最相似的一类之中，其中主要利用相应的距离公式计算出该对象与各类的距离[11]。

2) 实际应用

我们利用 R 语言 class 包中的 knn( )函数，通过训练集进行训练，测试集进行验证的方式，得到混淆矩阵如表 11 所示。

Table 11. KNN model performance results
表 11. KNN 模型性能结果

测试集	官网数据	测试集中分类			
		A	B	C	合计
官网分类	A	1	0	0	1
	B	0	3	1	4
	C	0	2	6	8
	合计	1	5	7	13

计算得表 11，13 个测试城市中有 3 个城市分类错误，该算法的准确率为 76.92%，准确率较高，说明该算法用于城市等级分类还是可行。从图中发现，测试集中的 13 个城市中，有一个一线城市错分为了新一线城市，两个二线城市错分为了新一线城市，其他城市分类正确，总体来看该模型性能还是较好的。

3.2.5. 加权 KNN 算法

1) 基本原理

核函数加权最近邻(KKNN)分类算法，是 KNN 算法的一种扩展，主要思想是：通过给每个点的距离加入一个权重，是的距离较近的点可以得到更大的权重，而距离较远的点则得到较小的权重。

2) 实际应用

我们利用 R 语言 kknn 包中的 kknn( )函数，通过训练集进行训练，测试集进行验证的方式，得到混淆矩阵如表 12 所示。

计算得知，该算法的准确率达为 76.92%，与 KNN 算法准确率相同，从理论来看，该算法会比 KNN 更精确，但是由于数据量较少等原因，在此数据中并未能体现出其优越性。

Table 12. KKNN model performance results
表 12. KKNN 模型性能结果

测试集	官网数据	测试集中分类			
		A	B	C	合计
官网分类	A	1	0	0	1
	B	0	3	1	4
	C	0	2	6	8
	合计	1	5	7	13

4. 结论与建议

由表 13 我们可以得知：决策树分类与神经网络效果较好，准确率高达 84.62%。综上所述：主要影

响指标为 Fa1，而 Fa1 (综合经济和信息化程度)的主要影响因素是 X3 (财政预算内收入)、X4 社会消费品零售总额(亿元)、X10 电话普及率(部/100 人)、X11 互联网用户数(万户)、X7 (金融机构年末存款余额)、X19 (人均公园绿地面积)，说明综合经济实力对城市综合竞争力有着极大影响。

从以上结论可以发现，由于我们通过因子分析，得到三个关键因子分别为：Fa1：综合经济和信息化程度、Fa2：城市环境与医疗服务水平因子、Fa3：经济增长效益。其中，Fa1 的解释能力最强，所以对数据的影响程度也就越大，再通过因子得分作为数据进行分类分析后，Fa1 中的 X3、X4、X7、X10、X11 占比较大，对数据的影响较大。而在四种分类算法中，Fa1 都起到主要的作用，对城市综合竞争力，有着至关重要的影响。这不仅说明我国国民经济对城市发展的基础性作用，更说明还有政府对市场配置的宏观调控能够显著影响城市进程的发展和进步，来提升城市的综合竞争力。因此中国应着眼于提高经济的核心竞争力，以提升城市的综合竞争力。

Table 13. Comparison of classification methods
表 13. 各分类方法比较

分类方法	主要影响指标	准确率
k-中心	-	69.23%
决策树算法	Fa1 (X3、X4、X7、X10、X11)	84.62%
BP 神经网络	Fa1	84.62%
KNN 算法	Fa1	76.92%
加权 KNN 算法	Fa1	76.92%

此外，在党的十九大报告中提出了“建设科技强国”。如何以习近平新时代中国特色社会主义思想作为指导，这是广大爱国人民正在思考的问题，也是各政府工作人员必须面对的问题。从分类结果中也可以得知：科技发展对于各个城市的综合竞争力起到重要的作用，在发展城市经济与科技的同时，保持环境质量也是提高城市竞争力不可或缺的因素。总而言之，需要针对每个城市不同的发展状况，具体问题具体分析，以现实情况为基础，坚持经济综合实力的提升为核心，发展各城市的潜藏能力，全面提升城市综合竞争力。

致　谢

在本次论文设计过程中，黄辉林老师对该论文从选题，构思到最后定稿的各个环节给予细心指引与教导，使我们得以最终完成论文设计。并且感谢众多老师的关心支持和帮助。在此，谨向老师们致以衷心的感谢和崇高的敬意！

基金项目

温州大学大学生创新创业项目(No.JWD2017078)。

参考文献

[1] 陶陶玉, 顾朝林, 涂英石. 新时期城市与城市综合竞争力[J]. 城市规划论坛, 2001(4): 12-17.

[2] 潘春彩, 吴国玺, 闫卫阳. 基于主成分分析的河南省城市综合竞争力评价[J]. 地域研究与开发, 2012, 31(6): 60-64.

[3] 曹清峰, 倪鹏飞, 沈立, 张洋子. 东亚主导下的亚洲城市体系——基于城市竞争力的分析[J]. 北京工业大学学报(社会科学版), 2018, 18(6): 43-52.

[4] 孙霞. 基于因子分析和聚类分析的城市竞争力综合评价[J]. 赤峰学院学报(自然科学版), 2013(22): 53-55.

[5] Liu, C.F. and Hu, B.M. (2010) Evaluation of Urban Comprehensive Competitiveness Based on FAHP. *Advanced Materials Research*, **108-111**, 421-425. https://doi.org/10.4028/www.scientific.net/AMR.108-111.421

[6] Dong, C., Wu, C.H. and Sun, X.L. (2008) Research on Evaluating Urban Comprehensive Competitiveness during Multi-Year and the Spatial Characters of Important Cities in China. *Fifth International Conference on Fuzzy Systems and Knowledge Discovery*, Los Alamos, 18-20 October 2008. https://doi.org/10.1109/FSKD.2008.667

[7] 中国统计局. 中国统计年鉴[M]. 北京: 中国统计出版社, 2016.

[8] 章永来, 周耀鉴. 聚类算法综述[J]. 计算机应用: 1-14[2019-04-30].

[9] 杨剑锋, 乔佩蕊, 李永梅, 王宁. 机器学习分类问题及算法研究综述[J]. 统计与决策, 2019, 35(6): 36-40.

[10] 常强, 赵伟, 赵仰杰. 基于神经网络的数据分类预测与实现[J]. 软件, 2018, 39(12): 207-209.

[11] 王德宝. 基于 KNN 算法的改进研究及其在数据分类中的应用[D]: [硕士学位论文]. 淮南: 安徽理工大学, 2018.