Software Engineering and Applications 软件工程与应用, 2016, 5(5), 296-302
Published Online October 2016 in Hans. http://www.hanspub.org/journal/sea
http://dx.doi.org/10.12677/sea.2016.55034

Hans汉斯

Urban Cluster Analysis of Yangtze River Delta Metropolis Circle Based on the Main Economic Indicators

Mingfang Zhu[1,2], Yanling Ren[3], Hongfen Jiang[2]

[1]Science Computing and Intelligent Information Processing of Guangxi Higher Education Key Laboratory, Guangxi Teachers Education University, Nanning Guangxi
[2]School of Computer Engineering, Jiangsu University of Technology, Changzhou, Jiangsu
[3]School of Electronic Information, Jiangsu University of Technology, Changzhou, Jiangsu
Email: mfzhu2009@jsut.edu.cn

Received: Oct. 10th, 2016; accepted: Oct. 25th, 2016; published: Oct. 28th, 2016

Abstract

The Yangtze River Delta metropolitan area in the development of national economy in our country has a very important position and role. For each China's economic growth of 1%, Shanghai-Su-Zhejiang accounts for about 1/3, therefore the Yangtze River Delta area economic development status and level embodies China's economic development situation and level. In this paper, based on the 15 cities' main economic indicators of the Yangtze River Delta in 2013, using R and Rattle two open source software, we has carried on the urban cluster. We have evaluated objectively the relative ranking of economic development of those cities in the Yangtze river delta economic development, providing decision-making reference to the further development of cities in the region and the improvement of urban competitiveness.

Keywords

Cluster Analysis, Economic Indicator, Yangtze River Delta Area, R Software

基于主要经济指标的长三角大都市圈的城市聚类分析

朱明放[1,2]，任艳玲[3]，蒋红芬[2]

文章引用: 朱明放, 任艳玲, 蒋红芬. 基于主要经济指标的长三角大都市圈的城市聚类分析[J]. 软件工程与应用, 2016, 5(5): 296-302. http://dx.doi.org/10.12677/sea.2016.55034

[1]广西师范学院科学计算与智能信息处理广西高校重点实验室，广西 南宁
[2]江苏理工学院计算机工程学院，江苏 常州
[3]江苏理工学院电气信息学院，江苏 常州
Email: mfzhu2009@jsut.edu.cn

收稿日期：2016年10月10日；录用日期：2016年10月25日；发布日期：2016年10月28日

摘 要

长江三角洲都市圈在我国国民经济发展中具有十分重要的地位和作用，中国经济每增长1个百分点，沪苏浙约占1/3，因此长江三角洲地区经济发展状况和水平体现着中国的经济发展状况和水平。本文根据2013年度的长江三角洲的16个城市的主要经济指标，运用R和Rattle两个开源软件，进行了城市聚类，客观评价了各个城市经济发展在长三角经济发展中的相对位置，为该地区各个城市的进一步发展和提高城市竞争力提供决策参考。

关键词

聚类分析，经济指标，长三角地区，R软件

1. 引言

大都市圈是城市群发展到成熟阶段的最高组织形式，是由地理、经济和历史文化因素交织而成的一种空间现象，其中地理条件是基础，经济条件是主体，历史传承是空间格局的文化因素。目前，长江三角洲都市圈已经被认为全球第六大城市圈，由苏州、无锡、常州、扬州、南京、南通、镇江、泰州、杭州、嘉兴、宁波、绍兴、舟山、湖州、台州等 15 个地级市与上海一起组成[1]。由于大都市圈的人口、经济和城镇高度密集，在国民经济发展中具有十分重要的地位和作用，就我国的经济发展状况而言，我国经济每增长 1 个百分点，其中沪苏浙约占 1/3 [2]。目前，沪苏浙 15 个城市以群星璀璨之势，形成明亮的星云团，成为世界最大的都市圈，成为中国经济社会发展的强有力“引擎”[2]。

聚类分析是数据挖掘技术重要研究内容之一[3]，通过聚类分析能够识别数据对象中稠密区域和稀疏区域，从而发现数据的全局分布模式和数据属性之间的相关关系，可以获得研究对象更多的信息和更深刻的认识，可以对研究对象有更加科学的决策。本文是根据长三角大都市经济圈中 16 个城市的 2013 年主要经济指标，运用 R 和 Rattle 两个开源软件中提供的聚类分析的方法，对它们进行聚类分析，对各个城市经济发展在长三角经济圈中的相对位置进行客观评价，可为该地区各个城市的进一步发展和提高城市竞争力提供决策参考。

2. 相关软件与数据挖掘简介

R 是一款开源免费的软件，在数据分析领域获得了越来越多的用户的青睐它包含了大量的用于机器学习的添加包[4]-[6]，为用户提供了极大方便。R 目前已经广泛应用到于统计分析、绘图、数据挖掘等领域，并且随着大数据、云计算的广泛应用和研究，其研究和应用领域更为广泛。Rattle 是建立在 R 基础上的专门用于数据挖掘的用户界面操作的开源软件[7] [8]。

2.1. R 软件

R 是一种函数式编程语言，用户一方面通过交互方式使用它，同时用户可以开发自己的包或者修改

已有的函数的功能，满足自己的研发需求。

R 的功能强大，是因为它拥有十分丰富的附加包(package)，这些包需要在安装 R 的基础环境后，根据需要自由下载和安装。为了帮助用户选择需要的 R 包，CRAN (Comprehensive R Archive Network)任务视图提供了不同任务的各种 R 包集的向导，与数据挖掘任务相关的任务视图有：机器学习与统计学习，聚类分析与有限混合模型，时间序列分析，多元统计分析，空间数据分析等。用户根据自己的研究和应用任务，在相应的任务视图(Task View)下，选择安装自己需要的包。

2.2. Rattle 与数据挖掘

Rattle 是建立在 R 语言基础上的一款基于图形用户界面(GUI)的数据挖掘软件[7] [8]，它通过 GUI 快速简单地完成数据挖掘项目的各个阶段的工作，包括数据挖掘各项任务的数据预处理、模型建立、模型评价等工作。虽然 Rattle 语言是建立在 R 基础上，但是使用 Rattle 进行数据挖掘项目，并不一定非要学习 R。R 是拥有执行数据挖掘任务功能强大的语言，可以超出 Rattle 图形界面下各种数据挖掘功能的限制，当我们需要进一步调整或者开发数据挖掘项目时，可以将 Rattle 的各项工作迁移到 R 环境下进行调整。

跨行业数据挖掘过程标准(Cross Industry Process for Data Mining, CRISP-DM)为数据挖掘提供了一个统一的标准工作过程框架，仿照软件生命周期的模式，将数据挖掘过程划分为六个阶段[7]：1) 业务理解阶段；2) 数据理解阶段；3) 数据准备(预处理)阶段；4) 数据建模阶段；5) 模型评估阶段和 6) 部署应用阶段。Rattle 图形用户界面根据这一标准过程，将数据挖掘任务按照 7 个依次执行的步骤：1) 数据导入；2) 数据选择；3) 数据探查；4) 数据变换；5) 数据建模；6) 模型评估和 7) 模型导出，另外，Rattle 的 Log 将记录着我们操作的每一步的 R 代码，以便重复数据挖掘过程和修改、调整代码，以适应我们的挖掘要求。

2.3. 聚类分析算法

聚类分析是按照一定的要求和规律，把一个没有类别标号的数据集分成若干个子集(类)，使相似的对象尽可能地归为一类，不相似的对象尽可能地划分到不同类中的数据分析方法，通过聚类分析，能有效地发现隐含在数据集中的数据分布特性，从而为进一步充分、有效地利用数据奠定良好的基础。聚类分析是数据挖掘的核心任务之一，已经涌现了大量的数据聚类算法和应用案例[9]-[11]。R 支持大量的聚类算法，有基于划分的聚类算法，如 K-means 函数；基于模型的聚类算法，如 mclust 函数；基于层次的聚类算法，如 hclust 函数、agnes 函数和 diana 函数等。本节主要介绍 Rattle 环境下的主要的 K-means 算法和层次聚类的 hclust 算法。

K-means 聚类算法将数据挖掘获得的模型(知识)表示为 k 个均值的集合，数据集中的每个对象与这 k 个均值最近的那个均值表示的类(知识)相关联，并由此将数据集划分为 k 个子集类。因此，关于 k-means 的研究将主要集中在数据集中每个属性的不同测量的均值的定义方法，聚类个数 k 的确定方法等。为了对排除对象属性的测量单位影响，一般需要在聚类算法实施之前对每个属性进行无量纲化处理。

层次聚类算法一般划分为从上到下划分聚类方法和从下到上的凝聚聚类方法，hclust 函数是一种凝聚方式的聚类算法。凝聚方式的聚类的基本思想是首先将数据集中最相近的两个对象合并成第一个类，并用这个类各个属性的均值表示这个类，再将这个类看作一个新的对象放回对象集中；接下来继续对着的继续上述过程，直到凝聚成一个类时算法结束，这一过程容易使用一个谱系图表示。我们将在实验部分看到这样的可视化聚类结果。

3. 各城市主要经济指标数据

本文采用的长江三角洲大都市圈中各个城市的经济指标数据直接来源于上海统计局网站[12]。因个别

城市的个别指标缺失，也到相应城市的统计局网站进行了收集和遴选。

考察以及国家或地区的经济状况的指标有很多，这里采用主要指标有：生产总值(季度数)、规模以上工业总产值，规模以上工业产品产销率，进出口总额，出口总额，外商直接投资实际到位金额，地方财政一般预算收入，金融机构本外币存款余额，金融机构本外币贷款余额。这九个指标大致可以分为四个方面：宏观层面的经济指标，如生产总值、工业总产值以及产品产销率；国际指标，如进出口总额、出口总额、国际资本净流入量；金融指标，如金融机构的本外币存款、贷款余额；政府指标，如财政预算收入等。

在这九个指标中，规模以上工业产品产销率是无量纲的百分比计量，其余指标均为货币单位计量，其中进出口总额，出口总额，外商直接投资实际到位金额三个指标的单位是“亿美元”，其余是“亿元”为计量单位。

4. 实验分析

本节主要阐述是将长三角大都市经济圈的 16 个城市的主要经济指标数据导入到 Rattle 环境下，对其进行数据预处理，建立聚类分析模型，并对模型进行一定的分析。

4.1. 数据导入与预处理

从参考文献[12]中获取 2013 年度的相关数据集，并对有缺失值的指标到相关城市的统计官网进行筛选予以填补。数据以 Excel 文件存储，为了导入数据方便，我们将其存储为 CSV 格式的文件。启动 Rattle，将该数据集导入。

对原始数据进行数据探查，发现数据集中各列数据的量值有较大的差异，且每一列数据存在较大的偏倚，因此有必要对数据集进行预处理，即进行规范化处理。根据数据分布特征，选择最小-最大规范化算法进行数据预处理，即采用公式(1)对数据进行处理。

其中，x_{ij} 是数据集中第 i 行第 j 列的数据值，y_{ij} 是其转换后的值，max 函数和 min 函数是分别取数据集的第 j 列的最大值和最小值。数据经过这样处理后，保证了所有数据的值为[0,1]之间的无量纲的值。

$$y_{ij} = \frac{x_{ij} - \min_j\left(x_{ij}\right)}{\max_j\left(x_{ij}\right) - \min_j\left(x_{ij}\right)} \tag{1}$$

4.2. 实验结果与分析

采用 K-means 进行数据集的聚类分析，首先需要确定聚类个数 k，Rattle 为最优聚类个数 k 的确定提供了一个迭代聚类(iterate cluster)选项。实验中选中自动最优选择聚类个数，获得了如图 1 所示的聚类个数与类内平方和的关系图，根据该图，最优的聚类数量为 5。

实验采用聚类个数为 5 的 k-means 聚类算法，其输出结果为：每个聚类的对象个数分别是 1，1，1，4，9，对应的聚类结果的数据对象分别是{上海}，{苏州}，{台州}，{南京，无锡，杭州，宁波}，{常州，南通，扬州，镇江，泰州，嘉兴，绍兴，舟山，湖州}。因篇幅限制，省略 5 个聚类的均值向量和其他相关输出结果。根据数据集的各项指标含义，结合聚类结果，对每一类附上其语义，即五个发展层次。上海独立为一类，是长江三角洲各个城市中绝对主导地位；第二层次为苏州；第三层次城市有南京，无锡，杭州，宁波，第四层次队伍最庞大有常州，南通，扬州，镇江，泰州，嘉兴，绍兴，舟山，湖州九个城市；第五层次城市分别是台州。

为了更直观地显示聚类分析的效果，我们使用层次聚类算法的凝聚层次聚类算法对该问题进行聚类分析。选择对象之间距离计算方法为欧几里德距离公式，见公式(2)，簇凝聚时采用平均距离为簇之间的

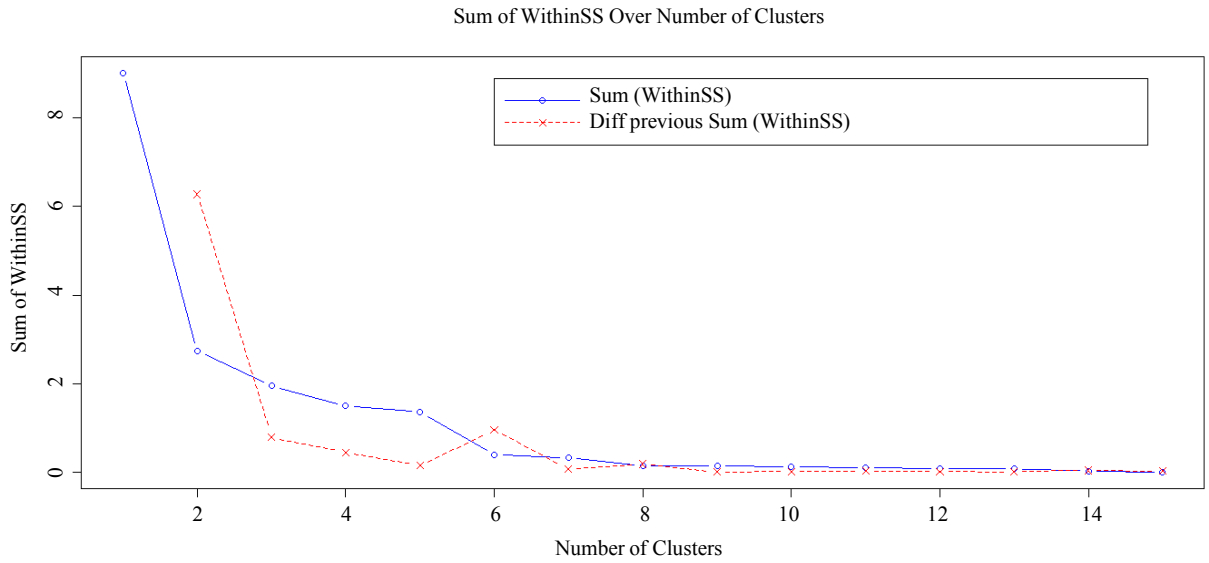

Figure 1. The relation diagram of the number of cluster vs. the WithinSS
图 1. 聚类个数与类内平方和关系图

距离度量算法，见公式(3)。公式(2)中 i，j 表示的对象的编号，n 是对象的属性个数。公式 3 中 C_i，C_j 分别是第 i 和 j 个簇集，n_i，n_j 分别是对应簇集的对象个数。

图 2 为 hclust 对该问题进行聚类分析结果的图显示，其纵坐标的数字表示的城市的编号，横坐标表示簇集之间的距离。在 R 环境下对 log 代码进行修改润色，以便输出图更加直观，聚类的效果图如图 3 所示，图中序号 1~16 分别对应上海，南京，无锡，常州，苏州，南通，扬州，镇江，泰州，杭州，宁波，嘉兴，湖州，绍兴，舟山和台州几个城市名。

$$d(i,j)=\sqrt{\sum_{k=1}^{n}\left(x_{ik}-x_{jk}\right)} \tag{2}$$

$$d_{agv}\left(C_i,C_j\right)=\frac{1}{n_i n_j}\sum_{p\in C_i,p'\in C_j}\left|p-p'\right| \tag{3}$$

从标准化后的数据可以看出，上海除了规模以上产品产销率这一负向指标的数值不为 1，其余 8 项指标均为最大值 1，可见上海的几乎全部经济指标是好于其他城市，聚类效果也发现了这一点。根据谱系图直观发现，16 个城市划分为五个层次类比较合适，也验证了 k-means 算法的最优聚类个数为 5 个。

从以上两个聚类分析发现，不同的聚类方法其聚类结果并不尽相同，且对每一簇集也很难给一个确定的概念来概括，因此关于聚类分析的研究还需要结合实际分析和人们的基本认识来确定簇集名称和聚类效果。

5. 结论及下一步工作

本文运用 R 和 Rattle 软件提供的数据挖掘方法，依据长三角洲经济圈的 16 个城市的 2013 年主要经济指标进行了聚类分析研究，获得一些有趣的信息，尤其对于一些地域相邻、文化相近，经济发展却存在差异城市，应该找出经济发展的途径，更好促进经济发展。比如我们居住的常州市，以前是与苏州、无锡的经济发展水平相当，现在已经被苏州和无锡甩开了，不属于一个梯队的城市了。

下一步我们将采取更多的经济指标，考虑到人口数量等指标，对长江三角洲 16 个城市经济发展进行更深入的研究，并对聚类效果进行显著性差异比较，为城市经济发展的决策者提供更科学的依据。

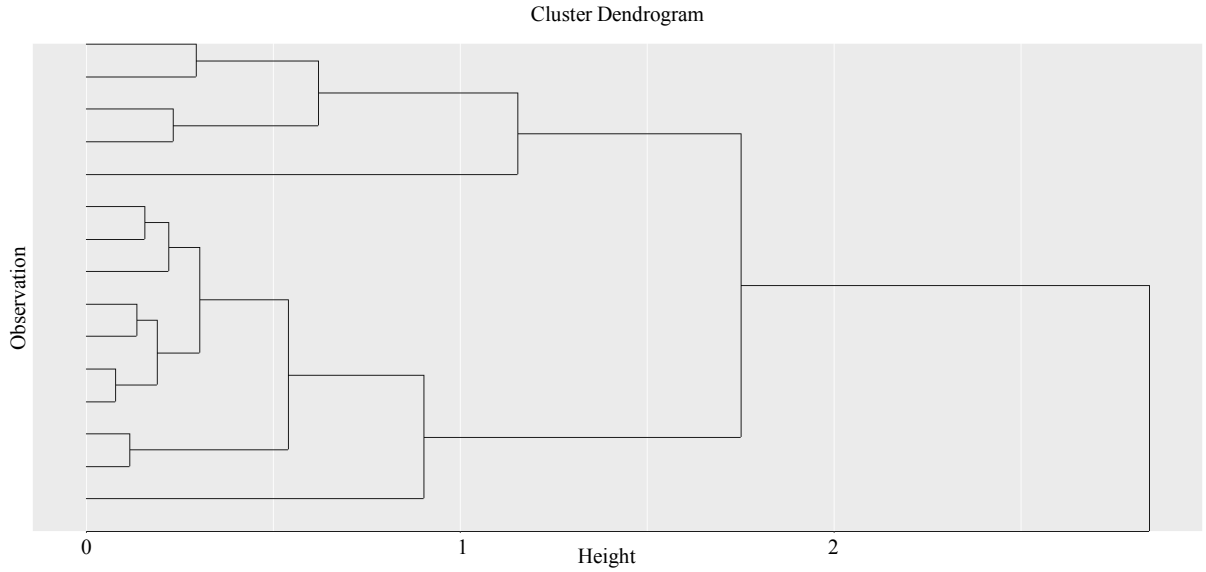

Figure 2. The output result chart of hierarchical clustering algorithm in Rattle
图 2. Rattle 中层次聚类算法输出结果图

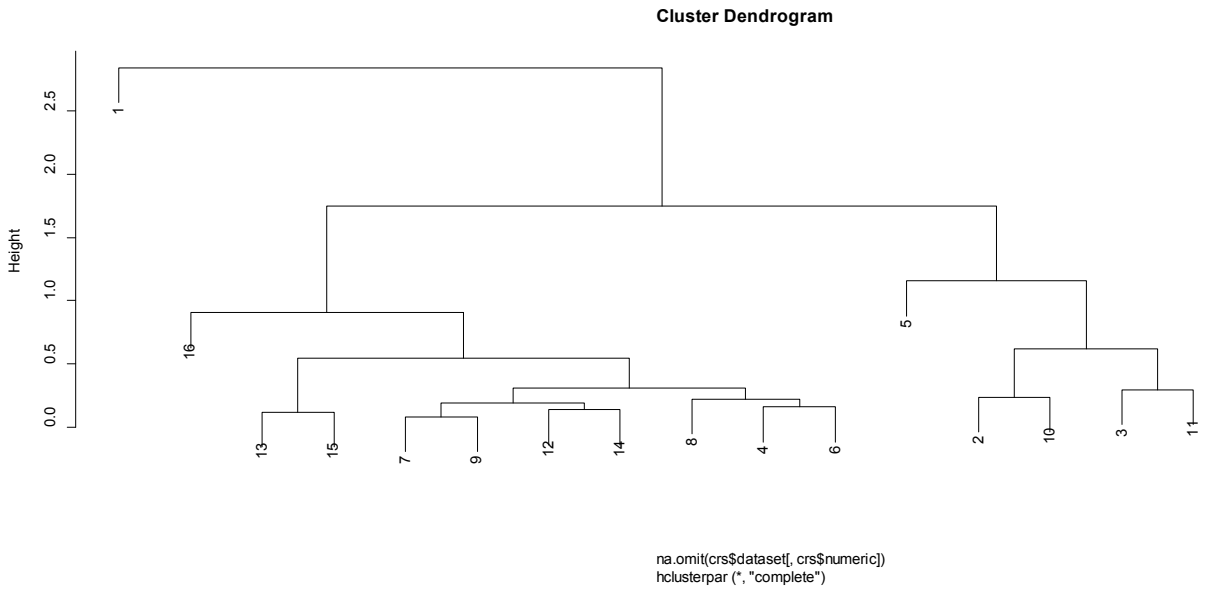

Figure 3. Modified cluster pedigree chart
图 3. 修饰后的聚类谱系图

基金项目

广西师范学院科学计算与智能信息处理广西高校重点实验室基金(GXSCIIP201408)；常州市云计算与智能信息处理重点实验室(CM20123004)；江苏理工学院自然科学预研项目(KYY13041)。

参考文献 (References)

[1] 谭晶荣, 颜敏霞, 邓强, 王健. 产业转型升级水平测度及劳动生产效率影响因素估测—以长三角地区 16 个城市为例[J]. 商业经济与管理, 2012, 1(5): 72-81.

[2] 卓勇良. 走向后长三角时代—长三角发展趋势与主要特征分析[J]. 浙江树人大学学报, 2005, 5(3): 32-37, 42.

[3] Han, J.W. and Kamber, M. 数据挖据概念与技术(第 3 版)[M]. 范明, 孟晓峰,译. 北京: 机械工业出版社, 2012.

[4] R Core Team (2014) R: A Language and Environment for Statistical Computing. http://www.R-project.org/

[5] Zhao, Y.C. (2012) R and Data Mining: Examples and Case Studies. Academic Press, New York.

[6] Torgo, L. (2010) Data Mining with R: Learning with Case Studies. CRC Press, Boca Raton. http://dx.doi.org/10.1201/b10328

[7] Williams, G. (2011) Data Mining with Rattle and R: The Art of Excavating Data for Knowledge Discovery. Springer, New York. http://dx.doi.org/10.1007/978-1-4419-9890-3

[8] Williams, G. (2009) Rattle: A Data Mining GUI for R. *The R Journal*, **2**, 45-55.

[9] 周涛, 陆惠玲. 数据挖掘中聚类算法研究进展[J]. 计算机工程与应用, 2012, 47(24): 100-111.

[10] 罗贤锋, 祝胜林, 陈泽健, 袁玉强. 基于 K-Medoids 聚类的改进 KNN 文本分类算法[J]. 计算机工程与设计, 2014, 35(11): 3864-3867.

[11] 何云斌, 肖宇鹏, 万静, 李松. 基于密度期望和有效性指标的 K-均值算法[J]. 计算机工程与应用, 2014, 49(24): 105-111.

[12] 上海统计局网站. 2013 年 1-12 月长江三角洲城市主要经济指标[DB/OL]. http://www.stats-sh.gov.cn/fxbg/201407/272121.html, 2014-07-31.