

基于主成分分析和 K -means 聚类的平行坐标可视化技术研究

马国峻^{1,2}, 王水波², 裴庆祺², 詹阳²

(1. 西安文理学院信息工程学院, 陕西 西安 710065;

2. 西安电子科技大学综合业务网理论及关键技术国家重点实验室, 陕西 西安 710071)

摘要: 为了解决多维数据的维度过高、数据量过大带来的平行坐标可视化图形线条密集交叠以及数据规律特征不易获取的问题, 提出基于主成分分析和 K -means 聚类的平行坐标 (PCA K P, principal component analysis and k -means clustering parallel coordinate) 可视化方法。该方法首先对多维数据采用主成分分析方法进行降维处理, 其次对降维后的数据采用 K -means 聚类处理, 最后对聚类得到的数据采用平行坐标可视化技术进行可视化展示。以统计局网站发布的数据为测试数据, 对 PCA K P 可视化方法进行测试, 与传统平行坐标可视化图形进行对比, 验证了 PCA K P 可视化方法的实用性和有效性。

关键词: 数据可视化; 平行坐标可视化; 主成分分析; K -means 聚类

中图分类号: TP301

文献标识码: A

doi: 10.11959/j.issn.2096-109x.2017.00189

Research on parallel coordinate visualization technology based on principal component analysis and K -means clustering

MA Guo-jun^{1,2}, WANG Shui-bo², PEI Qing-qi², ZHAN Yang²

(1. School of Information Engineering, Xi'an University, Xi'an 710065, China;

2. State Key Laboratory of Integrated Service Networks, Xidian University, Xi'an 710071, China)

Abstract: In order to solve the problem that parallel coordinate visualization graphic lines are intensive, overlap and rules of data is not easy to be obtained which caused by high dimension and immense amount of multidimensional data. Parallel coordinate visualization method based on principal component analysis and K -means clustering was proposed. In this method, the principal component analysis method was used to reduce the dimensionality of the multidimensional data firstly. Secondly, the data of the dimension reduction was clustered by K -means. Finally, the data of the clustering were visualized by parallel coordinate visualization. The PCA K P visualization method is tested with the data published by the Bureau of Statistics as the test data, and compared with the traditional parallel coordinate visualization graph, the validity and effectiveness of the PCA K P visualization method are verified.

Key words: data visualization, parallel coordinate visualization, principal component analysis, K -means clustering

1 引言

在大数据时代的今天, 人们身边无时无刻不充斥着纷繁复杂的数据。从国家政治到社会民生,

从尖端科学到百姓生活, 从 PC 端到移动设备, 数据时时刻刻都在产生, 大数据已上升为一个国家重要的战略资源。《Science》于 2011 年刊登了“Dealing with Data”^[1]一文, 该文章指出了如何

收稿日期: 2017-07-06; 修回日期: 2017-08-04。通信作者: 马国峻, 1578291722@qq.com

基金项目: 国家自然科学基金资助项目 (No.61373170)

Foundation Item: The National Natural Science Foundation of China (No.61373170)

利用数据资源来推动人类社会的发展。《大数据是国家战略资源》一文指出了大数据的产业价值,以及大数据的价值挖掘^[2]。在海量的数据中,如何迅速理解数据、如何挖掘数据中蕴含的价值信息、如何发现数据中的内在规律,是大数据时代发展中亟待解决的重要问题。

数据可视化技术作为一门新兴学科技术,通过将数据映射为可视化图形,为人们认识数据、理解数据、发现数据规律提供了一种方便、高效的途径。从数据维度的角度出发,可将数据分为低维数据和多维数据。其中低维数据包括一维数据和二维数据,多维数据往往是指数据维数不低于三维的数据。对于低维数据,可以采用折线图、柱状图、饼图等基本图表法进行可视化展示。多维数据可视化方法主要有基于几何的技术^[3]、基于图标的技术^[4]、基于层次的技术^[5]以及基于降维映射的技术^[6]等,其中最典型的多维数据可视化技术是基于几何的技术中的平行坐标可视化技术。针对于多维数据的维数过高、数据量过大带来的平行坐标可视化图形线条密集交叠以及数据规律特征不易获取的问题,本文提出基于主成分分析和K-means聚类的平行坐标可视化方法,并采用可视化案例加以测试验证。

2 平行坐标可视化技术

平行坐标可视化技术是多维空间在二维平面的表示^[7]。它的绘制过程如下:在二维平面中,根据多维数据的维数(假设为 N 维数据),绘制 N 条等间距竖直的平行坐标轴,每一条坐标轴分别表示多维数据中的一维属性,然后将多维数据中的每一维属性值分别映射到对应坐标轴上相应的刻度点上,将各个坐标轴上点依次连接形成一条折线。图1所示是采用平行坐标可视化技术在平面的可视化展示6维数据。

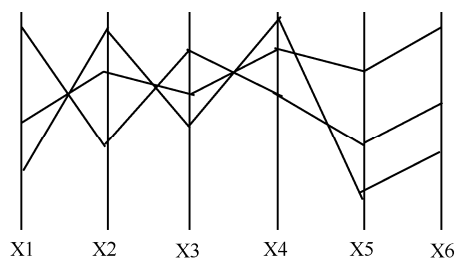


图1 平行坐标

为了在平行坐标可视化图形中更好地展示数据的规律和特征,往往采用平行坐标交互技术对图形进行操作从而发现隐藏的数据规律。常见的平行坐标交互技术主要有以下几种。

1) 维度缩放技术

在数据比较密集时,采用维度缩放技术,对用户关心的数据区域进行放大或者缩小观察,从而发现具体区域数据的取值或数据在某一范围内的变化趋势。

2) 刷技术

在研究平行坐标图形中,刷技术是研究特定属性数据在特定数值范围内的数据变化趋势的一种技术,通过“刷取”关心的数据区域,便可得到该区域数据可视化线条,进而对该区域的数据特征规律进行研究。

3) 交换坐标轴

在平行坐标图形中,当用户想研究2个维度属性之间的关系时,可以交换平行坐标轴的先后排列次序。将待研究的2个属性坐标轴交换到相邻的位置,观察交换后的2个属性之间的关系,从而发现两属性之间的内在关系。

4) 平行坐标维数控制

在采用平行坐标可视化图形分析数据时,有时需要单独研究某几个属性之间的关系。如果直接采用平行坐标图研究数据,其他的属性会对研究结果造成干扰和影响,故需要对暂时不研究的属性删除或者隔离。此时可以采取维数控制,即暂时删除某些属性坐标轴,只需将待研究属性的线条展示出来,这样既可以减小平行坐标图的复杂度,又有利于用户对特定数据进行分析,因此更容易观察出数据变化的规律。

以上是平行坐标交互技术,通过这些交互技术的操作,用户可以从平行坐标可视化图中更好地理解数据以及发现数据的规律和特征。

3 PCAKP 可视化技术研究

3.1 PCAKP 可视化技术提出的背景

在多维数据的维数和数据量适当的情况下,采用平行坐标可视化技术能够对多维数据的每一维属性数据的取值和分布有一个定性和定量的认识。然而在现实生活中往往存在维数过高、数据

量过大的多维数据,如证券市场数据、生物医学特征数据、航空航天数据等。如果在有限宽度的二维平面内对这些数据采用平行坐标可视化技术进行可视化分析,则会出现相邻两条平行坐标轴间距过窄、两条相邻的坐标轴之间过于拥挤、可视化线条交织甚至发生重叠的效应,造成视觉错乱,无法获取数据特征和规律。图2所示是数据维数较高、数据量较大的平行坐标可视化图形,从图中可以看出平行坐标轴的折线分布以及走势的可视化效果较差,无法快速获取数据特征和规律。

实际上,多维数据中的属性与属性之间往往具有线性相关性和相似性。本文先采用降维的思想对多维数据进行线性降维处理,然后对线性降维后的数据进行聚类处理,使同一类中的数据相似度高,不同类之间的数据相似度较低。本文采用主成分分析法(PCA, principle component analysis)对数据进行降维处理,对降维后的数据运用K-means聚类算法进行聚类处理,最后对聚类后的数据进行平行坐标可视化展示,将这种技术称为基于主成分分析和K-means聚类的平行坐标可视化技术,简称PCA-KP可视化技术。

3.2 主成分分析法的基本原理以及数据处理过程

主成分分析对数据降维处理的前提是基于多维数据的属性之间具有较强的相关性,如果属性之间相关性很低或彼此独立互不相关,那么采用主成分分析法对多维数据进行降维处理,则没有意义,得出的结果没有参考价值。如何知道数据样本是否具有相关性以及相关性的多少?统计学中往往常用KMO和Bartlett的球形度相关指标数据来检验统计数据的相关性以及作为采用主成分分析判断的重要依据^[8]。本文借助统计学中

的SPSS软件计算KMO值以及Sig大小。

主成分分析法是一种常用的线性降维方法。PCA的基本思想是采用一组互相独立的变量代表数据的统计性质,每一项变量都可能包含初始数据的多个属性,并且表现数据的某种统计特性^[9],其结果充分反映数据之间的个体差异。PCA采用线性变换方法,将数据变换到新的坐标系中,使任何数据点投影到第一个坐标(第一主成分)的方差为最大,在第二个坐标(第二主成分)的方差为次大,以此类推。因此,PCA方法可以减少数据的维数并保持对方差累计贡献最大的特征。

设有随机变量 $x_1, x_2, \dots, x_p$ 作标准变换

$$C_j = a_{j1}x_1 + a_{j2}x_2 + \dots + a_{jp}x_p, j=1, 2, \dots, p \quad (1)$$

则有

1) $C_1 = a_{11}x_1 + a_{12}x_2 + \dots + a_{1p}x_p$, 且使 $\text{Var}(C_1)$ 最大, 称 C_1 为第一主成分。

2) $C_2 = a_{21}x_1 + a_{22}x_2 + \dots + a_{2p}x_p$, 且使 $\text{Var}(C_2)$ 最大, 称 C_2 为第二主成分。

类似地, 有多个主成分, 两两主成分之间具有互不相关的特性。 $x_1, x_2, \dots, x_p$ 的相关矩阵为 R , 主成分的组合系数 $(a_{21}, a_{22}, \dots, a_{2p})$ 是 R 的第 i 个特征向量, 对应的特征值是这一主成分的方差, 各主成分的方差依次递减, 最重要的是经过变换后样本数据方差和变换后的主成分方差不会发生变化, 表明变换前后信息量没有发生变化。

在分析了PCA的基本原理之后, 下面重点介绍PCA对数据处理的过程, PCA对数据的处理主要有以下7个步骤。

Step1 从原始数据中采集 p 维随机向量, n 个样本 $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})^T, i=1, 2, \dots, n, n > p$ 构造样本矩阵。

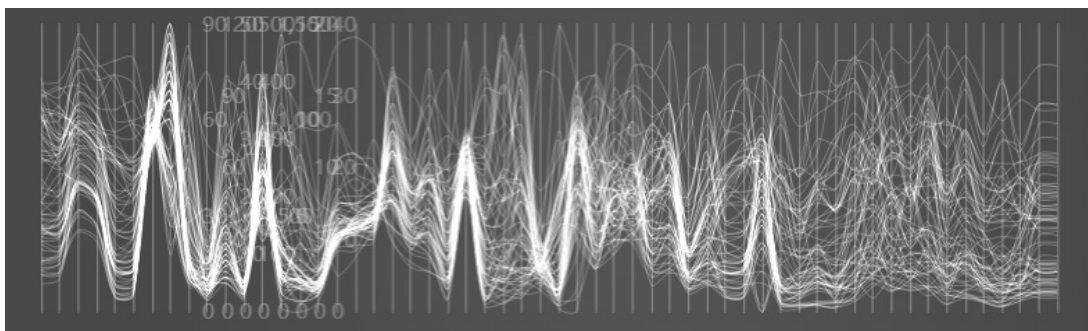


图2 维数过高、数据量过大的平行坐标

Step2 对样本矩阵进行如下标准化变换

$$Z_{ij} = \frac{x_{ij} - \bar{x}_j}{S_j}, i=1,2,\dots,n, j=1,2,\dots,p \quad (2)$$

其中, $\bar{x}_j$ 为第 j 个变量 n 个样本的均值, 即

$$\bar{x}_j = \frac{\sum_{i=1}^n x_{ij}}{n} \quad (3)$$

S_j^2 为 n 个样本的方差, 即

$$S_j^2 = \frac{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2}{n-1} \quad (4)$$

对标准化矩阵 Z , 求出它的相关系数矩阵 R 。

$$R = [r_{ij}]_{p \times p} = \frac{Z^T Z}{n-1} \quad (5)$$

$$r_{ij} = \frac{\sum_{k=1}^n z_{kj} z_{ki}}{n-1}, i, j=1,2,\dots,p \quad (6)$$

Step3 求解样本相关矩阵 R 的特征方程 $|R - \lambda I_p| = 0$, 得到 p 个特征根。

Step4 对每个选中的 $\lambda_j, j=1,2,\dots,m$, 解出方程组 $Rb = \lambda_j b$, 得到单位特征向量 b_j 。

Step5 将标准化后的指标变量转换为对应的主成分 $U_{ij} = z_i^T b_j, j=1,2,\dots,m$, 得到 $1 \sim p$ 主成分分别为: $U_1, U_2, \dots, U_p$ 。

Step6 计算单个主成分贡献率和累计贡献率, 单个主成分贡献率为

$$\frac{\lambda_i}{\sum_{k=1}^p \lambda_k} (i=1,2,\dots,p) \quad (7)$$

累计贡献率为

$$\frac{\sum_{k=1}^i \lambda_k}{\sum_{k=1}^p \lambda_k} (i=1,2,\dots,p) \quad (8)$$

Step7 选择 Step6 累计贡献率大于总方差的 85% 的前 k 个主成分, 按照式(9), 由 k 个主成分的特征向量 $w_i (i=1,2,\dots,k)$ 得到 k 个独立的线性组合新变量, 从而计算出各个主成分得分。

$$\begin{cases} U_1 = a_{11}x_1 + a_{12}x_2 + \dots + a_{1p}x_p \\ U_2 = a_{21}x_1 + a_{22}x_2 + \dots + a_{2p}x_p \\ \vdots \\ U_k = a_{k1}x_1 + a_{k2}x_2 + \dots + a_{kp}x_p \end{cases} \quad (9)$$

3.3 K-means 聚类算法

聚类是根据某种规则, 用特定方法将数据集划分为不同类别, 使同一个类别中的数据之间的相似度较高, 同类中的数据之间的相差较大^[10]。

在聚类中, 常采用距离表现样本数据之间的亲疏关系或相似度, 按照不同的计算方法, 距离的分类有欧式 (Euclidian) 距离、绝对值距离、切比雪夫 (Chebychev) 距离、马氏 (Mahalanobis) 距离等^[11], 最常用的距离是欧式距离, 图 3 是二维空间的数据点①和数据点②之间的欧式距离, 可用式(10)表示。

$$d_{21} = \sqrt{(x_{21} - x_{11})^2 + (x_{22} - x_{12})^2} \quad (10)$$

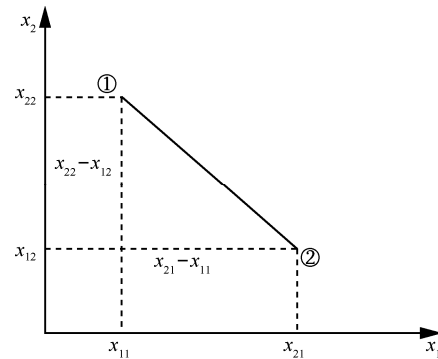


图3 二维平面欧式距离

多维空间两点 i, j 的欧式距离可以表示为

$$\begin{aligned} d_{ij} &= \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \dots + (x_{ip} - x_{jp})^2} \\ &= \left[\sum_{k=1}^p (x_{ik} - x_{jk})^2 \right]^{\frac{1}{2}} \end{aligned} \quad (11)$$

K-means 聚类算法思路简单且收敛速度快, 也称为快速聚类算法^[12], 是一种最常见的聚类算法之一, 它的主要思想是首先从 n 个数据样本集合中选取 k 个数据样本点作为初始聚类的中心点; 计算剩余样本点与初始中心点的距离, 并将剩余的样本点归于距离它们最近的初始中心点一类中; 重新计算每个聚类的平均值并作为聚类的中心点; 重复执行这一过程, 直到聚类中心不再变化为止。K-means 聚类算法流程如图 4 所示。

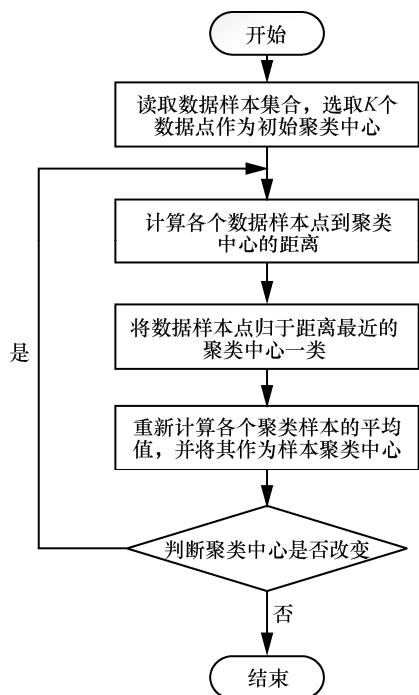


图 4 K-means 聚类流程

K-means 聚类算法基本步骤如下。

输入: n 个数据样本的集合以及聚类的数目 k

输出: k 个聚类结果

基本步骤如下。

Step1 在 n 个数据样本集合中选取 k 个数据样本点作为初始聚类的中心点。

Step2 计算各个样本点到 k 个聚类中心距离。

Step3 将各个样本点归于距离各自最近的聚类中心一类。

Step4 重新计算各个聚类的平均值, 并将聚类的平均值作为该类的聚类中心。

Step5 判断各个聚类中心是否发生改变, 如果聚类中心不再改变, 表明聚类已经完成。如果聚类中心发生改变, 则继续执行 Step2~Step4, 直到每个聚类中心不再改变或收敛为止。

3.4 PCAKP 可视化技术

3.2 节和 3.3 节中分别研究了主成分分析法以及 K-means 聚类算法, 本节重点介绍基于 PCAKP 可视化技术。

首先对数据进行 KMO 和 Bartlett 的球形度检验, 本文借助 SPSS 对数据进行相关性检验。假设数据满足线性相关性, 对数据进行主成分分析降维仿真处理, 对降维后得到的主成分得分采用 K-means 聚类算法进行仿真, 得到聚类结果。最后运用平行坐标可视化方法对不同聚类结果用不同的线型加以区分展示。本文对数据的主成分分析仿真和 K-means 聚类仿真均采用 SPSS 软件。

为了突出 PCAKP 的高维处理效果, 本文以国家统计局网站中《中国统计年鉴 2016》城镇单位就业人员中的数据为仿真实验数据, 该数据的特点是数据维度高, 且数据来源具有真实权威性, 结果可靠。图 5 是全国按行业分, 城镇单位就业人员数表部分截图。

图 5 所示的这张表中对就业人员按行业进行了详细的划分, 一共有 19 个行业, 分别是“农、

中国统计年鉴 2016																		
4-5 按行业分城镇单位就业人员数(年底数)																		
年份地区	合计	农、林、牧、渔业	采矿业	制造业	电力、热力、燃气及水生产和供应业	建筑业	批发和零售业	交通运输、仓储和邮政业	住宿和餐饮业	信息传输、软件和信息技术服务业	金融业	房地产业	租赁和商务服务业	科学研究和技术服务业	水利、环境和公共设施管理业	居民服务、修理和其他服务业	教育	其他
2003	10969.7	484.5	488.3	2980.5	297.6	833.7	628.1	636.5	172.1	116.8	353.3	120.2	183.5	221.9	172.5	52.8	1442.8	
2004	11098.9	466.1	500.7	3050.8	300.6	841.0	586.7	631.8	177.1	123.7	356.0	133.4	194.4	222.1	176.1	54.2	1466.8	
2005	11404.0	446.3	509.2	3210.9	299.9	926.6	544.0	613.9	181.2	130.1	359.3	146.5	218.5	227.7	180.4	53.9	1483.2	
2006	11713.2	435.2	529.7	3351.6	302.5	988.7	515.7	612.7	183.9	138.2	367.4	153.9	236.7	235.5	187.0	56.6	1504.4	
2007	12024.4	426.3	535.0	3465.4	303.4	1050.8	506.9	623.1	185.8	150.2	389.7	166.5	247.2	243.4	193.5	57.4	1520.9	
2008	12192.5	410.1	540.4	3434.3	306.5	1072.6	514.4	627.3	193.2	159.5	417.6	172.7	274.7	257.0	197.3	56.5	1534.0	
2009	12573.0	373.7	553.7	3491.9	307.7	1177.5	520.8	634.4	202.1	173.8	449.0	190.9	290.5	272.6	205.7	58.8	1550.4	
2010	13051.5	375.7	562.0	3637.2	310.5	1267.5	535.1	631.1	209.2	185.8	470.1	211.6	310.1	292.3	218.9	60.2	1581.8	
2011	14413.3	359.5	611.6	4088.3	334.7	1724.8	647.5	662.8	242.7	212.8	505.3	248.6	286.6	298.5	230.3	59.9	1617.8	
2012	15236.4	338.9	631.0	4262.2	344.6	2010.3	711.8	667.5	265.1	222.8	527.8	273.7	292.3	330.7	243.8	62.1	1653.4	
2013	18108.4	294.8	636.5	5257.9	404.5	2921.9	890.8	846.2	304.4	327.3	537.9	373.7	421.9	387.8	259.2	72.3	1687.2	
2014	18277.8	284.6	596.5	5243.1	403.7	2921.2	888.6	866.6	289.3	336.3	566.3	402.2	449.4	408.0	269.1	75.4	1727.3	
2015	18062.5	270.0	545.8	5068.7	396.0	2796.0	883.3	854.4	276.1	349.9	606.8	417.3	474.0	410.6	273.3	75.2	1736.5	
北京	777.3	3.9	5.3	92.2	8.2	45.3	77.1	60.0	29.8	68.0	47.2	42.2	80.1	59.3	10.2	9.0	47.3	
天津	294.8	0.5	6.5	110.8	4.5	29.5	17.8	15.0	5.2	4.4	12.1	7.3	8.2	11.3	4.1	11.0	18.0	
河北	643.6	4.2	24.6	140.9	18.7	84.4	27.1	29.2	5.8	8.8	29.9	11.0	13.4	14.8	11.8	1.7	89.1	
山西	440.3	1.8	95.3	65.4	11.7	33.3	17.5	24.1	4.2	5.5	16.8	3.6	8.9	7.5	9.6	0.6	51.7	
内蒙古	298.3	23.6	17.9	46.7	14.2	21.6	9.6	20.6	4.1	5.0	11.6	5.4	4.6	6.3	8.1	0.9	35.1	

图 5 城镇单位就业人员数表部分截图(来源:中国统计年鉴 2016)

林、牧、渔业”“采矿业”“制造业”“电力、热力、燃气及水生产和供应业”“建筑业”“批发和零售”“交通运输仓储和邮政业”“住宿和餐饮业”“信息传输和软件技术”“金融业”“房地产”“租赁和商务服务业”“科学研究和技术服务业”“水林、环境和公共设施管理”“居民服务、修理和其他服务业”“教育”“卫生和社会工作”“文化、体育、娱乐”“公共管理和社会保障和社会组织”。本文选取北京、天津、河北等 31 笔省市数据进行可视化研究，每一笔数据都有 19 维数据。对于 31 笔 19 维省市数据直接采用平行坐标可视化技术展示，得到的可视化图形如图 6 所示，这幅平行坐标可视化图形由于数据维数较多，出现了平行坐标轴较拥挤、可视化线条杂乱无章、可视化效果不理想、不能很快直接地从这幅图中获取数据的规律。

首先，本文借助 SPSS 软件对数据 KMO 和 Bartlett 检验进行相关性检验，得到数据 KMO 值为 0.845，Sig 值为 0，表明该数据非常适合 PCA 处理。然后，采用主成分分析法对数据进行处理，按照特征值大于 1 的提取方法提取 3 个主成分，3 个主成分累计贡献率达到 85.35%，从成分矩阵中可以看出，第一主成分与“批发和零售”“交通运输仓储和邮政业”“住宿和餐饮业”“信息传输和

软件技术”“金融业”“房地产”“租赁和商务服务业”“科学研究和技术服务业”“水林、环境和公共设施管理”“居民服务、修理和其他服务业”“教育”“卫生和社会工作”“文化、体育、娱乐”“公共管理和社会保障和社会组织”等 14 个变量相关性最高；第二主成分与“制造业”“电力、热力、燃气及水生产和供应业”相关性最高，第三主成分与“农林牧渔业”相关性最高。由此可知，第一主成分集中代表了第三产业的就业信息，第二主成分集中代表了第二产业的就业信息，第三主成分集中代表了第一产业的就业信息。不妨将第一主成分定义为第三产业，第二主成分定义为第二产业，第三主成分代表第一产业。对主成分的得分采用平行坐标技术进行展示，如图 7 所示。

从图 7 中可以看出可视化线条清晰，而且没有交织的现象，有效解决了杂乱无章的可视化效果不理想的问题。

完成主成分分析降维后，接下来，对降维后的主成分得分数据进行 *K*-means 聚类。本文将 31 笔省份数据聚为 4 类，即采用 *K*-means 算法聚类时设置聚类个数为 4。经过 *K*-means 聚类算法处理，聚类结果如下。第一聚类包括：北京、上海 2 笔数据。第二聚类包括：江苏、浙江、福建、广东、海南、重庆、西藏、青海、

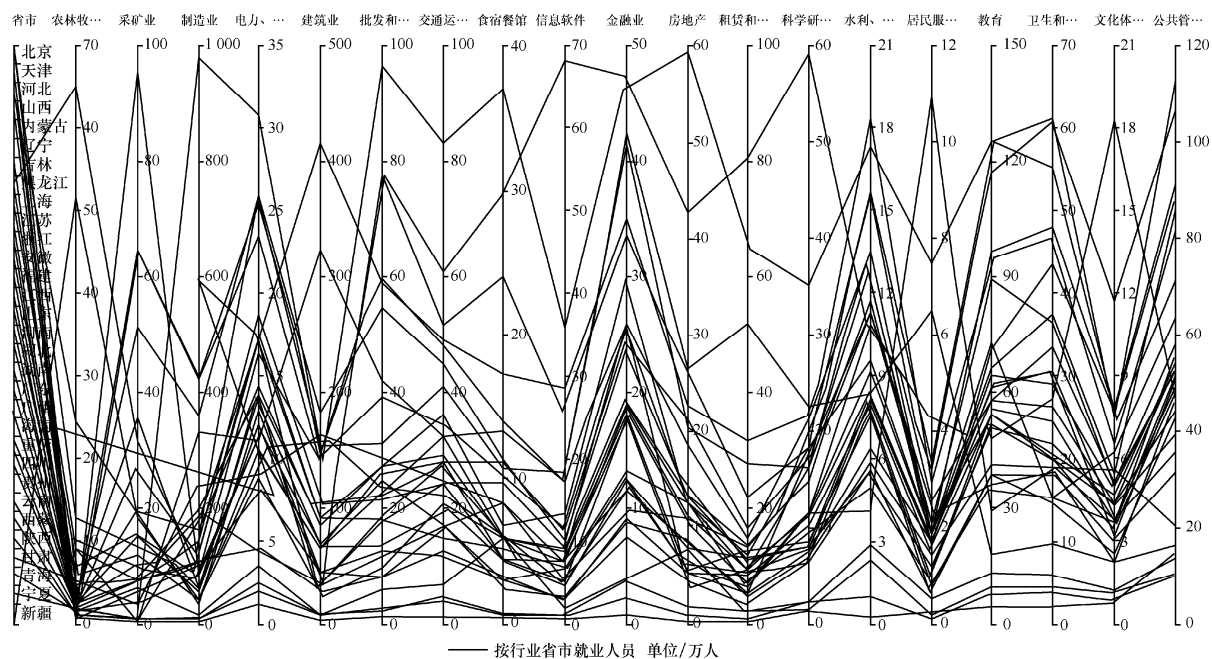


图 6 就业人员数平行坐标可视化图形

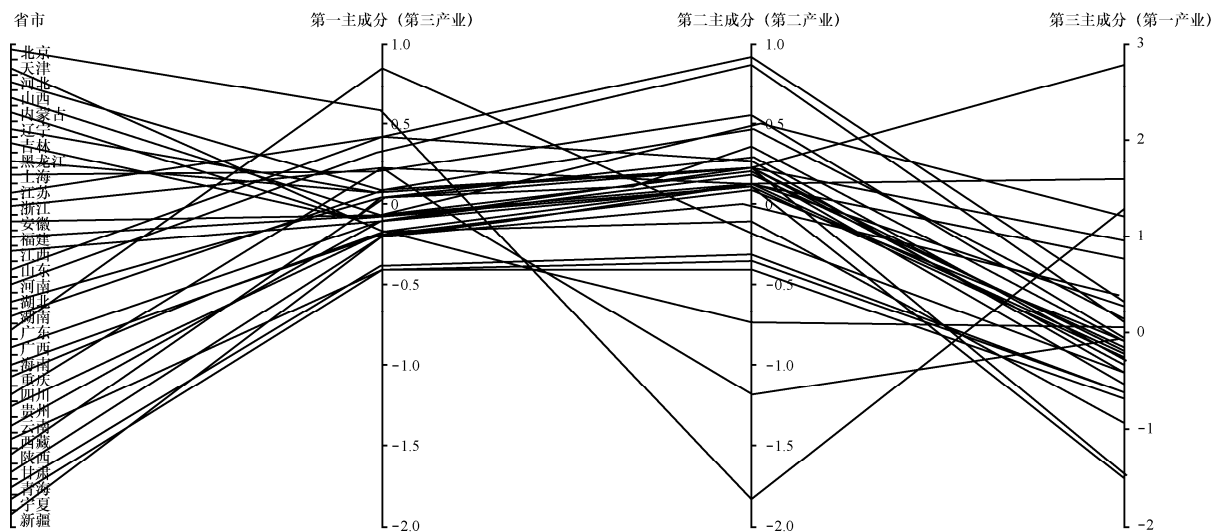


图7 降维后的平行坐标可视化图形

宁夏等9笔数据。第三聚类包括：天津、河北、内蒙古、吉林、安徽、江西、山东、河南、湖北、湖南、广西、四川、贵州、云南、陕西、甘肃16笔数据。第四聚类包括山西、辽宁、黑龙江、新疆4笔数据。对4个聚类数据绘制平行坐标可视化图形时，分别用不同的线型表示不同的聚类，如图8所示，经过K-means聚类后的数据在平行坐标图中展示它的层次规律明显，特点突出。

下面对4个聚类的可视化图形分别做如下分析说明。

图9是聚类1的数据平行坐标可视化图形，

可以看出，聚类1中的北京和上海数据在第二产业数值较小，在第三产业数值较大，表明北京和上海第三产业就业人员较多，第二产业就业人员较少，说明该地区第三产业比较发达。

图10是聚类2的数据平行坐标可视化图形，可以看出，聚类2中的数据在第三产业和第二产业取值较大，而且第三产业的数值较第二产业数值大，在第一产业取值较小，说明江苏、浙江、福建、广东等聚类2中的省份在第二产业和第三产业的就业人员较多，该地区第二产业最为发达。

图11是聚类3的数据平行坐标可视化图形，可以看出，聚类3中的数据在第三产业和第二产

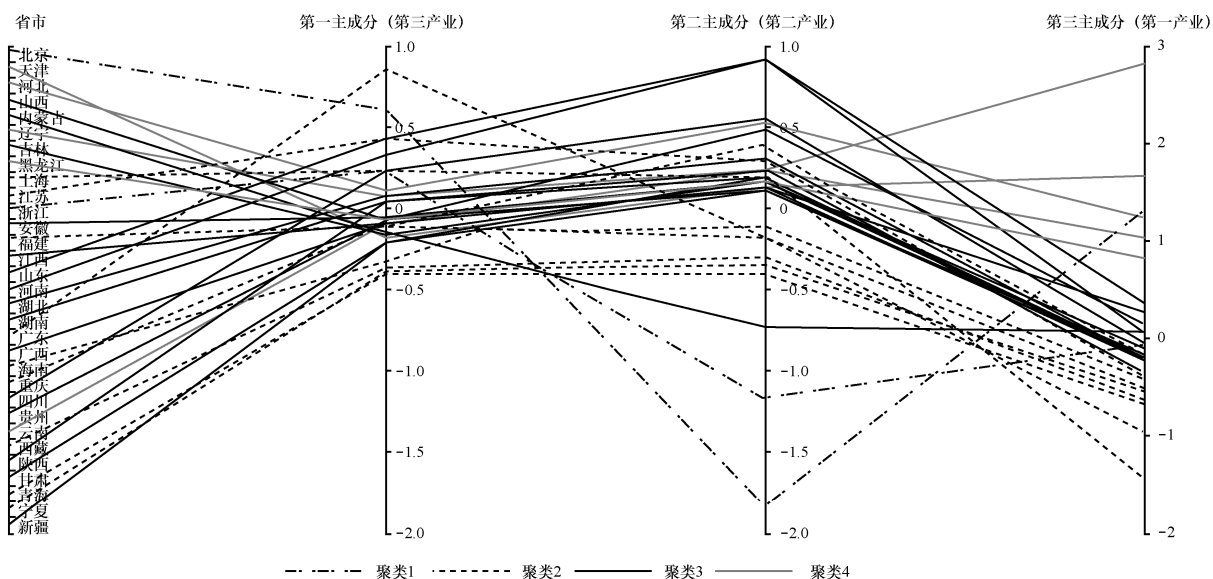


图8 PCAKP 可视化图形

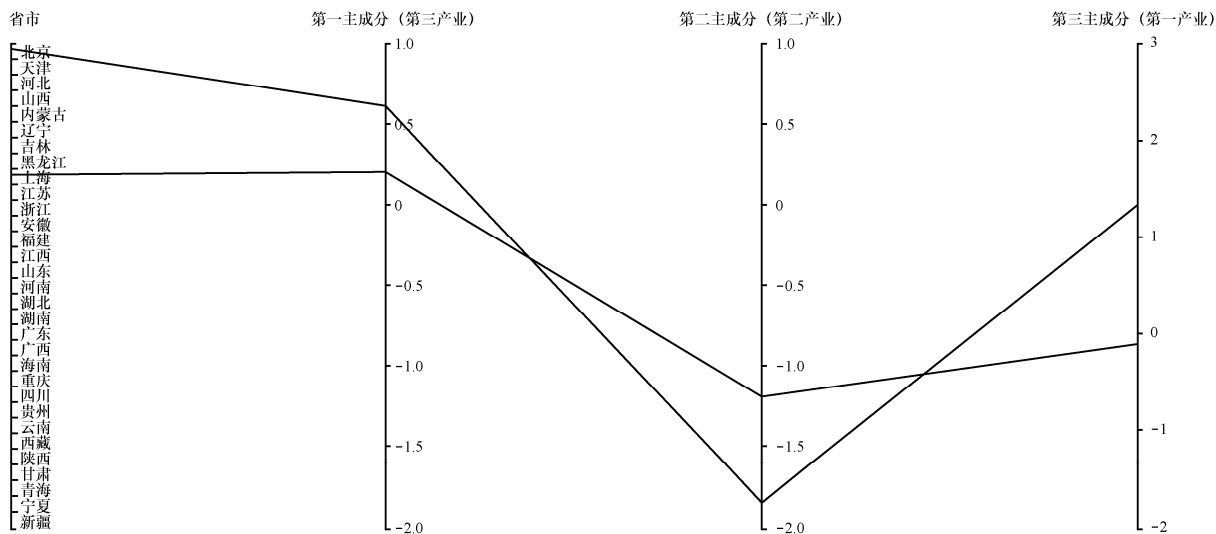


图 9 聚类 1 中的平行坐标可视化图形

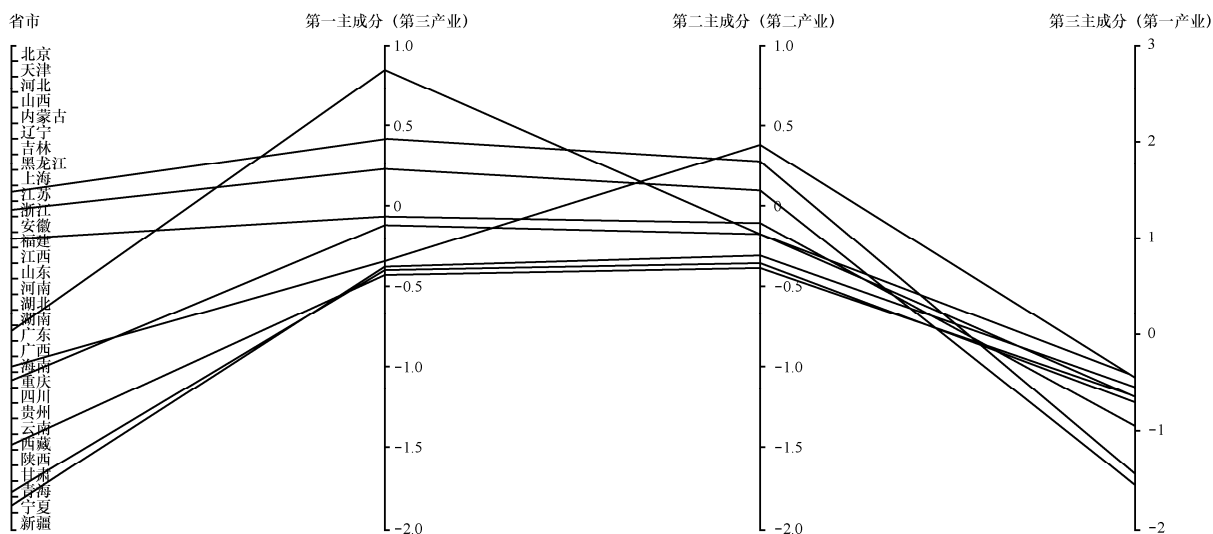


图 10 聚类 2 中的平行坐标可视化图形

业取值较大，而且第二产业的数值较第三产业数值大，在第一产业取值较小，说明了天津、河北、内蒙古、吉林等聚类 3 中的省市在第二产业和第三产业的就业人员较多，第二产业人员数大于第三产业人员数，而在第一产业就业人员较少，表明该地区第二产业最为发达，而第一产业的发展滞后。

图 12 是聚类 4 的数据平行坐标可视化图形，可以看出，聚类 4 中的数据在第一产业取值较大，而且第二产业的数值较第三产业数值小，说明山西、辽宁、黑龙江、新疆等聚类 4 中的省市在第一产业就业人员较多，第二产业人员数和第三产业人员数较少，表明该地区的第一产业最为发达，

第二、第三产业滞后。

以上就是对全国按行业就业人数采用的 PCAKP 可视化技术分析的结果。与图 5 中的原始数据相比，经过 PCAKP 可视化技术得到的图形直观反映了图 5 中数据的规律和特征；与直接采用平行坐标可视化技术得到的图 6 比较，PCAKP 可视化技术得到的图形可视化效果更加理想。这一仿真实验表明，通过 PCAKP 可视化技术得到的平行坐标可视化图形，有效解决了多维数据的维数较高、数据量较大带来的问题，用户更容易从平行坐标可视化图中获取数据的整体规律和特征，证明了该技术的有用性和实效性。

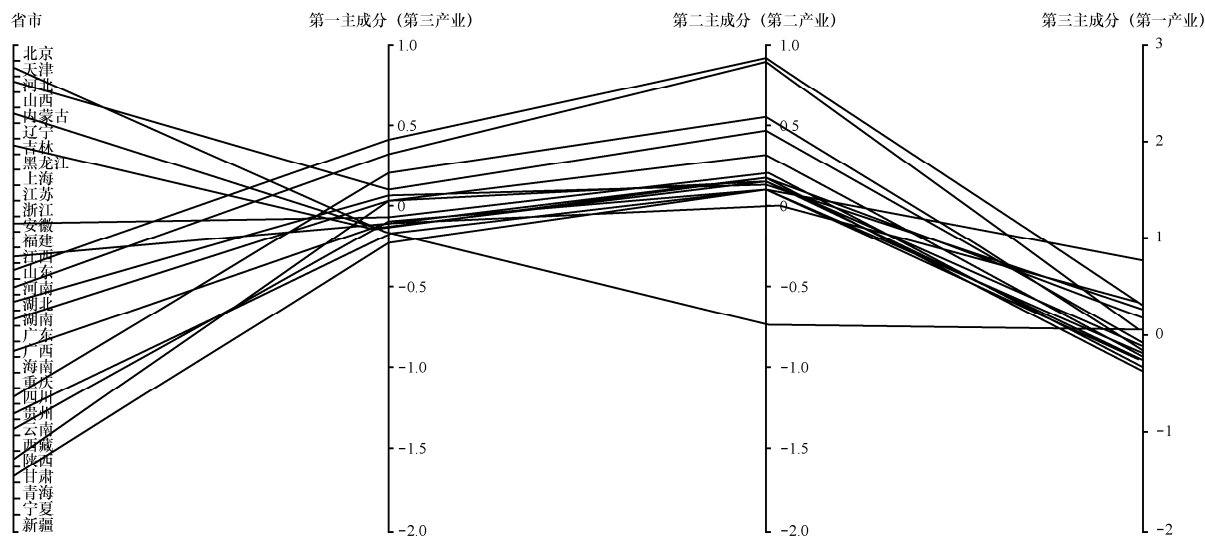


图 11 聚类 3 中的平行坐标可视化图形

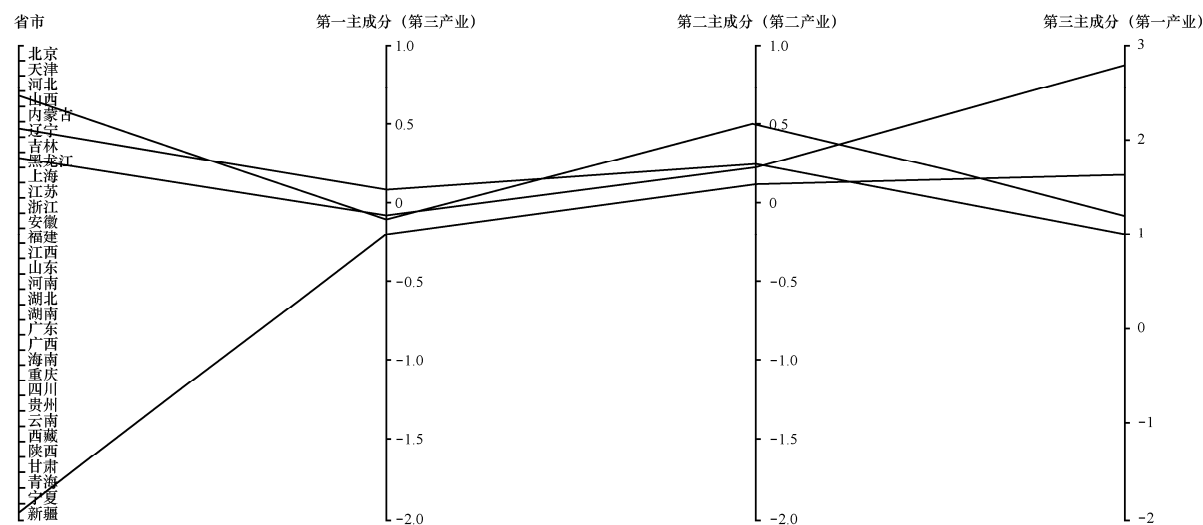


图 12 聚类 4 中的平行坐标可视化图形

4 结束语

本文首先研究了平行坐标可视化技术及其交互技术。针对多维数据的维数过高、数据量过大导致平行坐标可视化图形效果不理想以及无法快速获取数据规律特征的问题,本文提出了基于主成分分析和 K -means 聚类的平行坐标可视化技术。首先对多维数据采用主成分分析法进行降维处理,减小维度过高带来的可视化线条拥挤、交叠等可视化不理想情况;然后对降维后得到的主成分得分采用 K -means 聚类算法进行聚类,从而帮助用户快速获取数据规律特征。本文以国家统计局网站中《中国统计年鉴 2016》城镇单位就业人员中的数据为仿真实验数据,测试了 PCAKP

可视化技术的实用性和有效性。

参考文献:

- [1] REICHMAN O J, JONES M B, SCHILDHAUER M P. Challenges and Opportunities of open data in ecology[J]. Science, 2011, 331(6018): 703-705.
- [2] 怀进鹏. 大数据是国家战略资源[J]. 中国经济和信息化, 2013(8): 49-50.
HUALI J P. Big Data: national strategic resource[J]. China Economy & Information, 2013(8): 49-50.
- [3] 高芳. 平行坐标可视化技术的度量模型研究[D]. 北京: 北京交通大学, 2009.
GAO F. Research on measurement model of parallel coordinates visualization technique[D]. Beijing: Beijing Jiaotong University, 2009.
- [4] 胡俊, 黄厚宽, 任东怀. 一种基于图标技术的可视化方法与应用[J]. 北京交通大学学报, 2006, 30(5): 10-14.

- HU J, HUANG H K, REN D H. An icon_based visualization approach and its applications[J]. Journal of Beijing Jiaotong University, 2006, 30(5): 10-14.
- [5] ITOH T, YAMAGUCHI Y, IKEHATA Y, et al. Hierarchical data visualization using a fast rectangle-packing algorithm[J]. IEEE Transactions on Visualization & Computer Graphics, 2004, 10(3): 302-313.
- [6] CHEN P, JIAO L, LIU F, et al. Dimensionality reduction of hyper-spectral imagery using sparse graph learning[J]. IEEE Journal of Selected Topics in Applied Earth Observations & Remote Sensing, 2016, PP(99):1-17.
- [7] INSELBERG A. Parallel coordinates[J]. Encyclopedia of Database Systems, 2009, 43(4): 199-233.
- [8] 解坤, 张俊芳. 基于 KMO-Bartlett 典型风速选取的 PCA-WNN 短期风速预测[J]. 发电设备, 2017, 31(2): 12-36.
- XIE K, ZHANG J F. Short-term wind speed forecasting using PCA-WNN based on KMO-bartlett typical wind speed selection[J]. Power Equipment, 2017, 31(2): 12-36.
- [9] WANG J. Forecasting stock market indexes using principle component analysis and stochastic time effective neural networks[J]. Neuro Computing, 2015, 156(C): 68-78.
- [10] HASHEMINEJAD S M H, JALILI S. CCIC: clustering analysis classes to identify software components[J]. Information & Software Technology, 2015, (57): 329-351.
- [11] 段明秀. 层次聚类算法的研究与应用[D]. 长沙: 中南大学. 2009.
- DUAN M X. Research and application of hierarchical clustering algorithm [D]. ChangSha: Central South University. 2009.
- [12] JAIN A K. Data clustering: 50 years beyond K -means[J]. Pattern Recognition Letters, 2010, 31(8): 651-666.

作者简介:



马国峻(1978-), 男, 安徽无为, 人, 西安文理学院讲师, 主要研究方向为数字内容保护、智能移动应用开发、区块链应用与安全。



王水波(1990-), 男, 湖北黄梅人, 西安电子科技大学硕士生, 主要研究方向为 Web 前端。



裴庆祺(1975-), 男, 广西玉林人, 西安电子科技大学教授、博士生导师, 主要研究方向为信任管理、无线网络安全、区块链安全。

詹阳(1977-), 男, 陕西杨凌人, 西安电子科技大学讲师, 主要研究方向为信息安全、区块链应用。