学校代码 学号

分类号 密级

本科毕业论文（设计）

|  |
| --- |
| ClusterEasy:聚类可视化分析工具 |

院（系）名称 计算机学院

专业名称 计算机科学与技术

年 级 **2015**级

学生姓名 韩华旭

指导教师 王树良

**2019**年**4**月**18**日

ClusterEasy:聚类可视化分析工具

摘 要

在大数据爆发的信息时代，数据的价值被与来越多的人重视，科研人员研究出了许多方法可以在数据中挖掘出有价值的的信息，其中聚类作为一个被广泛应用的非监督学习方法，可以有效的将数据基于相似性聚合到一起，获得数据的结构化特征。在聚类的过程中，数据分析人员往往需要挑选合适的算法，尝试不同的参数来找到更好的聚类方案，却并没有一种可以快速聚类可视化、交互式、界面友好的工具可以满足数据分析人员的这种需求，为此，我们开发了ClusterEasy(CE)，致力于简化数据分析人员在聚类过程中所需要的操作，使其能够剖析数据内部结构，更加有效的对数据聚类，降低聚类分析的难度。CE是一个基于Web的聚类可视化分析工具，支持高维数据降维和聚类，为探索降维和聚类后的数据结构提供了可视化支持，并在交互式处理数据方面提供了友好的用户界面，具备数据动态变化效果，方便用户追踪数据变化、还可以帮助用户探寻不同维度之间相关性和不同维度对聚类或降维的影响，CE使用的前向投影算法和Brushing and linking技术是实现交互式分析数据的基础。

关键词：聚类，数据可视化，降维

ClusterEasy: Visual Analysis Tool for Clustering

Author:Han Huaxu

Tutor:Wang Shuliang

Abstract

In the information age of big data explosion, more and more people pay attention to the value of data. Researchers have developed many methods to mine valuable information in data. Clustering, as a widely used unsupervised learning method, can effectively aggregate data based on similarity and obtain structured features of data. In the process of clustering, data analysts often need to select appropriate algorithms and try different parameters to find better clustering schemes, but there is no fast clustering visualization, interactive, user-friendly tool to meet the needs of data analysts. Therefore, we developed Cluster Easy (CE) to simplify the clustering process of data analysts. The operation needed in the process enables it to analyze the internal structure of data, cluster data more effectively and reduce the difficulty of clustering analysis. CE is a Web-based Clustering Visualization Analysis tool, which supports dimensionality reduction and clustering of high-dimensional data, provides visualization support for exploring the data structure after dimensionality reduction and clustering, and provides a friendly user interface in interactive data processing. CE has the effect of dynamic data changes, facilitates users to track data changes, and helps users to explore the correlation between different dimensions. The influence of different dimensions on clustering or dimensionality reduction, the forward projection algorithm and Brushing and Linking technology used by CE are the basis of interactive data analysis.

Keywords：Clustering, Data Visualization, Dimension Reduction

目录

[1 绪论 1](#_Toc1675961854)

[1.1 研究背景和目的 1](#_Toc1029659026)

[1.2 国内外聚类可视化研究现状 2](#_Toc1321043494)

[1.3 论文研究内容及结构安排 3](#_Toc1933276132)

[2 ClusterEasy的设计 5](#_Toc797170191)

[2.1 设计理念 5](#_Toc506169470)

[2.1.1 支持快速调参聚类 5](#_Toc1955501335)

[2.1.2 直观的视觉编码 6](#_Toc2050578826)

[2.1.3 多种数据视图 6](#_Toc775048930)

[2.1.4 支持交互式探索数据 7](#_Toc702377673)

[2.2 视图 7](#_Toc998846215)

[2.2.1 表格 7](#_Toc1862238651)

[2.2.2 投影视图 8](#_Toc1817194626)

[2.2.3 聚类视图 8](#_Toc1308911204)

[2.2.4 数据统计视图 9](#_Toc133319567)

[2.2.5 编辑视图 10](#_Toc1396528476)

[2.3 交互功能 11](#_Toc841898523)

[2.3.1 特征值的隐藏与显示、删除与添加 11](#_Toc2085461870)

[2.3.2 行的添加与删除 12](#_Toc1661624365)

[2.3.3 数据的过滤 12](#_Toc1863113628)

[2.3.4 brushing and linking 13](#_Toc2136938517)

[2.4 降维 13](#_Toc1520121924)

[2.5 前向投影 14](#_Toc2128131572)

[2.6 变化曲线 15](#_Toc1864208437)

[2.7 视觉编码 16](#_Toc845450573)

[2.8 数据采样 17](#_Toc2084169477)

[2.9 系统信息 17](#_Toc72451660)

[结 论 19](#_Toc1917757837)

[参考文献 20](#_Toc1646403214)

# 1 绪论

## 1.1 研究背景和目的

聚类这个术语最早被Tryon[9]使用 ，聚类是一个基于数据之间的相似性或者按照距离远近，自动地将更相似的或者离得比较近的无标签数据聚合在一起的过程，最终能使整个数据集被划分成互不相交的几个集合，在每个集合中的数据要比集合之间的数据更要相似，数据分析人员通过聚类这一简单的操作，可以找出数据中隐藏的某种结构。聚类虽然是比较流行的一种非监督学习方法，但是聚类过程并不像运行一遍算法那样简单，数据分析人员不光要选择适合的算法，往往还需要尝试不同的参数，时刻了解聚类结果，来找到合适的聚类方案。这个过程简单地看来类似于做实验，不断的尝试新的参数，验证猜想，但是，聚类结果的好坏并不能通过数据直接反映出来，这时候，我们经常采用的一种方法就是可视化技术。可视化技术在数据分析中是非常受欢迎的一种解读信息方式，它可以将信息用颜色和图形元素编码在视觉上表达出来，使数据分析人员非常直观地感受到数据中有价值的信息，发现数据分布的特点，找到数据内的结构特征。

目前，已经有许多优秀的数据可视化工具，比如有需要依赖某种编程语言可视化的Matlab，R和python，也有可以直接导入数据可视化的WebDataRocks和Goole Charts，这些工具可以有效的可视化信息，满足我们大部分需求，但是这些工具并未针对聚类可视化做优化，不具有使数据分析人员快速聚类并可视化分析的功能。我们希望数据分析人员在聚类可视化分析中，能够快速的聚类数据、清晰到观察聚类结果，在交互式操作中，可以轻松获取数据的特征和结构等信息，为此，我开发了ClusterEasy (CE)。CE在快速聚类、可视化、交互和数据统计四个方面进行了设计，引入了excel表格来显示原生数据、散点图和热图来显示聚类结果、直方图和表格来显示统计的数据，为了使用户清晰的追踪到数据变化，提高用户交互分析数据的体验，工具使用了brushing and linking技术，可以让不同视图中的数据变化实时反映在其他视图上。

## 1.2 国内外聚类可视化研究现状

格式塔理论[10]表明人类的眼睛会自动的将具有相似性的对象进行聚类，通过人机交互，可以充分的利用人类的感知、推理以及知识，将图形符号所呈现来的数据结构特征及关联特征转化为可利用的、有用的信息，所以，在聚类分析的早期阶段，就已经有科研人员进行聚类和可视化技术结合的相关研究，并在数据的可视化方式上，有许多不同的尝试。第一种可视化数据的方式是采用树形图，比如J.Seo和B.Shneiderman等人开发的Hierarchical Clustering Explorer(HCE)[4]。HCE属于较早的一批使用了可视化技术的聚类分析工具，它使用了树形图和热图来编码基因数据的分层聚类的结果，树形图可以有效的展示不同簇之间的距离，但是树形图的可视化方式受限于屏幕和像素分辨率的大小，对于大数据的分析不太友好。第二种是采用平行坐标[11][17]，平行坐标在保留数据的高维特性上弥补了树形图的不足，在探索不同特征之间的相关性上非常有用，但是一方面对于许多用户来说，理解平行坐标仍然存在一些困难，另一方面，Holten[21]的研究表明，平行坐标在可视化多个类的时候表现的不太好。此外，还有散点图[15]和RadViz[16]方法被用来可视化聚类算法的结果。

由于人类的视觉系统并不擅长比对不同视图中的元素，研究人员提出了许多技术来比对不同算法产生的聚类结果[4][13][14][20]，DICON[12]精心设计了聚类图标来编码多个数据属性和统计信息，Van Long和Linsen[17]使用不同颜色在一个平行坐标中显示聚类树和精确值之间的关系，HCE和XCluSim[18]支持基因数据聚类结果的比对。

支持交互功能的可视化工具可以使数据分析人员更有效的完成数据分析工作，在聚类可视化分析工具的发展中，融入交互功能早已成为了研究人员的热门研究方向。Matchmaker[5]依赖[4]提供的技术实现了通过分组数据的维度来修改簇的功能。Henry的NodeTrix[22]支持用户选择和分组节点来生成一个邻接矩阵，高亮相关的模式。ClusterSculptor[6]和Cluster Sculptor [7]提供了有效的操作框架可以使用户合并、拆分和重新组织生成的簇。Schreck[8]则提出了借助用户的反馈信息来对数据更好的聚类。Sacha[19]等人开发的有关犯罪分析的工具能使用户交互的改变数据维度的权值。

ClustVis[26]和ClusterEasy相似，是一个支持PCA和聚类热图可视化的web项目，但是ClustVis不具备交互分析数据功能。以上研究人员在聚类可视化工具上的大量工作使我在可视化方式、功能和界面设计方面获得一些灵感。CE在可视化方式上采用了散点图和热图的组合，为了将高维数据投影到二维空间，CE引入了PCA和MDS等降维算法，聚类算法则使用了比较受欢迎的Kmeans，并对降维或者聚类后的数据用丰富颜色和图形编码。此外，CE注重于满足用户交互式分析数据的需求，在追踪数据变化和数据统计方面提供了支持。

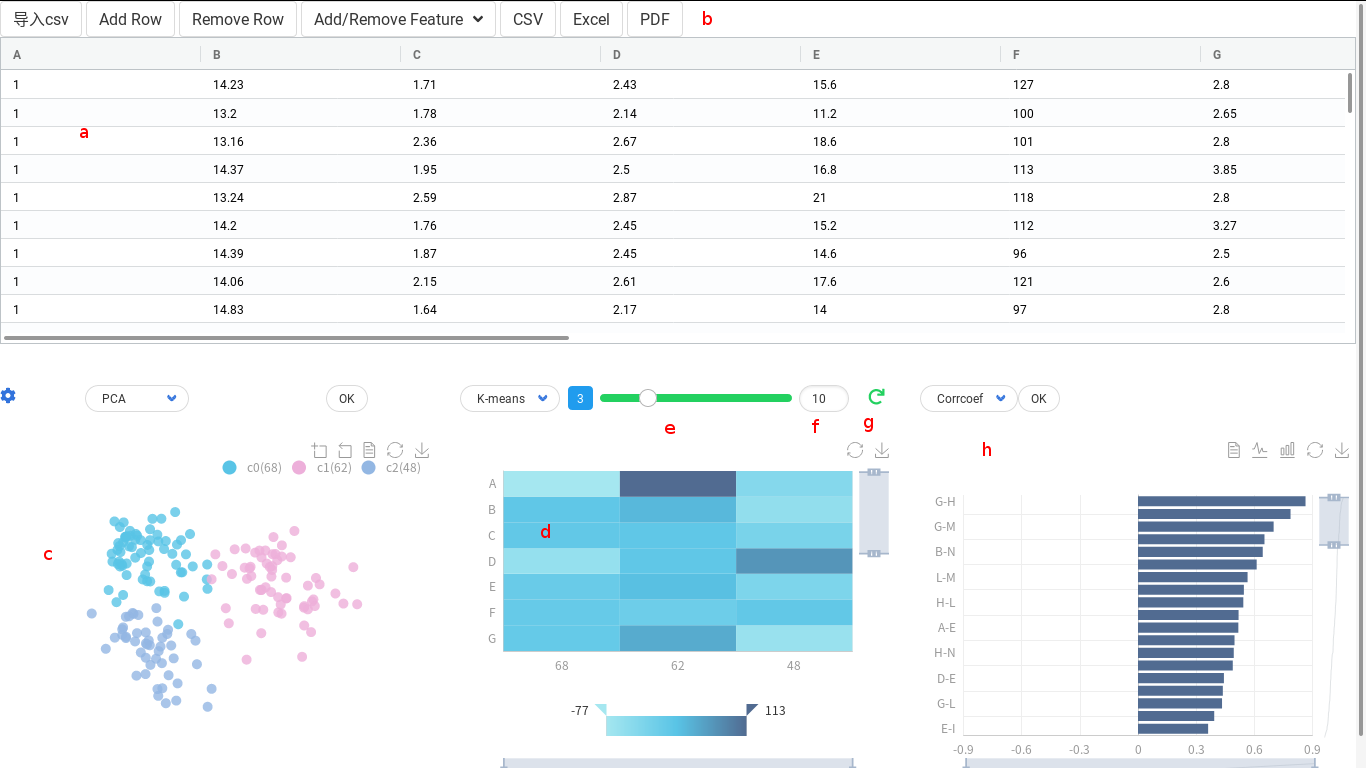
## 1.3 论文研究内容及结构安排

本文主要完成一个聚类可视化分析工具的设计。全文共分为两个章节，每章具体内容及结构安排如下。

第一章，绪论。本章主要介绍研究聚类可视化工具的研究背景和目的，分析国内外相关研究现状，并介绍本篇论文具体研究内容以及文章结构安排。

第二章，ClusterEasy的设计。本章主要从设计理念、视图和功能三方面剖析ClusterEasy，描述相关设计细节。

# 2 ClusterEasy的设计



ClusterEasy主要由表格、投影、聚类和数据统计视图组成 a)表格，显示原生数据 b)数据导出按钮 c)散点图 d)热图 e) 聚类数量滚动条 f)设置最大聚类数量 g)重新聚类按钮 h)数据统计视图

图 2.1 ClusterEasy主要界面

## 2.1 设计理念

ClusterEasy致力于为数据分析人员在聚类可视化方面提供可交互式、便于使用、界面友好、高效的体验，减少数据分析人员的工作量，提高工作效率。

### 2.1.1 支持快速调参聚类

聚类分析是一个不断调参的过程，在这个过程中，数据分析人员往往要试探不同的参数，试验自己的猜想，找到最优的聚类方案，就聚类来说，数据分析人员可能想要知道把数据集划分成多少簇是合适的，还可能想要了解数据集中不同特征对聚类的影响，本工具在设计过程中考虑到了用户快速调参、降低操作成本的需求，为用户设计了简单易操作的UI，使用户对参数的修改转化为对图形元素的操作，让用户像点击鼠标那样简单而轻松的完成聚类过程。

### 2.1.2 直观的视觉编码

作为一款可视化工具，能够简单有效的用颜色编码信息是最基本的要求，优秀的色彩搭配方案能提供给用户简单明了的视觉感受，让用户可以直观地观察到并理解要表达的信息，在本工具中，散点图和聚类热图作为聚类信息的输出窗口，需要用不同的颜色来标记不同的簇，所以颜色搭配上要协调不突兀、且颜色变化能明显区分不同的簇，保证向用户传递的信息一目了然。

### 2.1.3 多种数据视图

在聚类分析的过程中，数据分析人员往往需要更加深入的了解数据内部关系，发现数据内部隐藏的结构，所以如果能帮用户挖掘出更多的信息，对节约用户时间和提高聚类的速度是有帮助的。ClusterEasy提供了表格和散点图，分别用来显示原生数据、降维和聚类之后的数据，前者可以让用户了解到相关数据样本的精确数值，后者则能直观的让用户分辨出数据被划分之后产生的簇，及每个簇中数据点的分布情况。另外，ClusterEasy还提供了热图，通过颜色的种类和明亮程度来编码不同特征值在不同簇中的比重，可以帮助用户分析不同特征值对聚类的影响，结合ClusterEasy提供的ANOVA，还可以让用户了解到不同簇中同一特征值下数据的差异程度。

### 2.1.4 支持交互式探索数据

数据的价值来源于其内部隐藏的结构信息，聚类是挖掘这信息的方法之一，但是目前的聚类算法就像一个黑盒子，接受我们提供的原生数据，反馈给我们结果，除了提供给我们聚类信息之外，并没有告诉我们额外的有价值的信息，如果你还希望找寻数据被为什么被这样划分的原因，哪些特征值对数据的聚类产生了重要的影响，那么，能够探索某些数据点在数据集中扮演的角色对找到这些问题的答案比较有帮助。ClusterEasy赋予了用户操纵某些数据点和特征值的能力，并能即时响应用户的操作，实时渲染数据变化结果。

## 2.2 视图

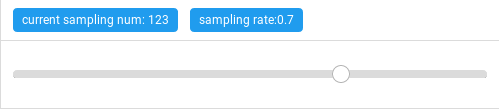
为了使数据分析人员对数据更好的完成聚类工作，本工具提供了以下几个视图接口。

### 2.2.1 表格

表格(图2.1a)是数据分析人员经常使用的工具，ClusetrEasy提供的表格可以帮助数据分析人员浏览数据、编辑数据、添加或删除行和列、过滤和排序数据，并支持导入CSV文件，可以根据需求选择以CSV/PDF/Excel格式导出数据(图2.1b)。另外，ClusterEasy的表格对大数据做了优化，程序内部采用了多线程的方式解析和加载数据，并在加载了部分数据之后就可以进行数据的显示，防止数据的加载阻塞主线程，造成用户等待全部数据加载完成导致冻结界面的情况，提高用户的体验。

### 2.2.2 投影视图

投影视图是用户观察总体数据分布和聚类结果的主要视图，该视图由散点图、降维参数控制面板、抽样设置视图三部分组成。在聚类的过程中，数据分析人员要处理的数据往往是高维的，这些数据无法直接以图形编码的方式显示出来，所以本工具提供了散点图(图2.1c)，结合维度缩减的功能，可以将高维数据通过PCA或者MDS算法降到两维，并以散点图的形式显示出来。对于聚类完成后的数据，会以不同的颜色标识，使用户对聚类信息有一个直观的视觉感受。在降维参数面板中，用户可以选择降维算法，也可以直接选择显示原生数据，这时候要设置好要显示的两个特征值分别作为散点图的x坐标和y坐标，这对观察特征值之间的相关性非常有用。用户可以通过滑动抽样设置视图(图2.2)中的滑动条来设置数据的抽样比率，随着抽样比率的变化，抽样视图中的统计数据和散点图中的抽样数据即时发生变化。



用户可通过滑动滑动条，修改抽样比率

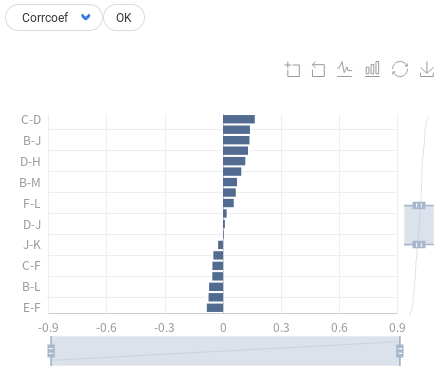
图 2.2 抽样视图

### 2.2.3 聚类视图

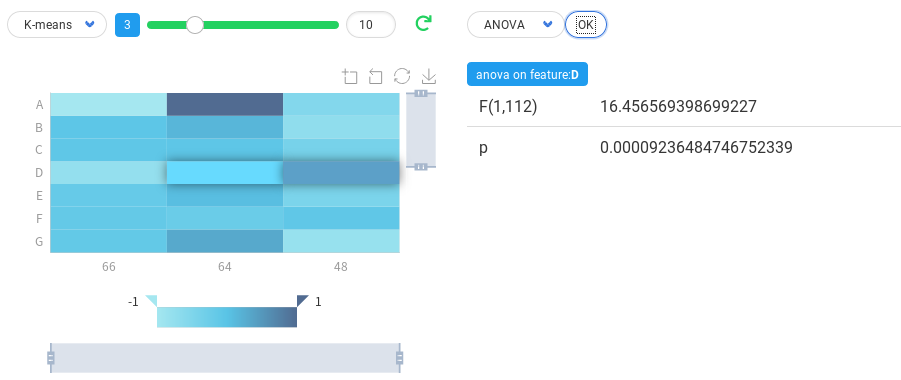
聚类视图(图2.1d)中的热图部分负责显示当前的聚类结果。热图的每一行对应于原始数据中的特征值，每一列对应于数据聚类之后的所划分的不同的簇，同时，列的显示顺序按照簇内数据的数量从左到右排列，列名对应为簇内数据的数量，热图的单元格的颜色编码由每一簇中不同特征值所对应数据的平均值的大小决定。为了让用户方便的调整聚类参数，在热图的上方引入了聚类参数设置面板，用户可以在这部分视图上通过输入框(Figure 2.1e)设置最大聚类数量，拖动滑动条(图2.1f)可以改变Kmeans的聚类数量，如果需要重新聚类，可以通过点击刷新按钮(图2.1g)完成。此外，用户可通过点击选取同一行不同列的两个单元格，来对两个簇中同一特征下的数据进行方差分析，分析的结果在数据统计视图(图 2.3b)显示。

### 2.2.4 数据统计视图

数据统计视图由两部分组成，一部分是用来显示对聚类数据进行单因素方差分析后的结果，该结果包括F值、p值，以表格的形式显示给用户(图2.3a)；第二个部分用来显示原生数据中每对特征值之间的相关系数(图2.3b)，该结果以直方图的形式呈现给用户，其中，直方图的y坐标值是由两个特征名组成的，这两个特征名所对应的柱状图表征相关系数的大小，并对y值按照对应的相关系数大小由上到下排序显示。



(a)相关系数视图，y坐标为每对特征值，x坐标为特征值之间的相关系数，柱状图按相关系数大小排序

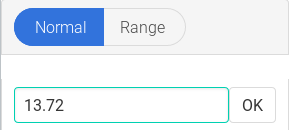


（b）在聚类热图上选取某一特征的两个单元格之后，选择anova算法，可以在统计数据视图获得结果

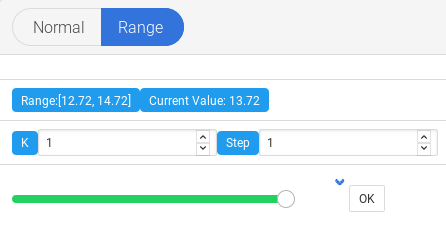
图 2.3 数据统计视图

### 2.2.5 编辑视图

编辑视图既是表格的一部分，也是表格与散点图数据联动的接口，编辑视图在双击表格单元格是启动。在此视图中有两种编辑数据的方式，分别是正常模式和区间模式(图2.4)，正常模式中用户可以赋予数据一个新的精确值，并在散点图上观察到一条连接原始值和修改值的直线来表示数据变化；区间模式中，用户需要设定两个值，k值和c值来设定一个区间，假设当前值为v，这个值所在的那列特征值的标准差为t，那这个区间表示为[v-kt, v+kt]，并以步长ct在区间上取值，绘制到散点图上的一条直线上，来显示该点的变化。



1. 正常编辑模式，用户可以输入精确值



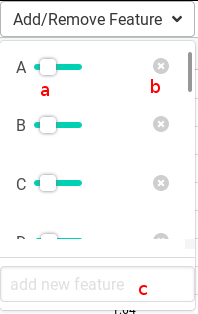
1. 区间编辑模式，用户可以设定区间，滑动条可以改变当前值

图 2.4 编辑视图

## 2.3 交互功能

### 2.3.1 特征值的隐藏与显示、删除与添加

数据中不同的特征值可能会对数据的的降维或者聚类产生不同的影响，这时候，找到对降维和聚类有重要影响的的特征值就比较重要。用户选择隐藏某一列或某几列后对数据进行投影或聚类，可以在散点图上清晰的看到数据变化，如果隐藏的特征值对数据的投影或聚类没有产生明显的影响，用户可以通过删除该特征值达到减少数据维度的目的。



a)控制特征的隐藏与显示 b)删除特征 c)添加新特征

图 2.5 特征值操作视图

### 2.3.2 行的添加与删除

用户在聚类过程中可以添加新的数据点，也可以将觉得不重要的数据点删除掉。

### 2.3.3 数据的过滤

用户可以在表格中选择多种过滤数据过滤方式，包括相等、包含、不包含、不相等、开始于和结束于，且支持“与”和“或”操作，可以快速的将目标数据呈现出来，供用户编辑。

### 2.3.4 brushing and linking

本工具的设计目标之一就是提供给用户极佳的交互体验，使用户可以迭代地完成数据的聚类操作，其中，brushing and linking是实现交互功能的基础。用户通过观察散点图，可能会对某个点的数据产生兴趣，这时，通过双击该数据点，表格就会以高亮数据点的方式帮你定位并显示出来。如果用户想分析该点数据中的特征值会对数据的投影和聚类产生的影响，可以通过双击该数据点的相应特征值对应的单元格，在弹出的编辑视图中改变数据，并在散点图上通过一条直线来观察到该数据点的移动，通过移动的大小和方向，我们可以得知该特征值对聚类产生的影响。借助brushing and linking，ClusterEasy还实现了表格数据更新和调整聚类参数后对散点图和热图的刷新。

## 2.4 降维

降维可以把高维空间的数据投影到低维空间，同时，被降维后的数据仍然保持着数据在高维空间中相似的某种内部联系。在可视化过程中，由于数据大多是高维的，我们不能直接从视觉上观察到聚类之后的结果，只能通过降维这种方法将数据投影到三维或者二维空间。

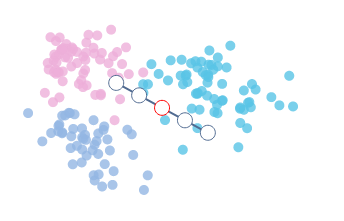
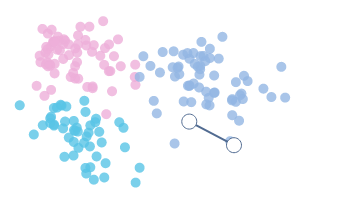
降维操作除了用于可视化之外，还可用于去除冗余特征、加快计算速度、减少存储空间等，所以对于降维算法的研究始终是一个热门领域，目前，有非常多的效果显著的优秀降维算法供我们选择使用，比如主成分分析（PCA），non – 多维数字缩放（MDS），局部线性嵌入（LLE），IClassical Multidimensional Scaling（CMDS），somap，和t分布随机邻域嵌入（t-SNE）。其中，PCA算法在机器学习中经常被使用，所以，在设计本工具时，我选用了PCA算法用作数据的降维，PCA将高维空间中的数据线性变换到低维空间，低维空间中的坐标为方差最大的几个主成分。为了在二维平面上可视化数据，我们需要通过PCA找到数据中方差最大的两个主成分e0和e1，并将数据线性变换到二维空间显示在散点图上。虽然可以有效的降维数据，但是PCA并不能告诉我们生成的两个主成分的意义，以及和特征变量的关系，为了弥补降维算法的不足，帮助用户挖掘数据之间的联系，ClusterEasy引入了前向投影的方法，使用户编辑数据时，可以通过观察散点图上数据点的实时变化，对数据特征值的重要性有一个大体的推断。

## 2.5 前向投影

前向投影是实现交互功能的技术之一，当用户在编辑视图中尝试修改数据时，该数据点所发生的变化，包括移动方向和距离大小，会清晰的的在散点图上呈现给用户，使用户在与数据的交互中，能够了解到所修改数据的特征与聚类后各个簇的关系，所修改数据的大小与数据点移动距离的关系。

前向投影技术的实现离不开out-of-sample extension[23]，借助于out-of-sample extension，PCA算法可以将已有的计算结果，直接应用到新的数据点，并将新数据点投影到低维空间，而无须再重新计算整个数据集，整个过程类似于深度学习中的模型的训练和使用，在使用模型时，不需要重新计算模型。所以，前向投影使得数据的变化可以实时渲染到散点图，并和其他未改变的数据形成鲜明的对比。

以PCA为例，假设E = [e0 e1]是我们已经计算好的主成分组成的向量，当前数据点为x，要修改的特征值为xi，在正常编辑模式(图2.4)中，当我们修改xi变成新数据点x'时，x'在二维投影中的数据点可以直接通过x'E得到，动态变化效果如(图2.6a)；在区间编辑模式中，当用户设置好k和c值，ClusterEasy生成区间[ xi – kσi, xi + kσi ]，并以cσi为步长在区间上取值，并生成2k/c个新的数据点，X' = [x'1 x'2 x'3 ... x'2k/c]，X'在二维投影空间的新数据点可以通过(X')tE算，动态变化效果如(图2.6b)。



(a)正常编辑模式的曲线变化

(b)区间编辑模式的曲线变化

图 2.6 变化曲线

## 2.6 变化曲线

聚类分析就像在做实验，通过改变实验条件（参数），产生不同的实验结果（聚类结果），验证自己的假设。数据分析人员经常想要了解修改某个数据点的值会对聚类结果产生的影响，前面我们提到的前向投影可以让我们在利用已有数据的情况下，仅仅通过实验的因变量（改变的数值）轻松获取新的数据点的投影位置，而不必重新计算整个数据集，变化曲线正是对前向投影的可视化方式。

变化曲线是一条由前向投影产生的数据点所连接而成的一条线(图2.6)，数据点来自由用户设定的区间[ xi – kσi, xi + kσi ]，其中xi为要观察数据点x的某一特征值，σi为整个数据集中该特征值的标准差，k和c为用户设定值，变化曲线会以步长cσi在用户设定的区间里取数据点，并通过前向投影算法获得这些数据点在投影空间的位置，在散点图上连接这些新的数据点绘制成变化曲线。

借助于前向投影，变化曲线可以即时渲染到散点图上，清晰明了的展现数据点的变化方向和变化距离，让用户在不改变原有数据集，不计算新模型的前提下，能够提前看到数据点的变化情况，是帮助用户解读数据的重要辅助手段。

## 2.7 视觉编码

可视化是一个将信息用颜色编码，让用户通过观察多种颜色搭配从视觉上来了解信息的过程。在颜色搭配的过程中，既要注意各种颜色的差异性，使表达的信息一目了然，又要注意各种颜色的的协调性，使用户不能感到突兀，且使信息表达的更加完善和优雅，所以，颜色的编码是用户获取信息的关键，本工具在设计过程中并没有简单的使用随机数生成的方式来生成不同的颜色标记不同的簇，而是使用了HSV和黄金比例来生成协调且对比鲜明的颜色搭配方案。

和常见的RGB相似，HSV也是由三部分组成，其中，H代表色相，即颜色种类，S代表饱和度，指颜色的纯度，V代表明度，即亮度。HSV弥补了RGB有时候会生成太暗或者太亮的颜色这一缺陷，我们只需要固定HSV中的纯度和明度两个参数，只修改色相值就可以获得可读性更高的颜色搭配方案。接下来，我们的目标是让颜色互相之间离得不那么近，以便人眼可以更直观的观察到颜色差异，使颜色分配的比较均匀的一种方法是使用黄金分割比例。

## 2.8 数据采样

当数据量过大时，会给我们的聚类分析工作带来两个困难，一个是过大的数据量会降低聚类速度，在交互式分析数据时增加了用户的等待时间；另一个是在可视化方面，过多的数据点要么会覆盖散点图的大部分区域，要么覆盖区域合适但是点的重复性比较高，不管出现哪一种情况，对用户清晰的识别簇的类别都造成了影响，而且聚类之后，数据内部的结构信息已经显现，每个簇中数据量的多少对我们分析聚类数据影响不大，如果对聚类之后的数据进行分层采样，可以在不影响数据分析的前提下，帮助我们解决大数据带来的困扰。

ClusterEasy在解决第二个困难中采用了对聚类数据分层随机抽样的方式，用户可以随意调节抽样比率，ClusterEasy会在聚类之后的每一簇中按照用户设置的抽样比率进行数据的抽取，减少的数据点并不会导致簇的种类和簇在散点图上的分布等信息的丢失。目前来说，第一个问题仍然未有好的解决方法，对整个数据集的抽样可能会出现偏差，导致某些数据特征的丢失，尽管Kim的工作可以在整个数据集上快速抽样并保证重要的可视化属性[24]，但是其工作并不适用于聚类信息的在散点图上的可视化，这个问题的解决将能在保证关键性信息的同时，减少数据量，加速数据的聚类。

## 2.9 系统信息

在设计工具时，考虑到平台兼容性和数据计算时的性能问题，ClusterEasy采用了客户端和服务器这一模型，并将该工具做成一个web项目，前端作为接口只负责和用户的交互，后端则负责要求比较高的数据计算，这样对用户来说，不用考虑硬件条件。在前端的构建过程中，用到了ag-grid.js[1]、vue.js[2]、echarts.js[3]等js库，分别用来实现表格、交互、散点图和热图功能。后端部分使用了spring mvc技术，tomcat作服务器，由于python在数据处理方面有着丰富的第三方库，所以借助于pytorch、numpy、scipy，实现了聚类、投影和数据统计三大模块 ，并通过java调用python脚本来处理数据，ClusterEasy总体架构如图2.7。

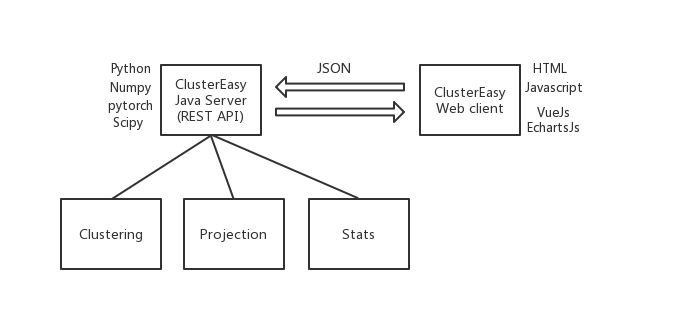


图 2.7 ClusterEasy架构图

# 结 论

本文的主要贡献是提出了一个新的聚类可视化分析工具ClusterEasy，在绪论中我们简单的了解了聚类的概念，认识到数据分析人员在聚类分析时需要一个易操作、可交互式的可视化辅助工具，之后我们看到和聚类可视化相关的研究在可视化方式、聚类结果比对和交互功能方面作出了探索，ClusterEasy以这些优秀的设计为基础，在交互式分析数据上做了进一步的探索，引入了前向算法和变化曲线，可以有效的帮助用户挖掘数据信息，ClusterEasy还实施了分层数据抽样来解决大数据带来的困扰，这些工作是对我对聚类可视化一次新的理解，试图从一个新的角度满足数据分析人员的需求，但ClusterEasy作为聚类可视化工具的一次尝试，仍然有许多不足，其所能解决的问题只是冰山一角，在聚类可视化领域还有许多问题亟待解决，仍然需要更多的学者投入更多的精力。

# 参考文献

[1] ag-grid.js https://www.ag-grid.com

[2] vue.js https://cn.vuejs.org/

[3] echarts.js https://echarts.baidu.com/

[4] J. Seo and B. Shneiderman. Interactively exploring hierarchical clustering results [gene identification]. Computer, 35(7):80–86, jul 2002.

[5] A. Lex, M. Streit, C. Partl, K. Kashofer, and D. Schmalstieg. Comparative analysis of multidimensional, quantitative data. IEEE Trans. Visual. Comput. Graphics, 16(6):1027–1035, nov 2010. doi: 10.1109/tvcg.2010.138

[6] E. J. Nam, Y. Han, K. Mueller, A. Zelenyuk, and D. Imre. ClusterSculptor: A visual analytics tool for high-dimensional data. In Proc. IEEE VAST’07, 2007.

[7] P. Bruneau, P. Pinheiro, B. Broeksema, and B. Otjacques. Cluster sculptor, an interactive visual clustering system. Neurocomputing, 150:627–644, 2015.

[8] T. Schreck, J. Bernard, T. von Landesberger, and J. Kohlhammer. Visual cluster analysis of trajectory data with interactive kohonen maps. Information Visualization,8(1):14–29, 2009.

[9] R. Tryon and D. Bailey. Cluster Analysis. McGraw-Hill, 1970.

[10] Arnheim R. The Gestalt theory of expression[J]. Psychological Review, 1949, 56(3): 156-171

[11] A. Inselberg, B. Dimsdale, "Parallel Coordinates: A tool for visualizing multi-dimensional geometry," IEEE Vis, pp. 361-378, 1990.

[12] N. Cao, D. Gotz, J. Sun, and H. Qu. Dicon: Interactive visual analysis of multidimensional clusters. IEEE Transactions on Visualization and Computer Graphics, 17(12):2581–2590, Dec 2011.

[13] S. L’Yi, B. Ko, D. Shin, Y.-J. Cho, J. Lee, B. Kim, and J. Seo. XCluSim: a visual analytics tool for interactively comparing multiple clustering results of bioinformatics data. BMC Bioinformatics, 16(11):1–15, 2015.

[14] A. Pilhofer, A. Gribov, and A. Unwin. Comparing clusterings using bertin’s idea. IEEE Trans. Visual. Comput. Graphics, 18(12):2506–2515, dec 2012.

[15] N. Elmqvist, P. Dragicevic, and J. Fekete. Rolling the dice: multidimensional visual exploration using scatterplot matrix navigation. IEEE Transactions on Visualization and Computer Graphics, 14(6):1141–1148, 2008.

[16] L. Nováková and O. Štepánková. Multidimensional clusters in RadViz. In Proceedings of WSEAS International Conference on Simulation, Modelling and Optimization, pages 470–475, 2006.

[17] T. V. Long and L. Linsen. MultiClusterTree: interactive visual exploration of hierarchical clusters in multidimensional multivariate data.Computer Graphics Forum, 28(3):823–830, 2009.

[18] S. L’Yi, B. Ko, D. Shin, Y.-J. Cho, J. Lee, B. Kim, and J. Seo. XCluSim: a visual analytics tool for interactively comparing multiple clustering results of bioinformatics data.BMC Bioinformatics, 16(11):1–15, 2015.

[19] D. Sacha, M. Kraus, J. Bernard, M. Behrisch, T. Schreck, Y. Asano,and D. A. Keim. Somflow: Guided exploratory cluster analysis with self-organizing maps and analytic provenance. IEEE Transactions on Visualization and Computer Graphics, 24(1):120–130, 2018.

[20] T. Metsalu and J. Vilo. Clustvis: a web tool for visualizing clustering of multivariate data using principal component analysis and heatmap. Nucleic acids research, 43(W1):W566–W570, 2015.

[21] D. Holten and J. Van Wijk. Evaluation of cluster identification performance for different PCP variants. Computer Graphics Forum, 29(3):793–802, 2010.

[22] N. Henry, J.-D. Fekete, and M. J. McGuffin. Nodetrix: a hybrid visualization of social networks. IEEE Transactions on Visualization and Computer Graphics, 13(6):1302–1309, 2007.

[23] L. Van der Maaten, E. Postma, and H. Van den Herik. Dimensionality reduction: A comparative review. Technical Report TiCC TR 2009-005, 2009.

[24] A. Kim, E. Blais, A. Parameswaran, P. Indyk, S. Madden, and R. Rubinfeld. Rapid sampling for visualizations with ordering guarantees. Proc. VLDB Endow., 8(5):521–532, jan 2015.