聚类数据挖掘可视化模型方法与技术

谢庆华[1]， 张宁蓉[1]， 宋以胜[1]， 王海波[2]， 岳振军[3]
(1. 解放军理工大学 国防工程学院，江苏 南京 210007；2. 南京军区 联勤部，江苏 南京 210016；
3. 解放军理工大学 通信工程学院，江苏 南京 210007)

摘 要：面向通用数据资源，研究聚类数据可视化方法与技术，旨在探索有效的数据处理方法，满足信息领域对高维数据处理的要求。通过对高维数据进行降维处理和可视化映射实现，建立 K 均值算法的聚类数据挖掘可视化系统模型，实现中间聚簇结果、聚类中心、收敛准则函数值三类要素的可视化。利用加利福利亚大学欧文分校(UCI)数据库中的 Iris 数据集、Wine 数据集、Seeds 数据集对可视化系统模型方法进行测试。结果表明，该模型实现了对数据集的有效聚类，能够将中间聚类、聚类中心、收敛准则函数值进行实时有效的可视化表达，达到了预期效果。

关键词：聚类数据挖掘；可视化；平行坐标法；K 均值算法
中图分类号：TP391 **DOI**：10.7666/j.issn.1009-3443.20140716002

Visualization methods and techniques of clustering data mining

XIE Qinghua[1]， *ZHANG Ningrong*[1]， *SONG Yisheng*[1]， *WANG Haibo*[2]， *YUE Zhengjun*[3]
(1. College of Defense Engineering, PLA Univ. of Sci. & Tech., Nanjing 210007, China;
2. Logistics Department of Nanjing Military Region, Nanjing 210016, China;
3. College of Communications Engineering, PLA Univ. of Sci. & Tech., Nanjing 210007, China)

Abstract: Visualization methods and techniques provide powerful tools for discovering hidden laws, helping decision making, and explaining the empirical phenomena. The objective of the research on clustering data mining model visualization methods and techniques is to explore effective data processing methods, and to meet the needs of efficient data processing in the field of information science. This proposal mainly focused on clustering data mining visualization technology, visualization techniques for high-dimensional data via dimension reduction, and visual mapping technology. It studied K-means algorithm for clustering data and visualization, and developed methods for visualizing intermediate clustered results, cluster centers, and convergence criterion functions. It investigated a number of visualization methods, such as clustering data process-oriented, integrated color ratio method, coordinate change, and dimension constraint, with the goal of achieving adequate visualization of clustering data mining and analysis and establishing a K-means algorithm mining visualization system model. Using the Iris data set, Wine data sets, Seeds data set in UCI database, it also systematically tested and verified our data mining visualization models, and analyzed the effects of visualization models on the clustering results and convergence criterion. The test shows that desired results have been adieved.

Key words: clustering data mining; visualization; parallel coordinate; K-means algorithm

随着国民经济的发展，许多领域出现大量类 型各异的数据集合，同时，数据处理技术一直保持

收稿日期：2014-07-16
基金项目：江苏省自然科学基金资助项目(BK2012511)
作者简介：谢庆华，博士，副教授，主要研究信息可视化技术，13605194898@163.com

高速发展。在需求推动和技术保证的前提下,可视化模型方法成为人们关注的热点。模型方法是人类借助抽象和过滤手段去认识事物、理解事物、描述事物的基本方法[1]。可视化模型的抽象程度高、规律性特征挖掘清楚、展现能力强,因而是数据资源管理与利用的有力工具。作为数据资源开发利用的创新成果,数据挖掘可视化方法与技术作为一个复合概念,由数据挖掘技术与可视化方法结合而生,数据挖掘可帮助工作人员更快地发现数据领域中感兴趣的信息或展现一些新颖的结论[2];可视化技术能够将复杂晦涩的数据直观化、简单化,两者的结合,保留数据挖掘技术高效数据处理能力的同时,可视化技术又可消除数据挖掘的"暗箱操作",从而催生一种既快速高效又与人脑认知能力相匹配的数据处理技术。

"基于平行坐标法的聚类数据挖掘可视化"的提法在国内外文献中出现较少,与之最为相近的研究为基于平行坐标法的数据挖掘可视化,部分数据挖掘可视化研究也涉及平行坐标法的应用。综合所查阅文献,以平行坐标法与数据挖掘过程、挖掘算法的结合深度为视角,平行坐标法在聚类数据挖掘可视化中的应用可分为2个层次。一是直观的聚类分析。这一层次平行坐标法的应用,原理简单,应用范围有限,主要集中在数据的可视化、数据挖掘结果的可视化,不涉及与具体算法的结合。具有代表性的平行坐标可视化工具有Parvis平行坐标工具和Xmdvtool可视化工具[3]。二是面向过程的数据挖掘可视化。平行坐标法在这一层次的应用已超越了简单的数据分析,转而面向过程,重点在于平行坐标法与具体数据挖掘算法的融合,旨在实现数据挖掘过程的可视化。目前,这一层次的研究取得了一定成果,诸如King Vis[4],VDM[5]等可视化数据挖掘工具及系统。从King Vis可视化数据挖掘系统、VDM可视化数据挖掘工具的实现效果来看,界面仅将数据挖掘过程中产生的中间聚簇以可视化图形的方式反馈给用户,而忽略了算法中其他关键因子的实时反馈。一方面,会导致关键信息的遗漏;另一方面,可视化效果不明显,形式过于单一。另外,面对不同的数据集、不同的使用者,挖掘算法与可视化技术的选择多种多样,导致挖掘算法与可视化技术的结合呈现多样化。

1 关键技术

本文数据的可视化主要借助平行坐标法来实现[6]。图形的可视化操作与分析则需借助软件中集成的相关可视化技术来完成,比如坐标轴交换技术、颜色比例法、维度约束技术等。

1.1 平行坐标法

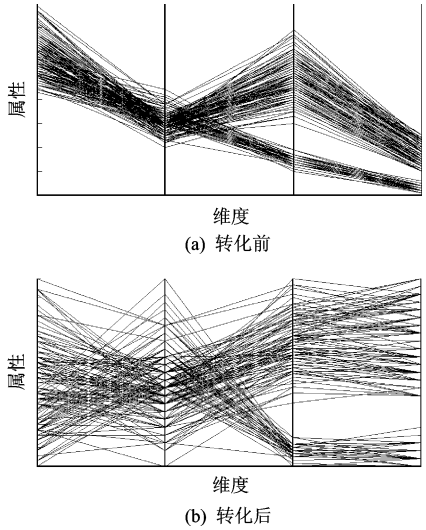
Matlab中,平行坐标系的建立方案如下:在激活的Axes控件中,以数据集的维度作为绘图依据,借助plot命令在坐标系中绘制垂直于y坐标轴的直线。这样,所作直线就与原有坐标系的x,y坐标轴一起,组成了平行坐标系。

平行坐标系绘制完成后,即可进行折线的绘制工作。数据集在利用平行坐标系进行可视化之前,通常要进行一定的转换,具体转换通过式(1)完成:

$$B_{ij}=D\frac{A_{ij}-\min_j}{\max_j-\min_j}。\tag{1}$$

式中:A_{ij}为转化之前的数据;B_{ij}为转化之后的数据;$\max_j$和$\min_j$分别为属性j的极大值与极小值;D为平行坐标系中属性轴的长度,由工作人员自行设定。

如图1所示,图1(b)为Iris数据集经转化后在平行坐标系中的显示效果,对比图1(a)可以看出,折线不再受原始属性值大小的影响,在平行坐标系中均匀分布,折线之间的间距合理,易于观察分析。



(a) 转化前

(b) 转化后

图1 Iris数据集的平行坐标可视化图形

Fig. 1 Parallel coordinate in Iris data visualization

1.2 颜色比例法

颜色比例法是平行坐标法的改进,为解决折线交叉问题而存在,所以平行坐标系的绘制是基础[7]。在绘制完成的平行坐标系中,另需完成两方面的工作。

(1) 选取搭配合理、区分鲜明的4种颜色。利用line命令对每条坐标轴的各个1/4等分进行着

色。本程序中，选取 R、G、Y、B 4 种颜色，这样整个平行坐标系就分为 4 个不同颜色区域，自下而上分别为红、绿、黄、蓝。

(2) 数据转化。与平行坐标法中数据转化的目的和方法相同，保证折线均匀分布于坐标系中。之后即可对折线进行分区间、分批次的处理。颜色比例法方案中，着色的定量性依据为折线与相邻 2 个属性轴的交点位置分布。在程序实现中，将坐标轴区间第一坐标轴(图 1(a)最左侧坐标轴)上的交点位置分布分为 5 种情况，如表 1 所示。属性轴区间均由左右 2 个属性轴构成，$A(i,j-1)$ 即代表折线与第一属性轴(左侧)的交点；$A(i,j)$ 代表折线与第二属性轴(右侧)的交点；本程序中 $D=10$。

表 1　第一属性轴上折线交点位置分布

Tab. 1　The first attribute axis positional distribution intersection polylines

序号	折线交点位置	说明
1	$A(i,j-1)=0$	交点位于属性轴最底端
2	$A(i,j-1)>0\&\&A(i,j-1)<=D/4$	交点位于 R 区域
3	$A(i,j-1)>D/4\&\&A(i,j-1)<=2*D/4$	交点位于 G 区域
4	$A(i,j-1)>2*D/4\&\&A(i,j-1)<=3*D/4$	交点位于 Y 区域
5	$A(i,j-1)>3*D/4\&\&A(i,j-1)<=D$	交点位于 B 区域

第二属性轴上的交点位置分为 4 种情况，如表 2 所示。

表 2　第二属性轴上折线交点位置分布

Tab. 2　The second attribute axis positional distribution intersection polylines

序号	折线交点位置	说明
1	$A(i,j)>0\&\&A(i,j)<=D/4$	交点位于 R 区域
2	$A(i,j)>D/4\&\&A(i,j)<=2*D/4$	交点位于 G 区域
3	$A(i,j)>2*D/4\&\&A(i,j)<=3*D/4$	交点位于 Y 区域
4	$A(i,j)>3*D/4\&\&A(i,j)<=D$	交点位于 B 区域

程序在对改进后的颜色比例法进行实现时，每个属性轴区间中折线两端与属性轴的交点位置包含 20 种子情况。

1.3　坐标轴交换

坐标轴交换的目的是方便工作人员按自己的意愿排列数据属性顺序，从而达到对数据集进行多角度观察的目的[8]。不同的数据集属性不同，为保证软件的通用性，在坐标轴交换交互对话框的实现上采用动态图形用户界面(gaphical user interface，GUI)设计，程序根据数据集的属性特征随时调整交互对话框设置，包括对话框的位置、大小、控件的个数等。

1.4　视图缩放

Matlab 本身具备强大的图形处理功能，对于图形处理中常见的操作，Matlab 都有集成，只需根据需求调用具体指令即可。在视图缩放的实现中，首先利用 getframe 指令获取坐标系中的图形信息，并在新打开的图形窗口中，将 cdata 数据进行显示，执行 zoom on 指令，即可实现视图的缩放。

1.5　维度约束

维度约束用以凸显工作人员感兴趣的维度，为保证软件的普遍性和通用性，同坐标轴交换一样，采用动态 GUI 设计[9]。维度约束实现中，使用 uicontrol 指令，为对话框添加 Static Text、PushButton、Check Box 等控件。

其中的关键在于 Check Box 控件 Value 这一属性的利用。Value 分为 0 和 1 两种状态，Value=1，Check Box 控件处于选中状态，否则为非选中状态。用户操作结束后，检查 Value=1 状态的 Check Box 控件的所在位置，即可得知用户的操作意图，进而将用户感兴趣的属性取出，实现对数据维度的约束。

2　K 均值算法的聚类数据挖掘可视化方法

K 均值算法通常用作其他算法执行之前的数据预处理，所以研究、改进 K 均值算法，不但能改善划分方法本身的性能，还能对结合的聚类方法提供良好的接口[10]。

2.1　流程

K 均值算法数据挖掘可视化的具体流程如图 2 所示。流程中对各关键因子的处理如下：

(1) 算法初始值设置。借助平行坐标法，将待处理数据集转化成可视化图形，通过对视图的分析，工作人员可对数据集的内部结构有一个初步的理解。之后借助交互对话框，对聚类个数 K 及初始聚类中心进行设置，从而在一定程度上减小初始值设置的随意性。

(2) 借助平行坐标法，将 K 均值算法每轮运算所得到的中间聚簇以可视化的形式进行实时显示，

同时弹出询问窗口,工作人员对当前聚类结果进行分析后作出选择。

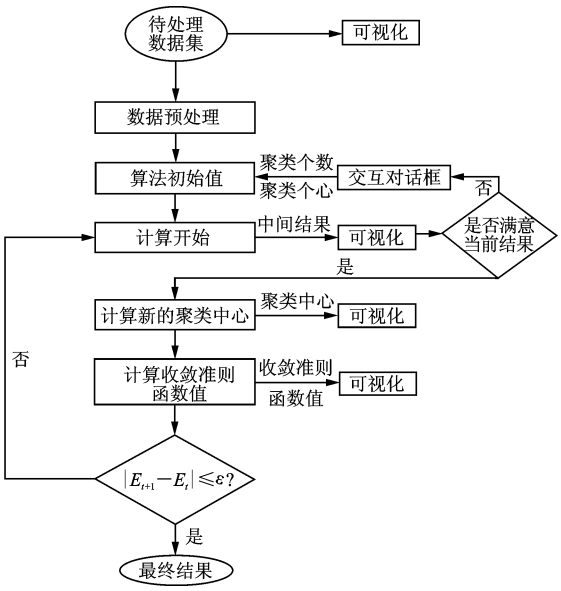


图2 K 均值算法的数据挖掘可视化流程图

Fig. 2 Data mining visualization Flowchart of K-mean algorithm

(3) 将 K 均值算法每轮运算得到的聚类中心进行实时可视化显示。聚类中心的可视化,采用图形叠加方案,即后续绘制的图形不覆盖已有视图。算法的运行过程中,聚类中心的可视化动态地反映聚类中心的变化。此外,聚类中心作为各聚簇的均值,通过观察聚类中心的变化,同样可对运算过程中各聚簇的分布动态有一个直观了解。

(4) 对于收敛准则函数值,其可视化放在算法运行结束之后。K 均值算法运行过程中,将每轮运算得到的收敛准则函数值保存在预定义的数组中。算法运算完毕,将数组中的数据取出,并将其转化成一幅可视化图形。通过该图形,工作人员可快速获知此次运算中算法的收敛速度、迭代次数、算法的收敛节点等信息。

2.2 软件模型

图3为聚类数据挖掘可视化过程模型软件设计的总体方案,分为4个模块。

(1) 工作人员。在聚类数据挖掘可视化中,工作人员往往是数据挖掘能否成功的关键因素之一,为突出工作人员的主动地位,将其单独视作一个主体。

(2) 交互界面。包括维度约束、算法初始值设置、视图缩放、视图保存、交换坐标轴、孤立点计算6个子要素。这是软件模型为工作人员提供的交互,保证工作人员主动性的发挥。

(3) 数据处理模块,该模块是软件的核心。数据处理模块主要包括聚类算法、可视化技术及可视化方法(平行坐标法)。数据处理模块的可视化技术包含颜色比例法、坐标轴交换技术、维度约束技术、数据抽象技术等,保证数据的充分可视化及为工作人员提供足够的视图分析工具。

(4) 数据库。用于源数据、中间处理结果、最终结果等数据的存储。

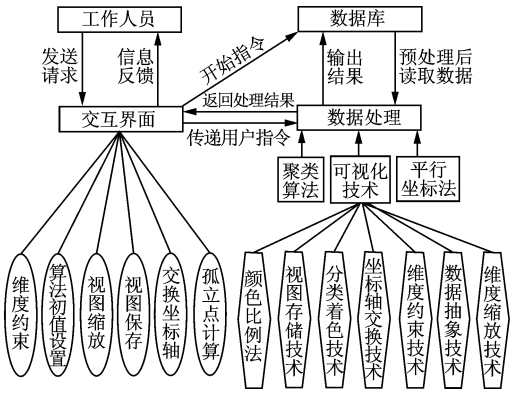


图3 聚类数据挖掘可视化过程模型软件设计的总体方案

Fig. 3 The overall scheme of clustering data mining visualization model

软件模型的基本工作流程为:工作人员通过交互界面发出开始指令(运用文件打开形式),数据库中的相应数据经过一定的预处理后,存入数据处理模块中预定义的储存空间(数组、矩阵),之后即等待工作人员的指令,工作人员通过交互界面发出什么样的指令,可视化技术模块即进行相应操作,并将处理结果通过交互界面反馈给工作人员。

3 具体实现

主要从数据预处理、算法的初始值设置、算法数据挖掘过程可视化的实现3个方面,对 K 均值算法数据挖掘可视化的实现进行描述。

3.1 数据的预处理

数据的预处理主要完成缺失数据的补充、冗余数据的清除、数据格式转换、数据规范化等。本软件模型中,主要考虑了孤立点[11]和数据的规范化[12]2个因素。

(1) 孤立点。检测孤立点的目的在于发现孤立点并进行有效处理,保证大部分数据点的正常运算。孤立点的发现采用基于距离和的思想,首先计算数据集中每个数据点与其他所有数据点的距离,并将

所有距离求和。

$$d=\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{i,j}-a_{(i+1),j}|。$$

式中：d 为数据点的距离和；$a_{i,j}$ 为折线与第一属性轴（左侧）的数据点；$a_{(i+1),j}$ 为折线与第二属性轴（右侧）的数据点。其次，设定一个阈值，将每个数据点的距离和同阈值作比较，若距离和大于阈值，则认为此数据点为孤立点，否则为正常数据点。

（2）数据的规范化。数据的规范化主要是完成数据的中心化和标准化。中心化保证数据集各属性的值具有相同的观察基点，变换之后各属性的均值为 0。标准化保证各属性的变化范围相同。对数据进行中心化处理，通常是在实际数值的基础上减去相应属性的均值。为避免孤立点的影响，各属性的均值用中位数代替[13]。

3.2　K 均值算法的初始值设置

K 均值算法的正常运行，需给定聚类个数 K、初始聚类中心。数据挖掘过程中，为保证用户快速理解中间挖掘结果，通常将各聚簇以不同颜色显示。在颜色方案的制定上，通常由用户来确定聚簇的颜色方案。这样，K 均值算法的初始值设置就包含 3 个部分：

（1）聚类个数 K。聚类个数 K 的确定通过一个 Edit Text 控件和 Push Button 控件实现，控制按钮提供的响应函数将输入的数据进行提取并传递给 K 均值算法。

（2）聚类中心。软件模型为用户提供 2 种输入方式：输入作为聚类中心的 K 行数据的编号、为用户提供直接的输入窗口。2 种方式的实现均采用动态 GUI 设计，交互对话框依据聚类个数的变化而变化。

（3）聚簇颜色方案。聚簇颜色方案设置与聚类中心的设置相同，采用交互对话框，由用户来确定聚簇颜色的搭配方案。

3.3　算法数据挖掘过程可视化的实现

K 均值算法的正常运行，需给定聚类个数 K 及初始聚类中心。数据挖掘过程中，为保证用户快速理解中间挖掘结果，通常将各聚簇用不同的颜色进行显示，在颜色方案的制定上，将主动权交给用户。这样，算法的初始值设置就包含聚类个数 K、聚类中心及聚簇颜色方案 3 个方面。

表 3 中列出了 K 均值算法运行过程中所涉及的变量。对照表 3，K 均值算法数据挖掘过程可视化的具体实现如下：

表 3　K 均值算法运行过程中所涉及的变量

Tab. 3　K-means algorithm running process variables involved

变量名	作用
A_{gld}	全局变量，存放经孤立点检测处理后的数据
A_{gfh}	全局变量，存放经规范化处理后的数据
C_{rownum}	全局变量，存放工作人员输入的聚类中心的行号
C_{center}	全局变量，存放工作人员直接输入的聚类中心
C_C	存储算法各轮运算计算出的聚类中心
C	存放各聚簇中所包含的数据点的编号
B	存储数据点属于哪一类，如属第二类，那么 B 中对应的数据为 2。
K	存放工作人员输入的聚类个数
E	全局变量，存放算法迭代过程中所计算出的收敛准则函数值
C_{color}	全局变量，存放工作人员输入的聚簇颜色信息
X_A	存放使用平行坐标法进行可视化显示的聚簇数据
X_{list}	存放使用信息列表进行显示的数据
D	存放各数据点到对应聚类中心的距离
d	用于收敛准则函数值的中间计算

（1）中间聚簇的可视化。K 均值算法是一个反复迭代求最优的过程，每轮运算都会产生一组满足初始条件设置的聚簇数据。在中间聚簇数据的可视化中，采用文本加图形的复合方式。K 均值算法每迭代运算一次，图 4 坐标系 1 中的图形、信息列表中的数据随之刷新 1 次，保持数据可视化与算法运行进程的同步。

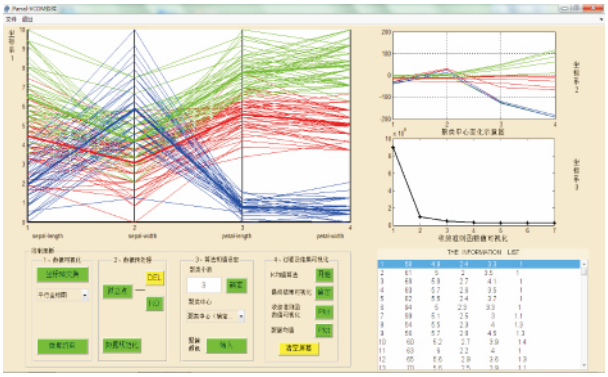


图 4　软件模型的最终实现效果

Fig. 4　Ultimately effects software model

（2）K 均值算法的每次迭代运算，均会产生一组新的聚类中心，并保存在变量 C_C 中。每轮迭代运算开始前，使用 plot 命令，将数组变量 C_C 中存储的聚类中心数据可视化显示在图 4 坐标系 2 中，并采用视图叠加方案，保证新绘制的图形不覆盖已有视图。算法运算结束后，便会得到一组关于算法运行过程中聚类中心的变化趋势图。

（3）为保证算法运行的有效性，避免错误或不恰当的初始值设置所引起的时间浪费，在完成中间聚簇数据的可视化之后，弹出询问窗口，该功能通过 Matlab 预定义的提问对话框（questdlg）实现。

(4) 算法每轮迭代运行完毕,便会计算收敛准则函数的数值,并通过前后2次数值的对比,判断算法是否收敛。数组变量 E 将算法每轮运算所得到的收敛准则函数值进行保存。算法运行完毕后,将数组 E 中的非零数据全部取出,并转化成为图4坐标系3中的图形,即得到收敛准则函数值的可视化图形。

软件最终实现的可视化效果如图4所示,为突出软件各组成部分的功用,将软件对Iris数据集进行处理时的工作状态进行展示。从所显示的效果来看,软件的实现基本达到了预期目标。

4 性能测试

对于软件的性能测试,选用UCI数据库[14]中的标准测试数据集。根据软件集成的算法,选取Iris、Wine及Seeds 3个数据集,利用软件对3个数据集进行相关实验,完成软件模型的性能分析。

4.1 软件模型的聚类效果

软件模型聚类效果的分析,主要是检验软件的设计方案是否有效,软件的各部分组件能否协同工作、完成对数据集的有效聚类。利用软件模型对表4所列出的3个数据集进行聚类,按照各数据集所提供的聚类标准对软件进行相应设置,通过实际聚类结果与理想聚类结果的对比,实现对软件聚类效果的分析。表4中,行号为选取数据集的哪几个样本点作为算法的初始聚类中心,是聚类中心设置的一种方式。符合规范化和标准化要求的以Iris数据集为例,如Setosa。理论上,这一类应为Iris数据的第1～50个样本点,对实际运算得到的Setosa类所包含的样本点进行分析,分析有多少个样本点为Iris数据集的第1～50个样本点,其余同此原理。

表4 软件对各数据集进行处理时的实验条件及实验结果

Tab.4 The experimental conditions and the data set for processing of the results

数据集	初始条件设置			聚类结果			
	聚类个数	聚类中心(行号)	是否经规范化处理	聚类	理想的样本点个数	实际的样本点个数	符合要求的个数
Iris	3	1、2、3	是	Setosa	50	50	50
				Versicolour	50	52	48
				Virginca	50	48	46
Wine	3	11、52、103	是	1	59	57	51
				2	71	60	54
				3	48	61	41
Seeds	3	1、2、3	否	Kama	70	67	57
				Rosa	70	61	60
				Canadian	70	82	70

由表4可以看出:

(1) 对于Iris数据集,在设置的初始条件下,软件能够将Setosa类完全挖掘出来。对于类Versicolour及类Virginca的挖掘,虽然未能到达理想效果,但与理想效果差别很小。

(2) 对于Wine数据集,类2的挖掘效果不好,类1、类3的挖掘效果与理想效果很接近。

(3) 对于Seeds数据集,Canadian类可以完全挖掘出来,Kama、Rosa两类的挖掘效果也与理想效果差别不大。

软件基本上可以对所选取的数据集进行有效地聚类挖掘,这说明,软件的方案设计是有效的,各组件能够协同工作,完成对数据集的有效聚类。

4.2 数据挖掘过程可视化效果

在对数据挖掘过程可视化的效果进行分析时,主要针对Iris数据集进行试验。另外,选取Wine数据集的部分试验结果,说明软件模型对高维数据的可视化效果。

Iris数据集的实验条件:

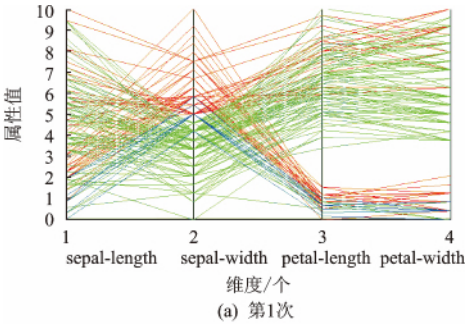
(1) 聚类个数:3;

(2) 初始聚类中心:[5.1,3.5,1.4,0.2],[4.9,3.0,1.4,0.2],[4.7,3.2,1.3,0.2];

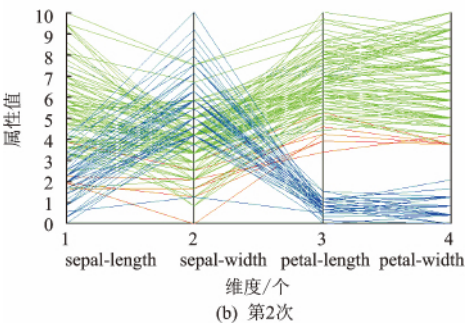
(3) 聚簇的颜色方案:'R','G','B';

(4) 数据进行规范化预处理。

此次实验,K 均值算法总共进行7次运算,分别将第1,2,3,7次运算得到的可视化图形取出,如图5所示。



(a) 第1次



(b) 第2次

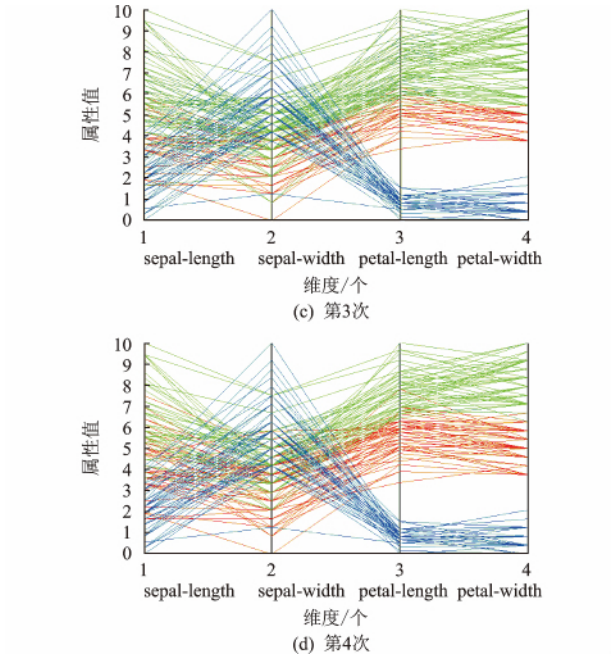


图 5　Iris 数据集的聚类数据挖掘过程可视化效果

Fig. 5　Clustering data mining process visualization of Iris data

由图 5 可以看出，平行坐标法将晦涩的数据转化成二维平面中的可视化图形，并辅以对比鲜明的颜色搭配，可视化效果良好。在聚类挖掘过程中，用户借助界面反馈的可视化图形，可实时了解挖掘的进程，消除了数据挖掘的不透明性。同时，可视化图形也加快了用户对挖掘结果的理解。

Wine 数据集的试验条件：

（1）聚类个数：3；

（2）初始聚类中心：

[14.23,1.71,2.43,15.6,127,2.8,3.06,0.28,2.29,5.64,1.04,3.92,1065]，

[13.2,1.78,2.14,11.2,100,2.65,2.76,0.26,1.28,4.38,1.05,3.4,1050]，

[13.16,2.36,2.67,18.6,101,2.8,3.24,0.3,2.81,5.68,1.03,3.17,1185]；

（3）聚簇颜色方案：'R'，'G'，'B'；

（4）数据经规范化预处理。

选取此次实验最终运算结果的可视化图形进行展示，如图 6(a)所示。由该图可以看出，由于 Wine 数据集维数较多(13 维)，数据转化成平行坐标系中的可视化图形之后，各属性轴之间排列过于紧密，不易观察折线的走向及分布。

为解决这一问题，软件中集成了维度约束技术，可对部分属性进行凸显，从而更好地对聚类结果进行观察。如图 6(b)所示，利用维度约束技术对前 5 个属性进行凸显。可以看出，凸显之后的视图，属性轴之间的间距增大，视图中折线的分布易于观察，可视化效果明显增强。

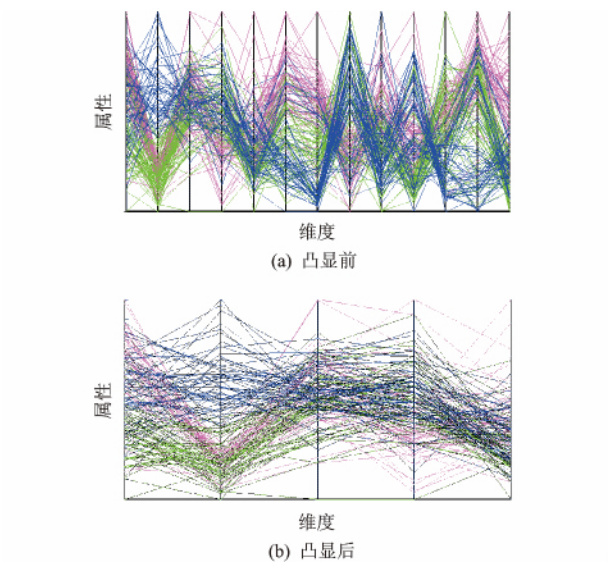


图 6　Wine 数据集的聚类结果可视化

Fig. 6　Wine data set clustering results visualization

4.3　收敛准则函数值的可视化效果

通过收敛准则函数值的可视化图形，可获取以下 3 方面信息：

（1）直观读出算法的运算次数；

（2）了解既定条件下算法的收敛速度；

（3）判断算法趋于收敛的节点，即算法在第几次运算后趋于收敛。

对收敛准则函数值可视化效果的分析，选用 Seeds 数据集作为实验数据，实验条件如下：

（1）聚类个数：3；

（2）初始聚类中心：

[15.26,14.84,0.871,5.763,3.312,2.221,5.22]，

[14.88,14.57,0.8811,5.554,3.333,1.018,4.956]，

[14.29,14.09,0.905,5.291,3.337,2.699,4.825]；

（3）聚簇颜色方案：'R'，'G'，'B'；

（4）数据不经规范化处理。

实验结束后，得到关于收敛准则函数值的可视化图形，如图 7 所示。由图 7 可以看出，此次试验中，在设置的初始条件下，算法总共迭代运算了 7 次，而且算法第 4 次运算得到的收敛准则函数值与第 3 次运算得到的差别不大。由此可以推断，算法在完成第 3 轮迭代后基本趋于收敛。

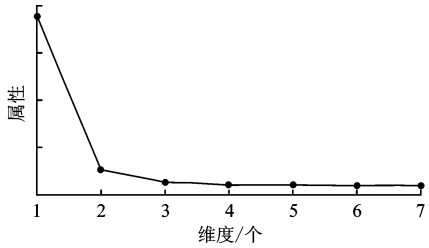


图7 Seeds数据集的收敛准则函数值可视化效果

Fig. 7 Convergence criterion function value Seeds data set visualization

这一结论可通过2幅可视化图形(第3次运算、最终结果)之间的对比进行验证。

图8(a)为算法完成第3轮运算后得到的可视化图形,图8(b)为本次实验最终结果的可视化图形。通过2幅图形的对比可以看出,两者之间仅存在细微差别。通过对聚类数据的分析获知,本次实验中算法的第4～7次运算,完成了非常少量的工作,主要将第3次运算得到的聚簇2(图6中的红色折线)中的第55,101,123,134,140行数据与聚簇3(图6中的蓝色折线)中的第13,30,66行数据取出,放入聚簇1中,即第3次运算的结果与最终结果已非常接近。由此可证明"算法在完成第3次迭代后基本趋于收敛"这一结论是正确的。

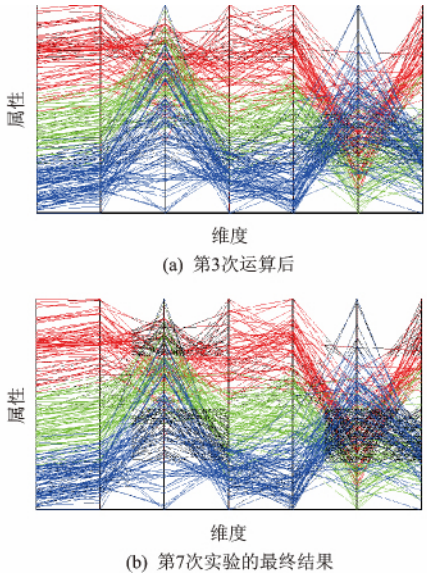


图8 对Seeds数据集进行实验时的部分可视化结果

Fig. 8 Seeds part visualizing the result of the experiment data sets

4.4 聚类中心可视化效果

§4.2对Iris数据集进行的实验,产生了一组聚类中心可视化图形,现将该组图形中的一部分列出,如图9所示。聚类中心的可视化采用图形叠加方案,这样易于对比临近2次运算中,后一组聚类中心相对于前组聚类中心的变化。聚类中心为各聚簇均值的均值,通过聚类中心的变化,可间接了解各聚簇的变化动态。

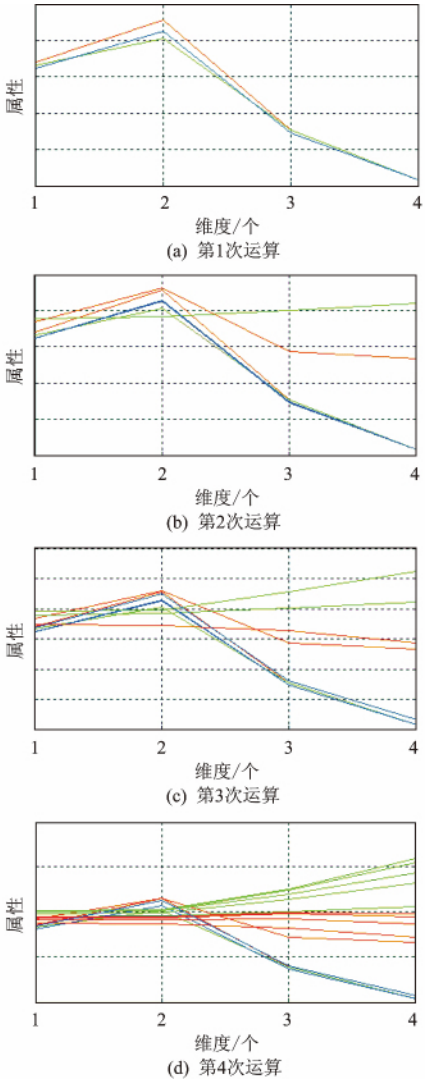


图9 Iris数据集的聚类中心可视化效果

Fig. 9 Cluster centers Iris data set visualization

由图9可以看出,本次实验中,以蓝色折线所对应数据为聚类中心的聚簇,在整个运算过程中,除第3次运算时聚类中心发生较小波动外,其余各次运算基本没有发生变化;而分别以红色、绿色折线所对应数据为聚类中心的2个聚簇,在第2次运算时,聚类中心已较初始聚类中心发生了较大变化,以绿色折线为甚。在之后的运算中,2个聚簇的聚类中心也一直在一定的范围内波动,直至本次实验结束。由此可以看出,以蓝色折线为聚类中心的聚簇,在实验过程中很快达到了稳定(第3次运算后),而分别以绿色、红色折线为聚类中心的聚簇,则一直处于波动状态。

5 结　语

面向通用数据资源，主要研究了聚类数据挖掘可视化理论方法与实现技术，实现了聚类算法与可视化技术的结合，做到聚类数据挖掘过程和可视化过程同时进行；设计了新的可视化隐喻（可视化机构），实现了格式化数据到可视化结构的映射。在此基础上，设计开发了 K 均值算法的聚类数据挖掘可视化软件模型，实现了 K 均值算法的聚类数据挖掘可视化；最后利用 UCI 数据库中的 Iris 数据集、Wine 数据集、Seeds 数据集对可视化软件模型进行性能测试，取得了良好的可视化效果。研究还需进一步解决平行坐标法在表示数据量较大的数据集时出现的折线重叠问题，拓展模型的聚类算法集成，使模型能够处理更多类型的数据集，同时进一步优化可视化效果。

参考文献：

[1] BOUGHRIRA A, FAY D, KHADIR M T. Kohonen map combined to the k-means algorithm for the identification of day types of algerian electricity load[C]. Andrienko G. Computer Information Systems and Industrial Management Applications. Alger: Computer Information Systems and Industrial Management Applications, 2008:78-83.

[2] WU Fangxiang. A genetic weighted k-means algorithm for clustering gene expression data[C] Yu Shidong. Second International Multi-Symposiums on Computer and Computational Sciences. Aomen: Second International Multi-Symposiums on Computer and Computational Sciences, 2007:68-75.

[3] LIU Yanli, LIU Xiyu, MENG Yan. Clustering analysis based on improved k-means algorithm and its application in HRM system[C] Niu Lian-qiang, Zhang Shengnan. First IEEE International Symposium on Information Technologies and Applications in Education. Kunming: First IEEE International Symposium on Information Technologies and Applications in Education, 2007:473-477.

[4] 翟旭君. 基于平行坐标法的可视化数据挖掘技术研究[D]. 北京：清华大学，2004.

[5] 李渊. 基于 K-means 算法的数据挖掘可视化技术的应用研究[D]. 北京：北京交通大学，2007.

[6] IWATA T, SAITO K. Visualization of anomalies using mixture models[J]. Journal of Intelligent manufacturing. 2005, 16(6): 635-643.

[7] 谢娟英，蒋帅，王春霞，等. 一种改进的全局 K 均值聚类算法[J]. 陕西师范大学学报：自然科学版，2010，38(2)：18-22.
XIE Juanyin, JIANG Shuai, WANG Chunxia, et al. An improved global K-means clustering algorithm [J]. Journal of Shanxi Normal University (Natural Science Edition), 2010, 38(2): 18-22. (in Chinese).

[8] 周爱武，于亚飞. K-Means 聚类算法的研究[J]. 计算机技术与发展，2011，21(2)：62-65.
ZHOU Aiwu, YU Yafei. The research about clustering algorithm of K-means [J]. Computer Technology and Development, 2011, 21(2): 62-65. (in Chinese).

[9] 谭桂龙，陈谊. 基于平行坐标的信息可视化方法的应用研究[J]. 北京工商大学学报：自然科学版，2008，26(2)：75-79.
TAN Guilong, CHEN Yi. The study on the visualization methods and techniques of data mining[J]. Journal of Beijing Technology and Business University (Natural Science Edition), 2008, 26(2): 75-79. (in Chinese).

[10] 雷君虎，杨家红，钟坚成，等. 基于 PCA 和平行坐标的高维数据可视化[J]. 计算机工程，2011，37(1)：48-50.
LEI Junhu, YANG Jiahong, ZHONG Jiancheng, et al. High-dimensional data visualization based on principal[J]. Computer Engineering, 2011, 37(1): 48-50. (in Chinese).

[11] SHIBATA T, FUJITA K, ITO K. A real-time learning processor based on k-means algorithm with automatic seeds generation[C]//Schaller A. International Symposium on System-on-Chip. Beijing: International Symposium on System-on-Chip, 2007: 165-169.

[12] HAUSER H, LEDERMANN F, DOLEISCH H. Angular brushing of extended parallel coordinates[C]// Siirtola H. IEEE Symposium on Information Visualization. Provence: IEEE Symposium on Information Visualization, 2012: 127-130.

[13] 罗建. 可视化数据挖掘方法的研究与实现[D]. 西安：电子科技大学，2009.

[14] 路燕梅. 基于平行坐标的可视化多维数据挖掘的研究[J]. 现代计算机，2011，56(20)：16-19.
LU Yanmei. Research on visual data mining based on parallel coordinate[J]. Modern Computer, 2011, 56(20): 16-19. (in Chinese).

（责任编辑：孙　威）