

多维数据集中高维数据可视化算法研究

张红军

(贺州学院 计算机科学与信息工程学院, 广西 贺州 542899)

摘 要: 多维数据集中包含海量的复杂高维数据, 如何对这些数据进行科学的理解、分析和使用, 是数据挖掘领域的一个重点和难点课题. 针对传统可视化算法无法解决高维数据多属性的可视化问题, 算法复杂、适用性较差的不足, 提出基于快速数据聚类的可视化算法研究. 算法构建了一种多维数据模型, 采用三角多项式建立映射关系; 对多维数据集做不同层次的划分, 并基于聚类的效果筛选最佳聚类数; 最后在平行坐标系基础上, 对高维数据进行时序多维分析, 提高算法鲁棒性. 实验证明提出算法简洁易用, 能够较好地实现对多维数据集中高维数据的理解和分析.

关键词: 多维; 数据集; 可视化; 平行坐标系

中图分类号: TP311

文献标识码: A

文章编号: 1000—7180(2017)05—0110—04

Cube High-Dimensional Data Visualization Algorithm Research

ZHANG Hong-jun

(College of Computer Science & Information Engineering Hezhou University, Hezhou 542899, China)

Abstract: Cube contains vast amounts of complex data, how to carry on the scientific understanding and analysis of the data and use, is an important and difficult topic in the field of data mining. For traditional visualization algorithm can't solve the problem of high-dimensional data visualization of multiple attribute, the deficiencies of the algorithm is less complex, applicability, visualization algorithm based on fast data clustering is proposed. Algorithm firstly builds a kind of multidimensional data model, trigonometric polynomial is adopted to establish the mapping relation; The cube for different levels of division, and based on clustering effect to screen the optimal clustering number; Finally based on parallel coordinates, temporal multi-dimensional analysis of high-dimensional data, improve the robustness of the algorithm. Experiments prove that the proposed algorithm is concise and easy to use, can well realize the cube understanding and analysis of high-dimensional data.

Key words: multidimensional; data set; visualization; parallel coordinates

1 引言

近年来,随着计算机科学技术的飞速发展和突破,系统的存储容量和运算能力都获得了极大的提高^[1-2],在社会经济生活统计、证券金融、工程制造、互联网等诸多领域的海量数据量呈现出几何速度的增长. 如何识别、理解、分析和应用这些海量的多维数据资源,成为数字时代的一个巨大挑战^[3-5]. 多维数据集内的高维数据分布极为复杂,而数据内部的特点、数据间的关联性及其变化的趋势是高维数据挖

掘的关键^[6-7]. 在多维数据集的数据类型中,时序数据占有较大的比例,数据集内的多个变量具有时变性、高维度、各不相关等特点,这些又增加了时序数据分析的难度^[8-10].

为了将类似于多媒体数据、人口统计数据、人体健康指标等抽象数据直观化、简洁化,信息可视化的概念在 20 世纪 90 年代首次被提出. 信息可视化^[11-12]主要目的是将抽象的高维数据转换为可以直观审视的、易于理解的数据,帮助数据使用者加强对高维数据的理解和分析,是一种可以化的数据交

互模式.作为一种应对海量高维数据的重要方法和手段,信息可视化在大数据分析的基础之上,通过全方位的数据分析和人机交互,可以识别出隐藏的高维数据时变特征和未知模式.

传统高维数据可视化算法,如平面直角坐标系、灰度直方图等算法在高维数据的可视化分析过程中,算法复杂、耗时长、而且在应对多属性可视化问题时,算法适用性较差^[13-14].本文提出基于快速数据聚类的可视化算法研究.针对高维数据的特点构建了一种多维数据模型,利用三角多项式建立数据映射的对应关系.基于聚类的效果筛选最佳聚类数,并对高维数据进行时序多维分析,提高算法的效率和精度.实验数据表明本文提出的基于快速数据聚类的可视化算法能够有效完成对多维数据集中高维数据的识别、理解和分析,算法简洁、实用性强.

2 多维数据集中高维数据可视化算法研究与实现

2.1 多维数据模型的构建

多维数据集中的高维数据量庞大、结构复杂,而当前的数据聚类手段和挖掘算法在可视化程度和鲁棒性方面亟待提高.本文在数据可视化处理过程中,引入了三角多项式并计算高维数据的鲁棒距离,构建多维数据模型,强化算法的有效性.设 ξ_{kij} 为高维数据立方体,其中 $k = 1, 2, 3, \dots, p$ 为成员数; $t = 1, 2, 3, \dots, q$ 为已知时间点; $i = 1, 2, 3, \dots, m$ 表示集合内数据量; $j = 1, 2, 3, \dots, n$ 为序列值数量.

为简化算法,令 $p = 1, m = 1$ 、在将序列值固定,数据立方体 ξ_{kij} 简化为序列 $\{\xi_i\}$,则矩阵 ξ_{kt} 可描述为

$$\xi_{kt} = \{\xi_{kij}\}_{m \times n} \quad (1)$$

引入变量 λ_n ,表示为

$$\lambda_n = \left\{ \frac{1}{n}, \frac{1}{n^2}, \dots, \frac{1}{n^m} \right\}^j \quad (2)$$

将式(1)左右两端分别乘上一个加权均线向量 λ_n 和中价向量 κ_m ,这时得到多维数据集中的高维数据序列可以表示为:

$$y_{kt} = \kappa_m \xi_{kt} \lambda_n \quad (3)$$

该数据挖掘算法可以最大限度地提高处理高维数据序列的速度.为提高可视化算法的有效性,本文引入三角多相式,并构建了多维数据模型,设在区间 $[-\pi, \pi]$ 内的正交函数系表示为

$$\{\sin t, \cos t, \sin 2t, \cos 2t, \dots\} \quad (4)$$

此时在高维数据立方体中,建立的映射关系可

以描述为

$$H = (h_1, h_2, \dots, h_m) \rightarrow f(t) = h_1 \sin t + h_2 \cos t + h_3 \sin 2t + h_4 \cos 2t \dots \quad (5)$$

$$t \in [-\pi, \pi]$$

设模型中变量 $\nu, \tau \in R^m$, ν 和 τ 间欧式距离 d 为

$$d = \sqrt{(\nu - \tau)^j (\nu - \tau)} \quad (6)$$

函数 $f_\nu(t)$ 和 $f_\tau(t)$ 均为在区间 $[-\pi, \pi]$ 的平方可积函数,两者间的欧式距离可以表示为

$$d_{f_\nu, f_\tau} = \sqrt{\int_{-\pi}^{\pi} |f_\nu(t) - f_\tau(t)|^2 dt} \quad (7)$$

在高维的条件下,提出的模型不会增加局部的差异,能够加速数据的聚类.

2.2 多维数据集层次划分及最佳聚类数的筛选

在多维数据模型构建的基础上,对数据集层次进行划分,并筛选出最佳的聚类数,提高可视化算法的鲁棒性.定义一个阈值 $t_i, i \in [1, d]$,若 x_i 和 y_i 之间欧式距离不大于 t_i ,则认为多维数据集中的数据点 X 和 Y 在第 i 上关于 t_i 相似.给定一个阈值向量 $T = \{t_1, t_2, \dots, t_k\}$,同样也可以理解为数据对象 X 和 Y 在全部维度上均不大于 t_i ,并且数据对象 X 和 Y 保持相似.

在寻找高维数据中的相似点和多维数据层次的划分时,需要保证阈值 t_i 呈现一种增加的趋势.假定多维数据集中的阈值增量为 Δt_i ,定义数据集 ξ_i 在第 i 维的稀疏度 ζ_i :

$$\zeta_i = \sqrt{\frac{\sum_{i=1}^n (x'_{ij} - c_i)^2}{n-1}} \quad (8)$$

式中, c_i 为第 i 维的中心, x'_{ij} 是一种规范化值,第 i 维属性的区间为 $[0, 1]$,则 c_i 和 x_{ij} 在该区间内可表示为

$$\begin{cases} c_i = \frac{1}{n} \sum_{i=1}^n x'_{ij} \\ x'_{ij} = \frac{x_{ij} - \min(x_{ij})}{\max(x_{ij}) - \min(x_{ij})} \end{cases} \quad (9)$$

那么, x_{ij} 是数据点 X 在第 j 维的属性值,由此可以推导出增量 Δt_i 的取值:

$$\Delta t_i = S \times \frac{\max\{\zeta_1, \zeta_2, \dots, \zeta_i\}}{\zeta_j} \quad (10)$$

在多维数据集 C 中,全部数据被分为 k 类,多维数据集的聚类情况可以表示为

$$C^k = \{C_1, C_2, \dots, C_m\} \quad (11)$$

则第 i 类中,第 j 个对象数据之间平均距离 D_{ij} 为

$$D_{ij} = \frac{1}{m-1} \sum (\sum \|x_{jh}^i - y_h^k\|) \quad (12)$$

式中, y_h^k 表示第 k 类数据、 h 维的质心, 则第 i 类中, 第 j 个对象数据类间距离为:

$$I(i, j) = \frac{1}{n-1} \sum (\sum \|x_{jh}^i - x_h^k\|) \quad (13)$$

为揭示多维数据集中聚类结构的类间分离性和紧凑性, 基于数据集的几何结构, 提出反映单类数据与其他数据及数据质心点关系的指标 $\varphi(i, j)$, 综合考虑数据对象的类间距离和类内距离, 基于有效性指标 $\varphi(i, j)$, 可以实现对两者的平衡.

为了使指标不受多维数据集内数据量的影响, 将类间距离和类内距离相加, 以二者的和对数据对象类间和类内差进行压缩, 确定指标的取值范围为 $[-1, 1]$. 当数据集内的类间距离无限小时, 此时 $\varphi(i, j)$ 趋近于 1 时, 为最佳的聚类点; 反之当 $\varphi(i, j)$ 趋近于 -1 时, 这时对高维数据的聚类效果最差.

由于多维数据集中存在孤立的噪声点, 影响数据聚类效果, 在最佳聚类数筛选时先剔除噪声点的干扰.

令 $C^* = \{C_1^*, C_2^*, \dots, C_k^*\}$, 其中 $|C_i^*|$ 为高维数据集中的对象数目, 基于 $|C_i^*|$ 生成一个新的数据序列 $\{C_1, C_2, \dots, C_k\}$, 分解该数据序列 C' 和 C'' :

$$\begin{cases} C' = \{C_1, C_2, \dots, C_m\} \\ C'' = \{C_{m+1}, C_{m+2}, \dots, C_k\} \end{cases} \quad (14)$$

设多维数据集的数据编码长度为 l_c , 利用编码长度进行数据集分层, 并求解最佳的聚类数:

$$l_c = f(v_1) + \sum_{i \leq m} f(v_1) (||C_j| - v_1|) + f(v_2) + \sum_{m+1 \leq k} f(v_2) (||C_j| - v_2|) \quad (15)$$

其中 v_1 和 v_2 可以分别表示为

$$\begin{cases} v_1 = [\sum_{i \leq m} \frac{|C_j|}{m}] \\ v_2 = [\sum_{m+1 \leq k} \frac{|C_j|}{k-m}] \end{cases} \quad (16)$$

一般认定孤立的噪声点位于数据集中数据对象数目较少的类中, 以此可以筛选出多维数据集的最佳聚类数.

2.3 数据集中高维数据可视化算法的实现

最佳数据聚类数的确定, 是实现高维数据可视化处理的关键因素之一, 而高维数据的可视化表现形式可以基于平行坐标系实现. 平行坐标系的主要原理是将多维数据集中记录的 N 个属性, 以 N 条平

行的坐标轴予以表示.

数据集中的属性值域决定坐标轴取值范围, 用连接 N 条属性轴的折线, 代表多维数据的记录, 维度上的取值由属性轴上的交点位置表示, 平行坐标系如图 1 所示.

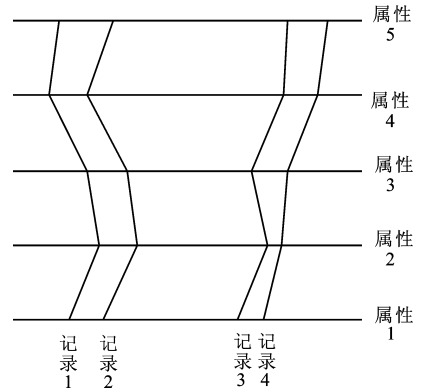


图 1 平行坐标系多维数据属性的表示

以图 1 中的平行坐标系为例, 是包含 4 条多维数据和 5 种属性的平行坐标系, 其中每条属性轴分别代表数据立方体不同的维度; 记录以并行的方式表示, 将属性集并行排列, 可以表示出属性集之间的关系, 能够基于平行坐标系对数据进行宏观的分析.

时间维是一种重要的维度, 包含时序的多维数据由多个时间点离散数据组合而成, 在各个独立的时间上采样, 形成数据记录, 时序多维数据可以表示为

$$T = \{\epsilon_1, \epsilon_2, \epsilon_3, \dots, \epsilon_q\} \quad (17)$$

式中, ϵ_q 表示单位时间点上的数据集采样, q 为采样次数. 任一采样记录对应多维数据集中的当前时间点的数据分析结果为

$$\epsilon_q = \{r_1, r_2, r_3, \dots, r_C\} \quad (18)$$

在数据立方体中增加时间维度后, 坐标系转化成了三维空间, 为多维数据的分析增加了新的视角, 又改善了数据的容量表现.

本文设计了一种基于快速数据聚类的可视化算法, 构建了多维数据模型, 建立数据映射的对应关系; 又基于聚类的效果和多维数据集的层析性, 确定最佳聚类数, 并对高维数据进行时序多维分析, 实现对多维数据集中高维数据的理解、分析和应用.

3 实验部分

为证明提出可视化算法的有效性, 本文进行了实验验证, 选取了标准的数据集, 多维数据集及参数设置, 如表 1 所示.

表 1 多维数据集参数设置

数据集	规模	维数	聚类数
A	135	4	4
B	603	10	10
C	1 850	9	6

采用本文提出的可视化算法分别对 A、B、C 三个数据集的聚类算法进行验证,具体的验证效果如图 2~4 所示。

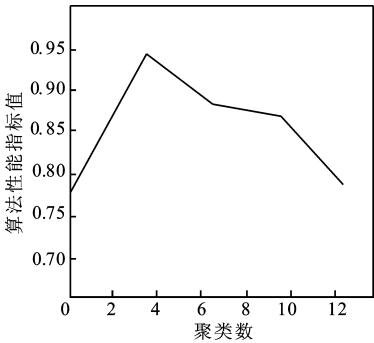


图 2 多维数据集 A 的最优聚类数

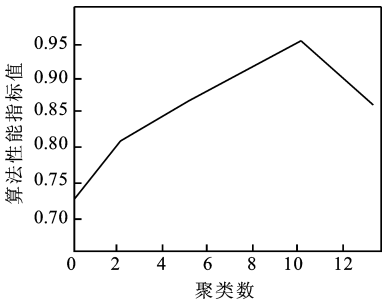


图 3 多维数据集 B 的最优聚类数

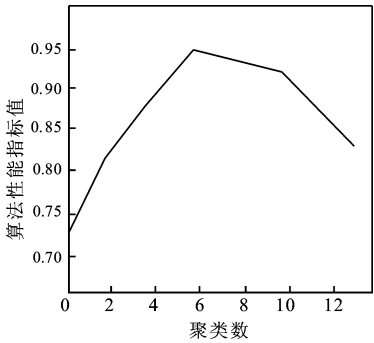


图 4 多维数据集 C 的最优聚类数

图 2~4 中显示的算法性能指标值曲线变化表明,在不同的多维数据集内,本文提出的可视化算法都可以获得正确的最优聚类数。

为直观地体现提出可视化算法的鲁棒性,将传统灰度直方图算法和本文算法在数据集 A 内的点

坐标变化予以对比,如图 5 所示。

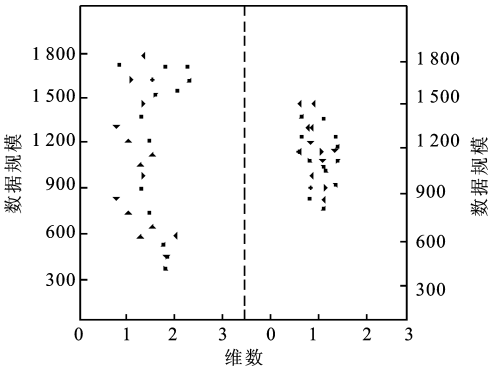


图 5 不同可视化算法下的点坐标变化对比

图 5 虚线左侧为传统直方图,右侧为本文基于快速数据聚类的可视化算法,从坐标点分布可以分析出,本文算法具有更好的收敛性能和更好的数据聚类效果,可以降低高维数据在识别、分析的难度,提高算法的实用性。

4 结束语

本文在当前数据大爆炸的背景下,提出了一种基于数据快速聚类的高维数据可视化算法研究,基于多维数据模型的构建及最优聚类数的筛选,降低高维数据分析的难度、提高海量数据的易用性,实验及对比数据证明了提出算法的有效性。

参考文献:

[1] 徐昇,夏德天,郑久寿. 高升力系统控制计算机容错技术研究[J]. 微电子学与计算机,2015,32(6):36-40.

[2] 南轶,李先国. 模板技术在内容管理系统中的应用[J]. 微电子学与计算机,2012,29(6):180-184.

[3] 田明银,兰一麟涛,钱伟. 基于分形算法的三维地形可视化应用[J]. 电子设计工程,2016,24(14):20-23.

[4] 张晓琳,霍伟,刘立新,等. 多维连续不确定 XML 数据查询处理算法[J]. 微电子学与计算机,2013,30(6):30-33.

[5] 朱欣颖. 云数据中心的多维资源调度策略探析[J]. 智能计算机与应用,2015,31(5):74-76.

[6] 陈佳,李敏. 用于多维数据实视图选择的分布估计算法[J]. 计算机工程,2012,38(11):45-47.

[7] 唐晓东. 基于关联规则映射的生物信息网络多维数据挖掘算法[J]. 计算机应用研究,2015,32(6):1614-1616.

[8] 郑龙,敖永红,梁莹莹,等. RPES:一种新的多维数据可视化方法[J]. 计算机工程与应用,2012,48(21):107-111.

4 结束语

本文设计了一种新的基于 Web 网络的身份识别系统,给出系统架构,介绍了浏览器端录制模块和 Web 服务器端声纹识别模块的设计过程.实验结果表明,所设计系统有很高的识别精度.

参考文献:

[1] 陈玉婷,范永胜. 基于身份识别的学校实验室网络安全设计与实现[J]. 绵阳师范学院学报, 2014, 33(11): 75-79.

[2] 郑浩. 基于 PHP 的网络在线投票系统的设计与实现[J]. 吉林师范大学学报(自然科学版), 2014, 35(1): 78-80.

[3] 冯贺平,杨敬娜,吴梅梅. 基于深度神经网络的身份识别研究[J]. 电脑知识与技术, 2016, 12(22): 161-162.

[4] 孙伟,刘晓敏,王浩宇,等. 基于三重人脸识别身份验证的门禁管理系统设计[J]. 计算机测量与控制, 2016, 24

(2): 225-227.

[5] 曲佳. 面向办公软件的身份认证系统的设计与实现[J]. 承德石油高等专科学校学报, 2015, 17(4): 35-39.

[6] 林超,叶振昀. 基于校园卡的大学生电子照片采集系统的设计与实现[J]. 微型电脑应用, 2015, 31(8): 46-47.

[7] 金丰,张悦,雍鹏. 基于双服务器的 B/S 模式监测系统的设计与实现[J]. 计算机仿真, 2014, 31(2): 201-205.

[8] 洪远泉,周永明,郑春龙,等. 基于距离判别的心电身份识别技术研究[J]. 科学技术与工程, 2014, 14(18): 252-255.

[9] 黄志威. 改进的 AdaBoost 检测算法在身份识别中的应用[J]. 科技通报, 2015, 31(10): 190-192.

[10] 赵卫. 一种基于网络安全的 WIFI 系统身份认证设计[J]. 电子设计工程, 2016, 24(14): 81-83.

作者简介:

陈文庆 男, (1981-), 硕士, 讲师. 研究方向为计算机网络与多媒体教学. E-mail: wwwsheying11@126. com.

(上接第 113 页)

[9] 李晓东,陈俊健,曾凡智. 基于 KDG-tree 的数据库多维索引技术[J]. 计算机应用与软件, 2013, 30(6): 162-165.

[10] 蒋华,张乐乾,王鑫. 基于多维评价模型及改进蚁群优化算法的云计算资源调度策略[J]. 计算机测量与控制, 2015, 23(7): 2559-2562.

[11] 周维,李敬辉,黄高明. 基于 VTK 的海战场电磁信息可视化技术研究[J]. 计算机与数字工程, 2012, 40(1): 121-124.

[12] 李伟波,李振中,何频. 移动位置服务的岩土信息可视化技术研究[J]. 计算机工程与应用, 2012, 48(5):

228-231.

[13] 董重,魏迎梅. 运用平行坐标系的多变元时序数据可视化方法[J]. 小型微型计算机系统, 2015, 36(10): 2408-2411.

[14] 叶学义,宋倩倩,高真,等. 基于直方图条件熵的水声数据分类算法[J]. 计算机工程, 2016, 42(11): 244-248.

作者简介:

张红军 男, (1975-), 硕士, 讲师. 研究方向为计算机算法与分析、软件工程、计算机教育与应用. E-mail: bspandsh@163. com.