●余肖生 周 宁 张芳芳

基于可视化数据挖掘的知识发现模型研究*

摘 要 基于可视化数据挖掘的知识发现模型,过程有 4 个步骤:数据的收集和存储;数据预处理,把数据转化成可以理解的形式;使用硬件和相关软件,产生一个可视化的数据表示;通过与数据的可视化表示进行交互,用户从中感知和挖掘知识。图 5。参考文献 14。

关键词 可视化技术 数据挖掘 知识发现 数据库
分类号 G250.74

ABSTRACT In a knowledge discovery model based on visualized data mining, there are four steps, i.e. data collection and storage, data preprocessing to convert the data into understandable forms, visualized data representation produced by hardware and related software applications, user's interaction with visualized data representation for understanding and knowledge mining. 5 figs. 14 refs.

KEY WORDS Visualization technology. Data mining. Knowledge discovery in databases.
CLASS NUMBER G250.74

数据库中的知识发现(KDD,knowledge discovery in the databases)目标,是从数据库中发现潜在的、有意义的、未知的关系、模式和趋势,并以易被理解的方式表示出来[1]。知识发现过程重要步骤之一的数据挖掘是采用自动方式完成的。对大多数用户而言,理解和解释仅仅由自动算法产生的结果可能有一定的困难。可视化数据挖掘是知识发现过程中的一种新方法,它利用可视化作为人机交流渠道。它将人集成到整个数掘挖掘过程且将人的随机应变能力、感知能力与计算机巨大的存储能力、计算能力结合起来[2-3]。人的感知能力使用户可以在短时间内分析复杂问题,认知重要的模式且得出比任何计算机更有效的结论[4]。

1 可视化数据挖掘

所谓可视化数据挖掘,是为了提高数据挖掘的准确性和用户的主动性,将可视化技术应用于数据挖掘的各个阶段,以便在知识发现过程中得到更符合用户需要的知识的一系列理论、方法和技术。

相对数据挖掘而言,可视化数据挖掘有不少优点[5]。

由于允许用户参与数据挖掘过程,且通过人机接口与数据挖掘过程进行交互,实时监测挖掘的中间结果,从而有效地提高挖掘结果的可信度,改变以往知识发现过程中数据挖掘仅仅是给出一个自动挖掘结果的"黑盒"的角色。

通过对数据和信息的可视化,充分利用人类认知能力,可显著提高数据挖掘结果的有效性和质量。

在可视化数据挖掘过程中采用了人机交互式的可视化用户界面。因此,如果用户是领域专家,他就能在整个过程中充分利用领域知识来约束算法的搜索过程,提高搜索效率。

数据挖掘可视化工具比较多,这里主要介绍有代表性的 3 种。

(1)平行坐标(Parallel Coordinates)。1981 年,Inselberg 首先提出平行坐标法来解决高维数据可视化问题。其后,Inselberg 和其他研究人员将它应用于统计学、计算机图形学、机器人技术等领域并获得成功[6]。这一方法在数据挖掘、系统优化设计等方面都得到了较好应用[7]。它的基本思想是在二维空间中,采用等距离的竖直的 n 个平行坐标轴表示 n 维空间,n 个变量值对应到 n 个平行坐标轴上,再将 n 个坐标轴上的点用连续线段连接起来表示一个空间点。例如,图 1 在二维空间上,用平行坐标法显示了含有四维和 150 个数据项的 Iris 数据集,每一个数据项和它们的聚类从图形显示中都能清楚地看到[8]。

* 本文系国家自然科学基金项目(70473068)和教育部哲学社会科学研究重大课题攻关项目(05JZD00024)的研究成果之一。

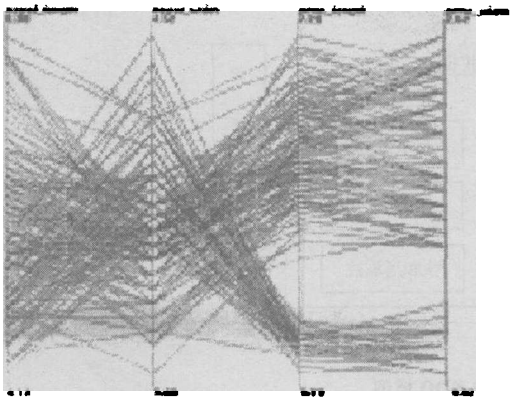

图1　在平行坐标中的 Iris 数据集

(2)d 维大旅行(d-dimensional Grand Tour)[9]。它是 Asimov 提出的二维大旅行的推广。d 维大旅行的基本观点是从所有可能的角度来观察数据云。d 维大旅行算法有两个关键因素:空间填充和连续性。空间填充允许数据分析人员从所有的角度来观察数据云,而连续性则允许人类视觉系统跟随数据云。为了达到这样的目的,Wegman 于 1991 年描述了 Asimov-Buja 算法的应用且对发现连续性的几种不同方法作了进一步讨论,2002 年,Wegman 和 Solka 又提出了空间填充大旅行。空间填充大旅行的关键思想是通过作为时间参数函数的所有旋转矩阵来发现空间填充路径。当一个旋转矩阵决定时,协同系统的标准基向量通过矩阵旋转将数据云映射到旋转的协同系统中,最后,映射的数据在平行坐标中显示出来。图 2 是九维立方体五维大旅行的结果图[10]。

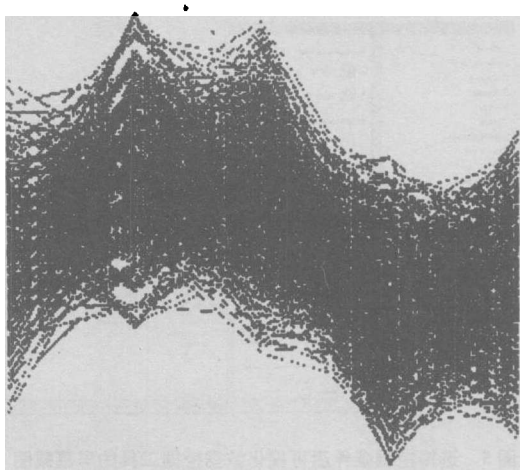

图2　九维立方体五维大旅行的结果图

(3)饱和度刷(Saturation Brushing)[11]。它是 Wegman 和 Luo 于 1997 年提出来旨在作为处理大型数据集的一种技术,是普通刷的推广。所谓普通刷,是为了数据的可视化,对不同的数据片段用不同颜色的刷。正常情况下,普通刷采用一个矩形框来完成这项工作。而当大型数据集中有大量重复数据时,在有大量重复的数据集中,普通刷可能容易使人产生误解,尤其是在有动画的区域如旋转或大旅行等。要区分一个像素点是表示一个观测点还是表示成百上千个观测点是比较困难的,而饱和度刷的关键思想是每个点被赋予一个高度不饱和色(接近黑色),且当点重叠时,它们颜色饱和度通过所谓的频道(a-channel)技术而增加。高度重叠的像素点有完全的饱和色,而少量重叠的像素点则保持接近黑色。

上述 3 种方法往往不单独使用,而是同时使用两种甚至 3 种。比如对大型高维数据集而言,平行坐标和大旅行相结合就非常有效。

2　基于可视化数据挖掘的知识发现模型

笔者提出的基于可视化数据挖掘的知识发现模型,充分利用了目前可视化技术的成果来改进以往的基于数据挖掘的知识发现模型,可以充分发挥用户的主观能动性,让用户积极参与到发现的全过程,有利于使用户得到真正想要的知识。

整个过程大致包括 4 个步骤。

(1)数据的收集和存储。在这个过程中,主要是从不同的源数据库中抽取相关的原始数据,经过适当的加工、整理,去除噪声数据,存放在数据仓库的内部数据库中。通过数据仓库访问工具,给用户提供一个集成的、能对数据进行综合分析、发现知识的环境。

(2)数据预处理,将数据转化成可以理解的形式。在此过程中,主要是对数据仓库中的数据进行抽取、清洗、转换、装载等操作,将数据仓库中的原始数据按不同的需求进行归类,得到相关的数据集。

(3)使用硬件和相关软件,产生一个可视化的数据表示。对不同的数据集采用不同的降维算法,将数据集的维度降到能可视化的程度。然后根据需要,选择适当的可视化方法对数据集进行可视化。

(4)通过与数据的可视化表示进行交互,用户从中感知和挖掘知识。用户对可视化结果进行评估,看是否是用户满意的知识,如果满意则整个过程结束,如果不满意,则返回步骤(2),重复(2)~(4),直到用户满意。

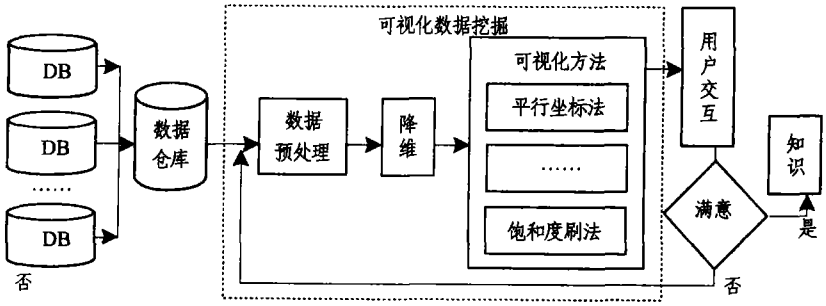

图3 可视化数据挖掘的KDD模型

3 实例研究

为了提高知识发现的准确性,利用可视化技术来改进知识发现过程,让用户充分参与到整个过程,已经成为这一领域专家的普遍共识。关于这方面的研究越来越受到人们重视,产生了一些有代表性的研究成果,INVISIP就是其中之一[12]。

INVISIP(Information Visualisation for Site Planning)是由欧盟委员会的相关部门提供资助,旨在创建这样一个框架,它能服务于场地规划过程所涉及的部门和个人:市政府、规划部门、数据提供者和市民。它使用不同的可视化技术用来改进有用信息的查找和分析效率,促进基于地理数据信息系统的决策。

初始化可视化数据挖掘工具后的情景如图4所示[13]。它显示了通过初始后激活的一些不同元数据变量的可视化情景,如:更新频率的平行图,参考日期的直方图等。从这些可视化图形中,用户能抽取关于现成地理数据集的有用信息,如:平行图暗示所有数据集都是完整的,但现成免费的数据集是不经常更新的。直方图显示一些数据集是在2月份更新的,而其他是在8月份更新的。用户想购买最新的数据集,他应该选择8月份生产的那些数据集。如果用户选择了这个挖掘条件,则得到挖掘结果的平行图,如图5所示[14]。很明显,用户可以知道最常用的数据集是连续更新的且价值100欧元。基于这个结果,用户立刻能决定这些数据集是否适合于他。如果这个结果数据集不符合他的标准,他能用不同的初始条件重新开始分析。

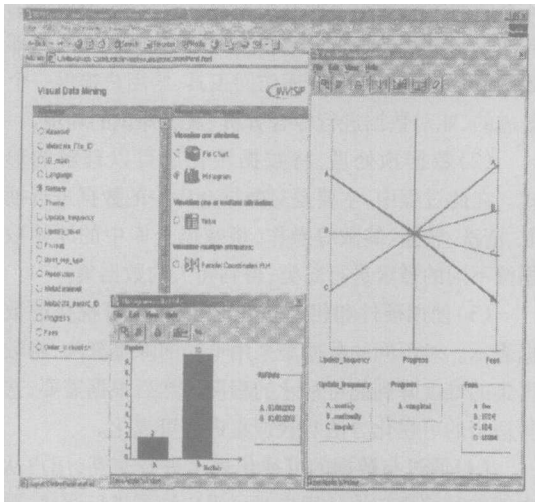

图4 初始化后可视化数据挖掘工具的屏幕截图

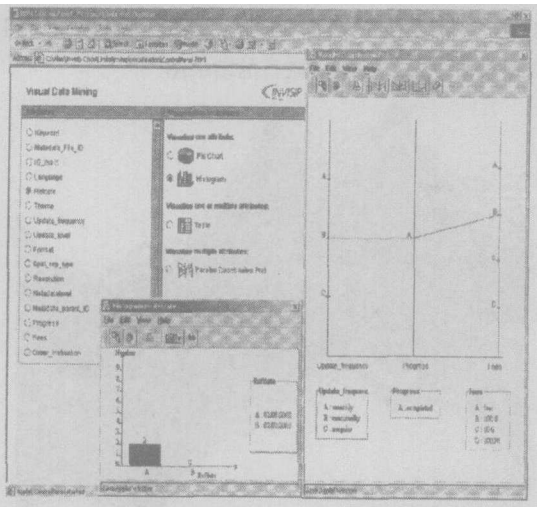

图5 选择挖掘条件后可视化数据挖掘工具的屏幕截图

(转第56页)

查询)
17 http://cite.cis.drexel.edu/#ConceptMap(2005-03-15查询)
19 http://cluster.cis.drexel.edu/~cchen/citespace/(2005-08-15查询)
20 刘玮等.基于文本的信息可视化方法研究.现代图书情报技术,2003(2)
21 罗龙艳.基于可视化技术的信息检索初探.现代图书情报技术,2002(4)
22 李君君.基于映射的信息检索逻辑模型.情报理论与实践,2004,27(4)
23 周宁,文燕平.检索结果的可视化研究.中国图书馆学报,2002(6)
24 周静怡,孙坦.信息可视化在数字图书馆中应用浅析.现代图书情报技术,2005(1)
25 李爱国,汪社教.信息检索可视化.现代图书情报技术,2004(2)
26 李春旺.信息检索可视化技术.现代图书情报技术,2003(6)
27 文燕平.WWW信息检索可视化实现原理研究.现代图书情报技术,2005(4)
28 文燕平.WWW信息检索可视化研究.武汉大学博士论文,2004
29 Lin,X.;White,H.D.;& Buzydlowski,J. Real-time author co-citation mapping for online searching. International Journal of Information Processing & Management,2003,39(5)

张学福 博士,黑龙江大学信息资源管理研究中心教授,黑龙江大学信息管理学院副院长,教授。通信地址:哈尔滨黑龙江大学信息管理学院。邮编150080。

(来稿时间:2005-12-05)

(上接第46页)

基于数据库的知识发现的关键,集中在与数据获取任务相关的问题。数据选择的错误将严重影响整个知识发现过程,导致失败。为了增强用户关于地理数据有效性的意识,INVISIP提出了一种基于可视化数据挖掘的知识发现方法。它允许用户完成确定的和探索性的分析,帮助他在所需要的数据和可用的数据之间找到折中。这样,通过使用可视化作为交流渠道,即使是一名不熟练的用户,也能较好地发现自己所需的知识。

参考文献

1 Badjio, E. F., Poulet, F. Dimension Reduction for Visual Data Mining. http://asmda2005.enst-bretagne.fr/IMG/pdf/proceedings/266.pdf (2005-12-10查询)
2 Daniel A. K. Information Visualization and Visual Data Mining. IEEE Transactions on Visualization and Computer Graphics, 2002(1)
3,9,11 Edward J. W. Visual Data Mining. http://www.galaxy.gmu.edu/stats/syllabi/infi979/VisualDataMining.pdf (2005-12-10查询)
4 Pak Chung Wong. Visual Data Mining. http://www.pnl.gov/infoviz/visual_data_mining.pdf (2005-12-12查询)
5 刘绪崇.基于OLAM的可视化数据挖掘技术研究.[学位论文].长沙:国防科学技术大学,2002
6 刘勘等.基于平行坐标法的可视数据挖掘.计算机工程与应用,2003(5)
7 欧海英等.平行坐标可视化技术在固体火箭发动机优化设计中的应用.固体火箭技术,2004(4)
8 Jing Y., Matthew O. W., etc. Visual Hierarchical Dimension Reduction for Exploration of High Dimensional Datasets. http://davis.wpi.edu/~xmdv/docs/vhdr.pdf(2005-12-14查询)
10 Dianne, C. Calibrate your Eyes to Recognize High-Dimensional Shapes from Their Low-Dimensional Projections. http://www.jstatsoft.org/v02/i06/paper.html(2005-12-21查询)
12,13,14 Albertoni, R., et al. Knowledge Extraction by Visual Data Mining of Metadata in Site Planning. http://www.ima.ge.cnr.it/irna/personal/albertoni/PersonalPage/src/ScanGIS2003.pdf(2005-12-21查询)

余肖生 武汉大学信息管理学院2004级博士研究生。通信地址:湖北武汉。邮编430072。
周 宁 武汉大学信息管理学院教授。通信地址同上。
张芳芳 武汉大学信息管理学院2004级博士研究生。通信地址同上。

(来稿时间:2006-01-10)