文章编号:1672 - 058X(2013)03 - 0058 - 04

可视化数据挖掘技术的研究与实现*

张　俊
(芜湖职业技术学院,安徽 芜湖 241000)

摘　要:为了充分发掘和利用信息资源的价值,数据挖掘技术应运而生;首先就可视化数据挖掘的概念和分类进行了阐述,然后探讨了可视化数据挖掘的一些主要技术,最后通过所开发的一个系统对经典的购物篮分析问题进行了可视化数据挖掘技术的实现探讨。

关键词:数据挖掘;可视化;挖掘技术
中图分类号:TP391　　**文献标志码**:A

传统的数据挖掘过程对用户而言是一个“黑盒子”,用户将数据集交给算法,然后自动地生成结果,挖掘的过程不可见,用户很难参与,挖掘出的结果也常常只有专业的数据挖掘人员能够理解,这些结果难以在实际应用中发挥作用,用户对挖掘的结果也难以信任,怎样改进这种挖掘过程一直是个难题? 考虑到图形和图像、颜色等表达方式的直观性和形象性,因而可以通过数据挖掘与可视化技术的结合,来弥补传统数据挖掘过程的缺陷,加强数据挖掘的处理过程。可视化数据挖掘正是数据挖掘和可视化技术的有机结合。这种结合强调的是以人为中心,一方面强调充分利用人类的知识领域和模式感知能力,另一方面强调用户对挖掘结果的理解和利用。可视化的方法使数据挖掘技术的应用更具形象性和直观性,挖掘的过程加入更多人类的参与和指导,可以有效地提高数据挖掘结果的可信度、可理解性和可用性。

1　可视化数据挖掘概述

可视化数据挖掘技术是可视化技术和数据挖掘技术的有机结合,是运用计算机图形学、图像处理技术等,将数据挖掘的源数据、中间结果和最终挖掘结果转换成直观、易于理解的图形或图像的方式,并进行交互处理的理论、方法和技术。按照可视化在数据挖掘中应用的不同阶段,可以将可视化数据挖掘划分为源数据的可视化、挖掘过程的可视化、结果的可视化。

(1) 源数据的可视化。目前对源数据的可视化方法已经有了很多种,就是在数据投入挖掘算法之前,将整个数据集以可视化的方式呈现给用户,目的是使用户能够快速地找到感兴趣的区域,从而有目的、有针对性地实施下一步的挖掘。

(2) 过程可视化。挖掘过程的可视化实现起来比较复杂,现阶段的可视化方法主要集中于对源数据和结果的可视化方法。挖掘过程的可视化有两种方法,一种方法是对挖掘过程中产生的中间结果进行可视化呈现,方便用户根据中间结果的反馈调整参数和约束条件;另一种方法是将整个数据挖掘的处理过程以图

收稿日期:2012 - 10 - 10;**修回日期**:2012 - 12 - 20.
* **基金项目**:安徽省高校优秀青年人才基金(2012SQRL260);安徽省教育规划课题(JG11373);2010 年芜湖职业技术学院院级教学研究项目.
作者简介:张俊(1980 -),安徽安庆人,讲师,硕士,主要从事数据挖掘研究.

标和流程图的形式显示,用户可以观察数据的来源,数据集成、清理和预处理的过程,挖掘结果的存储和可视化表示等等。

(3) 结果可视化。数据挖掘结果可视化是在挖掘过程结束之后,以图形和图像的形式描述挖掘的结果或知识, 以提高用户对结果的理解,使用户更好地评估和利用挖掘结果。

2　可视化数据挖掘主要技术

将数据挖掘技术与可视化技术相结合,其动机一方面是为了利用人类的知识领域来指导数据挖掘的过程,从而提高挖掘的质量;另一方面是为了帮助分析人员快速且最大限度地获得数据中隐含的信息,理解数据挖掘的过程和结果。可视化技术根据是否包含物理数据,可分为科学计算可视化和信息可视化,科学计算可视化的重点放在如何真实有效地反映三维坐标场,而信息可视化的研究重点则是通过选择和设计合适的表达方式来描述大型的多维数据之间的联系,以便于用户理解。数据挖掘技术的可视化主要定位于信息可视化。

被可视化的数据类型包括一维数据(如时序数据)、二维数据(如地理数据)、多维数据、文本/Web 数据(首先要将其转化为向量描述,然后才能应用可视化技术)、层次/图形数据、算法/软件的可视化。可视化的技术可分为标准 2D/3D 技术、几何转换技术、面向像素的技术、基于图标的技术、分层技术。还可以将可视化技术与一些变形与交互技术相结合,以实现更有效的数据挖掘。

(1) 标准 2D/3D 技术。标准 2D/3D 技术,如折线图、条形图、柱状图、饼图、散点图等,在统计应用中常用到,但是在表示多维数据方面存在缺陷。

(2) 几何转换技术。几何转换技术的基本思想是通过几何学的投影和转换方法,通过线性或非线性的投影和映射,把多维数据集转换成二维平面或三维空间可以表示的形式,从而实现对数据集的降维处理。目的是发现多维数据集的令人感兴趣的投影。几何转换技术适用于数据量不大,但维数较多的数据集。几何转换技术的具体实现方法有投影追踪、地形图、散点图矩阵和著名的平行坐标法等。

(3) 基于图标技术。基于图标技术的基本思想是用图标上的各个特征对应描述一个数据项的多维属性值,并将所有的图标依据一定的顺序进行排列。其中图标可以随意定制为一些三维几何对象,而且图标的各项属性包括图标的大小、颜色、形状等均可用来描述数据项的维。基于图标技术适用于维数不多,但具有某些代表特殊含义的属性的数据,用户可以更准确清晰地理解这些属性。基于图标技术的实现方法有表长法、契诺夫脸谱图法、彩色图标法、形状编码法、枝形图法。

(4) 面向像素技术。面向像素技术的基本思想是用屏幕上不同的独立子窗口分别表示数据集中不同的属性,并在各个独立的子窗口中用一个个彩色像素来表示各个数据项的一个属性值,面向像素技术可以非常有效地描述大型数据集,用户不仅可以观察自己感兴趣的局部区域,还可以获得对数据的整体认识。面向像素技术研究的重点在于考虑这些像素点如何在屏幕上排列的问题,应根据不同的目的使用不同的排列方式。根据不同的像素及窗口排列方法,面向像素技术的具体实现方法主要有递归模式技术、圆环分段技术、数据管道技术等。

(5) 分层技术。分层技术非常适用于层次型数据集的可视化,它的基本思想是根据数据集的层次特征将多维数据空间划分为若干个子空间,然后依据数据集中各层次的关系将这些子空间以层次结构的方式组织起来,最后转换成图形输出,常采用的方法就是利用树形结构,直接可视化层次型数据集,或者对数据维依据不同的标准进行划分,在不同层次上表示不同的属性值。分层技术的具体实现方法主要有层次轴、维嵌套、锥形树、双曲线树等。

3 可视化数据挖掘技术的实现

可视化数据挖掘主要利用 Swing 技术、AWT、Java2D、结合 JFreeChart 开源工具包和 Java3D 技术开发实现,本论文的验证数据集,是著名的购物篮分析数据集。

(1) 可视化交互的实现。系统中用户可以很好的与可视化图形交互,包括设置可视化图形的颜色、形状,对产生的关联规则进行筛选、排序,用户还可以对结果进行移动、缩放、旋转等操作,从而获得关联规则挖掘结果的多角度视图。

(2) 数据可视化技术的实现。本文的数据可视化部分主要包括两部分,一部分是对单个数据属性的二维展示,另一部分是对整个数据集的可视化呈现。系统中对单个数据属性的二维展示,主要采用了饼图和条形图两种方法,用饼图可以清晰地描述属性中各个属性值所占的比重,直方图可以比较不同数据对象中相同属性的值。如图 1 所示,饼图表现的是对购物篮数据集中各商品占总购买的比例。

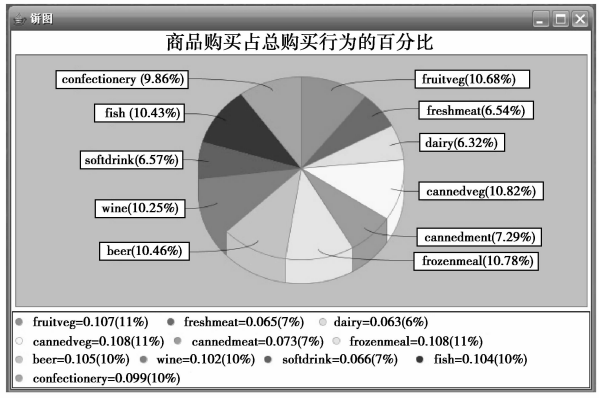

图 1 属性二维展示

系统中对数据集的可视化呈现包括散点图矩阵、平行坐标方法,这两种方法均可以有效地可视化高维数据集,平行坐标方法可以通过调整平行轴的顺序,很好地体现数据的功能依赖性。散点图矩阵方法可以很好地呈现数据的分布,方便用户发现孤立点。

(3) 过程可视化技术的实现。为了便于比较和验证本文所提出的过程可视化技术的有效性和优越性,本文对频繁项集的挖掘过程分别采用了文字化的表现方法和基于平行坐标的方法。基于平行坐标的方法是对传统平行坐标方法的一种改进,以平行坐标的每一条坐标轴表示对数据库的一次扫描,轴上均匀分布的是所有的1-项频繁集,第 i 条坐标轴和第 $i+1$ 条坐标轴之间的连线表示的是 $i+1-$项频繁集,各项集的支持度用轴间连线的粗细来描述,并且用不同的颜色将各频繁项集区分开,避免产生界面混乱的问题。用户可以根据中间结果的反馈来调整算法的参数和约束条件,从而改善挖掘结果,提高挖掘质量,并提升用户对挖掘结果的信赖度。对著名的购物篮分析数据集设置支持度阈值为 0.05 所得的平行坐标,如图 2 所示。

(4) 结果可视化技术的实现。本文对数据挖掘结果的可视化采用了基于三维坐标的方法,该方法可以更清晰直观的表示关联规则,并能够很好的避免界面紊乱、歧义、遮蔽的问题,也能够有效地表示多对多和多维的关联规则。根据 Apriori 算法的第二步,设置置信度阈值为 0.9,则所生成的强规则如表 1 所示。

用基于三维坐标的可视化方法表示这 7 条关联规则,其中 X 轴表示的是规则,Z 轴是所有的 1-项频繁集,X-Z 平面上各绿色方格对应规则的前项,红色方格对应规则的后项,Y 轴上红色的立方体表示规则的支持度,绿色的立方体表示的是规则的置信度,如图 3 所示。

由图 3 可以看出,基于三维坐标的关联规则可视化方法表达清晰准确,对于关联规则的参数也能直观地描述出来,界面不存在遮蔽的问题,且对于多对多的规则也能有效地呈现。

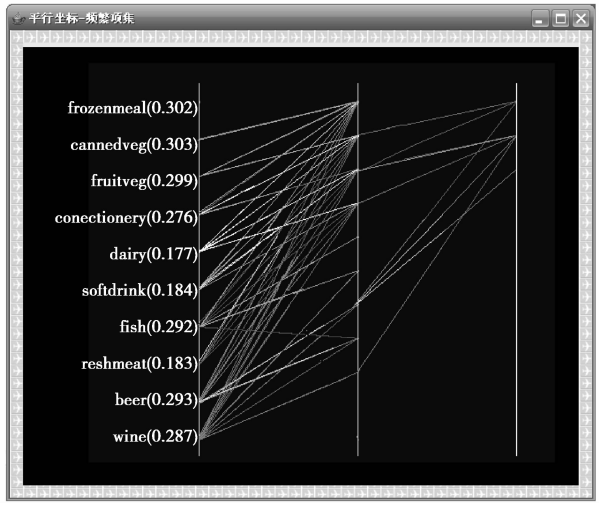

图 2　基于平行坐标的过程可视化方法

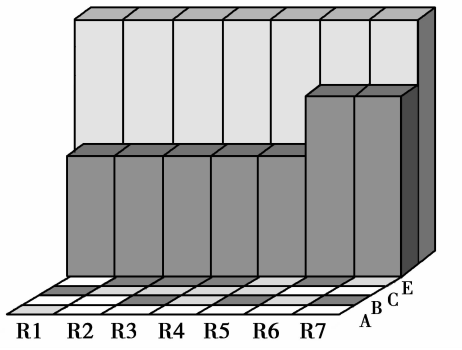

图 3　基于三维坐标的关联规则结果可视化

表 1　关联规则集

规则名	前项	后项	支持度	置信度
R1	A	C	0.5	1.0
R2	C	E	0.5	1.0
R3	C	B,E	0.5	1.0
R4	B,C	E	0.5	1.0
R5	C,E	B	0.5	1.0
R6	B	E	0.75	1.0
R7	E	B	0.75	1.0

4　结束语

在此主要就可视化数据挖掘技术的实现进行探讨,采用了一种基于改进的平行坐标技术的过程可视化方法来表示频繁项集的挖掘过程,实现了可视化交互、数据可视化、过程可视化和结果可视化。并通过用笔者所开发的原型系统对经典的购物篮分析问题进行了可视化数据挖掘,验证了本文所采用的可视化数据挖掘技术达到了预期目标,同时也突出了基于改进的平行坐标技术的过程可视化方法的有效性和优越性。

参考文献:

[1] 钟杨俊,文堂柳. 可视化数据挖掘方法与技术[J]. 福建电脑,2008,24(8):59,95

[2] XML Signature Working Group. XML-Signature Syntax and W3C Proposed Recommendation[S]. August 20,2001

[3] 刘玲. 基于数据挖掘系统的可视化技术研究[D]. 北京:北京工业大学,2010

[4] 罗文静. 数据挖掘中可视化技术研究与实现[D]. 成都:电子科技大学,2007

[5] 宁津生,郭金来. 地球重力场可视化数据挖掘平台 WHU-3Dgravity 的设计与实现[J]. 武汉大学学报:信息科学版,2007,32(11):945-949

[6]刘绪崇. 基于 OLAM 的可视化数据挖掘技术研究[D]. 国防科学技术大学,2002

[7] 陈霞,陈桂芬. 基于可视化的时空数据挖掘研究与应用[J]. 安徽农业科学,2012,40(17):9542-9545

[8] 胡俊. 数据挖掘可视化模型及其应用研究[D]. 北京交通大学,2009

(下转第 92 页)