数据挖掘工具的分类与挖掘

姚毓才[2],王本年[1,2]
(1.南京大学 计算机软件新技术国家重点实验室,江苏 南京 210093;
2.铜陵学院 计算机系,安徽 铜陵 244000)

摘 要:随着数据挖掘技术的发展,各种各样的数据挖掘工具不断开发出来,如何把握这些工具的功能、挖掘技术和未来发展趋势,是一个非常困难的事情。文中借助数据挖掘技术提出了数据挖掘软件工具的一个多维立方体分类模型,给出了一个具体分类实例,总结出数据挖掘工具的技术发展路线和未来发展趋势,并通过对三个不同阶段的数据挖掘工具的深入比较,进一步验证了文中的结论。

关键词:数据挖掘工具;分类;比较
中图分类号:TP18 **文献标识码**:A **文章编号**:1673-629X(2006)08-0006-04

Classifying and Mining for Data Mining Tools

YAO Yu-cai[2], WANG Ben-nian[1,2]
(1.National Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093, China;
2.Department of Computer Science, Tongling College, Tongling 244000, China)

Abstract: With the development of the data mining technology, a number of data mining tools were developed. It is a difficult thing for grasping functions, technology and developing trend of these tools. So gives a model of multi-dimension cub for classifying, and then puts forward a classified example. Presents the roadmap and trend of technology of data mining tools, which is supported by comparison of three tools from three different periods.

Key words: data mining tools; classifying; comparison

0 引 言

随着数字采集和存储技术的飞速发展和广泛应用,存储于数据库、数据库仓库和其它(如WWW等)储存库中数据已经越积越多,在如此庞大的数据背后隐藏着许多重要信息,如果人们不借助于强有力的工具,就很难或无法发现隐含在大量数据中各种关系和规则以及隐含在表象下的内在原因、机制和趋势规律。为了更好地利用这些数据,挖掘其背后隐藏的知识,需要有更高层次的分析技术和分析工具。在20世纪80年代,一个新的学科"数据挖掘"出现了,并得到了迅速发展[1]。在数据挖掘技术日益发展的同时,形形色色的数据挖掘的商业软件工具也竞相开发出来[2,3]。其实,要想对所有的数据挖掘软件工具进行综述介绍、比较,或给出一个科学合理的分类,也是一个非常困难的事情,这个工作的本身也是一个数据挖掘的过程[4]。

文中正是利用数据挖掘的技术试图给出这些软件工具的一个分类方法和分类实例,总结数据挖掘工具的技术发展路线和未来发展趋势,并通过对三个不同阶段的软件工具的比较挖掘,进一步验证文中的结论。

由于数据挖掘是在知识发现的基础上发展起来的,从不同的角度出发[4],也就有关于数据挖掘和知识发现的许多不同的定义,从而也就决定了它们之间的关系。对于数据挖掘文中倾向于这样的定义:数据挖掘是从巨量数据中获取有效的、新颖的、潜在有用的、最终可理解的模式的非平凡过程[5]。具有这种功能的软件工具即认为是数据挖掘工具。

1 数据挖掘工具的分类现状

关于数据挖掘工具的分类方法有很多。按使用领域划分,可分为:特定领域的数据挖掘工具和通用的数据挖掘工具。特定领域的数据挖掘工具是针对某个特定领域的问题提供解决方案,其针对性比较强,往往采用一些特殊的算法,可以处理特殊的数据,实现特殊的目的,因而发现知识的可靠度也比较高。通用的数据挖掘工具不区分具体数据的含义,采用通用的挖掘算法,处理常见的数据类型,挖掘什么、用什么来挖掘都由用户根据自己的应用来选择。按照软件所基于的平台划分,可分为:基于DOS的软件工具、基于Windows的软件工具、基于Linux的软件工具、基于Solaris的软件工具等。然而这样的划分方

收稿日期:2005-11-25
基金项目:安徽省教育厅自然科学基金项目(2005kj093)
作者简介:姚毓才(1976-),男,安徽桐城人,助教,研究方向为数据挖掘、机器学习。

法并没有给人们带来多少更深层的信息。从这一点考虑,需要有更好的分类方法,这里主要介绍如下两种。

1.1 基于特性的分类方案

基于特性的分类方案[2]主要是基于这样的考虑:不同的数据挖掘工具可能执行不同的数据挖掘任务和采用不同的方法来达到它们的目的,一些可能需要或支持更多地与用户交互,一些可能运行在单机上,而一些可能运行在C/S模式下,为了获取这些不同性,M. Goebel 和 L. Gruenwald 提出了利用特性分类的方案。

M. Goebel 和 L. Gruenwald 基于上述分类方案对当时的 43 种数据挖掘软件工具进行了分析,得到了 3 张非常直观、非常有说明性的特征表,并且较为正确地预测了数据挖掘软件工具的发展趋势。然而,随着数据挖掘软件工具的日新月异,这份调查已不能概括现今数据挖掘软件工具的发展。

1.2 KDnuggets 分类方案

KDnuggets. com[6]是关于数据挖掘、知识发现和决策支持方面的一个最主要的信息源,其中包含了对大量数据挖掘软件工具的 Website 链接和分类介绍、民意调查,以及供数据挖掘研究测试使用的数据集的 Website 链接和介绍等。

从 KDnuggets 所呈现的数据挖掘软件工具的 Website 链接和分类介绍可以清楚地看出,它实际给出了一个基于链接(文档)的树型层次结构关系的分类框架。这种分类方法的好处是层次清楚,然而其分类集则不是正交的,其交集中隐藏了许多信息,使人很难从整体上把握。

2 数据挖掘软件工具的多维立方体表示模型

借鉴多维数据立方体表示的概念[1],提出了数据挖掘软件工具的多维立方体分类方案,即用数据挖掘工具的不同属性类作为多维立方体的维,从而建立一个多维的立方体模型,然后在此基础上进行有关知识的挖掘。例如可以建立这样一个三维立方体模型,它的三个维分别是:数据源(包括:数据库、文本文件 Web 数据、多媒体数据等);方法(包括:统计方法、决策树、规则抽取、基于事例的推理、Bayes 网络、遗传算法、神经网络、模糊集、Rough 集等);功能(包括:预处理、关联分析、回归分析、分类、预测、聚类分析、异常分析、模型可视化等),如图 1 所示。

与其它方法相比,数据挖掘软件工具的多维立方体分类模型有如下优点:

(1)可以精心选择与要挖掘的知识相关的属性作为多维立方体的维,例如,可用时间(1990 年以前、1990～1995 年、1996～2000 年、2000 年以后)作为一维,这样可以利用时间序列分析技术挖掘出不同阶段主导产品、主流技术,甚至是发展趋势等知识。

(2)具有很好的可伸缩性,很容易将上述模型中的某些属性进行压缩和扩充,例如:可将基本分析扩充为{统计方法、决策树、规则抽取、基于事例的推理等};可以在方法维后追加{模糊集、Rough 集等};可以将数据源维中的{Web 数据、多媒体数据}合并;还可以对整个维进行压缩和扩充,例如:可以将数据源维整个地压缩掉,可以追加一个时间维;有时为了定量对一些软件工具进行评估,可以对某些维的属性赋以权值等。

(3)可以充分利用现有数据挖掘技术进行知识的挖掘。

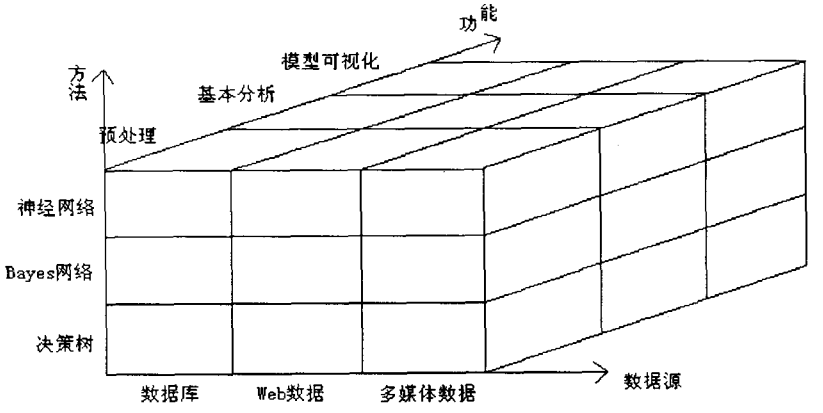

图 1　数据挖掘工具的三维立体表示

3 数据挖掘软件工具的多维立方体分类实例及结果分析

利用上述数据挖掘软件工具的多维立方体分类模型,可以方便地对数据挖掘软件工具进行分类。实际上,很容易用多维立方体分类方案(二维立方体分类方案)来表示基于特性的分类方案,即将产品名称作为一维,其它一般属性作为另一维;数据库联接属性作为一维;数据挖掘特征作为另一维。对于 KDnuggets 分类,也很容易给出二维立方体分类方案表示,即将挖掘方法作为一维,将法律状态作为另一维。因此,基于特性的分类方案和 KDnuggets 只是多维立方体分类方案的两个二维表示特例。

为了进一步说明多维立方体分类方案,这里给出另一种基于多维立方体分类方案分类实例:时间—产品—数据源/方法(如表 1 所示),其中:

(1)时间{1991～1995 年、1995～1999 年、2000～2003 年};

(2)产品:1991～2003 年主要产品;

(3)数据源{DB/TEXT = 数据库/文本文件、Web = Web 数据}、方法{ NN = 神经网络、GA = 遗传算法、FS = 模糊集、RS = Rough 集、St. = 统计方法、DT = 决策树、RI = 规则抽取、BN = Bayesian 网络、CBR = 基于事例的推理、CVM = 支持向量机、Ag = Agent}。

对表 1 分析可得到如下结果:

(1)各阶段主要技术及技术发展趋势:1991～1995 年,主要是基于统计、决策树和规则抽取等技术。发展至

表 1 基于多维立方体分类方案分类实例——时间－产品－数据源/方法

时间	数据源	产品名	方法										
			NN	GA	FS	RS	St.	DT	RI	BN	CBR	CVM	Ag
91～95[2]	DB/TEXT	CN2(Univ. of Texas)						X					
		Ecolweb (Tel Aviv Univ.)				X							
		BrainMaker (Cal. Sc. Software)	X										
		IND (Nasa)					X						
		C5.0 (RuleQuest)					X						
		Mobal (GMD)							X				
		Brute (Univ. of Washington)							X				
		AutoClass C (Nasa)					X						
		Darwin (Thinking Machines)	X	X				X					
		PVE (IBM)					X						
		Ripper (AT&T)							X				
		WizRule (WizSoft)								X			
96～99[2]	DB/TEXT	Kepler (GMD)	X				X	X	X		X		
		MLC++ (Silicon Graphics)					X	X	X		X		
		MSBN (Microsoft)								X			
		Sipina－W (Univ. of Lyon)					X	X					
		Bayesian Knowl. Disc. (Open U.)								X			
		Data Surveyor (Data Destilleries)	X				X	X	X	X			
		DBMiner (SFU)					X		X				
		IDIS (Information Discovery)			X				X				
		MineSet (Silicon Graphics)					X						
		ModelQuest (AbTech)	X				X	X	X				
		Rough Enough (Troll Data)				X			X				
		Scenario (Cognos)					X	X					
		SuperQuery (AZMY)					X		X				
		Weka (Univ. of Waikato)					X	X	X		X		
		Clementine (Integral Solutions Ltd.)	X					X	X				
		Intelligent Miner (IBM)	X					X					
		PolyAnalyst (Megaputer)		X			X		X				
		Rosetta (NTNU)		X		X	X		X				
		Decision Series (NeoVista)	X				X	X	X				
		KATE－Tools (AcknoSoft)						X			X		
2000 以后	DB/TEXT	CART 5.0						X					
		Classification Tree in Excel						X					
		Grobian				X							
		KINOsuite－PR	X						X				
		KXEN										X	
		LIBSVM										X	
		LS－SVMlab										X	
		PolyAnalyst,			X			X	X				
		SuperQuery							X				
		SvmFu 3										X	
		SVMlight										X	
		WINROSA			X		X		X				
		XpertRule Miner (Attar Software)						X	X				
	WEB	XML Miner and XML Rule											
		DeepMetrix Mining											
		Lumio Re:cognition suite											
		Neptunet											
		IBM SurfAid											

1996～1999 年,统计、决策树和规则抽取技术已成为基本技术方法,而 Bayes 网络、基于事例的推理技术已经得到了相当好的进展,甚至基于模糊集和 Rough 集的技术也开始得到应用;这一阶段的另一个主要特点是多方法的集成。到 2000 年以后,不仅模糊集和 Rough 集技术得到了进一步的应用,而且像支持向量机这样的新技术也颇受关注。

(2)对数据源的支持:1991～1995 年,主要是对文本文件和 Dbase 文件的支持,而到 1996～1999 年,由于大规模分布式数据库的广泛应用,使得对象 Sybase,Oracle 等这样大型的分布式数据库的支持,成为这一阶段主要特点之一;另一个主要特点是对多数据库的支持。到 2000 年以后,由于 Internet 技术的发展,Internet 本身构成了一个巨大的分布式数据库,因此基于 Internet 的数据挖掘和信息检索,是这一阶段的一个热点研究领域,另外就是多媒体数据的挖掘也得到了研究者的重视。

(3)从上述数据挖掘发展的路线图可预测数据挖掘软件工具的未来发展趋势:其一是对多数据源的混合支持和多方法的集成与溶合;其二是对 Web 挖掘技术的研究(特别是基于 Agent 技术的研究)和多媒体挖掘技术的研究。

4　三种数据挖掘软件工具的比较

为了进一步比较三个阶段数据挖掘工具软件的异同性,选取三个不同阶段的工具软件作深入比较,这三个软件分别是:See5,Rosetta 和 DeepMetrix Mining。

1)See5 是一个非常简单的、基于决策树和规则集的形式对数据进行分析和分类的工具软件。它对数据源的要求是必须符合一定格式的文本文件。它的所有功能主要集中在一个含有多个选项的对话框上,它的基本分类方法是基于决策树的方法,但也可以用基于规则集的方法。

2)Rosetta 是一个基于规则的工具,它也要求符合一定格式的文本数据源,但它不象 See5 只支持单文档窗口界面,它可以打开多个多文档窗口(因而它支持同时对几个数据源的操作),其结构化的数据(Structures)和可操作的算法(Algorithms)或功能集中一个多文档窗口的一个项目树中,项目树中的对象支持拖放操作,也可用鼠标右键弹出操作功能菜单形式,选择所要操作的功能项执行相关操作。在 Structures 节点中,包含打开的数据源及操作结果等结构对象,结构对象可用 View 查看,并可对有关对象执行统计处理。

显然在功能上 Rosetta 比 See5 要强得多,虽然它不显式地支持决策树,但它支持统计、遗传算法、甚至 Rough 集等处理技术,可以对不完全的数据进行填充,这些是 See5 所做不到的。

3)DeepMetrix Mining 是 DEEPMETRIX 公司开发的基于 Web 的数据挖掘工具,与 See5 和 Rosetta 截然不同的是:

(1)DeepMetrix Mining 是一个商业化软件工具;

(2)DeepMetrix Mining 是一个基于 Web 数据源的数据挖掘工具;

(3)DeepMetrix Mining 可以实现在线挖掘。

DeepMetrix Mining 通过对访问某个 Website(可以是基于 ASP 和 CGI)的客户进行跟踪,从而获得大量实时的数据,依此来分析客户的行为,并从各种角度挖掘信息,形成统计、比较和预测报告及分类报告,值得一提的是 DeepMetrix Mining 提供的非常直观图形报告,这是 See5 和 Rosetta 所没有的。

从上述三个不同阶段的工具的深入比较,可以进一步看出数据挖掘工具的发展路线和发展趋势。

5　总　结

随着计算机应用的普及,对数据挖掘工具的要求越来越高,而数据挖掘技术的飞速发展,则给数据挖掘工具的开发提供了技术上的支持,因而各种数据挖掘工具不断被开发出来。文中利用数据挖掘技术提出了数据挖掘软件工具的一个多维立方体分类模型,对一些主要的数据挖掘工具软件进行分类,通过实例分析,总结出数据挖掘工具的技术发展路线和未来发展趋势,并通过对三个不同阶段的数据挖掘工具的深入比较,进一步验证文中的结论。数据挖掘工具的发展离不开应用的需求和数据挖掘技术研究的背景,将数据挖掘工具与用户的需求和对数据挖掘技术研究结合起来进行分析,这将是下一步的研究工作。

参考文献:

[1] Han J, Kamber M. Data Mining: Concepts and Techniques [M]. San Francisco, CA: Morgan Kaufmann, 2001.

[2] Goebel M, Gruenwald L. A Survey of Data Mining and Knowledge Discovery Software Tools [J]. SIGKDD Explorations, 1999, 1: 20 - 33.

[3] Adriaans P, Zantinge D. Data mining [M]. London: Addison Wesley Longman, 1999. 40 - 100.

[4] Zhou Zhi - Hua. Three perspectives of data mining [J]. Artificial Intelligence, 2003, 143: 139 - 146.

[5] Fayyad U, Piatetsky - Shapiro G, Smyth P. Knowledge Discovery and Data Mining: Towards a Unifying Framework [A]. In: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD - 96) [C]. Portland, Oregon, CA: [s. n.], 1996. 82 - 88.

[6] Chang Chin - Chung, Chih Jenlin. LIBSVM: A Library for Support Vector Machines [M]. [s. l.]: [s. n.], 2004.

(上接第 5 页)

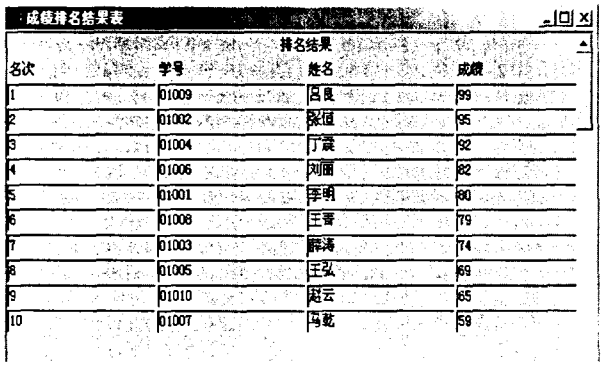

图 4　成绩排名结果表窗口

参考文献:

[1] 吴小东. Java 程序设计基础 [M]. 北京:清华大学出版社, 2002.

[2] 耿祥义,张跃平. Java 2 实用教程 [M]. 北京:清华大学出版社, 2001.

[3] 朱战立,沈　伟. Java 程序设计实用教程 [M]. 北京:电子工业出版社, 2005.

[4] 印　昱. Java 与面向对象程序设计教程 [M]. 北京:高等教育出版社, 2002.

[5] 张　晨,付　冰,赵　军. Java 2 应用编程 150 例 [M]. 北京:电子工业出版社, 2003.