一个面向大规模数据库的数据挖掘系统*

钱卫宁, 魏 藜, 王 焱, 钱海蕾, 周傲英

(复旦大学 计算机科学与工程系,上海 200433);
(复旦大学 智能信息处理开放实验室,上海 200433)
E-mail: {wnqian,lwei,ayzhou}@fudan.edu.cn
http://www.fudan.edu.cn

摘要: 数据挖掘融合了数据库技术、人工智能和统计学,是目前的研究热点.为了能够集成当前数据挖掘的主要技术并使它们协同工作,在进行数据挖掘基本算法研究的基础上研制开发了一个数据挖掘系统——Golden-Eye.系统实现了在数据挖掘研究中的一些最新成果,集成了泛化、数据清洗这两个数据准备操作以及关联规则发现、例外规则发现、时序模式发现、分类器构造、聚类分析等基本数据挖掘操作,并实现了对挖掘操作的基本管理和结果的图形化显示.整个框架设计充分体现了系统的完整性、协调性和高效性:自底向上将存储控制模块、数据预处理模块、挖掘操作模块、挖掘库管理模块有机地结合在一起,在底层实现了对包括中间结果在内的数据的统一管理,在上层为用户提供了可视化的界面.实验结果表明,该系统能够在大规模数据库上成功地完成用户所指定的数据挖掘操作.

关 键 词: 数据挖掘;系统;数据预处理;存储控制;挖掘库
中图法分类号: TP311 文献标识码: A

数据挖掘(data mining)又被称作数据库中的知识发现(knowledge discovery in databases),是指从数据库或数据仓库中提取隐含的、未知的和潜在的有用信息的非平凡过程.数据挖掘技术主要包括关联规则(association rule)发现、分类(classification)、聚类(clustering)分析、泛化(generalization)和预测(prediction)等.当前,数据挖掘的研究热点在于提高挖掘所得的知识的准确度和可理解性、提高数据挖掘操作的可伸缩性、集成数据挖掘操作和现有的数据存储和分析工具等.此外,作为数据挖掘准备工作的数据离散化、数据变换、数据清洗(data cleaning)和数据挖掘结果的可视化显示以及挖掘结果的评估等技术也属于数据挖掘研究的范畴.

虽然数据挖掘包含诸多方面的工作,但在实际运用中,这些方面的技术往往需要相互协作,共同完成某项挖掘任务.这就需要数据挖掘工具能够集成各方面的技术,使它们能够协同工作,并统一管理各个挖掘步骤以及结果.Golden-Eye 系统就是为此目的而开发的.本系统具有如下特点:

- 集成了泛化、数据清洗、关联规则发现、时序模式(sequential pattern)发现、分类、聚类等多种基本数据挖掘操作.
- 集成了一些新的操作和新的算法,比如改进的 DBSCAN 聚类算法以及例外规则发现、数据清洗、类别属性(categorical attribute)聚类这些数据挖掘领域里较新的操作.
- 能处理大规模的数据集,测试的最大记录数目达到了 1 000 000 条.
- 在系统框架的设计上充分考虑到了系统的完整性、协调性和高效性.

* 收稿日期: 2001-04-05; 修改日期: 2002-01-24
基金项目: 国家自然科学基金资助项目(60003016);国家重点基础研究发展规划 973 资助项目(G1998030414)

作者简介: 钱卫宁(1976－),男,浙江上虞人,博士生,主要研究领域为数据挖掘,聚类,Web 数据管理;魏藜(1978－),女,江西南昌人,硕士生,主要研究领域为数据挖掘技术;王焱(1977－),女,江苏镇江人,硕士,主要研究领域为数据挖掘,Web 数据管理;钱海蕾(1977－),女,上海人,硕士,主要研究领域为数据挖掘,聚类,Web 数据管理;周傲英(1965－),男,安徽宣城人,博士,教授,博士生导师,主要研究领域为 Web 数据管理,数据挖掘,Web 搜索.

- 有一个友好的用户界面.

1 系统结构

1.1 系统框架

如图 1 所示为 Golden-Eye 系统的框架.整个系统将不同的挖掘操作模块、数据预处理模块、存储控制模块、挖掘库及挖掘库管理模块、数据库和外部文件紧密地结合在一起,构成了一个层次结构.系统框架的设计主要基于以下几点考虑:

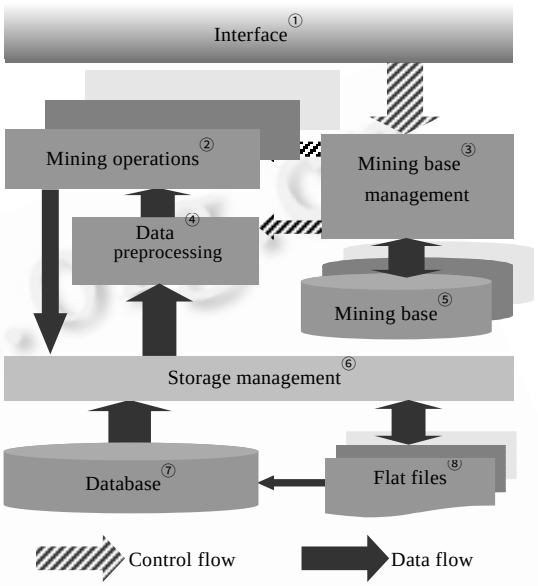
(1) 数据挖掘系统包括很多方面的操作,这些操作所要求的数据源形式不同、输出不同、所需参数不同,这就使得实现这些操作的各个挖掘操作模块之间必须相对独立.

(2) 数据挖掘系统作为一个整体,必须能够协调各个操作模块之间的工作.系统使用挖掘库提供统一的机制来管理各模块所使用的数据源、参数和挖掘结果.

(3) 数据挖掘的对象既可能存在于数据库或数据仓库中,也可能存在于文件中,系统应该分别提供处理它们的相应方法.

(4) 数据挖掘的结果需要保留.这一方面是因为数据挖掘的目的是支持决策分析;另一方面是为了方便重新挖掘、增量挖掘.

(5) 作为一个支持决策分析的系统,其使用者不是计算机工作者,而是决策者,系统应该提供友好的界面.



①界面,②挖掘操作模块,③挖掘库管理模块,④数据预处理模块,⑤挖掘库,⑥存储控制模块,⑦数据库,⑧外部文件.

Fig.1 System architecture

图 1 系统框架

1.2 功能模块

1.2.1 挖掘操作模块

不同的挖掘操作模块负责不同的数据挖掘操作.它们彼此之间相对独立,共同之处是都受到挖掘库管理模块的管理,通过存储控制模块获得数据,并把结果写入挖掘库.在下一节里我们将详细介绍各个操作模块.

1.2.2 数据预处理模块

数据预处理模块的主要功能是定义数据源、格式化数据源以及过滤数据源.该模块对整个系统的可用性非常重要,它可以分为以下几个子模块:

- 数据映射.将源表中的数据映射成 ID 形式,并生成对照表(ID 和原始值的对照).此功能的目的是把不同形式的数据映射成统一的、可供挖掘模块操作的形式.
- 类型映射.对源表中所列数据类型进行强制类型转换.之所以需要这个功能,是因为在数据库中不同的数据类型很多,数据挖掘算法只支持其中最基本的几种.
- 列映射.该子模块从源表中提取所需要的列,以减少数据量,提高系统的效率.

1.2.3 存储控制模块

系统假设数据源存放在数据库中,由存储控制模块对数据库统一进行操作.对于存放在外部文件中的数据,需要使用数据库管理系统提供的导入工具把数据导入数据库以后再进行挖掘操作.当前,系统的数据源存放在 DB2 UDB 5.2 中,从可移植性的角度考虑,我们使用 ODBC 作为底层的接口.我们对存储控制的封装高于 ODBC 对存储控制的封装,这是因为数据挖掘应用不同于一般的数据库应用程序,它对数据库的访问频繁,而每次对数据库的访问都会耗费一定的时间和资源.对于数据挖掘操作来说,对大数据量的处理能力和处理效率是一个根

本的问题,所以,由系统来进行缓冲和内存索引就非常重要.存储控制模块的功能主要体现在 3 个方面:

- 对连接数据库、管理外部文件以及交换外部文件和内存的内容等较为底层的操作进行封装.
- 负责缓冲管理.具体地说,该模块为数据源、数据挖掘中间结果以及挖掘结果分别申请缓冲区,并保证其驻留在内存中.
- 提供简单的数据格式转换.不同于数据预处理模块提供的数据格式转换,该功能主要弥补关系数据库不能存储不规则格式数据的问题:在向缓冲区中存放数据以前对事务记录进行重新拼接.

正是由于存储控制模块的存在,系统才获得了良好的可扩展性,各个算法的检测数据集规模都达到了 100 000 条以上,其中部分算法的检测数据集规模达到了 1 000 000 条.具体的实验数据参见相关论文.

1.2.4 挖掘库及挖掘库管理模块

挖掘库和挖掘库管理是整个系统的核心部分.挖掘库是一个逻辑概念.一个挖掘库存放用户所指定的一系列挖掘操作的所有信息.在系统中,所有的挖掘库都统一存放在数据库中,由系统统一管理.

挖掘库所保存的挖掘操作是指包括数据准备和数据挖掘在内的所有操作.在挖掘库中存放的这些操作信息是有顺序的(用户进行这些操作的顺序).这是因为一个数据挖掘操作在整个知识发现过程中往往不是孤立的,它所使用的数据源常常是另一个数据挖掘操作的结果,而它的挖掘结果又有可能是其他操作的数据源.所以,保留挖掘顺序实际上就是保留了挖掘操作之间的这种关系,这无论对用户理解挖掘结果还是以后重新进行挖掘都是有帮助的.除了操作的名称和顺序以外,挖掘库还保存数据源信息、挖掘操作的参数设置以及挖掘的结果.因此,我们的系统能够很方便地实现把一个挖掘操作的结果作为另一个挖掘操作的输入.

我们提供了一套管理挖掘库的操作,这些操作被封装成挖掘库管理模块.图形界面通过调用挖掘库管理模块来完成对挖掘库的管理.同时,挖掘库管理模块通过调用各个挖掘操作模块来实现挖掘操作.管理挖掘库的所有操作可以被分成以下 4 类.

- 对挖掘库的操作.这组操作主要提供对挖掘库整体的管理.包括连接挖掘库、断开挖掘库、打开挖掘库、增加挖掘库、存储挖掘库、删除挖掘库和查询挖掘库.任何对挖掘库的操作必须在打开了一个挖掘库以后才能进行,而系统的任意运行时刻最多只能打开一个挖掘库.
- 对数据源的操作.这组操作主要用于定义数据源.包括查询数据库信息、增加数据源、查询数据源信息等.
- 对挖掘操作的设置操作.包括增加挖掘操作、查询挖掘操作、设置挖掘操作参数、查询挖掘操作参数等.
- 对挖掘结果的操作.系统实现了对挖掘结果的查询操作.

1.2.5 界 面

Fig.2 The graphical interface of the decision tree and the background: the GUI of the system

图 2 分类操作树型结果的图形界面以及类 Explorer 风格的主图形界

Golden-Eye 提供 API 作为访问界面,并在此基础上开发了图形界面.

系统的主图形界面采取类 Explorer 的风格.我们使用不同形式的图形技术来表示不同形式的挖掘结果.如图 2 所示为分类操作树型结果的图形界面.目前,Golden-Eye 所使用的图形界面及其针对的挖掘操作如下:

- 表格.泛化和数据清洗的结果显示.
- 树型结构.决策树的显示.
- 2 维点/3 维点.聚类结果的图形显示.
- 文本.关联规则、例外规则以及时序模式的显示.

2 系统功能

2.1 数据准备

2.1.1 泛 化

泛化就是将相关数据或概念泛化到更高级的层次上.本系统集成的泛化算法是 GDBR[1].该算法的特点是:对比其他算法(如 LCHR,AOG 等),它有最好的时间复杂度 $O(n)$以及很好的空间复杂度 $O(c)$.

2.1.2 数据清洗

数据清洗的主要工作就是准确、高效地检测出数据库中的相似重复记录.系统使用一种基于 N-Gram[2]的检测相似重复记录的综合方法,能处理常见的拼写错误,如插入、删除、交换、替换和单词的交换等.为了消除基本算法在检测精度上的一些不足,系统采用了经过改进的算法[3],在实现中运用了统计学原理较好地去除了噪声,并综合应用了正向和逆向重复矩阵,提高了插入/删除错误的检测率.

2.2 数据挖掘操作

2.2.1 关联规则

关联规则发现可以分两步来完成:找出所有的频繁项集;由发现的频繁项集生成关联规则.在找出所有的频繁项集时,系统实现了 Agrawal 等人提出的 Apriori 算法[4].由频繁项集产生关联规则的基本思想是:对于每一个频繁项集 l,找出 l 的所有非空子集 a,如果 $support(l)/support(a)>\min conf$,则输出规则 $a\Rightarrow(l-a)$.系统对这种方法略作改进:如果频繁项集 l 的子集 a 不能产生出规则,则没有必要用 a 的子集来产生关联规则.

2.2.2 例外规则

广义的关联规则可分为 3 类:强规则、例外规则和随机规则.强规则(大部分数据服从的规则)可以帮助我们预料将来的情况.然而在一些特定的场合,我们需要的不仅仅是预测,而是我们还不知道的知识.这时,我们更感兴趣的是发现例外规则(小部分数据服从的高可信度规则)[5].

设 I 为数据集 D 中的所有记录,Y 为属性集.对于分类关联规则(CAR)$X\Rightarrow y(X\subset I,y\in Y)$,它在两种情况下会成为虚假的规则(SCAR):一是若 y 的支持度大于 $X\Rightarrow y$ 的支持度;二是若 $X'\subseteq X$ 且$(X'\Rightarrow y)$的支持度大于$(X\Rightarrow y)$的支持度.若 $X\Rightarrow y$ 为 CAR,例外分类关联规则(ECAR)有如下的形式:$X,Z\Rightarrow \tilde{y}$,其中 $X,Z\in I,X\cap Z=\varnothing,\tilde{y},y\in Y,\tilde{y}\neq y$.例外分类关联规则的可信度满足最小可信度的设置,但支持度低于最小支持度的设置.

我们的例外规则挖掘模块包含 3 个子模块:生成 CAR、删除 SCAR、生成 ECAR.算法的细节请参见文献[5].

2.2.3 时序模式

同关联规则一样,挖掘时序模式的问题也源于由销售记录组成的事务数据库 D,但时序模式主要是对物品(项)在时间上的关联性加以考虑.Golden-Eye 系统集成的时序模式挖掘算法是 Agrawal 等人提出的 AprioriAll 算法[6],关于算法的说明在此不作赘述.

2.2.4 分 类

分类的基本思想是:根据一些已定义好类别的数据的信息,产生一个可以描述数据类别或对未知类别的数据进行分类的分类器.本系统集成的分类算法最终生成的分类器被称为区间分类器(interval classifier)[7].该算法的特点在于与采用二叉树的决策树分类器相比,它的准确度较高,决策树的深度也不至于过深.

2.2.5 考虑综合因素的聚类方法

系统集成的考虑综合因素的聚类方法[8]吸收了一些现有聚类算法的优点,使用层次聚类方法的框架,综合考虑簇之间的距离和簇中对象的密度来决定两个簇是否应该合并.它吸收了 CURE 算法中采用多个代表点来表示簇的方法,因而能够有效地识别特殊形状的簇.为了增强处理大数据量的能力,在使用层次聚类法之前,算法将对象所分布的数据空间划分成数据单元,计算统计信息后得到初始的簇.最后,算法利用索引对数据库中的所有对象进行标记.该算法的主要步骤如下:(1) 取样;(2) 划分数据单元;(3) 消除噪声;(4) 利用距离和密度判断、合并簇;(5) 识别 outlier;(6) 标记数据.

2.2.6 改进的基于密度的聚类方法

本系统还包含了另一种聚类算法,即对 DBSCAN 算法[9]加以改进后得到的一种高效算法[10].我们通过以下 3 个方面对 DBSCAN 进行有效的改进:

首先使用快速算法.在选取下一步扩展的种子点时我们只选取具有代表性的一部分邻居对象,这样就提高了算法的速度.

其次,用数据分区改进.根据数据空间的分布特性,将整个数据空间划分为多个较小的分区,然后分别对这几个分区进行聚类,最后将各局部聚类进行合并.

最后,通过数据取样改进算法.对数据进行取样,并通过对取样数据的聚类计算来达到对整个数据库数据的聚类.

3 相关工作

国际上的一些数据挖掘研究者已经开发了一些集成的数据挖掘系统,例如 DBMiner[11],QUEST[12]等,这些系统大都是研究成果向产品的过渡.与这些系统不同,Golden-Eye 致力于提供一个数据挖掘技术研究的试验平台.一方面可以在集成的环境下对人工数据或者真实数据进行挖掘操作,检验新开发的挖掘技术的正确性和有效性;另一方面,系统以后可以集成更多的数据挖掘功能,并能投入到真实的数据挖掘应用中去.

一些国内的研究者也开发了集成的数据挖掘系统.其中有些系统尚处于初步实现阶段,侧重于系统结构的搭建,仍未考虑大数据量的处理等重要问题,例如,RoboMiner[13].另一些则侧重于数据挖掘的流程的管理,例如 Open Miner[14].与我们的系统不同,这些系统都只集成了一些现有的典型算法,其开发目的更偏向于应用.

4 总结与展望

我们开发的数据挖掘系统 Golden-Eye 成功地集成了数据挖掘和数据准备的几个方面的功能.从结构上看,系统利用挖掘库将各个挖掘操作松散且一致地结合起来,便于扩充新的挖掘操作模块;从功能上看,我们集成了一些新兴的数据挖掘操作;从实现上看,我们实现了一些自创或者经过改进的算法.

当然,本系统还存在着一些不足之处.首先,对各挖掘操作的集成还不够紧密,挖掘操作只能简单地按顺序进行.其次,系统并未考虑与 DBMS 和 OLAP 工具的集成.

我们还需要在以下几个方面做更多的工作:

- 在现有平台的基础上开发新的挖掘操作;
- 更紧密地集成各个数据挖掘操作;
- 集成简单的数据库操作和数据仓库操作.

测试真实数据,检验系统的可用性和有效性.

致谢 金文博士、周水庚博士对系统的设计和实现给予了细心的指导,邱越峰、范晔、胡江滔、俞舫、曹晶等同志对主要模块做了实现,在此一并表示感谢.

References:

[1] Carter, C.L., Hamilton, H.J. Efficient attribute-oriented algorithms for knowledge discovery from large databases. IEEE Transactions on Knowledge and Data Engineering, 1998,10(2):193~208.

[2] Kukich, K. Techniques for automatically correcting words in text. ACM Computing Surveys, 1992,24(4):377~439.

[3] Tian, Zeng-ping, Lu, Hong-jun, Ji, Wen-yun, *et al*. An *n*-gram-based pproach for detecting approximately duplicate database records. International Journal on Igital Library, 2001,5(3):325~331.

[4] Agrawal, R., Srikant, R. Fast algorithms for mining association rules in large databases. In: Proceedings of the VLDB. 1994. 487~499.

[5] Yu, Fang, Jin, Wen. An effective approach to mining exeption class association rules. In: Proceedings of the Web-Age Information Management 2000. 2000. 145~150.

[6] Agrawal, R., Srikant, R. Mining sequential patterns. In: Proceedings of the ICDE. 1995. 3~14.

[7] Agrawal, R., Ghosh, S., Imielinski, T., *et al*. An interval classifier for database mining applications. In: Proceedings of the VLDB. 1992. 560~573.

[8] Zhou, Ao-ying, Qian, Wei-ning, Qian, Hai-lei, *et al*. A hybrid approach to clustering in very large databases. In: Proceedings of the 5th PAKDD. 2001. 519~524.

[9] Ester, M., Kriegel, H.P., Sander, J., *et al*. A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of the KDD. 1996. 226~231.

[10] Zhou, Ao-ying, Zhou, Shui-geng, Cao, Jing, *et al*. Approaches for scaling DBSCAN algorithm to large spatial databases. Journal of Computer Science and Technology, 2000,15(6):509~527.

[11] Han, J., Fu, Y., Wang, W., *et al*. DBMiner: a system for mining knowledge in large relational databases. In: Proceedings of the KDD. 1996. 250~255.

[12] Agrawal, R., Mehta, M., Shafer, J.C., *et al*. The quest data mining system. In: Proceedings of the KDD. 1996. 244~249.

[13] Liu, Li-jun, Huang, Ya-lou, Xue, Bin, *et al*. The design and implementation of the data mining prototype system: RoboMiner. Computer Sciences, 2000,27(10):57~60 (in Chinese).

[14] Shao, Hua, Wan, Jia-hua, Wang, Jian-hu, *et al*. A user-centered data mining tool: open miner. Computer Sciences, 2000,27(10):68~72 (in Chinese).

附中文参考文献:

[13] 刘丽君,黄亚楼,薛彬,等.数据挖掘原型系统 RoboMiner 的设计和初步实现.计算机科学,2000,27(10):57~60.

[14] 邵华,万家华,王剑虎,等.一个以用户为中心的数据挖掘工具:Open Miner.计算机科学,2000,27(10):68~72.

A Data Mining System for Very Large Databases*

QIAN Wei-ning, WEI Li, WANG Yan, QIAN Hai-lei, ZHOU Ao-ying

(*Department of Computer Science and Engineering, Fudan University, Shanghai* 200433, *China*);

(*Laboratory for Intelligent Information Processing, Fudan University, Shanghai* 200433, *China*)

E-mail: {wnqian,lwei,ayzhou}@fudan.edu.cn

http://www.fudan.edu.cn

Abstract: Data mining is a hotspot that combines the techniques in databases, artificial intelligence and statistics areas. On the basis of the research on some data mining algorithms and their implementation, a data mining system, Golden-Eye, is developed to incorporate primary data mining techniques and coordinate their operations. As the integration of several existing techniques including some improved algorithms as well as some newly proposed operations in data mining area, the system implements a wide spectrum of data mining functions such as generalization, data cleaning, association rule mining, exception rule mining, sequential pattern mining, classification and clustering. By tightly integrating different functional modules such as storage management, data preprocessing, mining operations and mining base management, the system succeeds in managing all kinds of data including midterm results uniformly and providing a user-friendly, visualized interface, which makes Golden-Eye a complete and efficient system with good performance. Experimental results show that the system can successfully fulfill the mining tasks specified by users on very large databases.

Key words: data mining; system; data preprocessing; storage control; mining base

* Received April 5, 2001; accepted January 24, 2002

Supported by the National Natural Science Foundation of China under Grant No.60003016; the National Grand Fundamental Research 973 Program of China under Grant No.G1998030414