数据挖掘中数据预处理的研究与实现*

菅志刚, 金　旭
(北京科技大学 信息工程学院, 北京 100083)

摘　要: 数据预处理将原始的真实数据库转换成适于数据挖掘的挖掘数据库, 为挖掘算法更好的实现以及挖掘结果形象的显示打下了良好的基础。针对结构化数据讨论了数据预处理的两个目标: 消除现实数据库中的数据缺陷; 为数据挖掘做准备。并在此基础上, 介绍了数据挖掘软件 KDD 中数据预处理技术的实现。

关键词: 数据预处理; 数据分析; KDD(Knowledge Discover in Database)
中图法分类号: TP391　　**文献标识码:** A　　**文章编号:** 1001-3695(2004)07-0117-02

Research on Data Preprocess in Data Mining and Its Application

JIAN Zhi-gang, JIN Xu
(Dept. of Computer Science & Engineering, Beijing University of Science & Technology, Beijing 100083, China)

Abstract: In data mining, data preprocess converts the real database to the mining database. So the mining algorithms can run effectively and the mining results can get a better display. Aim at structural data, discusses two targets of the data preprocess. One is to eliminate the defects in real database. The other is to make prepare for the mining process. On this bases, we introduce its application in the KDD, a software of data mining.

Key words: Data Preprocess; Data Analysis; Data Mining; KDD

数据挖掘整体过程中, 原始数据库中的数据从现实中提取而来, 存在着各种各样现实中不可避免的缺陷。海量数据 GB 乃至 TB, 使得运行时间成为需要考虑的问题; 不同数据表中对相同属性的不同命名, 在表面上切断了数据之间联系; 数据表中总会有大量的空缺值, 甚至是错误的记录。这些问题形成了原始数据库与数据挖掘所需要的挖掘数据库之间一道鸿沟[1]。即使这些问题在一定程度上得以解决, 考虑到挖掘算法的有效性和运行时间的问题, 还需要对数据库中的数据做一定的处理。以上, 从原始数据库到挖掘数据库之间, 对数据进行的操作称为数据预处理。数据预处理一般分为四个步骤: 数据选取、数据表属性一致化、数据清理、数据离散化(数据归约) 。其中, 前三个步骤解决原始数据库中表面存在的问题, 已经有了相应的多种方法和技术[2,3]; 第四个步骤涉及到原始数据库中数据的内涵, 对下一步的挖掘工作起着决定性作用, 一般采用具有一定智能化的处理方法[4], 而为了避免挖掘出类似 "圣经密码"[5] 的无效知识, 领域专家的参与在该步骤是必不可少的[6]。

1　数据选取

数据选取是从用户的原始数据库中由用户指定选出用户感兴趣的、与知识发现任务相关的数据表项。用户在选择过程中可以通过查看所选数据表的记录数据, 来作出进一步的选择判断。通常用户都是对数据库中的数据包含的某个主题感兴趣, 希望通过数据挖掘工具对相关数据的操作来发现该主题下一些隐含的规律, 从而对所从事的行业行为有所指导。而数据库中的数据数量巨大, 涵盖范围也相对比较广泛。有些数据表格中的数据根本上是没有联系的。如果不对数据库进行简单筛选, 则会使无用数据参与挖掘过程, 造成各种资源上的浪费。更为严重的问题是, 由于一般挖掘算法仅对抽象的数据进行操作, 即使完全不相关的数据也会 "挖掘 "出 "规律 "。这种规律可以说毫无实际意义, 仅是数据海量造成的结果。

数据库操作人员对数据库中的数据有充分的了解, 由他们来选择待挖掘数据是很适合的。但是, 考虑到数据量的巨大, 如果完全由人来进行选取是不现实的。一般我们采取人机结合的方式。由人来选择较高概念层次上的数据类别, 而通过预先编制好的程序来选择数据库中具体的数据表格。如果数据挖掘在数据仓库的基础上进行, 那么操作起来会方便一些; 如果没有建立数据仓库, 在数据表选取的时候会遇到所谓 "实体识别 "[3] 问题, 即同一实体在不同数据表中由不同的属性来表示, 通常我们可以通过元数据的查询来解决这一问题。实体识别问题在数据表属性一致化中将得到根本解决。

2　数据表属性一致化

当待挖掘的数据表已经选取完毕时, 我们开始对这些数据表中的数据进行挖掘前的预处理。首先, 在数据表的属性这一层次上进行统一。主要解决上边提到的实体识别问题。具体来说, 一个在校学生数据库中, 学生成绩在一个数据表中可能记为 "学生成绩 ", 而在另一个数据表中可能以拼音来描述: "xscj "。如果不是数据库操作人员, 是不会了解其中联系的。作为挖掘前的准备, 需要根据数据字典对同一实体的不同命名

收稿日期: 2003-08-03; 修返日期: 2003-09-26
基金项目: 国家自然科学基金重点项目(69835001) ; 国家教育部科技重点项目(教技司[2000] 175)

表示来进行一致化,得到一个统一的、清晰的数据表示。具体实现方法可以以其中的某一个表示方式为准,更改其他的表示方式,或者重定义一个表示。需要注意的是,有时候同一属性的属性值有可能采用不同的度量单位,如学生成绩一般用百分制来表示,但也有时会采用五分制,或者“优、良、及格、差”等模糊的评判标准,我们可以根据需要来确定一个标准,并且规定一个转换方式,将非标准表示转换为标准表示。所有的更改需要记录下来,已备将来查阅或者数据更新时需要。

3 数据清理

前面两个步骤完成之后,我们认为挖掘数据库的框架和规格已经确定。下面将对其中的数据进行具体处理,主要解决的问题有:空缺值、错误数据、孤立点、噪声。其中空缺值和错误数据是这一步骤处理的重点,而后两者因为有可能在其中发现某些特殊规律,所以可以暂时不进行处理。

(1) 处理空缺值。可采取以下几种方法:①忽略,当一个元组的多个属性值空缺时,通常忽略它,即在数据表格中删除;②填补,当元组仅有少数属性值缺少,一般要对空缺值进行填补。填补有多种方式,人工填补、全局常量和所属属性下的平均值。还可以对该属性下的数据应用推导工具(判定树等),通过对其他数值的分析来得到最可能的填充值。

对于不同属性下的空缺值,我们需要不同的处理方法。通常认为应用推导工具分析出来的数值更加可靠和有实用价值。

(2) 处理错误数据。首先要能分辨出带有错误数据的元组,然后决定是更改数据还是忽略元组。通常在定义数据字典时,对数据有一个基本的规定。在这之上,现实世界中的事物都有其自身的约束,数据库中数据所系的实体亦然。譬如,学生考试分数是在 0 ~100 间的一个实数(其他的表示方式转换过来也应该满足这一要求)。这就是“学生成绩”属性下的一个约束,如果有哪一个元组的该属性下的值跳出这一范围,那么这是一个错误数据。当然并不是所有的约束都这么简单,但总可以找出一个函数来作为约束函数。这个函数有可能是属性自身相关的,也有可能是多属性相关的。

(3) 处理噪声数据。噪声数据,包括孤立点。对于一个变量的测量总会存在偏差,这些偏差就是噪声,如果偏差较大,就是孤立点。通常处理偏差的技术称为平滑技术。具体有以下几个方法:①分箱(Binning),即将数据平均分入几个箱中,对每个箱子里的数值进行转换,可以转换为箱中所有数值的平均值、中值或者边界值。转换后,数值的变化范围就相应缩小了。事实上,这是数据离散化的一种方式。②聚类(Clustering),聚类消除了噪声,同时可以发现孤立点,聚类分析有相应专门的技术,这里不赘述。③回归(Regression),线性回归和多线性回归分析可以应用到噪声的消除中。④人机结合,在数据挖掘整体过程中人机结合都是十分必要的,通常只是利用人工设置阈值的方式来辅助计算机识别孤立点。

4 数据离散化(数据归约)

数据清理工作完成后,由于海量数据等问题,需要进一步根据数据特征进行相应处理。这一步骤,可以看作是数据形式变换和数据压缩。

(1) 聚类、平滑、概念提升。这些都可以看作是数据离散化的某种具体方式。规范化是指将数据映射到某一个较小的特定区间,一般是 0.0 ~1.0 区间。这一变换适应于不同的挖掘算法,如果预先可以确定挖掘算法,那么这一工作是可以在数据选取时进行。注意要对规范化用到的映射函数存档。

(2) 数据归约。主要是通过变换数据的表示形式来得到可以保持原有数据完整性的相对较小的数据集,从而使数据挖掘变得可行。不同的数据集可以有不同的规约方式。如果建立了数据立方体,可以采用数据立方体上的聚集。对于数据集中包含属性较多的数据表格可以采用维归约的方式,即删除不相关的属性。那么如何选取有用数据属性是一个重要的问题,通常可以利用贪心算法或信息增益度量建立分类判定树。这些方法通常会占用一定的运行时间,我们认为要在算法的基础上加入人工指导、提高效率、缩短运行时间。如果数据集中的表格以稀疏矩阵为主,那么可以采用以下两种流行的数据压缩方法:小波变换(DWT)和主要成分分析(PCA)。数值归约通过选择替代的、较小的数据表示形式来减少数据量。数值归约技术可以是有参的,也可以是无参的。有参方法是使用一个模型来评估数据,只需存放参数,而不需要存放实际数据。

(3) 数据离散化。将连续的属性值划分为离散的几个区间,离散的属性值化分为不同的几个取值范围,从而减少属性值的数量,提高属性值的内涵,方便数据挖掘的过程以及挖掘结果的可视化展示。有许多具体的离散化方法:自然划分、分箱、直方图、聚类分析、基于数理统计、基于熵的离散化、人机结合等。自然划分易于实现,便于用户理解,可方便地应用于一些日常数据:年龄、收入等。分箱和直方图在平滑数据的过程中实现了数值的离散化。聚类分析将数据值分成离散的群,作为基本单位参与挖掘。通过数理统计得到的数值离散化结果相对更加可靠,适合分析相对抽象的数据值。基于熵的离散化用来递归的划分数值属性的值,产生分层的离散化,达到压缩数据量的效果。由于数据属性间较大的差异性,不同的属性下的数值需要用不同的离散化方法,而且许多离散化方法也需要认为设定一些参数,因此领域专家的参与十分重要和必要。

5 KDD 软件系统中数据预处理的实现

笔者参与开发的 KDD 软件系统是基于双库协同机制[6] 的数据挖掘软件,其数据预处理模块实现了上述四个步骤,并在数据离散化步骤中引入了语言值与语言场理论[7],为数据挖掘总双库协同机制的运行打下了基础。其模块如图 1 所示。

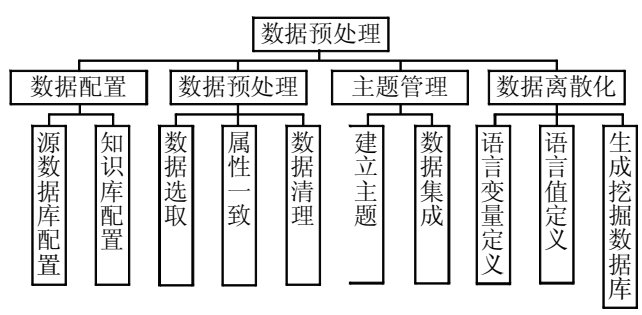

图 1 KDD 软件系统数据预处理功能模块示意图

其中,数据预处理和数据离散化子模块主要用来完成上述讨论的四个步骤。在数据选取、属性一致化、数据清理中集成了相应技术;语言值定义时集成了数据离散化中的自然划分、直方图、聚类分析、基于数理统计等方法,并充分考虑了领域专家在数据内容上的先验知识,给出了良好的 (下转第 157 页)

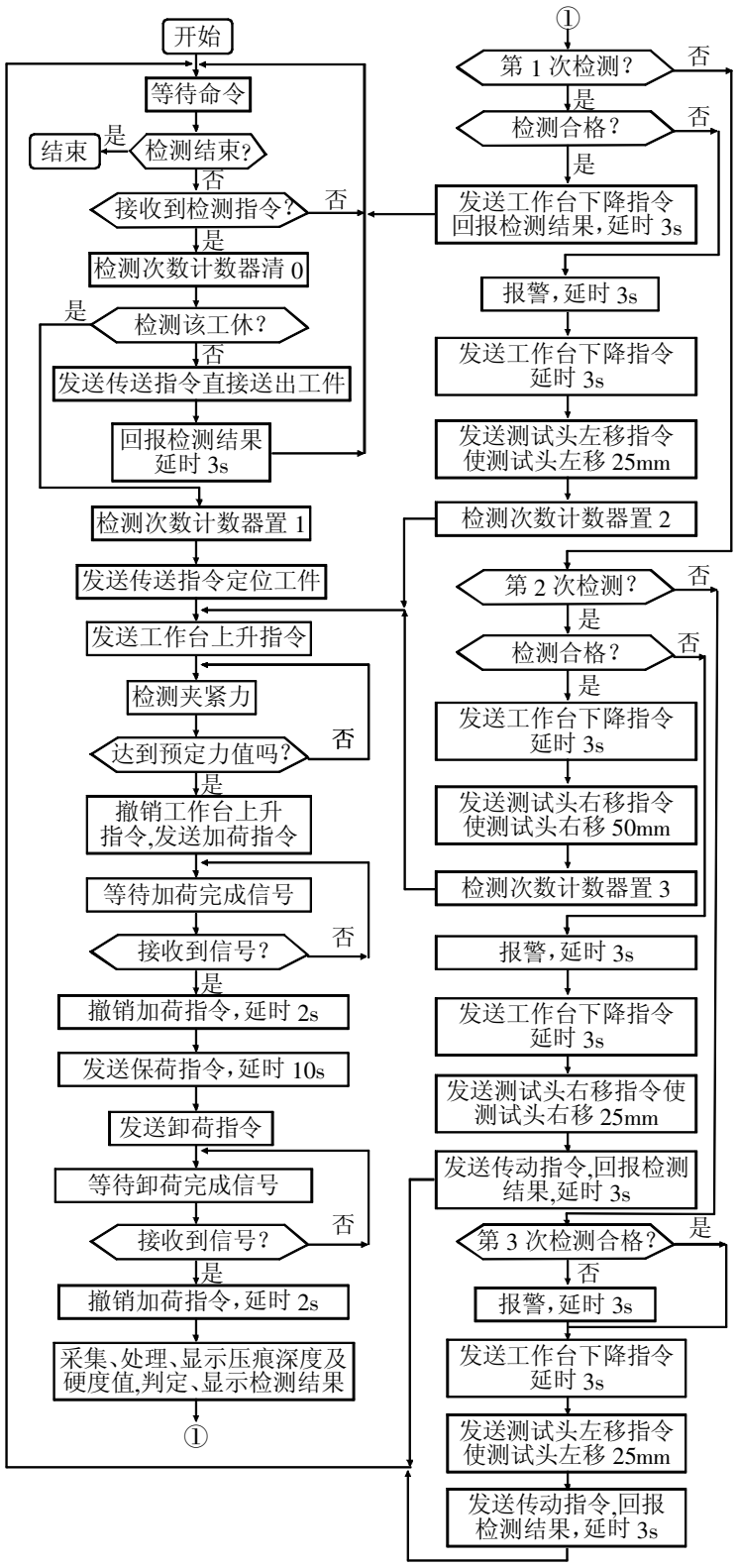

图 4　在线检测模块程序流程图

5　问题及措施

(1) 该系统需要在较恶劣的工作环境下运行, 必须考虑并解决好系统的抗干扰问题。在现场调试过程中, 我们发现由于继电器等外部电气的干扰, 夹紧力传感器和位移传感器的电压输入信号及 A/D 转换后的数字量不够稳定。为了解决问题, 我们通过在采集卡的接线端增加适当的电容等方法, 取得了满意的效果。

(2) 由于该系统需要长期连续不断地运行, 为切实解决好系统的稳定性及可靠性问题, 我们主要采取了以下措施: ①在系统硬件设备选择方面, 不过于追求性能价格比, 一定要求可靠性要高, 为此我们选择了研华公司的系列产品; ②在系统软件设计方面, 我们充分考虑到了一些硬件设备突然出现故障或电源中断时的预防措施, 以确保系统的安全性。系统软件还具有故障自诊断功能。

(3) 为了提供良好的用户界面, 我们除了采用菜单驱动及分栏显示方式之外, 还采用了直观的图形及动画提示, 如工作台上升与下降的 Flash 动画、开关量输入/输出信号指示灯等。

6　结束语

本文介绍了我们开发的“在线检测自动布氏硬度计计算机与控制系统”的主要技术路线及实现方法, 以便能为从事此类项目开发与研究的同行提供一些有益的借鉴。本系统已经在马鞍山钢铁公司的车轮产品检测线正式投入使用, 系统结构设计合理、抗干扰能力强、运行稳定可靠、功能齐全、操作简单, 深受用户青睐, 具有广泛的推广应用价值。

参考文献:

[1] 梁恩主. Visual Basic 6.0 编程与实例解析[M]. 北京: 科学出版社, 2000.

[2] 高金源, 等. 计算机控制系统——理论、设计与实现[M]. 北京: 北京航空航天大学出版社, 2001.

[3] 王春香, 翁新华, 杨汝清, 等. 基于 VB 的远程监控系统设计[J]. 计算机应用研究, 2002, 19(9): 110-111, 157.

作者简介:

石奋苏(1958-), 男, 副研究员, 主要从事计算机应用软件、实时管理与监控系统的开发研究。

(上接第 118 页) 人机交互接口。

KDD 软件系统已在多个项目中对真实数据库(庐江病虫害情况的数据库等) 进行挖掘, 得到了较好的挖掘结果, 数据预处理在其中起着十分重要的作用。

6　结论

本文介绍了数据挖掘过程中数据预处理的方法和技术, 并介绍了数据挖掘软件 KDD 在这方面的研究与实现。根据统计, 在一个完整的数据挖掘过程中, 数据预处理要花费 60% 左右的时间, 而后的挖掘工作仅占总工作量的 10% 左右[3]。因此, 不能简单地将数据挖掘中数据预处理过程孤立出来, 而应该从整体结构上考虑数据预处理过程。相信数据预处理将是数据挖掘研究与应用的热点问题之一。

参考文献:

[1] 杨炳儒. 知识工程与知识发现[M]. 北京: 冶金工业出版社, 2000. 584-586.

[2] A Famili, et al. Evangelos Simoudis. Data Preprocessing and Intelligent Data Analysis[J]. Intelligent Data Analysis, 1997, (1): 3-23.

[3] iawei Han, Micheline Kamber. Data Mining: Concepts and Techniques[M]. USA: Morgan Kaufmann Publishers, 2001. 70-95.

[4] avid J Hand. Intelligent Data Analysis: Issues and Opportunities[J]. Intelligent Data Analysis, 1998, (2): 67-69.

[5] McKay B Bar-Natan, D Bar-Hillel M, Kalai G. Solving the Bible Code Puzzle[J]. Statist. Sci, 1999, 14: 150-173.

[6] Yang Bingru, Shen Jiangtao. KDD Based on Double-base Cooperating Mechanism and Its Realization of Software[J]. Journal of System Engineering and Electronics, 1999, 10(4): 1-9.

[7] Yang Bingru. A Type of Language Field Integrated Algorithm Used for Analysis and Control of Complicated System[J]. Journal of System Engineering and Electronics, 1998, 9(1): 66-76.

作者简介:

菅志刚(1979-), 男, 河北人, 硕士研究生, 研究方向为结构化数据挖掘、Web 文本挖掘; 金旭(1979-), 女, 河北人, 硕士研究生, 研究方向为自然语言理解、Web 文本挖掘。