

数据挖掘

数据库

数据预处理

知识发现

(15)

54-57

数据挖掘中的数据预处理*)

Data Preprocessing in Data Mining

刘明吉 王秀峰 黄亚楼

TP311.13

(南开大学计算机与系统科学系 天津 300071)

Abstract Data Mining (DM) is a new hot research point in database area. Because the real-world data is not ideal, it is necessary to do some data preprocessing to meet the requirement of DM algorithms. In this paper, we discuss the procedure of data preprocessing and present the work of data preprocessing in details. We also discuss the methods and technologies used in data preprocessing.

Keywords Data mining, Tuple, Attribute, Knowledge-base, Rough-set, Genetic algorithm

1 引言

数据挖掘(Data Mining,简称DM),也称为数据库中的知识发现 KDD (Knowledge Discovery in Database),是近几年来随着数据库和人工智能发展起来的一门新兴的数据库技术。其处理对象是大量的日常业务数据,目的是为了从这些数据中抽取一些有价值的知识或信息。原始业务数据是知识和信息提取的源泉,对于数据挖掘就显得十分重要。目前所进行的关于数据挖掘的研究工作,大多着眼于数据挖掘算法的探讨,而忽视了对数据处理的研究^[5]。目前一些比较成

熟的算法对其处理的数据集合一般都有一定的要求,比如数据完整性好、数据的冗余性少、属性之间的相关性小。然而,实际系统中的数据一般都具有不完全性、冗余性和模糊性,很少能直接满足 DM 算法的要求。另外,海量的实际数据中无意义的成分很多,严重影响了 DM 算法的执行效率;而且由于其中的噪音干扰还会造成挖掘结果的偏差。因此,如何对不理想的原始数据进行有效的归纳和预处理,已经成为 DM 系统实现过程中的关键问题。

数据挖掘过程一般包括数据采集、数据预处理、数据开采以及知识评价和呈现,其过程可以用图1表示。

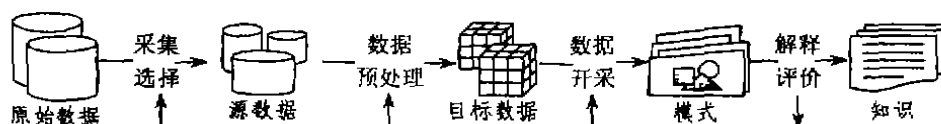


图1 数据挖掘过程图

数据预处理是 DM 的重要一环,而且必不可少。要使挖掘内核更有效地挖掘出知识,就必须为它提供干净、准确、简洁的数据。然而实际应用系统中收集到的原始数据是“脏”的^[2,4],通常存在以下几方面的问题:

·**杂乱性**。原始数据是从各个实际应用系统中获取的(多种数据库、多种文件系统),由于各应用系统的数

据缺乏统一标准和定义,数据结构也有较大的差异,因此各系统间的数据存在较大的不一致性,共享问题严重,往往不能直接拿来使用。

·**重复性**。是指对于同一个客观事物在数据库中存在其两个或两个以上完全相同的物理描述。由于应用系统实际使用中存在的一些问题,几乎所有应用系统中都存在数据的重复和信息的冗余现象。

*) 本论文的工作得到天津市自然科学基金项目“大型数据库的数据挖掘技术研究(983600411)”资助。刘明吉 博士研究生,研究领域为数据仓库和数据挖掘;王秀峰 博导,研究领域为遗传算法、计算智能技术;黄亚楼 博导,研究领域为机器人和数据挖掘。

·**不完整性**。由于实际系统设计时存在的缺陷以及一些使用过程中人为因素所造成的影响,数据记录中可能会出现有些数据属性的值丢失或不确定的情况,还可能缺少必需的数据而造成数据不完整。实际使用的系统中,存在大量的模糊信息,有些数据甚至还具有一定的随机性质。

一个完整的数据挖掘系统必须包含数据预处理模块。它以发现任务作为目标,以领域知识作为指导,用全新的“业务模型”来组织原来的业务数据,摒弃一些与挖掘目标不相关的属性,为数据挖掘内核算法提供干净、准确、更有针对性的数据,从而减少挖掘内核的数据处理量,提高了挖掘效率,提高了知识发现的起点和知识的准确度。

2 预处理的基本功能

数据挖掘中的预处理主要是接受并理解用户的发现要求,确定发现任务,抽取与发现任务相关的知识源,根据背景知识中的约束性规则对数据进行合法性检查,通过清理和归约等操作,生成供挖掘核心使用的目标数据,即知识基。知识基是原始数据库经数据汇集处理后得到的二维表,纵向为属性(Attributes 或 Fields)、横向为元组(Tuples 或 Records)。它汇集了原始数据库中所有数据的总体特征,是知识发现状态空间的基底,也可以认为是最初始的知识模板。数据预处理应该包括以下几方面的功能:

2.1 数据集成(Data Integration)

数据集成主要是将多文件或多数据库运行环境中的异构数据进行合并处理,解决语义的模糊性。该部分主要涉及数据的选择、数据的冲突问题以及不一致数据的处理问题。

用于进行知识发现的数据可能来自多个实际系统,因而存在着异构数据的转换问题。另外,多个数据源的数据之间还存在许多不一致的地方,如命名、结构、单位、含义等。因此,数据集成并非是简单的数据合并,而是把数据进行统一化和规范化处理的复杂过程。它需要统一原始数据中的所有矛盾之处,如字段的同名异义、异名同义、单位不统一、字长不一致等,从而把原始数据在最低层次上加以转换、提炼和聚集,形成最初始的知识发现状态空间。

另外,在数据集成中还应考虑数据类型的选择问题,应尽量选择占物理空间较小的数据类型。如在值域范围内用 tinyint 替代 int,每条记录可以节省3个字节。如果对于大规模数据集来说将会大大减少系统开销。

2.2 数据清理(Data Cleaning)

数据清理要去除源数据集中的噪声数据和无关数据,处理遗漏数据和清洗脏数据,去除空白数据域和知识背景上的白噪声,考虑时间顺序和数据变化等。主要

包括重复数据处理和缺值数据处理,并完成一些数据类型的转换。

数据清理可以分为有监督和无监督两类。有监督过程是在领域专家的指导下,分析收集的数据,去除明显错误的噪音数据和重复记录,填补缺值数据;无监督过程是用样本数据训练算法,使其获得一定的经验,并在以后的处理过程中自动采用这些经验完成数据清理工作。

数据清理的另一个重要内容是数据类型的转换,通常是指连续属性的离散化。一般来说,与类别无关的离散化方法有等距区间法、等频区间法和最大熵法。与类别有关的方法有划分法(splitting)和归并法(merging)等。通过离散化,可以有效地减少数据表的大小,提高分类的准确性。

2.3 数据变换(Data Transformation)

数据变换主要是找到数据的特征表示,用维变换或转换方法减少有效变量的数目或找到数据的不变式,包括规格化、归约、切换、旋转和投影等操作。

规格化指将元组集按规格化条件进行合并,也就是属性值量纲的归一化处理。规格化条件定义了属性的多个取值到给定虚拟值的对应关系。对于不同的数值属性特点,一般可以分为取值连续和取值分散的数值属性规格化问题;归约指将元组按语义层次结构合并。语义层次结构定义了元组属性值之间的 IS-A 语义关系。规格化和归约能大量减少元组个数,提高计算效率。同时,规格化和归约过程提高了知识发现的起点,使得一个算法能够发现多层次的知识,适应不同应用的需要。

我们还可以用多维数据立方(data cube)来组织数据,采用数据仓库中的切换、旋转和投影技术,把初始的知识状态空间按照不同的层次、粒度和维度进行抽象和聚集(即数据泛化),从而生成在不同抽象级别上的知识基。

2.4 数据简化(Data Reduction)

有些数据属性对发现任务是没有影响的,这些属性的加入会大大影响挖掘效率,甚至还可能导致挖掘结果的偏差。因此,有效地缩减数据是很有必要的。数据简化是在对发现任务和数据本身内容理解的基础上,寻找依赖于发现目标的表达数据的有用特征,以缩减数据规模,从而在尽可能保持数据原貌的前提下最大限度地精简数据量。它主要有两个途径:属性选择和数据抽样,分别针对数据库中的属性和记录。

属性选择包括针对属性进行剪枝、并枝、找方程和找相关等操作。剪枝就是去除对发现任务没有贡献或贡献率极低的属性域;并枝就是对属性进行主成分分析,把相近的属性进行综合归并处理;找方程就是发现两个或多个数值表示的属性之间的函数关系;找相关,即因子分析,在取值无序且离散的属性之间寻找依赖

关系,确定某个特定属性对其它属性依赖的强弱并进行比较。通过属性选择能够有效地减少属性,降低知识状态空间的维数。

数据抽样就是进行数据记录之间的相关性分析,用少量的记录基底的线性组合来表示大量的记录。它主要利用统计学中的抽样方法,如简单随机抽样、等距抽样、分层抽样等,具体进行统计运算,对于相同元组进行归并,并增加必要的支持度属性域。最简单的支持度属性域就是相同元组的数目,或者占总元组的百分比,也可以是置信度。最后去除那些支持度较低的元组(可视为例外或噪声)。

3 预处理的主要方法

预处理方法就是从大量的数据属性中提取出一部分对目标输出有重要影响的属性,即降低原始数据的维数,从而达到改善实例数据质量和提高数据挖掘速度的目的。换言之,我们要从 n 个输入变量 $x_1, x_2, \dots, x_n$ 的输入数据集合中,找出对于挖掘目标最为重要的 k 个输入变量($k < n$)。预处理方法可以分为以下几类:

(1) 基于粗糙集(RS)理论的约简方法

粗糙集理论是一种研究不精确、不确定性知识的数学工具,目前受到了KDD研究者的广泛重视,利用RS理论对数据进行处理是一种十分有效的精简数据维数的方法^[10,13~15]。我们所处理的数据一般存在信息的含糊性(vagueness)问题,含糊性有三种,术语的模糊性,如高矮;数据的不确定性,如噪声引起的;知识自身的不确定性,如规则的前后件间的依赖关系并不是完全可靠的。在KDD中,对不确定数据和噪声干扰的处理是粗糙集方法的特长^[3]。

RS理论的最大特点是无需提供问题所需处理的数据集合之外的任何先验信息,其基本思路是利用定义在数据集合 U 上的等价关系对 U 进行划分。对于数据表来说,这种等价关系可以是某个属性,或者是几个属性的集合。因此,按照不同属性的组合就把数据表划分成不同的基本类,在这些基本类的基础上进一步求得最小约简集。

采用RS理论作为数据预处理方法具有许多的优点:不需要预先知道额外信息;算法简单、易于操作。应用RS的属性约简可以有效地去除冗余的属性。另外,对于每个属性的值域出现的冗余现象,同样可以应用RS方法中的约简技术删除某些属性的多余值,从而使条件属性的个数和取值得到约简。但是,RS理论只能处理离散型属性,对于连续的属性必须先进行离散化才能再运用RS理论进行处理。

(2) 基于概念树的数据浓缩方法

在数据库中,许多属性都是可以进行数据归类,各属性值和概念依据抽象程度不同可以构成一个层次结构,概念的这种层次结构通常称为概念树^[12],概念树

一般由领域专家提供,它将各个层次的概念按一般到特殊的顺序排列。

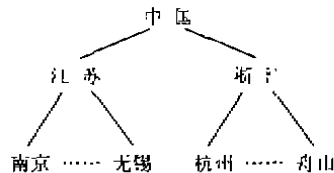


图2 “地区”概念树

基于概念树的数据预处理方法是一种归纳方法,其实是数据库中元组合并的处理过程,其基本思路如下:首先,一个属性的较具体的值被该属性的概念树中的父概念所代替,然后对相同元组进行合并,构成更宏观的元组,并计算宏元组所覆盖的元组数目(权重, T-weight),如果生成的宏元组数目仍然很大,那么用该属性的概念树中父概念去替代或者根据另一个属性进行概念树的提升操作,最后生成覆盖面更广、数量更少的宏元组。

在实际应用中,我们还可以根据经验和需要指定剪枝阈值(T-threshold),对那些偏离常规的、分散的极少量的宏元组($T\text{-weight} < T\text{-threshold}$)进行剪除,以消除噪声数据的不利影响,利用概念树提升的方法可以大大浓缩数据库的记录。基于概念树的数据浓缩实际上是一种概念泛化处理,可以使处理的数据体现不同的层次和汇聚密度,有利于关联规则等挖掘算法发现不同层次属性值之间的关系。

(3) 信息论思想和普化知识发现

特征知识和分类知识是普化知识的两种主要形式,其算法基本上可以分为两类:数据立方方法和面向属性归约方法。

普通的基于面向属性归约方法在归约属性的选择上有一定的盲目性。在归约过程中,当供选择的可归约属性有多个时,通常是随机选取一个进行归约,事实上,不同的属性归约次序获得的结果知识可能是不同的,根据信息论最大熵的概念,应该选用一个信息丢失最小的归约次序。

最大熵概念已经成功地运用在科学和工程方面。对于数据中的多个属性,通过比较各个属性归约后所得数据子空间熵的大小来确定最优的归约次序,归约属性的选择仅与相应数据子空间熵的相对大小有关,只需比较和估算相结合的方法,无需计算其准确值,大大减少了计算量。

(4) 基于统计分析的属性选取方法

我们可以采用统计分析中的一些算法来进行特征属性的选取^[2],比如主成分分析、逐步回归分析、公共因素模型分析等,这些方法的共同特征是用少量的特征元组去描述高维的原始知识基。

主成分分析的基本思想是:对于给定的输入数据

矩阵 X , 计算其相关系数矩阵 $R = X^T X$, 取与 R 中最大的几个特征值相应的特征向量作为主成分。其中数学准则是希望每次取得一个综合变量的方差, 在原变量的全部方差 (或剩下的全部方差) 中所占的比例最大。主成分方法的特点是将描述某一事物的多个变量压缩成描述该事物的少数几个综合变量或称主成分 (通常用原变量的线性组合表示), 旨在用新的少数几个综合变量代替原始变量, 并使这种替代所蒙受的损失最少。主成分分析法具有方差最优性、信息损失最小性、相关最优性和回归最优性, 使它得以成为多元降维的重要工具之一。

(5) 遗传算法 (Genetic Algorithms, GA)

遗传算法是一种基于生物进化论和分子遗传学的全局随机搜索算法^[9,11]。遗传算法的基本思想是将问题的可能解按某种形式进行编码, 形成染色体。随机选取 N 个染色体构成初始种群, 再根据预定的评价函数对每个染色体计算适应值, 选择适应值高的染色体进行复制, 通过遗传运算 (选择、交叉、变异) 来产生一群新的更适应环境的染色体, 形成新的种群。这样一代一代不断繁殖进化, 最后收敛到一个最适应环境的个体上, 从而求得问题的最优解。遗传算法应用的关键是适应度函数的建立和染色体的描述。在实际应用中通常将它和神经网络方法综合使用, 通过遗传算法来搜寻出更重要的变量组合。

遗传算法的高效搜索能力可以用来进行数据的聚类预处理。其基本思路是: 把一条具有 n 个属性的记录看作是 n 维空间中的一个点, 数据库中的数据记录就成为 n 维空间中的一组点群。这样对样本的聚类问题就转化为对点群的划分或归类问题。划分的原则是几何距离接近 (距离小于某设定阈值) 的点归为一类。每一类用参照点来标识, 参照点可以认为是代表该类的典型样本点, 即该类的中心。对于任意一个样本点, 它隶属的类由距离它最近的参照点来确定。这样对于样本集, 当给定一组参照点后每个样本都唯一地划分到其中一个类中 (距离两个参照点相等, 则随机选择一个参照点作为该样本的所属类), 即被聚类, 因此, 数据的聚类预处理就转变成寻找汇集中心的问题, 就转变为利用遗传算法搜寻全局 (样本空间) 最优解 (聚类中心) 的问题。这种方法仍局限于对数值型样本的处理, 非数值数据需要转换为数值型。

用遗传算法进行聚类预处理的最大的特点是聚类结果不依赖于初始聚类中心, 一般都能收到较好的效果。

结束语 数据预处理的方法很多, 主要目的是提高知识基的抽象概括程度, 缩小知识模板的物理尺寸,

为挖掘内核提供与发现任务有关的半成品数据集。如果数据挖掘建立在以主题来组织数据的数据仓库基础上, 则可以大大减少数据预处理的工作量。另外, 需要明确预处理并不是 DM 本身, 它应当是一个快速的数据处理过程, 如果过于强调预处理本身, 而忽略了它服务于 DM, 是不恰当的。

数据挖掘技术涉及领域广, 发展时间短, 理论体系不够成熟, 尤其是对数据预处理的研究还不够完善, 不过, 近年来, 对数据预处理的研究越来越引起广大学者的注意, 在 PAKDD'99 会议上, 香港科大的 Lu Hongjun 教授作了关于数据质量和有效数据挖掘的专题报告^[4], 引起了与会同行的广泛关注。相信今后对数据预处理的研究会越来越受到重视。

参考文献

- 1 Agrawal R, et al. Mining association rules between sets of items in large databases. In: Bunemuu P, Jajodia S, eds. Proc. of the 1993 ACM SIGMOD Conf. on Management of Data. New York, NY: ACM Press, 1993. 207~216
- 2 Fayyad U, et al. The KDD process for extracting useful knowledge from volumes of data. Communications of the ACM, 1996, 39(11): 27~35
- 3 Zaarko W, Shan N. KDD-R: A Comprehensive System for Knowledge Discovery in Databases Using Rough Sets. In: Lin T Y, Wildberger A M, eds. Soft Computing. San Diego, CA: Simulation Councils, Inc, 1995. 298~301
- 4 Lu Hongjun. Quality Data and Effective Mining. PAKDD-99
- 5 Lu Hongjun. On Preprocessing Data for Effective Classification. Department of Information Systems and Computer Science, NUS
- 6 Justin C. W. Feature Subset Selection within a Simulated Annealing Data Mining Algorithm. Journal of Intelligent Information Systems, 1997, 9: 57~81
- 7 张剑平, 任德官, 任福继. 大规模特征提取问题的算法研究. 清华大学学报, 1998, 38(7): 154~156
- 8 邹雯, 陈文伟. 数据开采中的遗传算法. 计算机世界报, 1997, 24期专题版
- 9 陈文伟, 等. 数据开采和数据仓库论文集. 信息与决策系统, 1998, 3(1): 24~29
- 10 王志海, 胡可云, 等. 基于粗糙集合理论的知识发现综述. 模式识别和人工智能, 1998, 11(2): 346~351
- 11 黄金才, 陈文伟. 遗传算法和模糊神经网络在数据挖掘中的应用. 清华大学学报, 1998, 38(7): 50~54
- 12 陈文伟. 智能决策技术. 电子工业出版社, 1998
- 13 王珏, 苗夺谦, 等. 关于 Rough Set 理论与应用的综述. 模式识别与人工智能, 1996, 9(4): 337~344
- 14 曾黄麟. 粗糙理论及其应用. 重庆: 重庆大学出版社, 1998
- 15 王珏, 王任, 等. 基于 Rough Set 理论的“数据浓缩”. 计算机学报, 1998, 21(5): 393~400