学校编码：10384　　分类号_____密级_____
学　　号：X2005223008　　UDC _____

厦门大学

硕 士 学 位 论 文

数据预处理之数据缩减方法应用实例研究

Research on Data Reduction Methods and Instances of Data Preprocessing

王国庆

指导教师姓名：　李　茂　青　教　授
专　业　名　称：　控制理论与控制工程
论文提交日期：　2010　年　4　月
论文答辩日期：　2010　年　　月
学位授予日期：　2010　年　　月

答辩委员会主席：_____
评　　阅　　人：_____

2010 年 4 月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下,独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外,该学位论文为(　　)课题(组)的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：

年　　月　　日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门大学保密委员会审查核定的保密学位论文，于　　年　　月　　日解密，解密后适用上述授权。

（　√　）2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：

年　　月　　日

摘　要

数据挖掘的处理对象是海量业务数据。在实际应用中，海量业务数据常常包含很多影响挖掘算法执行效率、干扰挖掘结果的因素，如：数据不完整、不一致、高冗余及噪音数据、脏数据等。数据预处理在数据挖掘之前，以领域知识为指导，以发现任务为目标，通过清理、集成、变换、缩减等操作，对源数据集进行处理，获得可供数据挖掘核心算法使用的目标数据，以减少数据处理量、提高挖掘效率，提高知识发现的起点和准确度。

数据预处理技术包括：数据清理(Data Cleaning)、数据集成(Data Integration)、数据变换(Data Transformation)与数据缩减(Data Reduction)。数据缩减是指在尽可能保持数据集原貌的情况下最大限度地精简数据量，在缩减后的数据集上进行分析与挖掘，将获得与在原有数据集上挖掘相同或近似相同的挖掘结果，而挖掘算法的执行效率更高。

典型的数据缩减方法包括：1）数据立方体聚合：通过在数据立方体上的聚合操作实现数据缩减；2）属性子集的选取：通过检测和消除不相关、弱相关或冗余的属性实现数据缩减；3）数据压缩：利用编码技术压缩数据集的大小；4）数值缩减：利用更简单的数据表达形式来取代原有的数据以实现数据缩减；5）离散化与数据概念分层：将属性的原始值用区间值或更高层次的概念来替换，缩小数据集的大小，同时概念分层可以帮助挖掘不同抽象层次的模式知识。

本文将对典型的数据缩减方法及其相关计算机技术算法的实现进行学习、归纳、总结与研究，包括：选取属性子集、离散化和数据概念分层、粗糙集算法等。文中阐述了这些典型方法的技术原理、使用特点及应用情况，并通过对训练集数据的应用对数据缩减方法进行分析与研究。

关键词：数据预处理；数据缩减；维数缩减；数据块缩减；离散化与概念分层

ABSTRACT

Data mining manipulates on service data of massive size. In practical use, service data of massive size always include many factors which influence mining algorithm efficiency and disturb mining result, such as incomplete, inconsistent, high redundant and noisy, dirty data. Data preprocessing applies before data mining and manipulates on source data set through cleaning, integration, transformation and reduction, thereby using field knowledge as guidance, aiming to finding tasks, helping to obtain target data for data mining key algorithm and reduce data manipulation capacity, improve efficiency of the mining and accuracy of the knowledge discovery.

Data preprocessing technique include data cleaning, data integration, data transformation and data reduction. Data reduction obtains a furthest reduced representation of the data set, yet closely maintains the integrity of the original data. Mining and analyzing on the reduced data set should be more efficient, yet produce the same (or almost the same) analytical results as mining on the original data.

Typical strategies for data reduction include data cube aggregation (where aggregation operations are applied to the data cube), attribute subset selection (where irrelevant, weakly relevant, or redundant attributes or dimensions may be detected and removed), data compression (where encoding mechanisms are used to reduce the data set size), numerosity reduction (where original data are replaced by much more simpler data expression form), generalization and discretization (where raw data values for attributes are replaced by ranges or higher conceptual levels to reduce the size of the data set. Concept hierarchies help mining pattern knowledge from multiple abstraction levels.) .

This paper focuses on these typical data reduction methods and implementation of correlative algorithms, thereby learning, concluding, summarizing and researching on attribute subset selection, discretization and concept hierarchy generation, rough set algorithm etc., describes the technical principles, application characters and status. Also, we do some experimental works on data reduction techniques through analysis and research on training data set.

Key words: data preprocessing, data reduction, dimension reduction, numerosity reduction, discretization, discretization and concept hierarchy generation

目 录

Contents

第一章 绪论

1.1 论文的研究背景及选题意义

近些年，数据挖掘成为学术界的研究热点，其应用领域迅速扩展并不断趋于广泛。数据挖掘在银行、保险、电信、商业等领域既有成功的典范，也有失败的案例。对失败案例的研究发现，很重要的一个失败因素就是忽略了数据挖掘的前期工作—数据预处理。

在数据挖掘的实际应用中，作为数据挖掘对象的海量业务数据常常包含很多影响挖掘算法执行效率、干扰挖掘结果的因素，如：数据不完整、不一致、高冗余及噪音数据、脏数据等。在数据挖掘开始之前，以领域知识为指导，以发现任务为目标，对源数据集进行预处理，获得可供数据挖掘核心算法使用的目标数据，可以减少数据处理量、提高挖掘效率，提高知识发现的起点和准确度。

数据预处理技术通常包括[1]：数据清理(Data Cleaning)、数据集成(Data Integration)、数据变换(Data Transformation)及数据缩减(Data Reduction)。据对数据挖掘成功案例的统计与分析，其数据预处理工作量一般占整个数据挖掘过程总工作量的60%以上。

数据清理(Data Cleaning)专门处理待挖掘数据集中的缺失数据、噪声数据、脏数据等。缺失数据指特征值没有被记录下来的数据（如：作为关键字的部门编码没有值或值已丢失）；噪声数据指数据中存在着错误或异常(偏离期望值)的数据（如：作为关键字的同一部门编码出现了不符合编码规则的值）；脏数据则指数据内涵已经过时的数据和数据内涵发生变化后相应产生的不一致数据（如：某一部门已撤销或与其他部门合并，作为关键字的部门编码未及时更新，部分记录仍延用旧值）。缺失数据、噪声数据、脏数据在现实世界中是非常普遍的情况。究其产生的原因主要有以下几个：数据库创建初期，一些数据被认为是不必要记录的，一些数据没有初始数值；信息员对信息与数据存在误解，编码规则发生变化等，导致信息员对数据的修改被忽略，或导致数据记录与其它记录内容不一致；数据录入过程和数据传输过程中发生了人为或计算机错误，导致一些数据没有被及时记录下来。

数据集成(Data Integration)将来自多个数据源的异构数据(如:来自 SQL SERVER 的数据库和 EXCEL 工作表数据)按照统一的格式进行合并处理，使来自多个数据源的现实世界实体能够相互匹配，消除数据值冲突等现象，形成比较完整的数据集合。进行数据集成时常常会引起数据再一次的不一致或冗余。例如：顾客编码的属性名在一个数据库中为“顾客编号”，在另一个数据库为“顾客号”；属性值“信息技术系”在另一个数据库中可能被简化为“信息系”。在数据集成之后，有时还需要再次进行数据清洗，以消除集成中出现的噪声数据、脏数据及数据冗余，以免影响挖掘速度、误导挖掘进程。

数据变换(Data Transformation) 对待挖掘数据进行规格化操作，找出数据的特征表示并转换或归并以构成一个适合挖掘的描述形式。如：使用神经网络算法或最近邻分类等算法进行数据挖掘时，必须将数据缩至特定的范围之内，如：[0，10]，以符合按算法的数据规格化要求。

数据缩减(Data Reduction)在尽可能保持数据集原貌、不影响(或基本不影响)最终挖掘结果的情况下最大限度精简数据量，缩小待挖掘数据集的规模；在缩减后的数据集上进行分析与挖掘，可以大幅度减少后续数据处理与数据分析所耗费的时间，并获得与在原有数据集上相同或近似相同的挖掘结果，而执行效率更高。典型的数据缩减方法包括：1）数据立方体聚合：通过在数据立方体上的聚合操作实现数据缩减；2）属性子集的选取：通过检测和消除不相关、弱相关或冗余的属性实现数据缩减；3）数据压缩：利用编码技术压缩数据集的大小；4）数值缩减：利用更简单的数据表达形式来取代原有的数据以实现数据缩减；5）离散化与数据概念分层：将属性的原始值用区间值或更高层次的概念来替换，缩小数据集的大小，同时概念分层还可以帮助挖掘不同抽象层次的模式知识。

这里需要特别指出：各种数据预处理方法并不是相互独立的，而是相互关联的。如：消除数据冗余既可以看成是一种形式的数据清洗，也可以认为是一种数据消减。数据预处理帮助改善数据的质量，进而帮助提高数据挖掘进程的有效性和准确性。数据预处理已经成为整个数据挖掘与知识发现过程中一个不可忽视的重要步骤。

尽管数据预处理工作量占整个数据挖掘过程总工作量的 60%，但是在数据挖掘过程中，人们对数据预处理的关注与相应的投入远不如对挖掘算法的关注和投

入大。本文分析并研究数据预处理之数据缩减方法，并结合对训练集数据的应用分析，向大家阐述一个事实，那就是：数据缩减技术可以最大限度地精简数据量，提高数据挖掘的执行速度与效率。开展数据预处理工作可以轻松地获得事半功倍的效果，是数据挖掘实现过程中值得重点关注并大力投入的关键环节。

1.2 论文研究的内容与组织

本文共分五章，组织结构如下：

第一章 绪论，介绍论文的大体结构与内容

第二章 介绍数据预处理的相关背景知识，及其常用分析方法。

第三章 介绍数据缩减的相关计算机技术算法，着重对数据缩减的相关计算机技术算法的实现进行介绍。

第四章 实验部分，应用多种方法展开实验，并对实验结果进行详细分析。

第五章 总结全文，并对今后进一步的研究工作提出一些设想。

第二章 数据预处理及其常用分析方法

2.1 数据预处理概述

数据挖掘 (Data Mining)是从存放在数据库、数据仓库或其他信息库中的大量的数据中获取有效的、新颖的、潜在有用的、最终可理解的模式的非平凡过程。不论数据挖掘的目标是分类、预测、聚类、关联分析或序列分析，其基本过程都可以划分为：问题定义、数据预处理、数据挖掘以及结果的分析与评估阶段。[2]

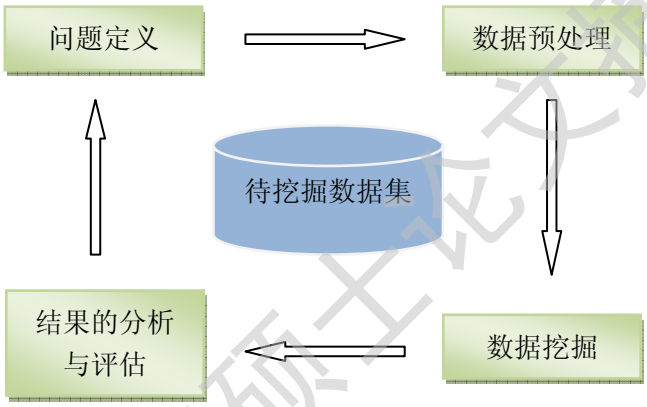

图 2-1 数据挖掘的基本过程

从这一过程可以看出，数据预处理是数据挖掘（知识发现）过程中一个十分重要的步骤。由于数据挖掘算法本身对数据有一定的要求，比如：数据相关性小、数据冗余性低等，而待挖掘的数据集直接来自于现实世界数据库或数据仓库，不可避免地具有不完整、含噪声或一致性较差等特点，常常表现为：(1)不完整性：数据属性值遗漏或不确定；(2)不一致性：由于原始数据的来源不同，数据定义缺乏统一标准，导致系统间的数据内涵不一致；(3)有噪声：数据中存在异常(偏离期望值)；(4)冗余性：数据记录或属性的重复。因此，在问题定义之后、数据挖掘开始之前，对待挖掘数据集进行数据预处理，将对接下来的数据挖掘产生直接的影响，包括：缩小待挖掘数据集的规模，改进挖掘算法的执行效率，获取更有价值、更有意义的挖掘结果。

数据预处理包括:数据清理(Data Cleaning)、数据集成(Data Integration)、数据变换(Data Transformation)、数据缩减(Data Reduction)。数据预处理的任务就是以领域知识为指导，用全新的“业务模型”来组织原有业务数据，清除

与挖掘目标无关的属性，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，提高数据挖掘效率，提高知识发现的起点和知识与规则的准确度。例如：一个负责公司销售数据分析的超级商场信息主管，在数据挖掘算法执行之前需要仔细检查公司数据库或数据仓库内容，精心挑选与挖掘任务相关数据对象的描述特征或数据仓库的维度，如：商品类型、价格、销售量等，精心检查数据库中数据记录的特征值有没有缺失，数据库中的数据记录有没有存在一些错误、不寻常、甚至是不一致情况。完成如上所述的一系列数据预处理工作，可以明显提高数据挖掘对象的质量，提高数据挖掘算法的执行效率，提高数据挖掘结果的准确度，并最终达到提高数据挖掘所获模式知识质量的目的。

2.2 数据预处理常用分析方法

数据预处理是一项繁杂的工程，常用分析方法有：数据清理(Data Cleaning)、数据集成(Data Integration)、数据变换(Data Transformation)、数据缩减(Data Reduction)等。

2.2.1 数据清理

现实世界的数据一般是不完整的、含噪声的和不一致的。数据清洗是指处理数据中的缺失数据、噪声数据和脏数据，主要包括：填补缺失的数据、消除数据中的噪声、剔除异常值以及纠正不一致数据等。

缺失数据、噪声数据、异常或不一致数据产生的原因很多，归纳起来主要有以下几个：(1)数据库创建初期，一些数据被信息员认为没有必要记录，还有一些数据在数据库创建初期并没有初始数值（如：销售事务数据中的顾客基本信息不全）；(2)信息员对信息与数据的编码及存储规则存在误解，导致数据存储有误；或者数据编码规则发生变化后，信息员忽略了对数据的更新与修改，导致数据记录内容不一致（如：销售事务数据中的部门设置状况发生变化后数据记录未能及时更新）；(3)数据录入过程或数据传输过程中发生了人为或计算机错误，导致一些数据没有被及时记录下来。

缺失数据的处理方法一般有：(1)忽略或删除该记录，即直接删除包含缺失数据的记录行。这种方法必须慎重采用，除非是无法确认、无法填补数据值的数

据记录，一般不能轻易删除属性值缺失的记录。(2)手工填补，即根据业务规则对包含缺失数据的数据记录行进行数据值填补。这种方法工作量大，耗时长，可操作性差。对于存在许多缺失情况的大规模数据集而言，有时难以实现或者根本无法实现。(3)根据已有数据值，采用默认值、平均值或者同类别平均值对缺失数据进行填补。这种方法有可能对数据挖掘产生误导。(4)通过回归分析、贝叶斯方法或决策树推断出该记录特定属性的最可能取值后，对缺失数据值进行填补。这类方法比较常用，与其他方法相比，它最大程度地利用现有的数据信息来推测缺失数据值，因而效果最好。

噪声数据是被测变量的随机错误或偏差，包括错误的值或偏离期望的孤立点。因此，可以用以下技术来平滑噪声数据，识别和删除孤立点：(1)分箱方法。将存储的值分布到一些箱中，通过考察“邻居”来局部平滑存储数据的值。可以采用按箱的平均值、中值或箱边界值进行平滑。(2)聚类。将类似的值组织成群或“聚类”，落在聚类集合之外的值被视为异常数据。对于异常数据，如果是垃圾数据，则予以清除，否则保留作为重要数据进行孤立点分析。(3)回归方法。利用拟合函数来平滑数据，帮助除去噪声。例如：线性回归、多元回归等。(4)人机结合检查方法。首先由计算机识别并输出那些差异程度大于某个阈值的数据，然后人工审核这些数据，确定孤立点。这种方法比单纯的人工检查要快。

现实世界中的数据库常常会有数据记录内容不一致的现象。出现数据不一致的原因很多，包括：属性命名不规范、数据更新不及时等等。不一致数据的处理一般通过数据与外部的关联手工处理，例如：数据录入中发生的错误或数据更新不及时造成的错误可以通过与原稿的校对来加以纠正。此外，可以使用一些例程或其他软件工具来帮助纠正使用编码时发生的错误。知识工程工具也可以帮助发现违反约束条件的数据。

2.2.2 数据集成

数据集成就是将多个数据源中的数据合并存放在一个同一的数据存储(如数据仓库、数据库等)的一种技术和过程，数据源可以是多个数据库、数据立方体或一般的数据文件。由于数据挖掘的处理对象是企业数据库和数据仓库，而在企业发展伴随着技术不断更新的过程中，企业数据库和数据仓库常常是多种不同类型数据库的组合（如：EXCEL 工作表、SQL SERVER 数据库、ORACLE 数据库）。因