数据挖掘中的数据预处理模型与算法研究

Study on the Model and Algorithms of Data Preprocessing in Data Mining

沈睿芳 郭立甫 （石家庄 河北经贸大学信息技术学院 050061）
时希杰 （天津大学管理学院 300072）

摘要：本文首先介绍了数据预处理的概念，然后将数据预处理无缝集成于数据仓库的构建过程中，提出了一种数据预处理过程模型。对于不同阶段所使用的数据预处理技术和算法，本文也进行了总结分析，并以粗糙集的属性约简为例给出了一种算法的具体实现步骤。

关键词：数据挖掘 数据预处理 模型 算法 属性约简

1 数据预处理简介

计算机以及存储技术的发展使得人们面对的信息越来越多，需要处理的信息量也呈指数形式增加。但是不可否认，并不是所有的信息对人们都是有用的。在这浩如烟海的信息中，可能存在着冗余甚至是错误的信息，严重影响着信息处理的效率和效果。如何在保证不失真的情况下对信息进行缩减、瘦身就成为信息技术领域一个非常具有现实意义的研究方向。这里提到的“不失真”是指在保证不丢失重要信息的情况下，去掉那些冗余的、次要的甚至是错误的信息，提高信息处理的效率和效果，这就是所谓的数据预处理（Data Preprocessing）问题。数据预处理可以分为两个方面，一个是 web 数据的预处理问题，另一个就是传统数据库中的数据预处理问题。web 数据格式的特点决定了前者有其单独的预处理技术和方法，本文重点讨论后者，即传统关系型数据库中的数据预处理问题。

数据预处理是数据挖掘应用中的重要工作，一般包括下列内容：

（1）数据清理：通过填写空缺值，平滑噪声数据，识别、删除孤立点，并解决“不一致”来“清理”数据；

（2）数据集成：将多个数据源合并成一致的数据存储，如将不同数据库中的数据集成入一个数据仓库中存储；

（3）数据变换：将数据转换成适合于挖掘的形式，如将属性数据按比例缩放，使之落入一个比较小的特定区间。这一点对那些基于距离的挖掘算法尤为重要；

（4）数据规约：在不影响挖掘结果的前提下，通过数值聚集、删除冗余特性的办法压缩数据，提高挖掘模式的质量，降低时间复杂度。

2 基于数据仓库的数据预处理过程模型

目前用于数据挖掘的数据主要还是以数值、字符型的结构化数据为主，并且大部分应用都是针对历史数据进行挖掘，因此作为存储结构化历史数据的主力军——数据仓库自然就成为目前数据挖掘应用的基础平台。

数据仓库的构建方法可分为三种：自顶向下、自底向上和混合方法。一般来讲，对于大部分信息化尚不充分的企业来说，自底向上方法应该是最合适的。其主要思想是：首先将分散的源数据按不同部门的业务功能（如市场部、研发部等）进行汇总，形成某个部门定制化使用的数据集市，然后按照主题从不同的角度进行整合、归并，最终形成数据仓库[1]。

事实上，数据预处理作为构建数据仓库过程中的一种基础性工作，完全可以融入数据仓库的构建过程，并将数据仓库的构建看作是数据挖掘的一个重要预处理步骤[2]，将两者有机合成，其过程模型如图 1 所示。

该图的核心思想可以概括为一条主线，两个过程，三个阶段。

（1）主线是数据流：原始数据—数据集市—数据仓库—待挖掘数据集。

（2）两个过程：上半部分为数据仓库的构建过程，下半部分为数据预处理过程。

（3）对应的三个阶段分别是清理阶段、集成阶段

和规约阶段。

清理阶段中，将原始数据按业务功能（如市场分析）进行汇总，形成部门级的数据集市。在此过程中需要处理一些原始数据中存在的问题，如空缺值、噪声数据干扰等。

集成阶段，将不同部门的数据集市按主题进行归并集成，形成企业级的数据仓库。数据集成时，由于各个数据集市存放数据的角度不同，因此在进入数据仓库系统中时有可能会产生冗余；而且由于它们在数据结构、数据编码和定义等方面的不一致性也会造成数据存在二义性等问题。因此，数据集成到数据仓库之后仍然需要进行去除冗余、解决不一致性等工作，甚至重新集成。至此，数据仓库构建完毕。

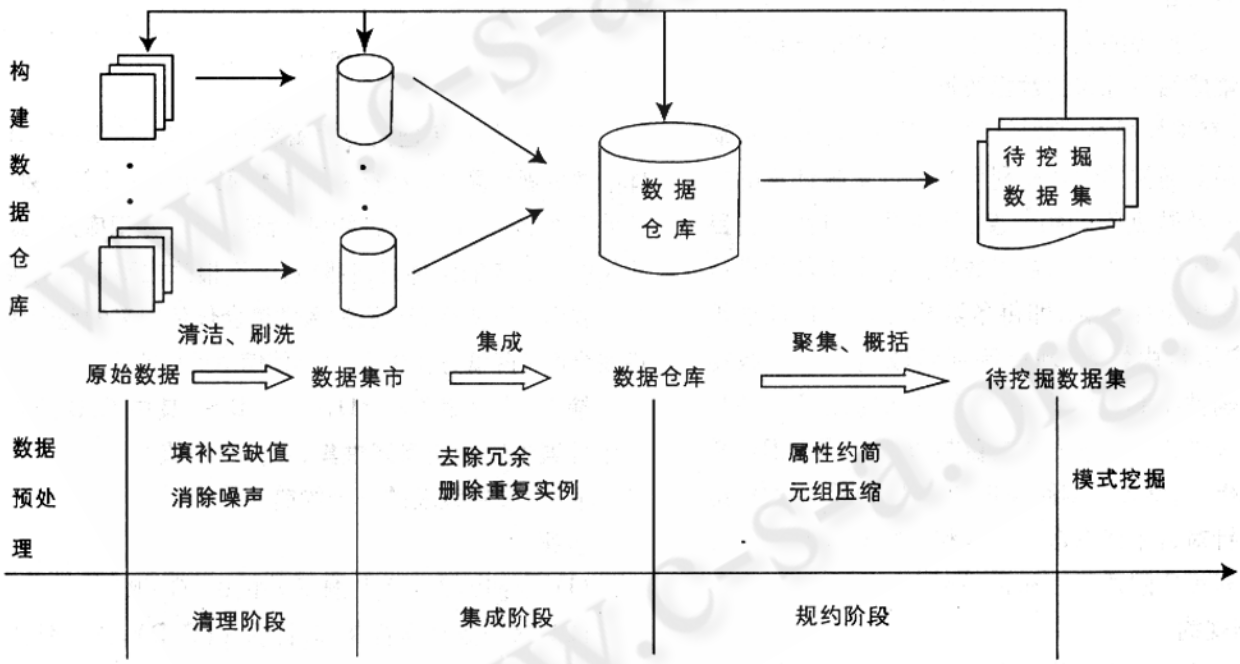

图 1　基于数据仓库的数据预处理过程模型

规约阶段则超出了构建数据仓库的范畴，而是为其后的数据挖掘做准备。数据仓库中的数据量过于庞大，必须采取"瘦身"策略对数据进行压缩、变换，但又不能脱离原始数据，必须把握住数据的"本质"。此阶段将用到数据压缩和属性约简技术。

从以上对过程模型的解释可以看到，数据预处理阶段并不局限于数据仓库的搭建过程，有可能在仓库建成以后，仍然需要对其中的规范化数据进行再处理，以满足特定数据挖掘任务的需要。而且数据预处理一次就达到期望效果的很少，更多的情况是上述各阶段要反复进行（如图中虚线箭头所示）。在构建数据仓库的过程中同时进行数据预处理工作，可以缩短数据挖掘应用的准备时间，也使构建过程各个阶段要做的预处理工作一目了然。将大量的数据预处理问题形成一种工作逻辑加以完成，不仅方便实施，而且得到的数据质量较高，为其后的挖掘工作打下了良好的基础。

3　数据预处理过程中的常用算法

3.1　数据清理

原始业务数据合成为数据集市阶段，要对数据进行"清洁"和"刷洗"，不仅要检查每个属性的存储格式，还要检验其实际内容是否符合规范，如处理空缺值，平滑噪声数据，识别、删除孤立点，删除某些重复记录、对属性值的有效性进行检验等。

在空缺值的处理上，目前最常用的方法是使用最可能的值填充空缺值。比如可以用回归、贝叶斯形式化方法工具或判定树归纳等确定空缺值。这类方法依靠现有的数据信息来推测空缺值，使空缺值有更大的机会保持与其他属性之间的联系。还有其他一些方法来处理空缺值，如用一个全局常量替换空缺值、使用属性的平均值填充空缺值或将所有元组按某些属性分类，然后用同一类中属性的平均值填充空缺值等。

在删除重复记录方面，目前重复记录识别算法中研究得最多的是基于距离的识别算法，如声音距离、编辑距离、输入距离等。它基于用户输入错误不可能太

多的事实，即在误差一定的情况下研究两个字符串是否等值。据此 Baeza 等提出了基于动态规划或过滤的算法[3]；李华旸等人则发展了动态规划算法，考虑到数据中的缩写情况，提出了缩写发现算法[4]；佘春红等人则提出了基于优先队列的增量式识别算法[5]。由于数据清理工作中的多样性和复杂性，这些算法都有各自严格的使用条件，并且清理效果也是千差万别，目前还不存在一种效果好、适用性强的算法。

3.2 数据集成

数据集市集成为数据仓库阶段中，要把不同数据集市的数据集成为一个紧密耦合的数据模型，结合为一个新的实体。这些数据来源往往遵守的不是同一套业务规则，在生成新数据时必须考虑到这一差异。数据集成中存在的一个主要困难是数据不一致，即同一数据在各个数据源的表现值不相同。这种不一致性严重影响了集成后目标库的数据质量。

目前，有关数据不一致性的理论研究较少，而且缺乏实践应用。通常的做法是利用相似性系数计算的方法求得不一致数据间的相似度，从而进行归并。但目前的相似性比较算法只适合对数据的单个特征项进行比较。如果信息多维化，即每条数据由多个特征项组成（如客户资料由姓名、地址等项组成），不一致数据归并的正确性不高。为此，张艳秋等人给出了一种加权组合的改进方法。该方法在目前单特征项的相似系数计算基础上，增加了对距离值的考虑，并利用加权组合的策略针对各相似系数加权求和。实验表明改进后的方法具有更广的适用性和正确率[6]。

3.3 数据规约

此过程可以看作是构建数据仓库的延续。数据规约技术可以用来得到数据集的规约表示，它接近于保持原数据的完整性，但数据量比原数据小得多。与非规约数据相比，在规约的数据上进行挖掘，所需的时间和内存资源更少，挖掘将更有效，并产生相同或几乎相同的分析结果。大多数数据仓库里的数据都要进行某种聚集和概括，将某一实体的记录数目压缩到易于驾驭的水平。例如商店把每日的销售额加在一起，生成按地区计算的月销售额（概括）；或者将不同业务元素加在一起成为一个公共总数（聚集）。聚集和概括还可以去除数据仓库中的过时细节，将过时数据以一种概括的形式存放，提高数据仓库的效率。此项工作有时也称为联机分析处理（OLAP）。

数据规约有两种，一种是数值规约，即通过选择替代的、较小的数据表示形式来减少数据量；另外一种是维规约，通过删除不相关的属性（或维）减少数据量，不仅压缩了数据集，还减少了出现在发现模式上的属性数目。如果将数据视为一张二维的表格，则数值规约就表示纵向数据量的压缩，而维规约就表示横向属性的压缩，如图 2 所示。

图 2

下文以粗糙集理论中的属性约简为例，给出一个具体的数据规约算法。粗糙集的基本理论请参考相关文献，此处不做赘述。所谓属性约简，就是在保持知识库分类或决策能力不变的条件下，删除其中不相关或不重要的属性。我们可以使用约简后的属性集合代替原来的整个属性集合而不降低分类效果，从而使信息更加精练。

输入：决策表 $T=<U,C\quad D>$，其中 C,D 分别为条件属性集和决策属性集，互信息阈值 δ。

输出：该决策表的一个约简 R

步骤：

(1) 令初始约简属性集 $R=\Phi$；$B=C-R$

(2) 对于条件属性集 C，计算 $\gamma(C,D)$，作为停止条件；

(3) For i=1 To m，以分类质量[6]作为属性重要度的定义，计算各条件属性的重要 $sig(c_i,B,D)$；

(4) If $\gamma(R,D)=\gamma(C,D)$ 则停止，得候选属性集 $R=\{c_1,c_2,\cdots c_l\}$；否则 i+1，转(5)继续执行；

(5) 从 B 中取最大值 $\max c_i=c_q$，$R=R\quad\{c_q\}$，$B=B-\{c_q\}$，转(3)继续执行；

(6) For i=1 To l，将 c_i 作为基准属性，计算 R 中其余属性 $c_i(\neq i)$ 相对于 c_i 的互信息[2] $I(c_i,c_i)$；

(7) If $I(c_i,c_i)>\delta$ 且 $\gamma(R-\{c_i\},D)=\gamma(C,D)$，则 $R=R-\{c_i\}$，否则 i+1，转(6)继续执行；则集合 R 是该决策表的一个约简。

（下转第 70 页）

(上接第 46 页)

4 结语

据测算,数据预处理阶段的工作量大约占到了全部数据挖掘过程的 60%,而且也是较难深入的部分。做好数据预处理工作,将在其后的决策过程得到高的回报。通过将数据预处理融入数据仓库的构建过程,并综合运用多种预处理技术,必将使整个数据挖掘系统更加集成、一体化,提高数据挖掘的起点和效果。

参考文献

1 Joyce Bischoff, Ted Alexander, 成栋、魏立原译, 数据仓库技术[M], 电子工业出版社, 1998。

2 刘明吉、王秀峰、黄亚楼, 数据挖掘中的数据预处理, 计算机科学[J], 2000, Vol. 27, No. 4: 54 ~ 57。

3 Baeza – Yates R, Perlegerg C. Fast and practical approximate pattern matching [J]. Information Processing Letters, 1996, 59: 21 – 27.

4 李华□、易宝林, 基于动态规划的缩写发现算法, 武汉大学学报: 工学版, 2004, 37(1): 128 – 131。

5 佘春红, 基于优先队列的增量式重复记录识别, 计算机应用, 2003, 23(9): 61 – 63。

6 张艳秋、徐六通, 数据集成中不一致性数据相似性比较的加权算法, 计算机科学, 2003, 30(8): 92。