数据挖掘技术在医院信息系统中的应用

韩 煜

（中山大学附属第五医院 珠海 519000）

〔摘要〕 介绍数据预处理、匿名化与标识转换等医学数据挖掘的关键技术，阐明医院信息系统中数据挖掘的基本过程，包括数据提取和预处理、运行挖掘算法、模式发现、知识表示和评价几个环节，并从模型建立、实现过程两方面详细论述数据挖掘技术在医院信息系统中的应用。

〔关键词〕 数据挖掘技术；医院信息系统；过程；应用

Application of Data Mining Technology in Hospital Information System *HAN Yu, Fifth Affiliated Hospital of Sun Yat-sen University, Zhuhai 519000, China*

〔**Abstract**〕 The paper introduces the key technologies in medical data mining such as data pretreatment, anonymization, identity transformation, etc. It elaborates the basic process of data mining in hospital information system, including data extraction and pretreatment, running the mining algorithm, model discovering, knowledge expression and evaluation. It also concretely discusses the application of data mining technology in hospital information system from two aspects: model construction and implementation process.

〔**Keywords**〕 Data mining technology; Hospital Information System (HIS); Process; Application

1 医学数据挖掘的关键技术

1.1 数据预处理

数据预处理是数据挖掘过程中的一个重要步骤，尤其当数据库中包含噪声、不完整，甚至是不一致的数据时，更需要数据的预处理[1]。在一个完整的数据挖掘过程中，数据预处理通常要花费 60% 左右的时间，而挖掘工作仅占整个过程的 10% 左右。数据预处理主要包括数据清洗、数据集成、数据转换和数据消减。

1.2 匿名化与标识转换

由于医学信息涉及到患者隐私信息的问题，医学数据需要进行特别的数据处理，即对患者记录进行匿名化和标识转换。匿名化是指从记录中去除患者的标识，或者用错误的标识代替正确的标识。匿名化之后，研究人员不可能通过观察记录知道有关患者的任何信息。标识转换与匿名化有一些细微的差别。变换后的标识可能仍然隐含着患者的真实信息，但是这些隐含的真实信息只有那些经过授权的研究人员才能获得。

1.3 医学文本数据挖掘

医学文本信息中，对影像、信号或者其它临床数据的解释是非标准化的，难以直接进行数据挖掘，需要进行标准化处理。目前通过计算机对医学文本数据进行标准化转换已经起到了一定的成效，主要包括 3 个步骤：分析源语句、转换、产生目标语句。转换的一个难点是源语句不是唯一的，因此需要无止尽地收集各种形式的源语句。目前的机器

〔收稿日期〕 2010-06-29

〔作者简介〕 韩煜，助理工程师，发表论文 3 篇。

转换只能处理小于 10 个单词的语句。XML（Extensible Markup Language）是一种结构化的语言，提供了文本数据标准化的另一途径。XML 不仅能创建包含结构化数据的文本，同时也可以共享和处理数据[2]，是数据挖掘和知识发现的关键技术。

1.4 影像数据挖掘

当前医学影像数据主要来自一些成像仪器（如 B 超、CT 等），它们已被越来越多的医学专家视为一种可靠的辅助诊断手段。因此，开发有效的影像数据挖掘工具也成为医学数据挖掘过程中的关键技术之一，这不仅仅与纯数字数据的挖掘方法不同，而且实现更加困难。医学影像数据挖掘主要包括：去除或降低影像噪音的影响，提高目标影像质量或对目标组织进行边缘提取；对目标组织进行概念描述，并概括这类对象的有关特征，从而获得或验证有关参数的动态范围；医学影像数据的管理与检索。目前，对 SPECT 影像的数据挖掘已取得了突破性进展。此外，研究快速的、优质的挖掘算法，确保挖掘所得知识的准确性和可靠性，都是医学数据挖掘的关键所在。

2 医院信息系统（HIS）中数据挖掘基本过程

2.1 医院信息系统数据挖掘过程（图 1）

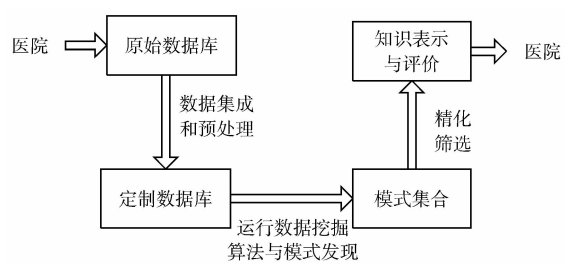

图 1 HIS 系统中数据挖掘过程

2.2 医院原始数据提取和预处理

也称数据准备阶段。医院决策部门根据某种决策需求制定挖掘任务后，需要从 HIS 系统的各种数据库、文件和非电子数据源中提取相关数据（可用数据库相关查询语言实现）[3]；然后对提取的数据进行净化和预处理（因为在提取过程中，它们往往易受噪声数据、空缺数据和不一致性数据的侵扰），正确剔除或修正错误数据，统一数据量纲，提取时间序列信息；最后进行数据变换和压缩，依据将要使用的分析方法，转换数据表示格式，提取数据特征来表示数据，降低数据的维数。该阶段为模式发现阶段提供高质量的输入数据。

2.3 运行数据挖掘算法

由于医院往往有多个不同的应用目标，因此要有与其相对应的挖掘任务。对于不同的挖掘任务，上述过程会形成多个由数据挖掘过程中使用到的信息组成的定制数据库，针对这些数据库有很多的数据挖掘算法。而每个算法都会提出一些诸如置信度、感兴趣度、新颖度等的统计属性作为对产生模式的评估标准，从而进一步决定哪些模式可以保留，哪些模式需要丢弃，更有效地找出潜在的有兴趣的模式。

2.4 模式发现

模式发现是数据挖掘过程的核心阶段，该阶段运用各种数据挖掘算法，通过对历史数据的分析，得到供决策使用的各种模式与规则。

2.5 知识表示和评价

将挖掘出来的模式与规则以直观、容易理解的方式呈现给用户。评价和筛选挖掘出来的模式与规则。

3 数据挖掘技术在医院信息系统中的应用

3.1 数据挖掘模型的建立

数据挖掘模型可以用多种方法来创建，利用 Analysis Services 模型向导，DSO 或者其它能够创建 Analysis Service 或客户端数据模型的应用软件。并且通过定位于数据透视表服务库的程序设计，还能创建出和永久模型一样好的会话级数据挖掘模型。本地数据挖掘模型的结构和关系数据库中的表很相

似。和表一样，数据挖掘模型是按 Column 来定义他们的内容的。然而又和 SQL Server 2005 中的表不一样，模型中的 Column 能够嵌套表[4]。SQL Server 2005 的 Analysis Services 支持两种数据挖掘模型：基于 OLAP 立方体模型和基于关系表的模型。

3.1.1 建立基于 OLAP 立方体的模型 本文利用了 CREATE OLAP MINING MODEL 语句来创建一个基于 OLAP 立方体的挖掘模型。其语句通式如下：

CREATE OLAP MINING MODEL < Mode l Name > FROM < Case Cube Name > （ < CubeMembers > ） USING < Algorithm Name >

其中：< Model Name > 指定了所构建模型的名字。这个模型的物理位置通过 Mining Location 特性来表示。如果 Mining Location 特性没有在连接串中说明，通过这句话所创造的挖掘模型将会有其连接域并且只在对话期间存在。< Case Cube Name > 是包含模型 < Cube Members > 的测试案例 Cube 的名字。< Algorithm Name > 包括了创建模型的算法的名字。

下面的例子创建一个医疗诊断 MyOlapModel 特性的 OLAP 挖掘模型。

```
CREATE OLAP MINING MODEL〔MyOlapMode l〕FROM〔Cure〕
(
 CASE
 DIMENSl0N〔userID〕/ * 病号号码 * /
 DIMENSION〔userSEX〕/ * 病号性别 * /
 DIMENSION〔userAGE〕/ * 病号年龄 * /
 LEVEL〔name〕/ * 诊断的病名 * /
 PROPERTY〔result〕, / * 诊断结果描述 * /
 PROPERTY〔inhospitaldate〕, PREDICT /就诊时间/
)
US IN G Microsoft – Decision – Trees
```

该模型定义了一个医疗诊断 Cure 的例子。

3.1.2 根据关系数据库的表来建立模型 通过指定模型中的 Column 来定义一个关系型的挖掘模型（确切的说是基于关系数据库表中的模型）。因为源数据的格式和结构是先前未知的，所以通过名字、数据类型、统计特性以及其查询中的可预见性来定义每一个 Column。创建关系性挖掘模型的通式如下所示：

CREATE MINING MODEL < Model Name > (< Column Members >) USING < Algorithm Name >

例如，考虑下面这个相关挖掘模型定义：

```
CREATE MINING MODEL〔MemberCards〕
(
〔userId〕LONG KEY,
〔userAGE〕INT,
〔inhospitaldate〕DATE,
〔result〕TEXT DISCRETE, PREDICT
)
USING Microsoft – Decision – Trees
```

在此例中，使用了 CREATE MINING MODEL 语句描述了一个名为 Member Cards 的挖掘模型。该语句的句法同 SQL 中的 CREATE TABLE 语句相似。命名该挖掘模型的各列，其类型用本身所包含内容的额外信息来描述。通过在其列的描述中使用 PREDICT 指定器，将 Result 列指定为可预测的。

3.2 在客户端访问数据库中的数据挖掘实现过程

3.2.1 过程概述 可以使用 JavaAPI 创建瘦客户端应用，访问 SQL 数据库中丰富的数据挖掘功能。ODM Java API 实现 SQL 对 JDMl. 0 的特殊扩展，该扩展遵从 JSR – 73 标准扩展框架。JDMl. 0 是 JCP（Java Community Process）制定的数据挖掘 Java API 的工业标准。它定义了供数据挖掘引擎使用的 Java 接口。JDM 接口支持的挖掘功能包括分类、回归分析、聚类、属性价值和关联分析，特定的挖掘算法包括朴素贝叶斯分类、支持向量机、决策树和 K 均值。本数据挖掘技术的实现由数据采集、数据预处理、执行引擎、执行结果处理组成，见图 2。

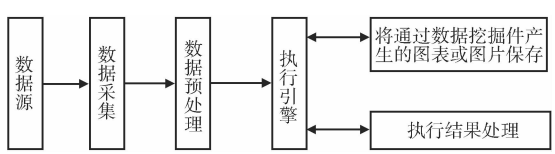

图 2 数据挖掘技术的实现

3.2.2 数据采集 通过上传文件方式接收用户提交的数据文件。用户可通过图形用户界面上传类似 Excel 的数据文件到服务器，由运行程序分析数据，

将第 1 行内容作为变量名，在数据分析中作为输入/输出的变量。

3.2.3 数据预处理 完成数据过滤，负责对待挖掘的源数据进行集成，包括数据清理与数据转换。由于不同的数据源中存在近似、重复的数据，以及对同一属性的不同表示，数据清理主要用于解决属性冗余和数据值冲突的问题。实现步骤：识别出标识同一特性的近似重复属性；将近似重复属性合并；从数据集中删除多余的属性。异构数据源中数据的表示方法、精度各不相同，需要格式化。数据转换对数据进行变换操作，将变换后的值作为新的变量存放在样本数据中。

3.2.4 执行引擎 是本实现技术的核心，体现在实现数据输入输出的控制、分析方法的灵活确定、数据挖掘脚本的生成等功能。运行程序实时创建一个执行脚本，传递给数据挖掘软件（本方案中选择 SPSS Clementine），让其根据脚本中的内容来执行每一步分析结果。一个脚本主要包括：数据来源、输入输出参数、分析模型、输出结果。

在数据挖掘分析过程中，有多个节点类型，每种类型又有不同的实现方式和个性，为了便于实现，需要抽象出节点的共性和分析算法的公用接口方法[5]。当用户在界面上操作时，可选择不同的节点组合在一起，选择时可设置某些节点的个性化参数，最后统一抽象执行该方法并将对象连接起来生成执行脚本，见图 3。

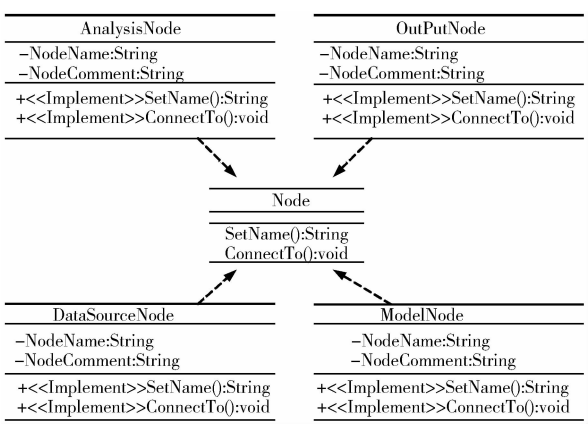

图 3 数据挖掘脚本实现 UML

执行引擎与数据挖掘软件 SPSS Clementine 之间主要是一个调用关系。执行引擎每次执行计算时，都从线程池（Thread P001）中取一个管理线程，赋予新的分析请求对象，该线程启动一个新的数据挖掘软件 SPSS Clementine 分析进程。复杂的计算和分析数据工作由进程完成，同时受到管理线程的监控。分析进程计算完成，管理线程将分析结果取回并保存到数据库中，若出错则将详细的错误信息通知用户。分析过程完成后，管理线程将分析请求对象删除，放到线程池中。见图 4。

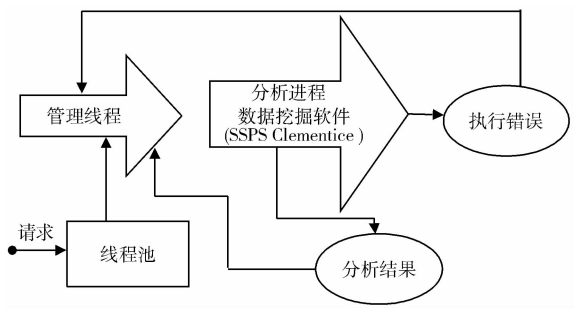

图 4 执行引擎与数据挖掘软件 SPSS Clementine 之间的调用关系

最后是执行结果处理，将数据挖掘软件 SPSS Clementine 产生的图表或图片保存起来，完成整个处理流程。

参考文献

1 周鸾杰，宋传军，周宝林，等. 数据挖掘可视化技术与医院管理〔J〕. 医疗设备信息，2006，(3)：23－24.
2 唐华松，姚耀文. 数据挖掘中决策树算法的探讨〔J〕. 计算机应用研究，2001，18（8）：96－98.
3 鲁为，王枞. 决策树算法的优化与比较〔J〕. 计算机工程，2007，33（16）：189－190.
4 刘波. 医院信息系统中数据库安全实现方法〔J〕. 计算机应用，2000，8（10）：20－21.
5 袁永革. 试析医院信息系统的设计思想及其实施〔J〕. 计算机与信息技术，2006，31（3）：182－183.