DOI:10.14107/j.cnki.kzgc.2005.01.001
2005年1月 第12卷第1期 控 制 工 程 Control Engineering of China Jan. 2005 Vol.12,No.1

文章编号:1671-7848(2005)01-0033-04

医疗诊断系统中的数据预处理

张思奇[1], 周淑文[2], 巩志国[1], 董名垂[1]
(1. 澳门大学 科技学院, 澳门 3001; 2. 东北大学 机械工程与自动化学院, 辽宁 沈阳 110004)

摘 要:针对长期积累的病历数据不仅数量庞大,记录方式、内容千差万别,而且噪声、缺省值大量存在的问题,提出了智能医疗诊断系统,利用以往的病历数据经过数据挖掘等技术来产生决策规则,以期为潜在的患者及早发现病情,获得早期诊治。总结了病历文本的记录内容和特点,对每一类的特点和形成原因进行了分析研究,并结合医疗诊断分别提出了针对它们的预处理方法,为下一步数据挖掘做好了准备。

关 键 词:数据挖掘;医疗数据;数据预处理
中图分类号:TP 274 文献标识码:A

Data Preprocess in Medical Diagnosis System

ZHANG Si-qi[1], *ZHOU Shu-wen*[2], *GONG Zhi-guo*[1], *DONG Ming-chui*[1]
(1. Faculty of Science and Technology, Macau University, Macau 3001, China;
2. College of Mechanical Engineering and Automation, Northeastern University, Shenyang 110004, China)

Abstract: Intelligent healthcare system data preprocess problem is discussed. It is to help prospective patient find his condition as soon as possible with decision rules produced based on medical cases by using data mining technology. Medical cases are not only huge in amount, but also different from each other in record styles and contents. The content and characteristic of medical cases is summarized. Combining with medical diagnosis, it aims at each kind to bring forward the idea of data preprocess for data mining respectively. The proposed method makes good preprocess for data mining.

Key words: data mining; medical data; data preprocess

1 引 言

数据挖掘(Data Mining, DM)是随着数据库技术和人工智能的发展而迅速兴起的边缘学科,它通过对海量历史数据的智能处理,可以揭示出反映事物内在规律和预测发展趋势的规则或模式。DM 在许多领域有着极其重要的作用,智能医疗诊断就是其中之一。

长期积累的病历文本的医疗数据是相当庞大的,对这些数据可以集中运用各种数据挖掘技术,了解各种疾病的典型症状、各种疾病的共性、各种疾病的发展规律等,为医疗诊断提供新的思路。此项工作对疾病的诊断、治疗和医学研究都是非常有价值的。利用 DM 技术对医学相关的研究很多,包括对心脏 SPECT 图像的数据挖掘[1],医学数据库中疾病模式的发现[2],可视化数据挖掘[3]等方面的研究。数据挖掘的各种技术和方法在医学领域都有广泛的应用,在今后的几年里,医学领域内的数据挖掘技术水平会更高,应用会更广。

2 在线智能医疗诊断系统简介

在线智能医疗诊断系统,是澳门大学科研委员会资助项目“Network-based, Intelligent, Home Healthcare System”。设计者尝试用数据挖掘对一些实际的医疗数据进行分析。从珠海某医院病案室保存的冠状动脉粥样硬化性心脏病(简称冠心病)病历中选取了一些数据作为挖掘的数据源。之所以选择冠心病作为研究对象是因为该病种与其他心脏病病种相比发病率较高,约占 80 %。该项目的目的是通过对这些病历数据的分析,得到一些有价值的诊断规则,帮助潜在的患者及早发现病情,以便获得最佳治病时机。在线智能医疗诊断系统如图 1 所示。

收稿日期:2004-08-10; 收修定稿日期:2004-09-18
基金项目:澳门大学科研委员会资助项目(RG049/02-03S/VM1/FST)
作者简介:张思奇(1978-),女,辽宁沈阳人,硕士研究生,主要研究方向为数据挖掘、数据仓库的理论与应用。

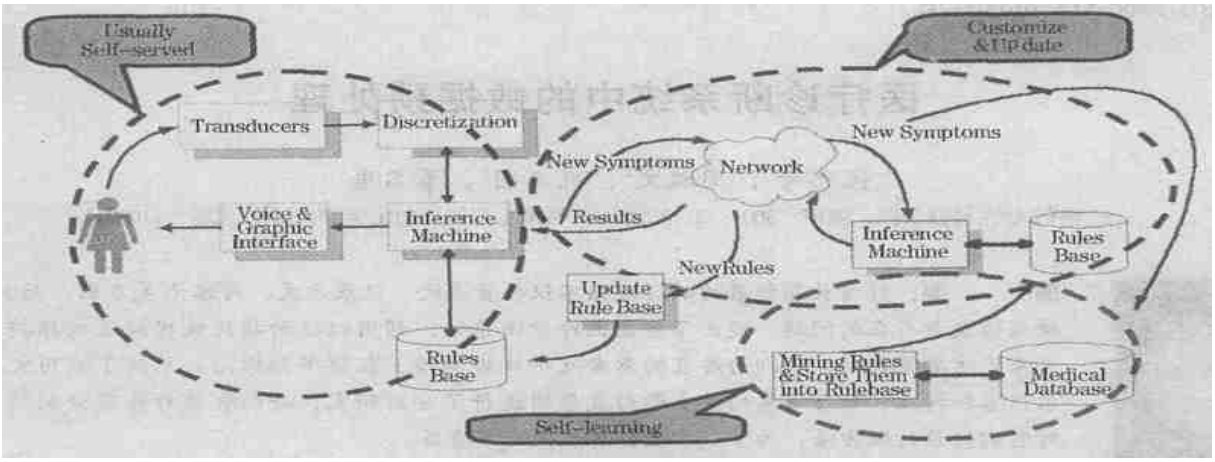

图 1 在线智能医疗诊断系统

3 数据预处理的定义

数据预处理包括实施数据挖掘算法前的所有工作。它实际上是一个转换 T，将现实世界中的原始数据向量 $\boldsymbol{X}_{ik}$ 转换为一系列新的数据向量 $\boldsymbol{Y}_{ij}$，$\boldsymbol{Y}_{ij}$ 由式(1)求得：

$$\boldsymbol{Y}_{ij}=T(\boldsymbol{X}_{ik}) \tag{1}$$

式中，$i=1, 2, \cdots, n$，n 为对象数目；$j=1, 2, \cdots, m$，m 为预处理后的特征数目，通常 $m\neq 1$；$k=1, 2, \cdots, l$，l 为预处理前的属性/特征的数目。

式(1)中，$\boldsymbol{Y}_{ij}$ 保存了 $\boldsymbol{X}_{ik}$ 中的“有价值的信息”。它消除了至少一个在 $\boldsymbol{X}_{ik}$ 中存在的问题。

通过上述关系处理，$\boldsymbol{Y}_{ij}$ 比 $\boldsymbol{X}_{ik}$ 更加有用。有价值的信息是那些存在于数据中的知识组件，而数据分析的目的就是发现它们并用某种方式将它们表示出来。有价值信息通常定义为以下 4 种属性：有效的、新颖的、潜在有用的和最终可被理解的。

4 医疗数据的特点

病历信息涵盖了医疗过程和医院活动的全部数据资源，包括临床医疗信息和医院管理信息。感兴趣的主要是临床医疗信息，只有这部分数据才能反映医学信息的独特之处，本项目的数据挖掘的主要对象也是针对这部分信息资源的。挖掘的医疗数据是心脏病病人的 4 600 页病历记录，共 640 个病历记录。病历记录由中医和西医组成，包括病人个人信息、病人自述、医生体测、化验和诊断结果等。必须把这些数据转换成适合数据挖掘算法的形式。但是不同的医生描述的风格和方式存在差异，需要在文本导入数据库和数据预处理阶段对它们统一和规范，这就需要对病历的描述和医疗知识有一定的了解。

1）噪声数据 由于病历是由医生手工记录的，这就会存在记录错误，它们多由笔误造成。而数据库录入人员为非医学专业人员，对于一些属性值的明显出入不能及时发现，同时录入时偶尔也会将原本正确的数据输错，使得数据具有噪声。带噪声的数据如果不处理则会影响知识发现的准确性。

2）记录形式不统一 不同的病历是由不同的医生撰写的，记录方式会有不同。

①记录用词 根据医生的习惯用词，例如有的医生喜欢用“双下肢浮肿”，有的用“两下肢浮肿”，而有的则用“双下肢中度可凹性浮肿”。②病症程度的描述 医生根据自己的主观感受和已往的经验进行描述，例如有的医生认为是“轻度浮肿”，有的认为是“中度浮肿”。不同的病人对相同的症状的主述也会出现不同，例如“头痛”，有的病人认为是“中等头痛”，有的认为是“剧烈头痛”。③描述的简略 一些常用医学用词或医生已在前面提及的用词常常略写，例如“烦渴”是“烦躁”和“口渴”的简称。

3）大量缺省值 在病历记录中存在大量的缺省值主要有以下 5 个原因：

①检测设备的发展 为了使医生获得更准确的诊断结果，很多测试设备正进行着改进。然而，因为一些在新的测试设备投入使用前得到的记录不会包括任何信息，从数据分析的角度来看，这些测试结果给数据挖掘带来更多的困难。②与医生诊断相关的缺省值 医生在做诊断时为了尽快地得到诊断结果，通常会选择体测和化验的项目。如果医生具有相当丰富的专业知识，感觉一些项目没有用，他们就不会做记录，那么这些测试值就会为空。③无

异常变化的症状　鉴于病历记录的特点，医生一般只在病状出现病变时才对它们做出相应的记录。因此无异常变化的症状将不做任何描述。④基本医学常识的省略　病历都是给医生、护士使用的，所以一些医学上的基本常识经常不写，例如，有的疾病名称的 ICD 号码。数据挖掘人员应该通过缺省数据对病症属性进行分类和处理。⑤检查项目不同　由于一些检测项目可以互相替代，并且根据医生的习惯、病人的经济能力，相同的病症因医生和病人的不同会进行不同项目的检测。因此在病历当中，这些项目的缺损率很高。

5　数据预处理

病历数据具有数量大、记录形式不够统一、大量缺省值存在、简略、存在易误字和易混字等特点，不利于挖掘人员迅速有效地发现所希望得到的信息。因此在进行数据挖掘之前需要在病历文本导入数据库时针对上述特点进行预处理，为数据挖掘内核算法提供干净、准确、更具有针对性的高品质数据，从而减少挖掘内核的数据处理量，提高数据挖掘效率，增加知识发现的起点和知识的准确度。

1）属性子集的选择和噪声数据的处理　最初采集的数据中并非所有的属性对于知识发现而言都是必要的。例如病历数据中的姓名、地址等，这些属性对病情没有任何实际意义，必须去除。根据医疗知识进行数据选择，建立 Access 数据库，病历为记录，症状为条件属性，疾病为决策属性。把属性分成两种类型：离散型（像呼吸困难）和连续型（像年龄，血压）。对于噪声数据的处理，可以采用以下 4 种方法。

①分箱（Binning）[4]　利用属性值的相邻性进行数据的平滑化，将这组属性值按照大小次序排成一个线性队列，再按照一定的步长将其分成若干个小组，最后就每个小组局部进行数据的平滑化，由于分箱方法参考相邻的值，可进行局部平滑。②聚类（Clustering）[5]　将一组数据按照某种相似性划分为若干个小组，如数据值的大小、数据语义的分类等，而那些遗留在所有小组之外的零散数据将被作为一种噪声数据而剔除。③人机结合检查　可以通过人工检查和计算机结合的办法来识别孤立点。④回归（Regression）[6]　定义一个回归函数来平滑数据。线性回归涉及找出适合两个变量“最佳”直线，使得一个变量能够预测另一个；多线性回归是线性回归的扩展，它涉及多个变量，数据要适合一个多维面。

2）空缺值处理　数据挖掘算法只适用于无缺省值的数据库，填充缺省值方法有：

①去除具有缺省值的样本　此方法优点是保留的数据全是真实数据，但只适用于具有少量缺省值的样本集。②以特定的值填充　对于数值属性，可以使用所有与该样本属于同一类别（同一疾病）的样本的平均值；对于非数值属性，可以使用所有与该样本属于同一类别的样本中出现最多的值。此方法不会造成任何有用信息的丢失，但会引进一定程度的误差。③中医症状　对于病历中属于中医的症状存在缺省值时，例如“晕厥”，“出汗”，根据中医的特点，认为是无异常变化的症状，将该症状的正常值与其填充。

3）连续属性的离散化　当遇到连续属性时，要把它进行离散化才能对样本进行数据挖掘。一种方法是一边建树一边离散，称它为动态离散化，例如 C4.5 中处理连续型属性的方法；另外一种是在建树前就将属性离散，称为静态离散化。根据离散化的执行方式，可分为划分法和归并法。划分法就是先将整个属性取值空间作为一个离散属性值，然后对该区间进行划分，一般是将它一分为二，每个区间对应一个离散值。这个过程循环下去直到满足停止条件为止。归并法是将整个属性取值区间看成由许多的小区间组成，然后根据某种标准将其中相近的合并在一起形成一个大区间，这个过程循环往复直到满足停止条件为止。如果离散化每次只对一个属性进行离散，则称为局部离散化。反之称为全局离散化。如果离散化过程中使用了分类信息，则称为监控离散化。例如 Holt 在 1993 年提出的 1R 离散法，反之称为非监控离散化。

离散化问题可以定义如下：

A 是数据的所有属性中的一个连续属性，A 的值域为区间$[a, b]$。$[a, b]$ 上的 k 个划分 π_k 是由 k 个区间构成的集合。π_k 由式（2）求得：

$$\pi_k = \{[a_0, a_1], [a_1, a_2], \cdots, [a_{k-1}, a_k]\} \quad (2)$$

式中，$a_0 = a$；$a_{i-1} < a_i$（$i = 1, 2, \cdots, k$）；$a_k = b$。

离散化就是在$[a, b]$产生 π_k 的过程。

根据所研究的课题，发现对某一属性离散化，如果要离散的连续属性与数据挖掘的目的之间有一定的关联度的话，一定可以发现一个或几个区间与决策属性的某一类别之间有较强的相关性，把它们称为特征区间。（下转第 66 页）

```
        ANL A, #0FEH
NEXT:   MOV R0, A
        MOV A, R1
        RL A
        MOV R1, A
        DJNZ B, CRC-LOOP ; 判断移位次数
        RET
```

……

4 CRC 校验在内燃机车柴油机内漏检测仪中的应用

内燃机车柴油机冷却水系统的作用是通过水的循环对其缸套、缸盖、增压器中的冷空气和润滑机油进行冷却，以保证各部位不会过热损坏。由于此冷却循环系统冷却部位多，联接管路复杂，因此极易发生冷却水内漏到柴油机汽缸内，造成机破临修，影响行车安全，所以日常检测柴油机汽缸是否内漏是十分必要的。柴油机汽缸冷却水一旦发生内漏，其汽缸内排放出来的气体的湿度必将发生显著的变化。可以检测出柴油机汽缸内气体在同一温度下排放时的湿度，然后根据同一台机车各个汽缸的湿度变化规律来判断是否有汽缸发生内漏。

依据这一原理采用 SHT71 做传感器来研制了内燃机车柴油机冷却水内漏检测仪，由于现场工作环境极其恶劣，微控制器与传感器之间的通讯数据常会发生无法预测的错误。为了防止错误所带来的影响，在数据通讯中采用了 CRC 校验的通讯规约。实践证明 CRC 法在微控制器与 SHT71 通讯的差错检验中应用是十分有效的，可以在实际中得到广泛应用。

5 结 语

对于数据通讯校验，相对于硬件的实现方法而言，用软件来实现 CRC 校验过程无疑是一种简单实用而又成本低廉的方法。本文设计 SHT71 的 CRC 值生成算法已经在内燃机车柴油机汽缸冷却水内漏检测仪得以实现，该仪器已在武昌机务段成功应用，获得良好的效果。

参考文献：

[1] 何立民. 单片机应用技术选编(4)[M]. 北京：北京航空航天大学出版社，1999.

[2] 戚俊，李季，张毅，等. CRC 校验在 DALLAS 单总线产品中的应用[J]. 量子电子学报，2001，18(6)：556-559.

[3] 张成君，等. 机车柴油机[M]. 北京：中国铁道出版社，1998.

[4] 梁寿愚. 分布式数据库通用查询对象模型[J]. 控制工程，2002，9(6)：18-21.

(上接第 35 页)

例如在对医疗数据进行分析时对年龄进行离散化，发现就冠心病来说，50～60 岁，70～80 岁的中老年人发病率相对较高，显然年龄区间[50，60] 和[70，80] 是两个与决策属性有较高相关性的特征区间。在特征区间之间的数据，可按照一定的标准(例如信息增益)进行二分，一部分归属于前一个特征区间，另一部分归属于后一个特征区间。特征区间可以通过统计的方法确定。

经过前面几步数据处理后，数据库中既无冗余数据又无缺省数据，只是数据的形式有数值型和布尔型等，这样混合型的数据库不满足知识发现算法的要求，因此必须将数据库全部转化为满足算法要求的布尔型数据库。对于离散型数据，每个不同的离散值用一个整数与之对应。显然如果属性值过多就会出现问题，但是这种情况在医疗数据中很少见。如果数据是连续的，根据类别对属性离散化分区间，每个不同的区间用一个整数与之对应。

6 结 语

医疗数据挖掘是计算机技术、人工智能、统计学等与现代医疗相结合的产物，为家庭医疗保健提供了强有力的支持。本文分析了病历文本的记录内容和特点，对每一类的特点和形成原因进行了研究，并结合医疗诊断分别提出了针对它们的预处理方法，为下一步医疗数据挖掘的成功进行奠定了基础。实践表明，本文所提供的方法可以在医学领域具有良好的应用前景。

参考文献：

[1] Sacha J P, Cios K J. Issues in automating cardiac SPECT diagnosis[J]. IEEE Engineering in Medicine and Biology, 2000, 19(4): 78-88.

[2] Ramirez J C G, Cook D J. Temporal pattern discovery in course-of-disease data[J]. IEEE Engineering in Medicine and Biology, 2000, 19(4): 63-71.

[3] Ankerst M. Visual data mining[D]. Germany: University of Munich, 2000.

[4] Pyle D. Data preparation for data mining[M]. San Francisco: Morgan Daufinann, 1999.

[5] Jain A K, Murty M N. Data clustering: a review[J]. ACM Computing Surveys (CSUR), 1999, 31(3): 264-323.

[6] Neter J, Kutner M H. Applied linear statistical model(4th ed)[M]. Chicago: Irwin, 1996.