糖尿病电子病历数据预处理*

杨美洁 浦科学 李 准

（重庆医科大学医学信息学院 重庆 400016） （重庆医科大学附属儿童医院 重庆 400014）

〔摘要〕 对糖尿病海量数据的电子病历进行处理和挖掘具有重要的意义。利用 SQL 语句和函数，基于 SQL Server 2008 平台对糖尿病电子病历的数据进行清洗、集成、变换和规约等数据预处理。消除噪声、不完整和不一致性的数据，实现非结构化文本数据到结构化数值数据的转换。

〔关键词〕 糖尿病；电子病历；预处理

〔中图分类号〕R－056 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2016.05.014

Data Preprocessing of Diabetes Electronic Medical Records *YANG Mei－jie, PU Ke－xue, Medical Informatics College, Chongqing Medical University, Chongqing 400016, China; LI Zhun, Children's Hospital of Chongqing Medical University, Chongqing 400014, China*

〔**Abstract**〕 It has great significance for the processing and mining of diabetes Electronic Medical Records (EMR) containing mass data. By use of SQL statements and functions, based on SQL Server 2008, the paper preprocesses data of diabetes EMR, including data cleaning, integration, transformation and reducing, etc. It eliminates noisy, incomplete and inconsistent data and transforms unstructured text data to structured numerical data.

〔**Keywords**〕 Diabetes; Electronic Medical Records (EMR); Preprocessing ; Data cleaning

1 引言

随着医院信息化水平的不断提高，含有医学信息和规律的糖尿病电子病历数据急剧增长，如何对此类数据进行处理和利用成为重要的问题。电子病历（Electronic Medical Record, EMR）是医务人员在医疗活动过程中，使用医疗机构信息系统生成的文字、符号、图表、图像、数据、影像等数字化信息，能实现存储、管理、传输和重现的医疗记录，是病历的一种记录形式[1]。对糖尿病电子病历的分析和挖掘，可以发现潜在有效的规律，为医疗、教学和科研等方面提供服务，为临床诊断和决策提供依据。而分析和挖掘的前提条件是对糖尿病电子病历数据进行预处理，实现数据的一致性和规范化。目前数据预处理技术的研究比较成熟，应用领域广泛，如生物医学、图书情报、物理化学、地质科学、电力[2]、机械等。在大数据背景下数据预处理技术的重要性更加突显，如在数据挖掘[3-4]、Web 日志挖掘[5]、文本挖掘[6-7]、医学数据挖掘[8]和论文相似性检测[9]等方面的应用[10-11]。曹洪欣等认为电子病历的数据具有医学特点并且可能来源于不同的医疗机构的不同的电子病历系统，因此其预处理方法有别于其他领域。主要对缺失数据和噪声数

〔修回日期〕 2016－03－10

〔作者简介〕 杨美洁，博士生，讲师，发表论文多篇；通讯作者：浦科学，副教授，博士。

〔基金项目〕 重庆医科大学医学信息学院助力计划（项目编号：2014A009）。

据的清洗、对异构数据的集成和重复数据的删除、数据的变换和规约等处理[12]。

本文基于糖尿病电子病历和医学数据的特点，利用 SQL Server 2008 平台和 SQL 中的相关处理语句和函数，对重庆市某医院的糖尿病电子病历 1 433 条进行研究。对糖尿病电子病历数据进行清洗、集成、转换和规约等预处理。实现数据的一致性和规范化，为后续的回归分析、聚类分析、分类分析和关联规则等数据挖掘奠定基础。

2 糖尿病电子病历数据预处理

2.1 概述

一份完整的病历主要包含病人的基本信息、入院记录、病程记录、检查检验和药品医嘱等信息。大多以文本的形式存储。其中入院记录有病人入院时病情状况的描述[13]，主要包含个人史、既往史、家族史、现病史和体格检查等。病程记录则记录病人在整个住院期间的各种检查结果、治疗过程和病情变化情况[13]，主要包含首次病程记录、日常病程记录、上级医师查房记录等。检查检验则记录病人的检查检验项目和检查检验结果。数据预处理主要是消除数据的噪声、不完整、冗余、不一致。一般包括数据清洗、数据集成、数据变换、数据规约 4 个方面[14]。数据预处理的基本过程，见图 1。其中存储在一系列二维表中与糖尿病相关的信息进行提取，然后对其进行数据预处理，将处理后的数据存放在结果数据库中，为后续的数据挖掘过程奠定基础。

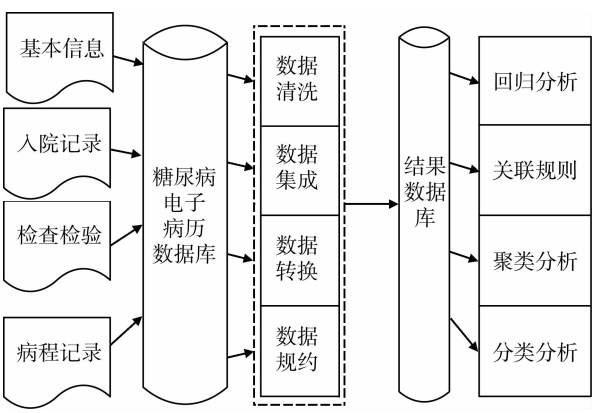

图 1 糖尿病电子病历数据预处理

2.2 数据清洗

糖尿病电子病历中的某些数据是不完整、有噪声和不一致的。数据清洗主要填充缺失值，识别非法值，纠正不一致性的数据。若该属性项的数据类型为数值型，对此类型的缺省值采用该属性的均值填充缺省值。电子病历中的属性项都有一定的取值范围，不在此范围的数据均视为非法数据。此类错误的出现主要是由于录入病历时疏忽，为保证数据的完整性和正确性，必须对此类数据进行修改。如“年龄 =500”、“舒张压 =50kpa” 等。电子病历数据还存在不一致的情况。某些属性项间的取值存在一定的相关性，可以用相关性寻找并纠正此类数据。

2.3 数据集成

数据集成是合并来自不同二维表的数据，存放在一个二维表中。本文将全部数据以住院号为外键进行关联，将相关的表分别通过住院号属性进行等值连接。部分 SQL 语句如下：

```
select 属性 1, 属性 2, …, 属性 n
from 表格 1 join 表格 2 on 表格 1. 住院号 = 表格 2. 住院号 join 表格 3 on 表格 2. 住院号 = 表格 3. 住院号
…
join 表格 n on 表格 n - 1. 住院号 = 表格 n. 住院号
order by 住院号
或者
select 属性 1, 属性 2, …, 属性 n
from 表格 1 , 表格 2 , …, 表格 n
where 表格 1. 住院号 = 表格 2. 住院号 and 表格 2. 住院号 = 表格 3. 住院号 …and 表格 n - 1. 住院号 = 表格 n. 住院号
order by 住院号
```

2.4 数据转换

数据变换将数据的类型或者取值范围变换成适合挖掘的形式。本文主要将糖尿病电子病历中的文本数据转换成适合挖掘和分析的数值数据[15]。

2.4.1 基本信息数据转换 例如对基本信息中的性别、年龄等属性进行数据转换。将性别中的男、女的属性取值修改为 1、2；年龄根据范围进行修改，如小于 50 岁取值 1，50 ~ 60 岁取值 2，60 ~ 70

岁取值为3，大于70 岁的取值为4。对此类信息进行转换后，可以对其进行单因素和多因素回归分析，从而得出糖尿病的影响因素。可以用 SQL 的 UPDATE 语句和 CASE – WHEN – ELSE – THEN – END 进行转换。年龄转换的部分代码如下：

```
UPDATE 基本信息
SET 年龄 =
(CASE
When 年龄 <50 then 1
When 年龄 between 50 and 60  then 2
When 年龄 between 60 and 70  then 3
Else 4
End )
```

转换结果，见表1。

表1 基本信息数据转换结果

住院号	性别	年龄
09006897	2	2
……	……	……
12002334	2	3

2.4.2 入院记录数据转换 入院记录主要包含个人史、既往史、家族史、现病史和体格检查等。如对既往史进行数据转换，提取与糖尿病相关的既往史疾病如肝炎[16]、高血压[17]、冠心病[18]、结核[19]等加入既往史表中作为属性。利用 SQL 技术编写相关函数，从既往史记录内容中查找是否存在既往史疾病（肝炎、高血压、冠心病、结核）的内容，若存在则该条记录相应的既往史疾病取值为1，否则取值为0。对此类信息进行转换后，可以对其进行单因素和多因素回归分析，从而得出影响糖尿病的既往史疾病；可以进行关联规则分析，得出糖尿病与其既往史疾病之间的规则，利用规则知道糖尿病的诊断与预防；可以根据个人史、既往史、家族史、现病史等数据采用不同的分类算法或者决策树分类算法[20]对糖尿病进行分类预测。

如既往史的记录内容中存在“无”、“不”、“没”、“未见”、“否”或不存在关键词则赋为0，代表否认该既往史；否则赋值为1。部分代码如下所示[21]（KeyWord 代表相关疾病如肝炎[16]、高血压[17]、冠心病[18]、结核等）：

```
UPDATE 既往史
SET [' + @ KeyWord + '] =0
WHERE dbo. getStr（记录内容，'+ @ KeyWord + '）
LIKE "% 无%"Or dbo. getStr（记录内容，'' + @ KeyWord + '）
LIKE"% 不%"Or dbo. getStr（记录内容，'+ @ KeyWord + '）
LIKE "% 未见%"Or dbo. getStr（记录内容，'' + @ KeyWord + '）
LIKE "% 没%"Or dbo. getStr（记录内容，'+ @ KeyWord + '）
LIKE "% 否%"Or dbo. getStr（记录内容，'+ @ KeyWord + '） =''
```

既往史数据转换结果，见表2。

表2 既往史数据转换结果

住院号	冠心病	高血压	肝炎	结核
09006897	0	1	1	0
……	……	……	……	……
12002334	0	0	0	0

2.4.3 病程记录数据转换 将症状体征的记录内容进行一致化处理。如利用 SQL 语句的 REPLACE 函数将记录内容进行转换，使得医学名词更加术语化、专业化。如将“上腹胀痛”替换为“腹痛”，“血压升高”替换为“高血压”，“湿罗音”替换为“湿啰音”等。然后将专业化的术语加入病程记录表中作为属性，若某条记录出现此术语，则对应的属性取值为1，否则取值为0。转换过程同既往史转换一致。对转换后的数据可以进行聚类分析，得出某类糖尿病病人的共同或者相似的症状体征。转换后的数据，见表3。

表3 病程记录数据转换结果

住院号	腹痛	高血压	湿啰音	……	乏力	气喘	眩晕
09006897	1	0	0	……	0	0	1
……	……	……	……	……	……	……	……
12002334	0	0	0	……	0	1	0

2.4.4 检查检验结果转换 检查检验结果分为高于正常值范围（h）、正常（z）、低于正常值范围（l）3 个水平。利用 UPDATE 语句对上述范围进行修改，取值为3、2、1，未做检查检验项目取值0。对此类信息进行转换后，可以进行聚类分析，得出某类糖尿病病人的共同或者相似的检查检验结果。转换代码如下所示转换结果，见表4。

```
UPDATE 检查检验
SET 检查检验结果 =3
WHERE 检查检验结果 = 'h'
```

表4 检查检验项目的数据转换结果

住院号	白蛋白	白细胞	脂蛋白
09006897	1	3	1
……	……	……	……
12002334	0	2	0

2.5 数据规约

糖尿病电子病历每个表格包含众多的属性，其中很多属性与挖掘任务不相关或冗余。维规约技术是通过删除不相关的属性（或维）来减少数据量[13]。本文采用逐步向后删除属性的方法进行维规约。即从整个属性集开始，每一步删除在属性集中无关的属性。如通过向后删除的维规约方法将基本信息的属性集中姓名和家庭住址等属性删除，从而得到挖掘或分析需要的属性集。

3 结语

电子病历具有数据量大、形式复杂多样等特点。电子病历的数据预处理是医学数据分析、挖掘和利用的基础和前提。在整个数据挖掘过程中，数据预处理占据重要地位。本文利用 SQL 技术和 SQL Server 2008 平台对糖尿病电子病历进行清洗、集成、转换和规约等预处理，将基本信息、入院记录、病程记录、检查检验记录中的文本数据转换成数值型数据，消除糖尿病电子病历中不完整、有噪声和不一致的数据。得到数据挖掘和数据分析的属性集，为后期的数据分析和回归分析、关联规则、聚类分析和分类分析等数据挖掘奠定基础。未来的工作将利用此方法对川崎病、高血压等电子病历进行数据预处理，或者对处理后的糖尿病电子病历进行关联规则、聚类、分类、回归分析等数据挖掘。

参考文献

1 谢栋梁. 电子病历初探［J］. 医学信息（上旬刊），2010，（6）：1781－1782.

2 Qu Z, Liu J. A New Method of Power Grid Huge Data Pre－processing［J］. Procedia Engineering, 2011, (15): 3234－3239.

3 Chen S, Shen B, Wee S, et al. Adaptive and Lazy Segmentation Based Proxy Caching for Streaming Media Delivery［C］. Proceedings of the 13th International Workshop on Network and Operating Systems Support for Digital Audio and Video, 2003: 22－31.

4 Han J, Kamber M, Pei J. Data Mining: concepts and techniques: concepts and techniques［M］. San Francisco: Elsevier, 2011.

5 刘立军，周军，梅红岩. Web 使用挖掘的数据预处理［J］. 计算机科学，2007，（5）：200－201，204.

6 Munková D, Munk M, Vozár M. Data Pre－processing Evaluation for Text Mining: transaction/sequence model［J］. Procedia Computer Science, 2013, (18): 1198－1207.

7 Haddi E, Liu X, Shi Y. The Role of Text Pre－processing in Sentiment Analysis［J］. Procedia Computer Science, 2013, (17): 26－32.

8 王华，胡学钢. 医学数据挖掘中的数据预处理 Apriori 算法改进［J］. 计算机系统应用，2009，（9）：94－97.

9 刘伙玉，王东波. 面向论文相似性检测的数据预处理研究［J］. 现代图书情报术，2015，（5）：50－56.

10 Guo P, Chen S S, He Y. Study on Data Preprocessing for Daylight Climate Data［M］. Berlin Heidelberg: Springer, 2012: 492－499.

11 Zhang N, Lu W F. An Efficient Data Preprocessing Method for Mining Customer Survey Data［C］. Industrial Informatics, 2007 5th IEEE International Conference on IEEE, 2007: 573－578.

12 曹洪欣，蔡海英，王侠，等. 基于 EMR 数据挖掘的临床路径构建中 EMR 数据预处理［J］. 中国医院管理，2013，33（3）：58－60.

（下转第 84 页）

mp. gov. cn/kjwxgyfw/201012/t20101202_ 2099. htm.

5 唐小利，孙涛涛，梅梅. 面向国家重大科技专项的信息服务探索与实践［J］. 中华医学图书情报杂志，2011，20（5）：1－4.

6 吴鸣，王丽. 嵌入式学科情报服务实践——以支持国家重大科技专项科研创新为例［J］. 图书情报工作，2013，57（22）：43－48，36.

7 高东平，方安，李扬，等. 知识服务平台的设计与应用—以重大传染病信息知识服务平台为例［J］. 情报理论与实践，2011，34（7）：111－115.

8 王桂枝，杨春华，李焱，等. 我国重大传染病专项信息管理与知识服务公共资源平台构建［J］. 军事医学科学院院刊，2010，（5）：473－475.

9 霍飞，徐娜，刘长娜，等. 微博等新媒体信息监测在重大传染病防控工作中的应用［J］. 职业与健康，2014（3）：411－413.

（上接第62 页）

13 庄军，郭平，周杨，等. 电子病历数据预处理技术［J］. 计算机科学，2007，（3）：141－144.

14 刘明吉，王秀峰，黄亚楼. 数据挖掘中的数据预处理［J］. 计算机科学，2000，（4）：54－57.

15 李准，冯思佳，杨美洁，等. 关联规则技术在冠心病电子病历中的应用［J］. 医学信息学杂志，2015，（1）：58－62.

16 张林杉. 住院2 型糖尿病患者非酒精性脂肪性肝病现状调查及危险因素分析［D］. 上海：复旦大学，2013.

17 王佳笑. 基于中医结构化住院病历数据的糖尿病合并高血压病证结合诊疗规律探讨［D］. 北京：中国中医科学院，2014.

18 高敏. 冠心病多因素相关性分析［D］. 石家庄：河北医科大学，2012.

19 李艳静，高微微，常占平. 肺结核合并糖尿病对抗结核药物血药浓度的影响［J］. 中国防痨杂志，2012，34（1）：23－25.

20 李怀庆. 决策树算法在医院数据挖掘中的应用探索［J］. 医学信息学杂志，2009，（8）：11－13.

21 李准. 基于冠心病电子病历的数据挖掘研究［D］. 重庆：重庆医科大学，2013.