第 33 卷 第 9 期 计 算 机 工 程 2007 年 5 月
Vol.33 No.9 Computer Engineering May 2007
·软件技术与数据库· 文章编号：1000—3428(2007)09—0103—03 文献标识码：A 中图分类号：TP311.5

2 型糖尿病发病危险因素及其特征提取技术

王 恒[1]，罗森林[1]，张铁梅[2]，韩怡文[2]
（1. 北京理工大学电子工程系，北京 100081；2. 卫生部北京老年医学研究所，北京 100730）

摘 要：综合选用数据挖掘中的 EM 聚类算法和 C4.5 分类算法，设计并进行了一系列以研究 2 型糖尿病发病危险因素与血糖变化关系为目的的实验。研究结果包括：发现了新的血糖门限值 5.26 和未发病门限值 4.28，发现和验证了影响最大的 8 个重要发病危险因素及其对应的一系列重要临界值，定性定量相结合地给出了各个重要发病危险因素的影响程度在血糖值不同预警门限值下的变化关系等。

关键词：EM 算法；C4.5 算法；2 型糖尿病；发病危险因素

Risk Factors and Feature Extraction Technology of Type 2 Diabetes

WAN Heng[1], LUO Senlin[1], ZHANG Tiemei[2], HAN Yiwen[2]
(1. Department of Electrical Engineering, Beijing Institute of Technology, Beijing 100081；
2. Beijing Institute of Geriatrics, Ministry of Health, Beijing 100730)

【Abstract】 This paper combines expectation maximization(EM) algorithm and C4.5 algorithm to build a Type 2 diabetes data processing system. With the system, a series of data mining experiment is designed to seek for important Type 2 diabetes risk factors and their relationships with blood glucose. Through a large quantity of experiments, some pathological knowledge of Type 2 diabetes is obtained, which includes 2 new blood glucose threshold——5.26 and 4.28, and 8 important Type 2 diabetes risk factors. Based on these factors and the results, the relationship between the functions of those risk factors and different blood glucose thresholds is studied and illustrated. And the relationship between important risk factors and blood glucose is analyzed.

【Key words】 Expectation maximization algorithm; C4.5 algorithm; Type 2 diabetes; Risk factors of disease

糖尿病是一种慢性非传染性疾病，近年来患病人数逐年增多，其中 90%以上为 2 型糖尿病病人。生物医学界对糖尿病的研究投入巨大，经过了 100 多年的研究和积累，目前已经准确定位了糖尿病的发病器官为胰腺，并进一步深入到胰岛细胞，胰岛内的β细胞；发现了胰岛素，知道了胰岛素的分子结构甚至可以人工合成；知道了胰岛素受体，确定了胰岛素与胰岛素受体的基因位点，及一些与特殊类型糖尿病发病相关的基因位点。筛选了几百个可能与 2 型糖尿病发病相关的基因，对这些基因指导合成的各种酶、受体、细胞因子的生理及病理生理作用进行了大量的研究。但当人们试图用这些已知的、分解到基因水平上的研究成果来阐述糖尿病的发生、发展过程时，发现这绝不仅仅是一个简单拼装的过程。另一方面，以往在应用统计学分析致病因素及预防手段时，大量采用经典的统计学分析方法，而在预测方面应用较多的是线性回归的方法。在最近的研究中，通过引入数据挖掘技术来处理大量的调查和体检，并取得了很好的效果。本文采用 EM 和 C4.5 算法相结合的方法，设计多组实验，从大量数据中发现了较有价值的研究结果。

1 2 型糖尿病数据处理系统设计

1.1 原始数据及预处理

本文所用实验数据来源于 2001 年 2 月~9 月北京市西城区和海淀区科研院所 17 946 人的整群抽样横断面健康调查，从中筛选出 17 072 条有效数据。其中每条记录（或样本）中数据项分为一般资料、生活和工作习惯、病史、体检数据和实验室检查数据 5 个部分：(1)一般资料部分主要是被调查人群的一般情况，包括年龄、性别、民族职业等；(2)生活和工作习惯部分主要是为了研究糖尿病与个人习惯之间的关系，主要包括抽烟及烟量、喝酒及酒量、工作压力及精神压力等方面；(3)病史部分主要是为了研究糖尿病与相关病史之间的关系，研究的相关病史主要包括高血压、糖尿病、高血脂等；(4)体检数据包括人的身高、体重、腰围等；(5)实验室检查数据包括胆固醇、甘油三脂、低密度脂蛋白、高密度脂蛋白等。

原始数据的预处理主要包括数据清理、纠正非法值、数据变换、数据规约等[1,2]。

1.2 数据处理框架及实验设计[3]

大量 2 型糖尿病实测数据处理的框架如图 1 所示。

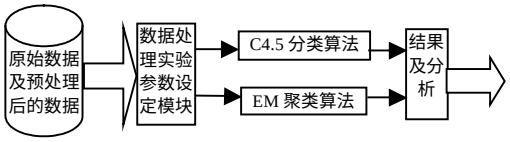

图 1 数据处理框架示意图

下面分别介绍功能结构图中的不同组成模块：

(1)数据处理实验参数设定模块。本模块的功能为根据用户要求选取算法、设定该算法的参数。要求同类试验的对比中算法、参数的实验条件要保持一致。

(2)C4.5 分类算法模块。本模块的功能为使用 C4.5 算法

基金项目：国家“十五”攻关基金资助项目（2001BA702B01）；国家自然科学基金资助项目（60671008）
作者简介：王 恒(1982－)，男，硕士生，主研方向：多维数据处理；罗森林，教授；张铁梅，研究员；韩怡文，实习研究员
收稿日期：2006-12-20 **E-mail：**bitluosenlin@sina.com

对输入的数据进行处理，建立分类决策树，取多个实验中准确率最高的训练结果作为最终的分类决策树。

(3)EM 聚类算法模块。本模块的任务是使用 EM 算法对输入数据进行聚类分析，建立聚类模型。本模块 EM 算法训练的最大迭代次数设为 100 000，确保算法收敛。

本课题所设计的实验主要包括 3 大组：全部数据组，性别对比组及 2 型糖尿病家族史对比组。每一组的实验流程相同，只是针对的数据不同。数据分析实验主要分为 2 类：聚类实验和聚类、分类组合实验。

(1)聚类实验

1)数据选择：预处理后的全部 67 维 13 781 条数据；

2)算法选择：EM 聚类；

3)聚类个数：3 类或 4 类；

4)实验过程：改变参与聚类的影响因素数量和种类，观察实验结果；

5)预期目标：得到能够较好反应出人群特点的聚类结果，记录参与聚类的影响因素。

聚类实验的流程如 2 所示。

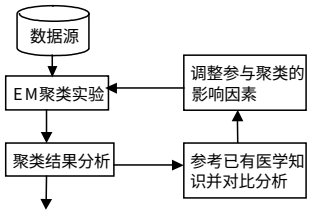

图 2 聚类实验流程

(2)聚类、分类组合实验

1)数据选择：预处理后的 67 维 13 781 条数据，门限值分别为 6.1、5.85、5.6 和 5.26；

2)算法选择：EM 聚类、C4.5 分类；

3)聚类个数：3 类或 4 类；

4)实验过程: 抚聚类再分类,得出对应不同人群特点的 2 型糖尿病发病与否的分类决策树；

5)预期目标: 得到不同健康特点人群所对应的分类决策树。

上述聚类、分类组合实验一共要做 4 组对应 4 个血糖门限值（6.1、5.85、5.6、5.26）、5 类对应不同数据集（全部数据、男性女性数据、有糖尿病家族史、糖尿病家族史不详数据）的共 10 个聚类实验和 140 个分类实验，每组实验的流程图如 3 所示。

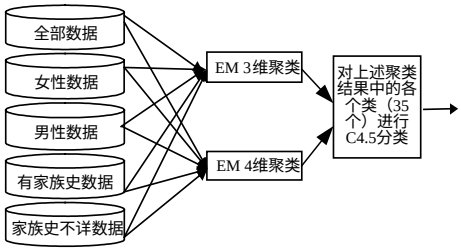

图 3 聚类、分类组合实验流程

针对数据处理结果，与同类实验、同组实验、医学知识等进行对照，找出不足并将调整的参数设置反馈到预处理和数据分析实验中以确保数据处理结果的有效性，而后围绕 2 型糖尿病发病危险因素与血糖变化关系这条线索分析得出有价值的结论。

2 实验结果及其结论[3]

2.1 血糖门限值的变化特征分析

在近些年国内外的研究中，2 型糖尿病的发病判断门限值越来越低成为一个显著特点。通过实验分析得出血糖所对应的一系列的发病门限值，如图 4 所示，分别为 6.1、5.85、5.6、5.26 和 4.28。其中，6.1 为医学界多年来使用的对 2 型糖尿病发病的判断门限，5.85 为本课题组在 2003 年的工作中提取出的血糖发病门限值[1,4,5]，5.6 为 2005 年国际上公布的血糖发病门限值，5.26 和 4.28 为本课题从大量实验结果中提取出来的新的血糖发病门限值及“未发病”血糖门限值。

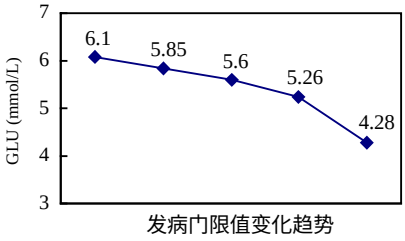

图 4 血糖门限值的变化

2.2 重要发病危险因素的提取

根据在全部的分类实验结果，统计在决策树的前 6 层中各个影响因素出现的次数，表 1 给出了全部实验结果中，按出现次数多少排列的影响因素；表 2 给出了在所有男性实验结果中影响因素按出现次数排列的结果；表 3 给出了在所有女性实验结果中影响因素按出现次数排列的结果。

表 1 全部实验结果

血糖	年龄	高密度脂蛋白	收缩压	胆固醇	体重系数	舒张压	腹围	甘油三酯
298次	196次	115次	115次	100次	95次	91次	91次	84次

表 2 男性实验结果

血糖	年龄	舒张压	身高	高密度脂蛋白	腹围	体重	胆固醇	体重系数	收缩压	甘油三酯
49次	35次	25次	25次	23次	20次	19次	14次	13次	10次	11次

表 3 女性实验结果

血糖	高密度脂蛋白	年龄	收缩压	舒张压	腹围	身高	体重	胆固醇
51次	24次	23次	16次	14次	12次	11次	11次	10次

由表 1~表 3 可知，对 2 型糖尿病发病影响较大的 9 个重要发病危险因素为血糖、年龄、高密度脂蛋白、收缩压、舒张压、胆固醇、体重系数、腹围、甘油三酯。其中，年龄起着特别重要的作用，其在全部数据和不同性别的实验中都明显地起到比其他因素更重要的作用。

2.3 重要发病危险因素的作用程度分析

在聚类、分类组合实验中的所有分类实验结果里，分别统计在 6.1、5.85、5.6、5.26 4 个血糖门限值下，重要属性在决策树中的影响程度。由 C4.5 算法可知，在决策树中，某属性越靠近根节点其所起的作用就越大。因此，统计分析在分类决策树的前 4 层中，各个重要属性所出现的位置，即可衡量出在分类实验中各重要属性的影响程度。为了把出现在不同位置的影响因素的作用统一起来，本文提出一个“层数系数”的概念，即层数系数 L：设在临界值 A 下，某影响因素 X 在第 i 层出现了 Y_i 次，则称在门限值 A 下 X 的层数系数为

$$L=\frac{1}{\sum Y_i}\sum\left(\frac{1}{2^i}Y_i\right)$$

根据上式，层数系数越大，表示该属性的位置越靠近决策树的根节点，也就是说该属性所起的作用越大，其层数系数与血糖门限关系如图 5 所示（收缩压和舒张压统一为血压

BP)。

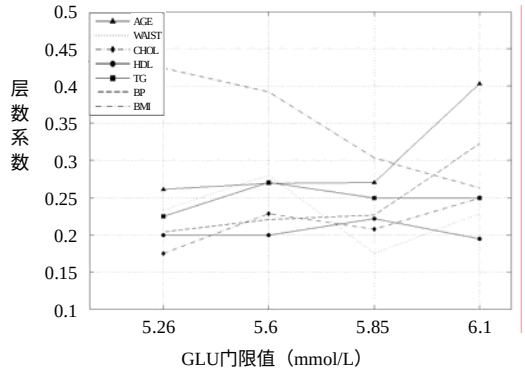

图 5 各重要属性的层数系数与门限值的变化关系

由此可以得出年龄是血糖门限值为 6.1 时的首要影响因素；门限值为 5.85、5.6 和 5.26 时，首要的影响因素为体重指数；体重指数随着门限值的降低发挥更大的影响；血压和年龄则随门限值的降低，所发挥的作用在减弱；腹围在门限值 5.6 时发挥的作用最大，在门限值 5.85 下发挥作用最小；甘油三酯、高密度脂蛋白和胆固醇的作用受门限值变化的影响不大。

2.4 重要发病危险因素与血糖门限值的变化关系

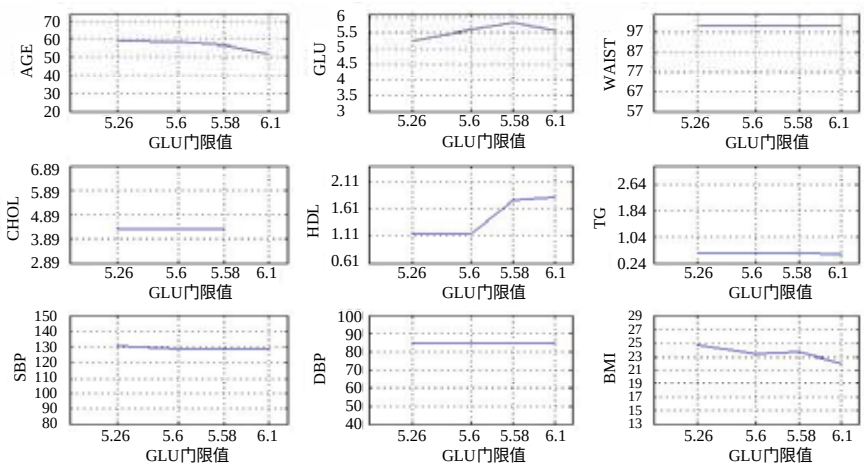

图 6 重要发病危险因素临界值与血糖门限值的变化关系

在聚类、分类组合实验中的所有分类实验结果中，根据对应血糖值为 6.1、5.85、5.6、5.26 四个门限值下，提取分类决策树前 4 层中各个重要发病影响因素的临界值的变化情况。当发病危险因素对应的临界值有多个时，使用其均值来衡量该属性的临界值与血糖门限值变化的关系。分析得到了重要发病危险因素临界值与血糖门限值的变化关系，如图 6 所示，其中 GLU 单位为 mmol/L。

由图 6 可以看出，年龄和体重指数的临界值随血糖门限值的降低有升高的趋势；高密度脂蛋白的临界值则随门限的降低而降低；血糖的临界值在 5.85 门限值下达到最大；腹围、胆固醇、甘油三酯和血压相对于门限值没有明显变化。

3 结束语

本文基于 EM 聚类和 C4.5 分类算法数据挖掘算法，通过设计大量实验和实验结果的分析，提炼得出了关于 2 型糖尿病发病的一些有价值的结论，不仅支持了关于 2 型糖尿发病的统计学领域和医学研究领域的经验知识，而且提出了一些新颖的具有更进一步研究价值的结论，内容包括 2 型糖尿病发病危险因素的重要发病危险因素的提取，重要发病危险因素的临界值属性以及重要发病危险因素与血糖的关系等。目前，2 型糖尿病的发病率在逐年上升，因而对 2 型糖尿病发病规律的研究对更好的控制和干预有着非常重要的作用。下一步的工作主要集中在以下几个方面：2 型糖尿病发病过程中各个发病危险因素之间的互相作用关系；基于高维大量 2 型糖尿病实测数据的样本状态预测技术的研究；基于高维大量 2 型糖尿病实测数据的发病规律的提取和仿真；建立疾病相关实测数据的泛化数据处理系统等。

参考文献

1 成　华. 数据挖掘在糖尿病数据中的应用研究[D]. 北京: 中国科学院软件研究所, 2003.

2 罗森林, 成　华, 张铁梅, 等. 多维 2 型糖尿病数据预处理技术[J]. 计算机工程, 2004, 30(17): 178-181.

3 王　恒. 2 型糖尿病发病危险因素与血糖变化关系的研究[D]. 北京: 北京理工大学, 2006.

4 罗森林, 成　华, 顾毓清, 等. 数据挖掘在 2 型糖尿病数据处理中的应用[J]. 计算机工程与设计, 2004, 25(11): 1888-1892.

5 罗森林, 成　华, 顾毓清, 等. C4.5 在 2 型糖尿病分类规则建立中的应用[J]. 计算机应用研究, 2004, 21(7).

（上接第 73 页）

菜单，通过此菜单可以将当前应用程序的状态保存到磁盘等存储介质上。需要注意的是，当前保存的应该只是 BeanProxy 实例和 ExtendedBeans 的状态，而不是 Beans 自身的状态。只要在程序中的构件没有发生改变，这些状态可被重新加载。

“编辑”菜单列出了当前选中构件所需要的接口和实现。用户可以通过该菜单生成接口与接口实现间的联系。列出的那些接口仅仅是提供一些信息，不作其它用途。

5 结束语

构件技术的主要目标就是通过构件使生成应用程序的过程相对简单化。构件组装技术是基于构件的软件开发的核心技术。构件必须经过组装才能形成应用系统，才能实现构件的复用价值。本文选取 Java Beans 构件模型作为研究对象，对基于 Java Beans 的组装工具 Bean Box 进行了扩展，对于如何满足扩展 Java Beans 模型进行了讨论，并给出了相应实现。

参考文献

1 杨芙清, 梅　宏, 李克勤. 软件复用与软件构件技术[J]. 电子学报, 1999, 27(2): 68-75.

2 梅　宏, 陈　锋, 冯耀东, 等. ABC: 基于体系结构、面向构件的软件开发方法[J]. 软件学报, 2003, 14(4).

3 Sun Microystem. JavaBeans (Version 1.01)[Z]. 1997-07-24. http://java.sun.com/beans/beans.101.pdf.

4 任洪敏, 钱乐秋. 构件组装及其形式化推导研究[J]. 软件学报, 2003, 14(6).