糖尿病电子病历分类算法性能研究*

杨美洁 邓 媛

（重庆医科大学医学信息学院 重庆 400016）

〔摘要〕 对糖尿病电子病历的基本信息、出入院记录和病程记录进行数据预处理，利用 Weka3.9 对处理后的数据分别进行决策树、人工神经网络、朴素贝叶斯和 K 最近邻分类，结果显示朴素贝叶斯分类法对此类数据的预测和分类更具优势，为糖尿病的分类和预测提供依据。

〔关键词〕 SQL；糖尿病；电子病历；Weka 3.9；朴素贝叶斯

〔中图分类号〕R－056 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2018.02.015

Classification Algorithm Performance Study on Diabetes Electronic Medical Records *YANG Mei－jie, DENG Yuan, Chongqing Medical Informatics College, Chongqing Medical University, Chongqing 400016, China*

〔**Abstract**〕 The paper preprocesses the data including basic information, admission and discharge record and progress note of diabetes Electronic Medical Records (EMR), implementing decision tree, Artificial Neural Network (ANN), Naive bayesian and K－Nearest Neighbor (KNN) classifications respectively on data that have been processed with Weka 3.9. The result shows that Naive bayesian classification, which is superior to the others in predicting and classifying such data, can provide basis for the classification and prediction of diabetes.

〔**Keywords**〕 SQL; Diabetes; Electronic Medical Records (EMR); Weka 3.9; Naive bayesian

1 引言

糖尿病是一系列以血糖升高为特征的代谢紊乱综合征，长期高血糖会导致心、脑、肾、足、眼睛以及周围神经病变。糖尿病已经发展为继心脑血管疾病、恶性肿瘤后的第 3 大威胁人类健康的慢性非传染性疾病[1]。电子病历是指在医疗活动过程中，医务人员使用信息系统生成的文字、符号、图表、图形、数据、影像等数字化的医疗信息资料[2]。如何从海量的电子病历数据中挖掘有价值的信息成为目前亟需解决的问题。

Jin Park 利用神经网络（Neural Network，NN）模型对健康风险评估数据进行挖掘，预测糖尿病的患病风险[3]，Barakat 等利用年龄、空腹血糖、体重指数等指标以及糖尿病家族史等数据挖掘出糖尿病的诊断规则[4]。Habibi 等收集糖尿病患者的年龄、性别、血压、家族史和体重指数等影响因素，利用 CART、C5 和 GRI 方法对糖尿病患者进行诊断[5]。Vijayanv V 等利用 EM、K 最邻近（K－Nearest Neighbor，KNN）、K－means、Amalgam KNN 和 ANFIS 5 种数据挖掘算法对糖尿病进行预测和诊断，对其性能进行比较[6]。肖文祥利用随机森林数据挖掘

〔修回日期〕 2017－11－28

〔作者简介〕 杨美洁，讲师。

〔基金项目〕 重庆市社会事业与民生保障科技创新专项（项目编号：cstc2015shms－ztzx10003）；重庆医科大学医学信息学院大学生创新实验（项目编号：2015C005）。

方法对体检数据中的空腹血糖数据进行分析，对体检患者是否患有糖尿病进行预测[7]。本研究对某医院的糖尿病电子病历进行数据预处理，利用 Weka 3.9 软件中的决策树算法 C4.5、朴素贝叶斯、人工神经网络（Artificial Neural Network，AAN）、KNN 4 种分类算法进行挖掘，从而找出最适合此资料的分类算法，为临床的疾病诊断提供决策依据。

2 资料与方法

2.1 资料来源

收集重庆市某综合医院近年主诊断为糖尿病患者的电子病历 1 433 份。

2.2 数据收集

研究糖尿病患者电子病历的基本信息、病程记录、入院记录 3 类表格。通过查阅糖尿病相关文献，找出年龄、性别、糖尿病家族史、吸烟、饮酒、高血压、高血脂、胸痛、胸闷、湿啰音、糖尿病分类 11 个属性[8-13]。分别从基本信息表中选取患者住院号以及性别、年龄、出院诊断属性；从入院记录表选取家族史 1 个属性；从入院记录个人史表中选取吸烟、饮酒 2 个属性；从入院记录既往史中选取高血压、高血脂 2 个属性；从病程记录表中选取胸痛、胸闷、湿啰音 3 个属性。

2.3 数据预处理

海量的电子病历数据结构化处理是临床数据分析的必要前提[14]，因此必须对电子病历的数据进行预处理。数据预处理包括数据清洗、集成和转换[14]。利用 SQL Server 2008 工具对数据进行处理，得到符合本研究的数据和模型。

2.3.1 基本信息数据预处理 对基本信息表中的性别和糖尿病分类进行处理，性别中的男、女分别取值 1、2；糖尿病分类的 I 和 II 型分别取值为为 1、2。其中部分代码和结果如下：

```
UPDATE 基本信息 SET 性别=1 where 性别='男'
UPDATE 基本信息 SET 性别=2 where 性别='女'
```

基本信息数据转换结果，见表 1。

表 1 基本信息数据转换结果

住院号	性别	年龄	糖尿病分类
10018802	1	71	2
……	……	……	……
12002334	2	68	2

2.3.2 入院记录数据预处理 （1）既往史选取与糖尿病密切相关的既往史指标高血压、高血脂。利用 SQL 语言编写递归函数，如记录内容中的既往史指标前面存在“无”、“不”、“没”、“未”、“否”等关键字或不存在关键词则取值为 0，代表否认该既往史[15]；否则取值为 1，代表有该既往史。既行史数据转换结果，见表 2。（2）个人史选取与糖尿病密切相关的个人史指标吸烟、饮酒等数据。当记录内容中有“无烟酒”则相应字段取值为 0；有“包”、“支”、“斤”、“瓶”等量词时，则取值为 1。处理方法与既往史相同，个人史数据转换结果，见表 3。（3）家族史处理过程与既往史相同，家族史数据转换结果，见表 4。

表 2 既往史数据转换结果

住院号	高血压	高血脂
10018802	1	0
……	……	……
12002334	0	1

表 3 个人史数据转换结果

住院号	吸烟	饮酒
10018802	0	0
……	……	……
12002334	0	0

表 4 家族史数据转换结果

住院号	糖尿病家族史
10018802	0
……	…
12002334	0

2.3.3 病程记录预处理 处理方法与既往史数据处理方法相同。病程记录数据转换结果，见表 5。

表 5 病程记录数据转换结果

住院号	胸痛	胸闷	湿啰音
10018802	0	0	0
……	……	……	……
12002334	0	0	0

2.4 数据集成

以住院号为关联字段，将基本信息表、入院记录和病程记录等 3 个表集成为 1 个表。数据集成，见表 6。

表 6 数据集成

住院号	性别	年龄	高血压	…	胸闷	湿啰音
10018802	1	71	1	…	0	0
……	……	……	……	……	……	……
12002334	2	68	0	…	0	0

2.5 数据属性资料

采用的糖尿病资料数据模型中包含 10 个自变量属性和 1 个因变量属性。糖尿病资料数据模型，见表 7。

表 7 糖尿病资料数据模型

属性	域（取值）	属性类别
性别	分男（1）、女（2）2 类	nominal
年龄	（15，93）	numeric
高血压	分有（1）、无（0）2 类	nominal
高血脂	分有（1）、无（0）2 类	nominal
糖尿病家族史	分有（1）、无（0）2 类	nominal
吸烟	分有（1）、无（0）2 类	nominal
饮酒	分有（1）、无（0）2 类	nominal
胸痛	分有（1）、无（0）2 类	nominal
胸闷	分有（1）、无（0）2 类	nominal
湿啰音	分有（1）、无（0）2 类	nominal
糖尿病分类	分为 I 型（1）、II 型（2）2 类	nominal

3 分类挖掘算法性能比较

3.1 概述

本研究采用 Weka 3.9 软件，利用决策树 C4.5、人工神经网络、朴素贝叶斯分类以及 K 最邻近分类 4 种分类算法和十折交叉验证方法（10 - fold Cross Validation）进行分析。相关评价指标采用灵敏度、特异性、精度、准确率、平均绝对误差和 AUC。

3.2 准确率

在分类问题中最常见的指标是准确率（Accuracy）[16]，表示模型预测正确的样本比例。公式如下：

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

3.3 灵敏度

灵敏度（Sensitivity）也称为召回率（Recall），指被预测为正类的样本中，其真正类别也为正类的样本所占的比例。计算公式为：

$$Sensitivity = \frac{TP}{TN + FN}$$

3.4 精度

精度（Precision）指预测为正类的样本中，真实类别为正类的样本所占的比例。计算公式为：

$$Precision = \frac{TP}{TP + FP}$$

3.5 特异度

特异度（Specificity）是在负样本中正确预测的概率，即负样本的召回率计算公式为：

$$Specificity = \frac{TN}{TN + FP}$$

3.6 平均绝对误差

平均绝对误差（Mean Absolute Error，MAE）[17]

是计算所有样本预测值与真实值差的绝对值的平均值。平均绝对误差能更好地反映模型的拟合程度。算法性能指标结果，见表 8。可看出朴素贝叶斯的性能较好，更适合此类数据的分类预测。

表 8 算法性能指标结果

算法名称	灵敏度	特异度	精度	准确度（%）	平均绝对误差
C4.5	0.985	0.015	0.971	98.52	0.029
朴素贝叶斯	0.985	0.015	0.971	98.52	0.027
ANN	0.985	0.015	0.971	98.52	0.028
KNN	0.978	0.015	0.970	97.78	0.029

3.7 AUC[18]

AUC 是评价模型平均分类性能好坏的指标。其中 AUC 表示 ROC 曲线下面积。AUC 的数值范围是[0，1]，一般认为模型分类性能越好，对应的 AUC 值越大[7]。AUC 结果，见表 9。可看出朴素贝叶斯的分类性能更好。

表 9 AUC 结果

算法	决策树 C4.5	朴素贝叶斯	ANN	KNN
AUC	0.097 7	0.670 7	0.621 3	0.587 4

3.8 性能比较

对处理后的数据采用决策树 C4.5、人工神经网络、朴素贝叶斯分类以及 K 最邻近等分类挖掘方法，对其进行评价。相关评价指标采用灵敏度、特异性、精度、准确率、平均绝对误差和 AUC。对指标的对比分析结果显示朴素贝叶斯分类法的灵敏度为 0.985，特异性为 0.015，精度为 0.971，准确率为 98.52%，平均绝对误差为 0.027，AUC 为 0.670 7，各指标均优于其他算法，更适合此类数据的分类和预测。为后期糖尿病患者的预测和诊断提供理论依据，也为其他电子病历数据的分类和预测提供方法学基础。

4 结语

本研究以糖尿病电子病历为研究对象，运用 SQL 数据库技术对电子病历中的基本信息、病程记录进行预处理，利用 Weka 3.9 中决策树 C4.5、人工神经网络、朴素贝叶斯分类以及 K 最邻近分类 4 种分类算法进行分析，挖掘出适合此类数据的算法，为糖尿病的分类和预测提供参考依据。

参考文献

1 刘力生. 中国高血压防治指南 2010 [J]. 中国医学前沿杂志（电子版），2011，3（5）：42－93.

2 李伟明. 电子病历档案应用现状及前景的探讨 [J]. 广东档案，2010，（3）：38－39.

3 Jin P, Edington D W. A Sequential Neural Network Model for Diabetes Prediction [J]. Artificial Intelligence in Medicine, 2001, 23 (3): 277－293.

4 Barakat M N, Barakat N, Diederich J, et al. Diagnosis of Diabetes Mellitus: a data mining approach [C]. International Conference of the Gulf Group for the Study of Diabetes, 2005: 42.

5 Vijayanv V, Ravikumar A. Study of Data Mining Algorithms for Prediction and Diagnosis of Diabetes Mellitus [J]. International Journal of Computer Applications, 2014, 95 (17): 12－16.

6 肖文翔. 基于电子病历分析的糖尿病患病风险数据挖掘方法研究 [D]. 青岛：青岛大学，2016.

7 方家追. 2 型糖尿病住院原因和慢性并发症患病率及其危险因素分析 [D]. 杭州：浙江大学，2015.

8 王银会. 北京市朝阳区居民高血压糖尿病及相关危险因素现况分析 [D]. 郑州：郑州大学，2015.

（下转第 77 页）

4.4 服务医疗核心，实现智慧医院

微信与医疗服务嵌合以后，有效分担医疗核心服务以外的其他配套和周边服务。如线上支付缩短排队等候的无效时间，报告查询减少患者往返医院的周折，候诊系统避免较多患者集中等候时的交叉感染风险，科普宣教提升患者的依从性。医疗微信通过科技信息的优势，减少原来非医疗核心和专业领域的服务流程和机械劳动，使医院将人力、物力和财力等资源更集中地投入医疗专业核心环节中。

4.5 把握科技脉搏，紧跟智能发展

随着科技迅猛地发展，微信将不断优化就医流程。信息技术助力建设友好医患关系也成为医院管理者和微信等信息技术运营商的努力方向。互联网资源输入背景下的医疗生态整合将是未来医院的发展方向。医院的云服务也将不再是一个遥不可及的构想，涉及支付、实名电子身份、数据处理、电子病历、远程协作、金融、周边、电商等全方位能力的 C 端服务系统已经初具雏形。医院管理者更应具备探索前沿科技的开拓精神，使信息技术发挥出提升医疗服务水平的巨大潜能。

5 结语

随着“互联网 +”概念在 2015 年政府工作报告中被正式提出，互联网医疗也逐渐成为一种新型的医疗健康服务业态。结合中国国情实际，公立医院是实施健康中国战略的主体，应当继续秉承公益性，进一步着力解决好“人民日益增长的对美好生活的需要与不平衡不充分的发展之间的矛盾”，促进优质医疗资源下沉，提高人民群众看病就医需求的可及性、体验度和获得感，提高健康科普“治未病”的积极作用。公立医院开通运营公众微信号，是突破传统医疗技术的服务范畴，与健康服务进行深度融合的有益载体，特别在非核心医疗服务中体现出效率和便捷。这种服务新模式顺应人民群众的现实需要，更是改革发展的必然趋势。

参考文献

1 谭德军. 微信公众平台在医院的应用实践研究 [J]. 现代医院，2016，16（2）：294 - 295.

2 沈良盛，王凤娟，李选治，等. “微信”在医患沟通中的作用研究 [J]. 中外医学研究，2016，14（26）：154 - 155，156.

3 姜红波，邵雪媛. 微信电子商务顾客忠诚度影响因素分析 [J]. 厦门理工学院学报，2015，23（4）：45 - 50.

（上接第 68 页）

9 乔晶. 大连市城乡居民糖尿病的现况调查 [D]. 大连：大连医科大学，2007.

10 胡傲容，郭淑霞，唐景霞，等. 社区居民糖尿病患病率及危险因素分析 [J]. 石河子大学学报（自科版），2007，25（4）：468 - 470.

11 王天歌. 中国成人糖尿病流行与控制现状及危险因素研究 [D]. 上海：上海交通大学，2014.

12 王午喜，屈宗杰，朱爱冬. 重庆市社区 10932 名普通居民糖尿病流行病学调查分析 [J]. 重庆医学，2013，（26）：3149 - 3150.

13 杨美洁，浦科学，李准. 糖尿病电子病历数据预处理 [J]. 医学信息学杂志，2016，37（5）：59 - 62，84.

14 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，2015，36（1）：58 - 62.

15 Metz C E. Basic Principles of ROC Analysis [J]. Seminars in Nuclear Medicine，1978，8（4）：283 - 298.

16 Willmott C J，Matsuura K. Advantages of the Mean Absolute Error（MAE）over the Root Mean Square Error（RMSE）in Assessing Average Model Performance [J]. Climate Research，2005，30（1）：79 - 82.

17 Fawcett T. An Introduction to ROC Analysis [J]. Pattern Recognition Letters，2006，27（8）：861 - 874.