文章编号：2095-4980(2020)04-0703-05

基于决策树的听力损失致病基因筛查方法

郭　宇，李凤美，陈雨行，洪凯程，赵也明，陈晓禾

(中国科学院 苏州生物医学工程技术研究所，江苏 苏州 215163)

摘　要：确定出生缺陷高危致病基因类型，推进遗传性疾病早期筛查和生育指导，对于先天性听力损失等出生缺陷的一级预防具有重要意义。本文采用通用数据挖掘工具，应用其决策树算法分析了近千例 GJB2 基因突变检测的临床数据，建立了听力出生缺陷的致病基因辅助筛查模型。通过研究模型树的结构和样本分类结果，发现模型树中有 5 组分支获得了纯净的听力损失阳性样本。此外，每个分支构成的基因位点的状态集合与临床研究证实的致病基因突变状态相一致。该决策树方法建立的筛查模型可以协助医生从临床大数据中快速筛选出致病基因的类型。

关键词：临床大数据；听力损失；决策树；基因筛查；数据挖掘；人工智能

中图分类号：TN911.7；TP181　　**文献标志码：**A　　**doi：**10.11805/TKYDA2018378

A method of genetic disease screening for hearing loss based on decision tree

GUO Yu，LI Fengmei，CHEN Yuhang，HONG Kaicheng，ZHAO Yeming，CHEN Xiaohe

(Suzhou Institute of Biomedical Engineering and Technology，Chinese Academy of Sciences，Suzhou Jiangsu 215163，China)

Abstract：It is important to determine the genetic types of high risk of birth defects and to promote early heredopathia screening and birth guidance for first class prevention of birth defect like congenital hearing loss. This research applies the decision tree algorithm to analyze the clinical data of nearly 1 000 cases of GJB2 genetic mutation record by using data mining tools. The model of assistant screening for pathogenic genes of hearing loss birth defects is established. By studying the structure of the model tree and samples' classification results, it is concluded that pure positive samples of hearing loss are obtained in five groups of branches in the model tree. Besides, the set of genetic loci states formed by each of these branches is consistent with the mutation state of pathogenic genes confirmed by clinical studies. The screening model established by the decision tree method can assist doctors to quickly screen out the types of pathogenic genes from the clinical big data.

Keywords：clinical big data；hearing loss；decision tree；genetic screening；data mining；artificial intelligence

2012 年 9 月 12 日，卫生部在北京发布了《中国出生缺陷防治报告(2012)》。报告指出，我国是出生缺陷高发国家。根据世界卫生组织估计，我国出生缺陷发生率与世界中等收入国家的平均水平接近，约为 5.6%，每年新增出生缺陷数约 90 万例[1]。其中，遗传性耳聋是最常见的遗传疾病之一，是影响我国出生人口素质的重要问题[2]。2006 年第二次残疾人抽样调查显示：中国目前听力障碍人群占残疾人总数的 1/3，并以每年 3 万聋儿的速度在增长，在耳聋人群中 60%与遗传因素有关[3]。因此，推进耳聋出生缺陷高危致病基因筛查方法研究，提升耳聋出生缺陷一级预防技术水平，对降低遗传性耳聋的发生率具有重要意义。

机器学习算法在数据挖掘和建立预测模型方面已有许多成熟的应用。常见的算法有：决策树、神经网络、支持向量机、贝叶斯分类和深度学习等。大多数复杂的机器学习算法由于计算过程晦涩难懂，无法展示或阐明处理过程的场景，因此并不适合用于基因筛查诊断的专家决策系统[4]。决策树算法具有结构简单，模型效率和分类精

收稿日期：2018-12-12；**修回日期：**2019-02-14

基金项目：科技部重点研发计划重点专项资助项目(2017YFC1001800)

作者简介：郭　宇(1989–)，男，在读博士研究生，助理研究员，主要研究方向为信号及电源完整性、电磁兼容与机器学习。email：guoyu@sibet.ac.cn

确度较高，结果直观，便于解析等优点[5]。该方法用于出生缺陷致病基因类型筛查可以直观地在模型树状结构中看到模型预测分类的过程，解析出高危致病基因的组成，为进一步分析研究提供了方便[6-10]。因此，本文选用决策树算法对搜集到的基因检测样本数据建立并分析相应的听力损失出生缺陷基因筛查模型。

1 数据准备

1.1 原始数据

研究的数据来自国家重点研发计划项目——《出生缺陷一级预防孕前检测技术设备及应用平台的研发》的牵头单位，上海交通大学医学院附属第九人民医院收集整理的临床基因检测及病例诊断数据。用于研究的数据属性列包括就诊人员的年龄、性别、基因检测结果和临床诊断结果等，并隐去了所有病例的个人信息。基因检测结果的内容以 GJB2 基因不同位点的突变信息为主，其中还有极少数病例样本记录了患者的线粒体和 SLC26A4 基因的突变信息。基因位点的突变信息主要分为未发生突变、纯合突变和杂合突变等 3 种类型。同一名患者的基因检测结果可能包括多个位点的突变。临床诊断结果是医生记录的对就诊人员听力情况的临床诊断，包括听力正常和多种听力损失的临床诊断。由于原始数据的内容大多采用文字记录，且存在数据丢失和表达不一致等情况，数据挖掘软件难以进行分析处理，因此必须对原始数据进行数据清洗、特征提取等预处理。

1.2 数据预处理

首先对数据进行净化处理。通过初步的数据统计，原始数据中线粒体和 SLC26A4 基因检测数据存在严重的信息缺失，包含有效信息的数据量不足 100 例。相比之下，记录 GJB2 基因突变信息的样本数超过 700 例。因此，本文只筛选出 GJB2 基因不同位点的突变信息属性作为主要研究筛查对象。此外，在临床诊断数据记录方面，原始数据也存在信息缺失的情况。部分样本缺少了临床诊断数据，使得决策树模型建立和验证过程缺少了关键的样本分类依据。因此，在数据净化过程中将这类样本删除。通过数据净化处理后，从原始数据中筛选出 984 例样本，样本包含的信息包括：临床诊断结果、GJB2 基因中位点突变检测结果和性别。

其次对数据进行特征提取。文本数据的特征提取本质是将数据转变成便于计算机分析和处理的形式。将原始数据中临床诊断数据简化成 2 种类型：听力正常和听力缺损，符号“N”表示听力正常；“H”表示听力缺损。由此可以将研究简化为二分类问题，便于决策树算法的分析处理。将原有的 GJB2 基因突变信息展开成不同位点的突变状态：“0”表示该位点没有突变；“1”表示该位点是纯合突变；“2”表示该位点是杂合突变。基因属性细分后，该数据集的每个样本共有 37 个 GJB2 基因位点的突变信息。对于性别属性，符号“M”表示男性；“F”表示女性。其他与基因无关或严重缺失的属性进行删除，预处理后的数据形式如表 1 所示。

表 1 预处理后的数据形式

Table1 Data form after preprocessing

No.	c.235delC	V271	…	V371	E114G	sex	class
1	1	0	…	0	0	F	H
2	2	0	…	0	0	M	N
3	0	2	…	0	2	M	H
4	0	2	…	2	2	M	H
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

2 数据挖掘

2.1 工具选择

目前国内外主流的开源数据挖掘软件有 WEKA,KNIME,Orange 和 RapidMiner 等。其中，RapidMiner 是一款世界领先的集数据挖掘、预测分析和商业智能为一体的软件工具，已成功用在许多不同的领域，包括文本挖掘、多媒体挖掘、功能设计、数据流挖掘、集成开发方法和分布式数据挖掘等[11]。该软件不仅具有丰富且强大的算法库以及更加多样全面的数据分析可视化效果，而且具备良好的可编程性和可扩展性[12]。RapidMiner 集成的决策树算法相对其他数据挖掘软件，更为全面，性能领先，可对名义和数值型属性的数据进行多种智能挖掘和分析操作。

2.2 挖掘步骤

建立预测分析模型，首先进行数据加载。数据加载过程中重要的一步是修改数据的类型，如图 1 所示。选择合适的数据类型将极大提升数据挖掘的效果。本文加载的数据集中除了年龄类别的数据为“real”类型，即实数型，其他类别的数据都是“polynominal”类型，即名义型。数据加载完成后，进入数据挖掘和模型建立环节。应用软件自带的算法库进行数据挖掘，主要步骤有：选择数据、选定任务、准备指标、选择输入量、选择算法和结果展示等，如图 2 所示。

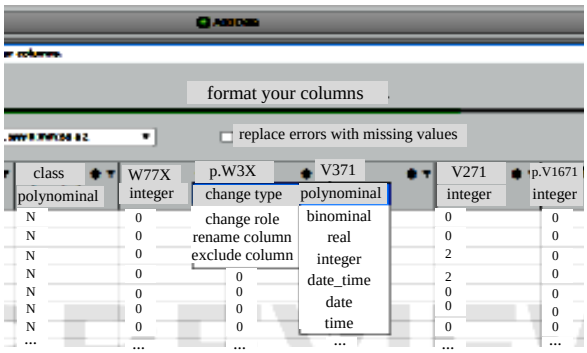

Fig.1 Property settings for loaded data
图 1　加载数据的属性设置

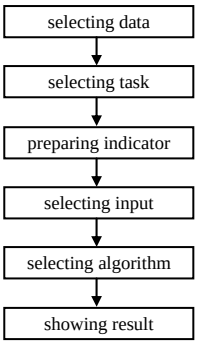

Fig.2 Data mining and modeling processes
图 2　数据挖掘与建模流程

首先，在已加载的数据中选取进行数据挖掘的数据集，即选定上述预处理后的数据集。数据集选定后，需要明确数据挖掘的任务类型。通用数据挖掘工具 RapidMiner 中共有 3 种数据挖掘任务类型：类别预测、数据聚类和异常值检测。本文进行数据挖掘的目标是建立耳聋出生缺陷致病基因的筛查模型，再通过模型智能分析基因检测数据，预测是否出现耳聋出生缺陷。因此，本文选用了数据类别预测的任务模式。该模式选定后，需进一步确定需预测的数据属性列，即选定数据集中临床诊断属性列为需预测类别的样本属性。在随后的准备指标环节中，可以选定模型建立时更关心的属性类别。本文将临床诊断为听力缺损，即标记为“H”的类别选定为预测模型建立时更关注的属性类别，以期提高模型的致病基因检出率。接下来是选择输入量，即筛选用于数据挖掘和模型建立的属性列。本文旨在研究基于基因检测结果建立预测耳聋出生缺陷的模型，因此选择了数据量充足的 GJB2 基因的 37 个位点突变属性和性别属性作为预测模型的输入量，其中性别属性也是一种与基因相关的属性。虽然，RapidMiner 的预测分类任务的算法库中集成了贝叶斯、深度学习、决策树和随机森林等主流算法。但本文建立的预测模型用于出生缺陷基因筛查，相当于一种专家系统；同时决策树模型具有简单直观、便于理解和验证的优势，因此本文选用数据挖掘软件中集成的决策树算法进行预测模型建立。

3　结果分析

数据挖掘和建模流程完成后，RapidMiner 可以快速生成各种可视化的数据分析结果。软件中数据相关性分析结果显示 c.235delC 基因位点的突变情况对预测样本为听力缺损，即“H”类型的重要性最强，如图 3 所示。图中深灰色长条表示相应属性类别对判定样本为“H”类型的重要性，白色长条表示该属性类别对排除样本为“H”类型的重要性。自上而下，各属性类的重要性依次降低。通过查阅文献可知，c.235delC 是亚裔人群最常见的突变位点[13]。这一统计结果和本文数据相关性分析结果相吻合。表 2 为权重值排名靠前的属性值。权重值越大，该属性在决策树模型中的分类作用越明显。因此，这些属性大多成为决策树模型的重要组成节点[14]。

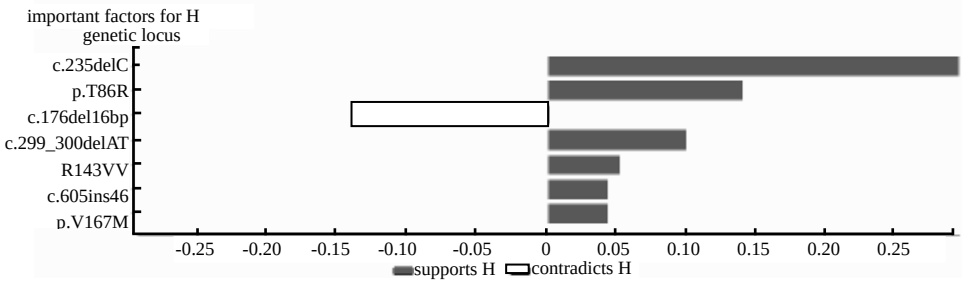

Fig.3 Importance of attribute classes to assistant screening models
图 3　属性类对辅助筛查模型的重要性

表 3 为数据挖掘软件建立的不同深度的决策树模型对应的分类表现，即样本属性预测的准确度。其中，模型深度为 5 层时，模型的准确度最高，分类表现最佳。此外，准确度相近时，决策树模型的深度较低，有助于提升模型的泛化性能。因此，软件最终采用了深度为 5 层的决策树模型作为建模的最终输出。

表 2 属性权重
Table2 Attribute value weight

attribute	weight
c.235delC=2	1.000
E114G=2	0.337
V371=2	0.320
c.507_510insAACG=0	0.298
V271=0	0.288
c.176del16bp=0	0.288
V271=2	0.269
c.176del16bp=2	0.268

表 3 不同深度模型的分类表现
Table3 Classification performance of models with different depths

maximal depth	performance
2	0.649
3	0.676
5	0.704
7	0.703
1	0.703
15	0.703

应用决策树算法对输入的基因检测数据集进行分析挖掘，建立了针对 GJB2 基因的耳聋出生缺陷致病基因筛查模型，如图 4 所示。图中叶节点上的字母(“H”或“N”)表示模型预测的样本临床诊断的属性值。红色条代表实际样本类别为“H”，即听力损失阳性样本；蓝色代表实际样本类别为“N”，即听力正常样本。从模型树的分支结构中可以看出，GJB2 基因的位点出现 c.235delC 为纯合突变。c.235delC 的杂合突变与 c.176del16bp, c.299_300delAT 或 c.507_510insAACG 的杂合突变同时出现，以及 c.176del16bp 和 c.299_300delAT 的杂合突变同时出现时，分类所得的样本都为听力损失阳性样本，纯净度为 100%。这表明上述基因位点突变的组合与听力损失的临床表现具有强关联性。这一结果也与现有的医学临床研究相一致[15]，同时印证了决策树建模分析和筛查出生缺陷致病基因的有效性。该模型可以快速处理 GJB2 的检测结果，辅助医生筛选出存在较高致病风险的病例样本。同时，从模型树的分类结果可以看出，其他分支没有获得较纯净的分类结果。现有的医学研究已经证明，先天性听力缺损的致病原因不只是 GJB2 基因位点的特定突变组合，还与 SLC26A4 基因和线粒体基因等其他因素有关，由此可以解释本文中的决策树模型的其他分支不能得到较为纯净的分类结果。

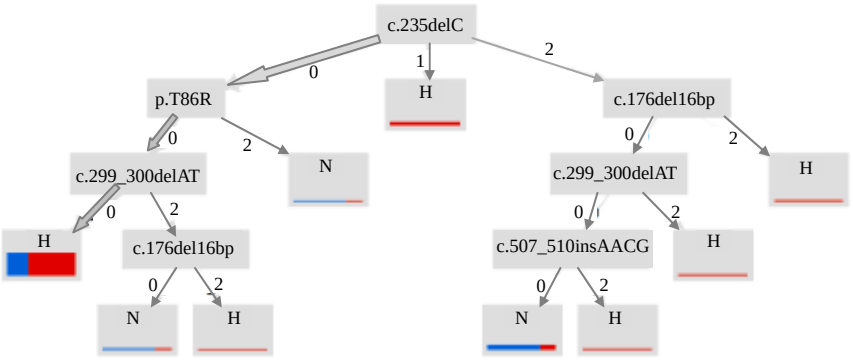

Fig.4 Decision tree model for screening deafness and birth defects
图 4 耳聋出生缺陷筛查决策树模型

4 结论

本文通过对一组临床统计的基因检测数据进行数据净化、特征提取等预处理，将原始数据转化成便于数据挖掘软件分析处理的数据集。经过前期调研，针对辅助医学研究和诊断的建模目标，本文采用了决策树算法挖掘数据信息，以建立可进一步分析和验证的分类预测模型。应用可视化的数据挖掘软件 RapidMiner 完成了数据的分析和挖掘，建立了耳聋出生缺陷筛查决策树模型。通过分析模型树的纯净分类的分支结果和属性类相关性计算结果，发现与现有医学研究相一致。该决策树方法建立的筛查模型可以协助医生从基因检测和临床诊断的医疗大数据集中快速筛选出遗传性疾病的致病基因类型。本文通过数据挖掘找到了最重要的亚裔耳聋出生缺陷的致病基因位点 c.235delC，以及几种致病的基因突变组合类型。由于耳聋出生缺陷可由多种基因突变造成，本文研究的数据只包含较多的 GJB2 基因突变信息，因此不能获得较高的分类准确度。今后通过采集更全面的样本信息合并其他临床表征数据，应用决策树算法将能够获得更加精确的耳聋出生缺陷致病基因辅助筛查模型。

参考文献：

[1] 中华人民共和国国家卫生和计划生育委员会. 卫生部发布《中国出生缺陷防治报告(2012)》[EB/OL]. (2012-09-12)[2018-12-12]. http://www.moh.gov.cn/wsb/pxwfb/201209/55840.shtml. (Ministry of Health of the People's Republic of China. The ministry of health released the report on prevention and treatment of birth defects in China(2012)[EB/OL].

(2012-09-12) [2018-12-12]. http://www.gov.cn/gzdt/2012-09/12/content_2223371.htm.)

[2] 袁慧军,曹菊阳,孙捍军,等. 一个常染色体显性遗传非综合征性耳聋巨大家系调查[J]. 解放军医学杂志, 2003,28(8):703-704. (YUAN Huijun,CAO Juyang,SUN Hanjun,et al. Investigation of a huge family with autosomal dominant hereditary non-syndromic hearing loss[J]. Medical Journal of Chinese People's Liberation Army, 2003,28(8):703-704.)

[3] 王艳丽,张颖. 遗传性耳聋常见致病基因及筛查方法的研究进展[J]. 国际生殖健康/计划生育杂志, 2013(4):309-312. (WANG Yanli,ZHANG Ying. Common morbid genes of hereditary deafness and screening methods[J]. Journal of International Reproductive Health/Family Planning, 2013(4):309-312.)

[4] 张椋,曹健. 面向大数据分析的决策树算法[J]. 计算机科学, 2016(z1):374-379,383. (ZHANG Yan,CAO Jian. Decision tree algorithms for big data analysis[J]. Computer Science, 2016(z1):374-379,383.)

[5] 季桂树,陈沛玲,宋航. 决策树分类算法研究综述[J]. 科技广场, 2007(1):9-12. (JI Guishu,CHEN Peiling,SONG Hang. Study the survey into the decision tree classification algorithms rule[J]. Science Mosaic, 2007(1):9-12.)

[6] YOKOYAMA J S,BONHAM L W,SEARS R L,et al. Decision tree analysis of genetic risk for clinically heterogeneous Alzheimer's disease[J]. BMC Neurology, 2015,15(1):47.

[7] KOKOl P,POHOREC S,PODGORELEC V. Evolutionary design of decision trees for medical application[J]. Data Mining & Knowledge Discovery, 2013,2(3):237-254.

[8] ALBU A. From logical inference to decision trees in medical diagnosis[C]// 2017 E-Health and Bioengineering Conference. Sinaia,Romania:IEEE, 2017:65-68.

[9] LIU Q,XU X,TAO Y,et al. An improved decision tree method based on RELIEFF for medical diagnosis[C]// 2016 6th International Conference on Digital Home(ICDH). Guangzhou,China:IEEE, 2016:133-138.

[10] HEIKKILÄ P,FORMA L,KORPPI M. High-flow oxygen therapy is more cost-effective for bronchiolitis than standard treatment—a decision tree analysis[J]. Pediatric Pulmonology, 2016,51(12):1393-1402.

[11] 杨振瑜,王效岳,白如江. 国外主要可视化数据挖掘开源软件的比较分析研究[J]. 图书馆理论与实践, 2013(5):89-93. (YANG Zhenyu,WANG Xiaoyue,BAI Rujiang. Comparative analysis and research on the main abroad visual data mining open source software[J]. Library Theory and Practice, 2013(5):89-93.)

[12] 郑茹菁,王晓晔,柴晓瑞,等. 数据挖掘开源平台性能分析[J]. 天津理工大学学报, 2015,31(4):33-38. (ZHENG Rujing, WANG Xiaoye,CHAI Xiaorui,et al. Analysis of data mining open platform[J]. Journal of Tianjin University of Technology, 2015,31(4):33-38.)

[13] 崔庆佳,黄丽辉. GJB2 基因突变与听力损失的关系[J]. 临床耳鼻咽喉头颈外科杂志, 2013(19):1099-1102. (CUI Qingjia, HUANG Lihui. Hearing loss associated with GJB2 gene mutation[J]. Journal of Clinical Otorhinolaryngology Head and Neck Surgery, 2013(19):1099-1102.)

[14] GARCIA-MAGARINO I,GRAY G,LACUESTA R,et al. Survivability strategies for emerging wireless networks with data mining techniques:a case study with NetLogo and RapidMiner[J]. IEEE Access, 2018(99):1.

[15] 崔庆佳,王国建,张媛,等. GJB2,SLC26A4 基因相关耳聋儿童的听力损失特点分析[J]. 听力学及言语疾病杂志, 2014(2):120-123. (CUI Qingjia,WANG Guojian,ZHANG Yuan,et al. Audiological characteristics in 832 deaf children with biallelic causative mutations in GJB2,SLC26A4 gene[J]. Journal of Audiology and Speech Pathology, 2014(2):120-123.)