

关联规则算法在医疗大数据中的应用探索

李 强, 陈东涛, 罗先录

(广东东软学院, 广东 佛山 528225)

摘 要: 随着中国互联网越来越快的发展, 互联网+、大数据和人工智能技术等最新科技技术也越来越多的渗透到医疗领域, 对于提升多种疾病的筛查和诊断效率作用明显, 而大数据处理技术上仍然面对着许多难题。本课题针对智慧医疗发展的难点之一, 将大数据处理技术和医疗海量数据相结合, 应用Apriori关联规则算法在医疗信息系统中海量医疗数据中的应用, 发现疾病与其他疾病之间的联系, 从而为人们的健康预警及医疗机构对疾病的诊断提供科学依据参考, 本文实验数据以心肌炎和胃癌为例, 发现与病症相关性强的病因。

关键词: 智慧医疗; 数据挖掘; 关联规则; 医疗大数据

中图分类号: TP301 **文献标识码:** A

Implementation of the Association Rule Algorithm in Medical Big Data

LI Qiang, CHEN Dongtao, LUO Xianlu

(Guangdong Neusoft University, Foshan 528225, China)

Abstract: With the rapid development of Internet in China, Internet+, big data, artificial intelligence and other latest technologies are increasingly involved in the medical field, promoting the screening and diagnosis of various diseases. However, there are still many challenges in big data processing techniques. Aimed at the development of smart medical care, this paper combines big data processing technology with massive medical data and adopts the Apriori algorithm in massive medical data in the medical information system, to identify the relationship between one disease and others. Therefore, it provides some scientific evidence and reference for health warning and disease diagnosis in medical institutions. The study is conducted with the experimental data of myocarditis and gastric cancer.

Keywords: smart medical care; data mining; association rules; medical big data

1 引言(Introduction)

随着物联网、云计算和大数据等各项新技术的高速发展, 国家颁布了各项政策以促进医疗服务的发展。促进医疗信息平台的转变, 以三项技术为核心, 以患者数据为连接点, 将医疗信息平台逐步向专业的、便于医患使用的智慧医疗信息平台转变。智慧医疗这个新兴的产业已经引起了政府、集团和许多公司的关注, 其中代表有推想科技的推想人工智能致力于进行肺癌辅助筛查, 腾讯医疗人工智能实验室推出了一项帕金森病AI辅助诊断新技术等。数据表明人工智能针对一些疾病的诊断效果已经达到甚至超越了传统的人工治疗方案。但至今仍未形成成熟的产业链, 其中, 医疗物联网已逐步走向产业化, 而大数据处理技术上仍然面对着许多

难题^[1]。

本课题就智慧医疗发展的难点, 基于Apriori关联规则算法在医疗信息系统中海量医疗数据中的应用, 研究疾病与疾病之间的关联, 为个人健康提供预警和为医疗诊断提供科学依据参考。

2 关联规则算法(The apriori algorithm)

Apriori算法是一种最有影响的挖掘布尔关联规则频繁项集的算法, 其核心是基于两阶段频集思想的递推算法。该关联规则在分类上属于单维、单层、布尔关联规则。在这里, 所有支持度大于最小支持度的项集称为频繁项集, 简称频集^[2]。

关联规则一般主要分为两个过程:

(1) 基于数据产生频繁项集, 对于每一个频繁项集来说,

该项集在数据集中所出现的频率必须满足一定的要求。项集在数据集中出现频率称之为支持度，所需满足的要求称之为最小支持度。支持度可以表示关联规则是否普遍适用，通过设置最小支持度来使最终的关联规则拥有更广的适用面，也使最终获得的结果更具有价值^[3]。

(2)产生关联规则，通过所发现的频繁项集产生规则，计算每一个规则的置信度，规则的置信度若满足最小置信度，那么这条规则就称之为关联规则。置信度的大小表示使用该关联规则对数据进行推理的准确性。

其中规则的支持度的数学表达式为：

$\text{support}(A \Rightarrow B) = P(A \cup B) = \text{supportCount}(A \cup B) / \text{count}(D)$ ，其中A和B表示不相交的项集，D表示数据集， $\text{supportCount}(A \cup B)$ 表示A项集和B项集的并集在数据集中出现的次数， $\text{count}(D)$ 表示数据集的总事务数。置信度的数学表达式为： $\text{confidence}(A \Rightarrow B) = P(A|B) = \text{supportCount}(A \cup B) / \text{supportCount}(A)$ 。如果置信度为100%，意味着在数据集中该规则总是准确的。

最小支持度和最小置信度的阈值一般由用户或者专家进行设定，满足这两个阈值的规则称为强规则。

例如：年龄 ≥ 45 岁 $\Rightarrow$ 糖尿病(支持度7%置信度60%)

则表示在所有患者的诊疗信息中有7%的患者年龄大于等于45岁并且患有糖尿病，其中年龄大于等于45岁的患者中有60%的患者患有糖尿病。由此可见，如果针对医疗信息系统中大量的医疗数据使用关联规则算法进行挖掘，并能得出一些有趣的规则，那么将对医疗机构关于各种疾病的决策方面有着很大的帮助。

Apriori算法的基本思想是：算法扫描一次数据集，计算每一个项的支持度，找出所有的频繁1-项集L1，基于L1，产生所有可能频繁2-项集C2，也就是候选2-项集，再基于C2统计支持度，找出频繁2-项集L2，如此反复循环，直至发现所有的频繁项集。

候选项集的产生中使用的方法主要是 $F_{(k-1)} \times F_{(k-1)}$ 方法：通过合并两个上一步获得的频繁(k-1)-项集来生成候选k-项集。令 $F_{(k-1)} = \{C_1, C_2, C_3, \dots, C_{(k-1)}\}$ 和 $L_{(k-1)} = \{D_1, D_2, D_3, \dots, D_{(k-1)}\}$ 为两个频繁(k-1)-项集，当它们的前k-2项相同，且k-1项不同，则合并两项，获得候选k-项集，这个方法能有效的避免产生重复的候选项集问题，以及能确保该方法生成的候选k-项集的其余k-2个子集均为频繁的^[4]。

3 应用与实现(Application and implementation)

3.1 数据的准备和预处理

实际的数据库极易受噪声、缺失值和不一致数据的侵扰，因为数据库太大，并且多半来自多个异种数据源。低质量的数

据将会导致低质量的挖掘结果，所以首先需要对原始数据进行数据预处理^[5]。主要根据数据分析的任务选择任务所需的数据对象和属性，以及对数据进行数据清洗等。本次研究将对系统中门诊就诊基础数据表，住院就诊基础数据表中的数据列，如身份证、疾病代码进行处理，提取所需的数据列数据。数据预处理的过程主要有：

(1)去重处理：为了提高数据分析结果的价值，将数据中完全重复的记录去掉，属于去噪的一种。

(2)异常值处理：检测和处理异常值，可以自行确定一个异常值的标准和方法。如：为避免由于错误的身份证信息而导致不同病人之间的疾病历史混合在同一个记录中，根据一代身份证位数为15位数，以及二代身份证位数为18位数的特征^[6]，删除数据中所有身份证不符合位数要求的数据，并将所有身份证为空的数据删除；创建区域与身份证前两位特征码的对应表，删除不符合身份代码对照表的身份证的记录。

(3)特征提取：根据身份证号码中的性别特征码的奇偶性对身份证属性进行特征提取，创建病人的性别属性，如图1所示。

ID	SFZ	DMS	XB
1	025001	15033 J00.03	MALE
2	0300011	03160060 R10.401	FEMALE
3	0300011	12080047 J00.03	FEMALE
4	0310018	15019 R10.401	MALE
5	045070	08111824 J00.03	FEMALE
6	052272	40928883 J00.03,J06.903	FEMALE
7	052272	91016868 J00.03,J06.903	FEMALE
8	100241	08145920 R10.401	FEMALE
9	110101	01124511 J06.903	MALE
10	1101142	08157712 J00.03,J06.903,J40.03,K30.02	MALE
11	1102211	07207748 J06.903,K29.502,K30.02,R10.401	FEMALE
12	1111111	11111111 J00.03,J06.903,K30.02	MALE
13	1201071	504152119 J06.903,J40.03,R10.401	MALE

图1 门诊就诊数据

Fig.1 Clinic data

(4)降维处理：创建concat_array存储函数，以便之后对疾病代码列进行行转列操作，降低数据的维度，减轻算法的负担^[7]。函数代码如图2所示。

```
SQL> CREATE OR REPLACE FUNCTION concat_array(p tab_varchar2) RETURN CLOB IS
2   l_result CLOB;
3   BEGIN
4     FOR cc IN (SELECT column_value FROM TABLE(p) ORDER BY column_value) LOOP
5       l_result := l_result || ' ' || cc.column_value;
6     END LOOP;
7     return l_result;
8   END;
9 /
```

图2 存储参数代码

Fig.2 Storage parameter code

3.2 数据挖掘

本文基于Python语言作为编程语言实现关联规则算法的应用。Python语言的语法清晰，易于操作纯文本文件，并且有着丰富的第三方库，包括numpy、scipy、matplotlib等，使用python语言使本课题关联规则算法的实现更加快捷简明，并使用了pycharm和anaconda工具进行开发，利用这两款工具有效的简化了本课题中的开发工作流程^[8,9]。整个程序

主要流程有:

(1)输入疾病名称,模糊查询确认对应的疾病代码,输入最小支持度和最小置信度。

(2)对所有患有或曾患有胃癌疾病的患者的记录根据疾病代码进行聚合,统计此类患者的各种疾病的记录数(count(dm)),降序排列,获取前十条记录,即获取此类患者患病较多的前十种疾病,并根据这十种疾病筛选数据,将数据根据身份证进行行转列。

(3)根据性别分别统计每种疾病的患病人数,以及输入的疾病名称的地区分布人数,对其进行可视化。

(4)利用Apriori算法对数据进行数据挖掘。

3.3 挖掘结果

程序以疾病名称和代码作为输入,并输出与之关联的计算结果。

例如:运行程序,输入心肌梗死,输入对应的疾病代码编号20,并设置最小支持度为0.001和最小置信度为0.6,进行数据挖掘和可视化。最终得到关联规则715条,其中的部分挖掘结果如图3所示。

```
frocnaset(['151.403']) -> frocnaset(['016.903']) conf: 0.71685597244556 | supp: 0.805844004037363
frocnaset(['125.101']) -> frocnaset(['016.903']) conf: 0.6828421451021046 | supp: 0.8146767955827325
frocnaset(['1810.401']) -> frocnaset(['016.903']) conf: 0.603497512079366 | supp: 0.11964446909116515
frocnaset(['929.703']) -> frocnaset(['016.903']) conf: 0.673829828755255 | supp: 0.220291511949703
frocnaset(['1854.562']) -> frocnaset(['016.903']) conf: 0.62838691471566 | supp: 0.976940432591305
frocnaset(['016.903']) -> frocnaset(['016.903']) conf: 0.647188089697979 | supp: 0.341363454646255
frocnaset(['1540.03']) -> frocnaset(['016.903']) conf: 0.7154435744201931 | supp: 0.2845275015120458
frocnaset(['1813.991']) -> frocnaset(['016.903']) conf: 0.68919446931595 | supp: 0.49137114667554056
frocnaset(['016.903']) -> frocnaset(['016.903']) conf: 0.656262191283323 | supp: 0.32045511963045733
frocnaset(['125.101', '929.703', '1854.562']) -> frocnaset(['016.903', '016.903']) conf: 0.6404712880347747 | supp: 0.01265489957600346
frocnaset(['125.101', '1813.991', '929.703']) -> frocnaset(['016.903', '016.903']) conf: 0.6144087525570771 | supp: 0.030774402486415017
frocnaset(['1540.03', '125.101', '929.703']) -> frocnaset(['016.903', '016.903']) conf: 0.620026741542888 | supp: 0.30439748555449216
frocnaset(['125.101', '1813.991', '929.703']) -> frocnaset(['016.903', '016.903']) conf: 0.651762031072376 | supp: 0.30326231330965129
frocnaset(['125.101', '1813.991', '1810.401']) -> frocnaset(['016.903', '016.903']) conf: 0.609001141440186 | supp: 0.30196240549285523
frocnaset(['125.101', '1813.991', '1854.562']) -> frocnaset(['016.903', '016.903']) conf: 0.616124028759951 | supp: 0.0121298979775408
frocnaset(['125.101', '1813.991', '1854.562']) -> frocnaset(['016.903', '016.903']) conf: 0.632444433801106 | supp: 0.01224282447918234
frocnaset(['1540.03', '125.101', '1854.562']) -> frocnaset(['016.903', '016.903']) conf: 0.634397207220263 | supp: 0.30283498233446025
frocnaset(['1540.03', '125.101', '1813.991']) -> frocnaset(['016.903', '016.903']) conf: 0.630429151959244 | supp: 0.30144134515458779
frocnaset(['125.101', '1813.991', '1810.401']) -> frocnaset(['016.903', '016.903']) conf: 0.6162815940642718 | supp: 0.0121288138020781
frocnaset(['1540.03', '1813.991', '929.703']) -> frocnaset(['016.903', '016.903']) conf: 0.6212754565677161 | supp: 0.311414782720724316
frocnaset(['1540.03', '929.703', '1854.562']) -> frocnaset(['016.903', '016.903']) conf: 0.621244427420708 | supp: 0.311913191383674
frocnaset(['1813.991', '929.703', '1854.562']) -> frocnaset(['016.903', '016.903']) conf: 0.61642275952706 | supp: 0.3102564474164105
frocnaset(['1540.03', '1813.991', '1854.562']) -> frocnaset(['016.903', '016.903']) conf: 0.6041610901502974 | supp: 0.311331359011804
frocnaset(['125.101', '1813.991', '1854.562', '929.703']) -> frocnaset(['016.903', '016.903']) conf: 0.649069462647443 | supp: 0.0145239440344853
frocnaset(['125.101', '1813.991', '1854.562', '929.703']) -> frocnaset(['016.903', '016.903']) conf: 0.614897335080212 | supp: 0.0113475947977778
frocnaset(['125.101', '929.703', '1854.562']) -> frocnaset(['016.903', '016.903']) conf: 0.6565034965034965 | supp: 0.0111286162003024
```

图3 心肌梗死关联规则结果

Fig.3 Associated rules with myocarditis

由于胃癌患者的数量较小,为获得与胃癌相关的关联规则,经过多次对支持度与置信度的调整,最终设置支持度为0.00009,置信度为0.5。最终得到关联规则715条,从中筛选与胃癌相关的关联规则如图4所示。

1	胃癌	->	上呼吸道感染	置信度: 0.5801749271137026 支持度: 9.49268061852208e-05
2	胃癌	->	男性	置信度: 0.5860058309037901 支持度: 9.58808444383386e-05
3	胃癌	->	胃炎	置信度: 0.6793002915451894 支持度: 0.0001111454546882233

图4 胃癌关联规则结果

Fig.4 Associated rules with gastric cancer

3.4 数据可视化

本文分别对男性与女性在胃癌相关的10项疾病的患病人数进行统计,根据数据调用pyecharts绘制柱状图和区域分布饼图,其中柱状图的数据包括疾病的名称,患各疾病的男性人数,患各疾病的女性人数,饼图的数据包括分析的疾病在各区域的人数。最终得到胃癌相关的各疾病男女性患病人数统计条形图,以及患病人群区域分布图如图5和图6所示。

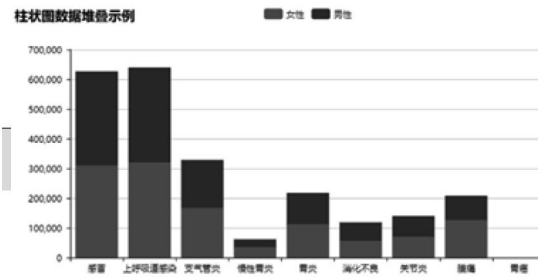


图5 胃癌相关的人数统计

Fig.5 Statistics of gastric cancer related patients

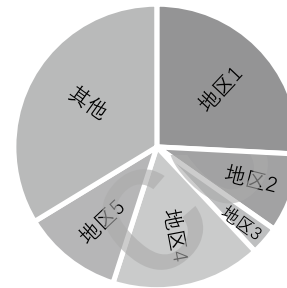


图6 胃癌人数区域饼图

Fig.6 Gastric cancer population area pie chart

3.5 数据分析

针对胃癌相关的关联规则结果进行分析可以得出以下结论:

规则第一条:患胃癌的患者性别为男性的置信度约为58.6%。该关联规则较为符合现今的医疗相关研究及实际医疗诊断情况:男性患胃癌的几率比女性高。

规则第二条:患胃癌的患者患胃炎的置信度约为67.9%。

胃癌是胃黏膜上皮的恶性肿瘤,胃炎是胃黏膜炎症,胃癌一般都是由胃炎发展而来的。根据数据可视化可以直观的看到各个疾病中男女性的患病人数比例。

由此可见,该关联规则算法的程序能够有效的研究各种疾病与性别,以及其他疾病之间的关联。

4 结论(Conclusion)

本课题利用python实现Apriori关联规则算法用于分析了医疗数据中疾病与疾病之间关系,发现了疾病之间的关联规则。通过本课题发现利用关联规则算法研究医疗数据中疾病之间的关系的的数据是有效的。关联规则算法能发现海量的医疗数据中蕴藏的信息并能得出关联规则的可信度。该结果能为医疗机构对疾病诊断提供参考,降低疾病的漏检误诊情况的发生。同时也证明了利用关联规则算法对医疗大数据进行数据挖掘所得出的结论有重要的参考价值^[10]。

同时由于源数据基于规模较小,分析结果并未能很好的展示疾病之间的明显关系。通过本课题的研究,探讨了Apriori关联规则算法在医疗数据中的应用,可更进一步的结合利用医疗数据中的个人信息和家庭信息数据对疾病相关数

据进行研究, 针对患者的年龄、地区、家族遗传、收入情况等数据进行分析, 寻找疾病与这些数据间的关联, 为疾病预防和治疗提供依据。

参考文献(References)

- [1] 降惠. 医学大数据可视分析研究[J]. 软件工程, 2017, 20(11): 1-3.
- [2] 程广, 王晓峰. 基于MapReduce的并行关联规则增量更新算法[J]. 计算机工程, 2016(02): 21-25; 32.
- [3] 李庆鹏, 张龙军, 耿新元. I-Apriori: 一种基于Spark平台的改进Apriori算法[J]. 科学技术与工程, 2017(12): 243-248.
- [4] 宋波, 杨艳利, 冯云霞. 基于关联规则Apriori算法的心肺性职业病情分析及预测[J]. 中国数字医学, 2017(04): 68-70.
- [5] 谢志明, 王鹏. 基于MapReduce架构的并行矩阵Apriori算法[J]. 计算机应用研究, 2017(02): 401-404.
- [6] 马继刚. 第二代居民身份证防伪特征的研究[D]. 中国人民公安大学学报(自然科学版), 2005.

(上接第26页)

(4) 打开菜单, 点击“PPT模式”, 系统转到PPT播放界面, 控制鼠标先打开要播放的ppt文件, 然后点击“从头开始放映”或“从当前开始放映”按钮, 点击“上一页”或“下一页”进行翻页, 点击“激光笔”实现激光笔效果, 再次点击“激光笔”按钮, 激光笔效果消失。

退出。先点击服务器端“停止”按钮, 关闭服务器; 然后打开手机客户端菜单, 点击“退出”按钮。

4 结论(Conclusion)

多媒体教学设备在一定程度上制约了教师在教学互动中的作用, 反而削弱了多媒体教学效果。本文针对这一问题, 利用Android智能手机及应用程序提出了无线鼠标的设计和实现, 主要针对学校日常多媒体教学, 利用手机来辅助教学, 使得教师可在教学中, 各种演讲活动中, 最大限度的发挥肢体语言的优势, 让教学或演讲更生动, 更完美, 解决以往在课堂或会议上使用鼠标键盘行动受限的状况。结合手机触控和硬件配置不断提高的特点, 程序开发出多种控制方式比如按钮控制、翻页控制和激光笔功能等, 使得对于课件控制自如, 远远超越传统键盘鼠标操控所带来的体验。智能无线鼠标易于使用、成本低、课前准备时间短, 有助于在课堂上拉近教师与学生间的距离、提高教学互动性。

参考文献(References)

- [1] Liu C H. A compatibility testing platform for android multimedia applications[J]. Multimedia Tools & Applications: 1-20.
- [2] 赵敏涯, 颜晓凯, 夏正航, 等. 基于Android的移动学习平台的设计与开发[J]. 软件工程, 2016, 19(5): 32-34.
- [3] Jeong E S, Kim I S, Lee D H. SafeGuard: a behavior based

- [7] 崔妍, 包志. 关联规则挖掘综述[J]. 计算机应用研究, 2016, 33(02): 330-334.
- [8] Pang-Ning Tan. 数据挖掘导论(完整版)[M]. 北京: 人民邮电出版社, 2011.
- [9] 曾勇. 基于关联规则的电子病历挖掘的应用研究[D]. 广州: 华南理工大学, 2012.
- [10] 邱书灵, 黄琳. 关联规则挖掘在研究生个人学习计划制定中的应用[D]. 石家庄: 石家庄铁道学院学报, 2007.

作者简介:

李 强(1976-), 女, 硕士, 副教授. 研究领域: 软件技术, 数据库技术, 大数据技术.

陈东涛(1995-), 男, 本科生. 研究领域: 软件技术, 大数据技术.

罗先录(1973-), 男, 硕士, 副教授. 研究领域: 软件技术, 人工智能.

real-time malware detection scheme for mobile multimedia applications in android platform[J]. Multimedia Tools & Applications, 2017, 76(17): 18153-18173.

- [4] 刘莎莎, 张哲. 基于Android平台的鼠标的设计与实现[J]. 电子器件, 2012, 35(1): 79-82.
- [5] 杨永平, 詹清林, 张东林. 一种Android手机远程控制计算机的系统实现[J]. 电脑知识与技术, 2015(11): 160-166.
- [6] 卢正军, 方勇, 刘亮, 等. 基于上下文信息的Android恶意行为检测方法[J]. 计算机工程, 2018(7): 156-161.
- [7] 罗志聪, 许昊, 叶廷东. 基于Android平台的模拟计算机输入设备研究[J]. 自动化与信息工程, 2018(2): 28-32.
- [8] 兰孝文, 刘江涛. 智能家居远程控制系统的设计与实现[J]. 计算机时代, 2014(5): 31-33.
- [9] 杨健, 谢晓方, 王彦, 等. 基于Android平台的多功能显示模拟器设计[J]. 火力与指挥控制, 2018(3): 129-133.
- [10] 朱士中, 陈凤, 先晓兵, 等. 基于移动终端和条码技术的实践教学管理平台研究[J]. 软件工程, 2017, 20(4): 22-25.
- [11] 张秋月, 吴琼, 苏全志, 等. 基于Android系统的RFID手持终端低功耗设计[J]. 电测与仪表, 2018(11): 112-115.
- [12] 周克良, 丛松楠, 邢素林. 基于物联网Android平台的智能多用户电表系统设计[J]. 现代电子技术, 2018(4): 116-118.

作者简介:

邓丽平(1987-), 女, 硕士, 助教. 研究领域: Web服务计算, 数据挖掘.

杨丽凤(1975-), 女, 博士, 副教授. 研究领域: 机器学习, 智能信息系统.