基于改进 Apriori 算法的肺癌致病因素研究

张 润,冯云霞
(青岛科技大学 信息科学技术学院,山东 青岛 266000)

摘 要:随着人民生活水平的不断提高,肿瘤疾病的人数在不断增多,其中肺癌是 21 世纪严重危害人类健康的重大疾病。面向肺癌电子病历如此庞大的数据量时,传统 Apriori 算法的串行计算方式需要频繁扫描数据库,会消耗巨大的内存占用量。对此,提出一种基于改进 Apriori 算法的肺癌风险评估因素分析的方法。运用 Hadoop 平台实现并行 Apriori 算法的优化,应用 HBase 文件存储系统对海量数据分布式存储以及 Map Reduce 框架进行分布式计算,最后给出基于 Hadoop 平台和 MapReduce 分布式计算模型的执行流程和测试结果。实验结果表明,改进算法在处理大数据及时有较好的执行效率以及良好的可扩展性,得出了肺癌的疾病模式与致病因素之间的隐匿规则,从而验证了改进后的 Apriori 算法对于辅助肺癌临床实验具有重要的意义。

关键词:关联规则;Apriori 算法;Hadoop;肺癌
中图分类号:TP391 **文献标识码**:A **文章编号**:1673-629X(2020)02-0143-05
doi:10.3969/j.issn.1673-629X.2020.02.028

Research on Pathogenic Factors of Lung Cancer Based on Improved Apriori Algorithm

ZHANG Run,FENG Yun-xia
(School of Information Science and Technology,Qingdao University of Science and Technology,
Qingdao 266000,China)

Abstract:With the continuous improvement of people's living standards,the number of cancer diseases is increasing,among which lung cancer is a serious threat to human health in the 21st century. Faced with such a large data volume in electronic medical records of lung cancer,the serial calculation method of traditional Apriori algorithm requires frequent scanning of the database,which will consume huge memory consumption. To this end,a method based on improved Apriori algorithm for lung cancer risk assessment factor analysis is proposed. The Hadoop platform is used to optimize the parallel Apriori algorithm. The HBase file storage system is used to distribute distributed data and the Map Reduce framework. Finally,the execution flow and test results based on Hadoop platform and MapReduce distributed computing model are given. The experiment shows that the improved algorithm has better execution efficiency and scalability in dealing with big data in time,and obtains the hidden rules between the disease pattern and the pathogenic factors of lung cancer,thus verifying the improved Apriori algorithm is of great significance for assisting clinical trials to assist lung cancer.

Key words:association rules;Apriori algorithm;Hadoop;lung cancer

0 引 言

肺癌是全球亟待解决的危害生命的最常见的癌症之一。2017 年,世界卫生组织的最新数据表示,仅仅 2015 年肺癌导致了约 170 万人死亡[1]。研究表明,肺癌早期患者的治愈率较高,而肺癌晚期患者的存活率仅为 15%[2]。主要原因是由于肺癌早期症状不明显,而中后期发病速度快,临床诊断时大多为中晚期[3]。

关联规则是反映出一个事务与其他事务之间相互关联或依赖的关系,看似不相关的事件的逻辑关系的知识,并运用于对同一事件中不同的特征之间的依存关系[4]。一份某种疾病的电子病历包含太多数据,若拿来直接用作预测或分类,会有多项不相关的因素干扰,因此,通过关联分析能寻找某类与该疾病存在密切相关的特征。其中,Apriori 算法是关联规则中经典的

收稿日期:2019-03-14 **修回日期**:2019-07-16 **网络出版时间**:2019-11-07
基金项目:国家自然科学基金(61572268)
作者简介:张 润(1994-),女,硕士生,CCF 会员,研究方向为医疗大数据应用技术;冯云霞,博士,副教授,研究方向为大数据应用技术、健康医疗信息工程。
网络出版地址:http://kns.cnki.net/kcms/detail/61.1450.TP.20191107.0912.046.html

算法。关联规则的运用最早追溯到 20 世纪 90 年代，Agrawal 首次提出关联规则问题，并将其运用在分析顾客交易数据中的关联因素[5]。此后，为了提高该算法的运行速率，不断有改进算法提出，如 Feng 等将 MapReduce 与 Apriori 算法相结合，设计了一种适用于大数据量的 MH-ACT 方法，把频繁项集分成几个互不相交的块，只对每个分块扫描一次，将结果合并到一起再计算所有项集的支持度[6]。Almaolegi 等为了降低数据库的规模，选用部分数据库中部分采样计算频繁项集，用数据库中剩余的数据来验证这些结果是否正确，这样大大降低了算法的时间复杂度[7]。Shah 等将哈希函数引入 Apriori 算法中，从而降低候选集的数量，提高了运算速率[8]。越来越多的改进算法，通过降低数据库的规模、分布式处理频繁项集、减少候选项集数目、引入 MapReduce 架构等方法，能够弥补传统 Apriori 算法的缺点。

1 Hadoop 平台与 Apriori 算法

1.1 Hadoop 平台

Hadoop 分布式广泛应用于多个软硬件平台，能够高速处理大规模数据的并行运算和存储问题，使用 Java 解决 PB 级的数据[9]。Hadoop 平台主要由并行编程模型 MapReduce、分布块 HDFS 和开源数据库 HBase 构成[10]。Hadoop 实现分布式并行数据处理时，若客户端想访问数据块，名称节点(NameNode)负责数据块的具体路径映射，并找到对应的数据节点(DateNode)，计算出数据的具体位置节点信息[11]。整个过程中，NameDode 作为 Hadoop 的主服务器节点，处理数据块到 Name Node 的映射关系，文件不通过其进行发送[12]。DataNode 完成对数据块进行创建、复制和删除的任务。工作流程如图 1 所示。

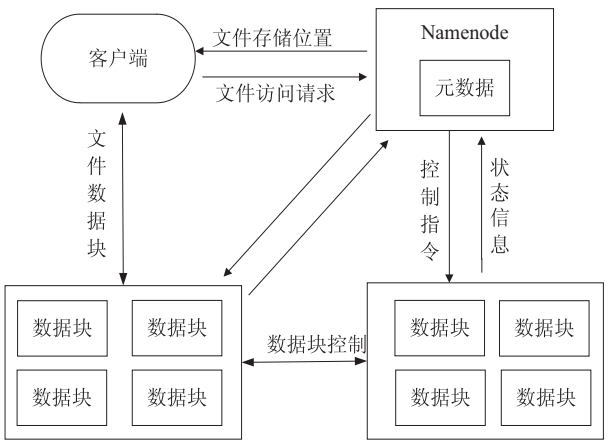

图 1 HDFS 的工作流程

1.2 Apriori 算法

作为数据挖掘中的重要工具，关联规则最早运用在分析顾客交易数据中的关联因素中[13]。最经典的算法即 Apriori 算法，首先按照最小支持度找到所有的频繁项集，然后产生强关联规则。Apriori 算法的挖掘过程包括挖掘频繁项集和关联规则的产生。

(1)频繁项集的挖掘。

设置最小支持度阈值(SUP_min)，在所有的候选项集中找出大于或等于 SUP_min 的项集，即频繁项集。

(2)强关联规则的产生。

根据第一步得出的频繁项集中，给定最小置信度(CONF_min)，若满足 CONF_min 的规则，称之为强关联规则。若存在项集 $\{I_3, I_4, I_5\}$，则规则为 $\{I_3 \rightarrow I_4, I_5\}$，$\{I_4 \rightarrow I_3, I_5\}$，$\{I_5 \rightarrow I_3, I_4\}$，$\{I_3, I_4 \rightarrow I_5\}$，$\{I_3, I_5 \rightarrow I_4\}$，$\{I_4, I_5 \rightarrow I_3\}$。

Apriori 算法是频繁挖掘项集中最重要的算法，是通过逐层迭代的候选生成方法[14]。核心是通过 K-项集挖掘(K+1)-项集[15]。衡量 Apriori 算法的两个重要标准：

(1)支持度(support)：描述关联样本中某个特征出现的频率。指对存在的项集 X、Y 和 B(X、Y 均属于项目集 B)，事物集 B 均包含事物集 X、Y 的百分比。存在如下关系：

$$\text{SUPPORT}(X \Rightarrow Y) = \frac{X \cup Y}{B}$$

(2)置信度(confidence)：描述两个特征之间相互关联的强度，指在事物 B 中包含 X、Y 事物数的百分比，关系如下：

$$\text{COMFIDENCE}(X \Rightarrow Y) = \frac{\text{SUPPORT}(X \cup Y)}{\text{SUPPORT}(X)}$$

支持度和置信度是 Apriori 算法中两个最重要的概念，两者通过 0% ~100% 的概率来衡量事务之间的紧密联系程度。最小支持度和最小置信度由人为设定，只有同时满足最小支持度和最小置信度才能称为两者具有强关联度。

2 基于 Hadoop 平台的 Apriori 算法并行化改进

基于 Hadoop 平台改进 Apriori 算法有两种主要方法：一种是数据集均匀分布在每个节点上，对局部并行挖掘频繁项集，收集全局频繁项集；第二种是使用 MapReduce 迭代挖掘频繁项集[16]。

本实验中，由于传统的 Apriori 算法的执行效率低、频繁扫描数据库，利用 Hadoop 平台结合 Apriori 算法迭代挖掘频繁项集，多次扫描数据库寻找候选项集。利用 MapReduce 将输入数据进行 Map 分块，在每次 Apriori 算法循环迭代时，对分布在每台计算机上的数据块进行累积求和，累计候选项集 C_k 的次数。在每个

分块数据中,通过求和运算计算候选项集 C_k 中属性的支持度,找出频繁项集 L_k。基于 Hadoop 平台的 Apriori 算法的并行化优化方法如下:

(1)执行 Apriori 算法得到候选-1 项集 C_1,将 C_1 与原始数据集对比,得到候选项集 C_1 中每个属性的支持度,通过 MapReduce 框架程序处理获得频繁项集 L_1。

(2)在分布于每个计算机上的 Map 块中,通过频繁项集 L_1,产生候选-2 项集 C_2,并逐次产生候选-k 项集 C_k。

(3)在 MapReduce 框架的 Reduce 进程中,通过求和运算对每个分布在 Map 节点上的 k 项候选项集的支持度累计,得到在 k 项时的全局支持度计数。比较全局支持度计数与最小支持度阈值,获得频繁项集 L_k。

(4)当前一台计算机计算出频繁项集后,后一台计算机启动 Map 进程并计算出频繁项集,以此步骤循环迭代,直到集合 L_k 为空,结束进程;否则继续执行步骤(2)。

(5)对处理完的数据块保存于 HDFS 中,并挖掘出相应的强关联规则。

图 2 为改进的算法流程。

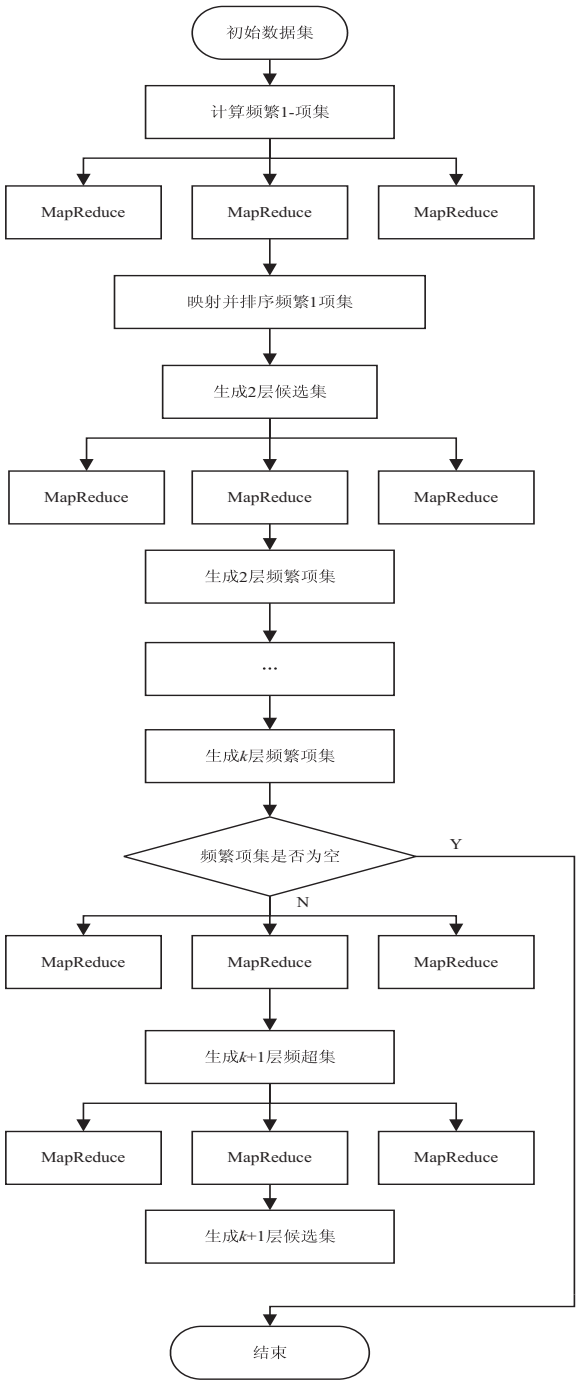

图 2　改进 Apriori 算法流程

3 改进 Apriori 算法在肺癌电子病历中的应用

3.1 实验环境

实验选用的 5 台 PC 机,包括一台名称节点 NameNode 和 4 台数据节点 DataNode,选取的计算机配置如表 1 所示。

表 1 计算机配置

CPU	内存	硬盘
Intel Core i7-3317 CPU @2.40 GHz	8 G	750 G

Hadoop 平台的一台计算器作为服务节点 NameNode Master,另外 4 台计算机作为服务节点 DataNode Slave,IP 分配如表 2 所示。

表 2 各个节点的 IP 分配

用户名	IP	职责
Master	192.168.10.10	master JobTracker
Slave1 ~ Slave4	192.168.10.11 ~ 192.168.10.14	slave TaskTracker

3.2 实验数据预处理

实验所使用的数据均来自本市某三甲级医院的肿瘤科电子病历,该电子病历记录患者从入院的身份数据、主诉、医嘱、检验数据到出院的各项数据。实验数据选取 2017 年 3 月至 2018 年 9 月的患者病历,以分析肺癌与吸烟、肺部疾病史、职业致病因子、咳嗽胸痛等信息之间的关联信息,以及症状之间的潜在规律。在进行关联分析之前,首先要对数据进行预处理,包括对数据合并、数据结构化、数据清洗以及数据转换等步骤。本次实验共选取肺部肿瘤患者共 18 个属性(包括性别、年龄、吸烟史、肺部疾病等信息)进行分析。

(1)数据合并:从医院 his 系统导出来的电子病历分为医嘱、诊断、检验等模块,需要根据患者唯一的 PID 标识进行关联,将患者的诊断、主诉、既往史、检验数据同步,所以运用 excel 表格对数据集成合并处理。

(2)数据结构化:使用 ICTCLAS 作为分词工具,建立医学用户词典,提取按词频分类结果的结构化属性表。

(3)数据清洗:提取特征属性的结构化电子病历存在异常数据、缺失值数据[17]。缺失值处理中,对数值型数据,选择均值代替;对字符型数据,选择众数代替。存在大量缺失值的数据,选择直接删除。异常值处理中,计算出每类数据所占比例,并画出正态分布,对于所占比例过低的数据判断为异常值[18]。异常值的处理方式与缺失值相同。

(4)数据转换:由于 Apriori 算法只能对离散化数据进行处理,所以在进行数据挖掘前,要对连续性数值进行离散化处理。以吸烟史为例,从未吸烟为 0,1 至 10 年为 1,10 至 20 年为 2 等。

3.3 改进算法对比分析

在本实验中,由于涉及的患者病历数据量较大,为了获得更加有价值的信息,将最小值支持度和最小置信度不断改进。在最小置信度固定为 0.6 的情况下,最小支持度为 0.06 时,挖掘的强关联规则太多,这种情况下的规则是无意义的。直到最小支持度提高到 0.1 时,挖掘的关联规则数量产生了明显的变化。通过多次实验对比,选定最小支持度为 0.1,最小置信度为 0.6,这是肺癌数据挖掘较为合适的参数设置。

Apriori 串行算法与改进并行算法在处理相同规模数据时所用时间的对比如表 3 所示。当数据规模不断增大时,串行算法所用的时间明显增多,直到提示内存不足。而改进的并行算法在处理大规模数据时完成能力较好。因此,在处理小规模数据量时,传统的串行算法比改进的并行算法效果好,这是因为 Hadoop 平台的节点启动运行需要一定的时间;在处理大规模数据量时,改进的并行算法的效率远远高于传统的串行算法。

表 3 改进算法与传统算法的比较

实验次数	数据/MB	时间 1/s	时间 2/s
1	3	10	86
2	7	22	97
3	18	46	120
4	48	137	155
5	61	内存不足	174

节点数选取 1、2、3、4,电子病历数据量大小分别为1 G、2 G、3 G。每次实验进行 3 轮取平均值,最终的运行时间结果如图 3 所示。

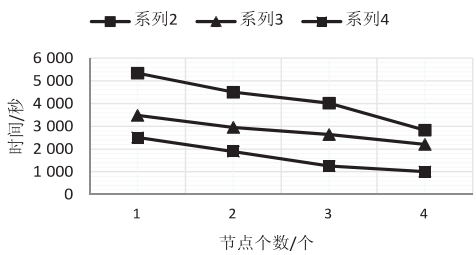

图 3 改进算法实验结果

从图 3 中可以看出,从下到上的折线分别为1 G、2 G、3 G 大小数据量的运行结果,随着数据规模的不断增大,运行所用时间随着节点数的增加而减少。因此,增加 Hadoop 集群的节点数可以显著提高数据处理能力,基于 Hadoop 平台改进的 Apriori 算法具有良好的性能,在处理肺癌电子病历具有较高的执行效率。

3.4 实验结果

最终得到强关联规则,部分如表 4 所示. 从表中的

关联结果,得到以下结论:

(1)在性别对比上,患肺癌男性远远高于女性,这与男性为吸烟主要群体有一定关系。在年龄对比上,60 至 70 岁的老年人患肺癌最多,但由于吸烟人数的增多,肺癌患者也呈现低龄化趋势,尤其是在年轻的男性中,患肺癌人数逐年增加。

(2)有胸闷憋气、胸痛、咳嗽、咳血等症状与肺癌有着密切的关系,对化验数据的关联规则挖掘,基于 CT 影像数据能够及时发现早期肺癌。

(3)吸烟对肺癌有着严重的影响,吸烟史与肺癌的发病率有极大的关系。

(4)肺癌与职业治病因子之间有一定关联规则,从事石油、粉末、煤炭等职业人群患肺癌的概率较大。

(5)肺部疾病患者中,有肺结核等疾病的患者癌变的可能性比较大。

改进的算法与临床医学结论相符合,能够挖掘疾病与病因之间的潜在规律和规则,这对肺癌疾病的分析与研究具有重要的意义。

表 4　关联规则结果

结果	支持度	置信度	兴趣度
{肺癌}→{吸烟}	0.134 5	0.654 5	1.35
{肺癌}→{肺部疾病史}	0.118 9	0.722 4	1.18
{肺癌}→{职业致病因子}	0.152 0	0.767 9	1.11
{肺癌}→{咳嗽,咳血}	0.127 6	0.417 8	1.21
{肺癌}→{胸痛,胸闷}	0.109 2	0.590 2	1.05

4　结束语

基于 Hadoop 平台改进的 Apriori 算法可以协助临床医生快速、准确、高效地做出判断,对肺癌早期预防、早期治疗具有重要的意义。通过实验结果可以看出,在处理大规模电子病历数据时,基于 Hadoop 平台改进的 Apriori 算法的执行效率远远高于传统 Apriori 算法。并且改进后的算法具有良好的可移植性,能够适用于肺癌电子病历的数据挖掘,及时有效地挖掘出肺部肿瘤疾病与症状之间的潜在规律,具有一定可行性。

参考文献:

[1] WOOD D E,KAZEROONI E A,BAUM S L,et al. Lung cancer screening,version 3. 2018,NCCN clinical practice guidelines in oncology[J]. Journal of the National Comprehensive Cancer Network,2018,16(4):412-441.

[2] LI W M,ZHAO S,LIU L X. The methods and clinical significance of early diagnosis of lung cancer[J]. Journal of Sichuan University:Medical Science Edition,2017,48(3):331-335.

[3] SAGAWA M,KOBAYASHI T,UOTANI C,et al. A survey about further work-up for cases with positive sputum cytology during lung cancermass screening in Ishikawa Prefecture,Japan:a retrospective analysis about quality assurance of lung cancer screening[J]. Japanese Journal of Clinical Oncology,2015,45(3):297-302.

[4] FLUSSER J,ZITOVA B,SUK T. Moments and moment invariants in pattern recognition[M]. [s. l.]:Wiley Publishing,2009.

[5] WILSON P W F,D'AGOSTINO R B,LEVY D,et al. Prediction of coronary heart disease using risk factor categories[J]. Circulation,1998,97(18):1837-1847.

[6] FENG D,ZHU L,ZHANG L. Research on improved Apriori algorithm based on MapReduce and HBase[C]//Advanced information management,communicates,electronic & automation control conference. Xi'an,China:[s. n.],2016:887-891.

[7] LIU X,LIU H. An improved Apriori algorithm for association rules[J]. Telkomnika Indonesian Journal of Electrical Engineering,2013,11(11):942-946.

[8] SHAH A. Association rule mining with modified Apriori algorithm using top down approach[C]//2016 2nd international conference on applied and theoretical computing and communication technology (iCATccT). Bangalore:IEEE,2016:747-752.

[9] 刘贤熳,宋　斌. 基于 Hadoop 的海量数据 TCP 报文重组技术[J]. 计算机工程,2016,42(10):113-117.

[10] BHANDARKAR M. MapReduce programming with apache Hadoop[C]//2010 IEEE international symposium on parallel & distributed processing (IPDPS). Atlanta,GA:IEEE,2010:1-1.

[11] AL-MAOLEGI M,ARKOK B. An improved Apriori algorithm for association rules[J]. International Journal on Natural Language Computing,2014,3(1):21-29.

[12] 王　姗. 基于 Hadoop 平台的一种 Apriori 算法改进方法[D]. 长春:吉林大学,2016.

[13] AGRAWAL R,SRIKANT R. Fast algorithms for mining association rules in large databases[C]//International conference on very large data bases. [s. l.]:Morgan Kaufmann Publishers Inc.,1993:487-499.

[14] YU W,WANG X,WANG F,et al. Notice of retraction:the research of improved Apriori algorithm for mining association rules[C]//2008 11th IEEE international conference on communication technology. Hangzhou:IEEE,2008:513-516.

[15] 卢　辉. 数据挖掘与数据化运营实战[M]. 北京:机械工业出版社,2013.

[16] 刘木林,朱庆华. 基于 Hadoop 的关联规则挖掘算法研究——以 Apriori 算法为例[J]. 计算机技术与发展,2016,26(7):1-5.

[17] 邹北骥. 大数据分析及其在医疗领域中的应用[J]. 计算机教育,2014(7):24-29.

[18] CHEN J,CHEN Y,DU X,et al. Big data challenge:a data management perspective[J]. Frontiers of Computer Science,2013,7(2):157-164.