基于电子病历信息大数据挖掘的患者就医行为分析

翟运开 武 戈

（郑州大学管理工程学院 郑州 450001）

〔摘要〕 在对就医行为和大数据挖掘及关联规则进行文献综述的基础上，介绍大数据挖掘及关联规则算法——Apriori 算法的基本内容。以郑州大学第一附属医院部分电子病历系统数据为基础，采用关联规则的 Apriori 算法对患者的性别、年龄、手术记录、危重记录、就诊距离、住院天数等因素进行挖掘分析，得出不同就诊距离患者的就医行为规律，提出相关政策和建议。

〔关键词〕 大数据挖掘；关联规则；Apriori 算法；就医行为

〔中图分类号〕 R－056 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673－6036.2017.07.003

Analysis on Medical Behaviors of Patients Based on Big Data Mining of Electronic Medical Records (EMR) Information

ZHAI Yun－kai, WU Ge, College of Management Engineering, Zhengzhou University, Zhengzhou 450001, China

〔**Abstract**〕 Based on the literature review of medical behaviors, big data mining and related rules, the paper introduces the basic contents of big data mining and related rule algorithm—Apriori algorithm, takes partial Electronic Medical Records (EMR) system data of the First Affiliated Hospital of Zhengzhou University as the basis, conducts mining analysis on factors such as the gender, age, operation records, critical ill records, visiting distance and hospital days of a patient by adopting Apriori algorithm of related rule, obtains the rules of medical behaviors of patients of different visiting distances, and puts forward relevant policies and suggestions.

〔**Keywords**〕 Big data mining; Association rules; Apriori algorithm; Health seeking behavior

1 引言

患者是医疗服务的利用者，患者的就医行为模式及就医决策反映了其选择医疗服务的意向，这一决策的产生受多种因素影响[1]，研究患者的就医行为及影响因素，对制定医疗服务规划、加强医院管理具有十分重要的意义。随着医疗信息化的普及和大数据技术的发展，面向医院电子病历系统的大数据挖掘将更能揭示影响患者就医行为的各种因素之间的作用关系。因此，本文针对目前省级医院门诊量井喷，患者首选“大医院”接受诊疗的现状，选择河南省三甲医院——郑州大学第一附属医院就诊患者的电子病历数据作为研究对象，通过大数据平台对电子病历进行大数据挖掘，分析患者的就医行为，以期进一步了解患者依赖“大医院”诊断的原因，为分析患者的医疗需求、合理分配国家医疗资源、正确制定医疗卫生政策提供支持和参考。

〔收稿日期〕 2017－03－29

〔作者简介〕 翟运开，博士，副教授，主任，硕士生导师，发表论文多篇；武戈，硕士研究生。

2 文献综述

2.1 就医行为

就医行为是个体以各种方式对身体的征兆做出反应，对体内状况进行监测，确定和解释躯体症状，寻找疾病原因，采取治疗措施，利用各种正式和非正式的保健资源，其可分为卫生服务利用行为、疾病反应或健康促进行为[2]。医学行为学认为，能够影响个体产生就医行为的因素很多，大致可以划分为 3 类：对症状的认知情况、医疗服务情况和社会经济情况。李淑玲等[3]通过发放调查问卷，运用结构方程模型论证了新农合制度对农民就医行为与认知产生了正面积极的影响，即农民可以更加自主、正确地选择适合自己的就医地点。钱东福等[4]采用 logistic 回归模型对影响甘肃省农村居民选择住院治疗机构的因素进行分析，得出其主要因素是疾病类型、住院天数、收入、职业和性别。刘晓莉等[5]对四川省农村居民进行随机抽样得出，农村居民对不同等级医院的关注度不同，其根据病情轻重选择不同医疗机构。黄建军等[6]采用随机抽样利用 logistic 回归模型验证得出医疗效果、费用、服务态度、就医过程的便捷性是患者做出就诊医院决策中的重要影响因素，而患者到医院的就诊距离的影响在模型中并不显著。综上，目前对患者就医行为的研究多侧重于医疗机构的选择，患者重点考虑疾病的严重程度、医疗服务质量及自身经济情况选择就医地点，从而导致患者“千里寻医”等依赖省级三甲医院诊疗的现象。在研究方法上，当前研究大多利用回归分析，只考虑了影响因素的主效应，或用 logit 模型进行叠加而忽略了各影响因素之间的交互作用。本研究取用郑州大学第一附属医院电子病历数据，从性别、年龄、科室、住院时间、疾病严重程度、手术记录、就诊距离几个层面分析患者在不同就诊距离的分布情况，利用大数据平台进行数据挖掘，探究各个因素与患者就医行为的关联程度。

2.2 大数据挖掘及关联规则

大数据挖掘是指从大量随机、模糊、有噪声的数据中提取出看似毫无关联的、未被发现的有用知识。关联规则可以从大量数据类中预测出任何属性之间存在的某种规律性关系。随着大数据技术的发展以及“互联网 +”与医疗的结合，关联规则在医学、药学等领域取得了突破性研究成果，如在疾病诊断与预测、药物研究、基因表达等领域[7-9]，应用向量支持机对各个因素之间的关联程度进行量化分析，揭示了因素间隐藏的作用关系。目前针对电子病历数据的研究大多对有价值的门诊信息字段进行聚类分析，得到医院目前门诊患者的年龄结构构成和主要就诊疾病类型等，为医院提供针对性服务指明了方向[10]。电子病历系统近几年于各大医院推广，病历数据的挖掘与应用面临着数据结构复杂、自然语言识别等技术难题，所以大数据挖掘在这一领域的应用尚处于起步阶段。

本研究基于 Hadoop 框架的大数据技术对数据进行，装载、转化、加载（Extract – Transform – Load，ETL）等逻辑操作，选择数据挖掘中经典的关联规则算法——Apriori 算法对医院的电子病历数据进行关联分析，提取医院电子病历系统中患者的基本信息、入院记录、就诊科室等字段，对患者性别、年龄、就诊科室、住院时间、就诊距离与疾病严重程度之间的联系进行关联性挖掘，描述不同就诊距离范围内患者的就医行为。

3 大数据挖掘及关联规则算法

3.1 大数据挖掘

进行大数据挖掘首先要针对所研究的内容准备原始数据，再对数据中的噪声值、缺省值、敏感信息等进行替换或填补，筛选出需要的字段构建数据仓库，然后结合选用的模型或者算法的需要将字段中包含的信息作无量纲化处理，得到大数据挖掘所需的数据集市，进而对其进行挖掘分析；最后根据挖掘结果做出合理的解释说明，得出隐藏的有用知识。其一般研究过程，见图 1。数据挖掘中的数据量越大，得到的信息越有研究意义；但是当信息量急剧增长时，利用传统方法对数据进行 ETL 等复杂逻辑运算所消耗的时间也将呈几何级数增长。因此

需要 MapReduce 的思想，采用基于 Hadoop 框架的大数据分析平台，利用 HDFS 存储海量电子病历信息，将 ETL 等复杂逻辑计算任务分配于多个节点同时执行，提高计算效率，使 TB 级甚至于 PB 级数据量的大数据挖掘成为可能。

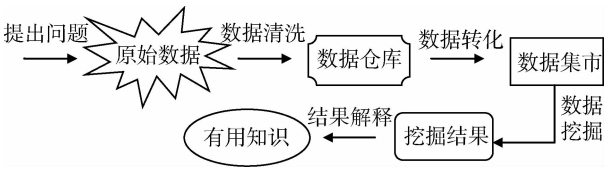

图 1 大数据挖掘流程

3.2 关联规则算法描述

本研究采用的 Apriori 算法是大数据挖掘中关联规则技术的核心算法，是由 Agrawal 和 Srikant 于 1994 年提出的。该算法利用了频繁项集的非单调性：如果一个 k - 项集是非频繁项集，则其超集一定是非频繁项集。Apriori 算法将发现关联规则的过程分为两个阶段：第 1 阶段，通过迭代，检索出事务数据库中的所有频繁项集，即支持度不低于预先设定的阈值的项集；第 2 阶段，利用频繁项集构造出满足要求的规则。在给定的一个含有 n 个事务（Transaction）的数据库 $D = \{t_1, t_2 \cdots t_n\}$ 中，有 m 个属性，这 m 个属性组成的项集（Itemset）为 $I = \{i_1, i_2 \cdots i_n\}$，那么其中每个事务 t 都是一个项集，且 $t \subseteq I$。关联规则的一般表示形式为 $X \rightarrow Y$，其中 X、Y 是项集，且 $X \subset I$，$Y \subset I$，$X \cap Y = \phi$ 称为规则前项（Antecedent），Y 称为规则后项（Consequent）。支持度（Support）即关联规则的普遍性，表示项集 X 和项集 Y 同时出现的概率，即条件概率 $P(X \cup Y)$，记作 $support(X \rightarrow Y) = P(X \cup Y)$。置信度（Confidence）代表关联规则的准确度，表示包含项集 X 的事务中同时也包含项集 Y 的概率，即条件概率 $P(Y \mid X)$，记作 $Confidence(X \rightarrow Y) = P(Y \mid X)$。提升度（Lift）是规则的置信度与规则后项支持度的比值，反映了 X 的出现对 Y 出现的影响程度，体现了关联规则的重要性。

4 电子病历信息大数据挖掘过程

4.1 原始数据

根据被调研医院数据研究方面的相关规定，研究采用郑州大学第一附属医院允许使用的 2013 年全年电子病历系统数据，共计 205 461 条。数据字段包括患者标识信息（患者标识号、就诊编号）、患者基本信息（性别、出生日期、出生地、国籍、民族、婚姻状况、职业、文化程度、籍贯、现住址、联系方式）、基本健康信息（身高、体重、入院主诉、现病史、既往史、过敏史、婚育史、家族史、入院检查、辅助检查、初步诊断）、卫生事件摘要（就诊机构、机构等级、机构类型、历史就诊时间、历史就诊机构、历史就诊科室）、治疗过程（门诊类别、入院科室、手术信息、药物过敏、查房记录、入院诊断、出院诊断、出院医嘱）。将全部电子病历信息导入 Hadoop 集群，建立以患者病历号为主索引的 HDFS 电子病历数据库，共包含 98 个特征字段，为数据仓库的建立做准备。

4.2 数据清洗

由于原始数据的数量庞大、结构复杂，因此在进行大数据挖掘之前需要进行预处理。利用 Map 函数和 Reduce 函数对数据进行 ETL 处理，发挥 HDFS 分布式文件的存储优势，调用 RDD 进行分布并行计算，缩短逻辑处理时间。对于隐私类信息，如患者基本信息中包含姓名、职业、住址、联系方式等个人隐私，为保障患者和医院的权益不受侵犯，同时确保数据的完整性，将此类真实的敏感信息加密替换；对于异常、错误、极端值，如性别（男、女）中出现其他字段、住院天数超过 365 天、年龄出现负值等错误信息，将此记录做删除处理。最终采用分析的数据仓库包含 190 932 条记录，选取字段包括性别、年龄、科室、手术记录、危重记录、入院日期、出院日期、归属地。

4.3 数据转化

数据清洗之后，选定符合大数据挖掘要求的样本，即数据集市，进而对样本各个字段的数据做无量纲化处理，主要包含数据的标准化与离散化。

（1）标准化。将“性别（sex）”分为 1 =“女”，2 =“男”；将“手术记录（oper）”分为 1 =“无”，2 =“有”；将“转院记录（tran）”分为 1 =“无”，2 =“有”；将“危重记录（dan）”分为 1 =“无”，2 =“有”。（2）离散化。针对连续型数据，结合样本分布情况，根据中国人口年龄分段标准将“年龄（age）”分为 10 个区间段：童年 1 = [0, 6)，少年 2 = [6, 17)，青春期 3 = [17, 28)，成熟期 4 = [28, 40)，壮年期 5 = [40, 48)，稳健期 6 = [48, 55)，调整期 7 = [55, 65)，初老期 8 = [65, 75)，中老期 9 = [75, 85)，年老期 10 = [85, -)；根据原数据中的“出院日期”与“入院日期”计算出患者“住院天数（day）”，将“住院天数”分为 5 个区间段：1 = [1, 10)，2 = [10, 20)，3 = [20, 40)，4 = [40, 80)，5 = [80, -)；调用空间数据分析函数，根据患者“现住址”与郑州大学第一附属医院的空间坐标得出“就诊距离（far）”，进而将其分为 4 个区间段：1 = [0, 100)，2 = [100, 200)，3 = [200, 300)，4 = [300, 500)。

4.4 大数据挖掘结果

预先设置最小支持度和最小置信度得出关联规则，调用 Apriori 函数得出关联规则，生成关联规则的子集矩阵后去掉冗余规则，最后查看有价值的关联规则进行分析。

4.4.1 关联后项为 {far =1} 的结果 患者居住地距就诊点小于 100 公里的关联结果，见表 1。规则 1 表明在该区间内，年龄于 28 ~ 40 岁的女性在生殖科就诊且无手术记录的患者占全部患者的 73.4%，是该范围内其他患者的 1.97 倍；规则 2 表明在该区间内，住院天数在 80 天以上且无危重记录的患者占全部患者的 50.9%；规则 3 表明在该区间内，就诊于康复科住院天数在 20 ~ 40 天之间且无危重记录的患者占全部患者的 48.9%。此距离内患者占总体的 37.21%，以无手术记录且无危重记录的患者居多，并且住院天数较长，以康复疗养及中老年慢性病为主，大多需要反复多次的体征观测进行长期治疗。

表 1 {far =1} 的关联规则结果

编号	前项 = > 后项	支持度	置信度	提升度
1	{sex = 1，age = 4，num = 生殖科，oper = 1} ⇒ {far = 1}	0.001 503	0.734 0	1.972 3
2	{day = 5，dan = 1} ⇒ {far = 1}	0.001 922	0.509 0	1.367 7
3	{num = 康复科，day = 3，dan = 1} ⇒ {far = 1}	0.001 047	0.488 9	1.313 9

4.4.2 关联后项为 {far =2} 的结果 患者居住地距就诊点距离 100 ~ 200 公里之间的关联结果，见表 2。规则 1 表明在该区间内，年龄于 6 ~ 17 岁在普外科就诊的患者占全部患者的 47.8%，是该范围内其他患者的 1.75 倍；规则 2 表明在该区间内，年龄于 6 ~ 17 岁在眼科就诊且无手术记录并住院 10 天以下的患者占全部患者的 42.9%，是该范围内其他患者的 1.57 倍；规则 3 表明在该区间内，年龄于 6 ~ 17 岁在耳鼻喉科就诊且无手术记录的患者占全部患者的 36.2%，是该范围内其他患者的 1.33 倍。此距离内患者占总体的 27.35%，以无手术记录或短期住院的患者为主，其中年龄在 6 ~ 17 岁且无手术记录的患者约占全部患者的 33.2%，多就诊于普外科、眼科与耳鼻喉科。

表 2 {far =2} 的关联规则结果

编号	前项⇒后项	支持度	置信度	提升度
1	{age = 2，num = 普外科} ⇒ {far = 2}	0.001 215	0.478 3	1.748 7
2	{age = 2，num = 眼科，day = 1，oper = 1} ⇒ {far = 2}	0.001 356	0.428 8	1.567 6
3	{age = 2，num = 耳鼻喉科，oper = 1} ⇒ {far = 2}	0.001 623	0.362 5	1.325 4

4.4.3 关联后项为 {far=3} 的结果 患者居住地距就诊点距离 200~300 公里之间的关联结果，见表 3。规则 1 表明在该区间内，手术后住院 40~80 天且无危重记录的男性患者占全部患者的 35%，是该范围内其他患者的 1.39 倍；规则 2 表明在该区间内，就诊于心内科且住院 20~40 天的患者占全部患者的 34.4%，是该范围内其他患者的 1.34 倍；规则 3 表明在该区间内，手术后住院 20~40 天年龄于 65~75 岁且无危重记录的女性患者占全部患者的 33.5%，是该范围内其他患者的 1.32 倍；规则 4 表明在该区间内，手术后住院 10~20 天就诊于肿瘤科的男性患者占全部患者的 33%，是该范围内其他患者的 1.3 倍。此距离内患者占总体的 25.28%，以入院接受手术治疗的老年患者为主，就诊科室多分布于心内科、肿瘤科这类慢性重症科室，住院天数也随之增加。

表 3 {far=3} 的关联规则结果

编号	前项⇒后项	支持度	置信度	提升度
1	{sex=2，day=4，oper=2，dan=1} ⇒ {far=3}	0.002 105	0.350 1	1.385 1
2	{num=心内科，day=3} ⇒ {far=3}	0.001 136	0.344 4	1.362 4
3	{sex=2，age=8，day=3，oper=2，dan=1} ⇒ {far=3}	0.001 262	0.334 7	1.323 9
4	{sex=1，num=肿瘤科，day=2，oper=2} ⇒ {far=3}	0.001 073	0.330 6	1.307 8

4.4.4 关联后项为 {far=4} 的结果 患者居住地距就诊点距离 300 公里以上的关联结果，见表 4。规则 1 表明在该区间内，手术后住院 10 天以下就诊于呼吸内科且无危重记录的患者占全部患者的 16.7%，是该距离内其他患者的 1.67 倍；规则 2 表明在该区间内，手术后住院 10 天以下年龄于 75~85 岁且无危重记录的男性患者占全部患者的 16.5%，是该范围内其他患者的 1.62 倍；规则 3 表明在该区间内，就诊于 ICU 内住院 10 天以下的患者占全部患者的 15.6%，是该范围内其他患者的 1.53 倍；规则 4 表明在该区间内，手术后住院 10 天以下就诊于肿瘤科且无危重记录的男性患者占全部患者的 15.2%，是该范围内其他患者的 1.5 倍。就诊距离大于 300 公里的患者人数最少，占总体的 24.7%，多就诊于呼吸内科、ICU 及肿瘤科，以急、重症患者为主，住院天数大多小于 10 天。

表 4 {far=4} 的关联规则结果

编号	前项⇒后项	支持度	置信度	提升度
1	{num=呼吸内科，day=1，oper=2，dan=1} ⇒ {far=4}	0.001 791	0.166 8	1.643 8
2	{sex=2，age=9，day=1，oper=2，dan=1} ⇒ {far=4}	0.001 120	0.165 1	1.627 0
3	{num=ICU，day=1} ⇒ {far=4}	0.001 089	0.155 6	1.534 0
4	{sex=2，num=肿瘤科，day=1，oper=2，dan=1} ⇒ {far=4}	0.001 042	0.152 4	1.502 5

5 结论

5.1 关联规则结果分析

根据医疗大数据挖掘结果可得出，患者的就诊距离与就诊科室、患病类型、住院天数、年龄和性别存在一定关联。数据统计显示，就诊距离越远人数随之递减，而患者中手术记录、危重记录人数随之增大。数据来源医院郑州大学第一附属医院为河南省内最大的一家三级甲等医院，其医疗服务水平和患者的认可程度均占省内医院榜首，这使得患者在某种程度上盲目依赖“大医院”的诊断，加之河南省医疗资源分配不均、基层医疗设施陈旧、医疗水平有限等原因，更加促成了患者生病后前往“大医院”就诊的习惯，造成了“看病难、看病贵”等一系列问题。从关联规则的结果看，就诊距离在 100 公里以内的住院患者占用了大量医疗资源；就诊距离在 100~200 公里之间的患者多为无手术记录

且无危重记录的普外科、眼科、耳鼻喉科患者，可以不必过多依赖“大医院”的诊疗；就诊距离在 200～300 公里内及 300 公里以上的患者，急、重症患者居多，大部分就诊于肿瘤科、心内科、ICU 等重症科室，对高质量医疗服务的需求增大，当地的医疗水平满足不了患者的医疗需求，故选择距离更远的“大医院”就诊。

5.2 相关对策及建议

为正确引导患者的就医路径，合理分配医疗资源，国家大力推进分级诊疗政策，鼓励远程医疗服务发展。通过推进分级诊疗政策，引导优质医疗资源下沉，提升基层医生诊断水平，以常见病、多发病、慢性病分级诊疗为突破口，培养科学合理的就医秩序，逐渐形成基层首诊、双向转诊、急慢分治、上下联动的分级诊疗模式。发展远程医疗服务，打破患者的地域和时间壁垒，提高医疗服务可及性，缩短偏远地区患者的就诊时间，同时也节省路途花费。基于远程医疗技术建立医疗联合体，提高医疗网络中专家医师、普通医生及医疗设备的利用率[11]，为我国医疗资源的合理配置提供新思路。

参考文献

1 高阔，甘筱青. 患者选择就医单位分布及其影响因素分析［J］. 中国卫生事业管理，2014（7）：516－517，545.

2 张容瑜，尹爱田，Shi Lisheng，等. 就医行为及政策影响因素研究进展［J］. 中国公共卫生，2012，28（6）：861－862.

3 李淑玲，乔钰涵，张伟. 新农合农民就医行为与认知影响因素的实证研究［J］. 工业工程与管理，2014，19（4）：98－103.

4 钱东福，尹爱田，孟庆跃，等. 甘肃省农村居民选择住院医疗机构的影响因素研究［J］. 中国卫生经济，2008，27（1）：40－43.

5 刘晓莉，段占祺，陈文，等. 四川省农村居民就医行为现状调查及对策分析［J］. 卫生软科学，2016，30（1）：30－33.

6 黄建军，曾玉和. 病人选择就诊医院影响因素的 logistic 回归分析［J］. 中国医院统计，2006，13（2）：119－121.

7 余辉，张力新，刘文耀. 计算机辅助医学知识发现系统研究——糖尿病并发症流行病学数据挖掘［J］. 生物医学工程，2008，25（2）：295－299.

8 刘尚辉，王露，郑德禄. Apriori 关联规则在甲状腺结节病案分析中的应用［J］. 中国卫生统计，2011，28（2）：178－179.

9 Montella A. Identifying Crash Contributory Factors at Urban Roundabouts and Using Association Rules to Explore Their Relationships to Different Crash Types［J］. Accident Analysis & Prevention，2011，（4）：1451－1463.

10 左篙，张雄，刘礼德. 基于数据挖掘的门诊信息资源分析［J］. 现代生物医学进展，2013，13（23）：4568－4592，4594.

11 赵杰，崔震宇，蔡雁岭，等. 基于远程医疗的资源配置效率优化研究［J］. 中国卫生经济，2014，33（10）：5－7.

（上接第 11 页）

17 Mohr DC，Burns MN，Schueller SM，et al. Behavioral Intervention Technologies：evidence review and recommendations for future research in mental health［J］. Gen Hosp Psychiatry，2013，35（4）：332－338.

18 Taylor RC. An Overview of the Hadoop/MapReduce/HBase Framework and Its Current Applications in bioinformatics［J］. BMC Bioinformatics，2010，11（Suppl 12）：S1.

19 Yao QA，Zheng H，Xu ZY，et al. Massive Medical Images Retrieval System Based on Hadoop［J］. Journal of Multimedia，2014，9（2）：216－222.

20 陆婷娟，戚小平. 基于 Hadoop 的医学影像数据平台应用研究［J］. 世界复合医学，2015，（3）：223－226.

21 李萍. 云计算与大数据时代医院信息化的三个转变［J］. 中国医院管理，2013，33（12）：80－81.