健康大数据的分析方法与应用研究

崔民权
对外经济贸易大学统计学学院，中国·北京　100029

【摘　要】为探寻健康大数据的高质量精准分析，以进一步有效应用这些数据，在本次研究中，结合实际需要，首先对健康大数据的分析方法体系框架进行搭建，明确了框架内各层级的职能；其次，对健康大数据的分析算法和流程进行了阐述，以此作为本次健康大数据分析应用的依据；最后，结合相关技术，对实际案例进行分析，分析结果表明，本次健康大数据分析方法取得了初步成功，有望在今后的类似工作中得到进一步的推广应用。

【关键词】健康大数据；数据分析；数据应用

Analysis Method and Application Research of Health Big Data

Cui Minquan
School of Statistics, University of International Business and Economics Beijing 100029

[Abstract] In order to explore the high quality and accurate analysis of health big data, according to the practical needs, the framework of the analysis method of health big data is established, and the functions of all levels of the health big data are defined. Secondly, the health big data analysis method is expected to be further applied in the future similar work.

[Keywords] Health big data; data analysis; data application

前言

近年来，为提升我国医疗健康行业水平，大数据技术在该行业中的应用得以进一步推进，基于大数据技术，能够辅助医务人员实现更为精准的医疗服务，有助于进一步提升医疗行业发展水平。当然，健康大数据的分析应用是一项综合性较强的工作，在实际工作中，仍需要从数据选取、分析方法、技术支持等多个角度同时入手，对这项工作展开深入探究。

1　健康大数据的分析方法体系框架

结合当前的实际需要，在健康大数据分析应用环节中，通常基于相关的分析系统展开对应的工作。在实际的系统设计工作中，通常采用多层级结构模型展开设计，其主要分为以下几个组成部分。

一是目标层，该层级是整个框架体系的基础部分，主要作用是对健康大数据进行目标性区分，其分类依据通常是数据复杂程度和数据价值，由此，目标层的数据分析结果也各不相同，部分数据仅用来描述，一些价值相对较高的数据则用于后续的分析预测。

二是类形层，该层级承担的主要功能是对数据类型进行细分，特别是对一些难以量化的数据进行处理，推动数据的“结构化”。

三是分析方法及工具层，该层级的细分功能模块相对较为繁杂，通常又可细分为以下几个组成部分：（1）分析方法模块：该模块在已经明确数据特征的数据集中，确定分析方法类别；（2）典型算法模块：根据实际功能需要，对具体的分析算法进行选择；（3）分析工具模块：根据实际功能需要，选择具体的分析工具，将数据集中的内容予以可视化处理。

四是应用层，该层级主要用于划分健康大数据的具体应用方向。

2　健康大数据的主要分析方法及技术流程

2.1 数据分析方法的选择

由于健康大数据的内容较为宽泛，且各类数据之间的关系错综复杂，其非线性的特点尤为突出，对此，为有效处理这些数量关系，在实际研究中通常采用支持向量机（SVM）模型进行数据分析。这种分析模型在泛化能力上具有突出优势，对于健康大数据的分析也较具作用。在应用SVM模型进行分析时，其通常基于以下几个流程进行：（1）在MATLAB环境下，选取Libsvm支持向量机集成工具包；（2）在Python编译环境下编写相关代码，对SVM模块进行导入处理，并选择RBF核函数；（3）使用已有的数据集，对SVM模型进行训练，并构造分类器，经过反复训练得到最优参数，建立最终的分析模型。

2.2 数据处理模块的设计

为提升健康大数据分析处理方面的效率，在数据处理模块中，多采用分布式计算模式进行设计。在这种模式下，利用多台计算机设备共同发挥作用，完成预期的数据分析任务。考虑到健康大数据分析环节对计算效率和算法效率均有较高的需求，因此在实际设计中，设计人员将分布式计算中常用的Hadoop和Spark两种计算模式予以整合，使之处于一个集群中，以打造基于Hadoop框架，融入Spark计算模式的Map/Reduce分布式计算模型。在Map/Reduce分布式计算模型当中，Spark计算模型的主要作用是：（1）在大数据的预处理环节中，调入部分原始数据进行数据清洗和加权处理；（2）在大数据分析计算过程中，利用多节点并行计算的方式，完成对算法的计算任务调度，将分割事务集和剪枝策略结合起来，以降低数据处理环节的繁琐程度。同时，在此环节当中，由于扫描Value的过程直接在Spark计算框架内完成，以往数据分析中的内存不足等问题也得到了有效的解决。

整体来看，在这种设计模式下，数据的分析处理环节主要分为以下两个步骤：其一，利用Hadoop框架下的Map/Reduce分布式编程模型，将整个数据库中的数据分解为数个连续的数据片段，每个数据片段分别对应一个储存设备；其二，以并行扫描的方式，统计各个储存设备的局部计算的数据集，每个数据集均由相对应的mapper进行处理，处理完成后的结果首先

保存在分布式缓存当中，而后各个分部处理的结果将集中映射到一个全局的节点，并按降序排序。

2.3 数据应用展现

为实现数据的可视化，在实际设计中，工作人员通过设置人机交互接口，构造数据应用展现模块，用户通过该交互接口，即可实现数据分析应用操作。在实际操作过程当中，医院管理人员可接受应用主题数据分析应用的可视化结果，如各数据的统计分析和医疗数据的共享查阅；同时，医务人员可输入患者数据和系统中最优治疗方案选择的分析计算结果进行对比分析，以选择最为合理的治疗方案，兼顾成本、效益和患者实际情况等多方面的需要。

3 健康大数据分析方法的实际应用

3.1 案例概况

为探寻本次健康大数据分析方法的实际应用效果，本次通过数据库导出方式，从某三甲医院的数据库中采集1000条数据，这些数据均为该医院收治的糖尿病患者数据。在获得数据后，对数据进行相应处理操作，处理结果显示，在本次所取得的数据中，共计存在64条住院记录，共采用治疗方案236种，分为药品、检查和检验三种类型。同时在所有入院记录中，共计检查127次，检验474次。

3.2 数据分析

在本环节的分析工作中，其基本思路如下：基于本次所采集到的数据，建立患者特征（主要包括就诊过程中的主诉、检查结果两方面内容）与患者最终的明确诊断，明确诊断所需的检查工作，以及能够兼顾成本和效益两方面的最优治疗方案，根据以上几项指标，建立统计模型分析各项指标之间的关系。基于该统计模型，医务人员可确定患者的诊断内容，并在此基础上进一步推导出最为合理的治疗方案。在实际操作时，由医务人员在系统上选择患者的特征情况，模型即可自动比对相应数据，给出初步诊断、应进行的检查内容和有效的治疗方案等多方面的内容。具体来看，本环节的分析的关键点则是对患者数据进行分类。

为实现分类目标，本次选择层次聚类分析法进行，其分类依据主要是治疗方案和住院时间两方面的指标。系统自动分析后显示，大部分患者的住院天数均在7-14天范围内，且治疗方案数目大部分位于20-45项的范围内。根据此推断结果可知，数据集中的患者在该医院就医时，分为四种情况：（1）患者在该医院进行了全流程（或绝大部分）的治疗环节；（2）患者在该医院仅进行了前期治疗；（3）患者在该医院仅进行了后期治疗；（4）无法归类到前三类的其他情形。根据以上四种情况可见，患者就医实际情况存在较多差异，这种差异也必将造成最终治疗方案上的差异，因此，在确定最终治疗方案前，对患者实际情况做进一步分析，本环节采用K均值聚类法进行分析，采用5类聚类法进行分析，分析结果显示，类别1-5所对应的样本数分别为4、15、19、17、9，具体的聚类中心数据则如表1所示。

表1 均值聚类法各类的类中心数据

类别	药品	检查	检验	治疗	护理	住院天数
1	25.5	2.75	11.25	20.75	1.5	8.75
2	11.6	3.5	5.8	8.1	1.1	6.2
3	14.4	3.5	7.2	12.7	1	9.5
4	5.9	2.8	6.4	6.9	1.1	6.9
5	2.2	2.9	10.7	2	0.7	4.9

从表1中的数据可见，各类患者在特点上均存在较为明显的差异，因此通过患者特点进行归类，再进行治疗方案的选取和评价是具有现实可行性的。

3.3 治疗方案的确定

根据SVM算法，对治疗方案的最大频繁项集进行求解，其求解结果如表2所示。

表2 各个分类中的频繁项集及支持度

类别	频繁项集	支持度
1	0.9%氯化钠注射液；12导联同步心电图检查；丹参川弓嗪注射液；糖化血红蛋白-1检验	0.2142857
2	0.9%氯化钠注射液；12导联同步心电图检查；HbA1C检验；大生化-1检验；肝胆脾胰检查；化验尿；下肢动脉检查；激素检验	0.2
3	0.9%氯化钠注射液；12导联同步心电图检查；HbA1C检验；颈动脉检查；下肢动脉检查；心脏检查；血常规；化验尿；尿微量白蛋白定量-1检验	0.1875
4	0.9%氯化钠注射液；12导联同步心电图检查；HbA1C检验；化验尿；大生化-1检验；心脏检查；血常规	0.1764706
5	ASO检验，ESR检验，hsCRP检验，RF检验；TNI检验，甲功五项和凝血四项检查；尿液分析；酶检验；激素检验	0.2857143

从表2中的数据信息不难看出，在不同患者类下，包含项数最多中支持度最高的频繁项集数量均为1，这表明健康大数据的分析可获得具有唯一性和一致性的结果，这种结果可初步认为是不同类患者的治疗方案选择推荐。同时，从本次实验当中可知，这种健康大数据的分析及应用方法能够有效展现出医疗健康大数据中的数据关联规律，对于辅助诊断治疗方面的作用也较为突出。

4 结束语

整体来看，健康大数据分析及应用是一项综合性较强的工作，其需要整合多方面的技术与功能模块，将具体分析和应用逐步细化，引入更多的业务内容，实现对数据的有效分析和应用，从本次研究中也不难发现，这种基于大数据技术的分析应用取得了初步的成绩。当然，在今后的工作中，仍需要加强对相关理论及技术的研究，从而不断提升数据分析质量。

参考文献：

[1]刘博,邓舒平,杨楠,等.智能网络下职业健康风险大数据分析方法[J].信息技术,2021(05):128-134.

[2]师小勤,赵杰,王琳琳,等.基于大数据分析技术的精准医疗应用综述[J].中国医院管理,2021,41(05):26-31.

[3]李军.基于疫情常态化的校园健康大数据分析与应用[J].电脑编程技巧与维护,2021(03):100-102.

作者简介：

崔民权（1987-）男，朝鲜族，吉林延边人，大学本科，研究方向：统计学、大数据科学与应用。对外经济贸易大学统计学学院在职人员高级课程研修班学员。