基于Spark的分布式健康大数据分析系统设计与实现

吴 磊,欧阳赫明

(北方工业大学 信息学院,北京 100144)

摘 要:随着各类医疗健康信息数量的增长,如何利用医疗健康大数据辅助临床诊疗和科研,已经成为各医疗科研机构普遍关注的问题。针对该问题,设计并实现了一种基于Spark的分布式健康大数据分析系统。系统采用大数据分析技术并基于随机森林模型,应用多个弱分类器将多个决策树获得的结果进行集成,基于该模型实现了睡眠质量预测,同时还研究了权重较高的影响因子。最终实验预测准确率达96.84%。实验结果对于睡眠质量分析具有一定参考意义,且系统能够较好地满足健康大数据的分析处理需求。

关键词:大数据;大数据分析;Apache Spark;智能健康;机器学习;随机森林

DOI:10.11907/rjdk.192546 **开放科学(资源服务)标识码(OSID):**

中图分类号:TP319 **文献标识码:**A **文章编号:**1672-7800(2020)007-0099-04

The Design and Implementation of Distributed Health Big Data Analysis System Based on Spark

WU Lei, OUYANG He-ming

(*School of Information*, *North China University of Technology*, *Beijing* 100144, *China*)

Abstract: With the growth of various types of medical health information, how to use medical health big data to assist clinical diagnosis and research has become a common concern of medical research institutions. Aiming at this problem, we propose a distributed health big data analysis system based on Spark. The system uses big data analysis technology based on the random forest model, and uses multiple weak classifiers to integrate the results obtained by multiple decision trees. Based on the model, the sleep quality prediction is realized, and the influence factors with higher weight are also studied. The final experimental prediction accuracy rate reached 96.84%. The experimental results have certain reference significance for the analysis of sleep quality, and the system can better meet the analysis and processing needs of healthy big data.

Key Words: big data; big data analysis; Apache Spark; smart health; machine learning; random forest

0 引言

近年来,医疗机构信息化程度不断提高,各类医疗健康信息在数量上有着惊人增长。健康大数据具有数据量大、多样性突出的独特性,如何利用健康大数据为临床医疗服务仍然是一个值得讨论的问题。大数据分析的核心问题是如何对这些数据进行有效表达、解释和学习[1]。

Spark[2]是加州大学伯克利分校AMP实验室开发的集群模式计算平台,其框架构建以内存计算为基础。而传统Hadoop中使用的计算平台是MapReduce[3]、MapReduce模型基于磁盘计算,运行计算作业时的磁盘读写有较大的时间和空间开销[4]。由于Spark模型基于内存计算,因而运行速度相比MapReduce更快,适合进行大规模数据处理。

Spark作为当前最流行的大数据处理平台之一,一直受到很多研究者的关注[5]。曹波等[6]在Spark平台上实现了FP-Growth算法的并行计算,利用车牌记录跟踪车辆;王虹旭等[7]在Spark平台上设计了一个并行数据分析系

收稿日期:2019-11-19

基金项目:北京市社会科学基金项目(18JYB015,18SRB003)

作者简介:吴磊(1963-),男,硕士,北方工业大学信息学院副教授、硕士生导师,研究方向为大数据应用、人工智能应用、数据采集与处理;欧阳赫明(1995-),男,北方工业大学信息学院硕士研究生,研究方向为大数据分析。本文通讯作者:吴磊。

统,该系统能够对海量数据进行高效分析。针对医疗健康大数据分析带来的多种挑战,很多研究者也进行了相关研究。罗辉等在[8]大数据环境下实现了科研专病数据库系统平台,对临床数据进行了集成整合与统计分析,但导入及处理数据的速度还有待提高;甘伟等[9]设计并实现了基于Hadoop分布式存储的大数据临床科研平台,并集成R语言实现了基本统计分析及高级挖掘算法,但机器学习结果的准确度较低。

本文设计并实现了一种基于Spark的分布式健康大数据分析系统,利用弹性分布式数据集RDD(Resilient Distributed Dataset)[10]对数据进行相应操作,选取Spark on YARN[11]集群模式运行,相比传统的Hadoop平台具有更高的容错性和更快的运算速度;系统实现的基于随机森林的睡眠质量预测方法具有较高的预测准确率,达到了96.84%。

1 分布式Spark集群搭建

系统构建实验中,使用1台物理机中的3台Linux虚拟机组成拥有3个节点的分布式集群,其中包括1个Master节点和2个Worker节点。Master节点用于分配任务以及维护状态,因此采用的配置相对于Worker节点而言较高。处理器配置方面,Master节点机器使用2个4核处理器,Worker节点机器使用1个4核处理器。内存配置方面,Master节点机器使用6GB内存,Worker节点机器使用4GB内存。网络连接方式全部采用NAT方式,以便对节点的IP地址进行配置和管理。物理机使用M.2 NVMe协议的固态硬盘,固态硬盘相比于机械硬盘具有读写效率高、寿命长等优势,对集群运行速度与工作质量有一定保障。由分布式集群所构成的系统,不仅对节点故障有一定容错性,而且能够依据需求调整节点数量。

1.1 Hadoop集群搭建

系统的Spark运行于YARN之上,需要预先安装Hadoop,而Hadoop又需要JDK的支持,因此首先需要安装JDK,然后配置各台虚拟机之间的ssh免密码登录及防火墙,最后解压Hadoop安装包并修改相关的配置文件。配置文件包括hadoop-env.sh、core-site.xml、hdfs-site.xml、mapred-site.xml、yarn-site.xml和slaves等。

1.2 Spark集群搭建

由于Spark运行需要Scala的支持,因此在安装Spark前需要先安装Scala环境,然后从Spark的官方网站上下载Spark的源码并使用Maven编译。Spark文件配置需要将所有Worker节点的主机名写入每台虚拟机中的Slaves文件中,并修改节点的Spark安装目录的Spark-env.sh文件。同时,集群所有节点的Spark-env.sh文件和Slaves文件的内容要保持完全一致[12]。完成以上配置后,使用Spark on YARN的方式启动Spark集群。启动后可以通过jps命令或在Master节点上使用浏览器访问localhost:8080查看启动情况,并可以通过Spark-submit提交一个Spark中的示例作业以测试集群运行情况。

1.3 Spark开发环境配置

系统的Spark应用程序使用的开发语言为Scala,因为Spark是Scala编写的,因而对Scala的支持性最好。IDE选择业内广泛使用的IntelliJ IDEA,该软件提供的Scala插件可以很好地支持Spark程序开发。调试时使用Spark Local模式运行Spark,可以直接在开发环境中调试而不必将作业提交到Spark集群之上。

2 睡眠质量预测实现

本文设计的数据分析系统通过Spark的MLlib实现了建模方法。系统首先将样本数据集分为训练数据集和测试数据集两部分,再通过连接各种功能函数的操作节点,并形成流程实现数据分析建模功能。

2.1 随机森林算法

系统采用随机森林算法实现了睡眠数据预测。选取睡眠数据的属性创建相应数据集,并提取相应特征向量建立分类模型。将系统采用的数据集分为两个部分:70%作为训练数据用于训练模型,30%作为测试数据用于测试模型。

随机森林(Random Forest,RF)[13]利用节点随机分裂技术和随机重采样技术构建多棵决策树,分类结果由投票决定。它具备了分析复杂相互作用分类特征的能力,且对于缺失值和噪声具有很好的鲁棒性。此外,随机森林的学习速度也较快。随机森林可以作为高维数据的特征选择工具[14],近年来已被广泛应用于各种分类及预测等问题中[15]。

单棵决策树普遍会存在过拟合现象,为避免这种现象,系统采用了随机森林算法,即利用机器学习的集成学习思想,通过构造多个弱分类器并最终合成为一个强分类器的方法,不仅有效减少了过拟合现象,而且提高了预测精度[16]。

随机森林是用多棵树对样本进行训练并预测的一种分类器,每颗树$h(X,\beta_k)$都有一票投票权以选择最终分类结果。分类决策如式(1)所示。

$$H(X)=\arg\max_{Y}\sum_{i=1}^{k}I(h_i(x)=Y) \tag{1}$$

其中,H(x)表示随机森林分类结果,I(·)为示性函数,$h_i(x)$表示单个分类结果,Y代表分类目标。该式取各决策树结果中的多数作为最终结果[16]。

2.2 模型构建

为计算出多个睡眠质量影响因子中权重较大的因子,首先需要构建预测模型,具体方法为:先设置K个弱分类器,其中类别纯度使用Gini系数[17]进行计算,再将相似样本放在同一个弱分类器中,最后使用K-means算法[18]进行训练,并使用均值组合方式。在模型训练完成后,使用另外一组构建好特征的样本,经过模型训练,最后评估模型。

建模过程分为训练和测试两个阶段,如图1所示。在

训练阶段，主要根据计算好特征的样本，划分好 K 个弱分类样本后，再进行随机森林训练。训练完成后，测试数据应用训练好的预测模型可得到预测值，将预测值与实际值进行运算可得到模型的精度值，从而评估模型性能。

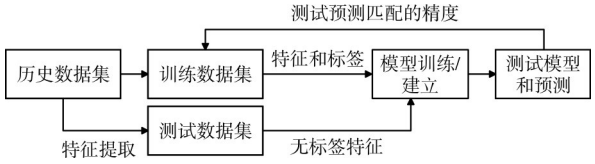

图 1　建模流程

系统针对睡眠质量的多个影响因素展开研究，数据采用 Kaggle 公司[19]提供的 Sleep Cycle 从 2014-2018 年的相关原始数据。Kaggle 公司于 2010 年创立，并于 2017 年被谷歌公司收购，主要是为数据科学家和开发商提供数据分享以及举办竞赛的平台[19]。目前，许多科学家和开发者都纷纷入驻这一平台。

系统采用的数据包含了睡眠相关的 8 个属性，主要有：Start、End、Heartrate 等，经过预处理后的数据如表 1 所示，其中用 Sleep Quality 属性值表示睡眠质量的好与差。

（1）构建影响因子特征向量。数据集中每条样本采用两个类别进行标记：-1（差）和 1（好），每个样本的特征包含如下字段：在数据属性中 Sleep quality 用来表示睡眠质量（-1 或 1）。

表 1　睡眠特征变量

序号	属性	示例	说明
1	Start	2014-12-29 22:57:49	睡眠开始时间
2	End	2014-12-30 07:30:13	睡眠结束时间
3	Time in bed	7:32	在床上的时间
4	Wake up	1	是否中途醒来
5	Stressful	1	压力自我备注
6	Heartrate	59	心率
7	Activity(steps/k)	0	运动步数（单位：千）
8	Sleep quality	-1	睡眠质量

特征向量选取原始数据的全部 8 个属性进行构建，特征：{“Start”，“End”，“Time in bed”，“Wake up”，“Stressful”，“Heartrate”，“Activity(steps/k)”，“Sleep quality”}，将 Start、End 和 Time in bed 中的时间提取出来并转换成小时，再对每个维度的特征做变换后返回 Dataframe，并增加标签列 label，其中数值 1 表示好，数值 0 表示差，如表 2 所示。

表 2　数据文件

	Start	End	Time in bed	Wake up	Stressful	Heartrate	Activity(steps/k)	Sleep quality	Features	Label
1	22.96	7.50	7.53	1	1	59	0	-1	[22.96,7.50,7.53,1,……]	0
2	23.71	6.41	7.30	1	1	54	6.4	-1	[23.71,6.41,7.30,1,……]	0
3	22.43	8.91	10.48	-1	-1	62	7.4	1	[22.43,8.91,10.48,1,……]	1
4	21.98	5.98	7.13	1	1	64	4.8	-1	[21.98,5.98,7.13,1,……]	0
5	22.52	6.35	7.84	1	1	61	0.2	-1	[22.52,6.35,7.84,1,……]	0
·	·	·	·	·	·	·	·	·	·	·

（2）训练随机森林分类器。系统训练随机森林分类器的主要参数有：maxDepth：树的最大深度；maxBins：最大分桶个数，用于决定每个节点如何分裂；auto：每个节点分裂时是否自动选择参与特征的个数；Seed：随机数生成种子。

系统采用的参数为：maxDepth：3、maxBins：20、auto：”auto”、Seed：4073。

2.3　实验分析

实验数据集共有 887 条数据，其中 Sleep quality 属性值表示睡眠质量的好/差，为了分析影响因素与睡眠质量之间的相关度，实验中选取了皮尔逊相关系数[20]进行表征，并将各属性的相关程度进行排序，如表 3 所示，各属性相关系数如图 2 所示。

表 3　属性相关度排序

排序	属性	相关系数
1	Start	0.732
2	Time in bed	0.711
3	Heartrate	0.672
4	End	0.647
5	Wake up	0.635
6	Stressful	0.597
7	Activity(steps/k)	0.421

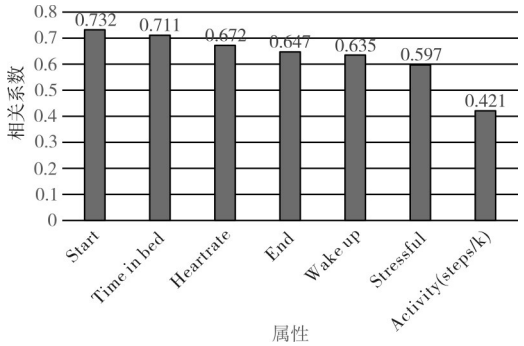

图 2　属性相关系数

通过计算各属性与睡眠质量的相关度可知，Start、Time in bed、Heartrate 影响因子对于睡眠质量的影响程度较大，对睡眠质量的相关研究具有一定借鉴意义。

系统采用管道学习训练模型，即 Pipeline。在机器学习中，通常有一系列的算法在数据中处理和学习。Spark MLlib 提供的机器学习算法 API，可以将多个算法组合成一个独立管道，之后管道会通过在参数网格上的不断爬行自动完成模型优化，最后系统进行预测时会使用通过管道训练得到的最优模型。预测结果中 Prediction 标签为最终预测结果，如表 4 所示。

表 4 实验预测结果

	Start	End	Time in bed	Wake up	Stressful	Heartrate	Activity(steps/k)	Sleep quality	features	label	prediction
1	22.96	7.50	7.53	1	1	59	0	-1	[22.96,7.50,7.53,1,……]	0	0
2	23.71	6.41	7.30	1	1	54	6.4	-1	[23.71,6.41,7.30,1,……]	0	0
3	22.43	8.91	10.48	-1	-1	62	7.4	1	[22.43,8.91,10.48,1,……]	1	1
4	21.98	5.98	7.13	1	1	64	4.8	-1	[21.98,5.98,7.13,1,……]	0	0
5	22.52	6.35	7.84	1	1	61	0.2	-1	[22.52,6.35,7.84,1,……]	0	0
·	·	·	·	·	·	·	·	·	·	·	·

将 Label 标签值与 Prediction 标签值进行比较得到模型的预测精度值是 96.84%，其中包含准确预测条数 859 条，如图 3 所示。预测结果表明，预测数据与原始数据拟合度较高。

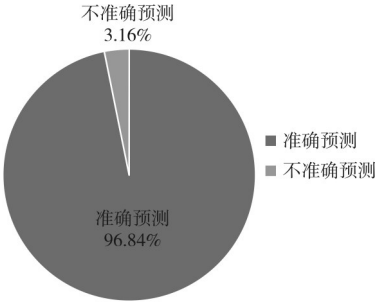

图 3 模型预测结果

系统采用基于 Spark 和随机森林算法的机器学习训练方法用于睡眠质量预测，取得较高准确率，证明了采用随机森林算法构建的睡眠数据预测机制较为成功，具有一定参考意义。

3 结语

本文基于 Spark 设计并实现了一种分布式健康大数据分析系统。系统采用基于随机森林模型的大数据分析技术，将多个决策树得出的结果进行分析集成，训练模型采用管道学习方法，并将其应用到睡眠质量预测场景中，实验分析得出该模型预测精度值为 96.84%。同时，通过相关度分析获得了与睡眠质量相关度较高的 3 个影响因素 Start、Time in bed、Heartrate，可以用作睡眠质量分析指标。同时，系统还有很多待改进之处，如集群运行参数、模型训练参数调优等。

参考文献：

[1] 程学旗，靳小龙，王元卓，等. 大数据系统和分析技术综述[J]. 软件学报，2014，25(9)：1889-1908.
[2] 李星，李涛. 基于 Spark 的推荐系统的设计与实现[J]. 计算机技术与发展，2018，28(10)：194-198.
[3] 高莉莎，刘正涛，应毅. 基于应用程序的 MapReduce 性能优化[J]. 计算机技术与发展，2015，25(7)：96-99，106.
[4] 于海浩. 基于 Spark 的抄袭检测云计算框架研究[J]. 计算机光盘软件与应用，2014，17(11)：110-112.
[5] 张恬恬，孙绍华. 基于 Spark 的云计算平台在实验室的应用与实现[J]. 软件导刊，2018，17(4)：191-193.
[6] 曹波，韩燕波，王桂玲. 基于车牌识别大数据的伴随车辆组发现方法[J]. 计算机应用，2015，35(11)：3203-3207.
[7] 王虹旭，吴斌，刘旸. 基于 Spark 的并行图数据分析系统[J]. 计算机科学与探索，2015，9(9)：1066-1074.
[8] 罗辉，薛万国，乔岫. 大数据环境下医院科研专病数据库建设[J]. 解放军医学院学报，2019(8)：713-718.
[9] 甘伟，徐明明，陈联忠，等. 大数据临床科研平台的设计与实现[J]. 中国数字医学，2019，14(2)：40-43.
[10] GENG Y S. Spark standalone mode process analysis and data skew solutions[C]. Proceedings of 2017 IEEE 2nd Information Technology, Networking, Electronic and Automation Control Conference, 2017：647-653.
[11] 杨玉，张远夏. Spark on Yarn 模式的电信大数据处理平台[J]. 福建电脑，2019，35(3)：34-38.
[12] 李艳红. 基于 Spark 平台的大数据挖掘技术分析[J]. 科技资讯，2018，16(27)：7-8.
[13] BREIMAN L. Random forests[J]. Machine Learning，2001，45(1)：5-32.
[14] STROBL C，BOULESTEIX A L，KNEIB T，et al.Conditional variable importance for random forests[J]. BMC Bioinformatics，2008.
[15] 姚登举，杨静，詹晓娟. 基于随机森林的特征选择算法[J]. 吉林大学学报(工学版)，2014，44(1)：137-141.
[16] 苗立志，刁继尧，娄冲，等. 基于 Spark 和随机森林的乳腺癌风险预测分析[J]. 计算机技术与发展，2019(8)：1-3.
[17] 刘星毅. 一种新的决策树分裂属性选择方法[J]. 计算机技术与发展，2008(5)：70-72.
[18] 唐浩，杨余旺，辛智斌. 基于 MapReduce 的单遍 K-means 聚类算法[J]. 计算机技术与发展，2017，27(9)：26-30.
[19] 邓仲华，刘斌. 数据挖掘应用热点研究——基于 Kaggle 竞赛数据[J]. 图书馆学研究，2019(6)：2-9，23.
[20] 姜亚斌，邹任玲，刘建，等. 表面肌电信号的下肢痉挛信号特征分析与识别[J]. 电子科技，2017，30(11)：38-41.

（责任编辑：孙 娟）