



专题:大数据

随机森林在运营商大数据补全中的应用

王铮,任华,方燕萍

(中国电信股份有限公司上海研究院,上海 200122)

摘要:电信运营商有大量数据,但是鉴于多种原因,数据的质量不够理想,出现大量数据不完整甚至缺失。对于已有数据的挖掘,必须在数据满足质量要求且达到足够采样比例的前提下开展。依托现有的全国日志留存系统,设计完整数据的模板样库,鉴别不能满足质量要求的数据,使用随机森林算法,找到最符合的相同或相关数据,补全数据并提升数据质量;用回溯反馈的方法优化并扩充模板样库。在全国日志留存系统中构建数据补全子系统,实现端到端的数据质量保障和提升,补全并改善历史数据甚至实时数据的质量,最终满足数据处理和挖掘的要求,提升运营商数据质量和价值。

关键词:大数据;随机森林;机器学习;数据补全

中图分类号: TN919.5

文献标识码: A

doi: 10.11959/j.issn.1000-0801.2016317

Application of random forest in big data completion

WANG Zheng, REN Hua, FANG Yanping

Shanghai Research Institute of China Telecom Co., Ltd., Shanghai 200122, China

Abstract: Telecom operators have a lot of data, but in view of a variety of reasons, the quality of the data is not ideal, there are a lot of data is not complete or even missing. For existing data mining, it is necessary to carry out the data to meet the quality of the data and to achieve sufficient sampling proportion. Relying on the country's existing log retention system, template library design data integrity, authentication could not meet the quality requirements of the data, using the random forest algorithm, the same data with or related data was found, data was completed and data quality was improved, and the template library was extended by optimization of feedback. The construction of completion data subsystem in the system log retained end-to-end data quality guaranteed and improved quality, completed and improved the real-time data and historical data, and ultimately met the requirements of data processing and mining operators, improved data quality and value.

Key words: big data, random forest, machine learning, data completion

1 引言

电信运营商是天然的大数据拥有者,拥有着基于用户的信令、上网、位置等多种类型的数据。随着近年大数据技术发展和应用推广,电信运营商也愈发重视数据这一战略

资产,研发了多种基于大数据技术的平台和系统,用来收集、存储、处理、开放和应用电信运营商的数据,体现电信运营商的数据价值。但由于历史的原因,电信运营商的各种设备和系统,设计和建设的年代不同、承建的厂商不同、不同省份的需求不同、后期的升级状况不同等,造成了数



据质量参差不齐。最常出现数据缺失、数据不全等现象,导致存储和收集的数据质量不高,甚至出现因为可采样数据比例过低,造成大批量的数据不可用。

在运营商全国日志留存系统的大数据集约运营应用中,数据采集、汇聚、上传在省一级实现,数据接收、关联入库、查询和开放共享则在全国一级实现,在接收后关联入库前会对收集的数据预做完整性检验。但是由于较多数据完整性不足,导致可用数据不多。本文从解决全国日志留存系统当前面临的数据完整性不够、数据质量差的问题出发,提出将随机森林算法引入数据补全中,以提升数据完整性的技术方案。

2 基于大数据架构的全国日志留存系统

用户网络行为信息是电信运营商数据信息的主要部分,也是进行用户行为画像的主要数据源,支持辅导预测、评估、决策等多项管理或商务行为,具有丰富价值和潜力。电信运营商的全国日志留存系统就是对相关信息进行收集、存储、挖掘和服务共享的系统,为基于电信运营商的大数据开展服务提供数据挖掘、共享等数据增值服务。

2.1 总体功能概述

(1) 数据采集

各省数据采集机需要采集上网日志数据以及业务网络数据并提供缓存,上网日志数据包括 WLAN 方式下,在固网 AAA 平台的 DPI 设备(或者类似设备)上取得的认证信息;分组域 DPI 上取得的认证信息和互联网访问信息;WAP 网关上取得的认证信息和互联网访问信息;WAP 网关防火墙、融合防火墙上取得的 NAT 信息;4G DPI 上取得的认证信息和互联网访问信息。本期工程只采集上网日志数据,业务网络平台及其他平台的数据采集未来逐步扩展增加。

(2) 数据传输

数据采集机定时扫描各数据源相关目录的数据,进行压缩传输等操作(如果已经压缩不必再次压缩,如果没有压缩,则进行压缩),并配有重传机制。

(3) 数据接收

全国接口机对上传的日志数据进行校验、稽核、去冗、清洗等操作后装载入 HDFS(Hadoop distributed file system, Hadoop 分布式文件系统)。

(4) 数据关联入库

对采集机上传的全量日志数据根据规则要求进行关

联入库,以供查询和共享。

(5) 查询功能

通过页面的方式提供相关数据查询操作。

(6) 数据共享

根据各共享数据需求平台的需求,通过特定接口方式提供数据共享功能。

2.2 系统架构

2.2.1 逻辑架构

全国日志留存系统从逻辑上可分为数据采集与传输层、数据接收层、数据处理层、数据服务层以及系统管理层,如图 1 所示。

数据采集与传输层采集各个数据源(DPI、WAP 网关、WAP 网关防火墙、融合防火墙、业务网络数据)上的原始数据并对数据进行压缩(已经压缩过的数据不用再压缩)、上传,并提供重传、缓存等功能。

数据接收层接收各省采集机的数据后对数据格式、规范性、关联性、完整性进行校验、稽核、去冗、清洗等操作并装载到 HDFS 中。接口机还可以将采集到的原始数据进行共享。

数据处理层对原始数据进行加工处理,包括数据管理、数据统计汇总、数据关联入库等功能。

数据服务层提供数据统一访问和共享服务,包括数据查询、数据共享、业务统计分析报表等功能。

系统管理层提供统一的接入访问管理、系统访问和数据安全管理、资源监控和分配管理、任务调度和监控管理、系统的运维、监控和日志管理等功能。

2.2.2 技术架构

全国日志留存系统从技术上可分为数据接收层、数据处理层、数据服务层以及系统管理层,如图 2 所示。

(1) 数据接收层

通过 FTP/SFTP 实现对数据的采集和传输,接收数据后进行文件的校验稽核、合并、切分等功能并装载到 HDFS。

(2) 数据处理层

提供海量数据的存储、查询、分析汇总等功能,用于支持业务的需求场景和应用。

- HDFS: 采用 HDFS 存储原始文件,读写吞吐量高。HDFS 存储包括原始入库数据(DPI、AAA、NAT等)、业务实时数据(HBase)、海量数据清洗和分析汇总(Hive)数据等。
- MapReduce/Yarn: 为 Hadoop 存储下的海量数据进

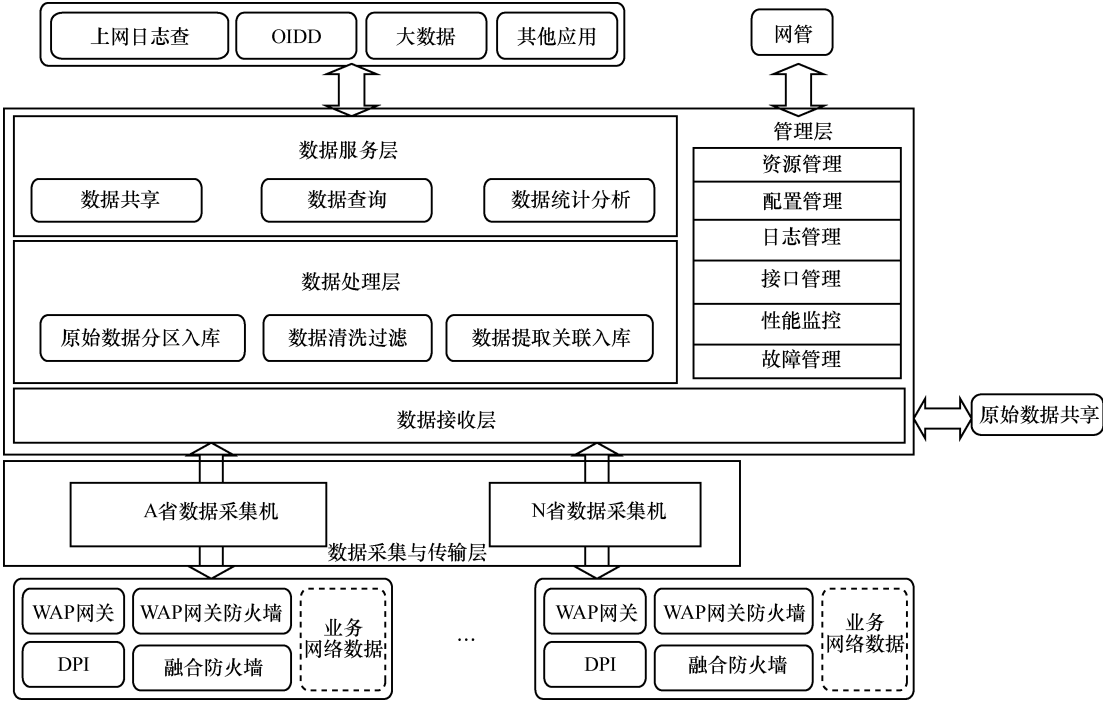


图 1 系统逻辑架构

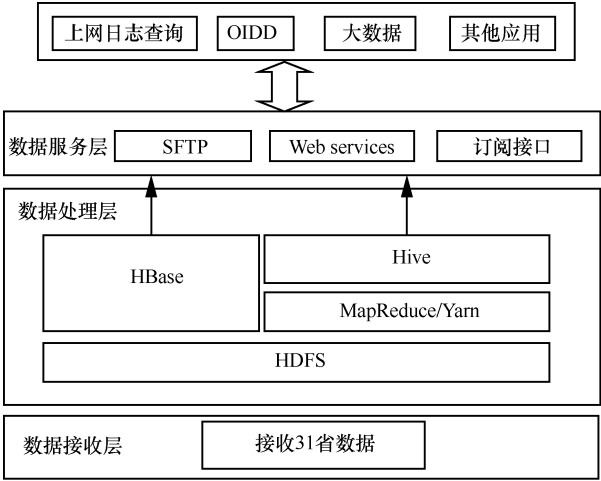


图 2 系统技术架构

行清洗、分析、关联和汇总的计算框架和资源管理。通过 MapReduce 并行运算框架实现日志关联任务的统一调度处理,充分利用集群内资源进行高性能处理;保持以省为单位进行日志关联处理。

- HBase:是一个针对结构化数据的可伸缩、高可靠、高性能、分布式和面向列的动态模式数据库。采用批量装载(BulkLoad)或文件复制的方式将关联后数据输出为 HBase 的内部数据格式,直接装载到 HBase 中,性能高,占用 CPU、网络资源少,用于提

供日志留存系统的对外数据实时查询。

- Hive:基于 Hadoop 的一个数据仓库工具,可以将结构化的数据文件映射为一张数据库表,并提供简单的 SQL 查询功能,可以将 SQL 语句转换为 MapReduce 任务进行运行。

(3)数据服务层

提供系统的对外服务接口,包括数据的文件共享、服务查询、数据分析等功能。

- 支持通过 SFTP/FTP 的批量数据的文件共享,包括 HBase 中实时查询数据文件接口方式的数据共享。
- 支持通用的 Web services 的消息接口,用于数据的访问、业务查询、数据分析等功能。
- 支持通过订阅方式进行数据共享(订阅接口需根据不同的应用开发)。

(4)系统管理层

- 服务管理:用于针对服务的接入管理、服务统计报表等功能。
- 安全管理:管理用户和服务在数据平台的安全管理,包括用户管理、访问授权、管理授权、数据授权等,支持对用户访问和数据的安全管理。
- 资源管理:管理 Hadoop 平台下的资源分配和资源



使用监控等。

- 任务管理:维护和监控服务发布任务的执行情况,支持任务的依赖、关联执行。
- 系统管理:包括平台的基本配置信息、服务和平台的日志信息、系统平台的监控和维护、平台告警、工单管理等功能。
- 数据管理:提供统一的前台数据管理功能,管理数据平台的数据全生命周期。

2.3 当前存在的问题

全国日志留存系统在清洗、处理收集的数据时,会检查数据的完整性。数据完整性检查,经常会发现一些关键数据不全,部分字段数据丢失(为空),导致数据检查不符合系统阈值,无法入库,最终导致入库数据过少,采样数据过少,影响到后继的数据分析、挖掘等工作。

根据现有全国日志留存系统发现的问题,重点要解决以下2个方面问题。

- 对于一些缺失但非关键数据字段,在清洗使用时,做一定的填充,提高数据完整性。
- 对于一些关键但不完整的数据字段,通过一些经验数据进行补全,提升数据完整性和质量。

3 随机森林技术及应用

随机森林(random forest, RF)是由Leo Breiman于2001年提出来的,是一个可处理高维度和非线性样本的分类器组合模型,并在众多领域得到了广泛应用^[1]。随机森林是一种综合性的分类方法,分类的准确率高^[2]。其实质是一个组合分类器,其中决策树是其核心。它不仅可以用来做分类,还可以用来做回归。随机森林相关的理论及应用实例极多,在此仅做简单介绍分析。

随机森林核心是随机样本、随机特征、决策树搭建、随机森林分类四大部分。其中,随机性体现在前两类:即在每次建树时,在样本全集中随机取样,训练集呈现的随机性^[3];在每个节点分裂时,从全集 M 个属性中选取 x 个属性,呈现特征属性的随机性^[4]。

(1) 随机样本

给定一个训练样本集,数量为 N ,本文使用有放回采样到 N 个样本,构成一个新的训练集。注意这里是有放回的采样,所以会采样到重复的样本。详细来说,就是采样 N 次,每次采样一个,放回,继续采样。即得到了 N 个样本。然后把这个样本集作为训练集,进入下一步。

(2) 随机特征

在构建决策树的时候,主要就是在一个节点上,计算所有特征的ID3(information gain)或者C4.5(gain ratio),然后选择一个最大增益的特征作为划分下一个子节点的走向。但是,在随机森林中,本文不计算所有特征的增益,而是从总量为 M 的特征向量中,随机选择 m 个特征,其中 m 可以等于 $\sqrt{M}$,然后计算 m 个特征的增益,选择最优特征(属性)。注意,这里的随机选择特征是无放回的选择!

(3) 决策树搭建

根据样本集,搭建决策树。用随机特征选择方法进行节点最优分类特征的计算。一般用ID3或者C4.5等作为选择特征的标准。

例如:集合 Y 包含 i 个类别的记录,那么其Gini指标为:

$$gini(Y)=1-\sum_{i=1}^m q_i^2 \quad (1)$$

其中, q_i 为类别 i 出现的频率。

如果集合 T 分成 n 部分 $M_1, M_2, \dots, M_n$ 。那么这个分割的Gini为:

$$gini_{split}(Y)=\frac{M_1}{M} gini(Y_1)+\dots+\frac{M_n}{M} gini(Y_n) \quad (2)$$

一般采用选择具有最小Gini的属性为分裂属性的选择规则。

(4) 随机森林分类

重复上述过程 N 次,就得到了 N 棵决策树。输入一个测试样本,用已有的每颗决策树对它分类,得到 N 个分类结果。最后,使用简单的投票机制获取平均值,得到最终分类结果^[5]。简单投票机制包括一票否决、一致表决、阈值表决、贝叶斯投票机制等。

随机森林是一个用随机方式建立的,包含多个决策树的分类器。其输出的类别是由各个树输出的类别的众数而定。当前,随机森林在各个领域都有应用,主要应用方法则偏向于模型建立、回归等。

4 采用随机森林算法的全国日志留存系统数据补全解决方案

最理想的情况是,数据补全应该在数据源头进行数据上传时进行,这样既可以保证上级系统汇集的数据都是高质量的有效数据,又可以减少无用数据的传输和存储,且降低并分担数据汇聚系统的数据处理量。但考虑到运营商网已有系统的改造量大,且初期需要尽量保证数据汇聚

平台中数据的原始性,所以数据补全可以在数据汇聚平台进行。

在现有全国日志留存系统中,当务之急是解决历史数据的完整性问题,通过筛选系统定义数据的关键和非关键字段,建立正确的模板作为采样集,并通过训练集和决策树构建和分类,最终选取最适当的形式对数据进行补全。

随机森林进行数据补全技术解决方案包括随机森林补全数据、不影响结果的反馈回归验证以及与现有平台的融合。

4.1 随机森林处理数据补全技术解决方案

在现有全国日志留存系统中或在整体体系架构中,定义数据补全子系统,如图3所示。其中,数据检测与修复模块中采用随机森林算法。

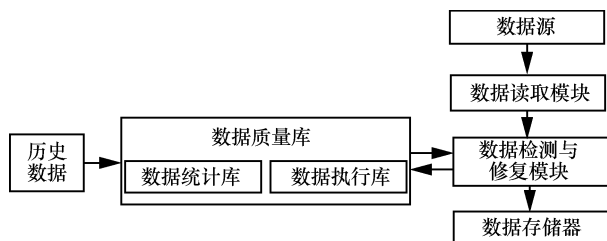


图3 数据补全子系统逻辑架构

数据源通过数据读取模块读入,数据检测与修复模块根据数据源的类型读入数据质量库中相关的数据执行表,并执行检测与修复任务,执行完毕保存到数据存储器。

数据质量库由数据统计表和数据执行表组成,数据统计表内存储每种业务历史数据的统计情况,数据执行表由具体业务需求确定,从数据统计表内抽取相关日期或相关种类业务数据统计表,生成数据执行表,执行数据检测和修复任务。

在与现有系统融合时,数据补全子系统可以以Spark 架构作为功能模块载入全国日志留存系统中,实现数据处理部分的功能。

4.2 数据补全效果反馈回归子系统技术方案

一般来说,随机森林算法通过现成的回归处理,可以对算法本身的正确性进行验证和测试。但是,电信运营商数据补全的最终目的,是提升数据采样率以及数据的质量,最后进行商用。所以,对于数据补全的效果反馈进行回归。本回归的目的,初期是解决当前数据采样率的问题,补全的数据不求提升大数据应用的效果或命中率。所以在回归的效果评定阈值,应该设置为100%。其含义为,数据补

全后百分百不影响数据应用的结果。简单的说,就是数据补全后,用更多的数据构建的模型计算出来的结果与不补全时计算的结果是一样的。

具体实现就是读取若干批次相同业务数据的数据应用结果统计表,采用融合权重且自适应的算法(式(3))计算出数据结果执行表的各执行参数,并据此验证数据补全后的效果。

$$X_i = \sum_{j=1}^m C_{ij} T_j / \sum_{k=1}^n C_{ij} T_j \sum_{l=1}^{i=n} C_{ij} T_k \quad (3)$$

这样既可以避免造成补全决策错误造成数据完整度高,但效果反而更差的情况;也尽量不要因为数据完整度高而效果提高,造成的决策正确的假象。总而言之,在数据补全的初期,先求数据利用率和价值的提升、不求数据应用结果的提升。

在数据补全任务结束后,效果回归子系统评估修复效果,与这批数据的无瑕疵部分的数据应用结果进行比较,越接近越优。针对静态数据可通过调整相关参数多次实验,针对流式数据可间隔一定时间重新生成数据执行表投入使用,以此不断完善数据质量库。

5 结束语

随机森林用于数据补全,优点是相当明显的,对于如全国日志留存系统这类数据噪声小的系统中,实现简单、效率高。但是,对电信运营商而言,在全国层面数据过多,且当数据应用实例扩充后,这样的数据补全及效果反馈回归子系统会导致更多的计算量和资源开销。所以,在数据补全系统逐步演进的过程中,对于初级的数据补全和效果反馈回归可以在升级数据上传时进行,而在全局层面则可以开展多系统融合、内外数据关联后的数据补全及效果反馈回归。这样既可以分担全国系统的工作压力提高效率,又可以全国和省两级的数据补全联动、质量提升分工。当然,在今后的补全中,不仅仅是随机森林,其他的分类聚类等各种算法都可以逐步引入,以适应不同的需求和场景。

参考文献:

- [1] BREIMAN L. Random forests[J]. Machine Learning, 2001, 45(1): 5-32.
- [2] 李慧. 一种改进的随机森林并行分类方法在运营商大数据的应用[D]. 成都: 电子科技大学, 2015.



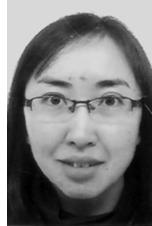
LI H. An improved random forest parallel classification method and its application to big data of telecom operators [D]. Chengdu: University of Electronic Science and Technology of China, 2015.

- [3] BREIMAN L. Bagging predictors [J]. Machine Learning, 1996, 24(1): 123-140.
- [4] DIETTERICH T. An experimental comparison of three methods for constructing ensembles of decision trees: bagging boosting and randomization [J]. Machine Learning, 2000 (40): 139-157.
- [5] 方匡南, 吴见彬, 朱建平, 等. 随机森林方法研究综述 [J]. 统计与信息论坛. 2011(3): 32-38.
- FANG K N, WU J B, ZHU J P, et al. A review of technologies on random forests [J]. Statistics & Information Forum, 2011(3): 32-38.
- [6] 曹正凤. 随机森林算法优化研究 [D]. 北京: 首都经济贸易大学, 2014.
- CAO Z F. Study on optimization of random forests algorithm[D]. Beijing: Capital University of Economics and Business, 2014.
- [7] 黄师师, 黄哲学. 随机森林理论浅析[J]. 集成技术, 2013, 2(1): 1-7.
- HUANG S S, HUANG Z X. A brief theoretical overview of random forests[J]. Journal of Integration Technology, 2013, 2(1): 1-7.

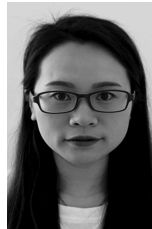
[作者简介]



王铮(1973-),男,中国电信股份有限公司上海研究院工程师,主要研究方向为大数据平台、应用及业务网络。



任华(1977-),女,中国电信股份有限公司上海研究院工程师,主要研究方向为大数据平台和业务平台。



方燕萍(1981-),女,中国电信股份有限公司上海研究院工程师,主要研究方向为大数据和移动互联网领域。