计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2018,27(11):27−34 [doi: 10.15888/j.cnki.csa.006634]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于 ADTree 改进算法的轮胎大数据质量分析①

许晓彬[1], 李敏波[1,2]

[1](复旦大学 软件学院, 上海 200433)
[2](复旦大学 上海市数据科学重点实验室, 上海 200433)
通讯作者: 李敏波, E-mail: limb@fudan.edu.cn

摘　要: 工业企业在生产制造过程中积累了大量的生产数据. 海量的工业数据蕴含了价值巨大的信息, 通过分析、挖掘这些工业数据能够提升企业数字化管理与质量数据分析能力. 本文以轮胎行业制造大数据的应用为背景, 分析了轮胎行业制造大数据的质量分析需求与数据特征, 将轮胎生产各个环节的多源异构数据有效整合, 经过数据预处理流程, 构建了结构化的生产制造与质量检测关联分析数据集. 针对传统 ADTree 算法性能较低的问题, 本文使用优化后的自底向上的归纳方法进行了改进, 充分利用已知数据, 减少了建树时分裂测试评估的计算量. 实验证明, 改进后的 ADTree 算法更适用于大数据量的数据挖掘. ADTree 的挖掘结果经过整理, 可以找出影响轮胎质量的重要因素.

关键词: 工业大数据; 质量分析; ADTree; 数据挖掘; 决策树

引用格式: 许晓彬,李敏波.基于 ADTree 改进算法的轮胎大数据质量分析.计算机系统应用,2018,27(11):27–34. http://www.c-s-a.org.cn/1003-3254/6634.html

Quality Data Analysis of Tyre Industry Based on Optimized ADTree Algorithm

XU Xiao-Bin[1], LI Min-Bo[1,2]

[1](Software School, Fudan University, Shanghai 200433, China)
[2](Shanghai Key Laboratory of Data Science, Fudan University, Shanghai 200433, China)

Abstract: Industrial enterprises have accumulated a large amount of production data. Massive industrial data contain valuable information. By analyzing and mining these industrial data, enterprises can enhance the ability of digital management and quality data analysis. This paper analyzes the demand and data characteristics of big data in tyre industry. First, the multi-source and heterogeneous data in every link of tyre production is integrated. After analyzing the data pre-processing process, we build the analysis data set of structured manufacturing and quality inspection. According to the low performance of the traditional ADTree algorithm, this study uses bottom induction method to make full use of the known data and reduce the amount of calculation. The experiment shows that the improved algorithm is more suitable for a large amount of data. After sorting out the results of ADTree, the important factors that affect the quality of the tires can be found.

Key words: industrial big data; quality analysis; ADTree; data mining; decision tree

1　引言

随着信息化融入工业化进程, 越来越多的工业企业已经完成了自动化、信息化建设[1], 企业产业链的各个环节都涉及到信息技术的应用, 如生产监控、成品检测、产品销售等. 传感器、RFID 等技术与 ERP、MES 等信息管理系统已经应用于制造企业生产经营管理中并积累大量的工业数据. 相比于互联网大数据, 工

① 基金项目: 国家自然科学基金 (61671157); 上海科技创新行动计划 (18511107800)
Foundation item: National Natural Science Foundation of China (61671157); Shanghai Technology Innovation Action Plan (18511107800)
收稿时间: 2018-04-10; 修改时间: 2018-04-28; 采用时间: 2018-05-08; csa 在线出版时间: 2018-09-30

业大数据的数据类型更丰富、来源更多样性[2]. 海量的工业大数据蕴含了价值巨大的生产制造与质量信息, 这些信息能为企业带来丰厚收益[3].

本文选取轮胎行业制造大数据作为工业大数据研究背景, 通过整合轮胎企业各个生产环节的多源异构数据, 构建结构化质量分析数据集; 对质量分析数据集进行决策树或关联分析挖掘, 可以帮助轮胎企业发现产品制造过程中的质量异常及其影响因素, 不仅能够精确定位质量问题, 还能帮助企业改善工艺流程参数, 降低产品的不合格率, 从而实现企业质量与效益的提升. 传统的 ADTree 算法不适用于大数据量的数据挖掘, 本文改进了 ADTree 决策树算法, 提升了其性能, 使其适用于轮胎大数据质量分析.

2 相关研究

随着大数据概念的火热, 国内外对工业大数据的研究也逐渐兴起. Yan 等提出了工业大数据问题的一种框架, 并介绍了智能制造、工业大数据带来的挑战, 如可靠性与安全性[4]. 张洁等[5]提出了一种大数据驱动的"关联+预测+调控"决策模式, 帮助企业深层次地挖掘工业生产规律, 提供精准决策. 杨枝雨使用决策树算法对工业印花质量问题进行了分析, 改善了印花质量的稳定性[6]. 国内外的研究虽然较为系统的阐述了工业大数据的背景、意义及解决方案, 但结合具体行业或企业工业大数据进行详细分析挖掘的实例并不多, 其中一个重要原因是工业大数据必须从工业企业处获得, 即工业大数据领域里, 真实数据的获取是制约学者们开展研究的一个难题[7].

针对制造企业质量异常数据分析, 可以采用 ADTree、FP-Growth[8]等算法. 本文选取的是 ADTree 算法, 在工业大数据应用场景下, 常规的 ADTree 算法在处理大数据方面稍显低效. Pfahringer 等[9]提出了 ADTree 的构建优化方案, 主要将 z 值改进为 Z_{pure}, 作为一种剪裁技术, 但这种方法需要在大量迭代后才有效果, 并且实验中数据集最多只有 50 000 条左右, 效果还有提升的空间. 杨碧姗等[10]提出了一种快速可拓展的 ADTree 优化构建算法 BICA (Bottom-up Induction for Constructing ADTree), 该算法设计了新的数据结构 AVW-set, 这个集合大小不受数据集大小制约. 同时, 该算法提出了自底向上的归纳算法, 避免了一些冗余计算, 提升了评估效率. 但是, 算法中 AVW-set 的生成与合并算法时间复杂度较高, 完全可以进一步优化. 此外, 生成算法中还存在修改零权重值的问题. 本文在 BICA 算法的基础上, 主要针对以上两点进行了改进, 使算法更为完善. 在应用方面,由于 ADTree 算法只能针对二分类问题, 所以将 ADTree 结合实际应用的研究较少, Watcharapasorn 等用 ADTree 算法对营养不良导致病人在手术中出现意外这一问题进行了分析[11]. 本文在改进 ADTree 算法的基础上, 将其应用于轮胎大数据质量分析, 实现算法与实际质量异常的影响因素分析问题相结合.

3 轮胎质量分析需求与数据集成

3.1 轮胎质量分析需求

随着工业市场竞争的越来越激烈, 制造企业要想得到客户的认可, 高质量的产品是不可或缺的[12]. 在大数据时代, 如何利用工业大数据的挖掘技术, 从海量生产制造数据中寻找影响质量的因素, 实现产品质量的有效控制与改善, 从而提高产品质量已经成为急需解决的问题, 这使得质量数据分析成为工业大数据的重要应用需求, 需求包括:

(1) 轮胎产品生产全过程的质量追溯;

(2) 轮胎生产过程的质量合格率统计分析;

(3) 轮胎质量异常的影响因素分析.

质量数据分析流程主要为数据获取、数据预处理、数据分析和分析整理步骤. 其中, 数据分析主要使用数据挖掘来进行, 采用多种算法进行分析可以确保分析的完整性, 起到互补的作用. 轮胎的质量分析可以采用关联分析的方法, 挖掘出轮胎生产环节中的特征指标 (例如主机手、设备、批次、工艺参数等) 与轮胎质量检测结果之间的显著关联关系, 实现对质量问题的追溯. 除了关联分析之外, 针对二分类问题 (如轮胎质量检测分为合格和不合格两种), 可以使用决策树中的 ADTree 算法进行分析, 这也是本文采用的挖掘算法.

总的来说, 产品质量异常数据分析有两个难点:

(1) 由于工业数据体量庞大, 使用传统 SPSS、WEKA 等分析工具效率较低, 一次处理数据量有限, 本文主要使用 HDFS+ Hive+Spark 作为工业大数据质量分析的技术支撑平台.

(2) 传统的 ADTree 算法效率有限, 不太适合大数据分析, 本文优化了 ADTree 算法, 提高了其性能.

3.2 轮胎质量数据集成

轮胎大数据涵盖了轮胎的整个生命周期, 种类较多, 轮胎企业非常看重其中的质量大数据. 轮胎在整个生产过程中重点是硫化与成型工序, 同时轮胎的动平

衡检测是轮胎质量检测中的关键一环[13]. 与动平衡检测结果相关的数据包括轮胎的硫化数据、成型数据.轮胎质量异常数据集中所包含硫化机的温度、压力等属性均是一系列时序数据, 对这些属性进一步细化抽取其统计指标作为辅助性特征, 这些统计特征包括平均值、方差、最大值、最小值等. 对轮胎生产中的时序型数据分别计算上述统计指标, 添加到质量异常数据追溯分析数据集中作为后续分析的基础.

总体来说,轮胎质量数据可以分为两大类数据, 分别是质量检测数据和质量生产数据. 质量检测数据是产品生产完成后进行的检测数据集, 主要包括产品编号、各个检测项目和检测结果, 其中动平衡检测结果包括三个指标 BAL_RANK, RO_RANK 与 UFM_RANK, 每个指标在 1 到 5 中取值, 只要三个指标中至少有一个指标为 4 或 5, 则产品为不合格品. 质量生产数据是产品在生产过程中产生的相关数据, 主要包括产品编号、各设备编号、生产时间、班组、各操作人员, 各工序的工艺参数集等. 以上两种数据可以用产品标号关联起来, 形成结构化的质量数据集.

轮胎生产制造的各种数据存储在企业的 MES、ERP 等不同系统中, 这些数据需要整合起来. 首先使用数据接口将这些数据存储在关系型数据库中. 然后, 利用 Sqoop 配置关系型数据库与 HDFS 之间的数据连接[14], 以增量导入的方式获取所有质量相关数据, 构建大数据存储中心来实现数据集中管理. 接下来进行数据预处理工作, 如重复数据的去除、数据缺失处理等[15]. 最后, 使用多表合并技术, 在 Hive 中集成前面获取到的所有质量数据, 去建立结构化质量分析数据集, 该数据集将应用于数据挖掘的进一步分析[16].

4 基于 ADTree 决策树的质量分析

4.1 轮胎质量大数据分析方法

图 1 展示了质量数据分析的流程, 其中数据获取和数据预处理已在 3.2 节阐述. 质量分析分为单因素分析与多因素分析. 单因素分析即使用统计的方式, 通过执行 HiveQL 查询语句, 得到单个因素与不合格率的关系. 对山东玲珑轮胎公司的千万级轮胎质量数据进行单因素分析, 可以得到一些初步结论, 例如不同物料编码的轮胎不合格率差异十分明显, 其中 21 种物料编码的轮胎占产品总数的 0.7%, 却产生了 13.3% 的不合格品. 单因素分析同样能排除一些影响因素, 例如轮胎硫化班组分早、中、晚班, 容易想到晚班的工人是否会因为精力不济导致不合格率增加, 但是统计结果表明三个班组的平均不合格率几乎相同.

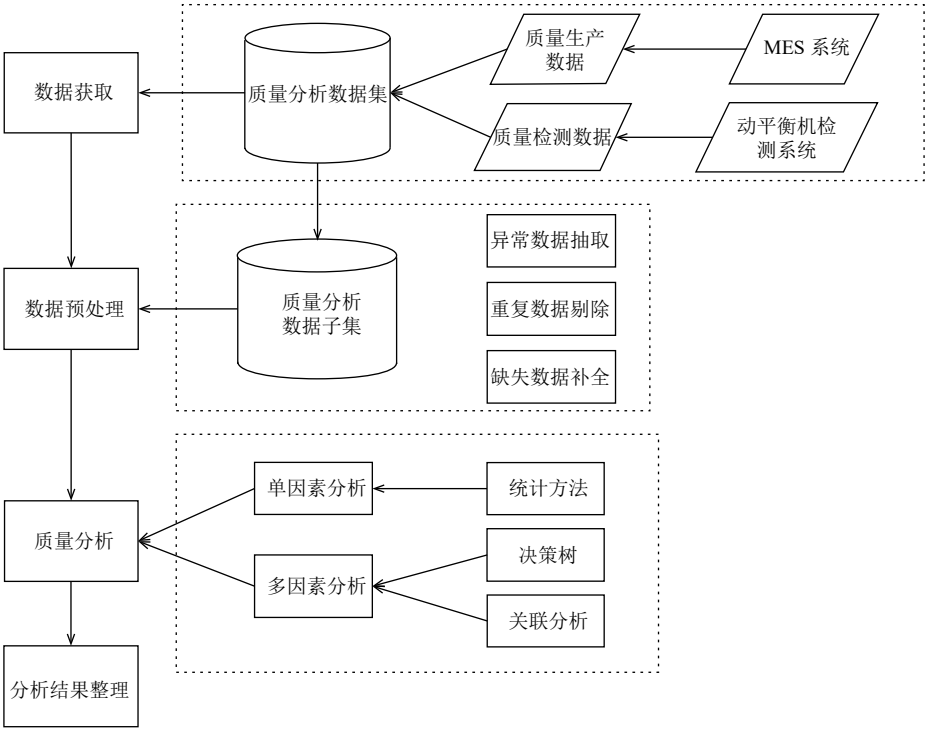

图 1 质量数据分析流程

产品质量的多因素分析使用数据挖掘的方法来找到造成不良品的影响因素. 本文将使用 ADTree 决策树作为轮胎质量分析的算法, 把轮胎生产过程中的硫化工序工艺参数特征值 (内温、内压、模温、板温的最大值、最小值、平均值、方差)、硫化操作人员 (CUR_ZJS_ID)、成型操作人员 (ZJS_ID)、各生产设备 (POT_ID, EQUIP_ID, EQUIP_CODE)、生产班次 (CLASS)、生产车间 (WOKR_SHOP_CODE)、生产模具 (MOLD_ID)、生产批次 (CUR_BATCH_ID) 作为 ADTree 算法的输入, 并将 qualified 字段设为标记字段,该字段为 1 代表产品合格, 如果为 2 则代表产品不合格. ADTree 算法将输出一个决策树作为挖掘结果. 由于传统的 ADTree 算法效率较低, 无法进行大数据下的分析, 因此本文将对 ADTree 算法进行改进.

4.2 传统 ADTree 算法

ADTree 算法由 Freund 和 Mason 提出[17], 其优点在于, 它的分类准确率往往比其他决策树算法要高, 可以同时处理离散型和数值连续型输入参数, 并且能够给出预测结果的置信度. ADTree 不仅能做分类工作, 其个别节点还可以评估自己的预测能力, 因此在轮胎质量分析问题中, 可以通过节点来分析导致最终质量不合格的潜在影响因素.

ADTree 算法适用于解决二分类问题, 例如轮胎质量分析中的合格与不合格就是典型的二分类情况. ADTree 的图形显示和传统决策树不同, 它包括两种节点: 预测节点和决策节点. 决策节点对应一个分裂测试, 训练集的样本经过分裂测试后被划分到相应预测节点中. 每个预测节点 p 对应一个预测值, 同时包括一部分样本, 划分到某个预测节点的样本集称为 $F(p)$.

传统 ADTree 算法的输入包括两个集合, 第一个集合里的每一个元素包括了属性向量和分类值, 其中分类值的取值为 1 或–1(也可以为 1 或 0), 在轮胎质量分析中分别代表不合格与合格. 第二个集合是权重 W_i(样本 i 的权重) 的集合. ADTree 的构建需要经过 T 次迭代, 每次迭代找到全局的最佳分裂测试, 然后生成相应的预测节点和决策节点. 最佳分裂测试通过 (1) 式取到最小值来获得:

$$z = 2\left(\sqrt{W_+(c)W_-(c)} + \sqrt{W_+(\neg c)W_-(\neg c)}\right) + W(\neg p) \quad (1)$$

其中, c 代表分裂测试, $W_+(c)$ 即预测节点样本中满足 c 的正标记权重和, $W(\neg p)$ 为不在预测节点里的样本权重和.

4.3 ADTree 改进算法

传统的 ADTree 算法受限于性能, 不适用于大数据问题. 学者 Pfahringer 提出了一个新的公式:

$$Z_{\text{pure}} = 2\left(\sqrt{W_+} + \sqrt{W_-}\right) + W(\neg p) \quad (2)$$

Z_{pure} 的计算不需要经过分裂测试, 只要累加 $F(p)$ 的正负权重和即可. z 和 Z_{pure} 经过拉普拉斯修正后, Z_{pure} 会成为 z 的下限. 如果根据 $F(p)$ 计算出来的 Z_{pure} 已经大于等于当前迭代的最小 z 值, 那么当前 $F(p)$ 的所有分裂测试评估值 z 都会大于等于当前迭代的最小 z 值, 所以这个节点不需要寻找更好的分裂测试, 可以直接跳过. 这种优化能提高传统 ADTree 算法的性能, 但效果有限. 杨碧姗等提出了 BICA 算法, 通过以空间换时间的策略, 降低了计算评估值 z 的复杂程度, 极大地提升了算法的性能. 本文在 BICA 算法的基础上做了进一步优化, 并修正了原算法中出现的零权重值问题, 提出了 ADTree 改进算法.

BICA 算法定义了新的数据结构 AVW-set(以下简称为 set), set 由 ADTree 算法需要处理的样本集生成. 表 1 是一个简单的样本集, 共有三条记录, 其中类别和权重是两个样本标识, 类别为 1 代表不合格, 类别为 –1 代表合格, 而权重一般初始都设为 1. 除去样本的标识, 每个样本有两个属性, 分别是操作人员和内温最小值. 样本的每个属性对应一个 set, 如本例中就有两个 set, 分别是操作人员的 set 和内温最小值的 set. 每个 set 有三个属性, 分别是属性名、正标记权重和与负标记权重和. 如果 set 记录的属性 attr 是连续型的,取所有属性值 v, 记录 $F(p)$ 中满足属性 attr≤v 的正标记权重和与负标记权重和; 如果 attr 是离散型的, 只记录 $F(p)$ 中满足属性 attr=v 的正标记权重和与负标记权重和.

表 1 预测节点 p 拥有的样本集

操作人员	内温最小值	类别	权重
20070488	95.5	1	1
20080001	120	–1	1
20080001	160.5	–1	1

以表 1 的样本集为例, 内温最小值属性是连续型的. 第一个值是 95.5, 在三个样本中只有一个样本的内温最小值小于等于该值, 同时该样本的类别为 1, 故取其权重 1, 算在正标记权重和里. 同理,第三个值是 160.5, 三个样本的内温最小值都小于等于该值, 统计这

三个样本的权重, 得到正标记权重和为 1, 负标记权重和为 2. 构建结果如表 2 所示.

表 2 内温最小值的 set

值	正标记权重和	负标记权重和
95.5	1	0
120	1	1
160.5	1	2

而操作人员属于离散型值, 且只有两种值. 20080001 有两个样本, 所以负标记权重和为 2. 操作手的 set 构建结果如表 3 所示.

表 3 操作人员的 set

操作人员	正标记权重和	负标记权重和
20070488	1	0
20080001	0	2

BICA 算法中的 set 在离散型属性的正标记权重和或负标记权重和为 0 时, 会赋一个自定义的较小值, 这是错误的. Pfahringer 在其论文[9]的第 3 节提到了权重和为 0 不会影响 ADTree 算法的结果, 从解释性来说, 主机手 20080001 操作了两个产品, 都是合格的, 如果正标记权重和不设为 0, 那么这个主机手的合格率就不是 100% 了, 这明显也不合理. 正确的做法是保留 0 这个值.

此外, BICA 算法构建连续型属性的 set 时, 采用先扫描样本集, 获得所有属性值, 然后对属性值排序, 再记录每个属性值的正负权重和的方式. 假设样本数量是 X, 不同属性个数是 Y, 那么时间复杂度是 $O(X)+O(Y\lg Y)+O(XY)$. 本算法在获取样本集所有属性值的同时, 直接记录每个属性值的权重和. 待属性值排序完毕后, 从小到大扫描一遍, 将权重和逐次累加即可, 前两步的时间复杂度不变, 第三步的时间复杂度从 $O(XY)$ 降到了 $O(Y)$, 从而减少构建 set 的时间.

所有属性的 set 都建立完成后, 将被统一放到 AVW-group (以下简称 group) 里作为一个集合.

在分裂测试中, 如果属性 attr 为连续型, 每个分裂测试为 $attr\leq(V_j+V_{j+1})/2$, 即每两个相邻数值的均值. 如果属性 attr 为离散型, 分裂测试较为简单, 直接是 $attr=V_j$. 这样设计后, set 起到的作用就是记录了预测节点 P 的每个分裂测试 c 的正负标记权重和. ADTree 中的内部预测节点的 set 可以根据下文介绍的自底向上的合并方法获得, 而传统 ADTree 算法在每个预测节点计算 z 时都要计算这两个值, 效率较低. 同时, set 的定义确保了该数据结构的大小和样本数量无关, 只和每个属性的不同取值个数有关. 这样, 在计算 Z_{pure} 时, 只需要扫描 set 的各个值即可, 不需要像传统 ADTree 算法一样扫描整个样本集. 设计 set 不仅减少了正负标记权重和的重复计算, 其容量一般也远小于样本数量, 所以 set 占的空间并不大.

BICA 算法的分裂测试评估改为自底向上的归纳来进行, 可以省去部分内节点的 group 计算. 每个预测节点都有对应的 group, 这涉及到 group 的合并问题.只要预测节点是 ADTree 的非叶子节点, 则取它的第一个决策子节点, 将其两个后代节点的 group 合并成本节点的 group. 由于每个 group 包含多个 set, 所以合并时根据同属性的 set 进行合并.

对于离散型属性的 set, 直接合并相同属性的正负权重和即可. 对连续型 set 合并, 设合并后的 set 为 P,待合并的 set 为 X, Y, 其中 X, Y 在构建时已经排序. 整个过程通过归并排序的算法持续进行, x, y, p 分别初始化为 X, Y, P 的末尾记录.

算法 1. 连续型 set 合并

输入: 待合并 set X, Y
输出: 合并后的 set P

当 x, y 没有全部指向 set 起始记录时:
1) $P[p].W_+=X[x].W_++Y[y].W_+$
2) $P[p].W_-=X[x].W_-+Y[y].W_-$
3) IF $X[x]$.value<$Y[y]$.value
$y=y-1$;
4) Else if $X[x]$.value>$Y[y]$.value
$x=x-1$;
5) Else
$x=x-1$, $y=y-1$;

同时, BICA 对连续型属性进行合并时, 会先扫描一遍两个待合并的 set, 得到新 set 里的属性值, 再扫描一遍两个待合并的 set, 计算出新 set 里的正负权重和.实际上, 只需要对两个待合并的 set 从后往前扫描一遍, 就可以生成新的 set, 如上文的算法所示, 这样能减少合并的时间. 通过 set 的合并, 可以充分利用已知信息, 不需要重复计算, 同时合并的时间复杂度是线性的. 而传统的 ADTree 算法在评估 z 值时, 需要对每个预测节点的样本的每个连续型属性进行排序, 在大数据量情况下开销巨大.

本文对 BICA 算法中的 ADTree 构建算法进行了

适当改进. 当算法遍历到叶子节点时, 如果叶子节点的正负标记权重和不全为正数, 那说明这个节点是完美分裂测试所生成的, 不需要再做处理. 原算法中缺少这一判断, 所以遍历到叶子节点后一定会进入算法的第 4 步, 这会增加算法的时间. 修改后的算法共 T 次迭代(即生成 T 个分裂测试), 每次迭代用后序遍历预测节点的方式, 通过得到最小的 z 找到最佳分裂测试, 生成新的预测节点 p. 算法不仅采用了 Pfahringer 等提出的 Z_{pure} 剪裁技术, 也结合了 BICA 自底向上归纳评估的思想, 分裂测试评估过程核心部分伪代码如算法 2.

算法 2. ADTree 评估算法

输入: 根节点 r 及根节点的 $F(p)$
输出: 节点的 group

1) 访问一个预测节点 p
2) 如果 p 是叶子
a) 根据 $F(p)$ 计算 group
b) 计算 p 的正负权重和、Z_{pure}, 如果正负权重和不全为正数, 直接返回
3) 否则 (即 p 是内节点)
a) 取 p 的第一个决策子节点 d, d 是 p 的决策子节点中的最佳分裂
b) 取 d 的两个预测子节点 q 与 r, 计算它们的 group, 然后分别作为输入, 递归调用本算法,这样就起到了后序遍历的作用
c) 将 q 与 r 的 group 合并为 p 的 group, 计算 p 的正负权重和、Z_{pure}
4) 如果当前 p 的 Z_{pure} 小于当前最小的 z, 因为 Z_{pure} 是 z 的下限, 那么可能存在 z 比当前最小的 z 还小, 所以对于 group 里的所有 set 的所有值 v
a) 计算分裂测试 c 的 z
b) 如果 z 比最小的 z 还小, 那么最小的 z 设为这个值, 并且将分裂测试 c 设为最佳测试, p 设为最佳分裂节点
5) 对于 p 除了第一个决策子节点 d 的其余子节点 d (如果存在的话)
a) 取 d 的两个子节点 q 与 r, 计算它们的 group,然后分别作为输入, 递归调用本算法

5 质量分析结果与算法性能实验

5.1 质量分析结果

本文对山东玲珑轮胎公司提供的千万级轮胎数据进行质量分析. 本例选取的轮胎物料代码是 221003794, 可用样本数为 308 880, 其中质检合格 306 471, 不合格 2409, 不合格率约为 0.78%. 借助生成迭代次数为 10 的 ADTree 图形进行分析, 如下图所示,每一个椭圆形的节点是分裂测试, 每个分裂测试有两个矩形的子节点, 节点上的数字代表置信打分, 本例中这个打分较高的话则代表该因素可能对质量不合格有重要影响.

根据 ADTree 的挖掘结果, 使用 Hive 数据库对质量数据进行追溯, 查询 ADTree 挖掘出的质量影响因素对产品不合格率的提升程度, 可以得到如下结论:

1) 成型主机手 20070488 负责的产品中, 合格 17 952 件, 不合格 1260 件, 不合格率高达约 6.6%. 这名主机手经手了约 6.2% 的产品, 却产生了约 52% 的不合格品, 可见其操作水平非常之低.

2) 其余主机手生产的轮胎, 在平均内压<1.776 时,合格 29 053 件, 不合格 413 件, 不合格率约 1.4%; 平均内压在[1.776, 1.817]时, 合格 238 090 件, 不合格 731 件,不合格率仅为约 0.3%; 当平均内压>1.817 时, 合格 21 376 件, 不合格仅 5 件, 不合格率几乎忽略不计. 由此可见, 轮胎硫化过程的硫化机平均内压对于最后的质检合格与否起到了重要影响.

3、硫化批次是 20161206 时, 合格 864 件, 不合格 401 件, 不合格率高达 31.7%. 其中, 经手成型主机手 20070488 的 951 件产品更是有 382 件不合格, 不合格率约为 40.1%; 剩余 314 件产品有 20 件不合格, 不合格率约为 6.4%, 也远高于平均不合格率. 因此, 该批次的生产出现了明显的问题.

以上挖掘结果反映出几个问题. 首先是成型主机手 20070488, 这名主机手的生产操作水平差得离谱, 严重影响了轮胎质量, 企业可以考虑对其进行技能培训, 或者调离岗位. 其次是轮胎加工中的平均内压, ADTree 反映该工艺参数对轮胎质量有较大影响, 企业需要对照自身制定的工艺参数, 确保轮胎生产时平均内压处于合理范围内. 最后, 硫化批次是 20161206 (即 2016 年 12 月 6 日) 时, 平均不合格率非常高, 企业需要排查当天的生产状况, 分析可能存在的问题.

由于不同轮胎物料代码经过的设备、操作人员、生产工艺参数等均不相同, 因此每种轮胎物料代码的挖掘结果存在差异. 但通过整理, 可以总结出影响轮胎质量的普遍规律:

1) 操作人员的水平好坏会影响轮胎质量, 个别操作人员经手的轮胎不合格率会非常高, 企业应该及时采取人员改进措施.

2) 轮胎生产过程中的平均内压对轮胎质量有明显影响, 一般来说, 如果平均内压偏低, 那么轮胎的整体不合格率会有提升. 因此, 企业需要提高生产技术, 确保硫化过程的平均内压在合理范围内.

3) 由于少量生产设备存在问题, 导致该设备生产的轮胎品种不合格率偏高. 企业应该及时维修设备或考虑购置新设备, 以此保证产品质量.

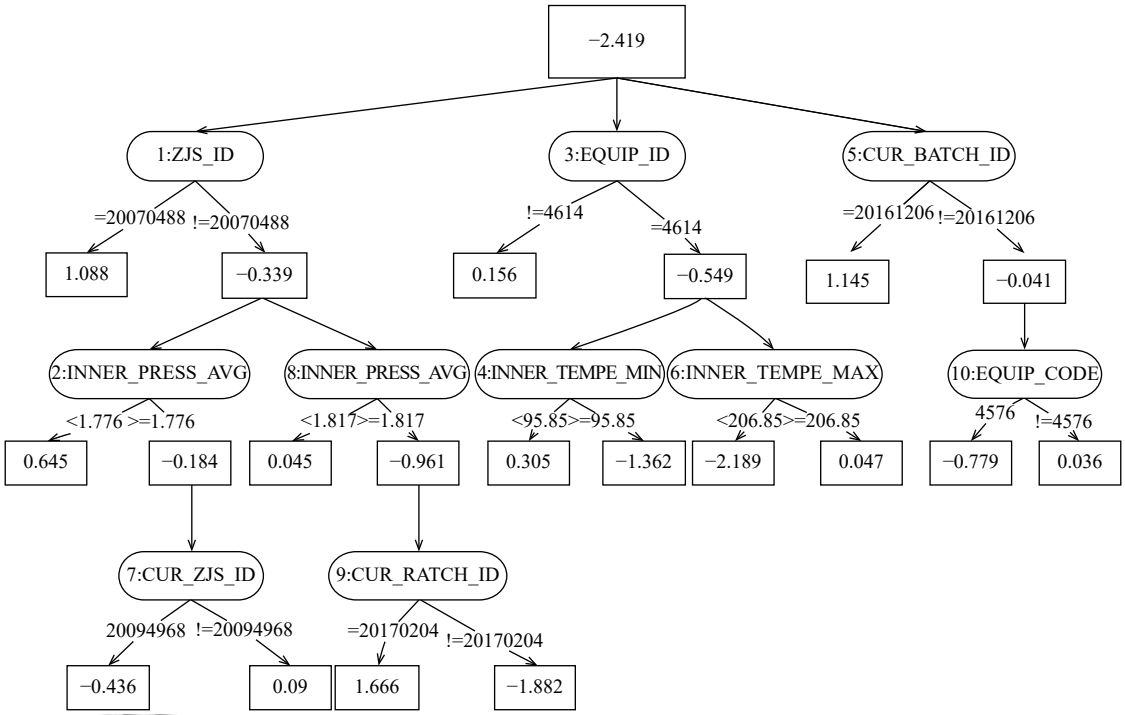

图 2 ADTree 算法挖掘结果图

5.2 算法性能实验

虽然现在已经有了较为成熟的大数据处理技术,但是算法本身的提升仍然会对整体性能有所提高. 以轮胎质量分析数据集物料编码 221005405、221003790 作为实验的数据集, 数据集大小为 379 010. 实验环境为 Intel i5 7000, 操作系统为 Centos 6.8, 4 台 24 GB 内存, 通过 Java 调用 Spark 并连接 Hive 进行实现. 实验比较结果见表 4 所示.

表 4 新算法实验结果比较

比较项目	BICA	传统 ADTree 算法	新算法
建树时间 (s)	14.2	102.5	11.6
内存占用 (MB)	21	8.8	21

由实验可见, BICA 算法相比于 Pfahringer 等提出的传统算法, 在建树时间上大大缩短了, 这是因为 BICA 用了 set 和自底向上的评估思路, 通过合并 group 这种利用已知数据的方法, 减少了 Z_{pure} 和 z 的计算量, 节省了排序次数; 在排序方面, Pfahringer 的算法在评估连续型属性时需要对整个数据集排序, 而 BICA 算法只对 set 中的属性值排序, 这也是性能提升的一方面. 本算法改进了 BICA 算法建立 set, 合并 group 的方式, 优化了时间复杂度, 并且对树的构建算法也做了适当改进, 在其基础上进一步提升了性能.

内存方面, 由于算法的整体思路是以空间换时间, 因此传统的 ADTree 算法内存占用较低, 但新算法的内存占用并不大, 是可以接受的.

6 总结

随着信息行业的快速发展, 很多工业企业正在大力建设工业信息化, 同时也积累了大量的工业数据. 在大数据时代的背景下, 如何利用这些数据成为了关键问题[18]. 通过分析、挖掘这些工业数据, 能够得到许多对企业有价值的信息, 使企业更好地发展.

本文选取轮胎行业大数据作为工业大数据研究的案例, 分析了轮胎行业大数据的需求与数据特征, 并开展了轮胎质量数据分析工作. 先利用大数据技术, 将轮胎生产各个环节的多源异构数据整合起来, 经过预处理等流程, 构建出大规模的结构化质量分析数据集. 本文重点介绍了使用改进后的 ADTree 算法进行轮胎质量多因素分析, 实验证明, 改进后的算法更适用于大数据背景下的数据挖掘. ADTree 的挖掘结果经过整理, 可以找出影响轮胎质量的重要因素, 这种精确定位出来的问题能够帮助企业改善工业流程, 降低产品的不合格率, 从而实现企业效益的提升.

参考文献

1 杨海成. 企业信息化建设与工业化进程融合的认识与思考. 中国机电工业, 2008, (7): 80.

2 李敏波, 王海鹏, 陈松奎, 等. 工业大数据分析技术与轮胎销售数据预测. 计算机工程与应用, 2017, 53(11): 100–109. [doi: 10.3778/j.issn.1002-8331.1609-0154]

3 宁宣凤, 吴涵. 浅析大数据时代下数据对竞争的影响. 汕头大学学报 (人文社会科学版), 2017, 33(5): 90–98. [doi: 10.3969/j.issn.1001-4225.2017.05.018]

4 Yan JH, Meng Y, Lu L, *et al*. Industrial big data in an industry 4.0 environment: Challenges, schemes, and applications for predictive maintenance. IEEE Access, 2017, 5: 23484–23491.

5 张洁, 高亮, 秦威, 等. 大数据驱动的智能车间运行分析与决策方法体系. 计算机集成制造系统, 2016, 22(5): 1220–1229.

6 杨枝雨. 基于大数据的印花质量影响因素分析方法研究 [硕士学位论文]. 上海: 东华大学, 2017.

7 李继安, 冯晓荣, 贾世准, 等. 工业大数据与服务大数据及其检测要点对比研究. 电子产品可靠性与环境试验, 2017, 35(S1): 1–6. [doi: 10.3969/j.issn.1672-5468.2017.S1.001]

8 Yang LP, Wang FZ, Wang T. Analysis of dishonorable behavior on railway online ticketing system based on k-means and FP-growth. Proceedings of 2017 IEEE International Conference on Information and Automation. Macau, China. 2017. 1173–1177.

9 Pfahringer B, Holmes G, Kirkby R. Optimizing the induction of alternating decision trees. Proceedings of the 5th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining. Hong Kong, China. 2001. 477–487.

10 杨碧姗, 王腾蛟, 常雷, 等. 一种快速可扩展的 ADTree 构建算法. 计算机研究与发展, 2007, 44(S3): 335–340.

11 Watcharapasorn P, Kurubanjerdjit N. The surgical patient mortality rate prediction by machine learning algorithms. Proceedings of the 13th International Joint Conference on Computer Science and Software Engineering. Khon Kaen, Thailand. 2016. 1–5.

12 俞燕. 工业企业生产成本核算中的内部控制措施. 中国经贸, 2014, (8): 169–170. [doi: 10.3969/j.issn.1009-9972.2014.08.108]

13 浦哲, 边慧光. 成型过程对全钢载重子午线轮胎动平衡的影响. 轮胎工业, 2013, 33(6): 364–366. [doi: 10.3969/j.issn.1006-8171.2013.06.012]

14 陈吉荣, 乐嘉锦. 基于 MapReduce 的 Hadoop 大表导入编程模型. 计算机应用, 2013, 33(9): 2486–2489.

15 Taleb I, Dssouli R, Serhani MA. Big data pre-processing: A quality framework. Proceedings of 2015 IEEE International Congress on Big Data. New York, NY, USA. 2015. 191–198.

16 Thusoo A, Sarma JS, Jain N, *et al*. Hive: A warehousing solution over a map-reduce framework. Proceedings of the VLDB Endowment, 2009, 2(2): 1626–1629. [doi: 10.14778/1687553]

17 Freund Y, Mason L. The alternating decision tree learning algorithm. Proceedings of the 16th International Conference on Machine Learning. San Francisco, CA, USA. 1999. 124–133.

18 孟辛澄. 大数据时代企业市场营销策略探索. 商场现代化, 2018, (2): 75–76.