(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 111930357 B
(45)授权公告日 2021.01.15

(21)申请号 202010981433.5
(22)申请日 2020.09.17
(65)同一申请的已公布的文献号
申请公布号 CN 111930357 A
(43)申请公布日 2020.11.13
(73)专利权人 国网浙江省电力有限公司营销服务中心
地址 311121 浙江省杭州市余杭区仓前街道云联路138号5幢
专利权人 国网浙江省电力有限公司
(72)发明人 张宏达 杜蜀薇 马亮 陈仕军 胡若云 王正国 裘炜浩 林森 叶方斌 欧阳柳 杨世旺 金王英
(74)专利代理机构 杭州华鼎知识产权代理事务所(普通合伙) 33217
代理人 项军
(51)Int.Cl.
G06F 8/20(2018.01)
G06F 16/84(2019.01)
G06F 16/901(2019.01)
G06F 16/25(2019.01)
G06F 16/2458(2019.01)
G06F 16/215(2019.01)
G06Q 10/04(2012.01)
G06Q 50/06(2012.01)

审查员 刘升

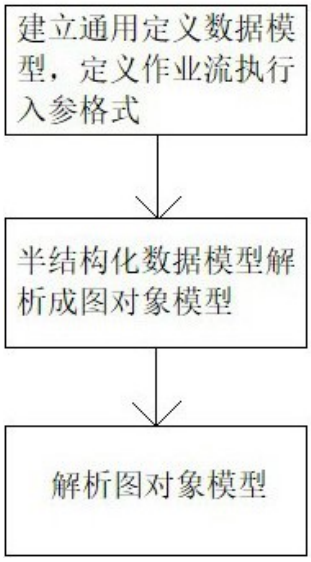
权利要求书2页 说明书6页 附图2页

(54)发明名称

一种可视化建模作业流调度引擎的构建方法

(57)摘要

本发明公开了一种可视化建模作业流调度引擎的构建方法,其包括:步骤一,系统依据大数据建模分析流程建立通用定义数据模型,定义作业流执行入参格式;步骤二,系统接收作业流执行入参,系统遵循通用定义数据模型的约束将作业流执行入参的半结构化数据模型解析成图对象模型;步骤三,系统将图对象模型作为作业流执行模块的入参,通过作业流执行模块对图对象模型进行解析,从而构建完成可视化建模作业流调度引擎。本发明可视化建模作业流调度引擎的构建方法为电力行业可视化建模方向、数据ETL方向的作业编排与调度,提供了技术依据,有较好的借鉴意义。



CN 111930357 B

1.一种可视化建模作业流调度引擎的构建方法,其特征是,包括以下步骤:

步骤一,系统依据大数据建模分析流程建立通用定义数据模型,定义作业流执行入参格式;

步骤二,系统接收作业流执行入参,系统遵循通用定义数据模型的约束将作业流执行入参的半结构化数据模型解析成图对象模型;

步骤三,系统将图对象模型作为作业流执行模块的入参,通过作业流执行模块对图对象模型进行解析,从而构建完成可视化建模作业流调度引擎;

所述的步骤二中作业流执行入参的半结构化数据模型解析成图对象模型具体为将半结构化模型中的节点对象抽象为图对象模型中的顶点,半结构化模型中的节点对象依赖关系抽象为图对象模型中的边,顶点和连接顶点的边构成有向无环图,有向无环图即为图对象模型;

所述的顶点包括虚拟顶点、分支顶点、循环顶点和执行顶点,虚拟顶点包括开始虚拟顶点和结束虚拟顶点,开始虚拟顶点为图对象模型数据处理的开始位置,结束虚拟顶点为图对象模型数据处理的结束位置,结束虚拟顶点输出可视化图表或根据可视化图表得出的分析评估结果;所述的可视化图表包括混淆矩阵,分析评估结果包括根据混淆矩阵的二分类评估结果;

系统检测混淆矩阵的预测结果的准确率A:

A=(a1+b2+c3+……+n)/N,

其中,a1表示混淆矩阵第一行第一个数,b2表示混淆矩阵第二行第二个数,c3表示混淆矩阵第三行第三个数,n表示混淆矩阵第n行第n个数,a1+b2+c3+……+n 为预测和结果一致的数量,N表示样本总量;

若A与预设的准确率值结果一致或两者的差值在设定的阈值内,则说明半结构化数据模型解析成图对象模型的过程正常;若A与预设的准确率值两者的差值不在设定的阈值内,则说明半结构化数据模型解析成图对象模型的过程不正常,系统发出报警并检测执行入参和解析过程是否正确。

2.根据权利要求1所述的一种可视化建模作业流调度引擎的构建方法,其特征是,所述的通用定义数据模型包括算法组件属性、算法组件输入输出属性和连接对象属性,其中算法组件属性包括数组的若干个节点对象,单个节点对象使用算法组件json定义,单个节点对象包括基础信息和算法参数;算法组件输入输出属性包含在节点对象中,单个节点对象含有算法组件唯一表示、输入输出类型和输入输出值;连接对象属性包括开始节点对象标识、开始节点输出位置、目标节点对象标识和目标节点输入位置。

3.根据权利要求1所述的一种可视化建模作业流调度引擎的构建方法,其特征是,选取a1、b2、c3……n中的一个数值,此数值大于剩余其他数值之和,则准确率A=ns/nsall,其中ns为大于剩余其他数值之和的数字,nsall为ns对应的列所有数值之和。

4.根据权利要求1所述的一种可视化建模作业流调度引擎的构建方法,其特征是,选取a1、b2、c3……n中的某一个或几个数值,此类数值均小于剩余其他数值,则准确率A=nt/ntall,其中nt为均小于剩余其他数值的一个或几个数值,ntall为此类数值对应的列所有数值之和。

5.根据权利要求1所述的一种可视化建模作业流调度引擎的构建方法,其特征是,所述

的步骤三中，解析的具体过程为：基于JGraphT图算法库，使用广度优先遍历算法进行顶点调度，对于同一层的顶点并行调用，对于存在依赖上下关系的顶点使用串行调用并把上一顶点执行结果根据依赖关系传递给下一顶点执行。

6.根据权利要求5所述的一种可视化建模作业流调度引擎的构建方法，其特征是，在顶点对象调度前、调度过程中以及调度后使用钩子接口松耦合方式不侵入作业流执行模块实现的情况下满足个性化需求定义。

一种可视化建模作业流调度引擎的构建方法

技术领域

[0001] 本发明涉及电网信息化技术领域，尤其是指一种可视化建模作业流调度引擎的构建方法。

背景技术

[0002] 随着电网行业信息化建设的发展，系统已积累了大量的营销业务、用电信息、客户服务、统计报表等各类海量数据，但如何挖掘海量数据的价值是面临的最大的挑战，数据价值应用存在很大的提升空间。

[0003] 可视化建模工具需求，以组件化、可视化的方式，按照大数据建模分析流程，提供从数据读取、数据清洗、数据处理、模型构建、模型固化、模型评估、模型部署等数据建模分析全流程、一体式可闭环组件，提供包括分类、聚类、回归、推荐等支撑大数据分布式并行计算的数据分析算法。因此可视化建模工具建设需要建设一种通用的、普适的作业流调度引擎，用来支撑大数据建模分析流程的定义与构建、执行与调度、以及运维监控。

[0004] 中国专利公开号CN108492006A，公开日2018年9月4日，名称为《一种层次控制模式的运维作业调度引擎》的发明专利中公开了一种层次控制模式的运维作业调度引擎，包括展示层、处理层和数据层，所述的展示层用于系统统一入口、用户操作、数据展示，用于给用户提供可视化界面，用户通过展示层制定作业引擎模板；所述的处理层用于给用户在展示层编排的任务在后台进行处理；所述的数据层用于实现运维对象终端功能的具体执行，完成展示层用户配置的作业任务，将作业任务执行的结果数据有序获取出来存储在数据库，给展示层提供基础数据。不足之处在于，该专利的作业调度引擎主要用于对运维作业的层次控制，不能满足对可视化建模的提供较好的作业流调度，使用上较为局限。

发明内容

[0005] 本发明的目的是克服现有技术中的缺点，提供一种可视化建模作业流调度引擎的构建方法。

[0006] 本发明的目的是通过下述技术方案予以实现：

[0007] 一种可视化建模作业流调度引擎的构建方法，包括以下步骤：

[0008] 步骤一，系统依据大数据建模分析流程建立通用定义数据模型，定义作业流执行入参格式；

[0009] 步骤二，系统接收作业流执行入参，系统遵循通用定义数据模型的约束将作业流执行入参的半结构化数据模型解析成图对象模型；

[0010] 步骤三，系统将图对象模型作为作业流执行模块的入参，通过作业流执行模块对图对象模型进行解析，从而构建完成可视化建模作业流调度引擎；

[0011] 所述的步骤二中作业流执行入参的半结构化数据模型解析成图对象模型具体为将半结构化模型中的节点对象抽象为图对象模型中的顶点，半结构化模型中的节点对象依赖关系抽象为图对象模型中的边，顶点和连接顶点的边构成有向无环图，有向无环图即为

图对象模型。

[0012] 顶点对应到业务作业就是算法组件，本方案对大数据建模分析流程的作业流进行了抽象，包含了算法组件属性，使用数组对象存储定义可配置N个算法组件节点，及提供弹性的节点配置普适不同的建模的作业流。由于一般建模的作业流往往需要多个步骤才能完成作业，例如需要经历数据清洗、数据贯彻、模型训练、模型评估等很多节点才能完成作业，本方案的设计正好满足了需求。

[0013] 作为一种优选方案，所述的通用定义数据模型包括算法组件属性、算法组件输入输出属性和连接对象属性，其中算法组件属性包括数组的若干个节点对象，单个节点对象使用算法组件json定义，单个节点对象包括基础信息和算法参数；算法组件输入输出属性包含在节点对象中，单个节点对象含有算法组件唯一表示、输入输出类型和输入输出值；连接对象属性包括开始节点对象标识、开始节点输出位置、目标节点对象标识和目标节点输入位置。在节点对象之间的连接时，需要考虑不同节点对象之间的个性化差异，即不同算法组件，需要定义的属性是不同的，如类型转换和归一化两个组件，算法配置的参数不同。应用到步骤二中，提取作业流中节点对象、节点对象的输入和输出，节点对象的上下依赖关系（下一个节点对象的执行需要依赖上一个节点对象执行的结果），将json半结构化转换成结构化对象即图对象模型。

[0014] 作为一种优选方案，所述的顶点包括虚拟顶点、分支顶点、循环顶点和执行顶点，虚拟顶点包括开始虚拟顶点和结束虚拟顶点，开始虚拟顶点为图对象模型数据处理的开始位置，结束虚拟顶点为图对象模型数据处理的结束位置，结束虚拟顶点输出可视化图表或根据可视化图表得出的分析评估结果。

[0015] 作为一种优选方案，所述的可视化图表包括混淆矩阵、分析评估结果包括根据混淆矩阵的二分类评估结果。

[0016] 作为一种优选方案，系统检测混淆矩阵的预测结果的准确率A：

[0017] A=(a1+b2+c3+……+n)/N，

[0018] 其中，a1表示混淆矩阵第一行第一个数，b2表示混淆矩阵第二行第二个数，c3表示混淆矩阵第三行第三个数，n表示混淆矩阵第n行第n个数，a1+b2+c3+……+n 为预测和结果一致的数量，N表示样本总量；

[0019] 若A与预设的准确率值结果一致或两者的差值在设定的阈值内，则说明半结构化数据模型解析成图对象模型的过程正常；若A与预设的准确率值两者的差值不在设定的阈值内，则说明半结构化数据模型解析成图对象模型的过程不正常，系统发出报警并检测执行入参和解析过程是否正确。

[0020] 混淆矩阵是表示精度评价的一种标准格式，用n行n列的矩阵形式来表示。主要用于比较分类结果和实际测得值，可以把分类结果的精度显示在一个混淆矩阵里面，其中横向行表示实际的测量结果，纵向列表示预测的结果。本设计通过对于混淆矩阵的检测，来判断半结构化数据模型解析成图对象模型的过程是否正常，进而判断可视化建模作业流调度引擎的构建是否正确。事先预设的准确率可由大数据得出，而由于数据会有一定的误差，因此实际准确率A与预设的准确率允许有一定的差值，但是差值过大时，说明半结构化数据模型解析成图对象模型的过程不正常。此外，由于执行入参可能也会出现错误导致准确率A的结果的偏差，因此需要人工或自动判断执行入参的数据是否正确。

[0021] 作为一种优选方案，选取a1、b2、c3……n中的一个数值，此数值大于剩余其他数值之和，则准确率A=ns/nsall，其中ns为大于剩余其他数值之和的数字，nsall为ns对应的列所有数值之和。

[0022] 当某一个数值大于剩余其他数值之和，则表示该值为混淆矩阵中最为常见的结果，即在混淆矩阵对应的执行入参中最为常见的结果，因此需要判断此数值对应的准确率。例如，在电力行业中，判断某一个企业的正常用电天数和非正常用电天数，对于一般的企业正常的生产情况下，正常用电天数远大于非正常用电天数，因此只需要判断正常用电天数的准确率A是否正常即可判断半结构化数据模型解析成图对象模型的过程正常。又例如，在判断某一个区域用电负荷的情况，分为正常负荷、超负荷和低于负荷三种情况，正常负荷的时间远大于超负荷和低于负荷的时间，而超负荷和低于负荷的时间由于样本数量较小准确率的误差会较大，因此只需要判断正常负荷的准确率的情况。

[0023] 作为一种优选方案，选取a1、b2、c3……n中的某一个或几个数值，此类数值均小于剩余其他数值，则准确率A=nt/ntall，其中nt为均小于剩余其他数值的一个或几个数值，ntall为此类数值对应的列所有数值之和。此设计根据实际情况灵活设计，例如，在电力行业中，需要估算电网发生故障停电的时间，而电网发生故障停电的时间占总时间的量较小，而根据实际需要要求分析电网故障的时间时，需要判断预测故障的准确率。

[0024] 作为一种优选方案，所述的步骤三中，解析的具体过程为：基于JGraphT图算法库，使用广度优先遍历算法进行顶点调度，对于同一层的顶点并行调用，对于存在依赖上下关系的顶点使用串行调用并把上一顶点执行结果根据依赖关系传递给下一顶点执行。

[0025] 作为一种优选方案，在顶点对象调度前、调度过程中以及调度后使用钩子接口松耦合方式不侵入作业流执行模块实现的情况下满足个性化需求定义。

[0026] 作为一种优选方案，在执行监控方面，执行状态包含等待执行、执行中、执行成功、执行失败、取消执行以及详细非结构化过程执行日志。

[0027] 本发明的有益效果是：可视化建模作业流调度引擎的构建方法为可视化建模的作业流构建提供了数据模型依据；同时可视化建模作业流调度引擎还能进行自我检查，判断构建的过程是否正确；可视化建模作业流调度引擎的构建方法为电力行业可视化建模方向、数据ETL方向的作业编排与调度，提供了技术依据，有较好的借鉴意义。

附图说明

[0028] 图1是本发明的一种流程图；

[0029] 图2是本发明的一种有向无环图。

具体实施方式

[0030] 下面结合附图和实施例对本发明进一步描述。

[0031] 实施例1：

[0032] 一种可视化建模作业流调度引擎的构建方法，如图1所示，包括以下步骤：

[0033] 步骤一，系统依据大数据建模分析流程建立通用定义数据模型，定义作业流执行入参格式；

[0034] 步骤二，系统接收作业流执行入参，系统遵循通用定义数据模型的约束将作业流

执行入参的半结构化数据模型解析成图对象模型；

[0035] 步骤三，系统将图对象模型作为作业流执行模块的入参，通过作业流执行模块对图对象模型进行解析，从而构建完成可视化建模作业流调度引擎；

[0036] 所述的步骤二中作业流执行入参的半结构化数据模型解析成图对象模型具体为将半结构化模型中的节点对象抽象为图对象模型中的顶点，半结构化模型中的节点对象依赖关系抽象为图对象模型中的边，顶点和连接顶点的边构成有向无环图，有向无环图即为图对象模型。

[0037] 顶点对应到业务作业就是算法组件，本方案对大数据建模分析流程的作业流进行了抽象，包含了算法组件属性，使用数组对象存储定义可配置N个算法组件节点，及提供弹性的节点配置普适不同的建模的作业流。由于一般建模的作业流往往需要多个步骤才能完成作业，例如需要经历数据清洗、数据贯彻、模型训练、模型评估等很多节点才能完成作业，本方案的设计正好满足了需求。

[0038] 所述的通用定义数据模型包括算法组件属性、算法组件输入输出属性和连接对象属性，其中算法组件属性包括数组的若干个节点对象，单个节点对象使用算法组件json定义，单个节点对象包括基础信息和算法参数；算法组件输入输出属性包含在节点对象中，单个节点对象含有算法组件唯一表示、输入输出类型和输入输出值；连接对象属性包括开始节点对象标识、开始节点输出位置、目标节点对象标识和目标节点输入位置。在节点对象之间的连接时，需要考虑不同节点对象之间的个性化差异，即不同算法组件，需要定义的属性是不同的，如类型转换和归一化两个组件，算法配置的参数不同。应用到步骤二中，提取作业流中节点对象、节点对象的输入和输出，节点对象的上下依赖关系(下一个节点对象的执行需要依赖上一个节点对象执行的结果)，将json半结构化转换成结构化对象即图对象模型。

[0039] Json是一种轻量级的数据交换格式。它基于 ECMAScript (欧洲计算机协会制定的js规范)的一个子集，采用完全独立于编程语言的文本格式来存储和表示数据。简洁和清晰的层次结构使得 JSON 成为理想的数据交换语言。同时json易于人阅读和编写，同时也易于机器解析和生成，并有效地提升网络传输效率。

[0040] 所述的顶点包括虚拟顶点、分支顶点、循环顶点和执行顶点，虚拟顶点包括开始虚拟顶点和结束虚拟顶点，开始虚拟顶点为图对象模型数据处理的开始位置，结束虚拟顶点为图对象模型数据处理的结束位置，结束虚拟顶点输出可视化图表或根据可视化图表得出的分析评估结果。

[0041] 所述的可视化图表包括混淆矩阵、分析评估结果包括根据混淆矩阵的二分类评估结果。

[0042] 如图2所示，是一种有向无环图的具体表达形式，包括13个顶点，即13个算法组件，根据有向无环图，定义了数据模型和分析流程，数据通过数据表进行读取，部分数据需要进行归一化处理后输出折线或柱状图，部分数据直接输出散点图和柱状图，部分数据输入后进行预测，得出混淆矩阵和二分类评估，其中，读数据表算法组件属性规范可定义如下：

[0043] “dom”: [{

[0044] “id”: “READ_TABLE_fjcexjiemuabrmyl”,

[0045] “label”: “读数据表”,

[0046] “dt_id”: “AI0001”,
[0047] “style”: {
[0048] “left”: “481px”,
[0049] “top”: “16px”
[0050] },
[0051] “desc”: “读取特征表”,
[0052] “status”:”SUCCESS”,
[0053] “viewdata”: [],
[0054] “prop”: {
[0055] “tableName”: “”
[0056] “columnDesc”: [],
[0057] “stepEngine”: “restapi”
[0058] }
[0059] }]

[0060] 其他算法组件与读数据表算法组件属性规范定义类似,根据每个算法组件不同的需求灵活调整。

[0061] 系统检测混淆矩阵的预测结果的准确率A:

[0062] A=(a1+b2+c3+……+n)/N,

[0063] 其中,a1表示混淆矩阵第一行第一个数,b2表示混淆矩阵第二行第二个数,c3表示混淆矩阵第三行第三个数,n表示混淆矩阵第n行第n个数,a1+b2+c3+……+n 为预测和结果一致的数量,N表示样本总量;

[0064] 若A与预设的准确率值结果一致或两者的差值在设定的阈值内,则说明半结构化数据模型解析成图对象模型的过程正常;若A与预设的准确率值两者的差值不在设定的阈值内,则说明半结构化数据模型解析成图对象模型的过程不正常,系统发出报警并检测执行入参和解析过程是否正确。

[0065] 混淆矩阵是表示精度评价的一种标准格式,用n行n列的矩阵形式来表示。主要用于比较分类结果和实际测得值,可以把分类结果的精度显示在一个混淆矩阵里面,其中横向行表示实际的测量结果,纵向列表示预测的结果。本设计通过对于混淆矩阵的检测,来判断半结构化数据模型解析成图对象模型的过程是否正常,进而判断可视化建模作业流调度引擎的构建是否正确。事先预设的准确率可由大数据得出,而由于数据会有一定的误差,因此实际准确率A与预设的准确率允许有一定的差值,但是差值过大时,说明半结构化数据模型解析成图对象模型的过程不正常。此外,由于执行入参可能也会出现错误导致准确率A的结果的偏差,因此需要人工或自动判断执行入参的数据是否正确。

[0066] 所述的步骤三中,解析的具体过程为:基于JGraphT图算法库,使用广度优先遍历算法进行顶点调度,对于同一层的顶点并行调用,对于存在依赖上下关系的顶点使用串行调用并把上一顶点执行结果根据依赖关系传递给下一顶点执行。Jgrapht图算法库是用java语言编写的算法库,适用于处理图数据结构的大多数算法,求最短路径等算法。

[0067] 在顶点对象调度前、调度过程中以及调度后使用钩子接口松耦合方式不侵入作业流执行模块实现的情况下满足个性化需求定义。

[0068] 在执行监控方面,执行状态包含等待执行、执行中、执行成功、执行失败、取消执行以及详细非结构化过程执行日志。本发明中,执行引擎默认支持For循环执行引擎、Restful API执行引擎、Spark执行引擎、Shell执行引擎、Python执行引擎,其中Spark执行引擎作为客户端角色使用Akka通信调用分布式机器学习算法服务。

[0069] 实施例2:一种可视化建模作业流调度引擎的构建方法,其原理和实施方法与实施例1基本相同,不同之处在于,在混淆矩阵的预测结果的准确率的计算中,选取a1、b2、c3……n中的一个数值,此数值大于剩余其他数值之和,则准确率A=ns/nsall,其中ns为大于剩余其他数值之和的数字,nsall为ns对应的列所有数值之和。

[0070] 当某一个数值大于剩余其他数值之和,则表示该值为混淆矩阵中最为常见的结果,即在混淆矩阵对应的执行入参中最为常见的结果,因此需要判断此数值对应的准确率。例如,在电力行业中,判断某一个企业的正常用电天数和非正常用电天数,对于一般的企业正常的生产情况下,正常用电天数远大于非正常用电天数,因此只需要判断正常用电天数的准确率A是否正常即可判断半结构化数据模型解析成图对象模型的过程正常,又例如,在判断某一个区域用电负荷的情况,分为正常负荷、超负荷和低于负荷三种情况,正常负荷的时间远大于超负荷和低于负荷的时间,而超负荷和低于负荷的时间由于样本数量较小准确率的误差会较大,因此只需要判断正常负荷的准确率的情况。

[0071] 实施例3:一种可视化建模作业流调度引擎的构建方法,其原理和实施方法与实施例1基本相同,不同之处在于,在混淆矩阵的预测结果的准确率的计算中,选取a1、b2、c3……n中的某一个或几个数值,此类数值均小于剩余其他数值,则准确率A=nt/ntall,其中nt为均小于剩余其他数值的一个或几个数值,ntall为此类数值对应的列所有数值之和。此设计根据实际情况灵活设计,例如,在电力行业中,需要估算电网发生故障停电的时间,而电网发生故障停电的时间占总时间的量较小,而根据实际需要要求分析电网故障的时间时,需要判断预测故障的准确率。

[0072] 以上所述的实施例只是本发明的一种较佳的方案,并非对本发明作任何形式上的限制,在不超出权利要求所记载的技术方案的前提下还有其它的变体及改型。

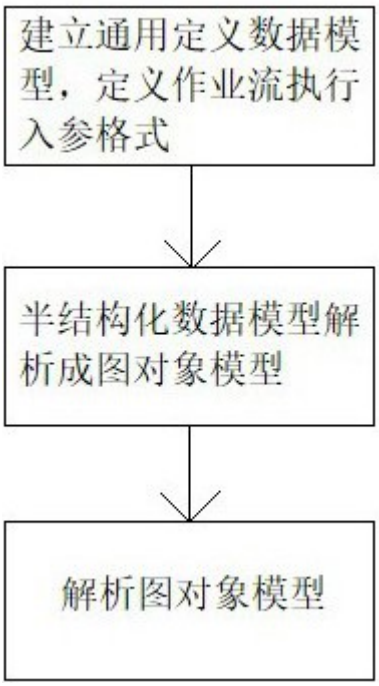


图1

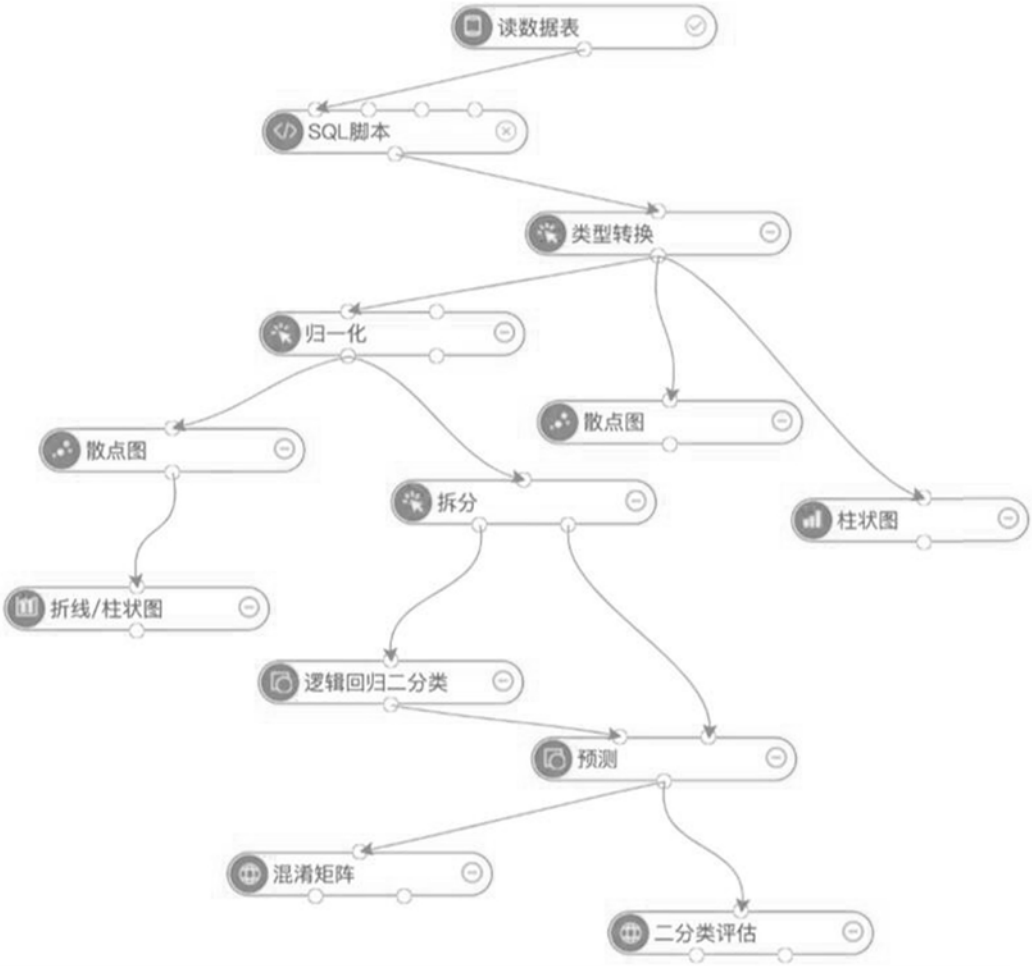

读数据表
SQL脚本
类型转换
归一化
散点图
散点图
拆分
柱状图
折线/柱状图
逻辑回归二分类
预测
混淆矩阵
二分类评估

图2