(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 115858622 A
(43)申请公布日 2023.03.28

(21)申请号 202211610454.1

(22)申请日 2022.12.12

(71)申请人 浙江大学
地址 310058 浙江省杭州市西湖区余杭塘路866号

(72)发明人 邓水光 王天笑 智晨 周小群 吴金杰

(74)专利代理机构 杭州天勤知识产权代理有限公司 33224
专利代理师 王琛

(51)Int.Cl.
G06F 16/2458(2019.01)
G06F 16/2455(2019.01)
G06F 16/242(2019.01)

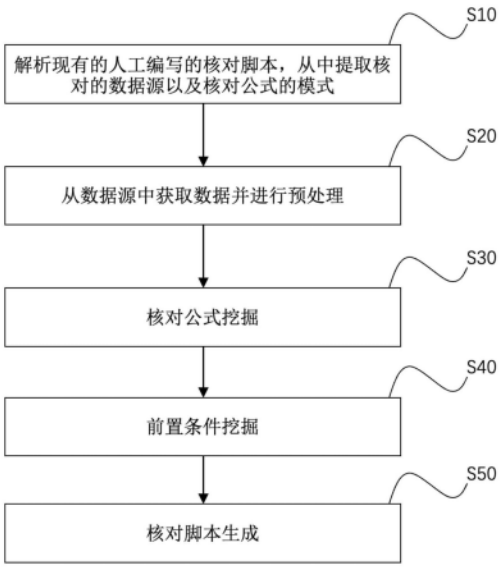
权利要求书1页 说明书6页 附图5页

(54)发明名称

一种业务数据核对脚本的自动化生成方法

(57)摘要

本发明公开了一种业务数据核对脚本的自动化生成方法，将回归方法、规则挖掘方法、SQL解析与生成技术相结合，针对大规模业务数据，自动挖掘核对公式和前置条件组成核对规则，并生成对应的核对脚本，极大地减少了核对监控部署的人工成本，提高了核对监控的覆盖范围。本发明从人工编写的核对脚本中解析获取多样化的核对数据源，进一步提高了挖掘的完备性；本发明使用改进的符号回归方法，在实验中对比基线方法Gplearn在不同的公式复杂度、公式支持度、数据复杂程度上，准确性和召回率都有明显提升；本发明通过字段语义分类和核对公式的模式，智能判断核对规则的业务价值，极大提高了方法在实际场景下的可用性。



CN 115858622 A

1.一种业务数据核对脚本的自动化生成方法,包括如下步骤:

(1)对人工编写的核对脚本进行解析,从中提取核对的数据源以及表达式的模式;

(2)从数据源中获取数据,并对数据进行预处理包括字段数据打平、字段分类、数据类型转换、特殊值处理、数据聚合;

(3)在预处理后的数据上挖掘核对公式,包括挖掘数值计算型的公式以及挖掘非数值一致型的公式;

(4)挖掘核对公式成立的前置条件;

(5)将核对公式和对应的前置条件转化为核对脚本。

2.根据权利要求1所述的自动化生成方法,其特征在于:所述步骤(1)的具体实现方式为:首先将SQL形式的核对脚本解析为AST形式;然后分析AST,获取表连接关系和聚合方式作为后续挖掘的数据源;通过分析AST,提取其中的表达式子树,构成表达式集,并归纳表达式的模式。

3.根据权利要求1所述的自动化生成方法,其特征在于:所述步骤(2)中字段数据打平包括对象类型字段以及键值对类型字段。

4.根据权利要求1所述的自动化生成方法,其特征在于:所述步骤(2)中字段分类包括对字段进行语义聚类,归纳字段类型,并设计分类规则;字段类型分为ID、数值、枚举、字符、时间、分区、版本7个大类,其中数值类型可细分为金额、数量和比例3个小类,枚举类型可细分为类型、状态和布尔标记3个小类。

5.根据权利要求4所述的自动化生成方法,其特征在于:所述字段分类所需要的语义信息包括字段名称、字段数据类型以及字段描述。

6.根据权利要求1所述的自动化生成方法,其特征在于:所述步骤(3)中使用改进的基于遗传的符号回归方法挖掘数值计算型的公式,具体地:将核对公式编码为公式语法树,结合从核对脚本中提取的表达式集和随机生成结果,初始化公式种群,利用适应度函数评估每个公式个体的适应度;根据字段分类评估公式个体的语义正确性,结合适应度选择优势个体;使用遗传和变异操作,生成新的公式树作为子代;重复适应度计算以及个体选择和子代生成的过程,直到出现公式满足条件或迭代次数达到上限。

7.根据权利要求1所述的自动化生成方法,其特征在于:所述步骤(3)中挖掘非数值一致型的公式,即将非数值类型的字段两两比较,需要保证两者的语义类型相同。

8.根据权利要求1所述的自动化生成方法,其特征在于:所述步骤(4)的具体实现方式为:首先在原始数据上验证公式是否成立,并用额外的标记字段记录结果;然后将每行数据构造成一个项集,在所有项集上执行关联规则挖掘,挖掘以公式成立为结果的条件项。

9.根据权利要求1所述的自动化生成方法,其特征在于:所述步骤(5)的具体实现方式为:首先定义核对脚本模板,使用核对公式和前置条件将核对脚本补充完整;然后检查脚本的语法是否正确,并测试运行。

10.根据权利要求1所述的自动化生成方法,其特征在于:该自动化生成方法的具体实现采用分布式框架,包括分布式数据仓库、分布式计算节点和分布式调度任务,实现大规模数据的并发处理。

一种业务数据核对脚本的自动化生成方法

技术领域

[0001] 本发明属于数据挖掘和代码生成技术领域，具体涉及一种业务数据核对脚本的自动化生成方法。

背景技术

[0002] 业务数据是指存在于业务系统数据库中的二维表结构数据，包括字段名称和具体值。为了及时发现异常数据，数据核对是直接有效的手段；数据核对通过定义核对规则，定期对业务数据进行正确性检查，核对规则R可以表示为蕴含式$P \Rightarrow F$，P和F都是命题，其中P表示前置条件，F表示核对公式，图1所示为一个实际案例。核对脚本是核对规则的可执行代码实现，通常是一段SQL语句，可以在核对监控系统中部署运行；由于业务规则的复杂性和变化性，现阶段核对脚本的编写依赖人工，需要消耗业务专家大量的时间和精力。因此，如何自动化地生成核对脚本是亟待解决的问题。

[0003] 自动化生成业务数据核对脚本要求方法能够自动挖掘存在于大规模数据中的核对规则，包括核对公式和前置条件，并生成对应的可执行代码。自动化核对脚本生成的主要技术突破点在于核对规则的准确高效挖掘，是一种融合了字段语义、符号回归方法和关联规则挖掘的新数据挖掘方法，其作为行业内的首创性技术，包括如下的技术问题：

[0004] 1.如何获取多样且尽可能完备的挖掘数据源。

[0005] 规则挖掘的第一步是从数据库中确定表、字段和数据行，构成一张二维表，作为分析数据源。在实际场景下，核对规则需要处理多种类型的数据源，包括但不限于单表数据源(数据来源于一张表)、多表数据源(数据涉及多张表，表之间通过外键相关联)以及聚合型数据源(数据包含聚合关系)，如何获取多样且尽可能完备的数据源是一个难点。

[0006] 2.如何保证挖掘的核对公式的准确性、召回率双高。

[0007] 核对公式被定义为一个等式，包括字段和运算符，一般的核对公式挖掘方法使用回归模型，例如线性回归。一方面这种方法只能发现特定的模型的公式，很难覆盖全部的公式类型；另一方面，回归分析通常需要指定字段范围以提高方法的准确率，因此很难做到完全的自动化。因此，保证挖掘得到的核对公式的准确性、召回率双高是一个困难的任务。

[0008] 3.如何评估核对规则的实际业务价值。

[0009] 自动化的挖掘任务往往只关注数据，而忽略了结果的实际业务价值。一方面，一些不合理的公式会降低方法的准确率；另一方面，挖掘的核对公式可能聚焦在一些没有核对必要的字段上，例如时间等。因此，需要通过某种方法来评估核对规则的业务价值，即理解其语义。

[0010] 4.如何高效处理大规模的数据。

[0011] 业务数据一般是大规模的，例如数据库的交易记录表中每天有超过百万甚至更高的记录数更新；在如此大规模的数据上进行挖掘需要消耗大量的计算资源，想要实现高效处理是极为困难的。

发明内容

[0012] 鉴于上述，本发明提供了一种业务数据核对脚本的自动化生成方法，结合先验知识和语义信息，从业务数据库中自动挖掘核对公式和前置条件组成核对规则，并生成对应的核对脚本。

[0013] 一种业务数据核对脚本的自动化生成方法，包括如下步骤：

[0014] (1)对人工编写的核对脚本进行解析，从中提取核对的数据源以及表达式的模式；

[0015] (2)从数据源中获取数据，并对数据进行预处理包括字段数据打平、字段分类、数据类型转换、特殊值处理、数据聚合；

[0016] (3)在预处理后的数据上挖掘核对公式，包括挖掘数值计算型的公式以及挖掘非数值一致型的公式；

[0017] (4)挖掘核对公式成立的前置条件；

[0018] (5)将核对公式和对应的前置条件转化为核对脚本。

[0019] 进一步地，所述步骤(1)的具体实现方式为：首先将SQL形式的核对脚本解析为AST(抽象语法树)形式；然后分析AST，获取表连接关系和聚合方式作为后续挖掘的数据源；通过分析AST，提取其中的表达式子树，构成表达式集，并归纳表达式的模式。

[0020] 进一步地，所述步骤(2)中字段数据打平包括对象类型字段以及键值对类型字段。

[0021] 进一步地，所述步骤(2)中字段分类包括对字段进行语义聚类，归纳字段类型，并设计分类规则；字段类型分为ID、数值、枚举、字符、时间、分区、版本7个大类，其中数值类型可细分为金额、数量和比例3个小类，枚举类型可细分为类型、状态和布尔标记3个小类。

[0022] 进一步地，所述字段分类所需要的语义信息包括字段名称、字段数据类型以及字段描述。

[0023] 进一步地，所述步骤(3)中使用改进的基于遗传的符号回归方法挖掘数值计算型的公式，具体地：将核对公式编码为公式语法树，结合从核对脚本中提取的表达式集和随机生成结果，初始化公式种群，利用适应度函数评估每个公式个体的适应度；根据字段分类评估公式个体的语义正确性，结合适应度选择优势个体；使用遗传和变异操作，生成新的公式树作为子代；重复适应度计算以及个体选择和子代生成的过程，直到出现公式满足条件或迭代次数达到上限。

[0024] 进一步地，所述步骤(3)中挖掘非数值一致型的公式，即将非数值类型的字段两两比较，需要保证两者的语义类型相同。

[0025] 进一步地，所述步骤(4)的具体实现方式为：首先在原始数据上验证公式是否成立，并用额外的标记字段记录结果；然后将每行数据构造成一个项集，在所有项集上执行关联规则挖掘，挖掘以公式成立为结果的条件项。

[0026] 进一步地，所述步骤(5)的具体实现方式为：首先定义核对脚本模板，使用核对公式和前置条件将核对脚本补充完整；然后检查脚本的语法是否正确，并测试运行。

[0027] 本发明方法的具体实现采用分布式框架，包括分布式数据仓库、分布式计算节点和分布式调度任务，实现大规模数据的并发处理。

[0028] 基于上述技术方案，本发明具有以下有益技术效果：

[0029] 1.本发明将回归方法、规则挖掘方法、SQL解析与生成技术相结合，针对大规模业务数据，自动挖掘核对公式和前置条件组成核对规则，并生成对应的核对脚本，极大地减少

了核对监控部署的人工成本,提高了核对监控的覆盖范围。

[0030] 2.本发明从人工编写的核对脚本中解析获取多样化的核对数据源,进一步提高了挖掘的完备性。

[0031] 3.本发明使用改进的符号回归方法,在实验中对比基线方法Gplearn在不同的公式复杂度、公式支持度、数据复杂程度上,准确性和召回率都有明显提升。

[0032] 4.本发明通过字段语义分类和核对公式的模式,智能判断核对规则的业务价值,极大提高了方法在实际场景下的可用性。

[0033] 5.本发明基于分布式架构,能够应对企业级的大规模数据场景。

附图说明

[0034] 图1为核对规则和核对脚本的实际案例示意图。

[0035] 图2为本发明业务数据核对脚本的自动化生成方法流程示意图。

[0036] 图3为解析核对脚本的流程示意图。

[0037] 图4为数据预处理的流程示意图。

[0038] 图5为核对公式挖掘的流程示意图。

[0039] 图6为前置条件挖掘的流程示意图。

[0040] 图7为核对脚本生成的流程示意图。

具体实施方式

[0041] 为了更为具体地描述本发明,下面结合附图及具体实施方式对本发明的技术方案进行详细说明。

[0042] 如图2所示,本发明业务数据核对脚本的自动化生成方法,包括如下步骤:

[0043] S10:解析现有的人工编写的核对脚本,从中提取核对的数据源以及表达式的模式,具体过程如图3所示:

[0044] S101.将SQL形式的核对脚本解析为抽象语法树AST形式;在本实施例中使用fastsql工具解析SQL语句。

[0045] S102.获取表连接关系和聚合方式,转化成挖掘的数据源。

[0046] 数据源是指挖掘核对规则的输入数据,其结构是包含列名的二维数据矩阵,需要在数据库中运行查询语句获取;数据源可以分为三类:单表数据源(数据来源于一张数据库表)、多表数据源(数据涉及多张数据库表,表之间通过外键相关联)、聚合型数据源(数据包含聚合操作)。遍历所有在步骤S101获取到的AST,获取Join、Group By、聚合函数(例如SUM)等节点,再解析获取表链接关系、外键、聚合方式,从而转化为上述三种类型的数据源。

[0047] S103.提取其中的表达式子树构成表达式集,并归纳表达式的模式。

[0048] 表达式是指由四则运算符和表字段构成的算术表达式,如price*number;通过遍历步骤S101中的所有AST,获取其中的表达式子树,去除重复的表达式构成表达式集。表达式的模式是指其合理的语义模式,即参与运算的字段、运算符以及结果之间的语义模式;最简单的二元运算,例如price*number等相似规则,可以被归纳为价格*数量=价格;更复杂的表达式可以视为上述过程的递归,例如(price*number)+price,可以先归纳出“价格*数量=价格”,并将表达式转换为price+price,再归纳出“价格+价格=价格”。本实施例中归

纳表达式模式的过程依赖业务领域相关知识与经验。

[0049] S20:从数据源中获取数据并进行预处理,具体过程如图4所示:

[0050] S201.字段数据打平;字段打平是指将复杂对象字段转化为多个简单键值对的形式,复杂对象包括JSON对象和键值对列表;对于多层JSON对象,使用点号分隔各层属性。

[0051] S202.字段分类;对所有数据库表字段进行语义聚类和人工分析,将字段按照语义分为了ID、数值、枚举、字符、时间、分区、版本7个大类,其中数值类型可以细分为金额、数量和比例3个小类,枚举类型可以细分为类型、状态和布尔标记3个小类。本实施例中通过定义每种类型的关键字,实现自动分类过程。

[0052] S203.数据类型转换;通过类型转换使得字段的数据类型和语义分类相一致,具体的对应关系为:ID(String类型)、数值(Integer、Float类型)、枚举(String类型)、字符(String)、时间(Datetime类型)、分区(String类型)、版本(String类型)。

[0053] S204.特殊值处理;数据中所有的数值型空值用NaN替换,字符型空值用空字符替换。

[0054] S205.数据聚合;当数据源为多表时,两张数据表通过外键连接后,记录之间可能存在一对多的关系,即主表的一条记录对应子表的多条记录,例如一件购物车包含多件商品;由于两表之间部分的数据关系无法在一行上表达,需要将子表记录按主表维度聚合。本发明通过表唯一主键对应关系来检查是否存在一对多映射的情况,若表主键未知,则计算每条记录的哈希值作为临时主键;在明确一对多关系后,使用聚合函数,按主表的主键维度聚合。本实施例中聚合函数包括求和、平均值、计数。

[0055] S30:核对公式挖掘;核对公式指数据表中字段之间的等式关系,核对公式分为两类:计算型公式(即变量是数值类型字段或常数,且包含四则运算符的等式)和一致性公式(变量是非数值类型字段,仅比较两个值是否相等的等式)。核对公式挖掘包含计算型公式挖掘和一致性公式挖掘两个过程。

[0056] 计算型公式挖掘,使用改进的符号回归方法挖掘计算型公式,具体流程如图5所示:

[0057] (1)将公式编码为语法树;指定某一字段(即某列数据)作为标签,定义为标签列,其他字段和常数将组合成一个算术表达式,将表达式部分表示为语法树,其中四则运算符是内部节点,变量和常数是叶节点(或称终止节点);子树可以被任何其他有效的表达式替换。

[0058] (2)基于先验的规则集和随机生成的混合方式初始化种群;利用从人工编写的核对脚本中获取的表达式集,在初始化种群是通过字段名称匹配的方式选择可用的表达式作为个体;同时,使用随机生成公式树的方式生成个体;最终的初始种群是上述两种方式1:1的混合。在本实施例中,随机生成公式树通过设置最大树深度、达到最大树深度之前节点包含子节点的概率、常数节点的取值范围,控制生成的结果。

[0059] (3)基于误差和相等率的个体适应度评估;适应度定义为平均绝对百分比误差(MAPE)和相等率(ER)的加权,如下式所示:

[0060] Fitness=0.2*MAPE+0.8*ER

[0061] $$MAPE = \frac{1}{m}\sum_{1}^{m}\left|\left(\frac{y_i - \hat{y}_i}{y_i}\right)\right|$$

[0062] $$ER = \frac{number(|y_i - \hat{y}_i| < p)}{number(y_i)}$$

[0063] 其中:m表示样本数,y_i表示实际值,$\hat{y}_i$表示估计值,p表示精度,number(y_i)为样本总数,$number(|y_i - \hat{y}_i| < p)$为预估值和实际值在某一精度p下相等的数量。

[0064] (4)基于适应度和公式语义的选择;利用步骤S103中归纳的表达式模式和步骤S202的字段分类结果,定义了表达式的语义正确性判定规则,具体如表1所示。

[0065] 表1

[0066]

加(减)法运算	金额 +/- 金额 = 金额
	数量 +/- 数量 = 数量
	比例 +/- 比例 = 比例
乘法运算	金额 * 数量 = 金额
	金额 * 比例 = 金额
	数量 * 比例 = 数量
	比例 * 比例 = 比例
除法运算	金额 / 金额 = 数量 或 比例
	金额 / 数量 = 金额
	金额 / 比例 = 金额
	数量 / 数量 = 比例
	比例 / 比例 = 比例

[0067] 特殊地,常数和未知类型可以被视为任意一种类型;若某个变量可能是多种类型,则在规则计算中需要考虑所有匹配的规则,具体的判定过程为:在表达式公式树上,将所有字段叶节点替换为语义分类类型,从叶节点自下而上的用表中所示的规则执行归并操作,最终得到表达式树根节点(即表达式)的语义类型,并判断与标签列的语义类型是否一致:若一致,则判定该公式语义正确,在遗传选择中赋予更高的优先级;若不一致,或在过程中遇到无法匹配规则的情况,例如"金额*金额",则判定公式语义不正确。

[0068] 另一方面,适应度计算的结果越小,公式在遗传选择中具有更高的优先级,最终的选择决策是适应度和公式语义正确性加权的结果。

[0069] (5)遗传、交叉、变异生成下一代种群;本实施例中采用GPlearn提供的五种遗传操作生成子代,包括Crossover、Subtree Mutation、Hoist Mutation、Point Mutation、Reproduction。

[0070] 一致性公式挖掘:首先两两比较非数值字段类型是否相同,若字段类型相同,则再比较字段值是否相同,相同的记录占总记录的比例即为公式成立的比例;若字段类型不相同,则认为公式的语义不正确,直接过滤。

[0071] 公式挖掘的结果将被存储到一个最小规则集中,通过去重保证每一个公式的唯一

性;若两个公式不同,但包含了相同的字段,且其中一个公式成立的记录范围大于另一个公式,则只保留前者。

[0072] S40:前置条件挖掘,具体过程如图6所示:

[0073] S401.在原始数据上验证公式是否成立,并用额外的标记字段记录结果。本实施例中在所有的原始输入数据上验证公式,用额外的一列数据FIT来标记公式是否在当前记录上成立,成立则FIT=1,反之FIT=0。

[0074] S402.将每行数据构造成一个项集。在本实施例中,每一项表示成“{字段名}={字段值}”的形式,包括FIT列。

[0075] S403.在所有项集上执行关联规则挖掘,挖掘以公式成立为结果的条件项。本实施例中,在所有项集上执行FPGrowth关联规则挖掘算法,挖掘以FIT=1为结果项的条件项,其中最小置信度阈值设置为1,保证前置条件成立时核对公式必然成立;最小支持度阈值是一个可变的参数,根据任务在0.1-0.3范围内设置;最后从结果中选择支持度最大,长度最小的一组条件项作为前置条件。

[0076] S50:核对脚本生成,具体过程如图7所示:

[0077] S501.定义核对脚本模板。在本实施例中,核对脚本模板对应步骤S102中的数据源类型,具体为:

[0078] 单表数据源:SELECT*FROM{表}WHERE{前置条件}AND{核对公式}

[0079] 多表数据源:SELECT*FROM{表1}JOIN{表2}ON{表外键关系}WHERE{前置条件}AND{核对公式}

[0080] 聚合型数据源:SELECT{聚合字段},{聚合函数}FROM{表}WHERE{前置条件}AND{核对公式1}GROUP BY{聚合字段}HAVING{核对公式2}

[0081] S502.使用核对公式和前置条件等将核对脚本补充完整。

[0082] S503.检查核对脚本的语法是否正确,并测试运行。在本实施例中SQL语法通过fastsql静态检查;在实际的数据库上执行SQL语句,查看能否正确运行并返回预期结果。

[0083] 上述对实施例的描述是为便于本技术领域的普通技术人员能理解和应用本发明,熟悉本领域技术的人员显然可以容易地对上述实施例做出各种修改,并把在此说明的一般原理应用到其他实施例中而不必经过创造性的劳动。因此,本发明不限于上述实施例,本领域技术人员根据本发明的揭示,对于本发明做出的改进和修改都应该在本发明的保护范围之内。

核对规则

前置条件	$trade\_order.delivery\_status = '9'$ **and** $trade\_order.actual\_fee \neq 0$ **and** $trade\_order.biz\_area = 'PH'$ **and** $pay\_plan.biz\_area = 'PH'$
核对公式	$\Longrightarrow trade\_order.pay\_status = pay\_plan.status$

核对脚本

```
SELECT *
FROM (
    SELECT *
    FROM x.trade_order
    WHERE delivery_status = '9'
    AND actual_fee != 0
    AND biz_area = 'PH'
) a
LEFT JOIN (
    SELECT *
    FROM x.pay_plan
    WHERE biz_area = 'PH'
) b
ON a.physical_package_no = b.pay_request_no
WHERE a.pay_status <> b.status
```

图1

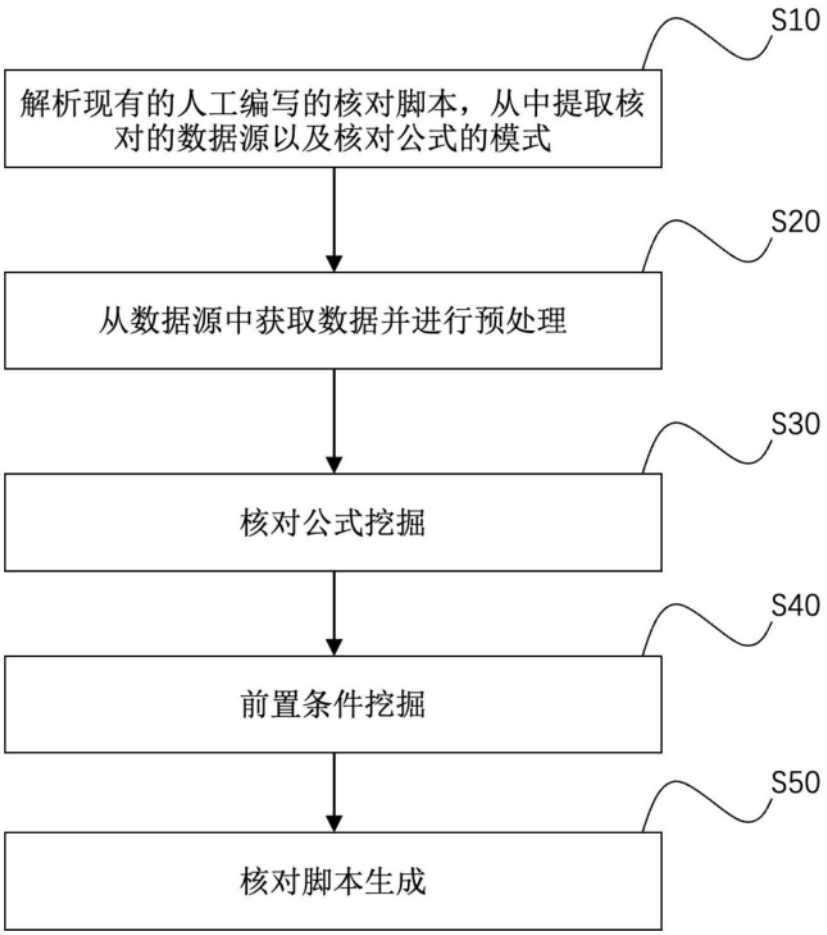

S10
解析现有的人工编写的核对脚本，从中提取核对的数据源以及核对公式的模式
S20
从数据源中获取数据并进行预处理
S30
核对公式挖掘
S40
前置条件挖掘
S50
核对脚本生成

图2

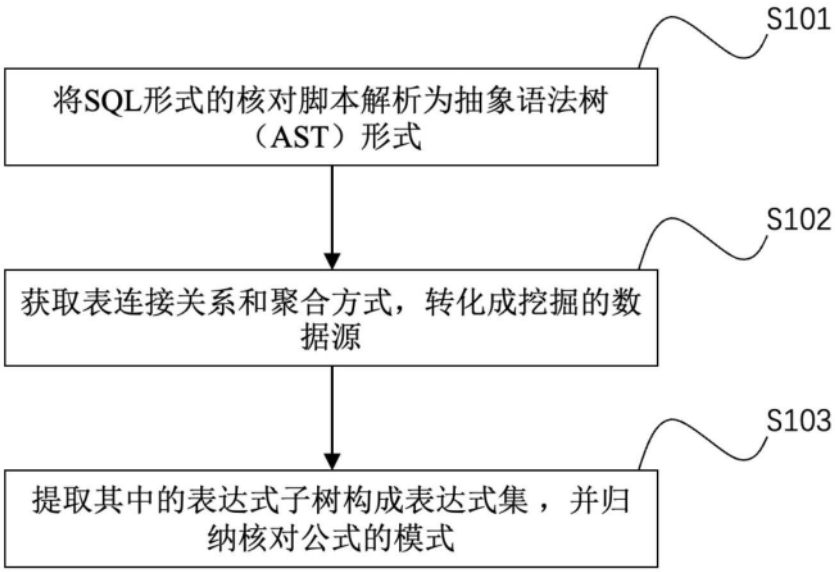

S101
将SQL形式的核对脚本解析为抽象语法树（AST）形式
S102
获取表连接关系和聚合方式，转化成挖掘的数据源
S103
提取其中的表达式子树构成表达式集，并归纳核对公式的模式

图3

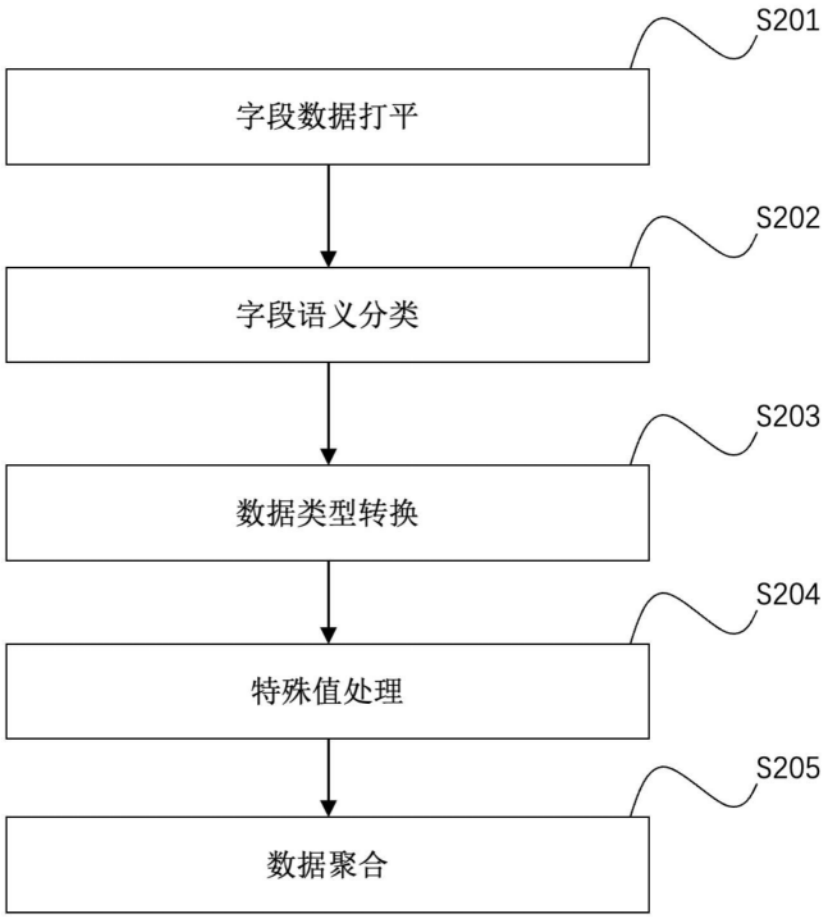


图4

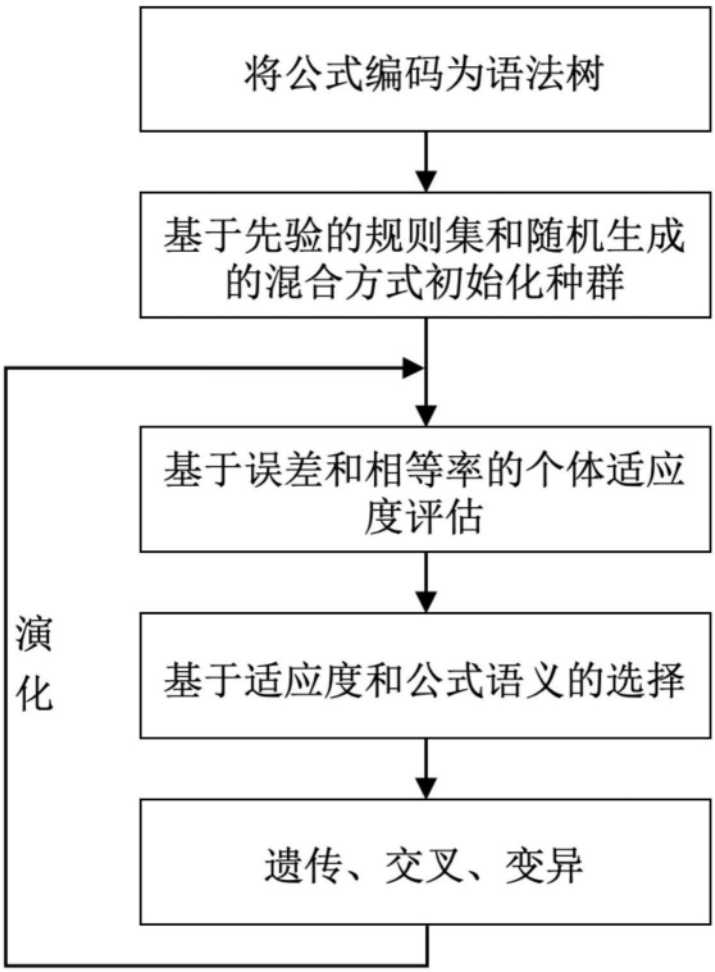

将公式编码为语法树
基于先验的规则集和随机生成的混合方式初始化种群
基于误差和相等率的个体适应度评估
基于适应度和公式语义的选择
遗传、交叉、变异
演化

图5

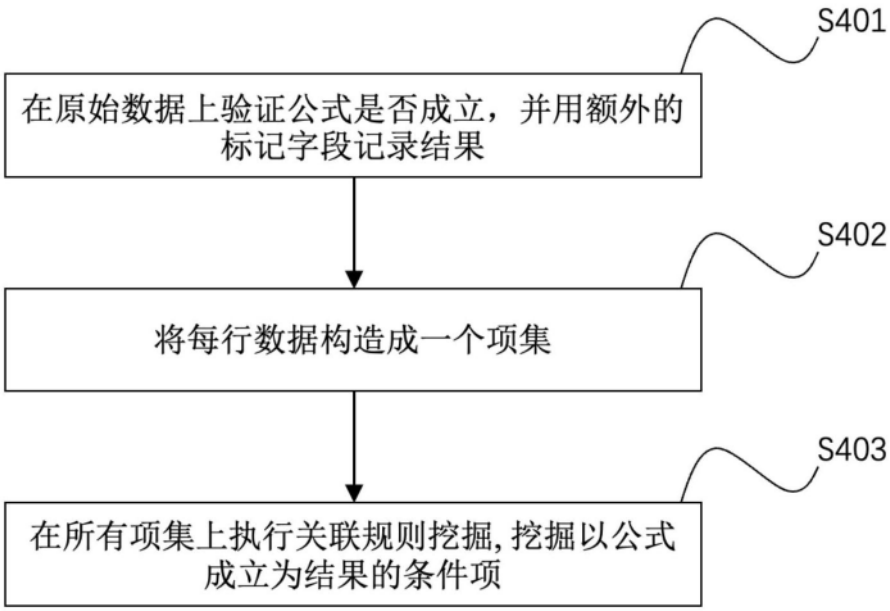

S401
在原始数据上验证公式是否成立，并用额外的标记字段记录结果
S402
将每行数据构造成一个项集
S403
在所有项集上执行关联规则挖掘,挖掘以公式成立为结果的条件项

图6

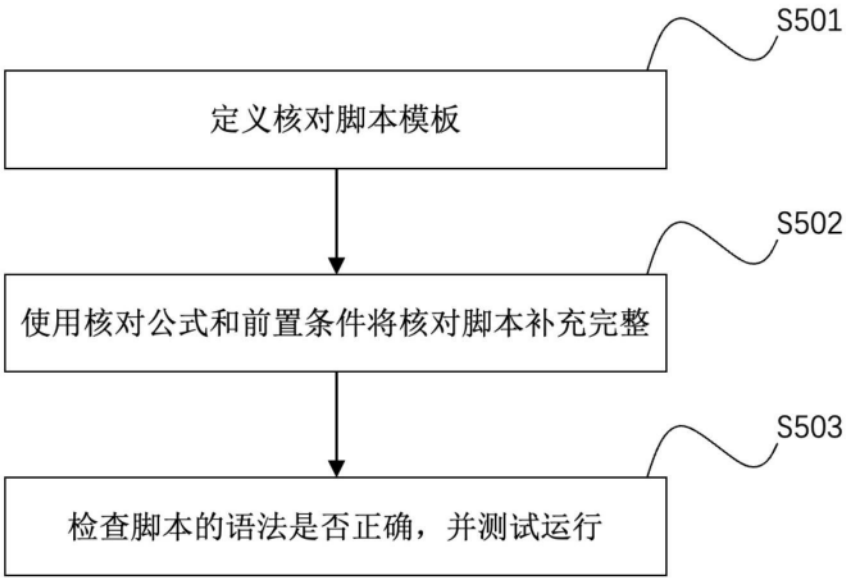

S501
定义核对脚本模板
S502
使用核对公式和前置条件将核对脚本补充完整
S503
检查脚本的语法是否正确，并测试运行

图7