(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 109408592 B
(45)授权公告日 2021.09.24

(21)申请号 201811190148.0

(22)申请日 2018.10.12

(65)同一申请的已公布的文献号
申请公布号 CN 109408592 A

(43)申请公布日 2019.03.01

(73)专利权人 北京聚云位智信息科技有限公司
地址 100101 北京市朝阳区北苑路甲13号院1号楼北辰泰岳大厦701

(72)发明人 张德辉

(74)专利代理机构 北京卓爱普专利代理事务所(特殊普通合伙) 11920
代理人 王玉松

(51)Int.Cl.
G06F 16/27(2019.01)
G06N 5/02(2006.01)

(56)对比文件
CN 107463564 A,2017.12.12
CN 108090516 A,2018.05.29
CN 106250987 A,2016.12.21
CN 102780264 A,2012.11.14
CN 101976375 A,2011.02.16
US 2017116524 A1,2017.04.27
WO 2018107128 A1,2018.06.14
CN 108008942 A,2018.05.08
黄引翔.网络流量分类中特征工程的研究.《中国优秀博硕士学位论文全文数据库(硕士)信息科技辑》.2018,(第02期),I139-114.
Tobias Schreck 等.Visual Feature Space Analysis for Unsupervised Effectiveness Estimation and Feature Engineering.《2006 IEEE International Conference on Multimedia and Expo》.2006,925-928.

审查员 何洋

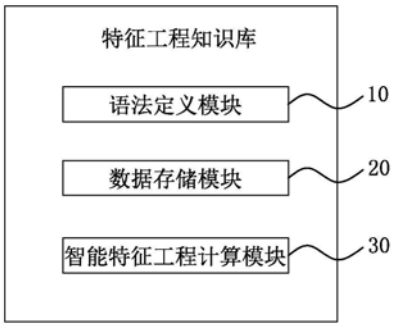
权利要求书3页 说明书7页 附图6页

(54)发明名称

一种决策型分布式数据库系统中AI的特征工程知识库及其实现方法

(57)摘要

本发明属于特征工程知识库,特别涉及一种决策型分布式数据库系统中AI的特征工程知识库及其实现方法;其一种决策型分布式数据库系统中AI的特征工程知识库,所述特征工程知识库包括:语法定义模块。本发明提供一种新的决策型分布式数据库系统中AI的特征工程知识库及其实现方法,该决策型分布式数据库系统中AI的特征工程知识库及其实现方法增加智能特征工程的功能,很大程度上降低了特征工程的门槛,即便是人工指定领域数据类型的情况下,由于数据分析师本身就掌握业务知识,识别领域数据类型相比较掌握各种特征工程处理函数以及其组合的适用场景而言已经是非常做到的事情了,不仅提高了特征工程的效率,还提高了整个AI项目实施的效率。



CN 109408592 B

1.一种决策型分布式数据库系统中AI的特征工程知识库,其特征在于,所述特征工程知识库包括:

语法定义模块(10),用于定义实现AI SQL的语法;

数据存储模块(20),用于存储自动化特征工程的领域数据类型、特征数据以及关联关系;

智能特征工程计算模块(30),用于支持智能特征工程的计算,并生成最后的特征向量;

所述数据存储模块(20)包括:

领域数据类型存储单元(201),用于存放系统内置的领域数据类型,包括年龄、地址、邮箱、性别、手机号以及身份证号;

关联关系存储单元(202),用于存放与领域数据类型相匹配的特征处理算法的关联关系;

特征数据存储单元(203),用于存放每个领域数据类型特征数据,每个领域数据类型的特征数据包括基本数据类型、能接受源数据类型列表及其转换器列表、简单匹配规则、典型样本数据和不属于此领域数据类型的样本数据以及类型识别模型中的一个或多个;

所述智能特征工程计算模块(30)包括:

特征列检查模块(301),用于取出一个尚未处理的列,对之进行特征列检查,判断该列是否存在领域数据类型的标记,若存在,则向处理模块(303)发送指令,若不存在,则向识别模块(302)发送指令;

识别模块(302),用于根据特征数据存储单元(203)自动识别其领域数据类型,同时,向处理模块(303)发送指令;

处理模块(303),用于对于已经存在领域数据类型的列,根据关联关系存储单元(202)找到其对应的特征处理算法,采用对应的算法对该列进行处理;

管理模块(304),用于判断是否还有未处理的列,若是,则向特征列检查模块(301)发送指令,若不是,则通过两两计算的方式,去除两个相关度高的列中一个,并生成最后的特征向量。

2.根据权利要求1所述的决策型分布式数据库系统中AI的特征工程知识库,其特征在于,所述语法定义模块(10)包括:

AI模型创建的语法定义单元(101),用于对AI模型创建的语法进行定义;

AI模型更新的语法定义单元(102),用于对AI模型更新的语法进行定义;

AI模型评估的语法定义单元(103),用于对AI模型评估的语法进行定义;

手动指令领域数据类型的语法定义单元(104),用于对手动指令领域数据类型的语法进行定义;

AI模型应用的语法定义单元(105),用于对AI模型应用的语法进行定义,其中AI模型的应用是预测函数根据用户输入的数据集合、选择的建好的模型生成预测结果,所述预测结果包括:分类结果、趋势、关联关系挖掘以及推荐结果;

AI SQL其他语法定义单元(106),用于对AL SQL的其他语法进行定义。

3.根据权利要求1所述的决策型分布式数据库系统中AI的特征工程知识库,其特征在于,所述识别模块(302)对于没有标记领域数据类型的列,采用简单匹配规则进行枚举、正则的简单匹配,得到领域数据类型或采用类型识别模型来高级匹配得到相应的置信度,并

根据不同的权重计算出整体的置信度，置信度最高的领域数据类型为所需得出的领域数据类型。

4.根据权利要求1所述的决策型分布式数据库系统中AI的特征工程知识库，其特征在于，所述特征工程知识库还通讯连接有AI SQL解析器(1)、关系表及AI模型元数据库(2)以及支持AI模型存储的分布式存储器(3)；

AI SQL解析器(1)，用于解析AI SQL并生成逻辑执行计划，在解析AI SQL过程中，调取与特征工程知识库相通讯的关系表及AI模型元数据库(2)内存储的信息对AI SQL进行除了语法格式之外的正确性验证及资源对象定位，在生成逻辑执行计划过程中，若包含特征工程计算则生成对应的特征工程处理算法的运算步骤，其中，若是采用智能特征工程计算，则构造一通过与数据存储模块(20)进行匹配的系列计算步骤；

关系表及AI模型元数据库(2)，用于存储元数据表信息；

支持AI模型存储的分布式存储器(3)，用于管理和存储关系表或AI模型的数据信息。

5.一种决策型分布式数据库系统中AI的特征工程知识库实现方法，其特征在于，所述方法包括：

S1:通过语法定义模块(10)定义实现AI SQL的语法；

S2:通过数据存储模块(20)存储自动化特征工程的领域数据类型、特征数据以及关联关系；

S3:通过智能特征工程计算模块(30)支持智能特征工程的计算，并生成最后的特征向量；

步骤S2包括：

S21:通过领域数据类型存储单元(201)存放系统内置的领域数据类型；

S22:通过关联关系存储单元(202)存放与领域数据类型相匹配的特征处理算法的关联关系；

S23:通过特征数据存储单元(203)存放每个领域数据类型特征数据；

步骤S3包括：

S31:通过特征列检查模块(301)取出一个尚未处理的列，对之进行特征列检查，判断该列是否存在领域数据类型的标记，若存在，则进行步骤S33，若不存在，则进行步骤S32；

S32:通过识别模块(302)根据特征数据存储单元(203)自动识别其领域数据类型，同时进行步骤S33；

S33:通过处理模块(303)对于已经存在领域数据类型的列，根据关联关系存储单元(202)找到其对应的特征处理算法，采用对应的算法对该列进行处理；

S34:通过管理模块(304)判断是否还有未处理的列，若是，则向特征列检查模块(301)发送指令，若不是，则通过两两计算的方式，去除两个相关度高的列中一个，并生成最后的特征向量。

6.根据权利要求5所述的决策型分布式数据库系统中AI的特征工程知识库实现方法，其特征在于，步骤S1包括：

S11:通过AI模型创建的语法定义单元(101)对AI模型创建的语法进行定义；

S12:通过AI模型更新的语法定义单元(102)对AI模型更新的语法进行定义；

S13:通过AI模型评估的语法定义单元(103)对AI模型评估的语法进行定义；

S14:通过手动指令领域数据类型的语法定义单元(104)对手动指令领域数据类型的语法进行定义;

S15:通过AI模型应用的语法定义单元(105)对AI模型应用的语法进行定义;

S16:通过AI SQL其他语法定义单元(106)对AL SQL的其他语法进行定义。

一种决策型分布式数据库系统中AI的特征工程知识库及其实现方法

技术领域

[0001] 本发明属于特征工程知识库,特别涉及一种决策型分布式数据库系统中AI的特征工程知识库及其实现方法。

背景技术

[0002] 现有的特征工程知识库在特征工程方面比较依赖用户自己决定处理方式,这就要求用户具备非常专业的AI技能的同时还需投入较多的精力完成建模所需的特征工程。这样导致AI计算相关项目实施风险高、周期长;这主要在于现实中存在各种来源的数据,除了数据库的数据外还包括各种千变万化的数据来源(如互联网、excel等等),而作为通用的AI计算软件包很难做出一些假设去自动化特征工程,以任意一个数值型字段为例,究竟是采用以x为底的对数函数还是采用开n次方来做规范化是很难自动决策的,事实上特征处理的可选的函数空间本身就是无限维度的超级空间;从而会降低特征工程以及整个AI项目的实施效率。

发明内容

[0003] 针对上述问题,本发明提供一种新的决策型分布式数据库系统中AI的特征工程知识库及其实现方法,该新的决策型分布式数据库系统中AI的特征工程知识库及其实现方法智能特征工程提高了特征工程的效率,从而提高了整个AI项目实施的效率。

[0004] 本发明具体技术方案如下:

[0005] 本发明提供一种决策型分布式数据库系统中AI的特征工程知识库,所述特征工程知识库包括:

[0006] 语法定义模块,用于定义实现AI SQL的语法;

[0007] 数据存储模块,用于存储自动化特征工程的领域数据类型、特征数据以及关联关系;

[0008] 智能特征工程计算模块,用于支持智能特征工程的计算,并生成最后的特征向量。

[0009] 本发明的有益效果如下:

[0010] 本发明提供一种新的决策型分布式数据库系统中AI的特征工程知识库及其实现方法,该决策型分布式数据库系统中AI的特征工程知识库及其实现方法增加智能特征工程的功能,很大程度上降低了特征工程的门槛,即便是人工指定领域数据类型的情况下,由于数据分析师本身就掌握业务知识,识别领域数据类型相比较掌握各种特征工程处理函数以及其组合的适用场景而言已经是非常做到的事情了,不仅提高了特征工程的效率,还提高了整个AI项目实施的效率。

附图说明

[0011] 图1为实施例1决策型分布式数据库系统中AI的特征工程知识库

[0012] 的结构框图；

[0013] 图2为实施例2语法定义模块的结构框图；

[0014] 图3为实施例2数据存储模块的结构框图；

[0015] 图4为实施例3智能特征工程计算模块的结构框图；

[0016] 图5为实施例4决策型分布式数据库的结构框图；

[0017] 图6为实施例5决策型分布式数据库系统中AI的特征工程知识库

[0018] 实现方法的流程图；

[0019] 图7为实施例6步骤S1的流程图；

[0020] 图8为实施例6步骤S2的流程图；

[0021] 图9为实施例7步骤S3的流程图。

具体实施方式

[0022] 下面结合附图和以下实施例对本发明作进一步详细说明。

[0023] 实施例1

[0024] 本发明实施例1提供一种决策型分布式数据库系统中AI的特征工程知识库，如图1所示，所述特征工程知识库包括：

[0025] 语法定义模块10，用于定义实现AI SQL的语法；

[0026] 数据存储模块20，用于存储自动化特征工程的领域数据类型、特征数据以及关联关系；

[0027] 智能特征工程计算模块30，用于支持智能特征工程的计算，并生成最后的特征向量。

[0028] 本发明增加智能特征工程的功能，很大程度上降低了特征工程的门槛，即便是人工指定领域数据类型的情况下，由于数据分析师本身就掌握业务知识，识别领域数据类型相比较掌握各种特征工程处理函数以及其组合的适用场景而言已经是非常做到的事情了，不仅提高了特征工程的效率，还提高了整个AI项目实施的效率；同时采用新的语法定义来实现AI SQL并且AI SQL这种SQL 2011标准的扩展SQL，减低了AI使用的门槛以及对AI编程开发人员的需求，节省了AI项目的成本，另外在充分利用整个集群整体资源(GPU，CPU和内存等等)的同时，对复杂的数据分析任务能够整体进行优化执行，相对于现有方案而言，消耗更少的硬件资源，从而节省了硬件成本。

[0029] 实施例2

[0030] 一种决策型分布式数据库系统中AI的特征工程知识库，如图2所示，与实施例1不同的是：所述语法定义模块10包括：

[0031] AI模型创建的语法定义单元101，用于对AI模型创建的语法进行定义，

[0032] <model definition>::＝CREATE[<model scope>]MODEL<model name>

[0033] AS<model constructor name><SQL argument list>

[0034] <model scope>::＝<global or local>TEMPORARY

[0035] <global or local>::＝GLOBAL|LOCAL

[0036] <model name>::＝<local or schema qualified name>

[0037] <local or schema qualified name>::＝同SQL 2011规范中定义

[0038] <model constructor name>::=[<schema name><period>]<qualified identifier>

[0039] <schema name>::=同SQL 2011规范中定义

[0040] <period>::=.

[0041] <qualified identifier>::=同SQL 2011规范中定义

[0042] <SQL argument list>::=同SQL 2011规范中定义;

[0043] AI模型更新的语法定义单元102,用于对AI模型更新的语法进行定义,

[0044] <update model definition>::=UPDATE[<model scope>]MODEL<model name>

[0045] AS<model constructor name><SQL argument list>;

[0046] AI模型评估的语法定义单元103,用于对AI模型评估的语法进行定义,

[0047] <evaluate model definition>::=SELECT<select list>FROM<model evaluation function name><SQL argument list>

[0048] <select list>::=同SQL 2011规范中定义

[0049] <model evaluation function name>::=[<schema name><period>]

[0050] <qualified identifier>

[0051] 其中<SQL argument list>必须至少包含一个MODEL的直接名称或构造表达式(比如采用调用DECISION_TREE_TRAIN构造的一个临时匿名的MODEL);

[0052] 手动指令领域数据类型的语法定义单元104,用于对手动指令领域数据类型的语法进行定义,

<alter column AI-domain type definition> ::= ALTER [COLUMN] <column name>

[0053] <alter column AI-domain type clause>

<alter column AI-domain type clause> ::= SET AI DOMAIN TYPE <AI-domain type>

[0054] <AI-domain type>::=<basic AI-domain type>|<user defined AI-domain type>

[0055] |<collection AI-domain type>

[0056] <basic AI-domain type>::=<qualified identifier>

[0057] <user defined AI-domain type>::=<qualified identifier>

[0058] <collection AI-domain type>::=<array AI-domain type>|<multiset AI-domain type>

[0059] <array AI-domain type>::=<AI-domain type>ARRAY

[0060] <multiset AI-domain type>::=<AI-domain type>MULTISET

[0061] 其中<basic AI-domain type>中均为系统内置在特征知识库中的领域数据类型,比如年龄、地址、邮箱、性别、手机号、身份证号、百分制评分、五分制评分、海拔等等。<user defined AI-domain type>是用户扩展的领域数据类型;

[0062] AI模型应用的语法定义单元105,用于对AI模型应用的语法进行定义,其中AI模型

的应用是预测函数根据用户输入的数据集合、选择的建好的模型生成预测结果，所述预测结果包括：分类结果、趋势、关联关系挖掘、推荐结果等等，这些结果也是一个集合，特殊情况下，这个集合可能只有一行数据，

[0063] <AI model apply definition>::=SELECT<select list>FROM<model apply function name><SQL argument list>

[0064] <model apply function name>::=[<schema name><period>]<qualified identifier>；

[0065] AI SQL其他语法定义单元106，用于对AL SQL的其他语法进行定义，其他和SQL 2011规范语法兼容保持一致。

[0066] 如图3所示，本实施例中，所述数据存储模块20包括：

[0067] 领域数据类型存储单元201，用于存放系统内置的领域数据类型，包括年龄、地址、邮箱、性别、手机号以及身份证号；

[0068] 关联关系存储单元202，用于存放与领域数据类型相匹配的特征处理算法的关联关系；

[0069] 特征数据存储单元203，用于存放每个领域数据类型特征数据，每个领域数据类型的特征数据包括基本数据类型、能接受源数据类型列表及其转换器列表、简单匹配规则、典型样本数据和不属于此领域数据类型的样本数据以及类型识别模型中的一个或多个。

[0070] 本发明中对AL SQL实现的语法进行具体的定义，并利用定义后的语法按照上述步骤存储自动化特征工程的内容。

[0071] 本发明中特征工程知识库一方面存放了系统内置在特征知识库中的领域数据类型，比如年龄、地址、邮箱、性别、手机号、身份证号、百分制评分、五分制评分、海拔等等。同时也支持注册用户扩展的领域数据类型。

[0072] 特征工程知识库还存放了和领域数据类型相匹配的特征处理算法的关联关系，以便通过领域数据类型快速找到其对应的特征处理算法。

[0073] 为支持自动识别领域数据类型的功能，特征工程知识库还存放了每个领域数据类型特征数据，每个领域数据类型的特征数据包括如下一种或多种组合：

[0074] 基本数据类型(比如年龄的基本数据类型是Integer)。

[0075] 能接受源数据类型列表及其转换器列表(即从源数据类型可以转换过来，比如字符串“1”可以通过转换器转换为数字1，从而符合年龄的基本数据类型)。

[0076] 简单匹配规则(比如正则表达式，范围约束等等)。

[0077] 典型样本数据和不属于此领域数据类型的样本数据，如果是可枚举的则样本数据可以是整个枚举全集。

[0078] 类型识别模型，该模型可以是直接从样本数据和不属于此领域数据类型的样本数据训练得到，也可以从已经训练好的模型导入。这些识别模型通常采用NER(Named-entity recognition，即命名实体识别)相关的算法进行训练，比如采用Apache OpenNLP中NER的算法进行模型训练。

[0079] 这样，对于一个没有标记领域数据类型的值，都可以通过上面的特征数据进行枚举、正则的简单匹配，也可以采用类型识别模型来高级匹配得到相应的置信度，然后根据不同权重计算出整体的置信度，最后选择置信度最高的那个领域数据类型，因而特征工程知

识库能够支持智能特征工程的计算。

[0080] 实施例3

[0081] 一种决策型分布式数据库系统中AI的特征工程知识库，如图4所示，与实施例2不同的是：所述智能特征工程计算模块30包括：

[0082] 特征列检查模块301，用于取出一个尚未处理的列，对之进行特征列检查，判断该列是否存在领域数据类型的标记，若存在，则向处理模块303发送指令，若不存在，则向识别模块302发送指令；

[0083] 识别模块302，用于根据特征数据存储单元203自动识别其领域数据类型，同时，向处理模块303发送指令；

[0084] 处理模块303，用于对于已经存在领域数据类型的列，根据关联关系存储单元202找到其对应的特征处理算法，采用对应的算法对该列进行处理；

[0085] 管理模块304，用于判断是否还有未处理的列，若是，则向特征列检查模块301发送指令，若不是，则通过两两计算的方式，去除两个相关度高的列中一个，并生成最后的特征向量。

[0086] 本实施例中所述识别模块302对于没有标记领域数据类型的列，采用简单匹配规则进行枚举、正则的简单匹配，得到领域数据类型或采用类型识别模型来高级匹配得到相应的置信度，并根据不同的权重计算出整体的置信度，置信度最高的领域数据类型为所需得出的领域数据类型。

[0087] 本发明中采用上述步骤进行智能特征工程的计算，很大程度上降低了特征工程的门槛，不仅提高了特征工程的效率，还提高了整个AI项目实施的效率。

[0088] 实施例4

[0089] 一种决策型分布式数据库系统中AI的特征工程知识库，如图5所示，与实施例1不同的是：所述特征工程知识库还通讯连接有AI SQL解析器1、关系表及AI模型元数据库2、支持AI模型存储的分布式存储器3、执行计划优化器4、AI算法库5以及支持AI计算的分布式执行器6；

[0090] AI SQL解析器1，用于解析AI SQL并生成逻辑执行计划，在解析AI SQL过程中，调取与特征工程知识库相通讯的关系表及AI模型元数据库2内存储的信息对AI SQL进行除了语法格式之外的正确性验证及资源对象定位，在生成逻辑执行计划过程中，若包含特征工程计算则生成对应的特征工程处理算法的运算步骤，其中，若是采用智能特征工程计算，则构造一通过与数据存储模块20进行匹配的系列计算步骤；

[0091] 关系表及AI模型元数据库2，用于存储元数据表信息；

[0092] 支持AI模型存储的分布式存储器3，用于管理和存储关系表或AI模型的数据信息；

[0093] 执行计划优化器4，用于将逻辑执行计划进行优化，并生成执行代价较小的物理执行计划；

[0094] AI算法库5，用于集成多种AI算法以及分布式计算引擎，其中数据在不同编程语言或AI库间高效的转换传输采用Apache Arrow作为公共数据层；

[0095] 支持AI计算的分布式执行器6，用于将物理执行计划分解成多个步骤进行运行。

[0096] 本发明由如上关键组件构成一个基于AI SQL和智能特征工程的决策型分布式数据库，从结构上看，一套决策型数据库就可以胜任了，这样数据不再需要从数据库导出来，

也不存在数据导出的安全隐患;从功能上看,决策型数据库具备智能特征工程的能力,这是现有分析型数据库加上二次开发的AI分析程序的方案所不具备的;从成本看,决策型数据库提供AI SQL这种SQL 2011标准的扩展SQL,减低了AI使用的门槛以及对AI编程开发人员的需求,节省了AI项目的成本,另外决策型数据库功能上覆盖了完整的数据/AI模型管理、数据传统OLAP分析,AI复杂分析全过程,在充分利用整个集群整体资源(GPU,CPU和内存等等)的同时,对复杂的数据分析任务能够整体进行优化执行,相对于现有方案而言,消耗更少的硬件资源,从而节省了硬件成本;从效率看,决策型数据库的AI SQL更容易使用,智能特征工程提高了特征工程的效率,从而提高了整个AI项目实施的效率。

[0097] 实施例5

[0098] 一种决策型分布式数据库系统中AI的特征工程知识库实现方法,如图6所示,所述方法包括:

[0099] S1:通过语法定义模块10定义实现AI SQL的语法;

[0100] S2:通过数据存储模块20存储自动化特征工程的领域数据类型、特征数据以及关联关系;

[0101] S3:通过智能特征工程计算模块30支持智能特征工程的计算,并生成最后的特征向量。

[0102] 本发明增加智能特征工程的功能,很大程度上降低了特征工程的门槛,即便是人工指定领域数据类型的情况下,由于数据分析师本身就掌握业务知识,识别领域数据类型相比较掌握各种特征工程处理函数以及其组合的适用场景而言已经是非常做到的事情了,不仅提高了特征工程的效率,还提高了整个AI项目实施的效率;同时采用新的语法定义来实现AI SQL并且AI SQL这种SQL 2011标准的扩展SQL,减低了AI使用的门槛以及对AI编程开发人员的需求,节省了AI项目的成本,另外在充分利用整个集群整体资源(GPU,CPU和内存等等)的同时,对复杂的数据分析任务能够整体进行优化执行,相对于现有方案而言,消耗更少的硬件资源,从而节省了硬件成本。

[0103] 实施例6

[0104] 一种决策型分布式数据库系统中AI的特征工程知识库实现方法,如图7所示,与实施例5不同的是:步骤S1包括:

[0105] S11:通过AI模型创建的语法定义单元101对AI模型创建的语法进行定义;

[0106] S12:通过AI模型更新的语法定义单元102对AI模型更新的语法进行定义;

[0107] S13:通过AI模型评估的语法定义单元103对AI模型评估的语法进行定义;

[0108] S14:通过手动指令领域数据类型的语法定义单元104对手动指令领域数据类型的语法进行定义;

[0109] S15:通过AI模型应用的语法定义单元105对AI模型应用的语法进行定义;

[0110] S16:通过AI SQL其他语法定义单元106对AL SQL的其他语法进行定义。

[0111] 如图8所示,本实施例中步骤S2包括:

[0112] 步骤S2包括:

[0113] S21:通过领域数据类型存储单元201存放系统内置的领域数据类型;

[0114] S22:通过关联关系存储单元202存放与领域数据类型相匹配的特征处理算法的关联关系;

[0115] S23:通过特征数据存储单元203存放每个领域数据类型特征数据。

[0116] 本发明中对AL SQL实现的语法进行具体的定义,并利用定义后的语法按照上述步骤存储自动化特征工程的内容。

[0117] 实施例7

[0118] 一种决策型分布式数据库系统中AI的特征工程知识库实现方法,如图9所示,与实施例5不同的是:步骤S3包括:

[0119] S31:通过特征列检查模块301取出一个尚未处理的列,对之进行特征列检查,判断该列是否存在领域数据类型的标记,若存在,则进行步骤S33,若不存在,则进行步骤S32;

[0120] S32:通过识别模块302根据特征数据存储单元203自动识别其领域数据类型,同时进行步骤S33;

[0121] S33:通过处理模块303对于已经存在领域数据类型的列,根据关联关系存储单元202找到其对应的特征处理算法,采用对应的算法对该列进行处理;

[0122] S34:通过管理模块304判断是否还有未处理的列,若是,则向特征列检查模块301发送指令,若不是,则通过两两计算的方式,去除两个相关度高的列中一个,并生成最后的特征向量。

[0123] 本发明中采用上述步骤进行智能特征工程的计算,很大程度上降低了特征工程的门槛,不仅提高了特征工程的效率,还提高了整个AI项目实施的效率。

[0124] 以上所述实施例仅仅是本发明的优选实施方式进行描述,并非对本发明的范围进行限定,在不脱离本发明设计精神的前提下,本领域普通技术人员对本发明的技术方案作出的各种变形和改进,均应落入本发明的权利要求书确定的保护范围内。

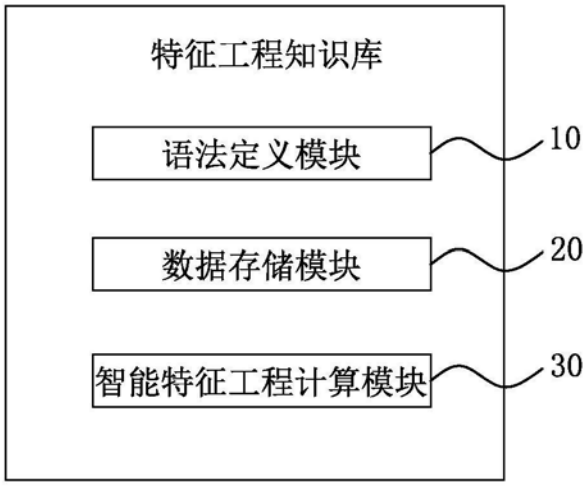

特征工程知识库
语法定义模块
10
数据存储模块
20
智能特征工程计算模块
30

图1

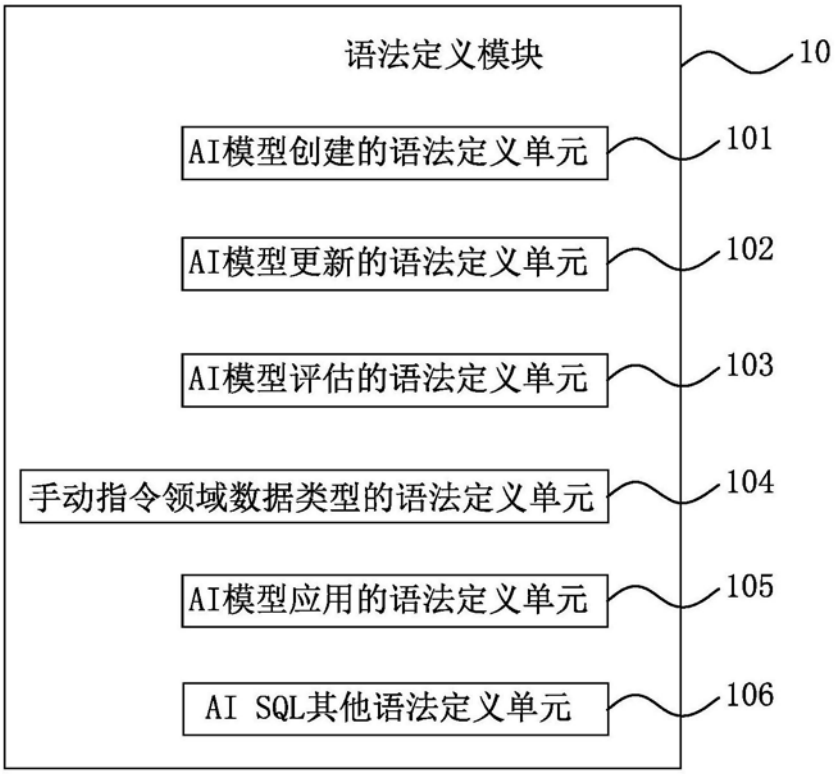

语法定义模块
10
AI模型创建的语法定义单元
101
AI模型更新的语法定义单元
102
AI模型评估的语法定义单元
103
手动指令领域数据类型的语法定义单元
104
AI模型应用的语法定义单元
105
AI SQL其他语法定义单元
106

图2

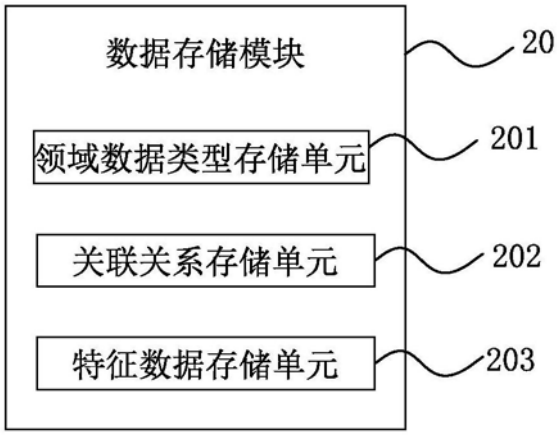

数据存储模块
20
领域数据类型存储单元
201
关联关系存储单元
202
特征数据存储单元
203

图3

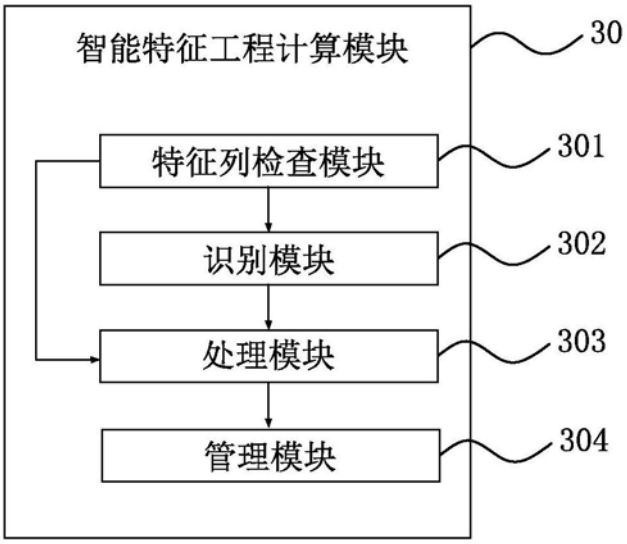

智能特征工程计算模块
30
特征列检查模块
301
识别模块
302
处理模块
303
管理模块
304

图4

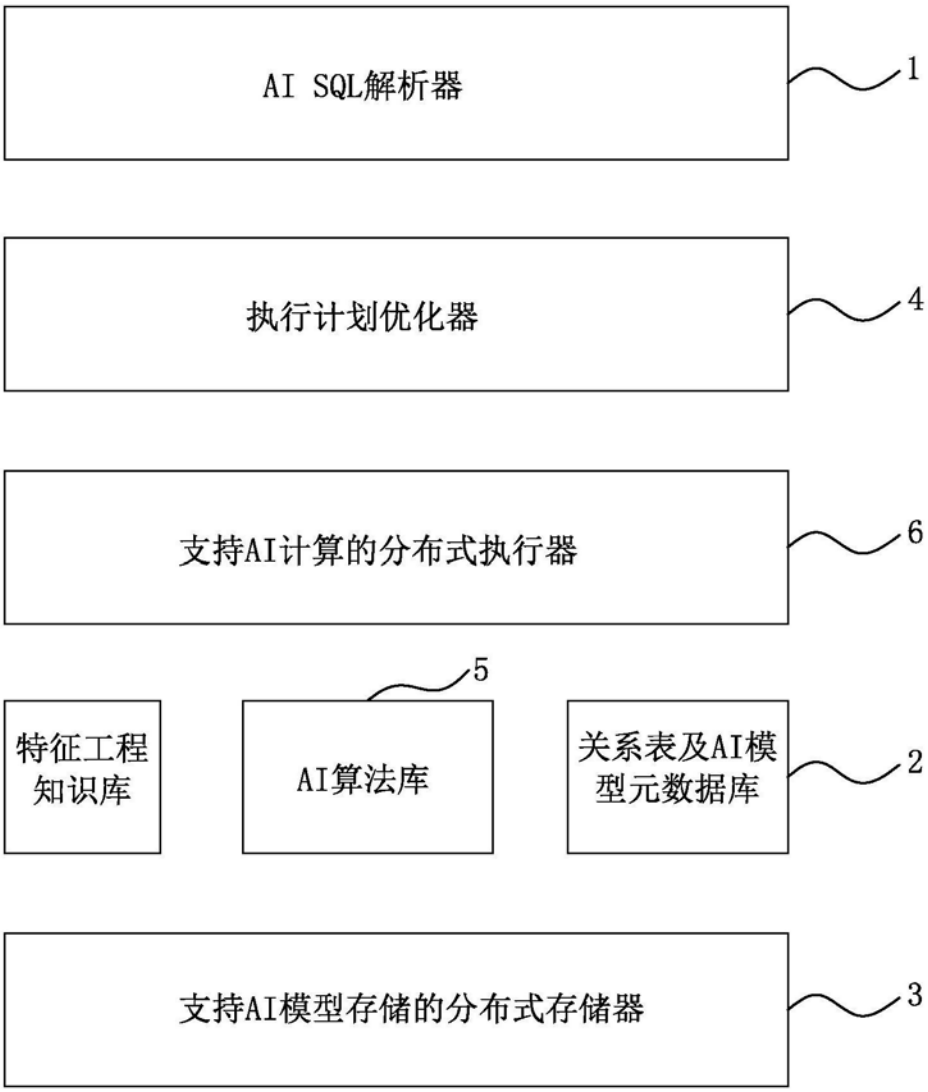

AI SQL解析器
1
执行计划优化器
4
支持AI计算的分布式执行器
6
5
特征工程知识库
AI算法库
关系表及AI模型元数据库
2
支持AI模型存储的分布式存储器
3

图5

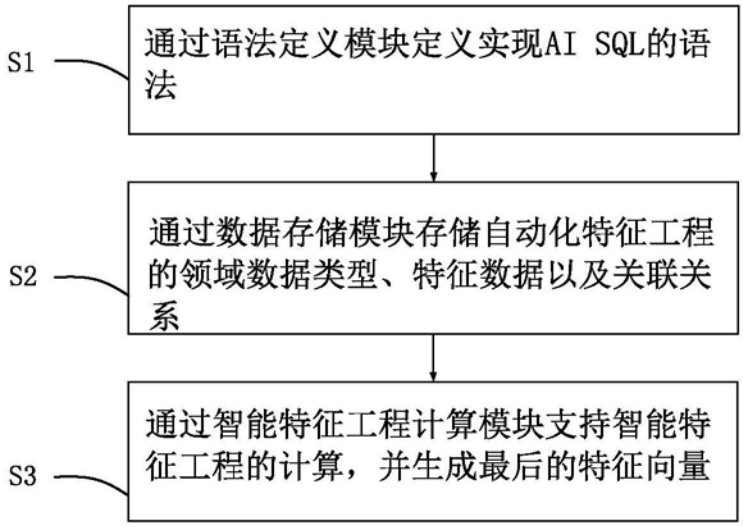

S1
通过语法定义模块定义实现AI SQL的语法
S2
通过数据存储模块存储自动化特征工程的领域数据类型、特征数据以及关联关系
S3
通过智能特征工程计算模块支持智能特征工程的计算，并生成最后的特征向量

图6

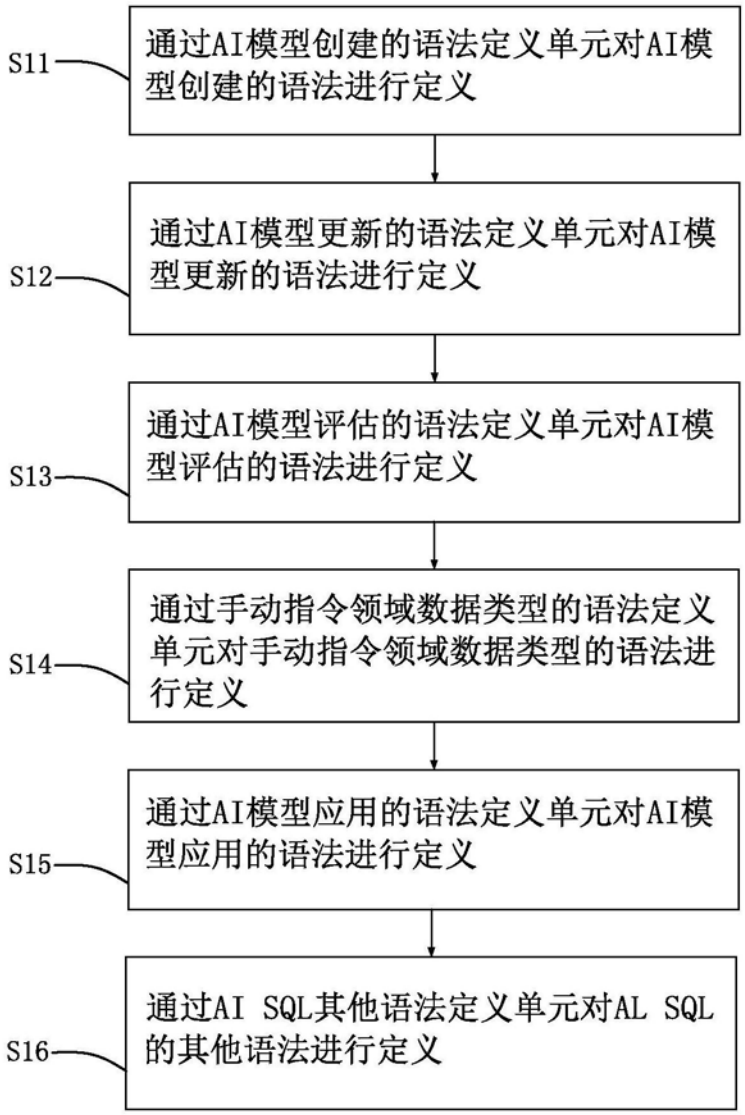

S11
通过AI模型创建的语法定义单元对AI模型创建的语法进行定义
S12
通过AI模型更新的语法定义单元对AI模型更新的语法进行定义
S13
通过AI模型评估的语法定义单元对AI模型评估的语法进行定义
S14
通过手动指令领域数据类型的语法定义单元对手动指令领域数据类型的语法进行定义
S15
通过AI模型应用的语法定义单元对AI模型应用的语法进行定义
S16
通过AI SQL其他语法定义单元对AL SQL的其他语法进行定义

图7

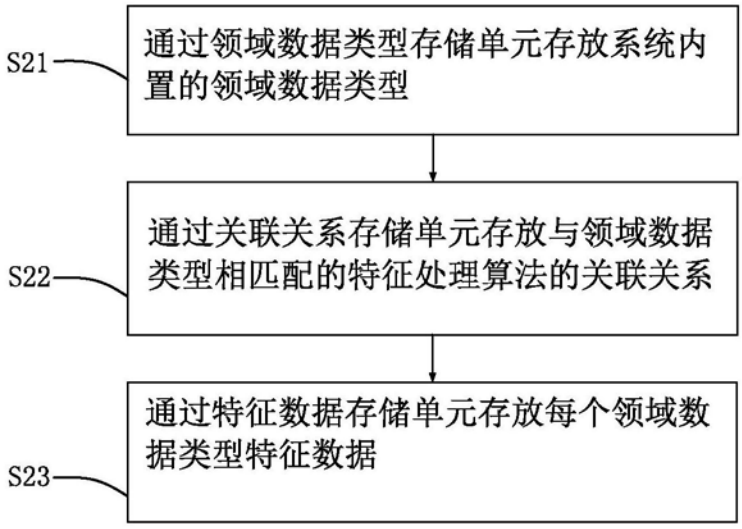

S21
通过领域数据类型存储单元存放系统内置的领域数据类型
S22
通过关联关系存储单元存放与领域数据类型相匹配的特征处理算法的关联关系
S23
通过特征数据存储单元存放每个领域数据类型特征数据

图8

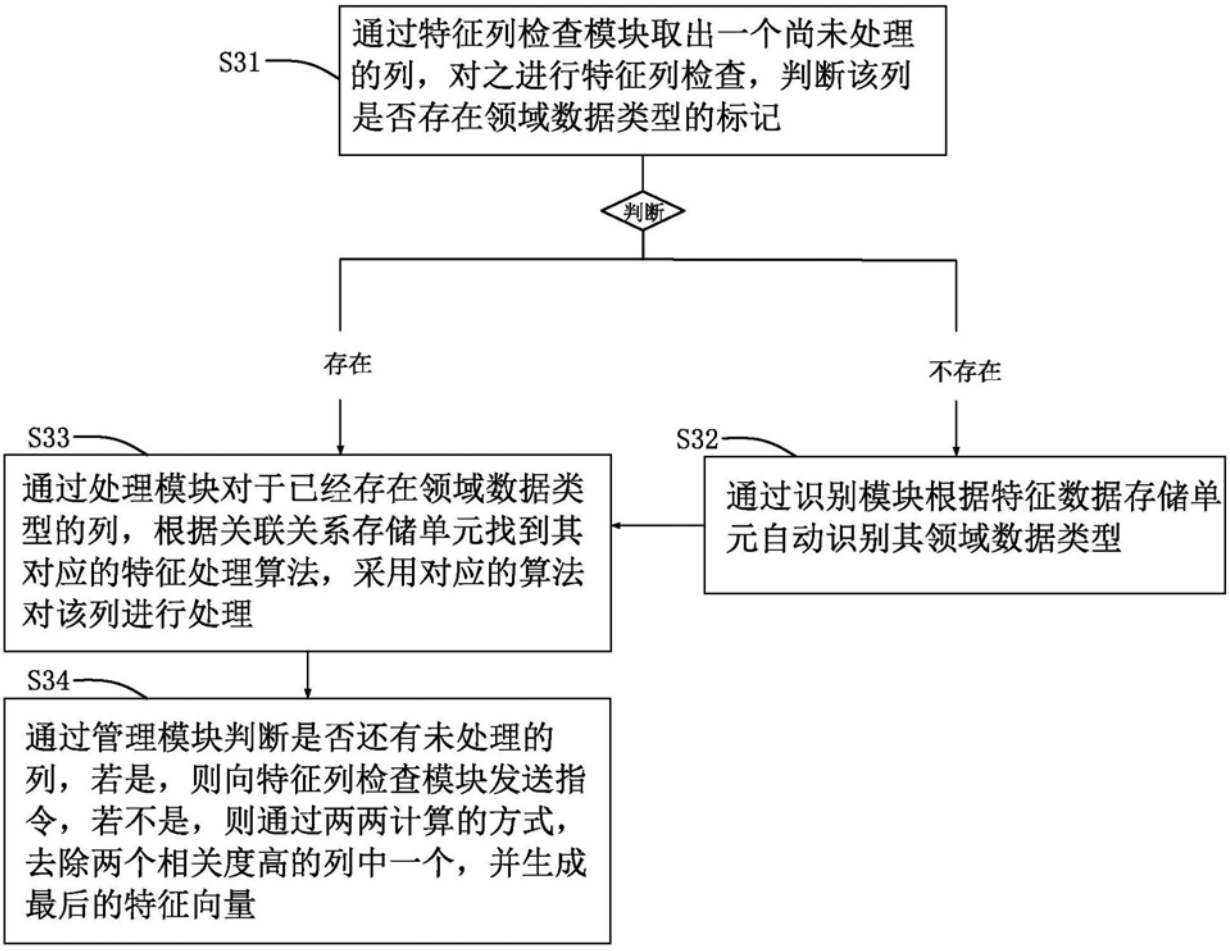

S31
通过特征列检查模块取出一个尚未处理的列，对之进行特征列检查，判断该列是否存在领域数据类型的标记
判断
存在
不存在
S33
通过处理模块对于已经存在领域数据类型的列，根据关联关系存储单元找到其对应的特征处理算法，采用对应的算法对该列进行处理
S32
通过识别模块根据特征数据存储单元自动识别其领域数据类型
S34
通过管理模块判断是否还有未处理的列，若是，则向特征列检查模块发送指令，若不是，则通过两两计算的方式，去除两个相关度高的列中一个，并生成最后的特征向量

图9