[19] Patents Registry
The Hong Kong Special Administrative Region
香港特別行政區
專利註冊處

[11] 40008317 B
CN 110362303 B

[12] **STANDARD PATENT (R) SPECIFICATION**
轉錄標準專利說明書

[21] Application no. 申請編號
19131785.8

[22] Date of filing 提交日期
01.11.2019

[51] Int. Cl.
G06F 8/35 (2018.01)

[54] A DATA HEURISTIC APPROACH AND SYSTEM
數據探索方法和系統

[43] Date of publication of application 申請發表日期
12.06.2020

[45] Date of publication of grant of patent 批予專利的發表日期
12.03.2021

CN Application no. & date 中國專利申請編號及日期
CN 201910636826.X 15.07.2019

CN Publication no. & date 中國專利申請發表編號及日期
CN 110362303 22.10.2019

Date of grant in designated patent office 指定專利當局批予專利日期
25.08.2020

[73] Proprietor 專利所有人
深圳市宇数科技有限公司
中國
518000 广东省深圳市福田区东园路向东围 63 号
501 室

[72] Inventor 發明人
林宇

[74] Agent and / or address for service 代理人及/或送達地址
HONGKONG TRUER IP LIMITED
香港
九龙旺角弥敦道 610 号
荷李活商业中心 1318-20 室

(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 110362303 B
(45)授权公告日 2020.08.25

(21)申请号 201910636826.X

(22)申请日 2019.07.15

(65)同一申请的已公布的文献号
申请公布号 CN 110362303 A

(43)申请公布日 2019.10.22

(73)专利权人 深圳市宇数科技有限公司
地址 518000 广东省深圳市福田区东园路向东围63号501室

(72)发明人 林宇

(74)专利代理机构 北京酷爱智慧知识产权代理有限公司 11514
代理人 袁克来

(51)Int.Cl.
G06F 8/35(2018.01)

(56)对比文件
CN 107578161 A,2018.01.12
CN 101110089 A,2008.01.23
CN 104123375 A,2014.10.29
CN 106605222 A,2017.04.26
US 10311442 B1,2019.06.04

审查员 吴阳

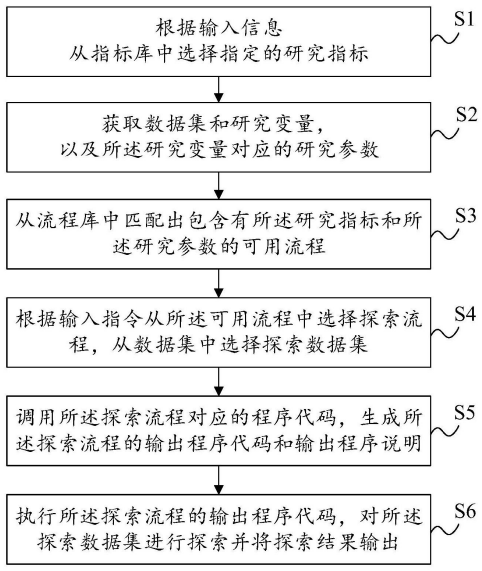
权利要求书3页 说明书12页 附图3页

(54)发明名称

数据探索方法和系统

(57)摘要

本申请涉及一种数据探索方法和系统，所述方法包括：根据输入信息从指标库中选择指定的研究指标；获取数据集、研究变量，以及所述研究变量对应的研究参数；从流程库中匹配出包含有所述研究指标和所述研究参数的可用流程；根据输入指令从所述可用流程中选择探索流程，从数据集中选择探索数据集；生成所述探索流程的输出程序代码和输出程序说明；执行所述探索流程的输出程序代码，对所述探索数据集进行探索并将探索结果输出。本申请的方案只需要提交待处理的数据集，并给出探索过程的研究指标、研究参数等，就能自动输出程序代码并调用输出程序代码对数据进行探索处理；极大免去研究人员编写代码的劳动，提高数据探索的效率。



CN 110362303 B

1.一种数据探索方法,其特征在于,包括:

根据输入信息从指标库中选择指定的研究指标;所述研究指标为用户所选择的预存储研究指标,所述预存储研究指标为统计方法得到的结果;

获取数据集、研究变量,以及所述研究变量对应的研究参数;

从流程库中匹配出包含有所述研究指标和所述研究参数的可用流程;

根据输入指令从所述可用流程中选择探索流程,从数据集中选择探索数据集;

调用所述探索流程对应的程序代码,生成所述探索流程的输出程序代码和输出程序说明;所述输出程序说明用于判断输出程序代码是否按照输出程序说明编写;

执行所述探索流程的输出程序代码,对所述探索数据集进行探索并将探索结果输出。

2.根据权利要求1所述的方法,其特征在于,所述生成所述探索流程的输出程序代码和输出程序说明,包括:

所述探索流程对应的程序代码调用所述探索流程中的方法或图像,调用所述方法对应的程序代码、所述图像对应的程序代码,生成输出程序代码和输出程序说明。

3.根据权利要求1所述的方法,其特征在于,还包括:

根据输入信息从设计库中选择指定的研究设计;

从包含有所述研究指标和所述研究参数的可用流程中,进一步匹配出包含有对应研究设计的可用流程。

4.根据权利要求1所述的方法,其特征在于,还包括:

根据输入指令从已选择的探索流程和数据集中选择一个或多个作为对比流程和对比数据集;

调用所述对比流程对应的程序代码,生成所述对比流程的输出程序代码和输出程序说明;

分别合并所述探索流程的输出程序代码和所述对比流程的输出程序代码、所述探索流程的输出程序说明和所述对比流程的输出程序说明,生成全部的输出程序代码和输出程序说明;

执行合并后的输出程序代码,对所述数据集进行探索;

将所述探索流程的探索结果与所述对比流程的探索结果进行比较并将比较结果输出。

5.根据权利要求4所述的方法,其特征在于,所述将所述探索流程的探索结果与所述对比流程的探索结果进行比较并将比较结果输出,包括:

获取探索流程的各个研究指标的探索结果;

获取对比流程的各个研究指标的探索结果;

将各个研究指标的探索流程的探索结果与对比流程的探索结果进行比较;

将不同的探索结果按照不同的格式进行显示输出。

6.根据权利要求1-5任一项所述的方法,其特征在于,所述获取数据集、研究变量,以及所述研究变量对应的研究参数,包括:

获取用户输入的数据集和研究变量;

获取用户指定的研究变量所对应的变量库中的变量;

获取用户指定的研究变量或所述用户指定的研究变量所对应的变量库中的变量所对应的参数信息;

根据参数信息从参数库中确定对应的研究参数。

7.一种数据探索系统,其特征在于,包括:

数据库,用于存储预设的指标库、参数库和流程库;所述指标库存储有多种不同的研究指标;所述参数库存储有多种不同的研究参数;所述流程库存储有多种不同的研究流程,以及每一种所述研究流程所对应的程序代码;所述研究指标为用户所选择的预存储研究指标,所述预存储研究指标为统计方法得到的结果;

指标选择模块,用于根据输入信息从指标库中选择指定的研究指标;

数据获取模块,用于获取数据集、研究变量,以及所述研究变量对应的研究参数;

流程匹配模块,用于从流程库中匹配出包含有所述研究指标和所述研究参数的可用流程;

流程选择模块,用于根据输入指令从所述可用流程中选择探索流程,从数据集中选择探索数据集;

程序输出模块,用于调用所述探索流程对应的程序代码,生成所述探索流程的输出程序代码和输出程序说明;所述输出程序说明用于判断输出程序代码是否按照输出程序说明编写;

探索输出模块,用于执行所述探索流程的输出程序代码,对所述探索数据集进行探索并将探索结果输出。

8.根据权利要求7所述的系统,其特征在于:

所述数据库还用于存储预设的变量库,所述变量库存储有多种不同的研究变量;

相应地,数据获取模块还用于:

获取所述研究变量对应的变量库中的变量;

所述数据库还用于存储预设的设计库,所述设计库存储有多种不同的研究设计;相应地,所述系统还包括:

设计选择模块,用于根据输入信息从设计库中选择指定的研究设计;

所述流程匹配模块还用于:

从包含有所述研究指标和所述研究参数的可用流程中,进一步匹配出包含有对应研究设计的可用流程。

9.根据权利要求7或8所述的系统,其特征在于,所述流程选择模块还用于:

根据输入指令从已选择的探索流程和数据集中选择一个或多个作为对比流程和对比数据集;

所述程序输出模块还用于:

调用所述对比流程对应的程序代码,生成所述对比流程的输出程序代码和输出程序说明;

分别合并所述探索流程的输出程序代码和所述对比流程的输出程序代码、所述探索流程的输出程序说明和所述对比流程的输出程序说明,生成全部的输出程序代码和输出程序说明;

所述探索输出模块还用于:

执行合并后的输出程序代码,对所述数据集进行探索;

将所述探索流程的探索结果与所述对比流程的探索结果进行比较并将比较结果输出。

10.一种计算设备,其特征在于,所述计算设备包括:处理器和存储器;

所述存储器用于存储计算机程序指令;

所述计算设备运行时,所述处理器执行所述存储器中的计算机程序指令,以执行权利要求1至6中任一项所述方法的操作步骤。

数据探索方法和系统

技术领域

[0001] 本申请涉及数据处理技术领域，具体涉及一种数据探索方法和系统。

背景技术

[0002] 随着现代科学技术的进步，信息技术的快速发展和应用，使得全行业信息化的程度全面提升，整个社会的数据正在以前所未有的速度快速增长，呈现出数量大、种类多、更新快的特点，逐渐成为各行各业的重要生产要素之一。云计算、物联网、移动终端及可穿戴设备高度发达与融合，使得这种发展趋势变得越来越快。

[0003] 与之对应的是，数据分析的难度和复杂度在不断增加。丰富的数据量蕴含着大量的宝贵信息，但这样的数据需要复杂的统计分析，才能从中提取到有意义的结果。它们不仅促进了统计学，尤其是多元统计分析的应用，如聚类和判别分析、主成分分析、因子分析等方法得到了广泛的应用。同时，它们也带来了机器学习、深度学习等人工智能(AI)技术的快速发展和应用，如各类深度学习网络框架：无监督预训练网络、卷积神经网络、循环神经网络、递归神经网络等。大量的多元统计方法、拥有大量参数和层的神经网络等方法的使用，使得统计方法变得越来越复杂和多样化，各类统计方法联合使用、相互比较的需求在不断增加。

[0004] 各类带有编程功能的相关程序，如Python、JAVA、R、Matlab、SAS、SQL、C、Perl等，逐渐成为科研与日常统计分析的工具，更精确的计算方法也得以大规模的应用。新的带有编程功能的相关程序，如面向科学计算的高性能动态语言Julia等，陆续出现。这些工具大量运用于科研和生产，让科研和生产线上的工作人员摆脱了手工计算，而且还可以快速发现过程异常，促进了科研和产品质量的提升。但是，这也导致工作人员需要花费大量的时间和精力来学习和提升，导致学习和应用成本大大的提高。

发明内容

[0005] 为至少在一定程度上克服相关技术中存在的问题，本申请提供一种数据探索方法和系统。

[0006] 根据本申请实施例的第一方面，提供一种数据探索方法，包括：

[0007] 根据输入信息从指标库中选择指定的研究指标；

[0008] 获取数据集、研究变量，以及所述研究变量对应的研究参数；

[0009] 从流程库中匹配出包含有所述研究指标和所述研究参数的可用流程；

[0010] 根据输入指令从所述可用流程中选择探索流程，从数据集中选择探索数据集；

[0011] 调用所述探索流程对应的程序代码，生成所述探索流程的输出程序代码和输出程序说明；

[0012] 执行所述探索流程的输出程序代码，对所述探索数据集进行探索并将探索结果输出。

[0013] 进一步地，所述生成所述探索流程的输出程序代码和输出程序说明，包括：

[0014] 所述探索流程对应的程序代码调用所述探索流程中的方法或图像，调用所述方法对应的程序代码、所述图像对应的程序代码，生成输出程序代码和输出程序说明。

[0015] 进一步地，所述方法还包括：

[0016] 根据输入信息从设计库中选择指定的研究设计；

[0017] 从包含有所述研究指标和所述研究参数的可用流程中，进一步匹配出包含有对应研究设计的可用流程。

[0018] 进一步地，所述方法还包括：

[0019] 根据输入指令从已选择的探索流程和数据集中选择一个或多个作为对比流程和对比数据集；

[0020] 调用所述对比流程对应的程序代码，生成所述对比流程的输出程序代码和输出程序说明；

[0021] 分别合并所述探索流程的输出程序代码和所述对比流程的输出程序代码、所述探索流程的输出程序说明和所述对比流程的输出程序说明，生成全部的输出程序代码和输出程序说明；

[0022] 执行合并后的输出程序代码，对所述数据集进行探索；

[0023] 将所述探索流程的探索结果与所述对比流程的探索结果进行比较并将比较结果输出。

[0024] 进一步地，所述将所述探索流程的探索结果与所述对比流程的探索结果进行比较并将比较结果输出，包括：

[0025] 获取探索流程的各个研究指标的探索结果；

[0026] 获取对比流程的各个研究指标的探索结果；

[0027] 将各个研究指标的探索流程的探索结果与对比流程的探索结果进行比较；

[0028] 将不同的探索结果按照不同的格式进行显示输出。

[0029] 进一步地，所述获取数据集、研究变量，以及所述研究变量对应的研究参数，包括：

[0030] 获取用户输入的数据集和研究变量；

[0031] 获取用户指定的研究变量所对应的变量库中的变量；

[0032] 获取用户指定的研究变量或用户指定的研究变量所对应的变量库中的变量所对应的参数信息；

[0033] 根据参数信息从参数库中确定对应的研究参数。

[0034] 进一步地，所述研究变量还包括：

[0035] 获取用户指定的研究变量所对应的变量库中的变量，使变量具有唯一的名称、实现变量快速重命名，以用于后续调用和整合多个数据集、生成输出程序代码和对比结果。

[0036] 根据本申请实施例的第二方面，提供一种数据探索系统，包括：

[0037] 数据库，用于存储预设的指标库、参数库和流程库；所述指标库存储有多种不同的研究指标；所述参数库存储有多种不同的研究参数；所述流程库存储有多种不同的研究流程，以及每一种所述研究流程所对应的程序代码；

[0038] 指标选择模块，用于根据输入信息从指标库中选择指定的研究指标；

[0039] 数据获取模块，用于获取数据集、研究变量，以及所述研究变量对应的研究参数；

[0040] 流程匹配模块，用于从流程库中匹配出包含有所述研究指标和所述研究参数的可

用流程；

[0041] 流程选择模块，用于根据输入指令从所述可用流程中选择探索流程，从数据集中选择探索数据集；

[0042] 程序输出模块，用于调用所述探索流程对应的程序代码，生成所述探索流程的输出程序代码和输出程序说明；

[0043] 探索输出模块，用于执行所述探索流程的输出程序代码，对所述探索数据集进行探索并将探索结果输出。

[0044] 进一步地，数据库还可用于存储预设的方法库、图像库。所述方法库存储有多种不同的统计方法；所述图像库存储有多种不同的图像；所述方法库和图像库分别存储有每一种所述统计方法和图像所对应的程序代码。

[0045] 进一步地，所述数据库还用于存储预设的变量库，所述变量库存储有多种不同的研究变量；

[0046] 相应地，数据获取模块还用于：

[0047] 获取所述研究变量对应的变量库中的变量；

[0048] 所述数据库还用于存储预设的设计库，所述设计库存储有多种不同的研究设计；相应地，所述系统还包括：

[0049] 设计选择模块，用于根据输入信息从设计库中选择指定的研究设计；

[0050] 所述流程匹配模块还用于：

[0051] 从包含有所述研究指标和所述研究参数的可用流程中，进一步匹配出包含有对应研究设计的可用流程。

[0052] 进一步地，所述流程选择模块还用于：

[0053] 根据输入指令从已选择的探索流程和数据集中选择一个或多个作为对比流程和对比数据集；

[0054] 所述程序输出模块还用于：

[0055] 调用所述对比流程对应的程序代码，生成所述对比流程的输出程序代码和输出程序说明；

[0056] 分别合并所述探索流程的输出程序代码和所述对比流程的输出程序代码、所述探索流程的输出程序说明和所述对比流程的输出程序说明，生成全部的输出程序代码和输出程序说明；

[0057] 所述探索输出模块还用于：

[0058] 执行合并后的输出程序代码，对所述数据集进行探索；

[0059] 将所述探索流程的探索结果与所述对比流程的探索结果进行比较并将比较结果输出。

[0060] 根据本申请实施例的第三方面，提供一种计算设备，所述计算设备包括：处理器和存储器；

[0061] 所述存储器用于存储计算机程序指令；

[0062] 所述计算设备运行时，所述处理器执行所述存储器中的计算机程序指令，以执行如上所述的任意一种方法的操作步骤。

[0063] 本申请的实施例提供的技术方案可以包括以下有益效果：

[0064] 本申请的方案只需要提交待处理的数据集和研究变量，并给出探索过程的研究指标、研究参数等，就能自动从预先构建的流程库中匹配出所需的探索流程，并调用预先存储的程序代码对数据进行探索程序代码生成和处理；本方案能够极大免去研究人员编写代码的劳动，提高数据探索的效率；并且探索过程完全标准化、可重复，还能方便地选取多种数据集和探索方法进行组合和对比结果。

[0065] 应当理解的是，以上的一般描述和后文的细节描述仅是示例性和解释性的，并不能限制本申请。

附图说明

[0066] 此处的附图被并入说明书中并构成本说明书的一部分，示出了符合本申请的实施例，并与说明书一起用于解释本申请的原理。

[0067] 图1是根据一示例性实施例示出的一种数据探索方法的流程图。

[0068] 图2是根据另一示例性实施例示出的一种数据探索方法的流程图。

[0069] 图3是根据一示例性实施例示出的一种数据探索系统的结构框图。

具体实施方式

[0070] 这里将详细地对示例性实施例进行说明，其示例表示在附图中。下面的描述涉及附图时，除非另有表示，不同附图中的相同数字表示相同或相似的要素。以下示例性实施例中所描述的实施方式并不代表与本申请相一致的所有实施方式。相反，它们仅是与如所附权利要求书中所详述的、本申请的一些方面相一致的方法和系统的例子。

[0071] 为至少在一定程度上克服相关技术中存在的问题，即数据量大、更新快，分析方法复杂多样且需要联合使用、相互比较，编程工具多样且学习和应用成本高的问题，本发明的目的之一在于，本申请的方案只需要提交待处理的数据集和研究变量，并给出探索过程的研究指标、研究参数等，就能自动从预先构建的流程库中匹配出所需的探索流程，并调用预先存储的程序代码对数据进行探索程序代码生成和处理。(1)本方案能够极大地减少研究人员编写程序代码的劳动，实现自动化或半自动化编程(获取数据部分可提交已编写好的程序代码)，提高数据探索的效率；(2)并且探索过程完全标准化、可重复，探索过程使用的方法、图像均可按照特定顺序进行，保证按要求准确实现探索过程；(3)同时还能方便地选取多种数据集、探索流程进行组合和对比，得到多个结果及对比结果；(4)能方便实现多种数据集整合、标准化，方便后续探索利用。

[0072] 图1是根据一示例性实施例示出的一种数据探索方法的流程图。该方法包括以下步骤：

[0073] 步骤S1：根据输入信息从指标库中选择指定的研究指标；

[0074] 步骤S2：获取数据集、研究变量，以及所述研究变量对应的研究参数；

[0075] 步骤S3：从流程库中匹配出包含有所述研究指标和所述研究参数的可用流程；

[0076] 步骤S4：根据输入指令从所述可用流程中选择探索流程，从数据集中选择探索数据集；

[0077] 步骤S5：调用所述探索流程对应的程序代码，生成所述探索流程的输出程序代码和输出程序说明；

[0078] 步骤S6:执行所述探索流程的输出程序代码,对所述探索数据集进行探索并将探索结果输出。

[0079] 本申请的方案只需要提交待处理的数据集和研究变量,并给出探索过程的研究指标、研究参数等,就能自动从预先构建的流程库中匹配出所需的探索流程,并调用预先存储的程序代码对数据进行探索程序代码的生成和处理;本方案能够极大免去研究人员编写代码的劳动,提高数据探索的效率;并且探索过程完全标准化、可重复,还能方便地选取多个数据集和多种探索方法进行组合。

[0080] 一些实施例中,所述生成所述探索流程的输出程序代码和输出程序说明,包括:

[0081] 所述探索流程对应的程序代码调用所述探索流程中的方法或图像,调用所述方法对应的程序代码、所述图像对应的程序代码,生成输出程序代码和输出程序说明。

[0082] 参照图2,一些实施例中,所述方法还包括:

[0083] 根据输入信息从设计库中选择指定的研究设计;

[0084] 从包含有所述研究指标和所述研究参数的可用流程中,进一步匹配出包含有对应研究设计的可用流程。

[0085] 一些实施例中,所述方法还包括:

[0086] 根据输入指令从已选择的探索流程和数据集中选择一个或多个作为对比流程和对比数据集;

[0087] 调用所述对比流程对应的程序代码,生成所述对比流程的输出程序代码和输出程序说明;

[0088] 分别合并所述探索流程的输出程序代码和所述对比流程的输出程序代码、所述探索流程的输出程序说明和所述对比流程的输出程序说明,生成全部的输出程序代码和输出程序说明;

[0089] 执行合并后的输出程序代码,对所述数据集进行探索;

[0090] 将所述探索流程的探索结果与所述对比流程的探索结果进行比较并将比较结果输出。

[0091] 一些实施例中,所述将所述探索流程的探索结果与所述对比流程的探索结果进行比较并将比较结果输出,包括:

[0092] 获取探索流程的各个研究指标的探索结果;

[0093] 获取对比流程的各个研究指标的探索结果;

[0094] 将各个研究指标的探索流程的探索结果与对比流程的探索结果进行比较;

[0095] 将不同的探索结果按照不同的格式进行显示输出。

[0096] 一些实施例中,所述获取数据集、研究变量,以及所述研究变量对应的研究参数,包括:

[0097] 获取用户输入的数据集和研究变量;

[0098] 获取用户指定的研究变量所对应的变量库中的变量;

[0099] 获取用户指定的研究变量或用户指定的研究变量所对应的变量库中的变量所对应的参数信息;

[0100] 根据参数信息从参数库中确定对应的研究参数。

[0101] 一些实施例中,所述获取数据集、研究变量,以及所述研究变量对应的研究参数,

包括：

[0102] 获取用户指定的研究变量所对应的变量库中的变量。

[0103] 如此能够使变量具有唯一的名称、实现变量快速重命名，以用于后续调用和整合多个数据集、生成输出程序代码和对比结果。

[0104] 为进一步阐明本申请的方案，对该方法各个步骤的细节做进一步说明。

[0105] 1.选择研究指标。获取用户所选择的预存储研究指标，得到一个或多个研究指标，记为r_i，i＝1，2，.....。

[0106] 进一步地，预存储研究指标与预存储流程呈现为一对多的关系，即任意一个预存储研究指标可以存在于多个预存储流程中。

[0107] 进一步地，预存储研究指标为统计方法得到的结果，包括但不限于，均数、标准差、AIC、BIC、各类回归模型系数和P值、各类比较的P值、曲线下面积、灵敏度、特异度等，这些统计方法得到的结果用于生成研究结果。

[0108] 2.选择研究设计。获取用户所选择的预存储研究设计，得到一个或多个研究设计，记为e_i，i＝1，2，.....。

[0109] 进一步地，预存储研究设计与预存储流程呈现为一对多的关系，即任意一个预存储研究设计可以存在于多个预存储流程中。

[0110] 进一步地，研究设计为统计学中的研究设计方案，包括但不限于，完全随机设计、系统分组设计、裂区设计、序贯试验设计、重复测量设计，用于匹配对应的预存储流程。

[0111] 3.获取数据集和研究变量，选择研究参数。获取用户所输入的数据集、研究变量，并选择研究变量对应的预存储参数，(1)得到一个或多个数据集，记为d_i，i＝1，2，.....；(2)得到一个或多个数据集中的一个或多个研究变量，记为v_ij，i＝1，2，.....；j＝1，2，.....；(3)得到一个或多个数据集中的一个或多个预存储参数，记为p_ik，i＝1，2，.....；k＝1，2，.....。(4)得到一个或多个数据集中的研究变量对应的预存储变量，记为v2ij，i＝1，2，.....；j＝1，2，.....。

[0112] 进一步地，预存储研究变量包含唯一名称、别名、序号、类型(连续数值、分类数值、等级数值)、分类信息，以用于后续调用和整合多个数据集、对比结果，使一个或多个数据集中相同的变量有唯一名称、实现快速重命名，避免因为变量名不规范导致程序代码无法运行。

[0113] 进一步地，多个数据集中共有的研究变量对应的预存储变量，可通过程序代码形式合并生成新的数据集，供后续探索和对比使用。

[0114] 进一步地，预存储参数将得到研究变量的任意组合，用于后续流程程序代码运行，即为流程程序代码的运行时需要的参数。也用于匹配得到包含这些参数的预存储流程。

[0115] 进一步地，任意一个研究变量可以对应一个预存储研究变量，用于生成包含有预存储研究变量的数据集，用于后续分析。

[0116] 进一步地，任意一个预存储参数可以包含一个或多个研究变量。

[0117] 进一步地，预存储参数与预存储流程呈现为一对多的关系，即任意一个预存储参数可以存在于多个预存储流程中。

[0118] 进一步地，数据集可以以文件形式、数据库形式、或者程序代码形式输入。

[0119] 进一步地，程序代码可以包含一种或多种编程语言，也可以由一种或多种编程语

言组合而成。

[0120] 进一步地,程序代码可以单独运行,也可以按特定顺序组合运行。

[0121] 4.匹配预存储流程。在选择研究指标、参数后,自动匹配包含有对应的研究指标和参数的预存储流程得到一个或多个预存储流程记为f_i,i=1,2,.....。

[0122] 进一步地,选择研究设计后,进一步匹配包含有对应研究设计的预存储流程。

[0123] 进一步地,预存储流程包括所述流程的信息、节点、路径、方法、图像、参数、指标和程序代码。

[0124] 信息是指对统计方法的原理和组合等说明、图像说明、数学计算公式、阈值判断公式等信息,用于输出程序说明。

[0125] 节点是指使用的统计方法和图像,包括方法的名称,用于得到研究指标结果;又或者是阈值判断;图像的名称,用于显示研究指标结果。路径是指从统计方法、图像节点或者阈值判断节点到下一个统计方法、图像节点或者阈值判断节点。程序代码是指用于完成统计方法节点的数学计算公式、图像节点的图像制作、阈值判断节点的判断公式,以及按照路径执行的程序代码,用于执行得到输出程序代码和输出程序说明。

[0126] 输出程序代码和输出程序说明,用于判断程序代码是否按照输出程序说明编写和处理。

[0127] 5.选择预存储流程和数据集。进一步地,可以选择一个或多个上述匹配后的预存储流程,记为f2_i,i=1,2,.....。数据集为输入得到一个或多个数据集,记为fd_i,i=1,2,.....。

[0128] 6.选择对比流程。进一步地,可以在一个或多个已选择的预存储流程中,选择一个或多个预存储流程作为对比流程,记为c_i,i=1,2,.....。

[0129] 7.生成输出程序代码和输出程序说明,获取预存储流程的操作、方法、图像和流程程序代码,生成并保存相应的输出程序代码和输出程序说明。

[0130] 进一步地,预存储流程包括信息、节点、路径、方法、图像、参数、研究指标、研究设计和流程程序代码,节点和路径组成操作,方法包括预存储方法,参数包括预存储参数,研究指标包含预存储研究指标,研究设计包含预存储研究设计,流程程序代码用于执行预存储流程。

[0131] 进一步地,预存储方法包括统计方法和方法程序代码,方法程序代码用于执行预存储方法。

[0132] 预存储统计方法是指统计方法的原理说明(文章、图表、视频等人可以接受的方式)、数学计算公式、统计量(研究指标),用于输出程序说明。

[0133] 方法程序代码是指用于完成统计方法,得到研究指标的数学计算公式的程序代码,用于被流程程序代码调用,生成输出程序代码,在输出程序代码执行后得到研究指标。

[0134] 预存储图像是指各类用显示研究指标的图像的原理说明(文章、图表、视频等人可以接受的方式)、数学计算公式、统计量(研究指标),用于输出程序说明。

[0135] 图像程序代码是指用于完成图像绘制,得到图像的程序代码,用于被流程程序代码调用,生成输出程序代码,在输出程序代码执行后得到图像。

[0136] 流程程序代码调用方法程序代码或图像程序代码,生成输出程序代码,得到用于处理数据的全部程序代码,即输出程序代码。

[0137] 进一步地，程序代码可以包含一种或多种编程语言，也可以由一种或多种编程语言组合而成。

[0138] 进一步地，程序代码可以单独运行，也可以按特定顺序组合运行。

[0139] 进一步地，生成输出程序说明，包括预存储流程的节点信息、路径信息、统计方法信息、图像信息，即得到用于处理数据的全部操作信息、统计方法信息，图像信息，可以以文字、图像等形式显示，用于判断输出程序代码是否按照输出程序说明编写。

[0140] 进一步地，流程研究参数用于与用户选择的预存储研究参数匹配。

[0141] 进一步地，流程研究指标用于与用户选择的预存储研究研究指标匹配。

[0142] 进一步地，流程研究设计用于与用户选择的预存储研究研究设计匹配。

[0143] 8.运行程序代码，运行输出程序代码，得到并保存相应的运行结果。

[0144] 9.显示运行结果，显示预存储流程、相应的输出程序代码、输出程序说明和相应的运行结果。

[0145] 进一步地，在运行结果中，各个研究指标结果自动与对比流程的研究指标结果进行比较，与对比流程研究指标结果不一致的，按照不同的格式显示，如通过改变颜色、改变字体，或者以文字、图像形式显示说明。

[0146] 进一步地，多个参照对比流程的，按特定顺序逐一选择其中一个对比流程，得到其他流程与选定的对比流程研究指标的结果差别。

[0147] 下面结合具体的应用场景，对本申请的方案进行拓展说明。

[0148] 场景一:数据预测

[0149] 1.选择研究指标。在数据预测、构建预测模型时，选择用于评价模型的研究指标。例如，在建立二分类预测模型时，选择预存储研究指标中的"AIC"、"曲线下面积"、"灵敏度"、"特异度"等研究指标来评价模型效果。

[0150] 2.获取数据集和研究变量，选择研究参数。获取用户所输入的数据集和研究变量，以及研究变量对应的预存储参数。例如，用户输入了数据集，同时输入了大量的研究变量，如"是否发生癌症"、"临床特征"、"影像学特征"、"基因特征"等几百个研究变量。用户将研究变量"是否发生癌症"对应于预存储研究参数"二分类因变量"，"临床特征"、"影像学特征"、"基因特征"等几百个研究变量对应于预存储参数"自变量"。

[0151] 3.匹配预存储流程。在选择研究指标和研究参数后，匹配预存储流程中包含有对应的研究指标和研究参数的流程，得到一个或多个预存储流程。例如，按照上述研究指标"曲线下面积"、"灵敏度"、"特异度"，参数"二分类因变量"和参数"自变量"，得到匹配的预存储流程有"Logistic回归模型"，"LASSO Logistic回归模型"，"随机森林模型"，"神经网络模型"等。

[0152] 4.选择预存储流程和数据集。用户选择其中的一个或多个，如选择"Logistic回归模型"，同时选择使用的数据集，如用户所输入的数据集。

[0153] 5.生成输出程序代码和输出程序说明，获取预存储流程的操作、方法、图像和流程程序代码，生成并保存相应的输出程序代码和输出程序说明。

[0154] 6.运行程序代码，运行输出程序代码，得到并保存相应的运行结果。

[0155] 7.显示运行结果，显示Logistic回归模型流程、相应的输出程序代码、输出程序说明和相应的运行结果。

[0156] 场景二:多个数据、多个流程、多个结果比较(也可用于多次修改数据后结果比较)

[0157] 1.选择研究指标。在数据预测、构建预测模型时,选择用于评价模型的研究指标。例如,在建立二分类预测模型时,选择预存储研究指标中的"AIC"、"曲线下面积"、"灵敏度"、"特异度"等研究指标来评价模型效果。

[0158] 2.获取数据集和研究变量,选择研究参数。获取用户所输入的数据集和研究变量,以及研究变量对应的预存储参数。例如,用户输入了数据集,同时输入了大量的研究变量,如"是否发生癌症"、"临床特征"、"影像学特征"、"基因特征"等几百个研究变量。用户将研究变量"是否发生癌症"对应于预存储研究参数"二分类因变量","临床特征"、"影像学特征"、"基因特征"等几百个研究变量对应于预存储参数"自变量"。

[0159] 进一步的,选择按照不同条件下的数据集。例如,全部性别数据集,男性数据集,女性数据集。又或者是第一次提供的数据集,第二次提供的数据集,第三次提供的数据集。

[0160] 3.匹配预存储流程。在选择研究指标和研究参数后,匹配预存储流程中包含有对应的研究指标和研究参数的流程,得到一个或多个预存储流程。例如,按照上述研究指标"曲线下面积"、"灵敏度"、"特异度",研究参数"二分类因变量"和研究参数"自变量",得到匹配的预存储流程有"Logistic回归模型","LASSO Logistic回归模型","随机森林模型","神经网络模型"等。

[0161] 4.选择预存储流程和数据集。用户选择其中的多个,如选择"Logistic回归模型","Logistic回归模型","LASSO Logistic回归模型","随机森林模型","神经网络模型"。数据集可以选择全部性别数据集、男性数据集,女性数据集中的一个或多个。

[0162] 5.选择对比流程。用户选择已选择的预存储流程(包含数据集)中一个或多个,如选择"Logistic回归模型"(男性数据集),"随机森林模型"(女性数据集)作为对比流程。

[0163] 6.生成输出程序代码和输出程序说明,获取预存储流程的操作、方法、图像和流程程序代码,生成并保存相应的输出程序代码和输出程序说明。

[0164] 7.运行程序代码,运行输出程序代码,得到并保存相应的运行结果。

[0165] 8.显示运行结果,显示"Logistic回归模型"(全部性别数据集),"Logistic回归模型"(男性数据集),"LASSO Logistic回归模型"(男性数据集),"随机森林模型"(女性数据集),"神经网络模型"(男性数据集)流程、相应的输出程序代码、输出程序代码说明和相应的运行结果。

[0166] 进一步地,运行结果包含多个数据,多个流程的研究指标,如以第一个流程Logistic回归模型(全部性别数据集)的"AIC"、"曲线下面积"、"灵敏度"、"特异度"作为参照,其他流程的"AIC"、"曲线下面积"、"灵敏度"、"特异度"与第一个的结果相同给予相同的样式,如颜色或者字体等显示。否则以不同的样式显示。也可以以文字或者图片方式说明显示。

[0167] 以第二个流程随机森林回归模型(女性数据集)的"AIC"、"曲线下面积"、"灵敏度"、"特异度"作为参照,其他流程的"AIC"、"曲线下面积"、"灵敏度"、"特异度"与第二个的结果相同给予相同的样式,如颜色或者字体等显示。否则以不同的样式显示。也可以以文字或者图片方式说明显示。

[0168] 场景三:研究设计结果模拟

[0169] 1.选择研究指标。在预测模型中,选择用于评价模型的研究指标。例如,研究设计

结果在特定条件下的一类错误概率和检验效能。此时可以选择可以研究指标"一类错误概率"、"检验效能"。

[0170] 2.选择研究设计。在预测模型中,选择用于模拟的研究设计。例如,选择"模拟正态分布两组均数比较"、"模拟二项分布两组均数比较"、"模拟贯序实验两组均数比较"。

[0171] 3.获取数据集和研究变量,选择研究参数。获取用户所输入的数据集、研究变量,以及选择研究变量对应的预存储研究参数。例如,用户输入了数据集,选择数据中的研究变量"均数1","均数2","标准差1","标准差2"。接着将研究变量"均数1","标准差1","均数2","标准差2"对应于预存储研究参数中的"均数1","标准差1","均数2","标准差2"。

[0172] 进一步地,这些数据可以以文件形式、数据库形式、或者程序代码形式读取。

[0173] 3.匹配预存储流程。在选择研究指标、研究设计、研究参数后,匹配预存储流程中包含有研究指标、研究设计和研究参数的预存储流程,得到一个或多个预存储流程。例如,按照上述研究指标"一类错误概率"、"检验效能",参数"均数1","标准差1","均数2","标准差2",匹配得到"模拟正态分布两样本均数比较"、"模拟指数分布两样本均数比较"的多个预存储流程。

[0174] 4.选择预存储流程和数据集。用户选择其中的一个或多个,如"模拟正态分布两样本均数比较"和输入的数据集。

[0175] 5.选择对比流程。用户选择已选择的预存储流程(包含数据集)。

[0176] 6.生成输出程序代码和说明,获取预存储流程的操作、方法、图像和流程程序代码,生成并保存相应的输出程序代码和输出程序说明。

[0177] 7.运行输出程序代码,运行上述流程的输出程序代码,得到并保存相应的运行结果。

[0178] 8.显示运行结果,显示相应的预存储流程、输出程序代码、输出程序说明和相应的运行结果。

[0179] 参照图3,本申请的实施例还提供了一种数据探索系统,包括:

[0180] 数据库,用于存储预设的指标库、参数库和流程库;所述指标库存储有多种不同的研究指标;所述参数库存储有多种不同的研究参数;所述流程库存储有多种不同的研究流程,以及每一种所述研究流程所对应的程序代码;

[0181] 指标选择模块,用于根据输入信息从指标库中选择指定的研究指标;

[0182] 数据获取模块,用于获取数据集、研究变量,以及所述研究变量对应的研究参数;

[0183] 流程匹配模块,用于从流程库中匹配出包含有所述研究指标和所述研究参数的可用流程;

[0184] 流程选择模块,用于根据输入指令从所述可用流程中选择探索流程,从数据集中选择探索数据集;

[0185] 程序输出模块,用于调用所述探索流程对应的程序代码,生成所述探索流程的输出程序代码和输出程序说明;

[0186] 探索输出模块,用于执行所述探索流程的输出程序代码,对所述探索数据集进行探索并将探索结果输出。

[0187] 进一步地,数据库还可用于存储预设的方法库、图像库。所述方法库存储有多种不同的统计方法;所述图像库存储有多种不同的图像;所述方法库和图像库分别存储有每一

种所述统计方法和图像所对应的程序代码。

[0188] 进一步地，所述数据库还用于存储预设的变量库，所述变量库存储有多种不同的研究变量；

[0189] 相应地，数据获取模块还用于：

[0190] 获取所述研究变量对应的变量库中的变量；

[0191] 所述数据库还用于存储预设的设计库，所述设计库存储有多种不同的研究设计；相应地，所述系统还包括：

[0192] 设计选择模块，用于根据输入信息从设计库中选择指定的研究设计；

[0193] 所述流程匹配模块还用于：

[0194] 从包含有所述研究指标和所述研究参数的可用流程中，进一步匹配出包含有对应研究设计的可用流程。

[0195] 进一步地，所述流程选择模块还用于：

[0196] 根据输入指令从已选择的探索流程和数据集中选择一个或多个作为对比流程和对比数据集；

[0197] 所述程序输出模块还用于：

[0198] 调用所述对比流程对应的程序代码，生成所述对比流程的输出程序代码和输出程序说明；

[0199] 分别合并所述探索流程的输出程序代码和所述对比流程的输出程序代码、所述探索流程的输出程序说明和所述对比流程的输出程序说明，生成全部的输出程序代码和输出程序说明；

[0200] 所述探索输出模块还用于：

[0201] 执行合并后的输出程序代码，对所述数据集进行探索；

[0202] 将所述探索流程的探索结果与所述对比流程的探索结果进行比较并将比较结果输出。

[0203] 关于上述实施例中的系统，其中各个模块执行操作的具体方式已经在有关该方法的实施例中进行了详细描述，此处将不做详细阐述说明。

[0204] 本申请的实施例还提供了一种计算设备，所述计算设备包括：处理器和存储器；

[0205] 所述存储器用于存储计算机程序指令；

[0206] 所述计算设备运行时，所述处理器执行所述存储器中的计算机程序指令，以执行如上所述的任意一种方法的操作步骤。

[0207] 根据本申请实施例的第四方面，提供一种计算机可读存储介质，其上存储有计算机程序，所述计算机程序被处理器执行如上所述的任意一种方法。

[0208] 可以理解的是，上述各实施例中相同或相似部分可以相互参考，在一些实施例中未详细说明的内容可以参见其他实施例中相同或相似的内容。

[0209] 需要说明的是，在本申请的描述中，术语"第一"、"第二"等仅用于描述目的，而不能理解为指示或暗示相对重要性。此外，在本申请的描述中，除非另有说明，"多个"的含义是指至少两个。

[0210] 流程图中或在此以其他方式描述的任何过程或方法描述可以被理解为，表示包括一个或更多个用于实现特定逻辑功能或过程的步骤的可执行指令的代码的模块、片段或部

分，并且本申请的优选实施方式的范围包括另外的实现，其中可以不按所示出或讨论的顺序，包括根据所涉及的功能按基本同时的方式或按相反的顺序，来执行功能，这应被本申请的实施例所属技术领域的技术人员所理解。

[0211] 应当理解，本申请的各部分可以用硬件、软件、固件或它们的组合来实现。在上述实施方式中，多个步骤或方法可以用存储在存储器中且由合适的指令执行系统执行的软件或固件来实现。例如，如果用硬件来实现，和在另一实施方式中一样，可用本领域公知的下列技术中的任一项或他们的组合来实现：具有用于对数据信号实现逻辑功能的逻辑门电路的离散逻辑电路，具有合适的组合逻辑门电路的专用集成电路，可编程门阵列(PGA)，现场可编程门阵列(FPGA)等。

[0212] 本技术领域的普通技术人员可以理解实现上述实施例方法携带的全部或部分步骤是可以通过程序来指令相关的硬件完成，所述的程序可以存储于一种计算机可读存储介质中，该程序在执行时，包括方法实施例的步骤之一或其组合。

[0213] 此外，在本申请各个实施例中的各功能单元可以集成在一个处理模块中，也可以是各个单元单独物理存在，也可以两个或两个以上单元集成在一个模块中。上述集成的模块既可以采用硬件的形式实现，也可以采用软件功能模块的形式实现。所述集成的模块如果以软件功能模块的形式实现并作为独立的产品销售或使用时，也可以存储在一个计算机可读取存储介质中。

[0214] 上述提到的存储介质可以是只读存储器，磁盘或光盘等。

[0215] 在本说明书的描述中，参考术语"一个实施例"、"一些实施例"、"示例"、"具体示例"、或"一些示例"等的描述意指结合该实施例或示例描述的具体特征、结构、材料或者特点包含于本申请的至少一个实施例或示例中。在本说明书中，对上述术语的示意性表述不一定指的是相同的实施例或示例。而且，描述的具体特征、结构、材料或者特点可以在任何的一个或多个实施例或示例中以合适的方式结合。

[0216] 尽管上面已经示出和描述了本申请的实施例，可以理解的是，上述实施例是示例性的，不能理解为对本申请的限制，本领域的普通技术人员在本申请的范围内可以对上述实施例进行变化、修改、替换和变型。

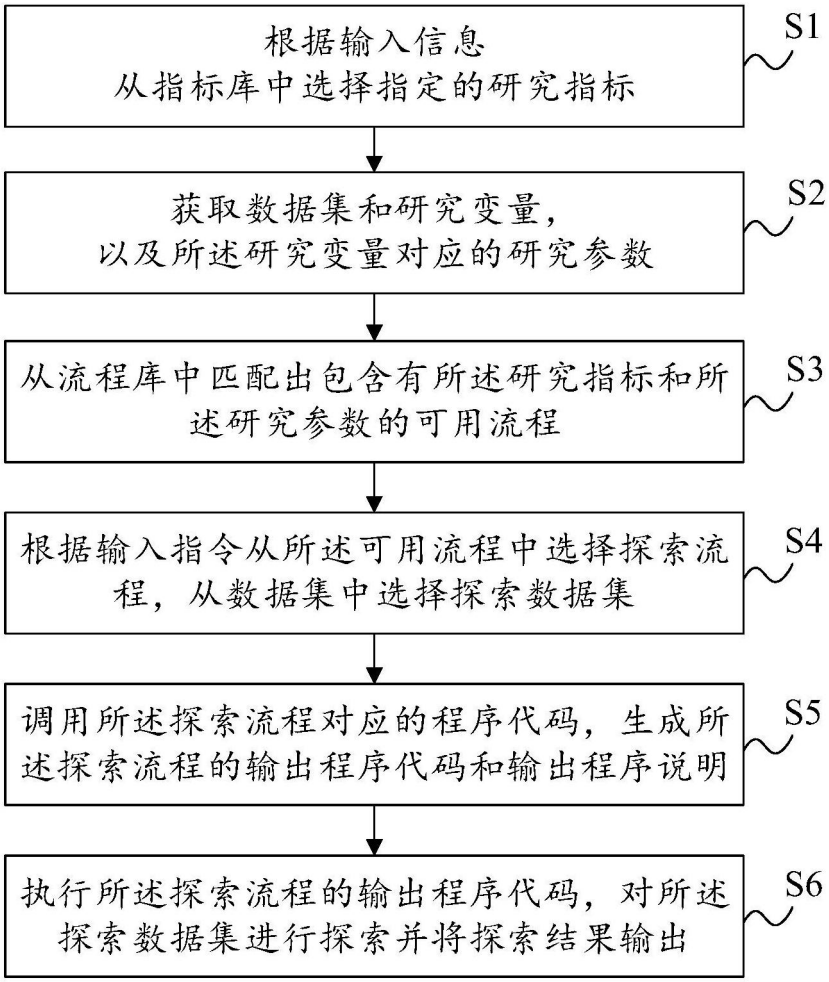

根据输入信息
从指标库中选择指定的研究指标
S1
获取数据集和研究变量，
以及所述研究变量对应的研究参数
S2
从流程库中匹配出包含有所述研究指标和所述研究参数的可用流程
S3
根据输入指令从所述可用流程中选择探索流程，从数据集中选择探索数据集
S4
调用所述探索流程对应的程序代码，生成所述探索流程的输出程序代码和输出程序说明
S5
执行所述探索流程的输出程序代码，对所述探索数据集进行探索并将探索结果输出
S6

图1

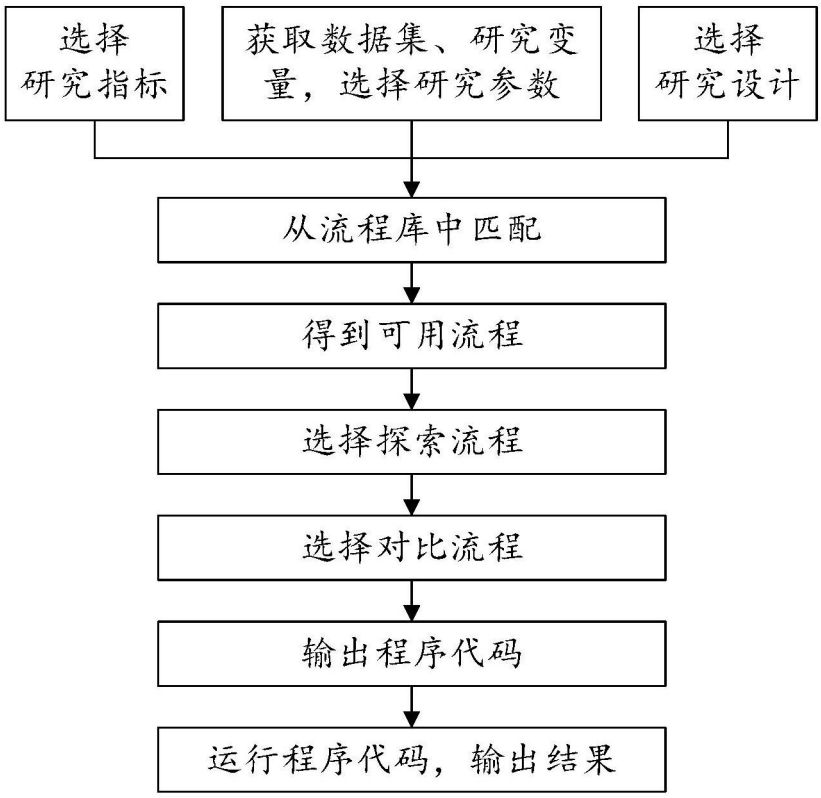

选择
研究指标
获取数据集、研究变
量，选择研究参数
选择
研究设计
从流程库中匹配
得到可用流程
选择探索流程
选择对比流程
输出程序代码
运行程序代码，输出结果

图2

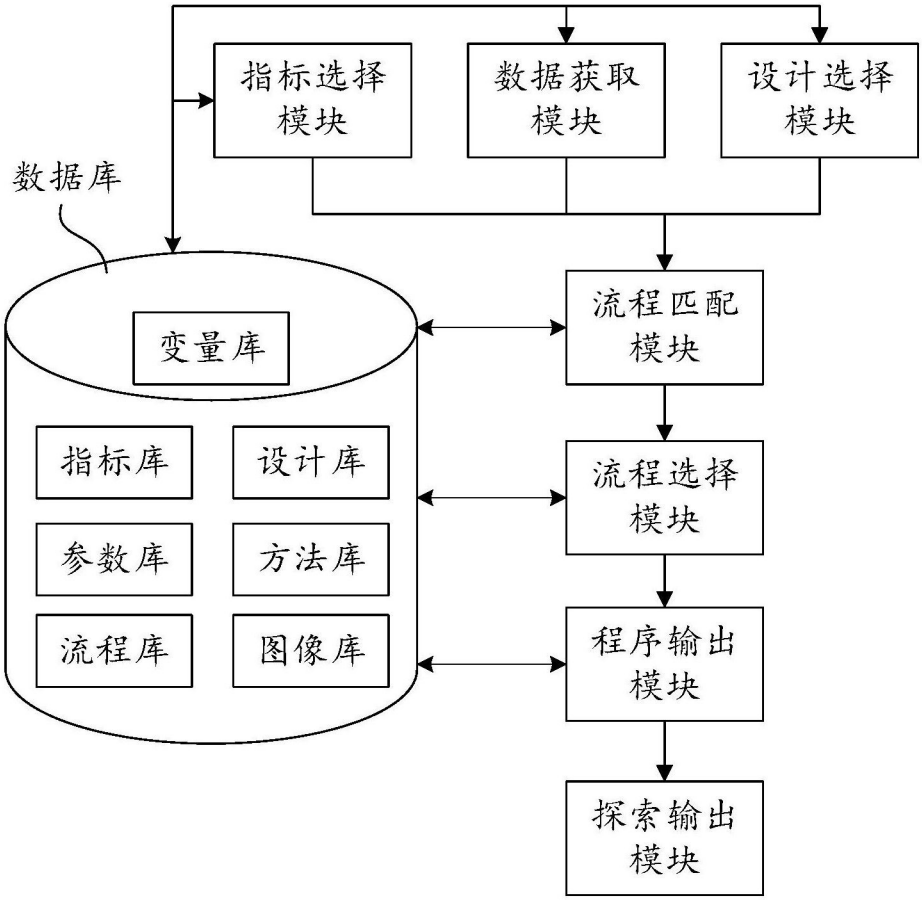

指标选择模块
数据获取模块
设计选择模块
数据库
变量库
指标库
设计库
参数库
方法库
流程库
图像库
流程匹配模块
流程选择模块
程序输出模块
探索输出模块

图3