(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国 际 局

(43) 国际公布日
2019 年 7 月 4 日 (04.07.2019)

WIPO | PCT

(10) 国际公布号
WO 2019/129060 A1

(51) 国际专利分类号：
006X 9/62 (2006.01) *0061√ 99/00* (2019.01)

(21) 国际申请号： ?01/CN2018/123910

(22) 国际申请日： 2018 年 12 月 26 日 (26. 12.2018)

(25) 申请语言： 中文

(26) 公布语言： 中文

(3○) 优先权：
20171 1445538.3 2017年 12月27日 (27. 12.2017) CN

(71) 申请人：第四范式（北京）技术有限公司 (THE FOURTH PARADIGM (BEIJING) TECH CO LTD) [CN/CN]：中国北京市海淀区上地东路35号颐泉汇大厦写字楼A座610室，Beijing 100085 (CN)。

(72) 发明人：杨强 (YANG, Qiang)；中国北京市海淀区上地东路35号颐泉汇大厦写字楼A座610室，Beijing 100085 (CN)。戴文渊 (DAI, Wenyuan)；中国北京市海淀区上地东路35号颐泉汇大厦写字楼A座610室，Beijing 100085 (CN)。陈雨强 (CHEN, Yuqiang)；中国北京市海淀区上地东路35号颐泉汇大厦写字楼A座610室，Beijing 100085 (CN)。孙迪 (SUN, Di)；中国北京市海淀区上地东路35号颐泉汇大厦写字楼A座610室，Beijing 100085 (CN)。杨慧斌 (YANG, Huibin)；中-北京市海淀区上地东路35号颐泉汇大厦写字楼A座610室，Beijing 100085 (CN)。刘守湘 (LIU, Shouxiang)；中国北京市海淀区上地东路35号颐泉汇大厦写字楼A座610室，Beijing 100085 (CN)。

(74) 代理人：北京展翼知识产权代理事务所（特殊普通合伙）(ZYX INTELLECTUAL PROPERTY LAW

(54) **Title** :METHOD AND SYSTEM FOR AUTOMATICALLY GENERATING MACHINE LEARNING SAMPLE

(54) 发明名称：自动生成机器学习样本的特征的方法及系统

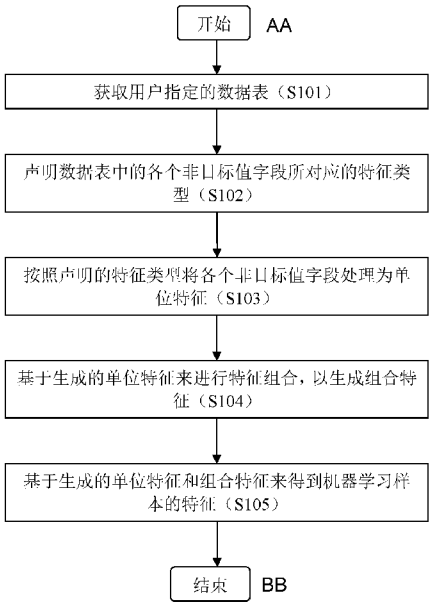

图 1

S101 Acquire a data table specified by a user
S102 State a feature type corresponding to each non-target value field in the data table
S103 Process each non-target value field as a unit feature according to the stated feature type
S104 Perform feature combination based on the generated unit feature, to generate a combined feature
S105 Obtain a feature of a machine learning sample based on the generated unit feature and the combined feature
AA Start
BB End

(57) Abstract: Provided are a method and system for automatically generating a machine learning sample. The method comprises: acquiring a data table specified by a user, wherein one row of the data table corresponds to one data record, and one column of the data table corresponds to one data field; stating a feature type corresponding to each non-target value field in the data table, wherein the feature type comprises a disperse feature, or comprises a continuous feature, or comprises a disperse feature and a continuous feature; processing each non-target value field as a unit feature according to the stated feature type; performing feature combination based on

WO 2019/129060 A1

[见续页]

FIRM)；中国北京市海淀区中关村东路18号财智国际大厦B-1503, Beijing 100083 (CN)。

(81) 指定国(除另有指明，要求每一种可提供的国家保护)：AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW。

(84) 指定国(除另有指明，要求每一种可提供的地区保护)：ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布：

- 包括国际检索报告(条约第21条(3))。

the generated unit feature, to generate a combined feature; and obtaining a feature of a machine learning sample based on the generated unit feature and the combined feature.

(57)摘要：提供一种自动生成机器学习样本的特征的方法及系统。所述方法包括：获取用户指定的数据表，其中，数据表的一行对应一条数据记录，数据表的一列对应一个字段；声明数据表中的各个非目标值字段所对应的特征类型；其中，所述特征类型包括离散特征，或包括连续特征，或包括离散特征和连续特征；按照声明的特征类型将各个非目标值字段处理为单位特征；基于生成的单位特征来进行特征组合，以生成组合特征；以及基于生成的单位特征和组合特征来得到机器学习样本的特征。

自动生成机器学习样本的特征的方法及系统

技术领域

本公开总体说来涉及人工智能领域，更具体地讲，涉及一种自动生成机器学习样本的特征的方法及系统。

背景技术

随着海量数据的出现，人们倾向于使用机器学习技术来从数据中挖掘出价值。

训练机器学习模型的基本过程主要包括：

1、导入包含历史数据记录的数据集（例如，数据表）；

2、完成特征工程，其中，通过对数据集中的数据记录的属性信息进行各种处理，以得到各个特征，这些特征构成的特征向量可作为机器学习样本；

3、训练模型，其中，按照设置的机器学习算法（例如，逻辑回归算法、决策树算法、神经网络算法等），基于经过特征工程所得到的机器学习样本来学习出模型。

在上述过程中，产生特征的处理很重要，它会影响模型的优劣。数据表中每条数据记录可包括多个属性信息（即，字段），而特征可指示各字段本身、或字段的组合等各种字段处理（或运算）结果，以便更好地反映数据分布以及字段间的内在关联与潜在含义，因此，特征工程质量的好坏直接决定了机器学习问题刻画的准确性，进而影响模型的优劣。

在现有的机器学习平台上，可采用基于图形界面的交互方式来完成机器学习模型训练流程，而不需要用户亲自编写程序代码。然而，在特征工程环节，却往往是将人为设定的特征生成方式手动地输入到平台系统中。也就是说，需要用户预先设定机器学习样本的特征，一方面，需要用户对业务场景有深刻的理解，即，用户凭借业务经验来设定特征；另一方面，一般在机器学习过程中，所使用数据的数据量都比较大，用户有时不能全面地分析数据，会导致设定一些无效的特征，为了提高机器学习样本的特征的效果，这就需要用户进行不断尝试，当面对大数据量和高维特征时，这样的工作需要花费较长的时间。这种情况下，不仅需要用户对业务场景有深刻的理解，增加了用户的工作量，还降低了机器学习的效率。

发明内容

本公开的示例性实施例在于提供一种自动生成机器学习样本的特征的方法及系统，以解决现有技术存在的不能便捷地生成机器学习样本的特征的问

题。

根据本公开的示例性实施例，提供一种自动生成机器学习样本的特征的方法，包括：(人)获取用户指定的数据表，其中，数据表的一行对应一条数据记录，数据表的一列对应一个字段；(6)声明数据表中的各个非目标值字段所对应的特征类型，其中，特征类型包括离散特征和/或连续特征；(C)按照声明的特征类型将各个非目标值字段处理为单位特征；(〇)基于生成的单位特征来进行特征组合，以生成组合特征；以及(£)基于生成的单位特征和组合特征来得到机器学习样本的特征。

根据本公开的另一示例性实施例，提供一种自动生成机器学习样本的特征的系统，包括：数据表获取装置，用于获取用户指定的数据表，其中，数据表的一行对应一条数据记录，数据表的一列对应一个字段；声明装置，用于声明数据表中的各个非目标值字段所对应的特征类型，其中，特征类型包括离散特征和/或连续特征；单位特征生成装置，用于按照声明的特征类型将各个非目标值字段处理为单位特征；组合特征生成装置，用于基于生成的单位特征来进行特征组合，以生成组合特征；以及特征获取装置，用于基于生成的单位特征和组合特征来得到机器学习样本的特征。

根据本公开的另一示例性实施例，提供一种用于自动生成机器学习样本的特征的计算机可读介质，其中，在所述计算机可读介质上记录有用于执行如上所述的自动生成机器学习样本的特征的方法的计算机程序。

根据本公开的另一示例性实施例，提供一种用于自动生成机器学习样本的特征的计算装置，包括存储部件和处理器，其中，存储部件中存储有计算机可执行指令集合，当所述计算机可执行指令集合被所述处理器执行时，执行如上所述的自动生成机器学习样本的特征的方法。

在根据本公开示例性实施例的自动生成机器学习样本的特征的方法及系统中，能够基于数据表自动生成机器学习样本的特征，既降低了特征工程的使用门槛，提高了特征工程的易用性，又提高了特征工程的效率。

将在接下来的描述中部分阐述本公开总体构思另外的方面和/或优点，还有一部分通过描述将是清楚的，或者可以经过本公开总体构思的实施而得知。

附图说明

通过下面结合示例性地示出实施例的附图进行的描述，本公开示例性实施例的上述和其他目的和特点将会变得更加清楚，其中：

图 1 示出根据本公开示例性实施例的自动生成机器学习样本的特征的方法的流程图；

图 2 示出根据本公开示例性实施例的由用户指定非目标值字段对应的特征类型的示例；

图 3 示出根据本公开的另一示例性实施例的自动生成机器学习样本的特征的方法的流程图；

图 4 示出根据本公开的另一示例性实施例的自动生成机器学习样本的特征的方法的流程图；

图 5 示出根据本公开的另一示例性实施例的自动生成机器学习样本的特征的方法的流程图；

图 6 示出根据本公开示例性实施例的用于训练机器学习模型的 DAG 图的示例；

图 7 示出根据本公开示例性实施例的自动生成机器学习样本的特征的系统的框图。

具体实施方式

现将详细参照本公开的实施例，所述实施例的示例在附图中示出，其中，相同的标号始终指的是相同的部件。以下将通过参照附图来说明所述实施例，以便解释本公开。

这里，机器学习是人工智能研究发展到一定阶段的必然产物，其致力于通过计算的手段，利用经验来改善系统自身的性能。在计算机系统中，“经验”通常以“数据”形式存在，通过机器学习算法，可从数据中产生“模型”，也就是说，将经验数据提供给机器学习算法，就能基于这些经验数据产生模型，在面对新的情况时，模型会提供相应的判断，即，预测结果。不论是训练机器学习模型，还是利用训练好的机器学习模型进行预测，数据都需要转换为包括各种特征的机器学习样本。机器学习可被实现为“有监督学习”、“无监督学习”或“半监督学习”的形式，应注意，本公开的示例性实施例对具体的机器学习算法并不进行特定限制。此外，还应注意，在训练和应用模型的过程中，还可结合统计算法等其他手段。

在此需要说明的是，在本公开中出现的“并且/或者”、“和/或”均表示包含三种并列的情况。例如“包括八和/或 3”表示如下三种并列的情况：(1)包括八；(2)包括氏 (3)包括八和 8。又例如“执行步骤一并且/或者步骤二”表示如下三种并列的情况：(1)执行步骤一；(2)执行步骤二；(3)执行步骤一和步骤二。

图 1 示出根据本公开示例性实施例的自动生成机器学习样本的特征的方法的流程图。这里，作为示例，所述方法可通过计算机程序来执行，也可由专门的自动生成机器学习样本的特征的系统或计算装置来执行。

作为示例，所述方法可通过启动与自动特征生成步骤相应的算子而自动执行。换言之，当与自动特征生成步骤相应的算子被启动时，将自动执行所述方法。进一步地，作为示例，所述算子对应于与机器学习流程相应的有向无环图(〇八〇图)中的节点。例如，与机器学习流程相应的DAG 图可包括特征生成节点，当运行整个 DAG 图时，在执行到所述特征生成节点时，将自动执行所述方法。下面，将结合图6来对根据本公开的示例性实施例的用于训练机器学习模型的DAG 图进行详细说明。

参照图 1，在步骤 S101 中，获取用户指定的数据表。这里，数据表的一行对应一条数据记录，数据表的一列对应一个字段。换言之，数据表中的每条数据记录具有与各个字段相应的字段值。作为示例，每条数据记录可被看作关于一个事件或对象的描述，对应于一个示例或样例，每个字段可用于描述事件或对象在一个方面的表现或性质 (例如，名字、年龄、职业等)。

作为示例，可向用户提供用于指定数据表的图形界面，并根据用户在该图形界面上执行的输入操作，来确定用户所指定的数据表。

在步骤 S102 中，声明数据表中的各个非目标值字段所对应的特征类型，其中，特征类型包括离散特征和/或连续特征。

这里，目标值字段即使用机器学习技术要预估的标记 (即，label)所对应的字段，该字段对应于有监督学习情况下的预测目标，而非目标值字段即数据表之中除目标值字段之外的字段。

在有监督学习的情况下，作为示例，非目标值字段可通过以下方式来获取：从数据表中的所有字段中去除用户指定的目标值字段。作为示例，可向用户提供用于指定目标值字段的图形界面，并根据用户在该图形界面上执行的输入操作，来确定用户所指定的目标值字段。进一步地，作为示例，所述算子在用户未指定目标值字段的情况下被启动时，可提供异常提醒，以提醒用户指定目标值字段。

此外，应该理解，数据表中可包括目标值字段，也可不包括目标值字段。

连续特征是与离散特征 (例如，类别特征)相对的一种特征，其取值可以是具有一定连续性的数值，例如，年龄、金额等。相对地，作为示例，离散特征的取值不具有连续性，例如，可以是“来自北京”、“来自上海”或“来自天津”、“性别为男”、“性别为女”等无序分类的特征。

作为示例，可自动或根据用户的指示，将所有非目标值字段声明为离散特征，或者，将各个非目标值字段声明为与其字段值数据类型相应的离散特征或连续特征。

作为示例，字段的字段值数据类型可为连续型 (例如，数值型 (例如，整型 int))或离散型 (例如，文本型 (例如，字符串型 string))○作为示例，将各个非目标值字段声明为与其字段值数据类型相应的离散特征或连续特征的步骤可包括：将数据表中的字段值数据类型为离散型的非目标值字段声明为离散特征，并将数据表中的字段值数据类型为连续型的非目标值字段声明为连续特征。

作为示例，可向用户提供用于指定非目标值字段对应的特征类型的图形界面，并根据用户在该图形界面上执行的输入操作，将所有非目标值字段声明为离散特征，或者，将各个非目标值字段声明为与其字段值数据类型相应的离散特征或连续特征。

下面结合图 2 来描述根据本公开示例性实施例的由用户通过图形界面来指定非目标值字段对应的特征类型的示例。如图 2 所示，用于指定非目标值字

段对应的特征类型的图形界面可显示单选按钮“全部离散”和单选按钮“离散+连续”(这两个按钮可被择一选中),可响应于用户对单选按钮“全部离散”的选择操作,将数据表中的所有非目标值字段声明为离散特征;可响应于用户对单选按钮“离散+连续”的选择操作,根据各个非目标值字段的数据类型将所述字段声明为相应的离散特征或连续特征,这里,可根据字段值的特性来自动判断出字段的数据类型,并进而根据数据类型为离散型还是连续型将字段声明为离散特征或连续特征。此外,所述图形界面中还可显示用于指定目标值字段的控件,用户可通过对该控件的操作来指定目标值字段。此外,所述图形界面的左侧还可显示数据表中的各字段的字段名及字段值数据类型。

参照回图 1,在步骤 8103 中,按照声明的特征类型将各个非目标值字段处理为单位特征。换言之,按照声明的特征类型分别将每个非目标值字段处理为一个单位特征。

作为示例,可对每一个字段值数据类型为连续型且被声明为离散特征的非目标值字段进行离散化处理,以得到一个单位特征。

应理解,这里的单位特征是指该特征对应于单个字段,其本身可根据取值的定义而具有一个或多个维度。可选地,可针对每一个字段值数据类型为连续型且被声明为离散特征的非目标值字段,执行一种或多种分桶运算以得到相应的一个或多个分桶特征,并将得到的分桶特征整体作为一个单位特征。

这里,分桶化$_{1111111¾}$)运算是指对连续型的字段进行分散化的一种特定方式,即,将连续型的字段的值域划分为多个区间(即,多个桶),并基于划分的桶来确定相应的分桶特征值。分桶运算大体上可划分为有监督分桶和无监督分桶,这两种类型各自包括一些具体的分桶方式,例如,有监督分桶可包括最小熵分桶、最小描述长度分桶等,而无监督分桶可包括等宽分桶、等深分桶、基于 k 均值聚类的分桶等。在每种分桶方式下,可设置相应的分桶参数,例如,宽度、深度等。

应注意,根据本公开的示例性实施例,对字段值数据类型为连续型且被声明为离散特征的非目标值字段执行的分桶运算不限制分桶方式的种类,也不限制分桶运算的参数,并且,相应产生的分桶特征的具体表示方式也不受限制。

作为示例,针对字段值数据类型为连续型且被声明为离散特征的非目标值字段执行的多种分桶运算可以在分桶方式和/或分桶参数方面存在差异。例如,所述多种分桶运算可以是种类相同但具有不同运算参数(例如,深度、宽度等)的分桶运算,也可以是不同种类的分桶运算。相应地,每一种分桶运算可得到一个分桶特征,这些分桶特征共同组成一个分桶组特征,该分桶组特征可体现出不同分桶运算,从而提升了机器学习素材的有效性,为机器学习模型的训练/预测提供了较好的基础。

也就是说,根据本公开的示例性实施例,可针对每一个字段值数据类型为连续型且被声明为离散特征的非目标值字段执行至少一种分桶运算而得到相应的至少一个分桶特征,将每一个分桶特征作为一个组成元素而得到与该字段

对应的特征，并将该特征作为单位特征。这里，应理解，分桶运算的执行使得字段值数据类型为连续型且被声明为离散特征的非目标值字段被分散化地置入相应的特定桶中，在转换后的多个分桶特征中，每个维度既可以指示桶中是否被分配了连续特征的离散值（例如，“0”或“1”），也可以指示具体的连续数值（例如，连续特征的实际特征值或其归一化值、所述桶中各连续特征的平均值、中间值、边界值等）。相应地，在机器学习中具体应用各个维度的离散值（例如，针对分类问题）或连续数值（例如，针对回归问题）时，可进行离散值之间的组合（例如，笛卡尔积等）或连续数值之间的组合（例如，算术运算组合等）。

在步骤 S104 中，基于生成的单位特征来进行特征组合，以生成组合特征。

作为示例，可对生成的全部单位特征进行各种组合来获取候选组合特征，或者，对生成的全部单位特征之中特征重要性较高的单位特征进行各种组合来获取候选组合特征；然后，可通过衡量与每个候选组合特征相应的机器学习模型的效果来从候选组合特征中筛选出组合特征。具体说来，可训练与每个候选组合特征相应的机器学习模型，由于相应的机器学习模型的效果能够反映候选组合特征的特征重要性（例如，预测力），从而可通过衡量与每个候选组合特征相应的机器学习模型的效果来从候选组合特征中筛选出组合特征，例如，机器学习模型的效果越好，相应的候选组合特征越容易被筛选为组合特征。作为示例，可使用指定的模型评价指标来评价与每个候选组合特征相应的机器学习模型的效果。作为示例，可自动或根据用户的指示，来指定模型评价指标。

作为示例，模型评价指标可以是 AUC（ROC（受试者工作特征，Receiver Operating Characteristic）曲线下的面积，Area Under ROC Curve）、MAE（平均绝对误差，Mean Absolute Error）或对数损失函数（logloss）等。

作为示例，可将全部单位特征之中特征重要性满足第一预设条件的单位特征进行各种组合来获取候选组合特征。例如，可将全部单位特征之中特征重要性处于第一预设阈值范围内的单位特征进行各种组合来获取候选组合特征，或者，按照单位特征的特征重要性由高到低将全部单位特征进行排序，并将前第一预定数量的单位特征进行各种组合来获取候选组合特征。

作为示例，可通过衡量与特征相应的机器学习模型的效果来确定单位特征的特征重要性，相应的机器学习模型的效果越好，单位特征的特征重要性越高。例如，可使用与特征相应的机器学习模型关于模型评价指标的评价值来衡量单位特征的特征重要性。这里，作为示例，可自动或根据用户的指示，来指定该模型评价指标。

在步骤 S105 中，基于生成的单位特征和组合特征来得到机器学习样本的特征。

作为示例，可将生成的全部单位特征和全部组合特征作为机器学习样本的特征。

作为另一示例，可将生成的全部单位特征和全部组合特征之中，特征重要

性较高的特征作为机器学习样本的特征。作为示例，可将全部单位特征和全部组合特征之中，特征重要性满足第二预设条件的特征作为机器学习样本的特征，例如，可将特征重要性处于第二预设阈值范围内的特征作为机器学习样本的特征，或者，按照特征的特征重要性由高到低将全部单位特征和全部组合特征共同进行排序，并将前第二预定数量的特征作为机器学习样本的特征。

作为另一示例，可将生成的全部单位特征之中特征重要性较高的单位特征和生成的全部组合特征，作为机器学习样本的特征。作为示例，可将全部组合特征连同特征重要性满足第三预设条件的单位特征作为机器学习样本的特征，例如，可将全部组合特征连同特征重要性处于第三预设阈值范围内的单位特征作为机器学习样本的特征，或者，按照单位特征的特征重要性由高到低将全部单位特征进行排序，并将前第三预定数量的单位特征连同全部组合特征作为机器学习样本的特征。

作为另一示例，可将生成的全部单位特征和生成的全部组合特征之中特征重要性较高的组合特征，作为机器学习样本的特征。作为示例，可将全部单位特征连同特征重要性满足第四预设条件的组合特征作为机器学习样本的特征，例如，可将全部单位特征连同特征重要性处于第四预设阈值范围内的组合特征作为机器学习样本的特征，或者，按照组合特征的特征重要性由高到低将全部组合特征进行排序，并将前第四预定数量的组合特征连同全部单位特征作为机器学习样本的特征。

此外，作为示例，根据本公开示例性实施例的自动生成机器学习样本的特征的方法还可包括：在步骤 8105 之后，向用户显示得到的机器学习样本的特征。进一步地，还可向用户显示每个特征的特征重要性。

作为示例，根据本公开示例性实施例的自动生成机器学习样本的特征的方法还可包括：在步骤 8105 之后，直接将得到的机器学习样本的特征应用于后续的机器学习步骤。例如，可直接基于得到的机器学习样本的特征来学习出模型。

图 3 示出根据本公开的另一示例性实施例的自动生成机器学习样本的特征的方法的流程图。

参照图 3，在步骤 8201 中，获取用户指定的数据表。

在步骤 8202 中，声明数据表中的各个非目标值字段所对应的特征类型。

在步骤 8203 中，按照声明的特征类型将各个非目标值字段处理为单位特征。

在步骤 8204 中，对生成的全部单位特征进行各种组合来获取候选组合特征，并通过衡量与每个候选组合特征相应的机器学习模型的效果来从候选组合特征中筛选出组合特征。

在步骤 8205 中，将生成的全部单位特征和全部组合特征作为机器学习样本的特征。

图 4 示出根据本公开的另一示例性实施例的自动生成机器学习样本的特

征的方法的流程图。

参照图4，在步骤8301中，获取用户指定的数据表。

在步骤8302中，声明数据表中的各个非目标值字段所对应的特征类型。

在步骤8303中，按照声明的特征类型将各个非目标值字段处理为单位特征。

在步骤8304中，对生成的全部单位特征之中特征重要性较高的单位特征进行各种组合来获取候选组合特征，并通过衡量与每个候选组合特征相应的机器学习模型的效果来从候选组合特征中筛选出组合特征。

在步骤8305中，将生成的全部单位特征之中特征重要性较高的单位特征和生成的全部组合特征作为机器学习样本的特征。

作为示例，可使用与特征相应的机器学习模型关于模型评价指标AUC的评价值来衡量特征的特征重要性，在步骤S304中，可对生成的全部单位特征之中对应的AUC值大于0.5且小于1的单位特征进行各种组合来获取候选组合特征，并且，在步骤S305中，可将生成的全部单位特征之中对应的AUC值大于0.5且小于1的单位特征和生成的全部组合特征作为机器学习样本的特征。

图5示出根据本公开的另一示例性实施例的自动生成机器学习样本的特征的方法的流程图。

参照图5，在步骤8401中，获取用户指定的数据表。

在步骤8402中，声明数据表中的各个非目标值字段所对应的特征类型。

在步骤8403中，按照声明的特征类型将各个非目标值字段处理为单位特征。

在步骤8404中，对生成的全部单位特征进行各种组合来获取候选组合特征，并通过衡量与每个候选组合特征相应的机器学习模型的效果来从候选组合特征中筛选出组合特征。

在步骤8405中，将生成的全部单位特征和全部组合特征之中，特征重要性较高的特征作为机器学习样本的特征。

作为示例，可使用与特征相应的机器学习模型关于模型评价指标AUC的评价值来衡量特征的特征重要性，在步骤S405中，可将生成的全部单位特征和全部组合特征之中，对应的AUC值大于0.5且小于1的特征作为机器学习样本的特征。

以上列出了一些自动生成机器学习样本的特征的示例性方法，然而，本领域技术人员应理解，本公开的示例性实施例并不受限于这些方法，而可以采用任何适当的特征(单位特征、候选组合特征或组合特征)生成或筛选方式。

根据本公开的示例性实施例，可通过有向无环图的形式来执行机器学习流程，该机器学习流程可涵盖用于进行机器学习模型训练、测试或预估的全部或部分步骤。例如，可针对机器学习模型训练来建立包括以下步骤之中的至少一个步骤的DAG图：历史数据导入步骤、数据拆分步骤、特征生成步骤、逻辑回归步骤和模型预测步骤。也即，上述各个步骤可作为DAG图中的节点而被

执行。

图 6 示出根据本公开示例性实施例的用于训练机器学习模型的 DAG 图的示例。

参照图 6，第一步：建立数据导入节点。作为示例，可响应于用户操作对数据导入节点进行设置以获取名称为“bank”的银行业务数据表 (即，将该数据表导入机器学习平台中)，其中，该数据表中可包含多条历史数据记录。

第二步：建立数据拆分节点，并将数据导入节点连接到数据拆分节点，以将上述导入的数据表拆分为训练集和验证集，其中，训练集中的数据记录用于转换为机器学习样本以学习出模型，而验证集中的数据记录用于转换为测试样本以验证学习出的模型的效果。可响应于用户操作对数据拆分节点进行设置以按照设置的方式将上述导入的数据表拆分为训练集和验证集。

第三步：建立两个特征生成节点，并将数据拆分节点分别连接到这两个特征生成节点，以对数据拆分节点输出的训练集和验证集分别进行特征生成，例如，默认数据拆分节点左侧输出的是训练集，右侧输出的是验证集。应理解，对于机器学习样本和测试样本而言，两者的特征生成方式是对应一致的。可响应于用户操作对特征生成节点进行设置，例如，可指定目标值字段、非目标值字段对应的特征类型、特征重要性的衡量指标等。

第四步：建立特点算法 (例如，逻辑回归)节点 (也即，模型训练节点)，并将左侧特征生成节点连接到逻辑回归节点，以利用逻辑回归算法基于机器学习样本来训练出机器学习模型。可响应于用户操作对逻辑回归节点进行设置以按照设置的逻辑回归算法来训练机器学习模型。

第五步：建立模型预测节点，并将逻辑回归节点和右侧特征生成节点连接到模型预测节点，以基于测试样本来验证训练出的机器学习模型的效果。可响应于用户操作对模型预测节点进行设置以按照设置的验证方式来验证机器学习模型的效果。

在建立包括上述步骤的 DAG 图之后，可根据用户的指示来运行整个 DAG 图。在执行到所述特征生成节点时，可自动执行上述示例性实施例的自动生成机器学习样本的特征的方法。

图 7 示出根据本公开示例性实施例的自动生成机器学习样本的特征的系统的框图。如图 7 所示，根据本公开示例性实施例的自动生成机器学习样本的特征的系统包括：数据表获取装置 10、声明装置 20、单位特征生成装置 30、组合特征生成装置 40 以及特征获取装置 50。

具体说来，数据表获取装置 10 用于获取用户指定的数据表，其中，数据表的一行对应一条数据记录，数据表的一列对应一个字段。

声明装置 20 用于声明数据表中的各个非目标值字段所对应的特征类型，其中，特征类型包括离散特征和/或连续特征。

作为示例，非目标值字段可通过以下方式来获取：从数据表中的所有字段中去除用户指定的目标值字段。

作为示例，声明装置 20 可自动或根据用户的指示，将所有非目标值字段声明为离散特征，或者，将各个非目标值字段声明为与其字段值数据类型相应的离散特征或连续特征。

单位特征生成装置 30 用于按照声明的特征类型将各个非目标值字段处理为单位特征。

作为示例，单位特征生成装置 30 可针对每一个字段值数据类型为连续型且被声明为离散特征的非目标值字段，执行一种或多种分桶运算以得到相应的一个或多个分桶特征，并将得到的分桶特征整体作为一个单位特征。

组合特征生成装置 40 用于基于生成的单位特征来进行特征组合，以生成组合特征。

作为示例，组合特征生成装置 40 可包括：候选组合特征获取单元 (未示出) 和组合特征筛选单元 (未示出)。

候选组合特征获取单元用于对生成的全部单位特征进行各种组合来获取候选组合特征，或者，对生成的全部单位特征之中特征重要性较高的单位特征进行各种组合来获取候选组合特征。

组合特征筛选单元用于通过衡量与每个候选组合特征相应的机器学习模型的效果来从候选组合特征中筛选出组合特征。

特征获取装置 50 用于基于生成的单位特征和组合特征来得到机器学习样本的特征。

作为示例，特征获取装置 50 可将生成的全部单位特征和全部组合特征作为机器学习样本的特征。

作为另一示例，特征获取装置 50 可将生成的全部单位特征和全部组合特征之中，特征重要性较高的特征作为机器学习样本的特征。

作为另一示例，特征获取装置 50 可将生成的全部单位特征之中特征重要性较高的单位特征和生成的全部组合特征，作为机器学习样本的特征。

作为另一示例，特征获取装置 50 可将生成的全部组合特征之中特征重要性较高的组合特征和生成的全部单位特征，作为机器学习样本的特征。

作为示例，根据本公开示例性实施例的自动生成机器学习样本的特征的系统还可包括：显示装置 (未示出)，显示装置用于向用户显示特征获取装置 50 得到的机器学习样本的特征。进一步地，作为示例，显示装置还可向用户显示每个特征的特征重要性。

作为示例，根据本公开示例性实施例的自动生成机器学习样本的特征的系统还可包括：应用装置 (未示出)，应用装置用于直接将特征获取装置 50 得到的机器学习样本的特征应用于后续的机器学习步骤。

作为示例，可通过启动与自动特征生成步骤相应的算子来使根据本公开示例性实施例的自动生成机器学习样本的特征的系统自动执行操作。

作为示例，所述算子可对应于与机器学习流程相应的有向无环图中的节点。

此外，作为示例，根据本公开示例性实施例的自动生成机器学习样本的特征的系统还可包括：提醒装置 (未示出)，提醒装置用于所述算子在用户未指定目标值字段的情况下被启动时，提供异常提醒。

应该理解，根据本公开示例性实施例的自动生成机器学习样本的特征的系统的具体实现方式可参照结合图1至图6描述的相关具体实现方式来实现，在此不再赘述。

根据本公开示例性实施例的自动生成机器学习样本的特征的系统所包括的装置可被分别配置为执行特定功能的软件、硬件、固件或上述项的任意组合。例如，这些装置可对应于专用的集成电路，也可对应于纯粹的软件代码，还可对应于软件与硬件相结合的模块。此外，这些装置所实现的一个或多个功能也可由物理实体设备 (例如，处理器、客户端或服务器等)中的组件来统一执行。

应理解，根据本公开示例性实施例的自动生成机器学习样本的特征的方法可通过记录在计算可读存储介质上的程序来实现，例如，根据本公开的示例性实施例，可提供一种存储指令的计算机可读存储介质，其中，当所述指令被至少一个计算装置运行时，促使所述至少一个计算装置执行：获取用户指定的数据表，其中，数据表的一行对应一条数据记录，数据表的一列对应一个字段；声明数据表中的各个非目标值字段所对应的特征类型，其中，特征类型包括离散特征和/或连续特征；按照声明的特征类型将各个非目标值字段处理为单位特征；基于生成的单位特征来进行特征组合，以生成组合特征；以及基于生成的单位特征和组合特征来得到机器学习样本的特征。

此外，当所述指令被至少一个计算装置运行时，还促使所述至少一个计算装置执行前述任一实施例中涉及的自动生成机器学习样本的特征的方法。

上述计算机可读存储介质中的计算机程序可在诸如处理器、客户端、主机、代理装置、服务器等计算机设备中部署的环境中运行，例如，由位于单机环境或分布式集群环境的至少一个计算装置来运行，作为示例，这里的计算装置可作为计算机、处理器、计算单元 (或模块)、客户端、主机、代理装置、服务器等。应注意，所述计算机程序还可用于执行除了上述步骤以外的附加步骤或者在执行上述步骤时执行更为具体的处理，这些附加步骤和进一步处理的内容已经参照图1至图6进行了描述，这里为了避免重复将不再进行赘述。

应注意，根据本公开示例性实施例的自动生成机器学习样本的特征的系统可完全依赖计算机程序的运行来实现相应的功能，即，各个装置与计算机程序的功能架构中与各步骤相应，使得整个系统通过专门的软件包 (例如，lib 库)而被调用，以实现相应的功能。

另一方面，根据本公开示例性实施例的自动生成机器学习样本的特征的系统所包括的各个装置也可以通过硬件、软件、固件、中间件、微代码或其任意组合来实现。当以软件、固件、中间件或微代码实现时，用于执行相应操作的程序代码或者代码段可以存储在诸如存储介质的计算机可读存储介质中，使得处理器可通过读取并运行相应的程序代码或者代码段来执行相应的操作。

例如，根据本公开示例性实施例，可提供一种包括至少一个计算装置和至少一个存储指令的存储装置的系统，其中，所述指令在被所述至少一个计算装置运行时，促使所述至少一个计算装置执行用于自动生成机器学习样本的特征的以下步骤：获取用户指定的数据表，其中，数据表的一行对应一条数据记录，数据表的一列对应一个字段；声明数据表中的各个非目标值字段所对应的特征类型；其中，所述特征类型包括离散特征，或包括连续特征，或包括离散特征和连续特征；按照声明的特征类型将各个非目标值字段处理为单位特征；基于生成的单位特征来进行特征组合，以生成组合特征；以及基于生成的单位特征和组合特征来得到机器学习样本的特征。

这里，所述系统可构成单机计算环境或分布式计算环境，其包括至少一个计算装置和至少一个存储装置，这里，作为示例，计算装置可以是通用或专用的计算机、处理器等，可以是单纯利用软件来执行处理的单元，还可以是软硬件相结合的实体。也就是说，计算装置可实现为计算机、处理器、计算单元（或模块）、客户端、主机、代理装置、服务器等。此外，存储装置可以是物理上的存储设备或逻辑上划分出的存储单元，其可与计算装置在操作上进行耦合，或者可例如通过 1/0 端口、网络连接等互相通信。

此外，例如，本公开的示例性实施例还可以实现为计算装置，该计算装置包括存储部件和处理器，存储部件中存储有计算机可执行指令集合，当所述计算机可执行指令集合被所述处理器执行时，执行自动生成机器学习样本的特征的方法。

具体说来，所述计算装置可以部署在服务器或客户端中，也可以部署在分布式网络环境中的节点装置上。此外，所述计算装置可以是 PC 计算机、平板装置、个人数字助理、智能手机、以七 应用或其他能够执行上述指令集合的装置。

这里，所述计算装置并非必须是单个的计算装置，还可以是任何能够单独或联合执行上述指令（或指令集）的装置或电路的集合体。计算装置还可以是集成控制系统或系统管理器的一部分，或者可被配置为与本地或远程（例如，经由无线传输）以接口互联的便携式电子装置。

在所述计算装置中，处理器可包括中央处理器（□□(；）、图形处理器（0?!；）、可编程逻辑装置、专用处理器系统、微控制器或微处理器。作为示例而非限制，处理器还可包括模拟处理器、数字处理器、微处理器、多核处理器、处理器阵列、网络处理器等。

根据本公开示例性实施例的自动生成机器学习样本的特征的方法中所描述的某些操作可通过软件方式来实现，某些操作可通过硬件方式来实现，此外，还可通过软硬件结合的方式来实现这些操作。

处理器可运行存储在存储部件之一中的指令或代码，其中，所述存储部件还可以存储数据。指令和数据还可经由网络接口装置而通过网络被发送和接收，其中，所述网络接口装置可采用任何已知的传输协议。

存储部件可与处理器集成为一体，例如，将 11^ 或闪存布置在集成电路微处理器等之内。此外，存储部件可包括独立的装置，诸如，外部盘驱动、存储阵列或任何数据库系统可使用的其他存储装置。存储部件和处理器可在操作上进行耦合，或者可例如通过 1/0 端口、网络连接等互相通信，使得处理器能够读取存储在存储部件中的文件。

此外，所述计算装置还可包括视频显示器 (诸如，液晶显示器) 和用户交互接口 (诸如，键盘、鼠标、触摸输入装置等)。计算装置的所有组件可经由总线和/或网络而彼此连接。

根据本公开示例性实施例的自动生成机器学习样本的特征的方法所涉及的操作可被描述为各种互联或耦合的功能块或功能示图。然而，这些功能块或功能示图可被均等地集成为单个的逻辑装置或按照非确切的边界进行操作。

根据本公开示例性实施例 ,用于自动生成机器学习样本的特征的计算装置可包括存储部件和处理器，其中，存储部件中存储有计算机可执行指令集合，当所述计算机可执行指令集合被所述处理器执行时，执行下述步骤：获取用户指定的数据表，其中，数据表的一行对应一条数据记录，数据表的一列对应一个字段；声明数据表中的各个非目标值字段所对应的特征类型，其中，特征类型包括离散特征和/或连续特征；按照声明的特征类型将各个非目标值字段处理为单位特征；基于生成的单位特征来进行特征组合，以生成组合特征；以及基于生成的单位特征和组合特征来得到机器学习样本的特征。

以上描述了本公开的各示例性实施例，应理解，上述描述仅是示例性的，并非穷尽性的，本公开不限于所披露的各示例性实施例。在不偏离本公开的范围和精神的情况下，对于本技术领域的普通技术人员来说许多修改和变更都是显而易见的。因此，本公开的保护范围应该以权利要求的范围为准。

权 利 要 求 书

1、一种由至少一个计算装置自动生成机器学习样本的特征的方法，包括：

获取用户指定的数据表，其中，数据表的一行对应一条数据记录，数据表的一列对应一个字段；

声明数据表中的各个非目标值字段所对应的特征类型；其中，所述特征类型包括离散特征，或包括连续特征，或包括离散特征和连续特征；

按照声明的特征类型将各个非目标值字段处理为单位特征；

基于生成的单位特征来进行特征组合，以生成组合特征；以及

基于生成的单位特征和组合特征来得到机器学习样本的特征。

2、根据权利要求1所述的方法，其中，所述方法通过启动与自动特征生成步骤相应的算子而自动执行。

3、根据权利要求2所述的方法，其中，所述算子对应于与机器学习流程相应的有向无环图中的节点。

4、根据权利要求3所述的方法，其中，非目标值字段通过以下方式来获取：从数据表中的所有字段中去除用户指定的目标值字段。

5、如权利要求4所述的方法，其中，所述算子在用户未指定目标值字段的情况下被启动时，提供异常提醒。

6、根据权利要求1-5中任一项所述的方法，其中，所述声明数据表中的各个非目标值字段所对应的特征类型包括：

自动或根据用户的指示，将所有非目标值字段声明为离散特征，或者，将各个非目标值字段声明为与其字段值数据类型相应的离散特征或连续特征。

7、根据权利要求1-5中任一项所述的方法，其中，所述基于生成的单位特征来进行特征组合，以生成组合特征包括：

对生成的全部单位特征进行各种组合来获取候选组合特征，或者，对生成的全部单位特征之中特征重要性较高的单位特征进行各种组合来获取候选组合特征；

通过衡量与每个候选组合特征相应的机器学习模型的效果来从候选组合特征中筛选出组合特征。

8、根据权利要求1-5中任一项所述的方法，其中，所述基于生成的单位特征和组合特征来得到机器学习样本的特征包括：

将生成的全部单位特征和全部组合特征作为机器学习样本的特征；

或者，将生成的全部单位特征和全部组合特征之中，特征重要性较高的特征作为机器学习样本的特征；

或者，将生成的全部单位特征之中特征重要性较高的单位特征和生成的全部组合特征，作为机器学习样本的特征；

或者，将生成的全部组合特征之中特征重要性较高的组合特征和生成的全部单位特征，作为机器学习样本的特征。

9、根据权利要求 1-5 中任一项所述的方法，还包括：

向用户显示得到的机器学习样本的特征。

10、根据权利要求 9 所述的方法，其中，在向用户显示得到的机器学习样本的特征时，还向用户显示每个特征的特征重要性。

11、根据权利要求 1-5 中任一项所述的方法，还包括：

直接将得到的机器学习样本的特征应用于后续的机器学习步骤。

12、根据权利要求 6 所述的方法，其中，所述按照声明的特征类型将各个非目标值字段处理为单位特征包括：

针对每一个字段值数据类型为连续型且被声明为离散特征的非目标值字段，执行一种或多种分桶运算以得到相应的一个或多个分桶特征，并将得到的分桶特征整体作为一个单位特征。

13、一种包括至少一个计算装置和至少一个存储指令的存储装置的系统，其中，所述指令在被所述至少一个计算装置运行时，促使所述至少一个计算装置执行用于自动生成机器学习样本的特征的以下步骤：

获取用户指定的数据表，其中，数据表的一行对应一条数据记录，数据表的一列对应一个字段；

声明数据表中的各个非目标值字段所对应的特征类型；其中，所述特征类型包括离散特征，或包括连续特征，或包括离散特征和连续特征；

按照声明的特征类型将各个非目标值字段处理为单位特征；

基于生成的单位特征来进行特征组合，以生成组合特征；以及

基于生成的单位特征和组合特征来得到机器学习样本的特征。

14、根据权利要求 13 所述的系统，其中，通过启动与自动特征生成步骤相应的算子来使所述系统自动执行操作。

15、根据权利要求 14 所述的系统，其中，所述算子对应于与机器学习流程相应的有向无环图中的节点。

16、根据权利要求 15 所述的系统，其中，非目标值字段通过以下方式来获取：从数据表中的所有字段中去除用户指定的目标值字段。

17、如权利要求 16 所述的系统，其中，所述指令在被所述至少一个计算装置运行时，促使所述至少一个计算装置还执行以下步骤：

在所述算子在用户未指定目标值字段的情况下被启动时，提供异常提醒。

18、根据权利要求 13-17 中任一项所述的系统，其中，所述声明数据表中的各个非目标值字段所对应的特征类型的步骤包括：

自动或根据用户的指示，将所有非目标值字段声明为离散特征，或者，将各个非目标值字段声明为与其字段值数据类型相应的离散特征或连续特征。

19、根据权利要求 13-17 中任一项所述的系统，其中，所述基于生成的单位特征来进行特征组合，以生成组合特征的步骤包括：

对生成的全部单位特征进行各种组合来获取候选组合特征，或者，对生成的全部单位特征之中特征重要性较高的单位特征进行各种组合来获取候选组合特征；

通过衡量与每个候选组合特征相应的机器学习模型的效果来从候选组合特征中筛选出组合特征。

20、根据权利要求 13-17 中任一项所述的系统，其中，所述基于生成的单位特征和组合特征来得到机器学习样本的特征的步骤包括：

将生成的全部单位特征和全部组合特征作为机器学习样本的特征；

或者，将生成的全部单位特征和全部组合特征之中，特征重要性较高的特征作为机器学习样本的特征；

或者，将生成的全部单位特征之中特征重要性较高的单位特征和生成的全部组合特征，作为机器学习样本的特征；

或者，将生成的全部组合特征之中特征重要性较高的组合特征和生成的全部单位特征，作为机器学习样本的特征。

21、根据权利要求 13-17 中任一项所述的系统，其中，所述指令在被所述至少一个计算装置运行时，促使所述至少一个计算装置还执行以下步骤：

向用户显示得到的机器学习样本的特征。

22、根据权利要求 21 所述的系统，其中，所述指令在被所述至少一个计算装置运行时，促使所述至少一个计算装置还执行以下步骤：

在向用户显示得到的机器学习样本的特征时，还向用户显示每个特征的特征重要性。

23、根据权利要求 13-17 中任一项所述的系统，其中，所述指令在被所述至少一个计算装置运行时，促使所述至少一个计算装置还执行以下步骤：

直接将得到的机器学习样本的特征应用于后续的机器学习步骤。

24、根据权利要求 18 所述的系统，其中，所述按照声明的特征类型将各个非目标值字段处理为单位特征的步骤包括：

针对每一个字段值数据类型为连续型且被声明为离散特征的非目标值字段，执行一种或多种分桶运算以得到相应的一个或多个分桶特征，并将得到的分桶特征整体作为一个单位特征。

25、一种存储指令的计算机可读存储介质，其中，当所述指令被至少一个计算装置运行时，促使所述至少一个计算装置执行如权利要求 1 至 12 中任一所述的自动生成机器学习样本的特征的方法。

26、一种用于自动生成机器学习样本的特征的系统，包括：

数据表获取装置，用于获取用户指定的数据表，其中，数据表的一行对应一条数据记录，数据表的一列对应一个字段；

声明装置，用于声明数据表中的各个非目标值字段所对应的特征类型；其中，所述特征类型包括离散特征，或包括连续特征，或包括离散特征和连续特征；

单位特征生成装置，用于按照声明的特征类型将各个非目标值字段处理为单位特征；组合特征生成装置，用于基于生成的单位特征来进行特征组合，以生成组合特征；以及

特征获取装置，用于基于生成的单位特征和组合特征来得到机器学习样本的特征。

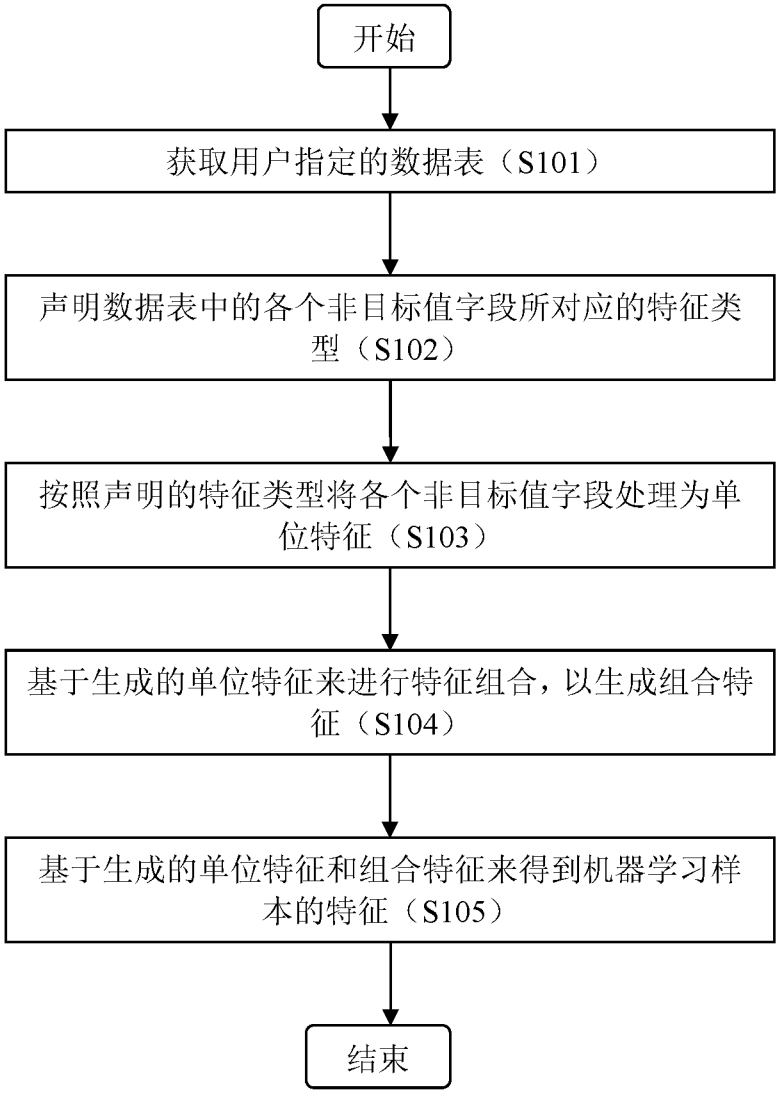
开始
获取用户指定的数据表（S101）
声明数据表中的各个非目标值字段所对应的特征类型（S102）
按照声明的特征类型将各个非目标值字段处理为单位特征（S103）
基于生成的单位特征来进行特征组合，以生成组合特征（S104）
基于生成的单位特征和组合特征来得到机器学习样本的特征（S105）
结束

图 1

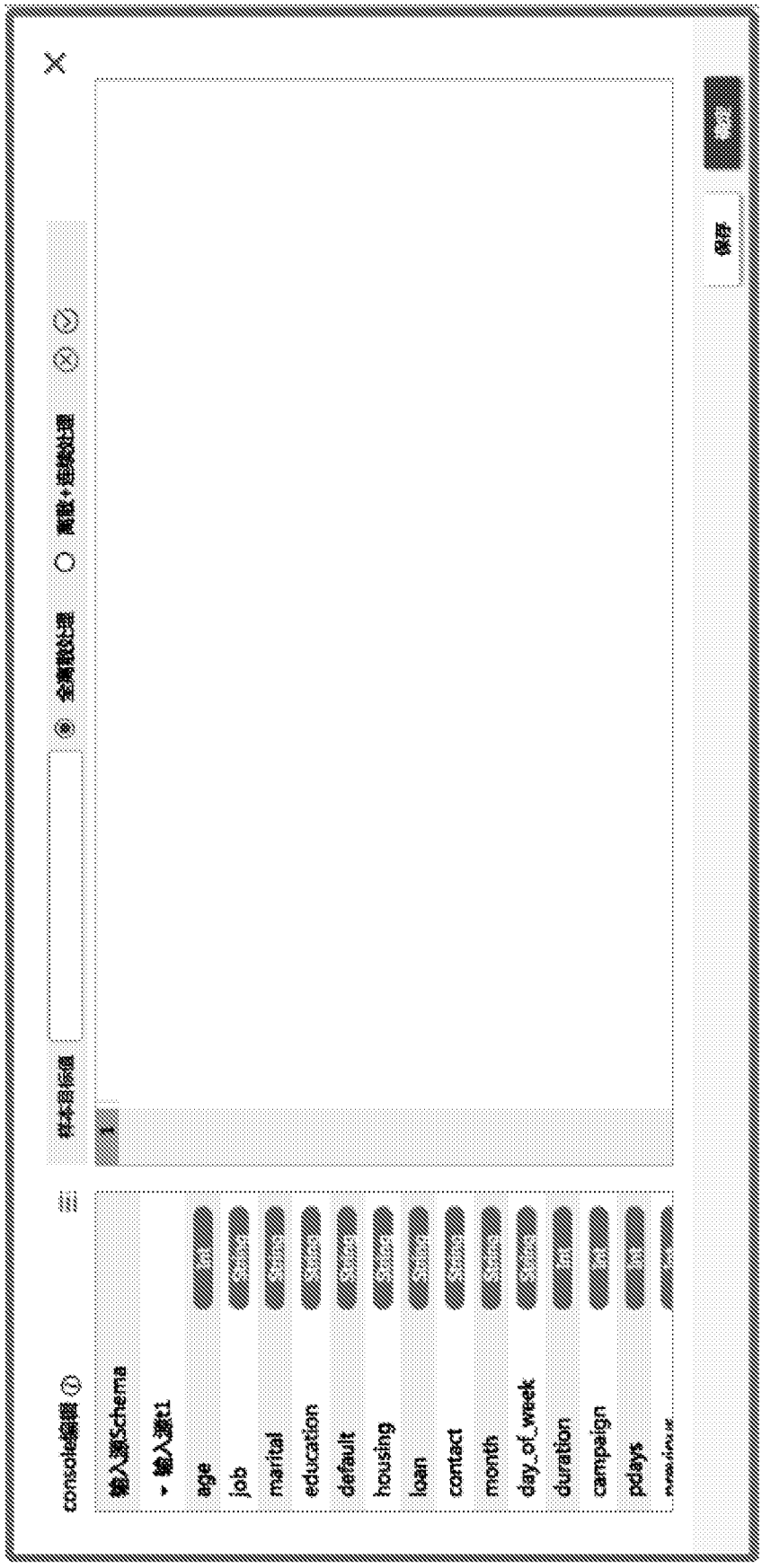
age
job
marital
education
default
housing
loan
contact
month
day_of_week
duration
campaign
pdays

图 2

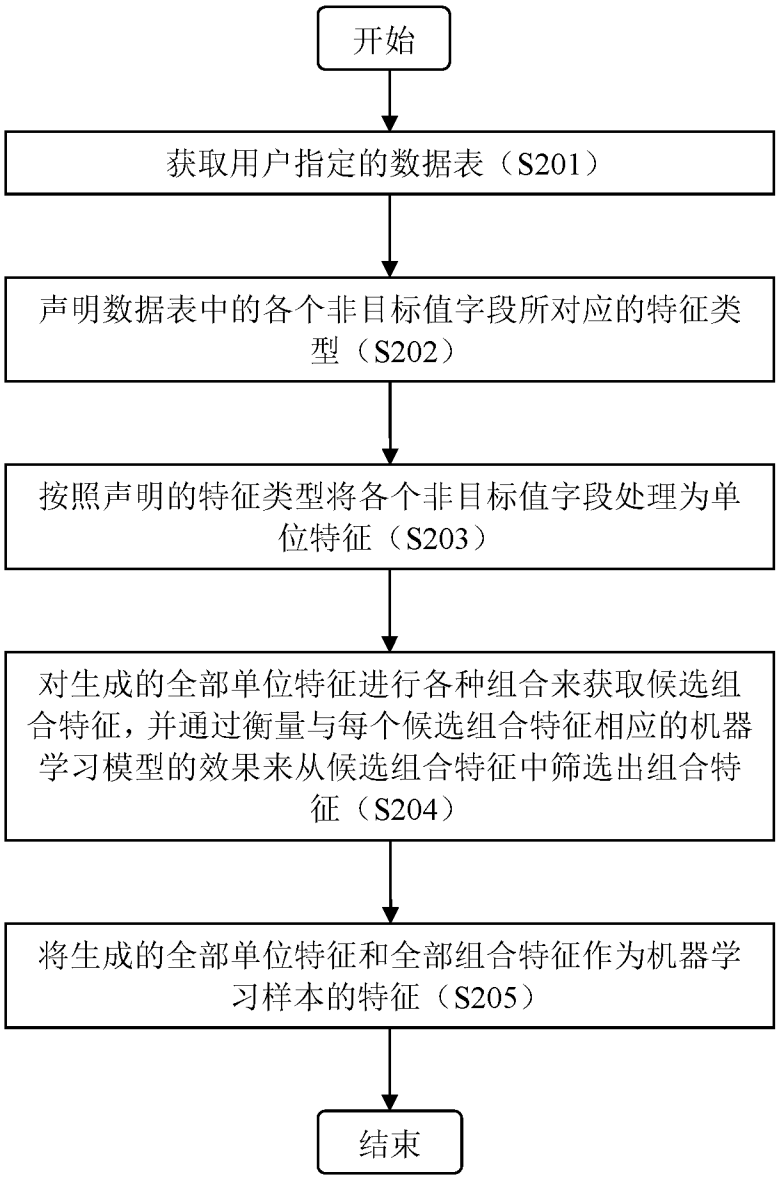
开始
获取用户指定的数据表（S201）
声明数据表中的各个非目标值字段所对应的特征类型（S202）
按照声明的特征类型将各个非目标值字段处理为单位特征（S203）
对生成的全部单位特征进行各种组合来获取候选组合特征，并通过衡量与每个候选组合特征相应的机器学习模型的效果来从候选组合特征中筛选出组合特征（S204）
将生成的全部单位特征和全部组合特征作为机器学习样本的特征（S205）
结束

图 3

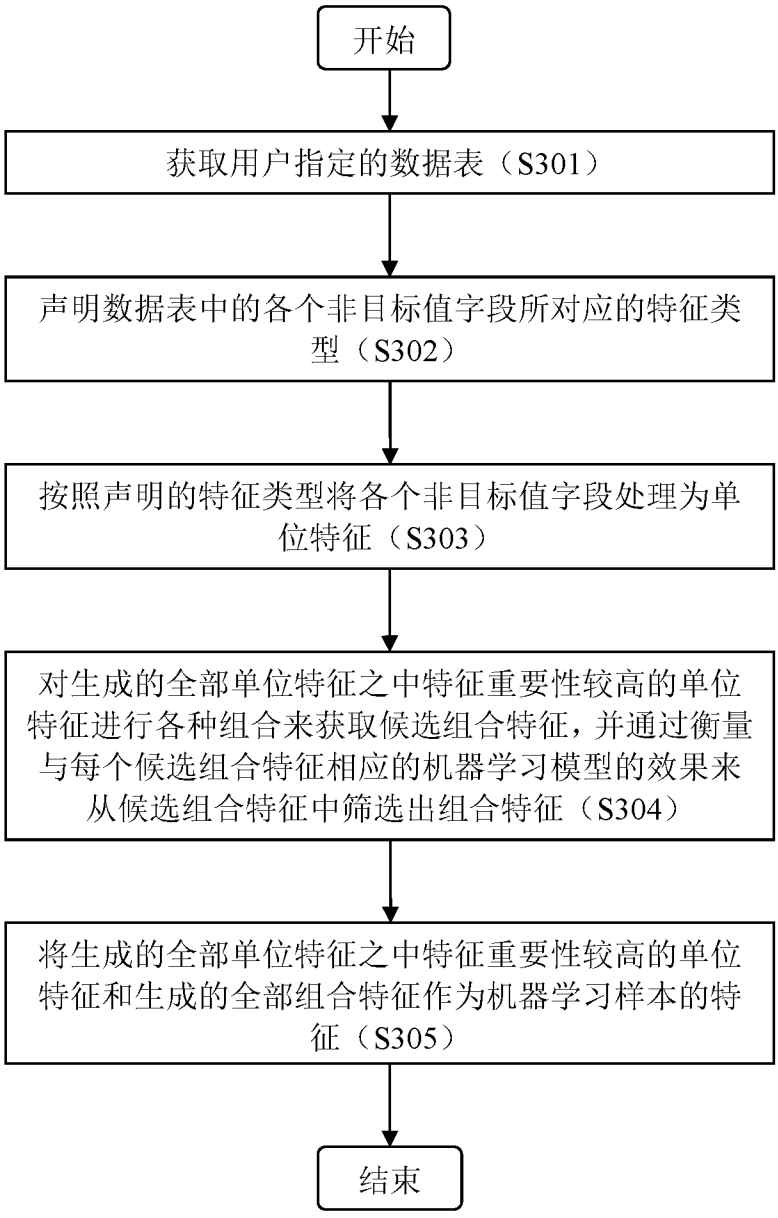
开始
获取用户指定的数据表（S301）
声明数据表中的各个非目标值字段所对应的特征类型（S302）
按照声明的特征类型将各个非目标值字段处理为单位特征（S303）
对生成的全部单位特征之中特征重要性较高的单位特征进行各种组合来获取候选组合特征，并通过衡量与每个候选组合特征相应的机器学习模型的效果来从候选组合特征中筛选出组合特征（S304）
将生成的全部单位特征之中特征重要性较高的单位特征和生成的全部组合特征作为机器学习样本的特征（S305）
结束

图 4

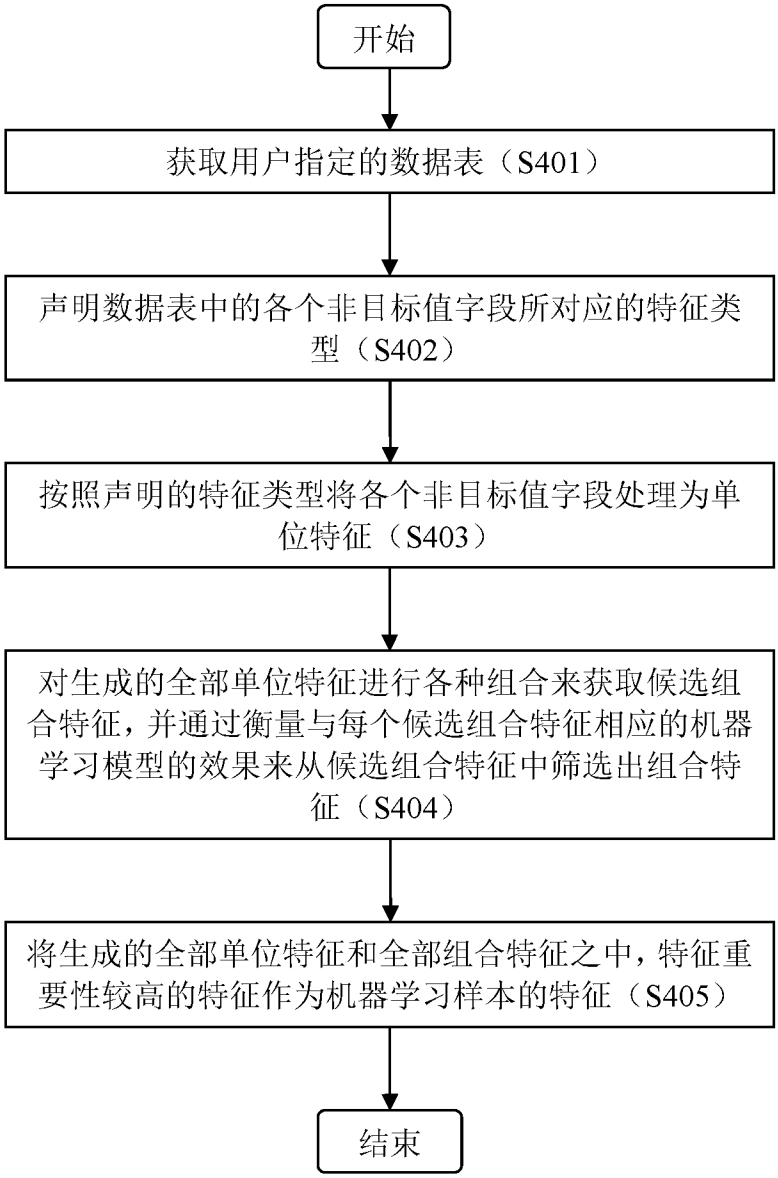
开始
获取用户指定的数据表（S401）
声明数据表中的各个非目标值字段所对应的特征类型（S402）
按照声明的特征类型将各个非目标值字段处理为单位特征（S403）
对生成的全部单位特征进行各种组合来获取候选组合特征，并通过衡量与每个候选组合特征相应的机器学习模型的效果来从候选组合特征中筛选出组合特征（S404）
将生成的全部单位特征和全部组合特征之中，特征重要性较高的特征作为机器学习样本的特征（S405）
结束

图 5

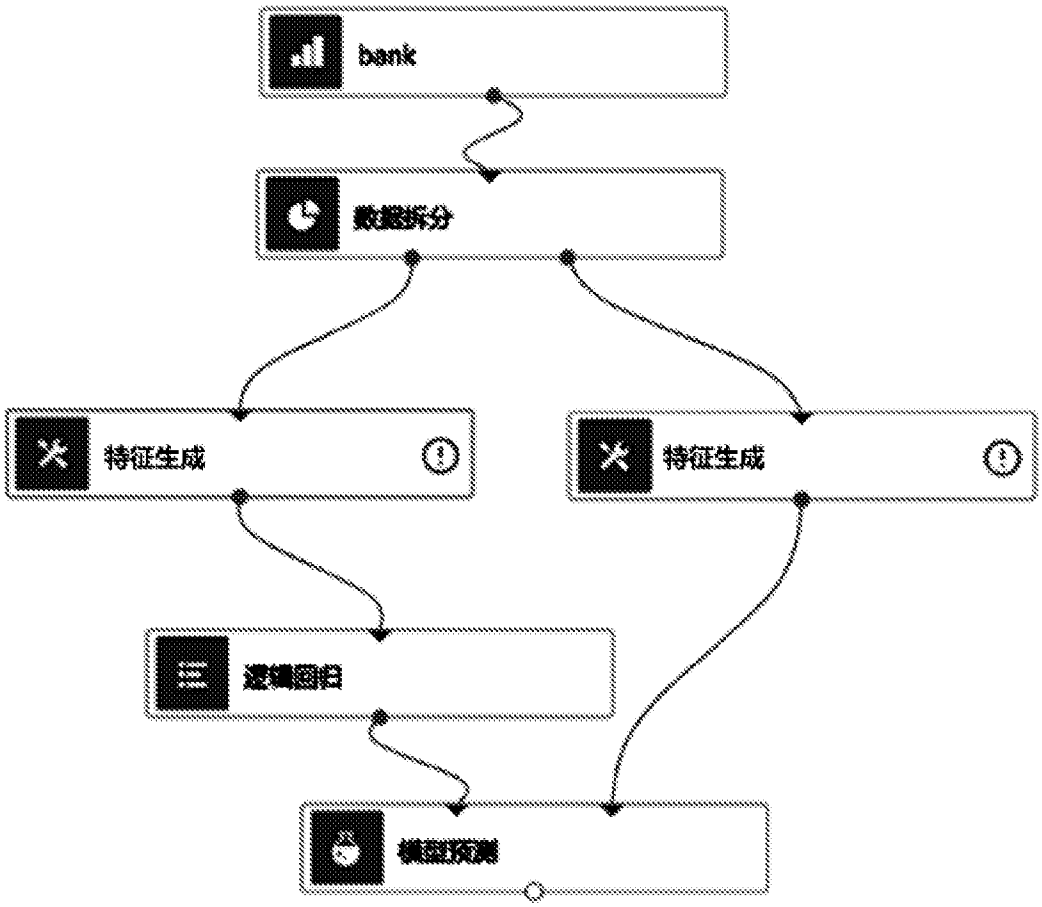
bank
特征生成
特征生成

图 6

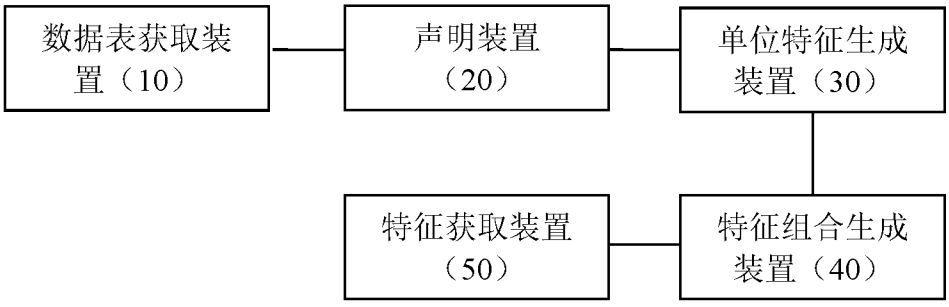
数据表获取装置（10）
声明装置（20）
单位特征生成装置（30）
特征获取装置（50）
特征组合生成装置（40）

图 7

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2018/123910

A. CLASSIFICATION OF SUBJECT MATTER

G06K 9/62(2006.01)i; G06N 99/00(2019.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06K; G06N

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNPAT, WPI, EPODOC, CNKI, IEEE: 机器,学习 样本,特征 ,记录 表 ,字段 ,离散 连续 组合 ,重要性 ,分桶 ,分箱 machine, learn+, sample, feature, record, table, field, discrete, continuous, composit+, important, binning

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 108090516 A (4PARADIGM (BEIJING) TECHNOLOGY CO., LTD.) 29 May 2018 (2018-05-29) claims 1-10, and description, paragraphs [0048]-[0079]	1-26
X	CN 107392319 A (4PARADIGM (BEIJING) TECHNOLOGY CO., LTD.) 24 November 2017 (2017-11-24) description, paragraphs [0006], [0089]-[0115] and [0138]	1-26
A	CN 107316082 A (4PARADIGM (BEIJING) TECHNOLOGY CO., LTD.) 03 November 2017 (2017-11-03) entire document	1-26
A	CN 105677353 A (BEIJING WUSI CHUANGXIANG TECHNOLOGY CO., LTD.) 15 June 2016 (2016-06-15) entire document	1-26
A	CN 107451266 A (BEIJING JINGDONG SHANGKE INFORMATION TECHNOLOGY CO., LTD.; BEIJING JINGDONG CENTURY TRADING CO., LTD.) 08 December 2017 (2017-12-08) entire document	1-26

☐ Further documents are listed in the continuation of Box C. ☑ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
06 March 2019	**27 March 2019**

Name and mailing address of the ISA/CN	Authorized officer
State Intellectual Property Office of the P. R. China **No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing** **100088** **China**	
Facsimile No. **(86-10)62019451**	Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2018/123910

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN 108090516 A	29 May 2018	None	
CN 107392319 A	24 November 2017	WO 2019015631 A1	24 January 2019
CN 107316082 A	03 November 2017	None	
CN 105677353 A	15 June 2016	None	
CN 107451266 A	08 December 2017	None	

国际检索报告

国际申请号
?01/CN2018/123910

八. 主题的分类

G06K 9/62(2006.01)i; G06N 99/00(2019.01)i

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

G06K; G06N

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称,和使用的检索词(如使用))

CNPAT, WPI, EPODOC, CNKI, IEEE: 机器,学习,样本,特征,记录,表,字段,离散,连续,组合,重要性,分桶,分箱, machine, learn+, sample, feature, record, table, field, discrete, continuous, composit+, important, binning

○. 相关文件

类 型*	引用文件，必要时，指明相关段落	相关的权利要求
PX	CN 108090516 A (第四范式北京技术有限公司) 2018年 5月 29日 (2018-05-29) 权利要求1-10，说明书第[0048]-[0079]段	1-26
X	CN 107392319 A (第四范式北京技术有限公司) 2017年 11月 24日 (2017-11-24) 说明书第[0006]，[0089]-[0115]，[0138]段	1-26
A	CN 107316082 A (第四范式北京技术有限公司) 2017年 11月 3日 (2017-11-03) 全文	1-26
A	CN 105677353 A (北京物思创想科技有限公司) 2016年 6月 15日 (2016-06-15) 全文	1-26
A	CN 107451266 A (北京京东尚科信息技术有限公司 北京京东世纪贸易有限公司) 2017年 12月 8日 (2017-12-08) 全文	1-26

☐ 其余文件在C栏的续页中列出。

☑ 见同族专利附件。

* 引用文件的具体类型:

"A" 认为不特别相关的表示了现有技术一般状态的文件

"E" 在国际申请日的当天或之后公布的在先申请或专利

"L" 可能对优先权要求构成怀疑的文件，或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件(如具体说明的)

"O" 涉及口头公开、使用、展览或其他方式公开的文件

"P" 公布日先于国际申请日但迟于所要求的优先权日的文件

"T" 在申请日或优先权日之后公布，与申请不相抵触，但为了理解发明之理论或原理的在后文件

"X" 特别相关的文件，单独考虑该文件，认定要求保护的发明不是新颖的或不具有创造性

"Y" 特别相关的文件，当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时，要求保护的发明不具有创造性

"&" 同族专利的文件

国际检索实际完成的日期	国际检索报告邮寄日期
2019年 3月 6日	2019年 3月 27日
ISA/CN的名称和邮寄地址 中国国家知识产权局(ISA/CN) 中国北京市海淀区蓟门桥西土城路6号 100088 传真号 (86-10)62019451	受权官员 李玉坤 电话号码 86-(10)-53961358

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2018/123910

检索报告引用的专利文件	公布日 (年/月/日)	同族专利	公布日 (年/月/日)
CN 108090516 A	2018年 5月 29日	无	
CN 107392319 A	2017年 11月 24日	WO 2019015631 A1	2019年 1月 24日
CN 107316082 A	2017年 11月 3日	无	
CN 105677353 A	2016年 6月 15日	无	
CN 107451266 A	2017年 12月 8日	无	