(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 110298679 A
(43)申请公布日 2019.10.01

(21)申请号 201810247666.5

(22)申请日 2018.03.23

(71)申请人 华为技术有限公司
地址 518129 广东省深圳市龙岗区坂田华为总部办公楼

(72)发明人 刘洋 蒋丰泽 赵晓东

(74)专利代理机构 北京同达信恒知识产权代理有限公司 11291
代理人 冯艳莲

(51)Int.Cl.
G06Q 30/02(2012.01)
G06K 9/62(2006.01)

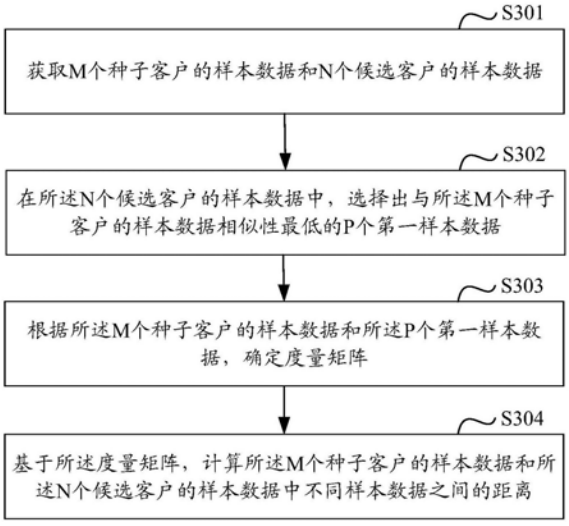
权利要求书3页 说明书18页 附图5页

(54)发明名称

一种计算样本数据之间的距离的方法及设备

(57)摘要

本申请公开了一种计算样本数据之间的距离的方法及设备，可以使计算出的不同样本数据之间的距离能够准确地体现样本数据之间的相似性。在该方案中，客户挖掘设备需要根据与种子客户样本数据相似性最低的P个候选客户的样本数据，以及种子客户的样本数据计算度量矩阵，且该度量矩阵满足：通过度量矩阵计算实际上相似的样本数据(种子客户的样本数据)之间的距离较小，实际上不相似的样本数据(P个候选客户的样本数据)之间的距离较大，显然，基于该度量矩阵计算得到的两个样本数据之间的距离可以更能体现这两个样本数据之间的相似度。



CN 110298679 A

1.一种样本数据相似性计算方法,其特征在于,包括:

获取M个种子客户的样本数据和N个候选客户的样本数据,M、N均为大于2的整数;

在所述N个候选客户的样本数据中,选择出与所述M个种子客户的样本数据相似性最低的P个第一样本数据,P为大于2的整数;

根据所述M个种子客户的样本数据和所述P个第一样本数据,确定度量矩阵;

其中,所述度量矩阵为半正定矩阵,所述度量矩阵中每个元素大于0,所述度量矩阵满足以下条件:基于所述度量矩阵计算的所述M个种子客户的样本数据中所有不同种子客户的样本数据之间的距离之和最小,且基于所述度量矩阵计算的所述P个第一样本数据中所有不同第一样本数据之间的距离之和大于设定距离阈值;

基于所述度量矩阵,计算所述M个种子客户的样本数据和所述N个候选客户的样本数据中不同样本数据之间的距离。

2.如权利要求1所述的方法,其特征在于,基于所述度量矩阵计算的不同样本数据之间的距离满足以下公式:

$$d(x,y)=\sqrt{(x-y)^{T}A(x-y)}$$

其中,d(x,y)为所述不同样本数据之间的距离,x为所述不同样本数据中的一个样本数据构成的向量,y为所述不同样本数据中的另一个样本数据构成的向量,A为所述度量矩阵。

3.如权利要求1或2所述的方法,其特征在于,在所述N个候选客户的样本数据中,选择出与所述M个种子客户的样本数据相似性最低的P个第一样本数据,包括:

计算所述M个种子客户的样本数据和所述N个候选客户的样本数据中不同样本数据之间的欧氏距离;

根据计算得到的不同样本数据之间的欧氏距离,以及预设的客户挖掘算法,确定所述N个候选客户的样本数据中每个候选客户的样本数据与所述M个种子客户的样本数据之间的相似性参数;

根据确定的每个候选客户的样本数据与所述M个种子客户的样本数据之间的相似性参数,在所述N个候选客户的样本数据中,选择所述P个第一样本数据。

4.如权利要求1-3任一项所述的方法,其特征在于,在计算所述M个种子客户的样本数据和所述N个候选客户的样本数据中不同样本数据之间的距离之后,所述方法还包括:

根据计算得到的不同样本数据之间的距离,以及预设的客户挖掘算法,确定所述N个候选客户的样本数据中每个候选客户的样本数据与所述M个种子客户的样本数据之间的相似性参数;

判断预设的停止迭代计算的条件是否满足;

若满足,则根据确定的每个候选客户的样本数据与所述M个种子客户的样本数据之间的相似性参数,在所述N个候选客户的样本数据中,选择出与所述M个种子客户的样本数据相似性最高的Q个第二样本数据,其中,Q为大于1的整数;

若不满足,则根据确定的每个候选客户的样本数据与所述M个种子客户的样本数据之间的相似性,在N个候选客户的样本数据中,选择新的P个第一样本数据;根据所述M个种子客户的样本数据和所述新的P个第一样本数据,确定新的度量矩阵;基于所述新的度量矩阵,重新计算所述M个种子客户的样本数据和所述N个候选客户的样本数据中不同样本数据

之间的距离;以及根据重新计算得到的不同样本数据之间的距离,以及所述客户挖掘算法,确定所述N个候选客户的样本数据中每个候选客户的样本数据与所述M个种子客户的样本数据之间的相似性参数,直至所述停止迭代计算的条件满足。

5.如权利要求4所述的方法,其特征在于,所述停止迭代计算的条件为以下至少一项:

迭代计算的次数达到设定次数;

所述N个候选客户的样本数据中至少一个候选客户的样本数据与所述M个种子客户的样本数据之间的相似性参数大于设定第一相似性阈值;

所述N个候选客户的样本数据中每个候选客户的样本数据与所述M个种子客户的样本数据之间的相似性参数大于设定第二相似性阈值。

6.如权利要求3-5任一项所述的方法,其特征在于,所述客户挖掘算法为密度传播算法。

7.一种客户挖掘设备,其特征在于,包括:

获取单元,用于获取M个种子客户的样本数据和N个候选客户的样本数据,M、N均为大于2的整数;

确定单元,用于在所述N个候选客户的样本数据中,选择出与所述M个种子客户的样本数据相似性最低的P个第一样本数据,P为大于2的整数;并根据所述M个种子客户的样本数据和所述P个第一样本数据,确定度量矩阵;

其中,所述度量矩阵为半正定矩阵,所述度量矩阵中每个元素大于0,所述度量矩阵满足以下条件:基于所述度量矩阵计算的所述M个种子客户的样本数据中所有不同种子客户的样本数据之间的距离之和最小,且基于所述度量矩阵计算的所述P个第一样本数据中所有不同第一样本数据之间的距离之和大于设定距离阈值;

处理单元,用于基于所述度量矩阵,计算所述M个种子客户的样本数据和所述N个候选客户的样本数据中不同样本数据之间的距离。

8.如权利要求7所述的客户挖掘设备,其特征在于,基于所述度量矩阵计算的不同样本数据之间的距离满足以下公式:

$$d(x,y)=\sqrt{(x-y)^{T}A(x-y)}$$

其中,d(x,y)为所述不同样本数据之间的距离,x为所述不同样本数据中的一个样本数据构成的向量,y为所述不同样本数据中的另一个样本数据构成的向量,A为所述度量矩阵。

9.如权利要求7或8所述的客户挖掘设备,其特征在于,所述确定单元,具体用于:

计算所述M个种子客户的样本数据和所述N个候选客户的样本数据中不同样本数据之间的欧氏距离;

根据计算得到的不同样本数据之间的欧氏距离,以及预设的客户挖掘算法,确定所述N个候选客户的样本数据中每个候选客户的样本数据与所述M个种子客户的样本数据之间的相似性参数;

根据确定的每个候选客户的样本数据与所述M个种子客户的样本数据之间的相似性参数,在所述N个候选客户的样本数据中,选择所述P个第一样本数据。

10.如权利要求7-9任一项所述的客户挖掘设备,其特征在于,所述处理单元,还用于:

在计算所述M个种子客户的样本数据和所述N个候选客户的样本数据中不同样本数据之间的距离之后,根据计算得到的不同样本数据之间的距离,以及预设的客户挖掘算法,确

定所述N个候选客户的样本数据中每个候选客户的样本数据与所述M个种子客户的样本数据之间的相似性参数；

判断预设的停止迭代计算的条件是否满足；

若满足，则根据确定的每个候选客户的样本数据与所述M个种子客户的样本数据之间的相似性参数，在所述N个候选客户的样本数据中，选择出与所述M个种子客户的样本数据相似性最高的Q个第二样本数据，其中，Q为大于1的整数；

若不满足，则根据确定的每个候选客户的样本数据与所述M个种子客户的样本数据之间的相似性，在N个候选客户的样本数据中，选择新的P个第一样本数据；根据所述M个种子客户的样本数据和所述新的P个第一样本数据，确定新的度量矩阵；基于所述新的度量矩阵，重新计算所述M个种子客户的样本数据和所述N个候选客户的样本数据中不同样本数据之间的距离；以及根据重新计算得到的不同样本数据之间的距离，以及所述客户挖掘算法，确定所述N个候选客户的样本数据中每个候选客户的样本数据与所述M个种子客户的样本数据之间的相似性参数，直至所述停止迭代计算的条件满足。

11.如权利要求10所述的客户挖掘设备，其特征在于，所述停止迭代计算的条件为以下至少一项：

迭代计算的次数达到设定次数；

所述N个候选客户的样本数据中至少一个候选客户的样本数据与所述M个种子客户的样本数据之间的相似性参数大于设定第一相似性阈值；

所述N个候选客户的样本数据中每个候选客户的样本数据与所述M个种子客户的样本数据之间的相似性参数大于设定第二相似性阈值。

12.如权利要求9-11任一项所述的客户挖掘设备，其特征在于，所述客户挖掘算法为密度传播算法。

13.一种计算机程序，其特征在于，当所述计算机程序在计算机上运行时，使得所述计算机执行权利要求1-6任一项提供的方法。

14.一种计算机存储介质，其特征在于，所述计算机存储介质中存储有计算机指令，所述计算机指令被计算机执行时，使得所述计算机执行权利要求1-6任一项提供的方法。

一种计算样本数据之间的距离的方法及设备

技术领域

[0001] 本申请涉及计算机技术领域，尤其涉及一种计算样本数据之间的距离的方法及设备。

背景技术

[0002] 实现精准营销是各类产品推销商推广其产品的目标。目前，产品推销商可以依托现代信息技术手段建立客户挖掘系统，利用客户挖掘系统对多个候选客户的样本数据进行分析，从而在所述多个候选客户中挖掘出潜在客户，实现精准营销。

[0003] 目前的客户挖掘系统一般是相似性(lookalike)算法实现的。仅需用户提供多个候选客户的样本数据，以及种子客户的样本数据，所述客户挖掘系统即可在所述多个候选客户中选择出潜在客户。其中，种子客户为一定会使用待推销产品的客户。

[0004] 传统的lookalike算法中需要采用K最近邻(K nearest neighbor，KNN)算法计算与某个样本数据最相似的K个样本数据，且目前常用欧氏距离计算不同样本数据之间的相似性。然而，在实际场景中，样本数据均具有多个维度的特征，且不同维度的特征的数据稀疏性可能较强。例如，某通信运行商开展流量包营销活动，那么每个样本数据如表1所示：

[0005] 表1

[0006]

候选客户	是否开通4G服务	是否为VIP客户	(归一化的)月平均流量
客户1	1	0	0.2
客户2	1	1	0.8
客户3	0	1	0.3

[0007] 其中，“是否开通4G服务”、“是否为VIP”两个特征是布尔型，取值为1或0；“月平均流量”为数值型(归一化后的取值范围为[0,1])。

[0008] 客户1的样本数据和客户2的样本数据之间的欧氏距离为：

[0009] $$S_{12}=\sqrt{(1-1)^2+(0-1)^2+(0.2-0.8)^2}=1.17$$

[0010] 客户1的样本数据和客户3的样本数据之间的欧氏距离为：

[0011] $$S_{13}=\sqrt{(1-0)^2+(0-1)^2+(0.2-0.3)^2}=1.42$$

[0012] 客户2的样本数据和客户3的样本数据之间的欧氏距离为：

[0013] $$S_{23}=\sqrt{(1-0)^2+(1-1)^2+(0.8-0.3)^2}=1.12$$

[0014] 通过不同客户的样本数据之间的欧氏距离可知，客户2的样本数据和客户3的样本数据之间的相似性最高，而客户1的样本数据和客户3的样本数据之间的相似性最低。

[0015] 从不同客户的样本数据之间的欧氏距离的计算结果可以看出，“是否开通4G服务”、“是否为VIP”这两个特征值的对计算欧氏距离的贡献更高。显然，由于上述三个特征的量纲以及类型不同，计算得到的相似性结果被布尔型的特征主导了，但是在实际业务中，月平均流量对客户挖掘的结果影响也较大。

[0016] 因此，由于样本数据的维度以及稀疏性等不确定因素，不同样本数据之间的欧氏距离无法准确地体现不同样本数据之间的相似性，进而导致客户挖掘算法计算的候选客户的样本数据与种子客户的样本数据之间的相似性不准确。

发明内容

[0017] 本申请提供一种计算样本数据之间的距离的方法及设备，可以使计算出的不同样本数据之间的距离能够准确地体现样本数据之间的相似性。

[0018] 第一方面，本申请实施例提供了一种样本数据相似性计算方法，该方法可以适用于客户挖掘设备，该方法包括以下步骤：

[0019] 所述客户挖掘设备获取M个种子客户的样本数据和N个候选客户的样本数据，M、N均为大于2的整数；然后，所述客户挖掘设备在所述N个候选客户的样本数据中，选择出与所述M个种子客户的样本数据相似性最低的P个第一样本数据，P为大于2的整数；并根据所述M个种子客户的样本数据和所述P个第一样本数据，确定度量矩阵；最终，所述客户挖掘设备基于所述度量矩阵，计算所述M个种子客户的样本数据和所述N个候选客户的样本数据中不同样本数据之间的距离。

[0020] 其中，所述度量矩阵为半正定矩阵，所述度量矩阵中每个元素大于0，所述度量矩阵满足以下条件：基于所述度量矩阵计算的所述M个种子客户的样本数据中所有不同种子客户的样本数据之间的距离之和最小，且基于所述度量矩阵计算的所述P个第一样本数据中所有不同第一样本数据之间的距离之和大于设定距离阈值。

[0021] 在上述方法中，由于度量矩阵满足以下条件：基于所述度量矩阵计算的所述M个种子客户的样本数据中所有不同种子客户的样本数据(即实际上相似的样本数据)之间的距离之和最小，且基于所述度量矩阵计算的所述P个第一样本数据中所有不同第一样本数据(即实际上不相似的样本数据)之间的距离之和大于设定阈值。该条件说明通过度量矩阵计算实际上相似的样本数据之间的距离较小，而实际上不相似的样本数据之间的距离较大，显然，基于该度量矩阵计算得到的两个样本数据之间的距离可以更能体现这两个样本数据之间的相似度。

[0022] 在一个可能的设计中，基于所述度量矩阵计算的不同样本数据之间的距离满足以下公式：

[0023] $$d(x,y)=\sqrt{(x-y)^{T}A(x-y)}$$

[0024] 其中，d(x,y)为所述不同样本数据之间的距离，x为所述不同样本数据中的一个样本数据构成的向量，y为所述不同样本数据中的另一个样本数据构成的向量，A为所述度量矩阵。

[0025] 通过该设计，所述客户挖掘设备可以基于该度量矩阵计算得到更能体现这两个样本数据之间的相似度距离算法。

[0026] 在一个可能的设计中，所述客户挖掘设备可以通过以下步骤，选择出与所述M个种子客户的样本数据相似性最低的P个第一样本数据：

[0027] 所述客户挖掘设备计算所述M个种子客户的样本数据和所述N个候选客户的样本数据中不同样本数据之间的欧氏距离；并根据计算得到的不同样本数据之间的欧氏距离，以及预设的客户挖掘算法，确定所述N个候选客户的样本数据中每个候选客户的样本数据

与所述M个种子客户的样本数据之间的相似性参数;以及根据确定的每个候选客户的样本数据与所述M个种子客户的样本数据之间的相似性参数,在所述N个候选客户的样本数据中,选择所述P个第一样本数据。

[0028] 通过上述设计,所述客户挖掘设备可以通过欧氏距离的距离算法,得到P个第一样本数据。

[0029] 第二方面,本申请实施例还提供了一种客户挖掘方法,该客户挖掘方法中包含第一方面的样本数据相似性计算方法中的步骤,该方法在计算所述M个种子客户的样本数据和所述N个候选客户的样本数据中不同样本数据之间的距离之后,还包括以下步骤:

[0030] 所述客户挖掘设备根据计算得到的不同样本数据之间的距离,以及预设的客户挖掘算法,确定所述N个候选客户的样本数据中每个候选客户的样本数据与所述M个种子客户的样本数据之间的相似性参数;然后,所述客户挖掘设备判断预设的停止迭代计算的条件是否满足;若满足,则根据确定的每个候选客户的样本数据与所述M个种子客户的样本数据之间的相似性参数,在所述N个候选客户的样本数据中,选择出与所述M个种子客户的样本数据相似性最高的Q个第二样本数据,其中,Q为大于1的整数;若不满足,则根据确定的每个候选客户的样本数据与所述M个种子客户的样本数据之间的相似性,在N个候选客户的样本数据中,选择新的P个第一样本数据;根据所述M个种子客户的样本数据和所述新的P个第一样本数据,确定新的度量矩阵;基于所述新的度量矩阵,重新计算所述M个种子客户的样本数据和所述N个候选客户的样本数据中不同样本数据之间的距离;以及根据重新计算得到的不同样本数据之间的距离,以及所述客户挖掘算法,确定所述N个候选客户的样本数据中每个候选客户的样本数据与所述M个种子客户的样本数据之间的相似性参数,直至所述停止迭代计算的条件满足。

[0031] 通过上述方法,所述客户挖掘设备可以通过多次度量优化,使基于最终确定的度量矩阵计算得到的两个样本数据之间的距离体现这两个样本数据之间的相似度的准确性更高,这样,所述客户挖掘设备根据基于该度量矩阵计算的不同样本数据之间的距离以及预设的客户挖掘算法,提高最终确定的每个候选客户的样本数据与所述M个种子客户的样本数据之间的相似性参数的准确度,进而提高挖掘的潜在客户的准确度,实现精准营销。

[0032] 在一个可能的设计中,所述停止迭代计算的条件为以下至少一项:

[0033] 迭代计算的次数达到设定次数;

[0034] 所述N个候选客户的样本数据中至少一个候选客户的样本数据与所述M个种子客户的样本数据之间的相似性参数大于设定第一相似性阈值;

[0035] 所述N个候选客户的样本数据中每个候选客户的样本数据与所述M个种子客户的样本数据之间的相似性参数大于设定第二相似性阈值。

[0036] 通过该设计,所述客户挖掘设备可以实现多次度量优化,使基于最终确定的度量矩阵计算得到的两个样本数据之间的距离体现这两个样本数据之间的相似度的准确性更高。

[0037] 在一个可能的设计中,所述客户挖掘算法为密度传播算法。

[0038] 第三方面,本申请实施例提供了一种客户挖掘设备,包括用于执行以上第一方面或第二方面各个步骤的单元或模块。

[0039] 第四方面,本申请实施例还提供了一种客户挖掘设备,包括处理器和存储器,所述

处理器用于调用并执行所述存储器中存储的程序指令，通过实现以上第一方面或第二方面提供的方法。

[0040] 第五方面，本申请提供一种客户挖掘设备，包括用于执行以上第一方面或第二方面的方法的至少一个处理元件(或芯片)。

[0041] 第六方面，本申请提供一种计算机程序，该计算机程序在计算机上运行时，使得所述计算机执行以上第一方面或第二方面的方法。

[0042] 第七方面，本申请提供一种程序产品，例如计算机可读存储介质，包括第六方面的程序。

[0043] 第八方面，本申请提供一种芯片，所述芯片用于读取并执行存储器中存储的计算机程序，以实现以上第一方面或第二方面的方法。

[0044] 第九方面，本申请实施例提供了一种芯片系统，该芯片系统包括处理器，用于支持客户挖掘设备实现上述第一方面或第二方面中所涉及的功能。在一种可能的设计中，所述芯片系统还包括存储器，所述存储器，用于保存该设备必要的程序指令和数据。该芯片系统，可以由芯片构成，也可以包含芯片和其他分立器件。

附图说明

[0045] 图1为本申请实施例提供的一种客户挖掘设备的结构示意图；

[0046] 图2为本申请实施例提供的一种客户挖掘设备的结构图；

[0047] 图3为本申请实施例提供的一种计算样本数据之间的距离的方法流程图；

[0048] 图4为本申请实施例提供的一种客户挖掘方法流程图；

[0049] 图5为本申请实施例提供的一种客户挖掘方法的示例流程图；

[0050] 图6为本申请实施例提供的一种客户挖掘设备的结构图。

具体实施方式

[0051] 本申请提供一种计算样本数据之间的距离的方法及设备，可以使计算出的不同样本数据之间的距离能够准确地体现样本数据之间的相似性。其中，方法和装置是基于同一发明构思的，由于方法及装置解决问题的原理相似，因此装置与方法的实施可以相互参见，重复之处不再赘述。

[0052] 本申请实施例提供的方案中，客户挖掘设备需要在N个候选客户的样本数据中选择P个与M个种子客户的样本数据相似性最低的P个第一样本数据，然后基于选择的P个第一样本数据和M个种子客户的样本数据确定用于计算不同样本数据之间距离的度量矩阵。由于度量矩阵满足以下条件：基于所述度量矩阵计算的所述M个种子客户的样本数据中所有不同种子客户的样本数据(即实际上相似的样本数据)之间的距离之和最小，且基于所述度量矩阵计算的所述P个第一样本数据中所有不同第一样本数据(即实际上不相似的样本数据)之间的距离之和大于设定阈值。该条件说明通过度量矩阵计算实际上相似的样本数据之间的距离较小，而实际上不相似的样本数据之间的距离较大，显然，基于该度量矩阵计算得到的两个样本数据之间的距离可以更能体现这两个样本数据之间的相似度。

[0053] 以下，对本申请中的部分用语进行解释说明，以便于本领域技术人员理解。

[0054] 1)、种子客户，为接受待推销产品的概率为100％的客户，或已经使用该待推销产

品的客户。

[0055] 2)、候选客户，为未使用待推销产品的客户。

[0056] 3)、样本数据，为客户的、与待推销产品相关的多个特征的取值。

[0057] 4)、度量矩阵，用于计算不同样本数据之间的距离。其中，每个样本数据均具有多个特征。可选的，基于所述度量矩阵计算的两个样本数据之间的距离可以但不限于满足以下公式：

[0058] $$d(x,y) = \|x-y\|_A = \sqrt{(x-y)^T A(x-y)}$$

[0059] 其中，d(x,y)为该两个样本数据之间的距离，x为所述两个样本数据中的一个样本数据构成的向量，y为所述两个样本数据中的另一个样本数据构成的向量，A为所述度量矩阵。

[0060] 需要说明的是，本申请中所涉及的多个，是指两个或两个以上。

[0061] 另外，需要理解的是，在本申请的描述中，“第一”、“第二”等词汇，仅用于区分描述的目的，而不能理解为指示或暗示相对重要性，也不能理解为指示或暗示顺序。

[0062] 5)、客户挖掘算法，用于计算候选客户的样本数据与种子客户的样本数据之间的相似性，从而通过计算结果，在候选客户的样本数据中，选择出与种子客户的样本数据之间的相似性最高的Q个候选客户的样本数据，通过该客户挖掘算法可以确定该Q个候选客户的样本数据所对应的Q个候选客户为挖掘出的潜在客户。

[0063] 客户挖掘算法，又可以称为lookalike算法。可选的，该客户挖掘算法可以是基于密度传播(density propagation，DP)模型实现的相似客户扩展方案。

[0064] 下面先对客户挖掘设备基于DP模型进行训练的过程进行说明：

[0065] 1、训练样本包含第一集合S和第二集合U。其中，S中包含的M个种子客户的样本数据，S_i表示S中的一个种子客户的样本数据，i为大于1且小于M的整数，U中包含N个候选客户的样本数据，U_j表示S中的一个候选客户的样本数据，j为大于1且小于N的整数。

[0066] 2、所述客户挖掘设备分别对第一集合S和第二集合U中的每个样本数据的密度进行初始化。其中，第一集合S中每个种子客户的样本数据的密度均设置为E_S，第二集合U中每个候选客户的样本数据的密度均设置为E_U。

[0067] 3、所述客户挖掘设备将第一集合S中包含的(种子客户的)样本数据和第二集合U中的包含的(候选客户的)样本数据组合为第三集合V，此时V中包含的样本数据的数量为M+N；所述客户挖掘设备确定该第三集合V中所述M+N个样本数据中不同样本数据之间的距离(例如传统算法中的欧氏距离)。

[0068] 4、所述客户挖掘设备根据得到的不同样本数据之间的距离，针对所述第一集合S中的每个样本数据S_i，在第三集合S+U中选取距离S_i最近的K个近邻；并将K个近邻中每个样本数据的密度的1/K传播给该S_i，此时，该S_i最终的密度ES(S_i)为该S_i的初始密度E_S与该K个近邻中每个样本数据的密度的1/K之和。本次密度传播后该K个近邻中每个样本数据的密度不变。

[0069] 5、所述客户挖掘设备根据得到的不同样本数据之间的距离，针对所述第二集合U中的每个样本数据U_j，在所述第三集合S+U中选取距离U_j最近的K个近邻；

[0070] 5a、针对第二集合U中的每个样本数据U_j，所述客户挖掘设备先执行以下过程，直至所述第二集合U中的每个样本数据均进行了第一次密度传播：

[0071] 所述客户挖掘设备将U_j的K个近邻中的每个种子客户的样本数据的当前最终密度$ES(S_i)$的1/K传播给该U_j，此时，该U_j更新后的密度$ES(U_j)$为该U_j的初始密度E_U与该K个近邻中每个种子客户的样本数据的最终密度的1/K之和。本次密度传播后，该U_j的K个近邻中每个种子客户的样本数据的最终密度不变。

[0072] 5b、上述步骤执行结束后，针对第二集合U中的每个样本数据U_j，所述客户挖掘设备再执行以下过程，直至所述第二集合U中的每个样本数据均进行了第二次密度传播：

[0073] 所述客户挖掘设备将U_j的K个近邻中的每个候选客户的样本数据更新后的密度的1/K传播给该U_j，此时，该U_j再次更新的密度$EU(U_j)$为该U_j更新后的密度$ES(U_j)$与该K个近邻中每个候选客户的样本数据更新后的密度的1/K之和。需要说明的是，针对所述第二集合U中的一个样本数据执行本次密度传播后，该样本数据的K个近邻中每个候选客户的样本数据更新后的密度不变，直至针对该K个近邻中每个候选客户的样本数据执行上述过程时密度才会发生变化。

[0074] 经过上述两个步骤后，所述第二集合U中的每个样本数据U_j的最终密度$EF(U_j)$满足以下公式：

[0075] $$EF(U_j)=\left(\frac{E_S}{\tanh(E_S*K)}\right)*\tanh(E_{u+}ES(U_j)+EU(U_j))$$

[0076] 其中，每个样本数据的密度可以体现该样本数据与种子客户的样本数据之间的相似性。

[0077] 通过以上训练过程进行密度传播后，若某个种子客户的样本数据的密度升高，则说明该种子客户的样本数据周围种子客户的样本数据的数量较多，即该种子客户的样本数据与其他种子客户的样本数据之间的相似性较高，因此，该种子客户属于实际的种子客户的准确率越高。

[0078] 同理，若某个候选用户的样本数据的密度升高，则说明该候选用户的样本数据周围种子客户的样本数据的数量较多，即该候选用户的样本数据与种子客户的样本数据之间的相似性较高，由此可以推断，对该候选客户进行产品营销成功的概率较高，显然，该候选客户属于潜在客户。

[0079] 在实际应用中，客户挖掘设备可以根据对训练样本进行多个迭代计算，并按照最后计算得到的每个候选客户的样本数据的密度从高到低进行排序，然后选择排序在前的设定数量的候选客户的样本数据，或者选择密度大于设定阈值的候选客户的样本数据，最终将选择出的候选客户的样本数据所对应的候选客户作为潜在客户。其中，所述设定数量或所述设定阈值可以根据实际的业务需求进行设定。

[0080] 通过以上训练的过程可以看出，为了保证挖掘的潜在客户更准确，实现精准营销，那么，在密度传播过程中，如何准确的确定出与某个样本数据的相似性最高的K个近邻是急需解决的重点问题，即需要所述客户挖掘设备确定不同样本数据之间的距离可以准确地体现出不同样本数据之间的相似性。然而，由于样本数据的维度以及稀疏性等不确定因素，不同样本数据之间的欧氏距离，已经无法准确地体现出不同样本数据之间的相似性，因此，研究可以准确地体现不同样本数据之间的相似性的距离计算方法，是数据分析领域技术人员亟待解决的问题。

[0081] 下面结合附图对本申请实施例做进行具体说明。

[0082] 图1示出了本申请实施例提供的计算样本数据之间的距离的方法以及客户挖掘方法适用的一种可能的客户挖掘设备。该客户挖掘设备可以部署在支持大数据平台预测分析服务的云平台中，并可以通过Hadoop或Spark计算框架实现。参阅图1所示，按照逻辑功能划分，所述客户挖掘设备可以包括以下模块：样本数据库101、客户挖掘服务模块102、个性化推荐模块103，以及算子库104。

[0083] 样本数据库101，用于存储客户挖掘服务模块102在数据挖掘过程中需要使用的种子客户的样本数据和候选客户的样本数据。其中，种子客户的样本数据和候选客户的样本数据中包含相同维度的特征。可选的，一般样本数据中包含多维度的特征，且不同维度的特征之间的稀疏性较高。

[0084] 客户挖掘服务模块102，主要与实现客户挖掘服务。具体的，客户挖掘服务模块102读取样本数据库101中的样本数据(包括种子客户的样本数据和候选客户的样本数据)，并计算不同样本数据之间的距离；以及根据计算得到的不同样本数据之间的距离，以及预设的客户挖掘算法，确定每个候选客户的样本数据与种子客户的样本数据之间的相似性参数。

[0085] 其中，本申请并不对所述客户挖掘服务模块102使用的客户挖掘算法进行限定。该客户挖掘算法可以为传统的或未来各种通过计算不同样本数据之间的距离，确定样本数据之间的相似性参数的算法。例如，该客户挖掘算法可以通过DP模型实现；通过以上对DP模型的训练过程的描述可知，在该情况下，每个候选客户的样本数据与种子客户的样本数据之间的相似性参数为训练后该候选客户的样本数据的最终密度。

[0086] 算子库104，用于对客户挖掘服务模块102进行除客户挖掘算法以外的其他算法的算法支持。如图可知，算子库中可以包括至少一个通用的算法，例如异常值处理算法，缺失值处理算法，归一化算法，特征选择算法等等。

[0087] 其中，异常值处理算法用于对所述客户挖掘服务模块102读取的某些样本数据中的异常值进行处理。所述缺失值处理算法，用于对所述客户挖掘服务模块102读取的某些样本数据中的缺失值进行处理。所述归一化算法，用于对所述客户服务模块102读取的所有样本数据中的某特征的数据进行归一化处理。所述特征选择算法用于所述客户服务模块102选择通过读取的样本数据的哪些特征的数据计算不同样本数据之间的距离。

[0088] 个性化推荐模块103，用于与客户挖掘设备的进行交互，从而实现让用户可以设置个性化推荐配置信息，例如，选择潜在客户的数量，和/或，选择潜在客户的条件等。所述个性化推荐模块103还可以将用户输入的个性化推荐配置信息发送给所述客户挖掘服务模块102，以使所述客户挖掘服务模块102基于所述个性化推荐配置信息，以及计算的每个候选客户的样本数据与种子客户的样本数据之间的相似性参数，筛选出满足所述个性化推荐配置信息的候选客户的样本数据，从而确定这些样本数据对应的候选客户为潜在客户。在所述客户挖掘服务模块102确定潜在客户后，可以通过所述个性化推荐模块103，将潜在客户展示给该用户。实际应用中，个性化推荐模块103可以基于人机显示界面实现，即各种个性化推荐配置信息可以在显示界面中显示，然后用户直接在显示的各种个性化推荐配置信息中选择相应信息推送给客户挖掘服务模块102，客户挖掘服务模块102可以通过显示界面将潜在客户信息展示给用户。

[0089] 需要说明的是，上述本申请实施例并不构成对所述客户挖掘设备中的功能和模块

分布的限定。可选的，该客户挖掘设备中还可以集成有其他功能模块。

[0090] 本申请实施例还提供了一种客户挖掘设备，其中，所述客户挖掘设备可以具有图1所示的客户挖掘设备的各种逻辑功能。参阅图2所示，所述客户挖掘设备200包括：处理器201、至少一个存储器202。

[0091] 所述处理器201和所述至少一个存储器202相互连接。可选的，所述处理器201和所述存储器202可以通过总线203相互连接；所述总线203可以是外设部件互连标准(peripheral component interconnect，PCI)总线或扩展工业标准结构(extended industry standard architecture，EISA)总线等。所述总线可以分为地址总线、数据总线、控制总线等。为便于表示，图2中仅用一条粗线表示，但并不表示仅有一根总线或一种类型的总线。

[0092] 可选的，所述客户挖掘设备200还包括人机交互接口204，用于实现个性化推荐模块103的功能。可选的，人机交互接口204中可以通过显示面板显示人机显示界面，或者通过音频电路输入输出音频信息，实现人机交互，这样用户可以实现设置个性化推荐配置信息，以及将处理器201筛选到的潜在客户的信息展示给用户。

[0093] 所述处理器201，具有客户挖掘服务模块102的功能，用于实现本申请实施例提供的计算样本数据之间的距离的方法，和/或，客户挖掘方法，具体实现过程可以参见本申请后续实施例对相应方法的具体描述，此处暂不赘述。

[0094] 其中，在所述处理器201实现客户挖掘方法时，可以从所述人机交互接口204的个性化推荐模块103中获取用户输入的个性化推荐配置信息，从而实现潜在客户的挖掘，并将挖掘的潜在客户信息通过所述人机交互接口204展示给用户。

[0095] 所述至少一个存储器202，用于存放程序指令，以及样本数据库101和算子库104。具体地，程序指令可以包括程序代码，该程序代码包括计算机操作的指令。存储器202可能包含随机存取存储器(random access memory，RAM)，也可能还包括非易失性存储器(non-volatile memory)，例如至少一个磁盘存储器。处理器201执行存储器202所存放的程序指令，并读取样本数据库101中的样本数据，实现上述功能，从而实现本申请实施例提供的方法。此外，样本数据库101和算子库104可以分别存储在不同的存储器202中，还可以存储在相同的存储器202中。

[0096] 可选的，在所述处理器201中的客户挖掘服务模块102需要除客户挖掘算法以外的其他算法的算法支持时，所述处理器201还可以从所述至少一个存储器202中的算子库104中读取相应的算法，以保证客户挖掘算法顺利进行。

[0097] 本申请实施例提供了一种样本数据相似性计算方法，该方法可以适用于如图2所示的客户挖掘设备200，该方法可以由图2所示的客户挖掘设备200中的处理器201执行。参阅图3所示，该方法的流程包括：

[0098] S301：所述处理器201中的客户挖掘服务模块102从存储器202中存储的样本数据库101中获取M个种子客户的样本数据和N个候选客户的样本数据，M、N均为大于2的整数。

[0099] 可选的，所述处理器201可以在接收到人机交互接口204发送端触发消息后启动执行S301，其中，该触发消息为用户通过所述人机交互接口204中的个性化推荐模块103输入的。此外，用户还可以通过所述人机交互接口204的个性化推荐模块103输入样本数据的信息，这样，所述处理器201就可以根据所述样本数据的信息，准确地读取到所述M个种子客户

的样本数据和所述N个候选客户的样本数据。

[0100] S302:所述客户挖掘服务模块102在所述N个候选客户的样本数据中,选择出与所述M个种子客户的样本数据相似性最低的P个第一样本数据,P为大于2的整数。

[0101] 可选的,所述客户挖掘服务模块102可以采用预设的客户挖掘算法,执行S302,具体包括以下步骤:

[0102] a、所述客户挖掘服务模块102计算所述M个种子客户的样本数据和所述N个候选客户的样本数据中不同样本数据之间的欧氏距离;

[0103] b、所述客户挖掘服务模块102根据计算得到的不同样本数据之间的距离,以及所述客户挖掘算法,确定所述N个候选客户的样本数据中每个候选客户的样本数据与所述M个种子客户的样本数据之间的相似性参数;

[0104] c、所述客户挖掘服务模块102根据确定的每个候选客户的样本数据与所述M个种子客户的样本数据之间的相似性参数,在所述N个候选客户的样本数据中,选择所述P个第一样本数据。

[0105] 当所述客户挖掘算法是基于DP模型实现时,在上述步骤b中,所述客户挖掘服务模块102可以参照上述对DP模型的训练过程中的描述,通过计算得到的不同样本数据之间的欧氏距离,确定每个样本数据的K个近邻,按照第2步的内容,对每个样本数据的能力进行初始化,以及按照第4、5a、5b几个步骤执行密度传播,最终可以确定每个候选客户的样本数据的密度。其中,每个候选客户的样本数据的密度即为相应的候选客户的样本数据与所述M个种子客户的样本数据之间的相似性参数。

[0106] 在所述客户挖掘服务模块102执行步骤c时,所述客户挖掘服务模块102可以选择密度最低的P个候选客户的样本数据为所述P个第一样本数据。

[0107] S303:所述客户挖掘服务模块102根据所述M个种子客户的样本数据和所述P个第一样本数据,确定度量矩阵。

[0108] 其中,所述度量矩阵为半正定矩阵,所述度量矩阵中每个元素大于0,所述度量矩阵满足以下条件:基于所述度量矩阵计算的所述M个种子客户的样本数据中所有不同种子客户的样本数据之间的距离之和最小,且基于所述度量矩阵计算的所述P个第一样本数据中所有不同第一样本数据之间的距离之和大于设定距离阈值。

[0109] 可选的,基于所述度量矩阵计算的样本数据1和样本数据2之间的距离满足以下公式:

[0110] $$d(e,f) = \|(e-f)\|_A = \sqrt{(e-f)^T A (e-f)}$$

[0111] 其中,d(e,f)为样本数据e和样本数据f之间的距离,e为所述不同样本数据中的一个样本数据构成的向量,f为所述不同样本数据中的另一个样本数据构成的向量,A为所述度量矩阵。

[0112] 因此,所述度量矩阵满足的条件还可以通过以下公式表示:

[0113] $$\min_A \sum_{(B_i,B_j)\in B} \left\|(B_i - B_j)\right\|_A, \text{且} \sum_{(C_k,C_h)\in C} \|(C_k - C_h)\|_A > R$$

[0114] 其中,B为M个种子客户的样本数据中每个种子客户的样本数据构成的向量组成的矩阵,B_i为矩阵B中一个种子客户的样本数据构成的向量,B_j为矩阵B中一个种子客户的样本数据构成的向量,i不等于j,且i和j均为小于或等于M的正整数;C为P个第一样本数据中每

个第一样本数据构成的向量组成的矩阵，C_k为矩阵C中的一个第一样本数据构成的向量，C_h为矩阵C中另一个第一样本数据构成的向量，k不等于h，且k和h均为小于或等于P的正整数；R为所述设定距离阈值。

[0115] 在实际应用中，所述客户挖掘服务模块102可以将上述度量矩阵满足的条件输入设定的函数优化算法，从而得到所述度量矩阵。还需要说明的是，所述度量矩阵的行数或列数等于样本数据包含的特征数。例如，当种子客户的样本数据和候选客户的样本数据均包含3个特征，那么，计算得到的度量矩阵为3*3矩阵。

[0116] 显然，该度量矩阵在满足P个第一样本数据中不同第一样本数据之间(即实际上不相似的样本数据之间)的距离之和在大于该设定距离阈值的情况下，还可以使得所述M个种子客户的样本数据中所有不同种子客户的样本数据(即实际上相似的样本数据)之间的距离之和尽可能小。

[0117] 理论上来说，不同种子客户的样本数据之间的相似度应该较高，而与所述M个种子客户的样本数据之间的相似度最不相似的不同第一样本数据之间的相似度应该较低。而通过该度量矩阵满足的条件可知，通过度量矩阵计算的实际上相似的样本数据之间距离较小，而实际上不相似的样本数据之间的距离较大。因此，通过以上描述可知，基于该度量矩阵计算得到的两个样本数据之间的距离可以更能体现这两个样本数据之间的相似度，因此，基于该度量矩阵计算不同样本数据之间的距离，可以实现度量优化。

[0118] S304：所述客户挖掘服务模块102基于所述度量矩阵，计算所述M个种子客户的样本数据和所述N个候选客户的样本数据中不同样本数据之间的距离。

[0119] 通过以上对所述度量矩阵的描述可知，基于所述度量矩阵计算的不同样本数据之间的距离满足以下公式：

[0120] $$d(x,y)=\sqrt{(x-y)^{T}A(x-y)}$$

[0121] 其中，d(x,y)为所述不同样本数据之间的距离，x为所述不同样本数据中的一个样本数据构成的向量，y为所述不同样本数据中的另一个样本数据构成的向量，A为所述度量矩阵。

[0122] 采用本申请实施例提供的方法，客户挖掘设备通过M个种子客户的样本数据和与所述M个种子客户的样本数据相似性最低的P个候选客户的样本数据，确定用于计算不同样本数据之间的距离的度量矩阵。由于通过该度量矩阵计算样本数据之间的距离可以保证：实际上相似的样本数据之间的距离较小，而实际上不相似的样本数据之间的距离较大。显然，基于该度量矩阵计算得到的两个样本数据之间的距离可以更能体现这两个样本数据之间的相似度，从而实现度量优化。

[0123] 基于以上实施例，本申请实施例还提供了一种客户挖掘方法，该客户挖掘方法中可以通过以上实施例中的样本数据相似性计算方法中的步骤，计算度量矩阵，并基于所述度量矩阵，计算M个种子客户的样本数据和N个候选客户的样本数据中不同样本数据之间的距离，具体过程可以参见图3所示的实施例中的描述，此处不再赘述。该方法也可以适用于如图2所示的客户挖掘设备200，该方法可以由图2所示的客户挖掘设备200中的处理器201执行。

[0124] 参阅图4所示，本申请实施例提供的客户挖掘方法中S401-S404与图3所示的实施例中S301-S304相同。在所述处理器201执行完S404之后还包括以下步骤：

[0125] S405:所述处理器201中的客户挖掘服务模块102根据计算得到的不同样本数据之间的距离,以及预设的客户挖掘算法,确定所述N个候选客户的样本数据中每个候选客户的样本数据与所述M个种子客户的样本数据之间的相似性参数。

[0126] 其中,为了保证计算一致性,所述客户挖掘服务模块102在S302中的步骤b中所采用的客户挖掘算法与本步骤中采用的客户挖掘算法是相同的。

[0127] 当所述客户挖掘算法是基于DP模型实现时,同S302中的步骤b,所述客户挖掘服务模块102可以参照上述对DP模型的训练过程中的描述,通过基于所述度量矩阵计算得到的不同样本数据之间的距离,确定每个样本数据的K个近邻,按照第2步的描述,对每个样本数据的能力进行初始化,以及按照第4、5a、5b几个步骤执行密度传播,最终可以确定每个候选客户的样本数据的密度。其中,每个候选客户的样本数据的密度即为相应的候选客户的样本数据与所述M个种子客户的样本数据之间的相似性参数。

[0128] 由于基于该度量矩阵计算得到的两个样本数据之间的距离可以更能体现这两个样本数据之间的相似度,因此,相对于传统方案中,根据基于该度量矩阵计算的不同样本数据之间的距离以及预设的客户挖掘算法,最终确定的所述N个候选客户的样本数据中每个候选客户的样本数据与所述M个种子客户的样本数据之间的相似性参数的准确度更高。

[0129] S406:所述客户挖掘服务模块102判断预设的停止迭代计算的条件是否满足;若满足,则所述处理器201执行S407;若不满足,则所述处理器201执行S408。所述停止迭代计算的条件可以为在所述处理器201中预设的。

[0130] 可选的,当所述客户挖掘服务模块102可以从人机交互接口204中的个性化推荐模块103获取所述停止迭代计算的条件。所述停止迭代计算的条件包含在用户通过个性化推荐模块103输入的个性化推荐配置信息中。

[0131] 其中,所述停止迭代计算的条件为以下至少一项:

[0132] 迭代计算的次数达到设定次数;

[0133] 所述N个候选客户的样本数据中至少一个候选客户的样本数据与所述M个种子客户的样本数据之间的相似性参数大于设定第一相似性阈值;

[0134] 所述N个候选客户的样本数据中每个候选客户的样本数据与所述M个种子客户的样本数据之间的相似性参数大于设定第二相似性阈值。

[0135] 根据图3所示的实施例中对S302的描述可知,在确定所述度量矩阵之前,所述客户挖掘设通过计算所述M个种子客户的样本数据和所述N个候选客户的样本数据中不同样本数据之间的欧氏距离,确定所述P个第一样本数据的。由于欧氏距离无法准确地体现不同样本数据之间的相似性,因此根据计算得到的不同样本数据之间的欧氏距离,确定的每个候选客户的样本数据与所述M个种子客户的样本数据之间的相似性参数的准确性较低,进一步会降低选择的所述P个第一样本数据的准确率,经过一系列误差累加,最终会影响上述S305中计算的每个候选客户的样本数据与所述M个种子客户的样本数据之间的相似性参数的准确性。

[0136] 为了进一步实现度量优化,在本申请实施例中,可以进行迭代计算,即基于上次计算确定每个候选客户的样本数据与所述M个种子客户的样本数据之间的相似性参数,重新确定与种子客户的样本数据相似性最低的P个第一样本数据,从而再次确定度量矩阵,如此反复,直至满足停止迭代计算的条件。通过这种方法,可以进一步提高所述处理器201最终

确定的每个候选客户的样本数据与所述M个种子客户的样本数据之间的相似性参数的准确性。

[0137] S407:所述客户挖掘服务模块102根据确定的每个候选客户的样本数据与所述M个种子客户的样本数据之间的相似性参数,在所述N个候选客户的样本数据中,选择出与所述M个种子客户的样本数据相似性最高的Q个第二样本数据,其中,Q为大于1的整数。

[0138] 其中,Q可以为所述客户挖掘服务模块102预设的,或者为所述客户挖掘服务模块102从人机交互接口204中的个性化推荐模块103获取的。此时,Q也包含在用户通过个性化推荐模块103输入的个性化推荐配置信息中。

[0139] 当所述客户挖掘服务模块102确定所述Q个第二样本数据后,可以确定这些第二样本数据对应的候选客户即为潜在客户。然后,所述客户挖掘服务模块102可以通过所述个性化推荐模块103,将这些潜在客户展示给该用户。

[0140] S408:所述客户挖掘服务模块102根据确定的每个候选客户的样本数据与所述M个种子客户的样本数据之间的相似性,在N个候选客户的样本数据中,选择新的P个第一样本数据;根据所述M个种子客户的样本数据和所述新的P个第一样本数据,确定新的度量矩阵;基于所述新的度量矩阵,重新计算所述M个种子客户的样本数据和所述N个候选客户的样本数据中不同样本数据之间的距离;以及根据重新计算得到的不同样本数据之间的距离,以及预设的客户挖掘算法,确定所述N个候选客户的样本数据中每个候选客户的样本数据与所述M个种子客户的样本数据之间的相似性参数,直至所述停止迭代计算的条件满足。

[0141] 参阅图4所示,所述客户挖掘服务模块102在执行S408的过程,即重新执行S402-S406的过程,具体描述可以以上实施例中相应步骤的描述,此处不再赘述。

[0142] 通过这种迭代计算的方法,可以进一步提高所述客户挖掘服务模块102最终确定的每个候选客户的样本数据与所述M个种子客户的样本数据之间的相似性参数的准确性。

[0143] 采用本申请实施例提供的方法,客户挖掘设备可以通过多次度量优化,使基于最终确定的度量矩阵计算得到的两个样本数据之间的距离体现这两个样本数据之间的相似度的准确性更高,这样,所述客户挖掘设备根据基于该度量矩阵计算的不同样本数据之间的距离以及预设的客户挖掘算法,提高最终确定的每个候选客户的样本数据与所述M个种子客户的样本数据之间的相似性参数的准确度,进而提高挖掘的潜在客户的准确度,实现精准营销。

[0144] 基于图3和图4所示的实施例,本申请实施例还提供了一种客户挖掘方法实例,该示例可以适用于如图1或图2所示的客户挖掘设备。参阅图5所示,该实例的流程包括:

[0145] S501:客户挖掘服务模块102从存储器202中的样本数据库101中获取M个种子客户的样本数据和N个候选客户的样本数据,M、N均为大于2的整数;计算所述M个种子客户的样本数据和N个候选客户的样本数据中不同样本数据之间的距离。

[0146] 需要说明的是,在首次计算过程中,所述客户挖掘服务模块102计算不同样本数据之间的欧氏距离,而在后续的迭代计算过程中,所述客户挖掘服务模块102基于确定的度量矩阵,计算不同样本数据之间的距离。

[0147] S502:所述客户挖掘服务模块102根据本次计算得到的不同样本数据之间的距离,以及所述客户挖掘算法,确定所述N个候选客户的样本数据中每个候选客户的样本数据与所述M个种子客户的样本数据之间的相似性参数。

[0148] 具体过程可以参见图3所示的实施例中对S302中的具体描述，或者图4所示实施例中对S405的描述，此处不再赘述。

[0149] S503：所述客户挖掘服务模块102判断预设的停止迭代计算的条件是否满足，若满足则执行S504；若不满足，则执行S505。

[0150] 所述停止迭代计算的条件可以为以下至少一项，本申请对此不作限定：

[0151] 迭代计算的次数达到设定次数；

[0152] 所述N个候选客户的样本数据中至少一个候选客户的样本数据与所述M个种子客户的样本数据之间的相似性参数大于设定第一相似性阈值；

[0153] 所述N个候选客户的样本数据中每个候选客户的样本数据与所述M个种子客户的样本数据之间的相似性参数大于设定第二相似性阈值。

[0154] S504：所述客户挖掘服务模块102根据确定的每个候选客户的样本数据与所述M个种子客户的样本数据之间的相似性参数，在所述N个候选客户的样本数据中，选择出与所述M个种子客户的样本数据相似性最高的Q个第二样本数据，其中，Q为大于1的整数。

[0155] S505：所述客户挖掘服务模块102根据确定的每个候选客户的样本数据与所述M个种子客户的样本数据之间的相似性，在N个候选客户的样本数据中，选择出与所述M个种子客户的样本数据相似性最低的P个第一样本数据，P为大于2的整数。

[0156] S506：所述客户挖掘服务模块102根据所述M个种子客户的样本数据和所述P个第一样本数据，确定度量矩阵。所述度量矩阵应该满足的条件参见图3所示的实施例对S303的相关描述，此处不再赘述。所述客户挖掘服务模块102确定所述度量矩阵后，继续执行S501，直至所述停止迭代计算的条件满足，最终选择Q个第二样本数据，该Q个第二样本数据对应的候选客户即为所述客户挖掘服务模块102挖掘出的潜在客户。然后，所述客户挖掘服务模块102可以通过所述个性化推荐模块103，将这些潜在客户展示给该用户。

[0157] 例如，在本实例中，所述种子客户的样本数据为表1所示，候选客户的样本数据为表2所示：

[0158] 表1

[0159]

客户标识	特征1	特征2	特征3	特征4
10	1	2	3	1
11	1	2	1	6
12	……	……	……	……
13	……	……	……	……

[0160] 表2

[0161]

客户标识	特征1	特征2	特征3	特征4
20	1	2	2	4
21	0	3	2	1
22	1	5	1	10
23	……	……	……	……
24	……	……	……	……

[0162] 首次计算后，所述客户挖掘服务模块102根据首次计算得到的不同样本数据之间的欧氏距离，以及基于DP模型实现的客户挖掘算法，确定每个候选客户的样本数据的密度为表3所示：

[0163] 表3

[0164]

客户标识	密度
20	0.8
21	0.87
22	0.5
23	0.67
24	0.9

[0165] 所述客户挖掘服务模块102在确定当前不满足停止迭代计算的条件后，根据每个候选客户的样本数据的密度，在该5个候选客户的样本数据中，选择出与表1所示的4个种子客户的样本数据相似性最低的3个第一样本数据，即选择出密度最低的候选客户的样本数据(客户标识为22、23、20的候选客户的样本数据)。

[0166] 然后，所述客户挖掘服务模块102根据该4个种子客户的样本数据和该3个第一样本数据，确定度量矩阵，所述度量矩阵为$\begin{matrix}3.245 & 3.286 & 0.081\\ 3.286 & 3.327 & 0.082\\ 0.081 & 0.082 & 0.002\end{matrix}$。

[0167] 所述客户挖掘服务模块102根据所述度量矩阵，继续执行S401，直至所述停止迭代计算的条件满足，末次迭代计算后，所述客户挖掘服务模块102根据末次迭代计算得到的不同样本数据之间的距离，以及基于DP模型实现的客户挖掘算法，确定每个候选客户的样本数据的密度为表4所示：

[0168] 表4

[0169]

客户标识	密度
20	0.83
21	0.92
22	0.3
23	0.56
24	0.95

[0170] 所述客户挖掘服务模块102根据末次确定的每个候选客户的样本数据的密度，在该5个候选客户的样本数据中，选择出与表1所示的4个种子客户的样本数据相似性最低的2个第二样本数据，即选择出密度最高的候选客户的样本数据(客户标识为24、21的候选客户的样本数据)。显然，客户标识为24、21的候选客户为所述客户挖掘服务模块102挖掘出的潜在客户。

[0171] 所述客户挖掘服务模块102可以通过所述个性化推荐模块103，将这些潜在客户的客户标识展示给该用户。

[0172] 表5示出了针对相同的样本数据，采用传统客户挖掘方法和本申请实施例提供的客户挖掘方法得出的潜在客户的命中率。其中在传统方案中，采用欧氏距离算法计算不同

样本数据之间的距离，而本申请实施例提供的方案中，采用度量优化的度量矩阵计算不同样本数据之间的距离。

[0173] 表5

[0174]

	传统方案	本申请实施例提供的方案
AUC	0.577	0.6264
Q＝5000	0.1394	0.2398
Q＝10000	0.1384	0.2417
Q＝20000	0.1364	0.2018
Q＝50000	0.1351	0.22
Q＝100000	0.1318	0.1891
Q＝200000	0.1259	0.1724

[0175] 通过表5可知，相对于传统的方案，采用本申请实施例提供的客户挖掘方法进行客户挖掘后，可以提高挖掘出的潜在客户的命中率，即提高挖掘的潜在客户的准确度。因此，通过本申请实施例提供的方案，可以提高待推销产品的营销成功率。

[0176] 基于以上实施例，本申请还提供了一种客户挖掘设备，所述客户挖掘算法用于实现以上实施例提供的方法。参阅图6所示，所述客户挖掘设备600包括：获取单元601，确定单元602，以及处理单元603，其中：

[0177] 获取单元601，用于获取M个种子客户的样本数据和N个候选客户的样本数据，M、N均为大于2的整数；

[0178] 确定单元602，用于在所述N个候选客户的样本数据中，选择出与所述M个种子客户的样本数据相似性最低的P个第一样本数据，P为大于2的整数；并根据所述M个种子客户的样本数据和所述P个第一样本数据，确定度量矩阵；

[0179] 其中，所述度量矩阵为半正定矩阵，所述度量矩阵中每个元素大于0，所述度量矩阵满足以下条件：基于所述度量矩阵计算的所述M个种子客户的样本数据中所有不同种子客户的样本数据之间的距离之和最小，且基于所述度量矩阵计算的所述P个第一样本数据中所有不同第一样本数据之间的距离之和大于设定距离阈值；

[0180] 处理单元603，用于基于所述度量矩阵，计算所述M个种子客户的样本数据和所述N个候选客户的样本数据中不同样本数据之间的距离。

[0181] 在一个实现方式中，基于所述度量矩阵计算的不同样本数据之间的距离满足以下公式：

[0182] $$d(x,y)=\sqrt{(x-y)^{T}A(x-y)}$$

[0183] 其中，d(x,y)为所述不同样本数据之间的距离，x为所述不同样本数据中的一个样本数据构成的向量，y为所述不同样本数据中的另一个样本数据构成的向量，A为所述度量矩阵。

[0184] 在一个实现方式中，所述确定单元602，具体用于：

[0185] 计算所述M个种子客户的样本数据和所述N个候选客户的样本数据中不同样本数据之间的欧氏距离；

[0186] 根据计算得到的不同样本数据之间的欧氏距离，以及预设的客户挖掘算法，确定

所述N个候选客户的样本数据中每个候选客户的样本数据与所述M个种子客户的样本数据之间的相似性参数；

[0187] 根据确定的每个候选客户的样本数据与所述M个种子客户的样本数据之间的相似性参数，在所述N个候选客户的样本数据中，选择所述P个第一样本数据。

[0188] 在一个实现方式中，所述处理单元603，还用于：

[0189] 在计算所述M个种子客户的样本数据和所述N个候选客户的样本数据中不同样本数据之间的距离之后，根据计算得到的不同样本数据之间的距离，以及预设的客户挖掘算法，确定所述N个候选客户的样本数据中每个候选客户的样本数据与所述M个种子客户的样本数据之间的相似性参数；

[0190] 判断预设的停止迭代计算的条件是否满足；

[0191] 若满足，则根据确定的每个候选客户的样本数据与所述M个种子客户的样本数据之间的相似性参数，在所述N个候选客户的样本数据中，选择出与所述M个种子客户的样本数据相似性最高的Q个第二样本数据，其中，Q为大于1的整数；

[0192] 若不满足，则根据确定的每个候选客户的样本数据与所述M个种子客户的样本数据之间的相似性，在N个候选客户的样本数据中，选择新的P个第一样本数据；根据所述M个种子客户的样本数据和所述新的P个第一样本数据，确定新的度量矩阵；基于所述新的度量矩阵，重新计算所述M个种子客户的样本数据和所述N个候选客户的样本数据中不同样本数据之间的距离；以及根据重新计算得到的不同样本数据之间的距离，以及所述客户挖掘算法，确定所述N个候选客户的样本数据中每个候选客户的样本数据与所述M个种子客户的样本数据之间的相似性参数，直至所述停止迭代计算的条件满足。

[0193] 在一个实现方式中，所述停止迭代计算的条件为以下至少一项：

[0194] 迭代计算的次数达到设定次数；

[0195] 所述N个候选客户的样本数据中至少一个候选客户的样本数据与所述M个种子客户的样本数据之间的相似性参数大于设定第一相似性阈值；

[0196] 所述N个候选客户的样本数据中每个候选客户的样本数据与所述M个种子客户的样本数据之间的相似性参数大于设定第二相似性阈值。

[0197] 在一个实现方式中，所述客户挖掘算法为密度传播算法。

[0198] 本申请实施例提供了一种客户挖掘设备，该客户挖掘设备通过M个种子客户的样本数据和与所述M个种子客户的样本数据相似性最低的P个候选客户的样本数据，确定用于计算不同样本数据之间的距离的度量矩阵。由于通过该度量矩阵计算样本数据之间的距离可以保证：实际上相似的样本数据之间的距离较小，而实际上不相似的样本数据之间的距离较大。显然，基于该度量矩阵计算得到的两个样本数据之间的距离可以更能体现这两个样本数据之间的相似度，从而实现度量优化。

[0199] 需要说明的是，本申请实施例中对模块的划分是示意性的，仅仅为一种逻辑功能划分，实际实现时可以有另外的划分方式，另外，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单元中，也可以是单独物理存在，也可以两个或两个以上单元集成在一个单元中。上述集成的单元既可以采用硬件的形式实现，也可以采用软件功能单元的形式实现。

[0200] 所述集成的单元如果以软件功能单元的形式实现并作为独立的产品销售或使用

时,可以存储在一个计算机可读取存储介质中。基于这样的理解,本申请的技术方案本质上或者说对现有技术做出贡献的部分或者该技术方案的全部或部分可以以软件产品的形式体现出来,该计算机软件产品存储在一个存储介质中,包括若干指令用以使得一台计算机设备(可以是个人计算机,服务器,或者网络设备等)或处理器(processor)执行本申请各个实施例所述方法的全部或部分步骤。而前述的存储介质包括:U盘、移动硬盘、只读存储器(Read-Only Memory,ROM)、随机存取存储器(Random Access Memory,RAM)、磁碟或者光盘等各种可以存储程序代码的介质。

[0201] 基于以上实施例,本申请提供一种计算机程序,当该计算机程序在计算机上运行时,使得所述计算机执行以上实施例提供的方法。

[0202] 基于以上实施例,本申请提供一种计算机存储介质,所述计算机存储介质中存储有计算机指令,所述计算机指令被计算机执行时,使得所述计算机执行以上实施例提供的方法。

[0203] 基于以上实施例,本申请提供一种芯片,所述芯片用于读取并执行所述存储器中存储的计算机程序,以实现以上实施例中的方法。

[0204] 基于以上实施例,本申请实施例提供了一种芯片系统,该芯片系统包括处理器,用于支持客户挖掘设备实现上述实施例中所涉及的相应的功能。在一种可能的设计中,所述芯片系统还包括存储器,所述存储器,用于保存所述客户挖掘设备必要的程序指令和数据。该芯片系统,可以由芯片构成,也可以包含芯片和其他分立器件。

[0205] 本领域内的技术人员应明白,本申请的实施例可提供为方法、系统、或计算机程序产品。因此,本申请可采用完全硬件实施例、完全软件实施例、或结合软件和硬件方面的实施例的形式。而且,本申请可采用在一个或多个其中包含有计算机可用程序代码的计算机可用存储介质(包括但不限于磁盘存储器、CD-ROM、光学存储器等)上实施的计算机程序产品的形式。

[0206] 本申请是参照根据本申请的方法、设备(系统)、和计算机程序产品的流程图和/或方框图来描述的。应理解可由计算机程序指令实现流程图和/或方框图中的每一流程和/或方框、以及流程图和/或方框图中的流程和/或方框的结合。可提供这些计算机程序指令到通用计算机、专用计算机、嵌入式处理机或其他可编程数据处理设备的处理器以产生一个机器,使得通过计算机或其他可编程数据处理设备的处理器执行的指令产生用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的装置。

[0207] 这些计算机程序指令也可存储在能引导计算机或其他可编程数据处理设备以特定方式工作的计算机可读存储器中,使得存储在该计算机可读存储器中的指令产生包括指令装置的制造品,该指令装置实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能。

[0208] 这些计算机程序指令也可装载到计算机或其他可编程数据处理设备上,使得在计算机或其他可编程设备上执行一系列操作步骤以产生计算机实现的处理,从而在计算机或其他可编程设备上执行的指令提供用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的步骤。

[0209] 显然,本领域的技术人员可以对本申请进行各种改动和变型而不脱离本申请的精神和范围。这样,倘若本申请的这些修改和变型属于本申请权利要求及其等同技术的范围

之内,则本申请也意图包含这些改动和变型在内。

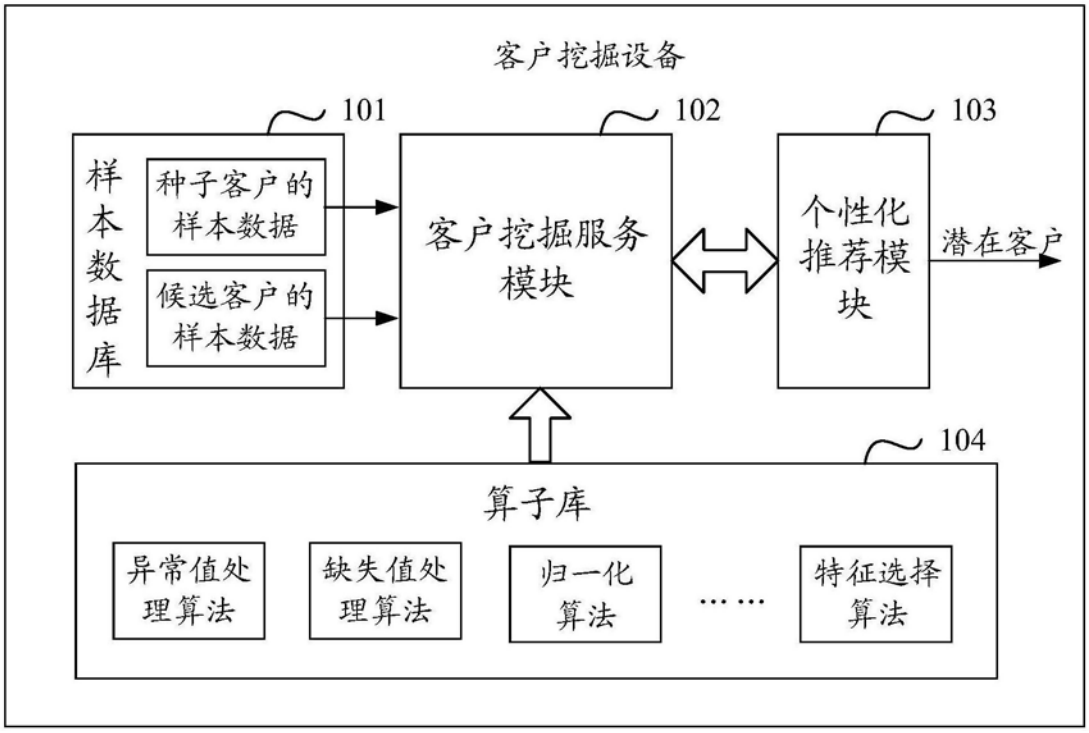

客户挖掘设备
101
102
103
样本数据库
种子客户的样本数据
候选客户的样本数据
客户挖掘服务模块
个性化推荐模块
潜在客户
104
算子库
异常值处理算法
缺失值处理算法
归一化算法
……
特征选择算法

图1

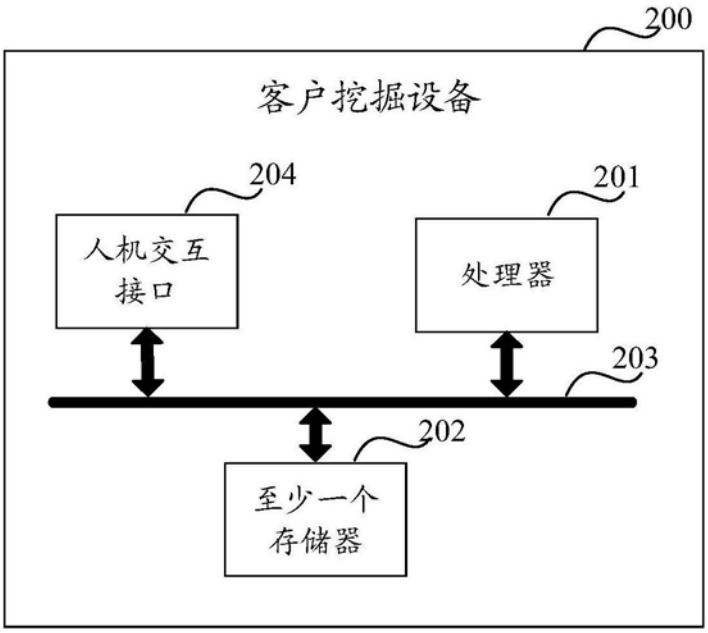

200
客户挖掘设备
204
201
人机交互接口
处理器
203
202
至少一个存储器

图2

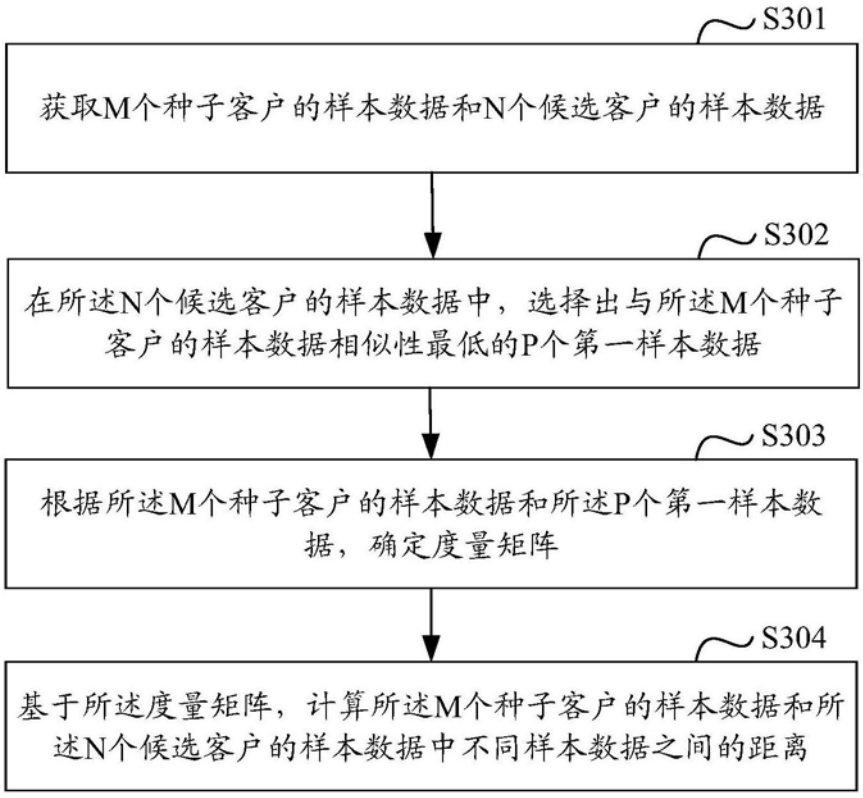

S301
获取M个种子客户的样本数据和N个候选客户的样本数据
S302
在所述N个候选客户的样本数据中，选择出与所述M个种子客户的样本数据相似性最低的P个第一样本数据
S303
根据所述M个种子客户的样本数据和所述P个第一样本数据，确定度量矩阵
S304
基于所述度量矩阵，计算所述M个种子客户的样本数据和所述N个候选客户的样本数据中不同样本数据之间的距离

图3

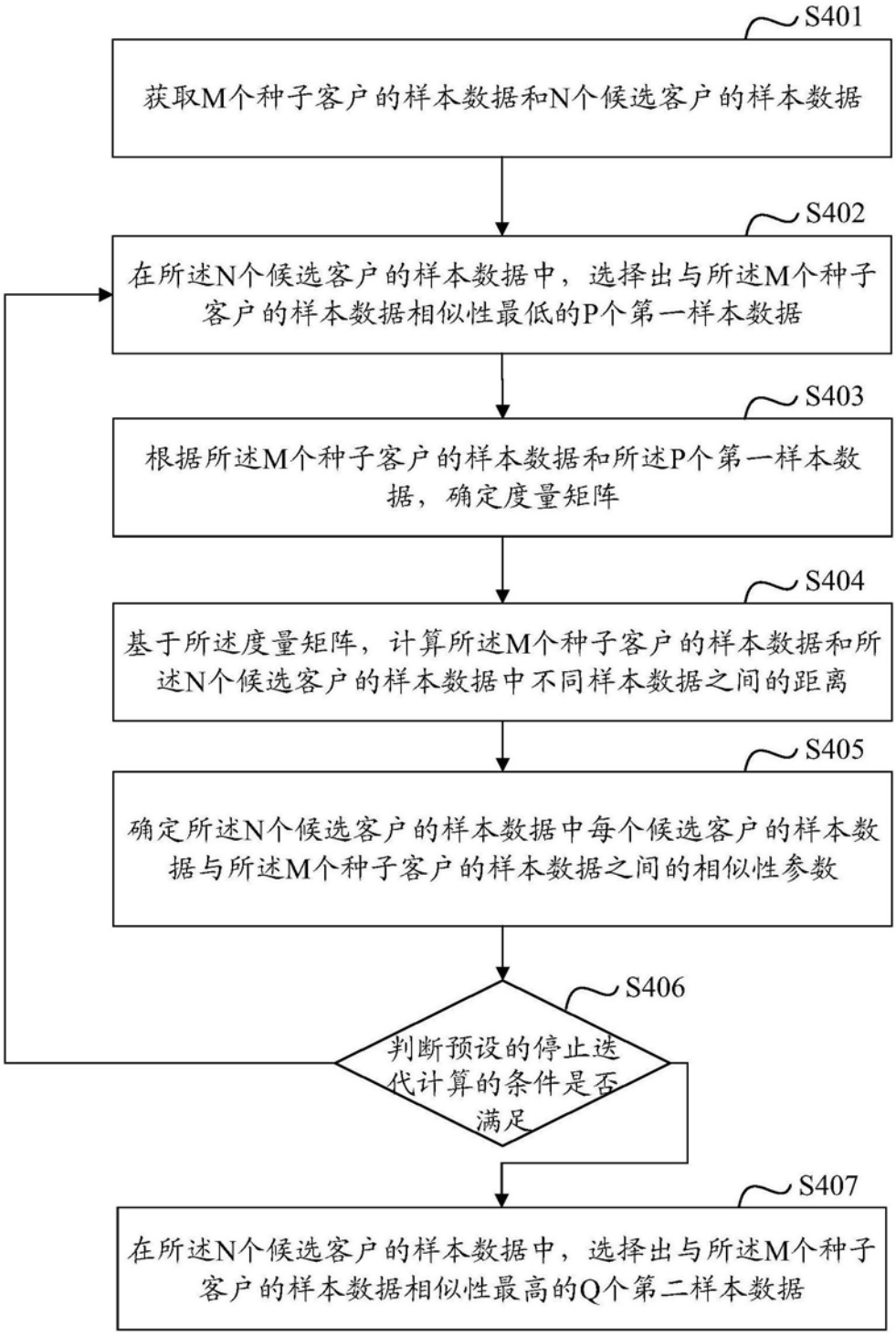

S401
获取M个种子客户的样本数据和N个候选客户的样本数据
S402
在所述N个候选客户的样本数据中，选择出与所述M个种子客户的样本数据相似性最低的P个第一样本数据
S403
根据所述M个种子客户的样本数据和所述P个第一样本数据，确定度量矩阵
S404
基于所述度量矩阵，计算所述M个种子客户的样本数据和所述N个候选客户的样本数据中不同样本数据之间的距离
S405
确定所述N个候选客户的样本数据中每个候选客户的样本数据与所述M个种子客户的样本数据之间的相似性参数
S406
判断预设的停止迭代计算的条件是否满足
S407
在所述N个候选客户的样本数据中，选择出与所述M个种子客户的样本数据相似性最高的Q个第二样本数据

图4

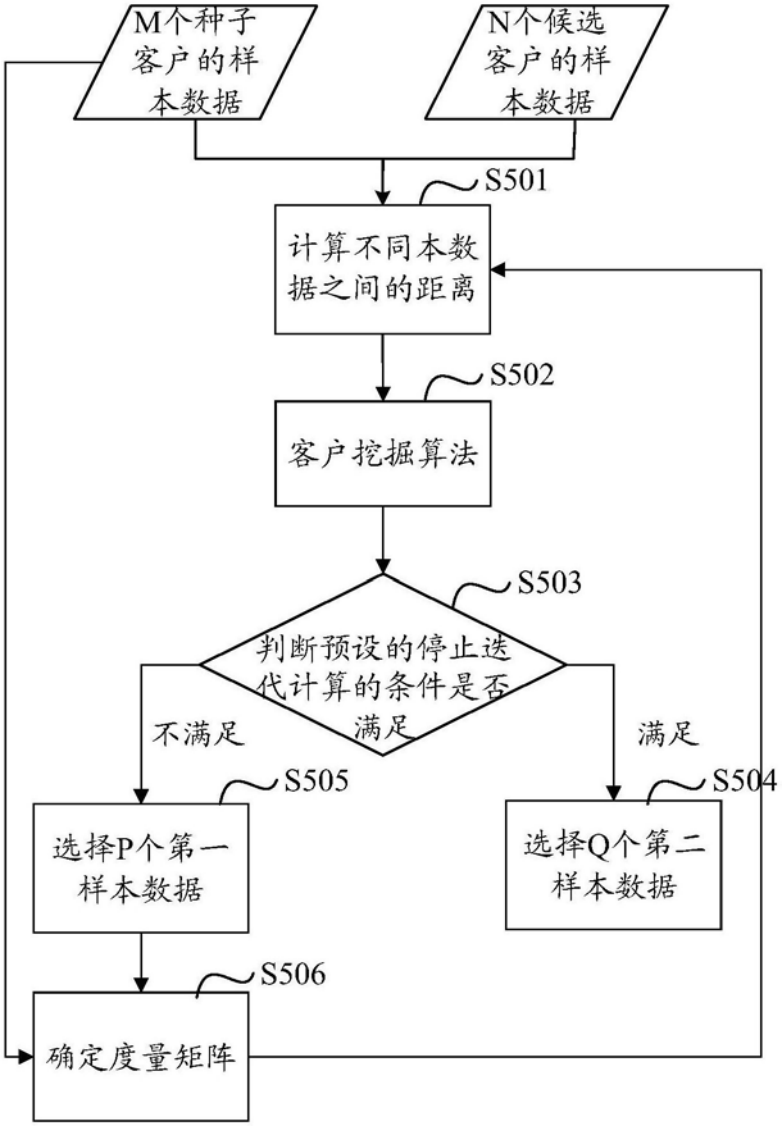

M个种子客户的样本数据
N个候选客户的样本数据
S501
计算不同本数据之间的距离
S502
客户挖掘算法
S503
判断预设的停止迭代计算的条件是否满足
不满足
满足
S505
选择P个第一样本数据
S504
选择Q个第二样本数据
S506
确定度量矩阵

图5

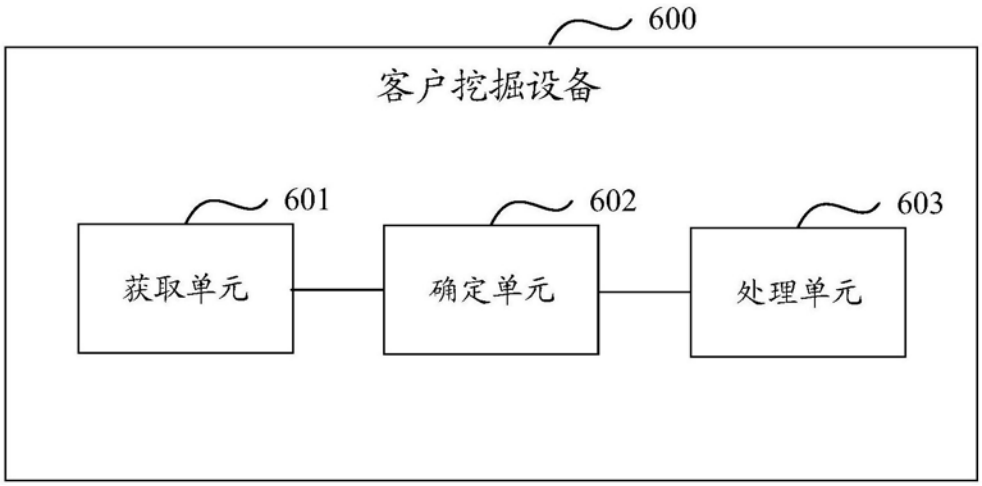


图6