(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 115099875 A
(43)申请公布日 2022.09.23

(21)申请号 202210836961.0

(22)申请日 2022.07.15

(71)申请人 平安科技(深圳)有限公司
地址 518000 广东省深圳市福田区福田街道福安社区益田路5033号平安金融中心23楼

(72)发明人 钱学广

(74)专利代理机构 深圳市世联合知识产权代理有限公司 44385
专利代理师 姜妍

(51)Int.Cl.
*G06Q 30/02*(2012.01)
*G06Q 40/08*(2012.01)
*G06K 9/62*(2022.01)
*G06F 16/2458*(2019.01)
*G06F 16/28*(2019.01)

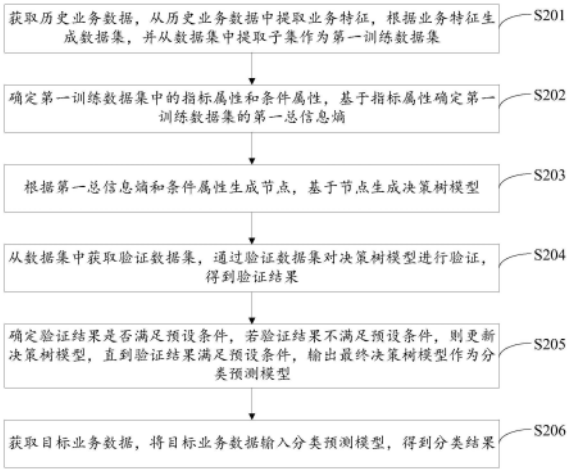
权利要求书2页 说明书15页 附图4页

(54)发明名称

基于决策树模型的数据分类方法及相关设备

(57)摘要

本申请实施例属于人工智能技术领域，涉及一种基于决策树模型的数据分类方法及相关设备，包括从获取的历史业务数据中提取业务特征生成数据集，并从数据集中提取子集作为第一训练数据集；确定第一训练数据集中的指标属性和条件属性，基于指标属性确定第一训练数据集的第一总信息熵；根据第一总信息熵和条件属性生成节点，基于节点生成决策树模型；通过验证数据集对决策树模型进行验证，得到验证结果，直到验证结果满足预设条件，输出最终决策树模型作为分类预测模型；将目标业务数据输入分类预测模型，得到分类结果。此外，本申请还涉及区块链技术，业务特征可存储于区块链中。本申请可以提高对业务数据的分类效率和分类准确性。




CN 115099875 A

1.一种基于决策树模型的数据分类方法,其特征在于,包括下述步骤:

获取历史业务数据,从所述历史业务数据中提取业务特征,根据所述业务特征生成数据集,并从所述数据集中提取子集作为第一训练数据集;

确定所述第一训练数据集中的指标属性和条件属性,基于所述指标属性确定所述第一训练数据集的第一总信息熵;

根据所述第一总信息熵和所述条件属性生成节点,基于所述节点生成决策树模型;

从所述数据集中获取验证数据集,通过所述验证数据集对所述决策树模型进行验证,得到验证结果;

确定所述验证结果是否满足预设条件,若所述验证结果不满足预设条件,则更新所述决策树模型,直到所述验证结果满足预设条件,输出最终决策树模型作为分类预测模型;

获取目标业务数据,将所述目标业务数据输入所述分类预测模型,得到分类结果。

2.根据权利要求1所述的基于决策树模型的数据分类方法,其特征在于,所述基于所述指标属性确定所述第一训练数据集的第一总信息熵的步骤包括:

确定所述指标属性中每个指标特征在所述第一训练数据集中的概率;

基于所述概率计算得到所述第一训练数据集的第一总信息熵。

3.根据权利要求1所述的基于决策树模型的数据分类方法,其特征在于,所述根据所述第一总信息熵和所述条件属性生成节点的步骤包括:

步骤A,根据所述第一总信息熵和所述条件属性的属性数据,计算得到每个所述条件属性的信息增益;

步骤B,获取所述条件属性的优化权值,通过所述优化权值优化对应的信息增益,得到优化信息增益;

步骤C,基于所述优化信息增益确定最优条件属性,并将所述最优条件属性作为节点;

步骤D,将所述节点之外的条件属性和所述指标属性组成第二训练数据集,根据所述指标属性计算所述第二训练数据集的第二总信息熵;

步骤E,循环步骤A至步骤D,直至所有的所述条件属性生成节点。

4.根据权利要求3所述的基于决策树模型的数据分类方法,其特征在于,所述根据所述第一总信息熵和所述条件属性的属性数据,计算每个所述条件属性的信息增益的步骤包括:

根据所述属性数据,计算每个所述条件属性中每个属性特征的属性信息熵;

基于所述属性信息熵计算得到对应所述条件属性的条件信息熵;

根据所述第一总信息熵和所述条件信息熵,计算得到信息增益。

5.根据权利要求3所述的基于决策树模型的数据分类方法,其特征在于,在所述根据所述第一总信息熵和所述条件属性的属性数据,计算得到每个所述条件属性的信息增益的步骤之前还包括:

确定所述属性数据是否存在异常数据;

若存在异常数据,则对所述异常数据进行修正。

6.根据权利要求5所述的基于决策树模型的数据分类方法,其特征在于,在所述获取所述条件属性的优化权值的步骤之前还包括:

确定所述异常数据对应的条件属性,统计所述异常数据在所述条件属性中的占比;

根据所述占比计算得到调整系数，根据所述调整系数调整所述条件属性的优化权值。

7.根据权利要求1所述的基于决策树模型的数据分类方法，其特征在于，所述通过所述验证数据集对所述决策树模型进行验证，得到验证结果的步骤包括：

将所述验证数据集输入所述决策树模型，输出预测结果；

根据所述预测结果计算预测准确度，将所述预测准确度作为验证结果。

8.一种基于决策树模型的数据分类装置，其特征在于，包括：

提取模块，用于获取历史业务数据，从所述历史业务数据中提取业务特征，根据所述业务特征生成数据集，并从所述数据集中提取子集作为第一训练数据集；

确定模块，用于确定所述第一训练数据集中的指标属性和条件属性，基于所述指标属性确定所述第一训练数据集的第一总信息熵；

生成模块，用于根据所述第一总信息熵和所述条件属性生成节点，基于所述节点生成决策树模型；

验证模块，用于从所述数据集中获取验证数据集，通过所述验证数据集对所述决策树模型进行验证，得到验证结果；

输出模块，用于确定所述验证结果是否满足预设条件，若所述验证结果不满足预设条件，则更新所述决策树模型，直到所述验证结果满足预设条件，输出最终决策树模型作为分类预测模型；

分类模块，用于获取目标业务数据，将所述目标业务数据输入所述分类预测模型，得到分类结果。

9.一种计算机设备，包括存储器和处理器，所述存储器中存储有计算机可读指令，所述处理器执行所述计算机可读指令时实现如权利要求1至7中任一项所述的基于决策树模型的数据分类方法的步骤。

10.一种计算机可读存储介质，其特征在于，所述计算机可读存储介质上存储有计算机可读指令，所述计算机可读指令被处理器执行时实现如权利要求1至7中任一项所述的基于决策树模型的数据分类方法的步骤。

# 基于决策树模型的数据分类方法及相关设备

## 技术领域

[0001] 本申请涉及人工智能技术领域，尤其涉及一种基于决策树模型的数据分类方法及相关设备。

## 背景技术

[0002] 由于信息技术的快速发展，在互联网上购买保险已经得到广泛应用。保险公司提供互联网平台，对外部第三方合作伙伴开放，第三方合作伙伴可在保险公司的互联网接口平台查询、购买保险产品。

[0003] 保险公司通常和很多合作伙伴进行业务对接，在合作伙伴业务对接的过程中，产生并存储大量接口访问数据，传统的数据库的数据查询、报表统计等功能，无法高效、快捷、精准的识别重要客户，无法发现数据中存在的关系和规则，无法根据现有的数据进行预期推测。

## 发明内容

[0004] 本申请实施例的目的在于提出一种基于决策树模型的数据分类方法及相关设备，以解决相关技术中无法高效、快捷、精准的识别重要客户，无法发现数据中存在的关系和规则，无法根据现有的数据进行预期推测的技术问题。

[0005] 为了解决上述技术问题，本申请实施例提供一种基于决策树模型的数据分类方法，采用了如下所述的技术方案：

[0006] 获取历史业务数据，从所述历史业务数据中提取业务特征，根据所述业务特征生成数据集，并从所述数据集中提取子集作为第一训练数据集；

[0007] 确定所述第一训练数据集中的指标属性和条件属性，基于所述指标属性确定所述第一训练数据集的第一总信息熵；

[0008] 根据所述第一总信息熵和所述条件属性生成节点，基于所述节点生成决策树模型；

[0009] 从所述数据集中获取验证数据集，通过所述验证数据集对所述决策树模型进行验证，得到验证结果；

[0010] 确定所述验证结果是否满足预设条件，若所述验证结果不满足预设条件，则更新所述决策树模型，直到所述验证结果满足预设条件，输出最终决策树模型作为分类预测模型；

[0011] 获取目标业务数据，将所述目标业务数据输入所述分类预测模型，得到分类结果。

[0012] 进一步的，所述基于所述指标属性确定所述第一训练数据集的第一总信息熵的步骤包括：

[0013] 确定所述指标属性中每个指标特征在所述第一训练数据集中的概率；

[0014] 基于所述概率计算得到所述第一训练数据集的第一总信息熵。

[0015] 进一步的，所述根据所述第一总信息熵和所述条件属性生成节点的步骤包括：

[0016] 步骤A,根据所述第一总信息熵和所述条件属性的属性数据,计算得到每个所述条件属性的信息增益;

[0017] 步骤B,获取所述条件属性的优化权值,通过所述优化权值优化对应的信息增益,得到优化信息增益;

[0018] 步骤C,基于所述优化信息增益确定最优条件属性,并将所述最优条件属性作为节点;

[0019] 步骤D,将所述节点之外的条件属性和所述指标属性组成第二训练数据集,根据所述指标属性计算所述第二训练数据集的第二总信息熵;

[0020] 步骤E,循环步骤A至步骤D,直至所有的所述条件属性生成节点。

[0021] 进一步的,所述根据所述第一总信息熵和所述条件属性的属性数据,计算每个所述条件属性的信息增益的步骤包括:

[0022] 根据所述属性数据,计算每个所述条件属性中每个属性特征的属性信息熵;

[0023] 基于所述属性信息熵计算得到对应所述条件属性的条件信息熵;

[0024] 根据所述第一总信息熵和所述条件信息熵,计算得到信息增益。

[0025] 进一步的,在所述根据所述第一总信息熵和所述条件属性的属性数据,计算得到每个所述条件属性的信息增益的步骤之前还包括:

[0026] 确定所述属性数据是否存在异常数据;

[0027] 若存在异常数据,则对所述异常数据进行修正。

[0028] 进一步的,在所述获取所述条件属性的优化权值的步骤之前还包括:

[0029] 确定所述异常数据对应的条件属性,统计所述异常数据在所述条件属性中的占比;

[0030] 根据所述占比计算得到调整系数,根据所述调整系数调整所述条件属性的优化权值。

[0031] 进一步的,所述通过所述验证数据集对所述决策树模型进行验证,得到验证结果的步骤包括:

[0032] 将所述验证数据集输入所述决策树模型,输出预测结果;

[0033] 根据所述预测结果计算预测准确度,将所述预测准确度作为验证结果。

[0034] 为了解决上述技术问题,本申请实施例还提供一种基于决策树模型的数据分类装置,采用了如下所述的技术方案:

[0035] 提取模块,用于获取历史业务数据,从所述历史业务数据中提取业务特征,根据所述业务特征生成数据集,并从所述数据集中提取子集作为第一训练数据集;

[0036] 确定模块,用于确定所述第一训练数据集中的指标属性和条件属性,基于所述指标属性确定所述第一训练数据集的第一总信息熵;

[0037] 生成模块,用于根据所述第一总信息熵和所述条件属性生成节点,基于所述节点生成决策树模型;

[0038] 验证模块,用于从所述数据集中获取验证数据集,通过所述验证数据集对所述决策树模型进行验证,得到验证结果;

[0039] 输出模块,用于确定所述验证结果是否满足预设条件,若所述验证结果不满足预设条件,则更新所述决策树模型,直到所述验证结果满足预设条件,输出最终决策树模型作

为分类预测模型。

[0040] 为了解决上述技术问题，本申请实施例还提供一种计算机设备，采用了如下所述的技术方案：

[0041] 该计算机设备包括存储器和处理器，所述存储器中存储有计算机可读指令，所述处理器执行所述计算机可读指令时实现如上所述的基于决策树模型的数据分类方法的步骤。

[0042] 为了解决上述技术问题，本申请实施例还提供一种计算机可读存储介质，采用了如下所述的技术方案：

[0043] 所述计算机可读存储介质上存储有计算机可读指令，所述计算机可读指令被处理器执行时实现如上所述的基于决策树模型的数据分类方法的步骤。

[0044] 与现有技术相比，本申请实施例主要有以下有益效果：

[0045] 本申请通过获取历史业务数据，从历史业务数据中提取业务特征，根据业务特征生成数据集，并从数据集中提取子集作为第一训练数据集；确定第一训练数据集中的指标属性和条件属性，基于指标属性确定第一训练数据集的第一总信息熵；根据第一总信息熵和条件属性生成节点，基于节点生成决策树模型；从数据集中获取验证数据集，通过验证数据集对决策树模型进行验证，得到验证结果；确定验证结果是否满足预设条件，若验证结果不满足预设条件，则更新决策树模型，直到验证结果满足预设条件，输出最终决策树模型作为分类预测模型；获取目标业务数据，将目标业务数据输入分类预测模型，得到分类结果；本申请通过获取的数据集提取训练数据集，并根据训练数据集的总信息熵确定节点，基于节点生成决策树模型，将根据生成的决策树模型对目标业务数据进行分类，可以提高对业务数据的分类效率和分类准确性，进一步高效准确地识别重要合作伙伴，并通过模型中的规则对业务人员预测合作伙伴特征具有一定的辅助作用。

## 附图说明

[0046] 为了更清楚地说明本申请中的方案，下面将对本申请实施例描述中所需要使用的附图作一个简单介绍，显而易见地，下面描述中的附图是本申请的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他的附图。

[0047] 图1是本申请可以应用于其中的示例性系统架构图；

[0048] 图2是根据本申请的基于决策树模型的数据分类方法的一个实施例的流程图；

[0049] 图3是根据本申请的基于决策树模型的数据分类装置的一个实施例的结构示意图；

[0050] 图4是根据本申请的计算机设备的一个实施例的结构示意图。

## 具体实施方式

[0051] 除非另有定义，本文所使用的所有的技术和科学术语与属于本申请的技术领域的技术人员通常理解的含义相同；本文中在申请的说明书中所使用的术语只是为了描述具体的实施例的目的，不是旨在于限制本申请；本申请的说明书和权利要求书及上述附图说明中的术语“包括”和“具有”以及它们的任何变形，意图在于覆盖不排他的包含。本申请的说明书和权利要求书或上述附图中的术语“第一”、“第二”等是用于区别不同对象，而不是用

于描述特定顺序。

[0052] 在本文中提及“实施例”意味着，结合实施例描述的特定特征、结构或特性可以包含在本申请的至少一个实施例中。在说明书中的各个位置出现该短语并不一定均是指相同的实施例，也不是与其它实施例互斥的独立的或备选的实施例。本领域技术人员显式地和隐式地理解的是，本文所描述的实施例可以与其它实施例相结合。

[0053] 为了使本技术领域的人员更好地理解本申请方案，下面将结合附图，对本申请实施例中的技术方案进行清楚、完整地描述。

[0054] 本申请提供了一种基于决策树模型的数据分类方法，涉及人工智能，可以应用于如图1所示的系统架构100中，系统架构100可以包括终端设备101、102、103，网络104和服务器105。网络104用以在终端设备101、102、103和服务器105之间提供通信链路的介质。网络104可以包括各种连接类型，例如有线、无线通信链路或者光纤电缆等等。

[0055] 用户可以使用终端设备101、102、103通过网络104与服务器105交互，以接收或发送消息等。终端设备101、102、103上可以安装有各种通讯客户端应用，例如网页浏览器应用、购物类应用、搜索类应用、即时通信工具、邮箱客户端、社交平台软件等。

[0056] 终端设备101、102、103可以是具有显示屏并且支持网页浏览的各种电子设备，包括但不限于智能手机、平板电脑、电子书阅读器、MP3播放器(Moving Picture Experts Group Audio Layer III，动态影像专家压缩标准音频层面3)、MP4(Moving Picture Experts Group Audio Layer IV，动态影像专家压缩标准音频层面4)播放器、膝上型便携计算机和台式计算机等等。

[0057] 服务器105可以是提供各种服务的服务器，例如对终端设备101、102、103上显示的页面提供支持的后台服务器。

[0058] 需要说明的是，本申请实施例所提供的基于决策树模型的数据分类方法一般由服务器/终端设备执行，相应地，基于决策树模型的数据分类装置一般设置于服务器/终端设备中。

[0059] 应该理解，图1中的终端设备、网络和服务器的数目仅仅是示意性的。根据实现需要，可以具有任意数目的终端设备、网络和服务器。

[0060] 继续参考图2，示出了根据本申请的基于决策树模型的数据分类方法的一个实施例的流程图，包括以下步骤：

[0061] 步骤S201，获取历史业务数据，从历史业务数据中提取业务特征，根据业务特征生成数据集，并从数据集中提取子集作为第一训练数据集。

[0062] 其中，历史业务数据可以从相应的业务数据库中获取，业务数据库可以是预先建立的专门存储业务数据的数据库，也可以是保险系统本身的存储数据库。

[0063] 在本实施例中，业务数据包括每天每家合作伙伴访问保险系统的访问次数、保单量以及保单信息等，其中，保单信息包括保单号、产品类型、投保额以及投保时间等。

[0064] 需要说明的是，在获取业务数据过程中，对连续型数据进行离散化处理；对缺失或异常数据，进行修正处理，如无数据或负数，则修正为0。

[0065] 基于业务规则从业务数据中提取业务特征，业务特征包括合作伙伴、访问次数、产品类型、保单量、投保总额以及是否关注合作伙伴等，基于上述业务特征，形成保单数据集。

[0066] 从保单数据集中提取子集作为第一训练数据集，可以避免数据量过大不容易收敛

的问题。在本实施例的一种具体实现方式中，第一训练数据集参见表1所示。

[0067] 表1第一训练数据集D1

[0068]

| 合作伙伴 | 产品 | 访问次数>1万次/天 | 保单量>1000/天 | 投保总额(万)/天 | 是否关注 |
|---|---|---|---|---|---|
| A | 财产险 | 否 | 是 | >100 | 是 |
| B | 财产险 | 否 | 是 | >100 | 是 |
| C | 财产险 | 是 | 否 | >10，<100 | 是 |
| D | 财产险 | 是 | 否 | >10，<100 | 是 |
| E | 财产险 | 否 | 否 | <10 | 否 |
| F | 意外险 | 否 | 否 | <10 | 否 |
| G | 意外险 | 否 | 否 | >10，<100 | 否 |
| H | 意外险 | 是 | 是 | >10，<100 | 是 |
| I | 意外险 | 否 | 是 | >100 | 是 |
| J | 意外险 | 否 | 是 | >100 | 是 |
| K | 宠物险 | 否 | 否 | (0) | 否 |

[0069]

| | | | | | |
|---|---|---|---|---|---|
| L | 宠物险 | 否 | 否 | >10，<100 | 否 |
| M | 宠物险 | 是 | 否 | >10，<100 | 是 |
| N | 宠物险 | 是 | 是 | <10 | 是 |
| O | 宠物险 | 否 | (0) | <10 | 否 |

[0070] 需要强调的是，为进一步保证业务特征的私密和安全性，上述业务特征还可以存储于一区块链的节点中。

[0071] 本申请所指区块链是分布式数据存储、点对点传输、共识机制、加密算法等计算机技术的新型应用模式。区块链(Blockchain)，本质上是一个去中心化的数据库，是一串使用密码学方法相关联产生的数据块，每一个数据块中包含了一批次网络交易的信息，用于验证其信息的有效性(防伪)和生成下一个区块。区块链可以包括区块链底层平台、平台产品

服务层以及应用服务层等。

[0072] 步骤S202,确定第一训练数据集中的指标属性和条件属性,基于指标属性确定第一训练数据集的第一总信息熵。

[0073] 在本实施例中,指标属性为决策属性,根据条件属性来确定数据集的决策属性,例如,决策属性为是否打篮球,根据条件属性如天气、温度、湿度、刮风等,来判断是否打篮球;决策属性为某一天零食小贩的收入是否良好,则依据条件属性天气、学生是否放假、保安是否打压、是否进行活动促销等来进行衡量。

[0074] 在本实施例中,基于指标属性确定第一训练数据集的第一总信息熵的步骤包括:

[0075] 确定指标属性中每个指标特征在第一训练数据集中的概率;

[0076] 基于概率计算得到所述第一训练数据集的第一总信息熵。

[0077] 具体地,第一总信息熵采用如下公式进行计算:

[0078] $$Ent(D1) = -\sum_{i=1}^{n} p_i \log_2 (p_i)$$

[0079] 式中,n为指标属性包括n个指标特征,$p_i$表示第i个指标特征在第一训练数据集所有样本中出现的概率。

[0080] 例如,表1所示的第一训练数据集D1中,确定的指标属性为是否关注合作伙伴,则依据条件属性产品类型、访问次数、保单量、投保总额来判断是否关注合作伙伴,其中,是否关注合作伙伴包括两个指标特征:是和否,则第一数据集的第一总信息熵为:

[0081] $$Ent(D1) = -\sum_{i=1}^{n} p_i \log_2 (p_i) = -\frac{9}{15}\log_2\left(\frac{9}{15}\right) - \frac{6}{15}\log_2\left(\frac{6}{15}\right) = 0.9710$$

[0082] 需要说明的是,通过计算信息熵可以反映数据分布的混乱程度,依据数据分布的混乱程度可以适用于高维数据的分类。

[0083] 步骤S203,根据第一总信息熵和条件属性生成节点,基于节点生成决策树模型。

[0084] 在本实施例中,利用信息熵原理选择信息增益最大的条件属性作为分类属性,递归地拓展决策树的分枝,完成决策树的构造。

[0085] 根据第一总信息熵和条件属性确定节点,具体过程如下:

[0086] 步骤A,根据第一总信息熵和条件属性的属性数据,计算得到每个条件属性的信息增益;

[0087] 步骤B,获取条件属性的优化权值,通过优化权值优化对应的信息增益,得到优化信息增益;

[0088] 步骤C,基于优化信息增益确定最优条件属性,将最优条件属性作为节点;

[0089] 步骤D,将节点之外的条件属性组成第二训练数据集,根据指标属性计算第二训练数据集的第二总信息熵;

[0090] 步骤E,循环步骤A至步骤D,直至所有的条件属性生成节点。

[0091] 其中,上述根据第一总信息熵和条件属性的属性数据,计算每个条件属性的信息增益的步骤包括:

[0092] 根据属性数据,计算每个条件属性中每个属性特征的属性信息熵;

[0093] 基于属性信息熵计算得到对应条件属性的条件信息熵；

[0094] 根据第一总信息熵和条件信息熵，计算得到信息增益。

[0095] 在本实施例中，每个条件属性包括m个属性特征，例如，条件属性为产品类型时，属性特征包括财产险、意外险和宠物险等，条件属性为访问次数＞1万次/每天，属性特征包括是、否。

[0096] 确定每个属性特征中每个指标特征的属性概率，基于属性概率计算得到该属性特征的属性信息熵，根据每个属性特征的属性信息熵，计算得到对应的条件属性的条件信息熵，计算第一总信息熵和条件信息熵的差值，得到该条件属性的信息增益。

[0097] 具体地，条件属性的信息增益采用如下计算公式：

[0098] Gain(条件属性)＝Ent(D1)-Ent(D1|条件属性)

[0099] 式中，Ent(D1|条件属性)表示某一条件属性的条件信息熵，条件信息熵的计算公式如下：

[0100] $$\mathrm{Ent}(\mathrm{D1}|\text{条件属性}) = p_j \sum_{j=1}^{m} \mathrm{Ent}(\mathrm{D1}|C_j)$$

[0101] 式中，$p_j$为第j个属性特征$C_j$在该条件属性的所有样本中出现的概率，$\mathrm{Ent}(\mathrm{D1}|C_j)$表示属性特征$C_j$的属性信息熵，$\mathrm{Ent}(\mathrm{D1}|C_j) = -\sum_{i=1}^{n} p_{ij} \log_2(p_{ij})$，其中，$p_{ij}$表示第i个指标特征在属性特征$C_j$所有样本中出现的概率。

[0102] 需要说明的是，信息增益越大说明根据该条件属性划分样本子集的同类性更高，更有利于分类。

[0103] 基于信息增益选取节点生成决策树模型，使得决策树的分类结果具有较好的解释性，并且提高了数据分类的准确性。

[0104] 在一些可选的实现方式中，在上述根据第一总信息熵和条件属性的属性数据，计算得到每个条件属性的信息增益的步骤之前还包括：

[0105] 确定属性数据是否存在异常数据；

[0106] 若存在异常数据，则对异常数据进行修正。

[0107] 在本实施例中，若训练数据集中存在异常数据，对其进行修正，例如，条件属性为保单量＞1000/天的属性数据存在异常数据0值，将其转换为相似值，即将0值转换为“否”；条件属性为投保总额(万)/天的属性数据中存在异常数据0，将其转换为“＜10”。

[0108] 修正之后再进行信息增益计算，可以降低干扰，提高准确度。

[0109] 在本实施例中，计算得到每个条件属性的信息增益后，要对其进行加权修正，获取预设的优化权值优化对应的信息增益。

[0110] 其中，条件属性对应的优化权值为预先配置，可以由业务人员根据重要性进行设置，也可以使用预先训练好的权重生成模型获得，具体根据实际需要进行选择，在此不做限制。

[0111] 在本实施例中，如果属性数据存在异常数据，并对其进行修正，相应的，则需要调整异常数据所对应的条件属性的优化权值，具体步骤如下：

[0112] 确定异常数据对应的条件属性，统计异常数据在条件属性中的占比；

[0113] 根据占比计算得到调整系数，根据调整系数调整条件属性的优化权值。

[0114] 具体地，调整系数的公式如下：

[0115] $$g=\left|1-\frac{\sum_{x\in\tilde{d}}k}{\sum_{x\in D1}K}\right|$$

[0116] 式中，$\sum_{x\in\tilde{d}}k$表示某一条件属性中异常数据的数量，$\Sigma_{x\in D1}K$表示该条件属性中所有样本的数量。假设条件属性A的初始优化权值为$w_1$，调整系数为$g_1$，则调整后的优化权值为$w_1\times g_1$。

[0117] 在本实施例中，通过优化权值优化对应的条件属性的信息增益，得到优化信息增益，将条件属性各自对应的优化信息增益进行比较，选取最大优化信息增益对应的条件属性作为最优条件属性，即最优划分特征，将其作为决策树的根节点。

[0118] 需要说明的是，通过信息增益选取节点可以使得生成的决策树具有较好的分类效果，同时，对信息增益进行加权修正，可以进一步降低干扰，提高分类准确度。

[0119] 在本实施例中，生成根节点后，还需进一步生成子节点和叶子节点，其中，子节点为决策树中间的节点，叶子节点为最底部的节点，也是决策结果，即指标属性。

[0120] 将根节点之外的条件属性和指标属性组成第二训练数据集，根据指标属性计算第二训练数据集的第二总信息熵，计算方法同上述第一训练数据集的第一总信息熵，在此不再赘述。

[0121] 然后，根据第二总信息熵和第二训练数据集中的条件属性的属性数据，计算得到每个条件属性的信息增益，即依次循环步骤A至步骤D，直到生成叶子节点，再根据所有节点生成决策树模型。

[0122] 举例说明，仍以上述表1作为例子进行说明。

[0123] 首先，分别计算四个条件属性产品类型、访问次数、保单量和投保总额对应的属性信息熵，具体如下：

[0124] $$Ent(D1|产品=财产险)=-\frac{4}{5}\log_2(\frac{4}{5})-\frac{1}{5}\log_2(\frac{1}{5})=0.7219$$

[0125] $$Ent(D1|产品=意外险)=-\frac{3}{5}\log_2(\frac{3}{5})-\frac{2}{5}\log_2(\frac{2}{5})=0.9710$$

[0126] $$Ent(D1|产品=宠物险)=-\frac{2}{5}\log_2(\frac{2}{5})-\frac{3}{5}\log_2(\frac{3}{5})=0.9710$$

[0127] 则产品类型的条件信息熵为：

[0128] $$Ent(D1|产品类型)=\frac{5}{15}\times Ent(D1|产品=财产险)+\frac{5}{15}\times Ent(D1|产品=意外险)+\frac{5}{15}\times Ent(D1|产品=宠物险)=0.8880$$

[0129] 进一步计算产品类型的信息增益：

[0130] Gain(产品类型)＝Ent(D1)-Ent(D1|产品类型)＝0.9710-0.8880＝0.083；

[0131] 依照上述公式，依次计算访问次数、保单量和投保总额对应的信息增益：

[0132] Gain(访问次数)＝Ent(D1)-Ent(D1|访问次数)＝0.9710-0.647＝0.324；

[0133] Gain(保单量)＝Ent(D1)-Ent(D1|保单量)＝0.9710-0.551＝0.420；

[0134] Gain(投保总额)＝Ent(D1)-Ent(D1|投保总额)＝0.9710-0.608＝0.363。

[0135] 条件属性“保单量”和“投保总额”中存在异常数据，则对保单量和投保总额的信息增益进行加权修正。

[0136] 需要说明的是，条件属性权重初始值默认为1，则保单量的调整系数为$g=1-\frac{1}{15}=\frac{14}{15}$，投保总额的调整系数为$g=1-\frac{1}{15}=\frac{14}{15}$，则修正后的信息增益为：

[0137] $$Gain'(保单量)=g\times Gain(保单量)=\frac{14}{15}\times 0.420=0.392;$$

[0138] $$Gain'(投保总额)=g\times Gain(投保总额)=\frac{14}{15}\times 0.363=0.339。$$

[0139] 加权修正后，保单量的信息增益最大，将保单量作为第一分类特征，即将其作为根节点。

[0140] 将保单量之外的条件属性产品类型、访问次数和投保总额以及指标属性是否关注合作伙伴组成第二训练数据集D2，第二训练数据集包括两个子集，即保单量>1000/天为是的子集和保单量>1000/天为否的子集，发现“保单量>1000/天”的值为“是”时，“是否关注”值也为“是”，则信息增益Ent(D1|“保单量>1000/天”＝是)＝0，则将对应的属性删除后形成第二训练数据集D2，参见表2所示。

[0141] 也就是说，在组成新的训练数据集之前，先对属性数据进行处理，剔除信息增益为0的属性数据，以提高数据处理效率。

[0142] 表2第二训练数据集D2

[0143]

| 合作伙伴 | 产品 | 访问次数 >1 万次/天 | 投保总额(万)/天 | 是否关注 |
|---|---|---|---|---|
| C | 财产险 | 是 | >10，<100 | 是 |
| D | 财产险 | 是 | >10，<100 | 是 |
| E | 财产险 | 否 | <10 | 否 |
| F | 意外险 | 否 | <10 | 否 |
| G | 意外险 | 否 | >10，<100 | 否 |
| K | 宠物险 | 否 | <10（0） | 否 |
| L | 宠物险 | 否 | >10，<100 | 否 |

[0144]

| | | | | |
|---|---|---|---|---|
| M | 宠物险 | 是 | >10，<100 | 是 |
| O | 宠物险 | 否 | <10 | 否 |

[0145] 计算第二训练数据集D2的第二总信息熵：

[0146] $$Ent(D2)=-\frac{3}{9}\log_2\left(\frac{3}{9}\right)-\frac{6}{9}\log_2\left(\frac{6}{9}\right)=0.9183$$

[0147] 计算第二训练数据集D2中每个条件属性的属性信息熵，计算方法同上，条件属性“产品类型”的属性信息熵计算如下：

[0148] $$Ent(D2|产品=财产险)=-\frac{2}{3}\log_2\left(\frac{2}{3}\right)-\frac{1}{3}\log_2\left(\frac{1}{3}\right)=0.9183$$

[0149] $$Ent(D2|产品=意外险)=-\frac{2}{2}\log_2\left(\frac{2}{2}\right)=0$$

[0150] $$Ent(D2|产品=宠物险)=-\frac{1}{4}\log_2\left(\frac{1}{4}\right)-\frac{3}{4}\log_2\left(\frac{3}{4}\right)=0.8113$$

[0151] 则产品类型的条件信息熵为：

[0152] $$Ent(D2|产品类型)=\frac{3}{9}\times Ent(D2|产品=财产险)+\frac{2}{9}\times Ent(D2|产品=意外险)+\frac{4}{9}\times Ent(D2|产品=宠物险)=0.6667;$$

[0153] 进一步计算产品类型的信息增益：

[0154] Gain(D2|产品类型)＝Ent(D2)-Ent(D2|产品类型)＝0.9183-0.6667＝0.2516

[0155] 条件属性“访问次数”的属性信息熵：

[0156] $$Ent(D2|访问次数=是)=-\frac{3}{3}\log_2\left(\frac{3}{3}\right)-0=0$$

[0157] $$Ent(D2|访问次数=否)=-0-\frac{6}{6}\log_2\left(\frac{6}{6}\right)=0$$

[0158] 则访问次数的条件信息熵为：

[0159] $$Ent(D2|访问次数)=\frac{3}{9}\times Ent(D2|访问次数=是)+\frac{6}{9}\times Ent(D2|访问次数=否)=0;$$

[0160] 则访问次数的信息增益为：

[0161] Gain(D2|访问次数)＝Ent(D2)-Ent(D2|访问次数)＝0.9183

[0162] 条件属性“投保总额”的属性信息熵计算如下：

[0163] $$Ent(D2|投保总额=“<10”)=-0-\frac{4}{4}\log_2\left(\frac{4}{4}\right)=0$$

[0164] $$Ent(D2|投保总额=“>10，<100”)=-\frac{3}{5}\log_2\left(\frac{3}{5}\right)-\frac{2}{5}\log_2\left(\frac{2}{5}\right)=0.9710$$

[0165] 则投保总额的条件信息熵为：

[0166] $$\mathrm{Ent}(D2|\text{投保总额})=\frac{4}{9}\times\mathrm{Ent}(D2|\text{投保总额}=\text{“}<10\text{”})+\frac{5}{9}\times\mathrm{Ent}(D2|\text{投保总额}=\text{“}>10,\ <100\text{”})=0.5394$$

[0167] 则投保总额的信息增益为：

[0168] Gain(D2|投保总额)＝Ent(D2)-Ent(D2|投保总额)＝0.9183-0.5394＝0.3789

[0169] 因为投保总额中存在异常数据，则对投保总额的信息增益进行加权修正：

[0170] $$\mathrm{Gain}'(D2|\text{投保总额})=g\times\mathrm{Gain}(D2|\text{投保总额})=\frac{8}{9}\times0.3789=0.3368$$

[0171] 对比各条件属性的信息增益，条件属性访问次数的信息增益最大，即将访问次数作为第二分类特征，访问次数为根节点之下的中间节点。

[0172] 将访问次数之外的条件属性产品类型和投保总额以及指标属性是否关注合作伙伴组成第三训练数据集，循环上述方法，直到所有的条件属性都形成节点，基于生成的节点得到决策树模型。

[0173] 步骤S204，从数据集中获取验证数据集，通过验证数据集对决策树模型进行验证，得到验证结果。

[0174] 在本实施例中，从数据集中提取子集作为验证数据集，将验证数据集的数据输入决策树模型，输出预测结果，计算预测结果的预测准确度，预测准确度作为验证结果。

[0175] 在本实施例中，通过对生成的决策树模型进行验证，可以提升决策树模型分类的准确性。

[0176] 步骤S205，确定验证结果是否满足预设条件，若验证结果不满足预设条件，则更新决策树模型，直到验证结果满足预设条件，输出最终决策树模型作为分类预测模型。

[0177] 在本实施例中，验证结果为预测准确度，则满足预设条件为预测准确度大于等于预设阈值，不满足预设条件的情况为预测准确度小于预设阈值。

[0178] 如果不满足预设条件，则重新建立决策树模型，重新从数据集中提取第一训练数据集，重复步骤S202至步骤S203，以得到新的决策树模型；如果满足预设条件，输出当前决策树模型作为分类预测模型。

[0179] 步骤S206，获取目标业务数据，将目标业务数据输入分类预测模型，得到分类结果。

[0180] 在本实施例中，将获取到的目标业务数据输出分类预测模型，得到分类结果，根据分类结果进行后续业务决策。

[0181] 业务人员可以根据分类预测模型判断合作伙伴的访问属性、行为信息等，分析合作伙伴的保险产品类型、投保比例等，较为准确地识别重要合作伙伴，例如，新合作伙伴A1连续多天访问保险系统，接口访问次数在10000次/天，根据接口访问次数预测到需要关注此合作伙伴A1，即将其作为重点用户，提醒业务人员持续关注合作伙伴A1业务情况，可针对性的进行业务洽谈和客服支持。

[0182] 本实施例通过分类预测模型对目标业务数据进行决策性判断，提高分类准确性和决策效率。

[0183] 本申请通过获取的数据集提取训练数据集，并根据训练数据集的总信息熵确定节点，基于节点生成决策树模型，将根据生成的决策树模型对目标业务数据进行分类，可以提

高对业务数据的分类效率和分类准确性，进一步高效准确地识别重要合作伙伴，并通过模型中的规则对业务人员预测合作伙伴特征具有一定的辅助作用。

[0184] 本申请可用于众多通用或专用的计算机系统环境或配置中。例如：个人计算机、服务器计算机、手持设备或便携式设备、平板型设备、多处理器系统、基于微处理器的系统、置顶盒、可编程的消费电子设备、网络PC、小型计算机、大型计算机、包括以上任何系统或设备的分布式计算环境等等。本申请可以在由计算机执行的计算机可执行指令的一般上下文中描述，例如程序模块。一般地，程序模块包括执行特定任务或实现特定抽象数据类型的例程、程序、对象、组件、数据结构等等。也可以在分布式计算环境中实践本申请，在这些分布式计算环境中，由通过通信网络而被连接的远程处理设备来执行任务。在分布式计算环境中，程序模块可以位于包括存储设备在内的本地和远程计算机存储介质中。

[0185] 本领域普通技术人员可以理解实现上述实施例方法中的全部或部分流程，是可以通过计算机可读指令来指令相关的硬件来完成，该计算机可读指令可存储于一计算机可读取存储介质中，该程序在执行时，可包括如上述各方法的实施例的流程。其中，前述的存储介质可为磁碟、光盘、只读存储记忆体(Read-Only Memory，ROM)等非易失性存储介质，或随机存储记忆体(Random Access Memory，RAM)等。

[0186] 应该理解的是，虽然附图的流程图中的各个步骤按照箭头的指示依次显示，但是这些步骤并不是必然按照箭头指示的顺序依次执行。除非本文中有明确的说明，这些步骤的执行并没有严格的顺序限制，其可以以其他的顺序执行。而且，附图的流程图中的至少一部分步骤可以包括多个子步骤或者多个阶段，这些子步骤或者阶段并不必然是在同一时刻执行完成，而是可以在不同的时刻执行，其执行顺序也不必然是依次进行，而是可以与其他步骤或者其他步骤的子步骤或者阶段的至少一部分轮流或者交替地执行。

[0187] 进一步参考图3，作为对上述图2所示方法的实现，本申请提供了一种基于决策树模型的数据分类装置的一个实施例，该装置实施例与图2所示的方法实施例相对应，该装置具体可以应用于各种电子设备中。

[0188] 如图3所示，本实施例所述的基于决策树模型的数据分类装置300包括：提取模块301、确定模块302、生成模块303、验证模块304、输出模块305以及分类模块306。其中：

[0189] 提取模块301用于获取历史业务数据，从所述历史业务数据中提取业务特征，根据所述业务特征生成数据集，并从所述数据集中提取子集作为第一训练数据集；

[0190] 确定模块302用于确定所述第一训练数据集中的指标属性和条件属性，基于所述指标属性确定所述第一训练数据集的第一总信息熵；

[0191] 生成模块303用于根据所述第一总信息熵和所述条件属性生成节点，基于所述节点生成决策树模型；

[0192] 验证模块304用于从所述数据集中获取验证数据集，通过所述验证数据集对所述决策树模型进行验证，得到验证结果；

[0193] 输出模块305用于确定所述验证结果是否满足预设条件，若所述验证结果不满足预设条件，则更新所述决策树模型，直到所述验证结果满足预设条件，输出最终决策树模型作为分类预测模型；

[0194] 分类模块306用于获取目标业务数据，将所述目标业务数据输入所述分类预测模型，得到分类结果。

[0195] 基于上述基于决策树模型的数据分类装置，通过获取的数据集提取训练数据集，并根据训练数据集的总信息熵确定节点，基于节点生成决策树模型，将根据生成的决策树模型对目标业务数据进行分类，可以提高对业务数据的分类效率和分类准确性，进一步高效准确地识别重要合作伙伴，并通过模型中的规则对业务人员预测合作伙伴特征具有一定的辅助作用。

[0196] 在本实施例中，确定模块302包括确定子模块和计算子模块，确定子模块用于确定所述指标属性中每个指标特征在所述第一训练数据集中的概率；计算子模块用于基于所述概率计算得到所述第一训练数据集的第一总信息熵。

[0197] 通过计算信息熵可以反映数据分布的混乱程度，依据数据分布的混乱程度可以适用于高维数据的分类。

[0198] 在本实施例中，生成模块303包括第一计算子模块、优化子模块、确定子模块、第二计算子模块以及循环子模块，其中：

[0199] 第一计算子模块用于根据所述第一总信息熵和所述条件属性的属性数据，计算得到每个所述条件属性的信息增益；

[0200] 优化子模块用于获取所述条件属性的优化权值，通过所述优化权值优化对应的信息增益，得到优化信息增益；

[0201] 确定子模块用于基于所述优化信息增益确定最优条件属性，并将所述最优条件属性作为节点；

[0202] 第二计算子模块用于将所述节点之外的条件属性和所述指标属性组成第二训练数据集，根据所述指标属性计算所述第二训练数据集的第二总信息熵；

[0203] 循环子模块用于循环步骤A至步骤D，直至所有的所述条件属性生成节点。

[0204] 在本实施例中，第一计算子模块进一步用于：

[0205] 根据所述属性数据，计算每个所述条件属性中每个属性特征的属性信息熵；

[0206] 基于所述属性信息熵计算得到对应所述条件属性的条件信息熵；

[0207] 根据所述第一总信息熵和所述条件信息熵，计算得到信息增益。

[0208] 通过信息增益选取节点生成决策树模型，使得决策树的分类结果具有较好的解释性，并且提高了数据分类的准确性。

[0209] 在一些可选的实现方式中，确定模块302还包括判断子模块和修正子模块，判断子模块用于确定所述属性数据是否存在异常数据；修正子模块用于若存在异常数据，则对所述异常数据进行修正。

[0210] 修正之后再进行信息增益计算，可以降低干扰，提高准确度。

[0211] 在本实施例中，确定模块302还包括调整子模块，用于：

[0212] 确定所述异常数据对应的条件属性，统计所述异常数据在所述条件属性中的占比；

[0213] 根据所述占比计算得到调整系数，根据所述调整系数调整所述条件属性的优化权值。

[0214] 通过信息增益选取节点可以使得生成的决策树具有较好的分类效果，同时，对信息增益进行加权修正，可以进一步降低干扰，提高分类准确度。

[0215] 在本实施例的一些可选的实现方式中，验证模块304进一步用于：

[0216] 将所述验证数据集输入所述决策树模型，输出预测结果；

[0217] 根据所述预测结果计算预测准确度，将所述预测准确度作为验证结果。

[0218] 通过对生成的决策树模型进行验证，可以提升决策树模型分类的准确性。

[0219] 为解决上述技术问题，本申请实施例还提供计算机设备。具体请参阅图4，图4为本实施例计算机设备基本结构框图。

[0220] 所述计算机设备4包括通过系统总线相互通信连接存储器41、处理器42、网络接口43。需要指出的是，图中仅示出了具有组件41-43的计算机设备4，但是应理解的是，并不要求实施所有示出的组件，可以替代的实施更多或者更少的组件。其中，本技术领域技术人员可以理解，这里的计算机设备是一种能够按照事先设定或存储的指令，自动进行数值计算和/或信息处理的设备，其硬件包括但不限于微处理器、专用集成电路(Application Specific Integrated Circuit，ASIC)、可编程门阵列(Field－Programmable Gate Array，FPGA)、数字处理器(Digital Signal Processor，DSP)、嵌入式设备等。

[0221] 所述计算机设备可以是桌上型计算机、笔记本、掌上电脑及云端服务器等计算设备。所述计算机设备可以与用户通过键盘、鼠标、遥控器、触摸板或声控设备等方式进行人机交互。

[0222] 所述存储器41至少包括一种类型的可读存储介质，所述可读存储介质包括闪存、硬盘、多媒体卡、卡型存储器(例如，SD或DX存储器等)、随机访问存储器(RAM)、静态随机访问存储器(SRAM)、只读存储器(ROM)、电可擦除可编程只读存储器(EEPROM)、可编程只读存储器(PROM)、磁性存储器、磁盘、光盘等。在一些实施例中，所述存储器41可以是所述计算机设备4的内部存储单元，例如该计算机设备4的硬盘或内存。在另一些实施例中，所述存储器41也可以是所述计算机设备4的外部存储设备，例如该计算机设备4上配备的插接式硬盘，智能存储卡(Smart Media Card，SMC)，安全数字(Secure Digital，SD)卡，闪存卡(Flash Card)等。当然，所述存储器41还可以既包括所述计算机设备4的内部存储单元也包括其外部存储设备。本实施例中，所述存储器41通常用于存储安装于所述计算机设备4的操作系统和各类应用软件，例如基于决策树模型的数据分类方法的计算机可读指令等。此外，所述存储器41还可以用于暂时地存储已经输出或者将要输出的各类数据。

[0223] 所述处理器42在一些实施例中可以是中央处理器(Central Processing Unit，CPU)、控制器、微控制器、微处理器、或其他数据处理芯片。该处理器42通常用于控制所述计算机设备4的总体操作。本实施例中，所述处理器42用于运行所述存储器41中存储的计算机可读指令或者处理数据，例如运行所述基于决策树模型的数据分类方法的计算机可读指令。

[0224] 所述网络接口43可包括无线网络接口或有线网络接口，该网络接口43通常用于在所述计算机设备4与其他电子设备之间建立通信连接。

[0225] 本实施例通过处理器执行存储在存储器的计算机可读指令时实现如上述实施例基于决策树模型的数据分类方法的步骤，通过获取的数据集提取训练数据集，并根据训练数据集的总信息熵确定节点，基于节点生成决策树模型，将根据生成的决策树模型对目标业务数据进行分类，可以提高对业务数据的分类效率和分类准确性，进一步高效准确地识别重要合作伙伴，并通过模型中的规则对业务人员预测合作伙伴特征具有一定的辅助作用。

[0226] 本申请还提供了另一种实施方式，即提供一种计算机可读存储介质，所述计算机可读存储介质存储有计算机可读指令，所述计算机可读指令可被至少一个处理器执行，以使所述至少一个处理器执行如上述的基于决策树模型的数据分类方法的步骤，通过获取的数据集提取训练数据集，并根据训练数据集的总信息熵确定节点，基于节点生成决策树模型，将根据生成的决策树模型对目标业务数据进行分类，可以提高对业务数据的分类效率和分类准确性，进一步高效准确地识别重要合作伙伴，并通过模型中的规则对业务人员预测合作伙伴特征具有一定的辅助作用。

[0227] 通过以上的实施方式的描述，本领域的技术人员可以清楚地了解到上述实施例方法可借助软件加必需的通用硬件平台的方式来实现，当然也可以通过硬件，但很多情况下前者是更佳的实施方式。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(如ROM/RAM、磁碟、光盘)中，包括若干指令用以使得一台终端设备(可以是手机，计算机，服务器，空调器，或者网络设备等)执行本申请各个实施例所述的方法。

[0228] 显然，以上所描述的实施例仅仅是本申请一部分实施例，而不是全部的实施例，附图中给出了本申请的较佳实施例，但并不限制本申请的专利范围。本申请可以以许多不同的形式来实现，相反地，提供这些实施例的目的是使对本申请的公开内容的理解更加透彻全面。尽管参照前述实施例对本申请进行了详细的说明，对于本领域的技术人员来而言，其依然可以对前述各具体实施方式所记载的技术方案进行修改，或者对其中部分技术特征进行等效替换。凡是利用本申请说明书及附图内容所做的等效结构，直接或间接运用在其他相关的技术领域，均同理在本申请专利保护范围之内。

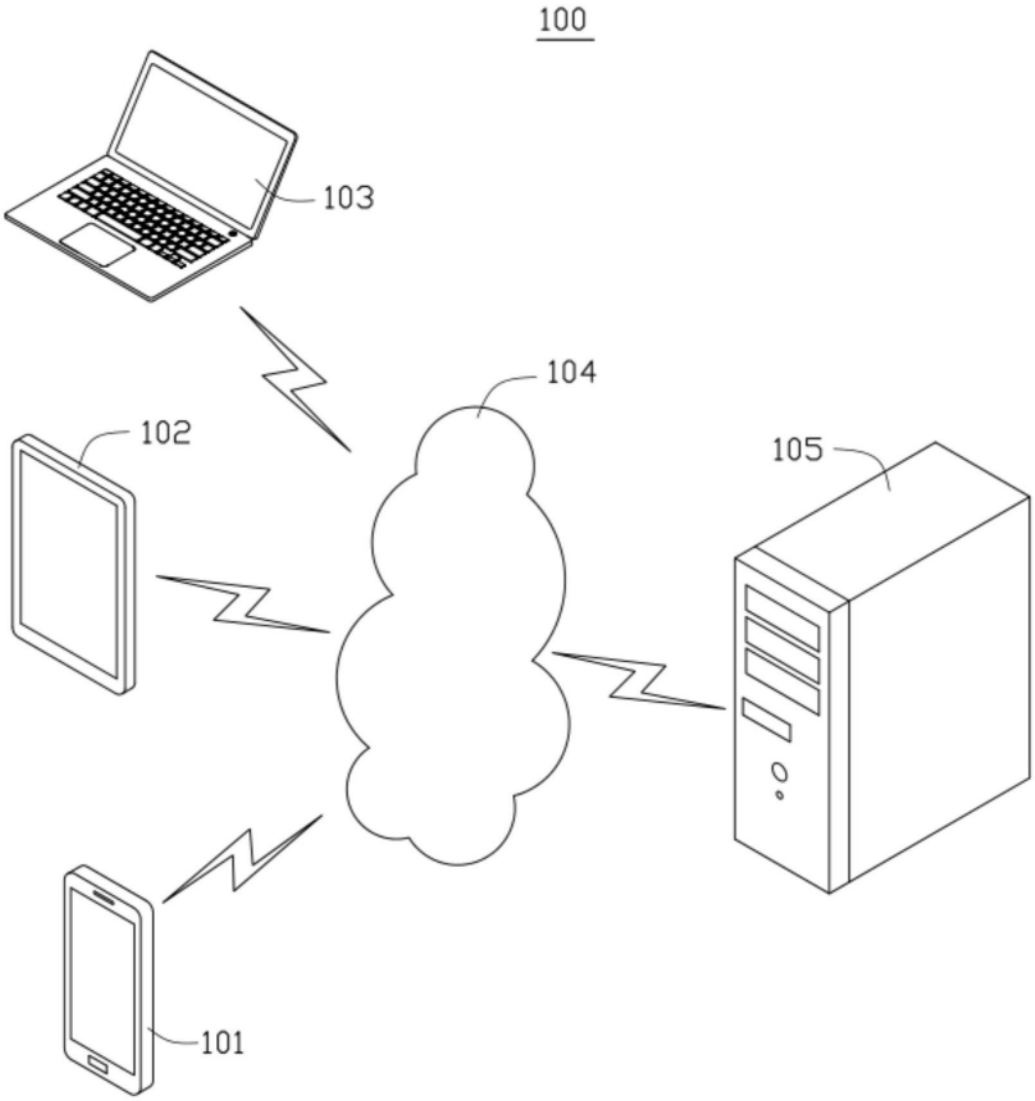



图1

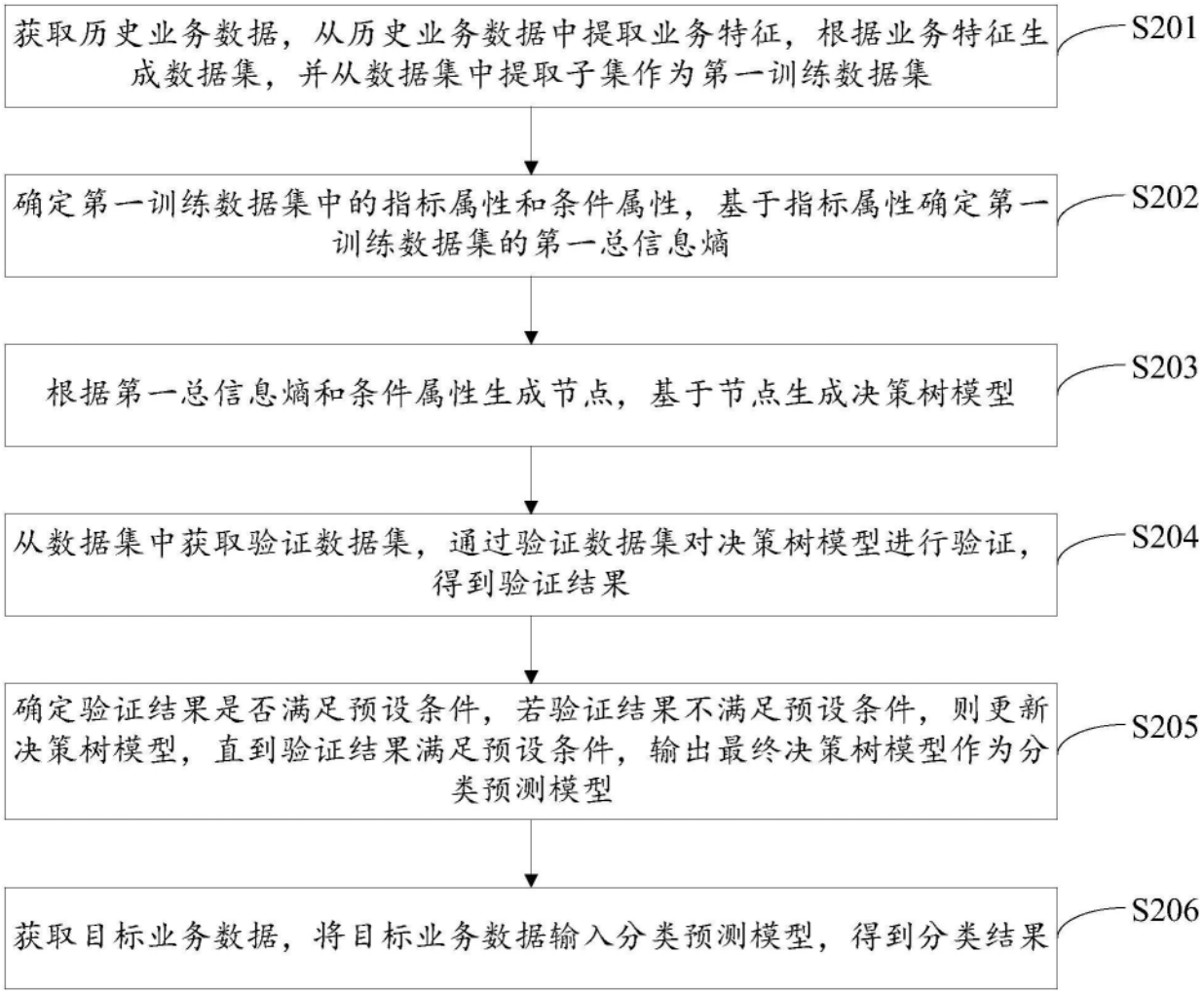

获取历史业务数据，从历史业务数据中提取业务特征，根据业务特征生成数据集，并从数据集中提取子集作为第一训练数据集
S201
确定第一训练数据集中的指标属性和条件属性，基于指标属性确定第一训练数据集的第一总信息熵
S202
根据第一总信息熵和条件属性生成节点，基于节点生成决策树模型
S203
从数据集中获取验证数据集，通过验证数据集对决策树模型进行验证，得到验证结果
S204
确定验证结果是否满足预设条件，若验证结果不满足预设条件，则更新决策树模型，直到验证结果满足预设条件，输出最终决策树模型作为分类预测模型
S205
获取目标业务数据，将目标业务数据输入分类预测模型，得到分类结果
S206

图2

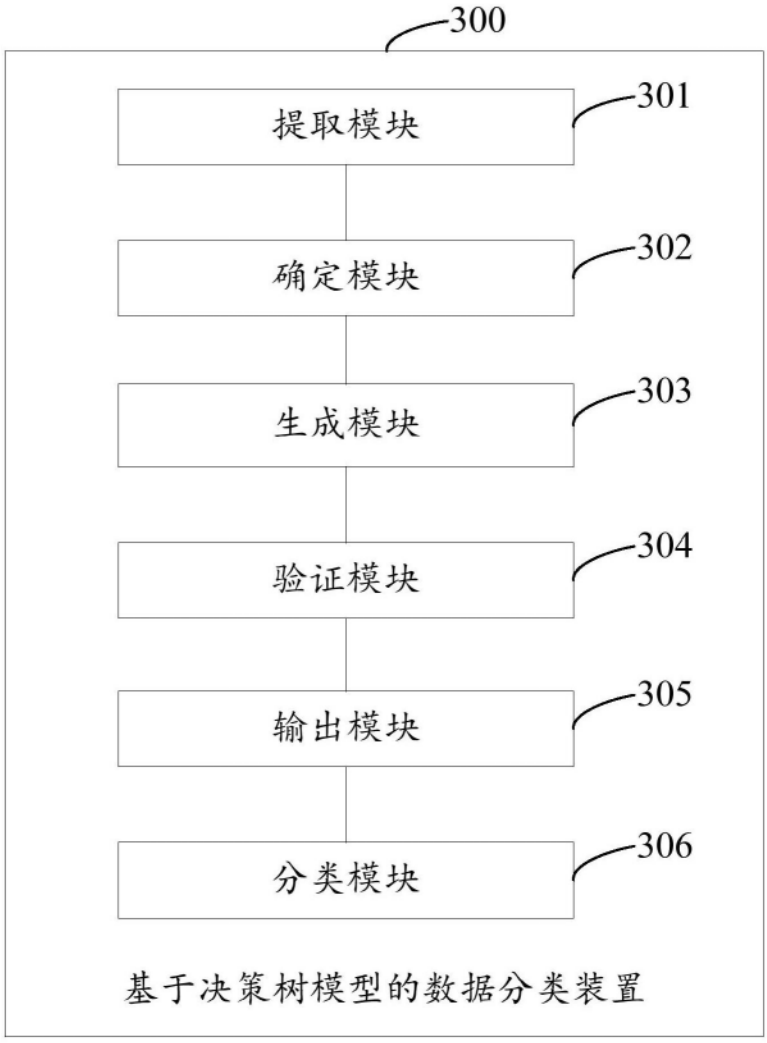

300
提取模块
301
确定模块
302
生成模块
303
验证模块
304
输出模块
305
分类模块
306
基于决策树模型的数据分类装置

图3

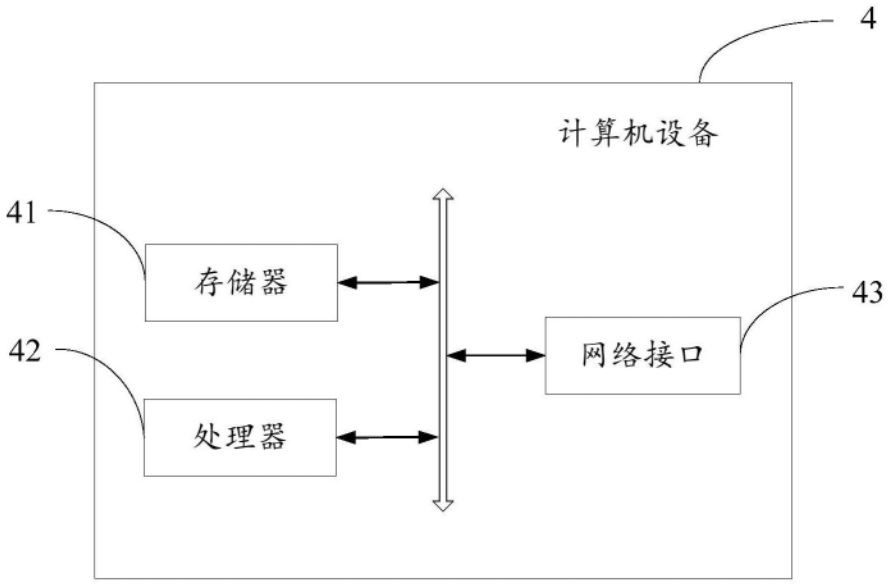



图4