

(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



(10) 国际公布号

WO 2020/048048 A1

(43) 国际公布日

2020 年 3 月 12 日 (12.03.2020)

(51) 国际专利分类号:

G06F 16/35 (2019.01)

(21) 国际申请号:

PCT/CN2018/123208

(22) 国际申请日:

2018 年 12 月 24 日 (24.12.2018)

(25) 申请语言:

中文

(26) 公布语言:

中文

(30) 优先权:

201811018913.0 2018 年 9 月 3 日 (03.09.2018) CN

(71) 申请人: 平安科技(深圳)有限公司(PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) [CN/CN];

中国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, Guangdong 518033 (CN)。

(72) 发明人: 于修铭(YU, Xiuming); 中国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, Guangdong 518033 (CN)。

汪伟(WANG, Wei); 中国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, Guangdong 518033 (CN)。肖京(XIAO, Jing); 中国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, Guangdong 518033 (CN)。

(74) 代理人: 广州华进联合专利商标代理有限公司(ADVANCE CHINA IP LAW OFFICE); 中国广东省广州市天河区珠江东路 6 号 4501 房

(部位: 自编 01-03 和 08-12 单元)(仅限办公用途), Guangdong 510623 (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,

(54) Title: UNBALANCED SAMPLE DATA PREPROCESSING METHOD AND APPARATUS, AND COMPUTER DEVICE

(54) 发明名称: 不平衡样本数据预处理方法、装置和计算机设备

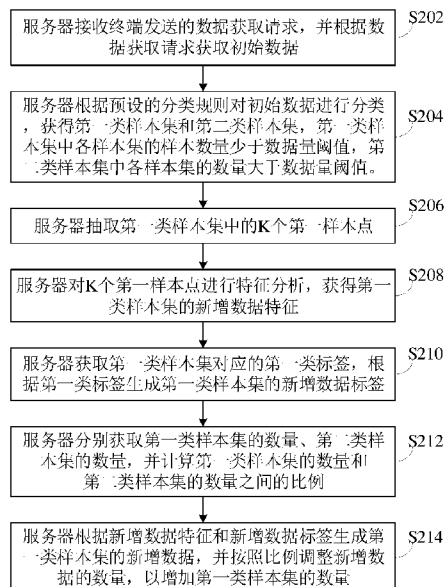


图 2

S202 A server receives a data acquisition request sent by a terminal, and acquires initial data according to the data acquisition request
S204 The server classifies the initial data according to a preset classification rule, and obtains a first type sample set and a second type sample set, wherein the number of samples in each of the first type sample sets is less than a data amount threshold, and the number of each sample set in the second type sample set is greater than the data amount threshold
S206 The server extracts K first sample points in the first type sample set
S208 The server performs feature analysis on the extracted K first sample points to obtain a newly added feature of the first type sample set
S210 The server acquires a first type label corresponding to the first type sample set, and generates a newly added data label of the first type sample set according to the first type label
S212 The server respectively acquires the number of the first type sample sets and the number of the second type sample sets, and calculates a ratio between the number of first type sample sets and the number of second type sample sets
S214 The server generates newly added data of the first type sample set according to the newly added data feature and the newly added data label, and adjusts the amount of the newly added data in proportion to increase the number of the first type sample sets

(57) Abstract: A data resource-based unbalanced sample data preprocessing method, comprising: receiving and acquiring initial data according to a data acquisition request; classifying the initial data according to a preset classification rule, and obtaining a first type sample set and a second type sample set; obtaining a newly added data feature of the first type sample set by performing feature analysis on extracted K first sample points; generating a newly added data label of the first type sample set according to the first type label corresponding to the first type sample set; calculating a ratio between the number of first type sample sets and the number of second type sample sets; generating newly added data of the first type sample set according to the newly added data feature and the newly added data label; and adjusting the amount of the newly added data in proportion to increase the number of the first type sample sets.

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))。

(57) 摘要: 一种基于数据资源的不平衡样本数据预处理方法, 包括: 接收并根据数据获取请求获取初始数据, 根据预设的分类规则对初始数据进行分类, 获得第一类样本集和第二类样本集。对抽取的K个第一样本点进行特征分析, 获得第一类样本集的新增数据特征。根据第一类样本集对应的第一类标签, 生成第一类样本集的新增数据标签。计算第一类样本集的数量和第二类样本集的数量之间的比例。根据新增数据特征和新增数据标签生成第一类样本集的新增数据, 按照比例调整新增数据的数量, 以增加第一类样本集的数量。

不平衡样本数据预处理方法、装置和计算机设备

相关申请的交叉引用

本申请要求于 2018 年 9 月 3 日提交中国专利局，申请号为 2018110189130，

- 5 申请名称为“不平衡样本数据预处理方法、装置和计算机设备”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

技术领域

本申请涉及一种不平衡样本数据预处理方法、装置、计算机设备和存储介质。

10

背景技术

随着计算机技术和社会信息化发展，在实际生产和应用中，需要对海量数据进行数据处理，而获取到的原始数据大多属于非均衡数据，即在整个数据集样本空间中其中一类的样本和其余一类或几类样本在数量上存在巨大的差异。

15

目前对于样本不平衡的情况，即有两类标签对应的数据量差异非常大时，对于样本量少的数据集，通常使用过采样来平衡样本。传统的过采样算法，如 SMOTE 算法，生成的新数据，根据 $X_{\text{new}} = X + \text{rand}(0,1) * (X_{\text{mean}} - X)$ ，得到的新数据是在 (0,1) 之间。

然而，发明人意识到，目前对于有样本不平衡的传统处理方法，生成的新数据不能满足实际数据标准，仍存在样本不平衡的问题。

20

发明内容

根据本申请公开的各种实施例，提供一种不平衡样本数据预处理方法、装置、计算机设备和存储介质。

一种不平衡样本数据预处理方法包括：

25

接收数据获取请求，并根据所述数据获取请求获取初始数据；

根据预设的分类规则对所述初始数据进行分类，获得第一类样本集和第二类样本集，所述第一类样本集中各样本集的样本数量少于数据量阈值；所述第二类样本集中各样本集的数量大于数据量阈值；

抽取所述第一类样本集中的 K 个第一样本点；

30

对 K 个所述第一样本点进行特征分析，获得所述第一类样本集的新增数据特征；

获取第一类样本集对应的第一类标签，并根据所述第一类标签生成所述第一类样本集的新增数据标签；

分别获取所述第一类样本集的数量、所述第二类样本集的数量，并计算所述第一类样本集的数量和所述第二类样本集的数量之间的比例；及

根据所述新增数据特征和所述新增数据标签生成所述第一类样本集的新增数据，并按照所述比例调整所述新增数据的数量，以增加所述第一类样本集的数量。

一种不平衡样本数据预处理装置包括：

初始数据获取模块，用于接收数据获取请求，并根据所述数据获取请求获取初始数据；

5 分类模块，用于根据预设的分类规则对所述初始数据进行分类，获得第一类样本集和
第二类样本集，所述第一类样本集中各样本集的样本数量少于数据量阈值；所述第二类样
本集中各样本集的数量大于数据量阈值；

抽取模块，用于抽取所述第一类样本集中的 K 个第一样本点；

10 新增数据特征获取模块，用于对 K 个所述第一样本点进行特征分析，获得所述第一类
样本集的新增数据特征；

新增数据标签生成模块，用于获取第一类样本集对应的第一类标签，并根据所述第一
类标签生成所述第一类样本集的新增数据标签；

比例计算模块，用于分别获取所述第一类样本集的数量、所述第二类样本集的数量，
并计算所述第一类样本集的数量和所述第二类样本集的数量之间的比例；及

15 新增数据生成模块，用于根据所述新增数据特征和所述新增数据标签生成所述第一类
样本集的新增数据，并按照所述比例调整所述新增数据的数量，以增加所述第一类样本集
的数量。

一种计算机设备，包括存储器和一个或多个处理器，所述存储器中储存有计算机可读
指令，所述计算机可读指令被所述处理器执行时，使得所述一个或多个处理器执行以下步
20 骤：

接收数据获取请求，并根据所述数据获取请求获取初始数据；

根据预设的分类规则对所述初始数据进行分类，获得第一类样本集和第二类样本集，
所述第一类样本集中各样本集的样本数量少于数据量阈值；所述第二类样本集中各样本集
的数量大于数据量阈值；

25 抽取所述第一类样本集中的 K 个第一样本点；

对 K 个所述第一样本点进行特征分析，获得所述第一类样本集的新增数据特征；

获取第一类样本集对应的第一类标签，并根据所述第一类标签生成所述第一类样本集
的新增数据标签；

30 分别获取所述第一类样本集的数量、所述第二类样本集的数量，并计算所述第一类样
本集的数量和所述第二类样本集的数量之间的比例；及

根据所述新增数据特征和所述新增数据标签生成所述第一类样本集的新增数据，并按
照所述比例调整所述新增数据的数量，以增加所述第一类样本集的数量。

一个或多个存储有计算机可读指令的非易失性计算机可读存储介质，计算机可读指令
被一个或多个处理器执行时，使得一个或多个处理器执行以下步骤：

35 接收数据获取请求，并根据所述数据获取请求获取初始数据；

根据预设的分类规则对所述初始数据进行分类，获得第一类样本集和第二类样本集，所述第一类样本集中各样本集的样本数量少于数据量阈值；所述第二类样本集中各样本集的数量大于数据量阈值；

抽取所述第一类样本集中的 K 个第一样本点；

5 对 K 个所述第一样本点进行特征分析，获得所述第一类样本集的新增数据特征；

获取第一类样本集对应的第一类标签，并根据所述第一类标签生成所述第一类样本集的新增数据标签；

分别获取所述第一类样本集的数量、所述第二类样本集的数量，并计算所述第一类样本集的数量和所述第二类样本集的数量之间的比例；及

10 根据所述新增数据特征和所述新增数据标签生成所述第一类样本集的新增数据，并按照所述比例调整所述新增数据的数量，以增加所述第一类样本集的数量。

本申请的一个或多个实施例的细节在下面的附图和描述中提出。本申请的其它特征和优点将从说明书、附图以及权利要求书变得明显。

15 附图说明

为了更清楚地说明本申请实施例中的技术方案，下面将对实施例中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅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它的附图。

图 1 为根据一个或多个实施例中不平衡样本数据预处理方法的应用场景图。

20 图 2 为根据一个或多个实施例中不平衡样本数据预处理方法的流程示意图。

图 3 为根据一个或多个实施例中对 K 个第一样本点进行特征分析，获得新增数据特征的流程示意图。

图 4 为根据一个或多个实施例中不平衡样本数据预处理装置的框图。

图 5 为根据一个或多个实施例中计算机设备的框图。

25

具体实施方式

为了使本申请的技术方案及优点更加清楚明白，以下结合附图及实施例，对本申请进行进一步详细说明。应当理解，此处描述的具体实施例仅仅用以解释本申请，并不用于限定本申请。

30 本申请提供的不平衡样本数据预处理方法，可以应用于如图 1 所示的应用环境中。终端 102 与服务器 104 通过网络进行通信。服务器 104 接收终端 102 发送的数据获取请求，并根据数据获取请求获取初始数据。根据预设的分类规则对初始数据进行分类，获得第一类样本集和第二类样本集。第一类样本集中各样本集的样本数量少于数据量阈值，第二类样本集中各样本集的数量大于数据量阈值。抽取第一类样本集中的 K 个第一样本点，对 K
35 个第一样本点进行特征分析，获得第一类样本集的新增数据特征。获取第一类样本集对应

的第一类标签，并根据第一类标签生成第一类样本集的新增数据标签。分别获取第一类样本集的数量、第二类样本集的数量，并计算第一类样本集的数量和第二类样本集的数量之间的比例。根据新增数据特征和新增数据标签生成第一类样本集的新增数据，并按照比例调整新增数据的数量，以增加第一类样本集的数量。终端 102 可以但不限于是各种个人计算机、笔记本电脑、智能手机、平板电脑和便携式可穿戴设备，服务器 104 可以用独立的服务器或者是多个服务器组成的服务器集群来实现。

在其中一个实施例中，如图 2 所示，提供了一种不平衡样本数据预处理方法，以该方法应用于图 1 中的服务器为例进行说明，包括以下步骤：

S202，服务器接收终端发送的数据获取请求，并根据数据获取请求获取初始数据。

10 具体地，不同数据获取请求与不同初始数据对应，服务器可根据预设的数据获取请求和初始数据之间的对应关系，获取与终端发送的数据获取请求对应的初始数据。数据获取请求携带数据标识，服务器可根据数据标识和初始数据之间的对应关系，获取与数据标识对应的初始数据。

以建模阶段为例，终端需要获取大量样本数据实现模型建立，服务器接收到终端发送的建模数据获取请求，并提取建模数据获取请求携带的建模数据标识，获取预设的建模数据标识和初始数据之间的对应关系，获取与建模数据标识对应的初始数据。

服务器根据数据获取请求从数据库中获取原始数据，并获取原始数据对应的数据类型。获取预设的数据类型和处理方式之间的对应关系，根据处理方式，对各数据类型对应的原始数据进行数据处理，获得初始数据。

20 S204，服务器根据预设的分类规则对初始数据进行分类，获得第一类样本集和第二类样本集，第一类样本集中各样本集的样本数量少于数据量阈值，第二类样本集中各样本集的数量大于数据量阈值。

具体地，预设的分类规则为根据数据量阈值对初始数据进行分类的规则，服务器将样本数量少于数据量阈值的初始数据分为第一类样本集，将样本数量多于数据量阈值的初始数据分为第二类样本集。

以建模阶段为例，可获取到的初始数据大多属于非均衡数据，即在整个数据集样本空间中其中一类的样本和其余一类或几类样本在数量上存在巨大的差异。在本实施例中，服务器可根据预设的数据量阈值，将所获得的初始数据进行分类，得到第一类样本集，即样本数量少于数据量阈值的初始数据。服务器还可将样本数量多于数据量阈值的初始数据，分为第二类样本集。

30 第一类样本集包括多个第一类样本，第一类样本中有多个第一样本点，第一类样本的数量少于预设的数据量阈值，因此，第一类样本集也可称为少类样本集，其对应的样本也可称为少类样本，第一样本点也可称为少类样本点。第二类样本集包括多个第二类样本，第二类样本包括多个第二样本点，第二类样本的数量多于预设的数据量阈值，因此，第二类样本集也可称为多类样本集，其对应的样本也可称为多类样本，第二样本点也可称为多

类样本点。

S206, 服务器抽取第一类样本集中的 K 个第一样本点。

具体地, 第一类样本集中包括多个第一类样本, 第一类样本中包括多个第一样本点。服务器从第一类样本集任意抽取一个样本集, 得到第一样本集。分别计算第一样本集中的全部第一样本点, 至第一类样本集中其他样本集的各个样本点的距离, 并将第一样本点按照距离进行排序, 获取前 K 个第一样本点。

服务器分别计算第一样本集中的全部第一样本点, 至第一类样本集中各个样本点之间的欧氏距离, 并将欧式距离按照从小到大的顺序进行排序, 根据欧式距离的排序, 获取与前 K 个欧氏距离对应的第一样本点。

欧氏距离在 m 维空间中两个点之间的真实距离, 或者向量的自然长度, 即该点到原点的距离。而在二维和三维空间中的欧氏距离就是两点之间的实际距离。将计算得到的欧氏距离按照从小到大的顺序排列, 并获取排序前 K 个欧氏距离对应的第一样本点, 即所获取的第一样本点为与前 K 个最小欧氏距离对应的样本点。

S208, 服务器对 K 个第一样本点进行特征分析, 获得第一类样本集的新增数据特征。

具体地, 服务器提取 K 个第一样本点的特征, 并对特征进行分析处理, 获得特征属性。根据特征属性分别对 K 个第一样本点进行特征提取, 分别获取 K 个第一样本点的公共特征。根据公共特征形成对应的公共特征组合, 并计算公共特征组合所包括的公共特征的数量, 将公共特征组合按照公共特征的数量进行排序, 获取最大数量对应的公共特征组合。根据最大数量对应的公共特征组合, 生成新增数据特征。特征属性表示对第一样本点的所具备的特征进行分析处理后, 得到的关于第一样本点的特征的属性, 包括特征所属类别以及特征的定义。特征提取指的是使用计算机提取图像信息, 决定每个图像的点是否属于一个图像特征。

服务器根据所获得的公共特征, 形成对应的公共特征组合。计算不同公共特征组合所包括的公共特征的数量, 将公共特征组合按照计算得到的公共特征的数量的大小进行排序, 获取最大数量对应的公共特征组合。比如, 不同公共特征组合包括的公共特征的数量为 4、5、3、6、2, 排序后得到的公共特征数量为 6、5、4、3、2, 获得的最大数量对应的公共特征组合, 即为公共特征数量为 6 对应的公共特征组合。服务器获取计算得到的最大数量对应的公共特征组合, 并获取该公共特征组合包括的公共特征, 将多个公共特征作为新增数据特征, 新增数据特征为新增的第一类样本集包括的样本数据对应的特征。

S210, 服务器获取第一类样本集对应的第一类标签, 并根据第一类标签生成第一类样本集的新增数据标签。

具体地, 服务器获取第一类样本集中第一类样本, 根据预设的第一类样本和第一类标签之间的对应关系, 获取与第一类样本对应的第一类标签。第一类样本集包括多个第一类样本, 不同第一类样本与不同第一类标签对应, 服务器可根据第一类样本和第一类标签之间的对应关系, 分别获取与不同第一类样本对应的第一类标签, 根据所获得的第一类标签

生成新增数据标签。新增数据标签表示针对第一类样本集新增的数据携带的标签，可根据新增数据标签和新增数据特征生成第一类样本集的新增数据。

S212，服务器分别获取第一类样本集的数量、第二类样本集的数量，并计算第一类样本集的数量和第二类样本集的数量之间的比例。

- 5 S214，服务器根据新增数据特征和新增数据标签生成第一类样本集的新增数据，并按照比例调整新增数据的数量，以增加第一类样本集的数量。

具体地，服务器获取第一样本集的数量，也就是第一样本集中样本点的数量，并获取第二类样本集的数量，即第二样本集中样本点的数量，并计算第一样本集中样本点的数量和第二样本集中样本点的数量之间的比例，根据所计算得到的比例，调整新增数据的数量，
10 以增加第一类样本集中样本点的数量，达到第一类样本集和第二类样本集之间的平衡状态。服务器根据新增数据特征和新增数据标签生成第一类样本集的新增数据，也就是说，新增数据通过新增数据特征和新增数据标签生成。

当服务器获取的第一样本集中样本点的数量为 20，获取的第二样本集中样本点的数量为 50，计算第一类样本集的数量和第二类样本集的数量之间的比例为 2:5，根据获得的
15 比例，调整新增数据的数量，以生成 30 单位的新增数据来实现第一类样本集和第二类样本集之间的平衡，同时，新增数据通过新增数据特征和新增数据标签生成。

上述不平衡样本数据预处理方法中，服务器根据预设的分类规则对初始数据进行分类，获得第一类样本集和第二类样本集。抽取 K 个第一样本点并进行特征分析，根据获得的第一类样本集的新增数据特征生成第一类样本集的新增数据标签。计算第一类样本集的数量和
20 第二类样本集的数量之间的比例。通过根据新增数据特征和新增数据标签生成第一类样本集的新增数据，并按照比例调整新增数据的数量，以增加第一类样本集的数量，可使得第一类样本集和第二类样本集的数量处于平衡状态，解决了样本不平衡的问题。

在其中一个实施例中，如图 3 所示，提供了一种对 K 个第一样本点进行特征分析，获得新增数据特征的步骤，包括：

- 25 S302，服务器提取 K 个第一样本点的特征。

S304，服务器对特征进行分析处理，获得特征属性。

具体地，服务器通过对 K 个第一样本点的特征进行提取，获得 K 个第一样本点的特征之后，对特征进行分析处理，获得特征的属性，特征的属性包括特征所属类别和特征的定义。以 K 个第一样本点中的任意一个第一样本点为例，服务器提取该第一样本点的特征，
30 对特征进行分析处理，获得相应的特征的属性，包括不同特征所属的类别，以及不同特征各自的定义。

S306，服务器根据特征属性分别对 K 个第一样本点进行特征提取，分别获取 K 个第一样本点的公共特征。

具体地，特征提取指的是使用计算机提取图像信息，决定每个图像的点是否属于一个
35 图像特征。服务器可通过特征提取，获得 K 个第一样本点各自的公共特征。通过对特征所

属类别进行判断，并将不同第一样本点根据特征所属类别进行分类。服务器提取不同特征的定义，并根据特征的定义获取不同特征的交叉点，提取该交叉点，获得不同特征之间的公共特征。

5 S308，服务器根据公共特征形成对应的公共特征组合，并计算公共特征组合所包括的公共特征的数量。

S310，服务器将公共特征组合按照公共特征的数量进行排序，获取最大数量对应的公共特征组合。

具体地，服务器根据所获得的公共特征，形成对应的公共特征组合。计算不同公共特征组合所包括的公共特征的数量，将公共特征组合按照计算得到的公共特征的数量的大小
10 进行排序，获取最大数量对应的公共特征组合。比如，不同公共特征组合包括的公共特征的数量为 2、7、4、6、9，排序后得到的公共特征数量为 9、7、6、4、2，获得的最大数量对应的公共特征组合，即为公共特征数量为 9 对应的公共特征组合。

S312，服务器根据最大数量对应的公共特征组合，生成新增数据特征。

具体地，服务器获取计算得到的最大数量对应的公共特征组合，并获取该公共特征组
15 合包括的公共特征，将多个公共特征作为新增数据特征，新增数据特征为新增的第一类样本集包括的样本数据对应的特征。

上述对 K 个第一样本点进行特征分析，获得新增数据特征的步骤中，服务器对 K 个第一样本点的特征，进行分析处理，获得特征属性，并根据特征属性分别对 K 个第一样本点进行特征提取，分别获取 K 个第一样本点的公共特征。根据公共特征形成对应的公共特
20 征组合，并计算公共特征组合所包括的公共特征的数量，将公共特征组合按照公共特征的数量进行排序，获取最大数量对应的公共特征组合。根据最大数量对应的公共特征组合，生成新增数据特征，可使得生成的新增数据特征，最大程度包括第一类样本集对应的特征，保证新增数据与第一样本集中的样本点的相似程度。

在其中一个实施例中，提供了一种接收数据获取请求，并根据数据获取请求获取初始
25 数据的步骤，包括：

服务器根据数据获取请求从数据库中获取原始数据；获取原始数据对应的数据类型；获取预设的数据类型和处理方式之间的对应关系；根据处理方式，对各数据类型对应的原始数据进行数据处理，获得初始数据。

具体地，不同原始数据对应不同的数据类型，数据类型包括数字型、字节型和文本
30 型等，对应的原始数据包括数字型原始数据、字节型原始数据和文本型原始数据。服务器可根据预设的数据类型和原始数据之间的对应关系，获取原始数据对应的数据类型。不同数据类型对应不同的处理方式，数据类型包括数字型、字节型和文本型等，对应的处理方式为判断处理、赋值处理和声明处理。

进一步地，针对数字型原始数据，执行判断处理，获取预设的取值范围，将预设的
35 取值范围和数字型原始数据的取值进行比对，判断数字型原始数据的取值是否符合预设的

取值范围,提取符合预设取值范围的数字型原始数据,生成数字型初始数据。针对字节型原始数据,执行赋值处理,判断字节型原始数据的取值是否符合预设取值,当字节型原始数据的取值不符合预设取值时,将预设取值赋给对应的字节型原始数据,并根据被赋值后的字节型原始数据,生成字节型初始数据。针对文本型原始数据,执行声明处理,获取文本型原始数据的组成成分,根据组成成分和预设组成成分进行比对,当文本型原始数据的组成成分与预设组成成分不一致时,将文本型原始数据声明为预设组成成分。

上述接收数据获取请求,并根据数据获取请求获取初始数据的步骤,通过按照数据类型对原始数据进行数据处理,按照不同数据类型有针对性地生成初始数据,提高了工作效率。

在其中一个实施例中,提供了一种根据预设的分类规则对初始数据进行分类,获得多个第一类样本集和第二类样本集的步骤,包括:

服务器根据初始数据对应的数据类型,得到各数据类型对应的样本集;计算样本集的初始数据的数据量;将各样本集的数据量与预设的数据量阈值进行比对,当数据量超出数据量阈值时,根据对应的样本集,得到第二类样本集;当数据量低于数据量阈值时,根据对应的样本集,得到第一类样本集。

具体地,服务器根据预设的初始数据和数据类型之间的对应关系,获取与初始数据对应的数据类型,并获取各数据类型对应的样本集。服务器计算样本集的初始数据对应的数据量,并分别将各样本集的数据量与预设的数据量阈值进行比对,判断各样本集的数据量是否超出预设的数据量阈值。当数据量超出数据量阈值时,表示数据量超出数据量阈值对的样本集为第二类样本集。当数据量未超出数据量阈值时,表示数据量未超出数据量阈值对的样本集为第一类样本集。

数据类型包括数字型、字节型和文本型等,对应的样本集为数字型样本集、字节型样本集和文本型样本集。服务器可根据数字型初始数据获取对应的数字型数据类型,并获取对应的数字型样本集。同样地,服务器可分别根据字节型初始数据、文本型初始数据,获取对应的字节型数据类型以及文本型数据类型,进而获取对应的字节型样本集和文本型样本集。

进一步地,可将数据量阈值设置成100个,获得的各样本集的数据量分别为80个、95个、108个和120个,则其中数据量为80个和95个的样本集为第一类样本集,而数据量为108个和120个的样本集为第二类样本集。

上述根据预设的分类规则对初始数据进行分类,获得多个第一类样本集和第二类样本集的步骤,通过对样本集根据数据量多少进行预先分类,获得第一类样本集和第二类样本集,尽快获知需要新增数据的样本集,加快数据新增进程,提高工作效率。

在其中一个实施例中,提供了一种抽取第一类样本集中K个第一样本点的步骤,包括:

服务器从第一类样本集任意抽取一个样本集,得到第一样本集;分别计算第一样本集

中的全部第一样本点，至第一类样本集中其他样本集的各个样本点的距离；将第一样本点按照距离进行排序，获取前 K 个第一样本点。

具体地，第一类样本集中包括多个第一类样本，第一类样本中包括多个第一样本点。服务器从第一类样本集任意抽取一个样本集，得到第一样本集。分别计算第一样本集中的全部第一样本点，至第一类样本集中其他样本集的各个样本点的距离，并将第一样本点按照距离的从小到大进行排序，获取前 K 个第一样本点。

上述抽取第一类样本集中 K 个第一样本点的步骤，通过计算第一样本集中的全部第一样本点，至第一类样本集中其他样本集的各个样本点的距离，并获取前 K 个距离对应的第一样本点，保证所提取的第一样本点为前 K 个距离最小的样本点，提高了样本点提取的精确度。

在其中一个实施例中，提供了一种抽取第一类样本集中 K 个第一样本点的步骤，还包括：

服务器分别计算第一样本集中的全部第一样本点，至第一类样本集中各个样本点之间的欧氏距离；将欧式距离按照从小到大的顺序进行排序；根据欧式距离的排序，获取与前 K 个欧氏距离对应的第一样本点。

具体地，欧氏距离表示在 m 维空间中两个点之间的真实距离，或者向量的自然长度，即该点到原点的距离。服务器通过将计算得到的欧氏距离按照从小到大的顺序排列，并获取排序前 K 个欧氏距离对应的第一样本点，即所获取的第一样本点为与前 K 个最小欧氏距离对应的样本点。

上述抽取第一类样本集中 K 个第一样本点的步骤，服务器分别计算第一样本集中的全部第一样本点，至第一类样本集中各个样本点之间的欧氏距离，并将欧式距离按照从小到大的顺序进行排序。根据欧式距离的排序，获取与前 K 个欧氏距离对应的第一样本点，进一步地保证所提取的第一样本点为前 K 个欧式距离最小的样本点，提高了样本点提取的精确度。

应该理解的是，虽然图 2-3 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，这些步骤可以以其它的顺序执行。而且，图 2-3 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，这些子步骤或者阶段的执行顺序也不必然是依次进行，而是可以与其它步骤或者其它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。

在其中一个实施例中，如图 4 所示，提供了一种不平衡样本数据预处理装置，包括：初始数据获取模块 402、分类模块 404、抽取模块 406、新增数据特征获取模块 408、新增数据标签生成模块 410、比例计算模块 412 和新增数据生成模块 414，其中：

初始数据获取模块 402，用于接收数据获取请求，并根据数据获取请求获取初始数据。

分类模块 404, 用于根据预设的分类规则对初始数据进行分类, 获得第一类样本集和
第二类样本集。第一类样本集中各样本集的样本数量少于数据量阈值, 第二类样本集中各
样本集的数量大于数据量阈值。

抽取模块 406, 用于抽取第一类样本集中的 K 个第一样本点。

- 5 新增数据特征获取模块 408, 用于对 K 个第一样本点进行特征分析, 获得第一类样
本集的新增数据特征。

新增数据标签生成模块 410, 用于获取第一类样本集对应的第一类标签, 并根据第一
类标签生成第一类样本集的新增数据标签。

- 10 比例计算模块 412, 用于分别获取第一类样本集的数量、第二类样本集的数量, 并计
算第一类样本集的数量和第二类样本集的数量之间的比例。

新增数据生成模块 414, 用于根据新增数据特征和新增数据标签生成第一类样本集
的新增数据, 并按照比例调整新增数据的数量, 以增加第一类样本集的数量。

- 15 上述不平衡样本数据预处理装置, 服务器根据预设的分类规则对初始数据进行分类,
获得第一类样本集和第二类样本集。抽取 K 个第一样本点并进行特征分析, 根据获得的第
一类样本集的新增数据特征生成第一类样本集的新增数据标签。计算第一类样本集的数
量和第二类样本集的数量之间的比例。通过根据新增数据特征和新增数据标签生成第一
类样本集的新增数据, 并按照比例调整新增数据的数量, 以增加第一类样本集的数量, 可
使得第一类样本集和第二类样本集的数量处于平衡状态, 解决了样本不平衡的问题。

- 20 在其中一个实施例中, 提供了一种初始数据获取模块, 还用于: 根据数据获取请求
从数据库中获取原始数据; 获取原始数据对应的数据类型; 获取预设的数据类型和处理
方式之间的对应关系; 根据处理方式, 对各数据类型对应的原始数据进行数据处理, 获
得初始数据。

上述初始数据获取模块, 通过按照数据类型对原始数据进行数据处理, 按照不同数
据类型有针对性地生成初始数据, 提高了工作效率。

- 25 在其中一个实施例中, 提供了一种分类模块, 还用于: 根据初始数据对应的数据
类型, 得到各数据类型对应的样本集; 计算样本集的初始数据的数据量; 将各样本集的
数据量与预设的数据量阈值进行比对, 当数据量超出数据量阈值时, 根据对应的样本
集, 得到第二类样本集; 当数据量低于数据量阈值时, 根据对应的样本集, 得到第一
类样本集。

- 30 上述分类模块, 通过对样本集根据数据量多少进行预先分类, 获得第一类样本
集和
第二类样本集, 尽快获知需要新增数据的样本集, 加快数据新增进程, 提高工作效率。

在其中一个实施例中, 提供了一种抽取模块, 还用于: 从第一类样本集任意抽取一
个样本集, 得到第一样本集; 分别计算第一样本集中的全部第一样本点, 至第一类样
本集中其他样本集的各个样本点的距离; 将第一样本点按照距离进行排序, 获取前 K
个第一样本点。

- 35 上述抽取模块, 通过计算第一样本集中的全部第一样本点, 至第一类样本集中其他

样本集的各个样本点的距离，并获取前 K 个距离对应的第一样本点，保证所提取的第一样本点为前 K 个距离最小的样本点，提高了样本点提取的精确度。

在其中一个实施例中，提供了一种抽取模块，还用于：分别计算第一样本集中的全部第一样本点，至第一类样本集中各个样本点之间的欧氏距离；将欧式距离按照从小到大的顺序进行排序；根据欧式距离的排序，获取与前 K 个欧氏距离对应的第一样本点。

上述抽取模块，服务器分别计算第一样本集中的全部第一样本点，至第一类样本集中各个样本点之间的欧氏距离，并将欧式距离按照从小到大的顺序进行排序。根据欧式距离的排序，获取与前 K 个欧氏距离对应的第一样本点，进一步地保证所提取的第一样本点为前 K 个欧式距离最小的样本点，提高了样本点提取的精确度。

在其中一个实施例中，提供了一种新增数据生成模块，还用于：提取 K 个第一样本点的特征；对特征进行分析处理，获得特征属性；根据特征属性分别对 K 个第一样本点进行特征提取，分别获取 K 个第一样本点的公共特征；根据公共特征形成对应的公共特征组合，并计算公共特征组合所包括的公共特征的数量；将公共特征组合按照公共特征的数量进行排序，获取最大数量对应的公共特征组合；根据最大数量对应的公共特征组合，生成新增数据特征。

上述新增数据生成模块，服务器对 K 个第一样本点的特征，进行分析处理，获得特征属性，并根据特征属性分别对 K 个第一样本点进行特征提取，分别获取 K 个第一样本点的公共特征。根据公共特征形成对应的公共特征组合，并计算公共特征组合所包括的公共特征的数量，将公共特征组合按照公共特征的数量进行排序，获取最大数量对应的公共特征组合。根据最大数量对应的公共特征组合，生成新增数据特征，可使得生成的新增数据特征，最大程度包括第一类样本集对应的特征，保证新增数据与第一样本集中的样本点的相似程度。

关于不平衡样本数据预处理装置的具体限定可以参见上文中对于不平衡样本数据预处理方法的限定，在此不再赘述。上述不平衡样本数据预处理装置中的各个模块可全部或部分通过软件、硬件及其组合来实现。上述各模块可以硬件形式内嵌于或独立于计算机设备中的处理器中，也可以以软件形式存储于计算机设备中的存储器中，以便于处理器调用执行以上各个模块对应的操作。

在其中一个实施例中，提供了一种计算机设备，该计算机设备可以是服务器，其内部结构图可以如图 5 所示。该计算机设备包括通过系统总线连接的处理器、存储器、网络接口和数据库。其中，该计算机设备的处理器用于提供计算和控制能力。该计算机设备的存储器包括非易失性计算机可读存储介质、内存储器。该非易失性计算机可读存储介质存储有操作系统、计算机可读指令和数据库。该内存储器为非易失性计算机可读存储介质中的操作系统和计算机可读指令的运行提供环境。该计算机设备的数据库用于存储不平衡样本数据。该计算机设备的网络接口用于与外部的终端通过网络连接通信。该计算机可读指令被处理器执行时以实现一种不平衡样本数据预处理方法。

本领域技术人员可以理解，图 5 中示出的结构，仅仅是与本申请方案相关的部分结构的框图，并不构成对本申请方案所应用于其上的计算机设备的限定，具体的计算机设备可以包括比图中所示更多或更少的部件，或者组合某些部件，或者具有不同的部件布置。一种计算机设备，包括存储器和一个或多个处理器，存储器中存储有计算机可读指令，计算机可读指令被处理器执行时实现本申请任意一个实施例中提供的不平衡样本数据预处理方法的步骤。

一个或多个存储有计算机可读指令的非易失性计算机可读存储介质，计算机可读指令被一个或多个处理器执行时，使得一个或多个处理器实现本申请任意一个实施例中提供的不平衡样本数据预处理方法的步骤。

10 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，所述的计算机可读指令可存储于一非易失性计算机可读存储介质中，该计算机可读指令在执行时，可包括如上述各方法的实施例的流程。其中，本申请所提供的各实施例中所使用的对存储器、存储、数据库或其它介质的任何引用，均可包括非易失性和/或易失性存储器。非易失性存储器可包括只读存储器
15 (ROM)、可编程 ROM (PROM)、电可编程 ROM (EPROM)、电可擦除可编程 ROM (EEPROM) 或闪存。易失性存储器可包括随机存取存储器 (RAM) 或者外部高速缓冲存储器。作为说明而非局限，RAM 以多种形式可得，诸如静态 RAM (SRAM)、动态 RAM (DRAM)、同步 DRAM (SDRAM)、双数据率 SDRAM (DDRSDRAM)、增强型 SDRAM (ESDRAM)、同步链路 (Synchlink) DRAM (SLDRAM)、存储器总线 (Rambus) 直接
20 RAM (RDRAM)、直接存储器总线动态 RAM (DRDRAM)、以及存储器总线动态 RAM (RDRAM) 等。

以上实施例的各技术特征可以进行任意的组合，为使描述简洁，未对上述实施例中的各个技术特征所有可能的组合都进行描述，然而，只要这些技术特征的组合不存在矛盾，都应当认为是本说明书记载的范围。

以上所述实施例仅表达了本申请的几种实施方式，其描述较为具体和详细，但并不能因此而理解为对发明专利范围的限制。应当指出的是，对于本领域的普通技术人员来说，在不脱离本申请构思的前提下，还可以做出若干变形和改进，这些都属于本申请的保护范围。因此，本申请专利的保护范围应以所附权利要求为准。

权利要求书

1、一种不平衡样本数据预处理方法，包括：

接收数据获取请求，并根据所述数据获取请求获取初始数据；

根据预设的分类规则对所述初始数据进行分类，获得第一类样本集和第二类样本集，所述第一类样本集中各样本集的样本数量少于数据量阈值；所述第二类样本集中各样本集

5 的数量大于数据量阈值；

抽取所述第一类样本集中的 K 个第一样本点；

对 K 个所述第一样本点进行特征分析，获得所述第一类样本集的新增数据特征；

获取第一类样本集对应的第一类标签，并根据所述第一类标签生成所述第一类样本集的新增数据标签；

10 分别获取所述第一类样本集的数量、所述第二类样本集的数量，并计算所述第一类样本集的数量和所述第二类样本集的数量之间的比例；及

根据所述新增数据特征和所述新增数据标签生成所述第一类样本集的新增数据，并按照所述比例调整所述新增数据的数量，以增加所述第一类样本集的数量。

2、根据权利要求 1 所述的方法，其特征在于，所述接收数据获取请求，并根据所述

15 数据获取请求获取初始数据，包括：

根据所述数据获取请求从数据库中获取原始数据；

获取所述原始数据对应的数据类型；

获取预设的数据类型和处理方式之间的对应关系；及

根据所述处理方式，对各数据类型对应的原始数据进行数据处理，获得初始数据。

20 3、根据权利要求 1 所述的方法，其特征在于，所述根据预设的分类规则对所述初始数据进行分类，获得多个第一类样本集和第二类样本集，包括：

根据所述初始数据对应的数据类型，得到各数据类型对应的样本集；

计算与所述样本集的初始数据的数据量；

将各样本集的所述数据量与预设的数据量阈值进行比对，当数据量超出所述数据量阈

25 值时，根据对应的样本集，得到第二类样本集；及

当所述数据量低于所述数据量阈值时，根据对应的样本集，得到第一类样本集。

4、根据权利要求 1 至 3 任意一项所述的方法，其特征在于，所述抽取所述第一类样本集中 K 个第一样本点，包括：

从所述第一类样本集任意抽取一个样本集，得到第一样本集；

30 分别计算所述第一样本集中的全部第一样本点，至第一类样本集中其他样本集的各个样本点的距离；及

将所述第一样本点按照所述距离进行排序，获取前 K 个第一样本点。

5、根据权利要求 4 所述的方法，其特征在于，还包括：

分别计算所述第一样本集中的全部第一样本点，至第一类样本集中各个样本点之间的

欧氏距离；

将所述欧式距离按照从小到大的顺序进行排序；及

根据所述欧式距离的排序，获取与所述前 K 个欧氏距离对应的第一样本点。

6、根据权利要求 1 至 3 任意一项所述的方法，其特征在于，所述对 K 个所述第一样

5 本点进行特征分析，获得新增数据特征，包括：

提取 K 个所述第一样本点的特征；

对所述特征进行分析处理，获得特征属性；

根据所述特征属性分别对 K 个所述第一样本点进行特征提取，分别获取 K 个所述第一样本点的公共特征；

10 根据所述公共特征形成对应的公共特征组合，并计算所述公共特征组合所包括的公共特征的数量；

将所述公共特征组合按照所述公共特征的数量进行排序，获取最大数量对应的公共特征组合；及

根据所述最大数量对应的公共特征组合，生成新增数据特征。

15 7、一种不平衡样本数据预处理装置，包括：

初始数据获取模块，用于接收数据获取请求，并根据所述数据获取请求获取初始数据；

分类模块，用于根据预设的分类规则对所述初始数据进行分类，获得第一类样本集和第二类样本集，所述第一类样本集中各样本集的样本数量少于数据量阈值；所述第二类

20 样本集中各样本集的数量大于数据量阈值；

抽取模块，用于抽取所述第一类样本集中的 K 个第一样本点；

新增数据特征获取模块，用于对 K 个所述第一样本点进行特征分析，获得所述第一类样本集的新增数据特征；

25 新增数据标签生成模块，用于获取第一类样本集对应的第一类标签，并根据所述第一类标签生成所述第一类样本集的新增数据标签；

比例计算模块，用于分别获取所述第一类样本集的数量、所述第二类样本集的数量，并计算所述第一类样本集的数量和所述第二类样本集的数量之间的比例；及

新增数据生成模块，用于根据所述新增数据特征和所述新增数据标签生成所述第一类样本集的新增数据，并按照所述比例调整所述新增数据的数量，以增加所述第一类样本

30 集的数量。

8、根据权利要求 7 所述的装置，所述初始数据获取模块，还用于：

根据所述数据获取请求从数据库中获取原始数据；

获取所述原始数据对应的数据类型；

获取预设的数据类型和处理方式之间的对应关系；及

35 根据所述处理方式，对各数据类型对应的原始数据进行数据处理，获得初始数据。

9、根据权利要求 7 所述的装置，其特征在于，所述分类模块还用于：

根据所述初始数据对应的数据类型，得到各数据类型对应的样本集；

计算与所述样本集的初始数据的数据量；

将各样本集的所述数据量与预设的数据量阈值进行比对，当数据量超出所述数据量阈

5 值时，根据对应的样本集，得到第二类样本集；及

当所述数据量低于所述数据量阈值时，根据对应的样本集，得到第一类样本集。

10、一种计算机设备，包括存储器及一个或多个处理器，所述存储器中储存有计算机可读指令，所述计算机可读指令被所述一个或多个处理器执行时，使得所述一个或多个处理器执行以下步骤：

10 接收数据获取请求，并根据所述数据获取请求获取初始数据；

根据预设的分类规则对所述初始数据进行分类，获得第一类样本集和第二类样本集，所述第一类样本集中各样本集的样本数量少于数据量阈值；所述第二类样本集中各样本集的数量大于数据量阈值；

抽取所述第一类样本集中的 K 个第一样本点；

15 对 K 个所述第一样本点进行特征分析，获得所述第一类样本集的新增数据特征；

获取第一类样本集对应的第一类标签，并根据所述第一类标签生成所述第一类样本集的新增数据标签；

分别获取所述第一类样本集的数量、所述第二类样本集的数量，并计算所述第一类样本集的数量和所述第二类样本集的数量之间的比例；及

20 根据所述新增数据特征和所述新增数据标签生成所述第一类样本集的新增数据，并按照所述比例调整所述新增数据的数量，以增加所述第一类样本集的数量。

11、根据权利要求 10 所述的计算机设备，其特征在于，所述处理器执行所述计算机可读指令时还执行以下步骤：

根据所述数据获取请求从数据库中获取原始数据；

25 获取所述原始数据对应的数据类型；

获取预设的数据类型和处理方式之间的对应关系；及

根据所述处理方式，对各数据类型对应的原始数据进行数据处理，获得初始数据。

12、根据权利要求 10 所述的计算机设备，其特征在于，所述处理器执行所述计算机可读指令时还执行以下步骤：根据所述初始数据对应的数据类型，得到各数据类型对应的

30 样本集；

计算与所述样本集的初始数据的数据量；

将各样本集的所述数据量与预设的数据量阈值进行比对，当数据量超出所述数据量阈值时，根据对应的样本集，得到第二类样本集；及

当所述数据量低于所述数据量阈值时，根据对应的样本集，得到第一类样本集。

35 13、根据权利要求 10 至 12 任一项所述的计算机设备，其特征在于，所述处理器执行

所述计算机可读指令时还执行以下步骤：

从所述第一类样本集任意抽取一个样本集，得到第一样本集；

分别计算所述第一样本集中的全部第一样本点，至第一类样本集中其他样本集的各个样本点的距离；及

- 5 将所述第一样本点按照所述距离进行排序，获取前 K 个第一样本点。

14、根据权利要求 13 所述的计算机设备，其特征在于，所述处理器执行所述计算机可读指令时还执行以下步骤：

分别计算所述第一样本集中的全部第一样本点，至第一类样本集中各个样本点之间的欧氏距离；

- 10 将所述欧式距离按照从小到大的顺序进行排序；及

根据所述欧式距离的排序，获取与所述前 K 个欧氏距离对应的第一样本点。

15、根据权利要求 10 至 12 任一项所述的计算机设备，其特征在于，所述处理器执行所述计算机可读指令时还执行以下步骤：

提取 K 个所述第一样本点的特征；

- 15 对所述特征进行分析处理，获得特征属性；

根据所述特征属性分别对 K 个所述第一样本点进行特征提取，分别获取 K 个所述第一样本点的公共特征；

根据所述公共特征形成对应的公共特征组合，并计算所述公共特征组合所包括的公共特征的数量；

- 20 将所述公共特征组合按照所述公共特征的数量进行排序，获取最大数量对应的公共特征组合；及

根据所述最大数量对应的公共特征组合，生成新增数据特征。

16、一个或多个存储有计算机可读指令的非易失性计算机可读存储介质，所述计算机可读指令被一个或多个处理器执行时，使得所述一个或多个处理器执行以下步骤：

接收数据获取请求，并根据所述数据获取请求获取初始数据；

根据预设的分类规则对所述初始数据进行分类，获得第一类样本集和第二类样本集，

- 25 所述第一类样本集中各样本集的样本数量少于数据量阈值；所述第二类样本集中各样本集的数量大于数据量阈值；

抽取所述第一类样本集中的 K 个第一样本点；

对 K 个所述第一样本点进行特征分析，获得所述第一类样本集的新增数据特征；

- 30 获取第一类样本集对应的第一类标签，并根据所述第一类标签生成所述第一类样本集的新增数据标签；

分别获取所述第一类样本集的数量、所述第二类样本集的数量，并计算所述第一类样本集的数量和所述第二类样本集的数量之间的比例；及

根据所述新增数据特征和所述新增数据标签生成所述第一类样本集的新增数据，并按

照所述比例调整所述新增数据的数量，以增加所述第一类样本集的数量。

17、根据权利要求 16 所述的存储介质，其特征在于，所述计算机可读指令被所述处理器执行时还执行以下步骤：

根据所述数据获取请求从数据库中获取原始数据；

5 获取所述原始数据对应的数据类型；

获取预设的数据类型和处理方式之间的对应关系；及

根据所述处理方式，对各数据类型对应的原始数据进行数据处理，获得初始数据。

18、根据权利要求 16 所述的存储介质，其特征在于，所述计算机可读指令被所述处理器执行时还执行以下步骤：

10 根据所述初始数据对应的数据类型，得到各数据类型对应的样本集；

计算与所述样本集的初始数据的数据量；

将各样本集的所述数据量与预设的数据量阈值进行比对，当数据量超出所述数据量阈值时，根据对应的样本集，得到第二类样本集；及

当所述数据量低于所述数据量阈值时，根据对应的样本集，得到第一类样本集。

15 19、根据权利要求 16 至 18 任一项所述的存储介质，其特征在于，所述计算机可读指令被所述处理器执行时还执行以下步骤：

从所述第一类样本集任意抽取一个样本集，得到第一样本集；

分别计算所述第一样本集中的全部第一样本点，至第一类样本集中其他样本集的各个样本点的距离；及

20 将所述第一样本点按照所述距离进行排序，获取前 K 个第一样本点。

20、根据权利要求 16 至 18 任一项所述的存储介质，其特征在于，所述计算机可读指令被所述处理器执行时还执行以下步骤：

分别计算所述第一样本集中的全部第一样本点，至第一类样本集中各个样本点之间的欧氏距离；

25 将所述欧式距离按照从小到大的顺序进行排序；及

根据所述欧式距离的排序，获取与所述前 K 个欧氏距离对应的第一样本点。

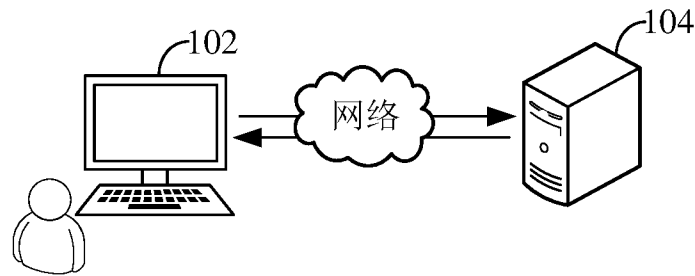


图 1

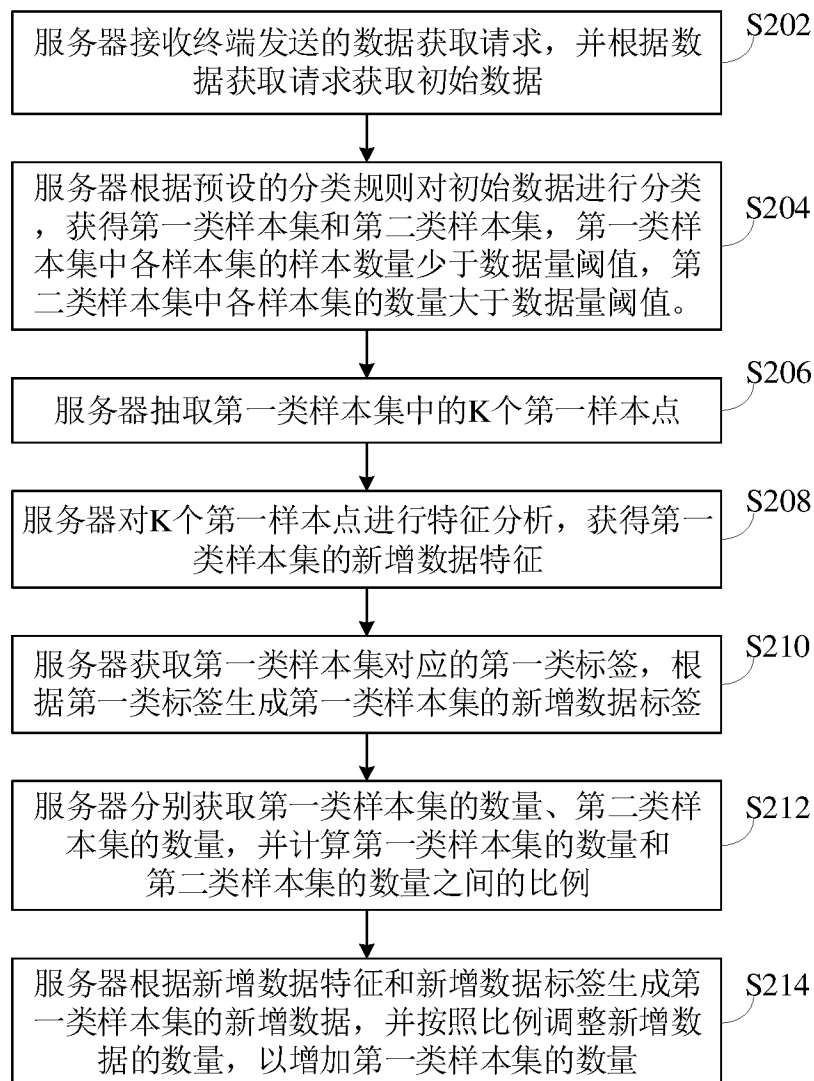


图 2

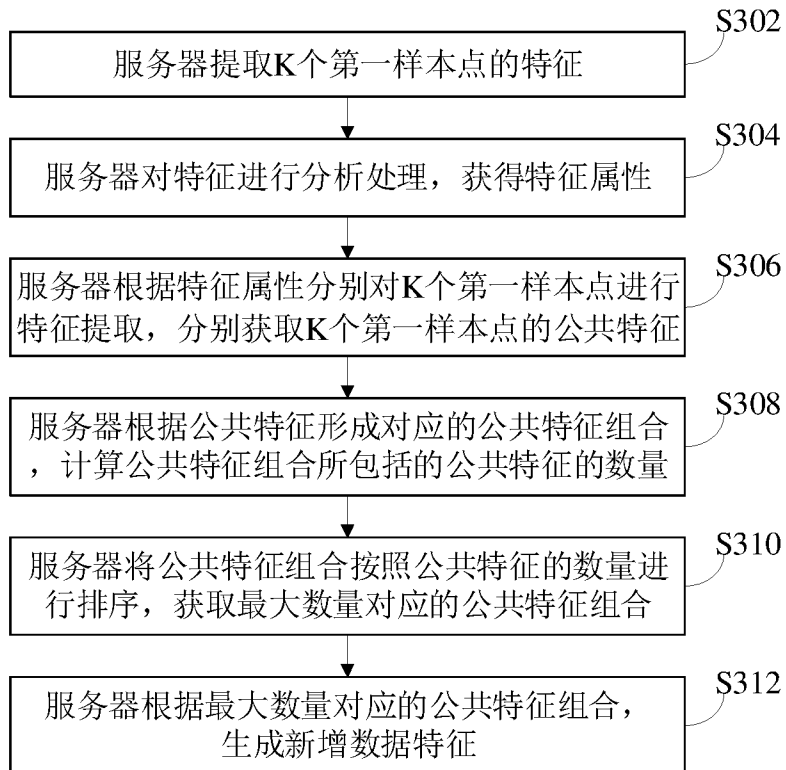


图 3

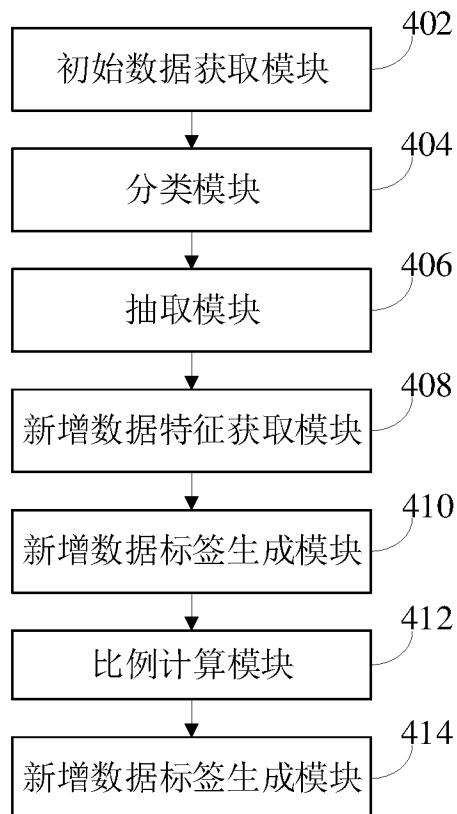


图 4

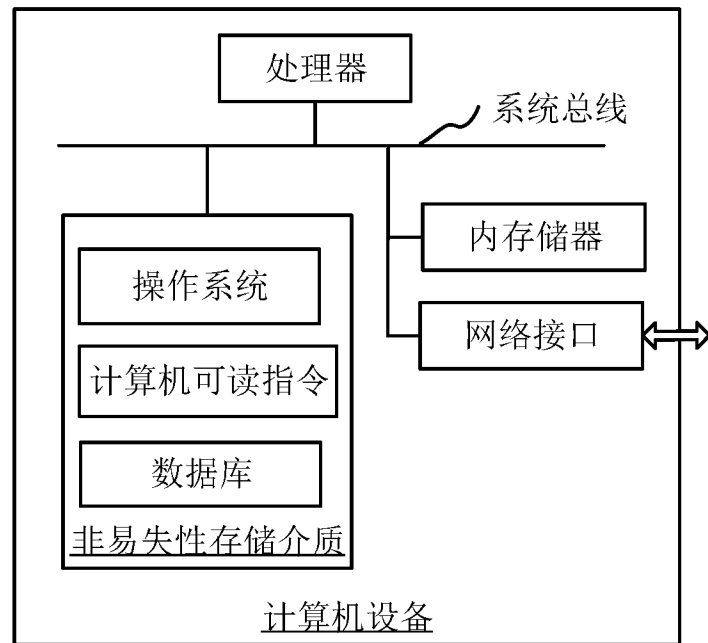


图 5

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2018/123208

A. CLASSIFICATION OF SUBJECT MATTER

G06F 16/35(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)

G06F

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)

CNKI, CNPAT, WPI, EPODOC, GOOGLE, IEEE: 平衡, 均衡, 平均, 样本, 分类, 阈值, 标签, 新增, 增加, 比例, unbalanc+, non-balanc+, balanc+, sample?, type?, tag?, increas+

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 105786970 A (SHENZHEN INSTITUTES OF ADVANCED TECHNOLOGY) 20 July 2016 (2016-07-20) description, paragraphs [0020]-[0038], and figures 1 and 2	1-20
A	CN 107992905 A (NEUSOFT CORPORATION) 04 May 2018 (2018-05-04) entire document	1-20
A	CN 101799748 A (CHINA MOBILE COMMUNICATIONS CORPORATION) 11 August 2010 (2010-08-11) entire document	1-20
A	CN 106599913 A (CHONGQING UNIVERSITY OF POSTS AND TELECOMMUNICATIONS) 26 April 2017 (2017-04-26) entire document	1-20
A	US 2007294223 A1 (TECHNION RESEARCH AND DEVELOPMENT FOUNDATION LTD.) 20 December 2007 (2007-12-20) entire document	1-20

☐ 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

13 May 2019

Date of mailing of the international search report

04 June 2019

Name and mailing address of the ISA/CN

National Intellectual Property Administration, PRC (ISA/CN)
No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088
China

Facsimile No. (86-10)62019451

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2018/123208

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN	105786970	A	20 July 2016	None	
CN	107992905	A	04 May 2018	None	
CN	101799748	A	11 August 2010	None	
CN	106599913	A	26 April 2017	None	
US	2007294223	A1	20 December 2007	None	

国际检索报告

国际申请号

PCT/CN2018/123208

A. 主题的分类

G06F 16/35 (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)

G06F

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))

CNKI, CNPAT, WPI, EPODOC, GOOGLE, IEEE: 平衡, 均衡, 平均, 样本, 分类, 阈值, 标签, 新增, 增加, 比例, unbalanc+, non-balanc+, balanc+, sample?, type?, tag?, increas+

C. 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
A	CN 105786970 A (深圳先进技术研究院) 2016年 7月 20日 (2016 - 07 - 20) 说明书第[0020]-[0038]段, 附图1、2	1-20
A	CN 107992905 A (东软集团股份有限公司) 2018年 5月 4日 (2018 - 05 - 04) 全文	1-20
A	CN 101799748 A (中国移动通信集团公司) 2010年 8月 11日 (2010 - 08 - 11) 全文	1-20
A	CN 106599913 A (重庆邮电大学) 2017年 4月 26日 (2017 - 04 - 26) 全文	1-20
A	US 2007294223 A1 (TECHNION RESEARCH AND DEVELOPMENT FOUNDATION LTD.) 2007年 12月 20日 (2007 - 12 - 20) 全文	1-20

☐ 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

2019年 5月 13日

国际检索报告邮寄日期

2019年 6月 4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

受权官员

吴瑶

传真号 (86-10) 62019451

电话号码 86-10-53961339

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2018/123208

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	105786970	A	2016年 7月 20日	无	
CN	107992905	A	2018年 5月 4日	无	
CN	101799748	A	2010年 8月 11日	无	
CN	106599913	A	2017年 4月 26日	无	
US	2007294223	A1	2007年 12月 20日	无	