(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 111178435 B
(45)授权公告日 2022.03.22

(21)申请号 201911398087.1
(22)申请日 2019.12.30
(65)同一申请的已公布的文献号
申请公布号 CN 111178435 A
(43)申请公布日 2020.05.19
(73)专利权人 山东英信计算机技术有限公司
地址 250001 山东省济南市高新区浪潮路1036号浪潮科技园S05号楼北三层北区
(72)发明人 王刚锋
(74)专利代理机构 北京集佳知识产权代理有限公司 11227
代理人 郄晨芳
(51)Int.Cl.
G06K 9/62(2022.01)
G06F 21/56(2013.01)
(56)对比文件
CN 108596199 A,2018.09.28
CN 107341497 A,2017.11.10
US 2014032450 A1,2014.01.30
CN 110163261 A,2019.08.23
Haibo He等.ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning.《2008 International Joint Conference on Neural Networks (IJCNN 2008)》.2008,第1322-1328页.
曹鹏等.基于决策准则优化的不均衡数据分类.《小型微型计算机系统》.2014,第35卷(第5期),第961-966页.
审查员 刘娜

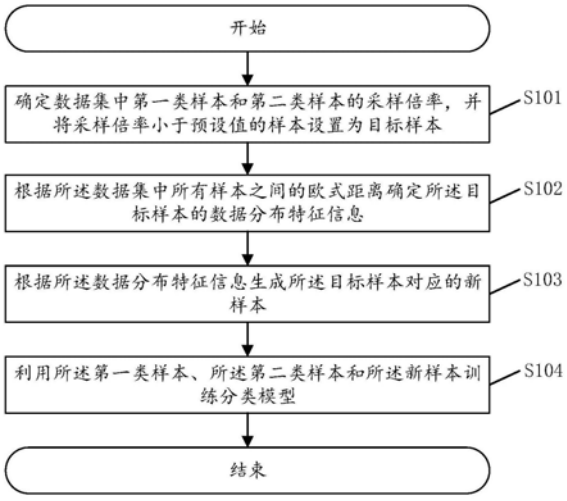
权利要求书3页 说明书8页 附图3页

(54)发明名称

一种分类模型训练方法、系统、电子设备及存储介质

(57)摘要

本申请公开了一种分类模型训练方法,所述分类模型训练方法包括确定数据集中第一类样本和第二类样本的采样倍率,并将采样倍率小于预设值的样本设置为目标样本;根据所述数据集中所有样本之间的欧式距离确定所述目标样本的数据分布特征信息;其中,所述数据分布特征信息为描述近邻样本中同类样本数量的信息,所述近邻样本为欧式距离小于预设距离的两个样本;根据所述数据分布特征信息生成所述目标样本对应的新样本;利用所述第一类样本、所述第二类样本和所述新样本训练分类模型。本申请能够均衡数据集中的各种类样本的数量,提高分类模型的预测准确度。本申请还公开了一种分类模型训练系统、一种电子设备及一种存储介质,具有以上有益效果。



CN 111178435 B

1.一种分类模型训练方法,其特征在于,包括:

确定数据集中第一类样本和第二类样本的采样倍率,并将采样倍率小于预设值的样本设置为目标样本;

根据所述数据集中所有样本之间的欧式距离确定所述目标样本的数据分布特征信息;其中,所述数据分布特征信息为描述近邻样本中同类样本数量的信息,所述近邻样本为欧式距离小于预设距离的两个样本;

根据所述数据分布特征信息生成所述目标样本对应的新样本;

利用所述第一类样本、所述第二类样本和所述新样本训练分类模型;

其中,根据所述目标样本之间的欧氏距离确定所述目标样本的数据分布特征信息包括:

利用第一公式计算任意两个近邻目标样本之间的优良性比值,并将所述优良性比值作为所述数据分布特征信息;其中,所述近邻目标样本为欧式距离小于所述预设距离的两个目标样本;

其中,所述第一公式为$Rat_{im}=Numx_i / Numx_{im}$,$Rat_{im}$为样本$x_i$与样本$x_{im}$之间的优良性比值,$x_i$为所述目标样本中的任一样本,$x_{im}$为样本$x_i$的k个同类近邻样本中第m个近邻样本,$Numx_i$为样本$x_i$的k个近邻样本中目标样本的个数,$Numx_{im}$为样本$x_{im}$的k个近邻样本中目标样本个数;

其中,根据所述数据分布特征信息生成所述目标样本对应的新样本,包括:

当所述优良性比值小于1时,利用第二公式生成所述目标样本对应的新样本x_{newim};其中,所述第二公式为$x_{newim}=x_{im}+rand(0,1)*Rat_{im}*(x_i-x_{im})$;

当所述优良性比值大于1时,利用第三公式生成所述目标样本对应的新样本x_{newim};其中,所述第三公式为$x_{newim}=x_i+rand(0,1)/Rat_{im}*(x_{im}-x_i)$;

当所述优良性比值等于1时,利用第四公式生成所述目标样本对应的新样本x_{newim};其中,所述第四公式为$x_{newim}=x_i+rand(0,1)*(x_{im}-x_i)$;

其中,所述第一类样本为病毒文件样本,所述第二类样本为非病毒文件样本,所述分类模型为文件类型检测模型;

其中,在根据采样结果对所述分类模型执行训练操作之后,还包括:

利用训练后的文件类型检测模型对未知文件执行检测操作生成检测结果,以便根据检测结果判定所述未知文件是否为病毒文件。

2.根据权利要求1所述分类模型训练方法,其特征在于,利用所述第一类样本、所述第二类样本和所述新样本训练分类模型包括:

对所述第一类样本、所述第二类样本和所述新样本执行采样操作,并根据采样结果对所述分类模型执行训练操作。

3.根据权利要求1或2所述分类模型训练方法,其特征在于,确定数据集中第一类样本和第二类样本的采样倍率包括:

根据所述数据集中的样本数量比例确定数据集中第一类样本和第二类样本的采样倍率。

4.一种分类模型训练系统，其特征在于，包括：

目标样本设置模块，用于确定数据集中第一类样本和第二类样本的采样倍率，并将采样倍率小于预设值的样本设置为目标样本；

分布特征确定模块，用于根据所述数据集中所有样本之间的欧式距离确定所述目标样本的数据分布特征信息；其中，所述数据分布特征信息为描述近邻样本中同类样本数量的信息，所述近邻样本为欧式距离小于预设距离的两个样本；

新样本生成模块，用于根据所述数据分布特征信息生成所述目标样本对应的新样本；

模型训练模块，用于利用所述第一类样本、所述第二类样本和所述新样本训练分类模型；

进一步的，分布特征确定模块具体用于利用第一公式计算任意两个近邻目标样本之间的优良性比值，并将所述优良性比值作为所述数据分布特征信息；其中，所述近邻目标样本为欧式距离小于所述预设距离的两个目标样本；

其中，所述第一公式为$Rat_{im}=Numx_i / Numx_{im}$，$Rat_{im}$为样本$x_i$与样本$x_{im}$之间的优良性比值，$x_i$为所述目标样本中的任一样本，$x_{im}$为样本$x_i$的k个同类近邻样本中第m个近邻样本，$Numx_i$为样本$x_i$的k个近邻样本中目标样本的个数，$Numx_{im}$为样本$x_{im}$的k个近邻样本中目标样本个数；

进一步的，新样本生成模块包括：

第一生成单元，用于当所述优良性比值小于1时，利用第二公式生成所述目标样本对应的新样本x_{newim}；

其中，所述第二公式为$x_{newim}=x_{im}+rand(0,1)*Rat_{im}*(x_i-x_{im})$；

第二生成单元，用于当所述优良性比值大于1时，利用第三公式生成所述目标样本对应的新样本x_{newim}；

其中，所述第三公式为$x_{newim}=x_i+rand(0,1)/Rat_{im}*(x_{im}-x_i)$；

第三生成单元，用于当所述优良性比值等于1时，利用第四公式生成所述目标样本对应的新样本x_{newim}；

其中，所述第四公式为$x_{newim}=x_i+rand(0,1)*(x_{im}-x_i)$；

其中，所述第一类样本为病毒文件样本，第二类样本为非病毒文件样本，所述分类模型为文件类型检测模型；

其中，还包括：

病毒检测模块，用于在根据采样结果对所述分类模型执行训练操作之后，利用训练后的文件类型检测模型对未知文件执行检测操作生成检测结果，以便根据检测结果判定所述未知文件是否为病毒文件。

5.一种电子设备，其特征在于，包括存储器和处理器，所述存储器中存储有计算机程序，所述处理器调用所述存储器中的计算机程序时实现如权利要求1至3任一项所述分类模型训练方法的步骤。

6.一种存储介质，其特征在于，所述存储介质中存储有计算机可执行指令，所述计算机可执行指令被处理器加载并执行时，实现如上权利要求1至3任一项所述分类模型训练方法

的步骤。

一种分类模型训练方法、系统、电子设备及存储介质

技术领域

[0001] 本申请涉及机器学习技术领域，特别涉及一种分类模型训练方法、系统、一种电子设备及一种存储介质。

背景技术

[0002] 基于非均衡的数据集构建的预测模型，会对数据集中占比大的类表现出更大的倾向，造成明显的预测误差。目前，针对非均衡数据普遍采用欠缺采样处理或过采样处理以使得非均衡数据类别平衡。欠采样的基本原理是主动丢弃非均衡数据集中类别占比较大的数据，以达到类别占比均衡，但欠采样处理大多会造成数据特性遗失，给最终的预测模型造成预置的误差；传统的过采样模型，往往采取单纯的数据复制，又会使得数据特征偏移，数据分布边缘化加重和增加噪声等问题，这虽然能够均衡数据集，但会使得最终的预测模型产生过拟合和泛化能力差的问题。

[0003] 因此，如何均衡数据集中的各种类样本的数量，提高分类模型的预测准确度是本领域技术人员目前需要解决的技术问题。

发明内容

[0004] 本申请的目的是提供一种分类模型训练方法、系统、一种电子设备及一种存储介质，能够均衡数据集中的各种类样本的数量，提高分类模型的预测准确度。

[0005] 为解决上述技术问题，本申请提供一种分类模型训练方法，该分类模型训练方法包括：

[0006] 确定数据集中第一类样本和第二类样本的采样倍率，并将采样倍率小于预设值的样本设置为目标样本；

[0007] 根据所述数据集中所有样本之间的欧式距离确定所述目标样本的数据分布特征信息；其中，所述数据分布特征信息为描述近邻样本中同类样本数量的信息，所述近邻样本为欧式距离小于预设距离的两个样本；

[0008] 根据所述数据分布特征信息生成所述目标样本对应的新样本；

[0009] 利用所述第一类样本、所述第二类样本和所述新样本训练分类模型。

[0010] 可选的，根据所述目标样本之间的欧氏距离确定所述目标样本的数据分布特征信息包括：

[0011] 利用第一公式计算任意两个近邻目标样本之间的优良性比值，并将所述优良性比值作为所述数据分布特征信息；其中，所述近邻目标样本为欧式距离小于所述预设距离的两个目标样本；

[0012] 其中，所述第一公式为$Rat_{im}=Numx_i/Numx_{im}$，$Rat_{im}$为样本$x_i$与样本$x_{im}$之间的优良性比值，$x_i$为所述目标样本中的任一样本，$x_{im}$为样本$x_i$的k个同类近邻样本中第m个近邻样本，$Numx_i$为样本$x_i$的k个近邻样本中目标样本的个数，$Numx_{im}$为样本$x_{im}$的k个近邻样本中目标样本个数。

[0013] 可选的，根据所述数据分布特征信息生成所述目标样本对应的新样本，包括：

[0014] 当所述优良性比值小于1时，利用第二公式生成所述目标样本对应的新样本x_{newim}；其中，所述第二公式为$x_{newim}=x_{im}+rand(0,1)*Rat_{im}*(x_i-x_{im})$；

[0015] 当所述优良性比值大于1时，利用第三公式生成所述目标样本对应的新样本x_{newim}；其中，所述第三公式为$x_{newim}=x_i+rand(0,1)/Rat_{im}*(x_{im}-x_i)$；

[0016] 当所述优良性比值等于1时，利用第四公式生成所述目标样本对应的新样本x_{newim}；其中，所述第四公式为$x_{newim}=x_i+rand(0,1)*(x_{im}-x_i)$。

[0017] 可选的，利用所述第一类样本、所述第二类样本和所述新样本训练分类模型包括：

[0018] 对所述第一类样本、所述第二类样本和所述新样本执行采样操作，并根据采样结果对所述分类模型执行训练操作。

[0019] 可选的，所述第一类样本为病毒文件样本，第二类样本为非病毒文件样本，所述分类模型为文件类型检测模型。

[0020] 可选的，在根据采样结果对所述分类模型执行训练操作之后，还包括：

[0021] 利用训练后的文件类型检测模型对未知文件执行检测操作生成检测结果，以便根据检测结果判定所述未知文件是否为病毒文件。

[0022] 可选的，确定数据集中第一类样本和第二类样本的采样倍率包括：

[0023] 根据所述数据集中的样本数量比例确定数据集中第一类样本和第二类样本的采样倍率。

[0024] 本申请还提供了一种分类模型训练系统，该分类模型训练系统包括：

[0025] 目标样本设置模块，用于确定数据集中第一类样本和第二类样本的采样倍率，并将采样倍率小于预设值的样本设置为目标样本；

[0026] 分布特征确定模块，用于根据所述数据集中所有样本之间的欧式距离确定所述目标样本的数据分布特征信息；其中，所述数据分布特征信息为描述近邻样本中同类样本数量的信息，所述近邻样本为欧式距离小于预设距离的两个样本；

[0027] 新样本生成模块，用于根据所述数据分布特征信息生成所述目标样本对应的新样本；

[0028] 模型训练模块，用于利用所述第一类样本、所述第二类样本和所述新样本训练分类模型。

[0029] 本申请还提供了一种存储介质，其上存储有计算机程序，所述计算机程序执行时实现上述分类模型训练方法执行的步骤。

[0030] 本申请还提供了一种电子设备，包括存储器和处理器，所述存储器中存储有计算机程序，所述处理器调用所述存储器中的计算机程序时实现上述分类模型训练方法执行的步骤。

[0031] 本申请提供了一种分类模型训练方法，包括确定数据集中第一类样本和第二类样本的采样倍率，并将采样倍率小于预设值的样本设置为目标样本；根据所述数据集中所有样本之间的欧式距离确定所述目标样本的数据分布特征信息；其中，所述数据分布特征信息为描述近邻样本中同类样本数量的信息，所述近邻样本为欧式距离小于预设距离的两个样本；根据所述数据分布特征信息生成所述目标样本对应的新样本；利用所述第一类样本、所述第二类样本和所述新样本训练分类模型。

[0032] 本申请将采样倍率小于预设值的第一类样本或第二类样本设置为目标样本，目标样本为数据集中占比较小的一类样本，若直接使用数据集中的样本训练分类模型就将会导致分类模型对于数据集中占比较大的类别的具有更大的识别倾向，影响识别效果。本申请基于各个样本之间的欧氏距离确定目标样本的数据分布特征信息，根据数据分布特征信息动态的生成与目标样本类别相同的新样本，进而使得数据集中的各个类别的样本数量均衡，避免出现由于样本种类不均衡带来的模型训练效果较差的情况。可见，本申请能够均衡数据集中的各种类样本的数量，提高分类模型的预测准确度。本申请同时还提供了一种分类模型训练系统、一种电子设备和一种存储介质，具有上述有益效果，在此不再赘述。

附图说明

[0033] 为了更清楚地说明本申请实施例，下面将对实施例中所需要使用的附图做简单的介绍，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他的附图。

[0034] 图1为本申请实施例所提供的一种分类模型训练方法的流程图；

[0035] 图2为本申请实施例所提供的一种非均衡数据集的采样方法的流程图；

[0036] 图3为本申请实施例所提供的一种新样本的倾向性示意图；

[0037] 图4为本申请实施例所提供的一种分类模型训练系统的结构示意图。

具体实施方式

[0038] 为使本申请实施例的目的、技术方案和优点更加清楚，下面将结合本申请实施例中的附图，对本申请实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例是本申请一部分实施例，而不是全部的实施例。基于本申请中的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例，都属于本申请保护的范围。

[0039] 下面请参见图1，图1为本申请实施例所提供的一种分类模型训练方法的流程图。

[0040] 具体步骤可以包括：

[0041] S101：确定数据集中第一类样本和第二类样本的采样倍率，并将采样倍率小于预设值的样本设置为目标样本；

[0042] 其中，本步骤中所提到的数据集中可以包括第一类样本和第二类样本，具体的第一类样本可以为正样本，第二类样本可以为负样本，本实施例可以根据所述数据集中的样本数量比例确定数据集中第一类样本和第二类样本的采样倍率，具体的数量比例越大的样本的采样倍率越大。可以理解的是，采样倍率除了与样本自身数量相关，也与训练分类模型时设置的参数相关。

[0043] 本实施例将采样倍率小于预设值的样本设置为目标样本，例如当预设值为1时，若第一类样本的采样倍率小于1时将第一类样本设置为目标样本，若第二类样本的采样倍率小于1时将第二类样本设置为目标样本，当然预设值可以根据实际应用场景灵活设置，此处不进行限定。本步骤的目的是将数据集中占比较少的一类样本设置为目标样本，以便在后续步骤中生成同样类别的新样本，进而均衡数据集中的样本比例。

[0044] S102：根据所述数据集中所有样本之间的欧式距离确定所述目标样本的数据分布特征信息；

[0045] 其中，在本步骤之前可以存在计算数据集中各个样本之间的欧式距离的操作，具体的可以包括第一类样本之间的欧氏距离，可以包括第二类样本之间的欧式距离，还可以包括第一类样本和第二类样本之间的欧氏距离，欧氏距离即欧几里得距离。根据所有样本之间的欧式距离可以得到目标样本的数据分布特征，所述数据分布特征信息为描述近邻样本中同类样本数量的信息，所述近邻样本为欧式距离小于预设距离的两个样本。本实施例将距离某一样本欧式距离小于预设值的所有样本作为该样本的近邻样本，一个样本的近邻样本中可以包同类别的样本也可以包括不同类别的样本。

[0046] S103：根据所述数据分布特征信息生成所述目标样本对应的新样本；

[0047] 其中，本实施例在已经得到数据分布特征信息的基础上生成目标样本对应的新样本。具体的，目标样本分布越密集的区域其样本的噪声越小，同时边缘化问题越小，因此本实施例根据可以根据数据分布特征在目标样本分布越密集的区域对应的新样本。可以理解的是，本步骤用于根据数据集中占比较少的目标样本生成新样本，进而均衡数据集中各个类别的样本数量，作为一种可行的实施方式，本实施例可以根据数据集中第一类样本和第二类样本的样本数量差生成相应数量的新样本，使得将新样本添加至数据集后第一类样本和第二类样本的处于数量均衡的状态。具体的，数量均衡状态指第一类样本与第二类样本的样本数量差在预设范围内的状态。

[0048] S104：利用所述第一类样本、所述第二类样本和所述新样本训练分类模型。

[0049] 其中，在得到了新样本的基础上，本实施可以将新样本添加至数据集中，进而利用数据集中的样本训练分类模型。本实施例中所提到的分类模型可以为人脸识别模型，进而在向分类模型输入一张图片后分类模型能够判定图片中是否包括人脸图像；该分类模型还可以为病毒检测模型，进而在向分类模型中输入未知文件后分类模型能够判定未知文件是否为病毒文件。

[0050] 本实施例将采样倍率小于预设值的第一类样本或第二类样本设置为目标样本，目标样本为数据集中占比较小的一类样本，若直接使用数据集中的样本训练分类模型就将会导致分类模型对于数据集中占比较大的类别的具有更大的识别倾向，影响识别效果。本实施例基于各个样本之间的欧氏距离确定目标样本的数据分布特征信息，根据数据分布特征信息动态的生成与目标样本类别相同的新样本，进而使得数据集中的各个类别的样本数量均衡，避免出现由于样本种类不均衡带来的模型训练效果较差的情况。可见，本实施例能够均衡数据集中的各种类样本的数量，提高分类模型的预测准确度。

[0051] 作为对于图1对应实施例的进一步介绍，图1对应实施例中S102的操作可以具体为利用第一公式计算任意两个近邻目标样本之间的优良性比值，并将所述优良性比值作为所述数据分布特征信息。其中，所述近邻目标样本为欧式距离小于所述预设距离的两个目标样本，上述优良性比值为描述一对近邻目标样本之间区域优良性的信息，本实施例将距离特定样本预设距离内的同类样本数量作为区域优良性的评价标准，数量越多该样本所在的区域优良性越高，该样本所在的区域指距离该样本预设距离内的所有区域范围。例如，样本A存在10个相同类别的近邻样本，样本B存在20个相同类别的近邻样本，此时可以判定样本A的区域优良性高于样本B所在的区域。

[0052] 具体的，上述第一公式为$Rat_{im}=Numx_i/Numx_{im}$，$Rat_{im}$为样本$x_i$与样本$x_{im}$之间的优良性比值，$x_i$为所述目标样本中的任一样本，$x_{im}$为样本$x_i$的k个同类近邻样本中第m个近邻

样本，$Numx_i$为样本x_i的k个近邻样本中目标样本的个数，$Numx_{im}$为样本x_{im}的k个近邻样本中目标样本个数。

[0053] 若将上述将优良性比值作为数据分布特征信息的方法与图1对应的实施例相结合，图1中S103生成新样本的操作可以包括以下步骤：

[0054] 当所述优良性比值小于1时，利用第二公式生成所述目标样本对应的新样本x_{newim}；其中，所述第二公式为$x_{newim}=x_{im}+rand(0,1)*Rat_{im}*(x_i-x_{im})$；

[0055] 当所述优良性比值大于1时，利用第三公式生成所述目标样本对应的新样本x_{newim}；其中，所述第三公式为$x_{newim}=x_i+rand(0,1)/Rat_{im}*(x_{im}-x_i)$；

[0056] 当所述优良性比值等于1时，利用第四公式生成所述目标样本对应的新样本x_{newim}；其中，所述第四公式为$x_{newim}=x_i+rand(0,1)*(x_{im}-x_i)$。

[0057] 在上述实施方式中可以根据数据集中第一类样本和第二类样本的分布特点和趋势，使得新样本产生在更加优良的区域，进而提升分类模型的训练效果。Rand函数指返回大于等于0且小于1的均匀分布随机实数。

[0058] 作为对于图1对应实施例的进一步介绍，S104中训练分类模型的操作可以包括：对所述第一类样本、所述第二类样本和所述新样本执行采样操作，并根据采样结果对所述分类模型执行训练操作。

[0059] 进一步的，图1对应实施例中的第一类样本可以为病毒文件样本，第二类样本可以为非病毒文件样本，分类模型可以为文件类型检测模型。相应的，在根据采样结果对文件类型检测模型执行训练操作之后，还可以利用训练后的文件类型检测模型对未知文件执行检测操作生成检测结果，以便根据检测结果判定所述未知文件是否为病毒文件。

[0060] 下面通过在实际应用中的实施例说明上述实施例描述的流程。请参见图2，图2为本申请实施例所提供的一种非均衡数据集的采样方法的流程图。本实施例描述的了一种非均衡数据集的采样方法，根据非均衡数据的分布特征和采样倍率，基于已有数据集的数据分布特征，动态地生成少数类样本，通过控制样本的生成方式，使新样本产生在更加优良的区域，从而降低样本边缘化加剧的情况，也能减少新样本为噪声的可能性。

[0061] 本实施例向根据样本所在的区域对样本进行分类和评估，根据样本k近邻中同类样本的占比作为分类标准。在构建新样本时，使其对k近邻样本中同类样本占比更大的样本表现出更大的倾向性，从而是新样本产生在更优良和更合理的区域。本实施例的基本思想如下：计算所有少数类样本的k近邻样本；统计该少数类样本的k近邻样本中，同类样本的数量占比，作为评估该样本优良性的标准；统计该样本同类样本中的k近邻样本；根据采样倍率，选取k近邻样本中的N个样本，作为辅助样本；计算该样本和其辅助样本的值，根据值，按照计算规则生成新样本的各个特征值，组合得到新增样本；将新增样本加入到数据集，得到最终的均衡数据集。具体本实施例可以包括以下步骤：

[0062] 步骤1：确定采样倍率；

[0063] 若采样倍率N<＝1，直接按采样倍率N对原少数类样本集随机抽样，以随机抽样结果，作为Tency-SMOTE算法的输出结果；若采样倍率N>1，对采样倍率取整处理并执行下一步。

[0064] 步骤2：计算样本所在区域优良性；

[0065] 样本所在区域优良度是根据该样本k近邻样本中同类样本占比来衡量的，样本所

在区域优良度如下：

[0066] 对于少数类的一个样本x_i，x_{im}表示样本x_i的k个同类近邻中的第m(m<=k)个近邻。$Numx_i$表示样本x_i在同时考虑两类样本的情况下，k个近邻样本中少数类样本的个数，$Numx_{im}$代表样本x_{im}在同时考虑两类样本的情况下，k个近邻中少数类样本的个数。x_{newim}表示通过样本x_i与样本x_{im}扩充的新样本。$Rat_{im}=Numx_i/Numx_{im}$定义为样本$x_i$与样本$x_{im}$优良性比值。如果$Ratx_{im}<1$说明样本$x_{im}$周边的少数类样本比样本$x_i$周边少数类样本分布的更多，也就是样本$x_{im}$所在的区域比样本$x_i$所在的区域更加优良，即通过$Ratx_{im}$值来某样本和其辅助样本的优良性关系。

[0067] 步骤3：依据样本所在区域优良性的不同，采取不同的生成策略；

[0068] 基于上述定义，生成新样本时，应该使新样本对样本x_{im}(或样本x_{im}所在的区域)有更大倾向性，请参见图3，图3为本申请实施例所提供的一种新样本的倾向性示意图。

[0069] 对于样本x_i和其近邻样本x_{im}，由于样本x_i近邻样本中少数类样本占比(或数量)比样本x_{im}近邻样本中负类样本占比高，为了让新样本x_{newim}产生在更优良的区域，所以新产生的样本x_{newim}应该有更大的倾向性偏于样本x_i，即在图3中，新样本x_{newim}应该有更大可能性在直线的左侧。即根据某样本和其辅助样本的$Ratx_{im}$值，采取以下不同的新样本生成策略：

[0070] $$\mathbf{X}new_{im}=\begin{cases} x_{im}+rand(0,1)*\mathrm{Rat}_{im}*(x_i-x_{im}) & ,\ \mathrm{Rat}_{im}<1 \\ x_i+rand(0,1)/\mathrm{Rat}_{im}*(x_{im}-x_i) & ,\ \mathrm{Rat}_{im}>1 \\ x_i+rand(0,1)*(x_{im}-x_i) & ,\ \mathrm{Rat}_{im}=1 \end{cases}。$$

[0071] 对于上述样本生成方法详细分析如下：

[0072] (a)如果$Ratx_{im}<1$，这样的样本x_i可能出现在Boundary classes类样本或Sensitive class类样本中。按照新样本应该处于更优良的少数类区域的原则，此时新扩充的样本x_{newim}应该对样本x_{im}表现出更大的倾向性，即：

[0073] $x_{newim}=x_{im}+rand(0,1)*Rat_{im}*(x_i-x_{im})$；

[0074] (b)如果$Ratx_{im}>1$，Boundary classes类样本或Sensitive class类样本都可能出现这种情况。同理，此时新扩充的样本x_{newim}应该对样本x_i表现出更大的倾向性，即：

[0075] $x_{newim}=x_i+rand(0,1)/Rat_{im}*(x_{im}-x_i)$；

[0076] (c)如果$Ratx_{im}=1$，Boundary classes类样本、Sensitive class类样本都可能出现这种情况，对于所有的Safety class类均满足此条件。同理，此时新扩充的样本x_{newim}应该对样本x_i和样本x_{im}表现出同等的倾向性，即：

[0077] $x_{newim}=x_i+rand(0,1)*(x_{im}-x_i)$；

[0078] 需要说明的是，以上公式也是原始的SMOTE算法公式。

[0079] 步骤4、根据不同的策略，生成新样本；

[0080] 依次遍历某样本与其辅助样本的特征属性，按步骤2中的公式中的某一策略，依次生成新样本的特征值，最终得到新样本。

[0081] 步骤5、完成过采样，输出采样结果。

[0082] 本实施例先得到要进行数据处理的数据集，统计其样本特征的维度和特征值类型。遍历数据集中少数类样本点，并得到每个少数类样本的k近邻样本点，在特征值均衡化

后的基础上，使用python数据处理工具sklearn得到样本的k近邻样本点；根据采样倍率，随机选择N个样本点，作为辅助样本；分别计算该样本点和其辅助样本点的$Ratx_{im}$值，确定新样本的偏移量，根据$Ratx_{im}$值，独立获取样本的各个特征值，各个特征值得到后，组合得到新增样本。最后将所有生成的新样本加入到数据集中，至此得到最终的类均衡数据集。本实施例解决了传统过采样方法中新样本加剧分布边缘化和增加噪声的问题，增强了过采样中新样本产生的合理性，提高了最终模型的准确性和泛化能力等性能。

[0083] 请参见图4，图4为本申请实施例所提供的一种分类模型训练系统的结构示意图；

[0084] 该系统可以包括：

[0085] 目标样本设置模块100，用于确定数据集中第一类样本和第二类样本的采样倍率，并将采样倍率小于预设值的样本设置为目标样本；

[0086] 分布特征确定模块200，用于根据所述数据集中所有样本之间的欧式距离确定所述目标样本的数据分布特征信息；其中，所述数据分布特征信息为描述近邻样本中同类样本数量的信息，所述近邻样本为欧式距离小于预设距离的两个样本；

[0087] 新样本生成模块300，用于根据所述数据分布特征信息生成所述目标样本对应的新样本；

[0088] 模型训练模块400，用于利用所述第一类样本、所述第二类样本和所述新样本训练分类模型。

[0089] 本实施例将采样倍率小于预设值的第一类样本或第二类样本设置为目标样本，目标样本为数据集中占比较小的一类样本，若直接使用数据集中的样本训练分类模型就将会导致分类模型对于数据集中占比较大的类别的具有更大的识别倾向，影响识别效果。本实施例基于各个样本之间的欧氏距离确定目标样本的数据分布特征信息，根据数据分布特征信息动态的生成与目标样本类别相同的新样本，进而使得数据集中的各个类别的样本数量均衡，避免出现由于样本种类不均衡带来的模型训练效果较差的情况。可见，本实施例能够均衡数据集中的各种类样本的数量，提高分类模型的预测准确度。

[0090] 进一步的，分布特征确定模块200具体用于利用第一公式计算任意两个近邻目标样本之间的优良性比值，并将所述优良性比值作为所述数据分布特征信息；其中，所述近邻目标样本为欧式距离小于所述预设距离的两个目标样本；

[0091] 其中，所述第一公式为$Rat_{im}=Numx_i/Numx_{im}$，$Rat_{im}$为样本$x_i$与样本$x_{im}$之间的优良性比值，$x_i$为所述目标样本中的任一样本，$x_{im}$为样本$x_i$的k个同类近邻样本中第m个近邻样本，$Numx_i$为样本$x_i$的k个近邻样本中目标样本的个数，$Numx_{im}$为样本$x_{im}$的k个近邻样本中目标样本个数。

[0092] 进一步的，新样本生成模块300包括：

[0093] 第一生成单元，用于当优良性比值小于1时，利用第二公式生成目标样本对应的新样本x_{newim}；其中，第二公式为$x_{newim}=x_{im}+rand(0,1)*Rat_{im}*(x_i-x_{im})$；

[0094] 第二生成单元，用于当优良性比值大于1时，利用第三公式生成目标样本对应的新样本x_{newim}；其中，第三公式为$x_{newim}=x_i+rand(0,1)/Rat_{im}*(x_{im}-x_i)$；

[0095] 第三生成单元，用于当优良性比值等于1时，利用第四公式生成目标样本对应的新样本x_{newim}；其中，所述第四公式为$x_{newim}=x_i+rand(0,1)*(x_{im}-x_i)$。

[0096] 进一步的，模型训练模块400具体用于对所述第一类样本、所述第二类样本和所述

新样本执行采样操作，并根据采样结果对所述分类模型执行训练操作。

[0097] 进一步的，所述第一类样本为病毒文件样本，第二类样本为非病毒文件样本，所述分类模型为文件类型检测模型。

[0098] 进一步的，还包括：

[0099] 病毒检测模块，用于在根据采样结果对所述分类模型执行训练操作之后，利用训练后的文件类型检测模型对未知文件执行检测操作生成检测结果，以便根据检测结果判定所述未知文件是否为病毒文件。

[0100] 进一步的，目标样本设置模块100包括：

[0101] 采样倍率确定单元，用于根据所述数据集中的样本数量比例确定数据集中第一类样本和第二类样本的采样倍率。

[0102] 由于系统部分的实施例与方法部分的实施例相互对应，因此系统部分的实施例请参见方法部分的实施例的描述，这里暂不赘述。

[0103] 本申请还提供了一种存储介质，其上存有计算机程序，该计算机程序被执行时可以实现上述实施例所提供的步骤。该存储介质可以包括：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。

[0104] 本申请还提供了一种电子设备，可以包括存储器和处理器，所述存储器中存有计算机程序，所述处理器调用所述存储器中的计算机程序时，可以实现上述实施例所提供的步骤。当然所述电子设备还可以包括各种网络接口，电源等组件。

[0105] 说明书中各个实施例采用递进的方式描述，每个实施例重点说明的都是与其他实施例的不同之处，各个实施例之间相同相似部分互相参见即可。对于实施例公开的系统而言，由于其与实施例公开的方法相对应，所以描述的比较简单，相关之处参见方法部分说明即可。应当指出，对于本技术领域的普通技术人员来说，在不脱离本申请原理的前提下，还可以对本申请进行若干改进和修饰，这些改进和修饰也落入本申请权利要求的保护范围内。

[0106] 还需要说明的是，在本说明书中，诸如第一和第二等之类的关系术语仅仅用来将一个实体或者操作与另一个实体或操作区分开来，而不一定要求或者暗示这些实体或操作之间存在任何这种实际的关系或者顺序。而且，术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含，从而使得包括一系列要素的过程、方法、物品或者设备不仅包括那些要素，而且还包括没有明确列出的其他要素，或者是还包括为这种过程、方法、物品或者设备所固有的要素。在没有更多限制的状况下，由语句“包括一个……”限定的要素，并不排除在包括所述要素的过程、方法、物品或者设备中还存在另外的相同要素。

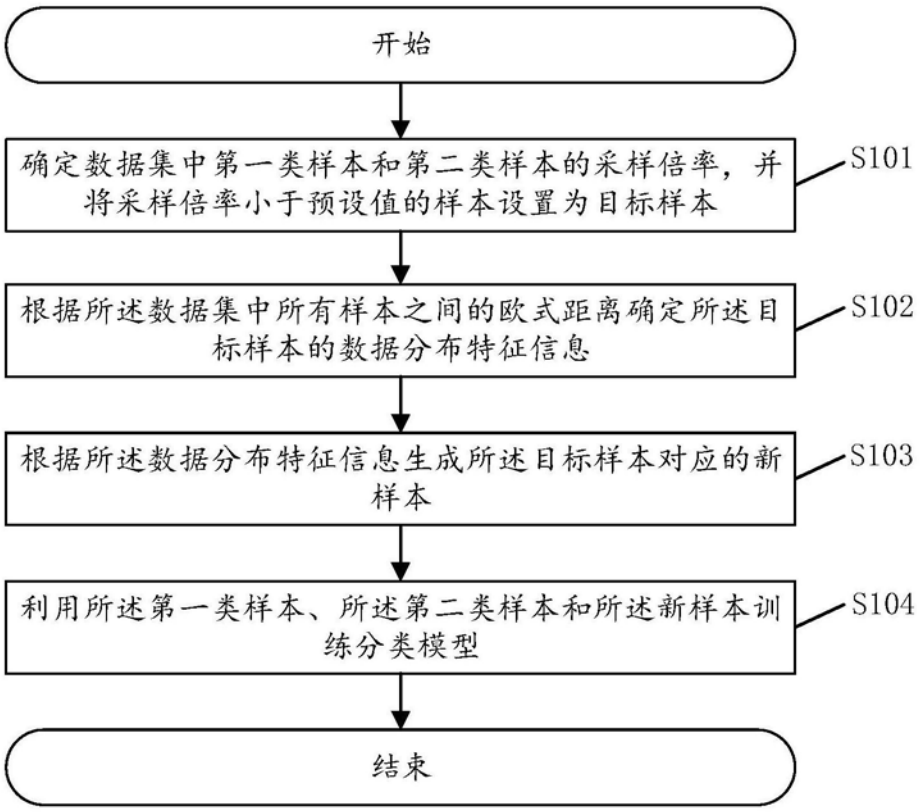

开始
确定数据集中第一类样本和第二类样本的采样倍率，并将采样倍率小于预设值的样本设置为目标样本
S101
根据所述数据集中所有样本之间的欧式距离确定所述目标样本的数据分布特征信息
S102
根据所述数据分布特征信息生成所述目标样本对应的新样本
S103
利用所述第一类样本、所述第二类样本和所述新样本训练分类模型
S104
结束

图1

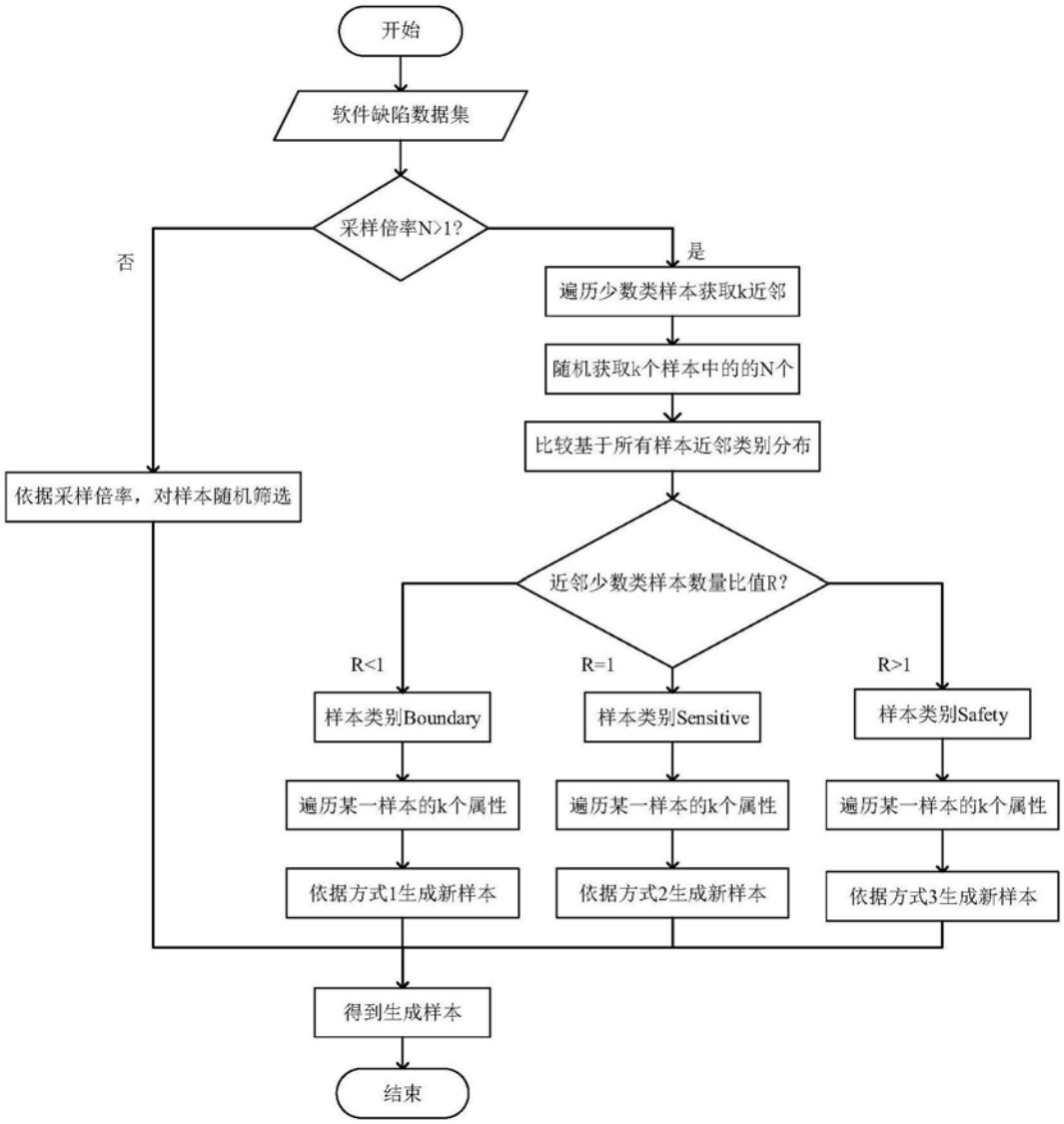

开始
软件缺陷数据集
采样倍率N>1?
否
是
遍历少数类样本获取k近邻
随机获取k个样本中的的N个
比较基于所有样本近邻类别分布
依据采样倍率，对样本随机筛选
近邻少数类样本数量比值R?
R<1
R=1
R>1
样本类别Boundary
样本类别Sensitive
样本类别Safety
遍历某一样本的k个属性
遍历某一样本的k个属性
遍历某一样本的k个属性
依据方式1生成新样本
依据方式2生成新样本
依据方式3生成新样本
得到生成样本
结束

图2

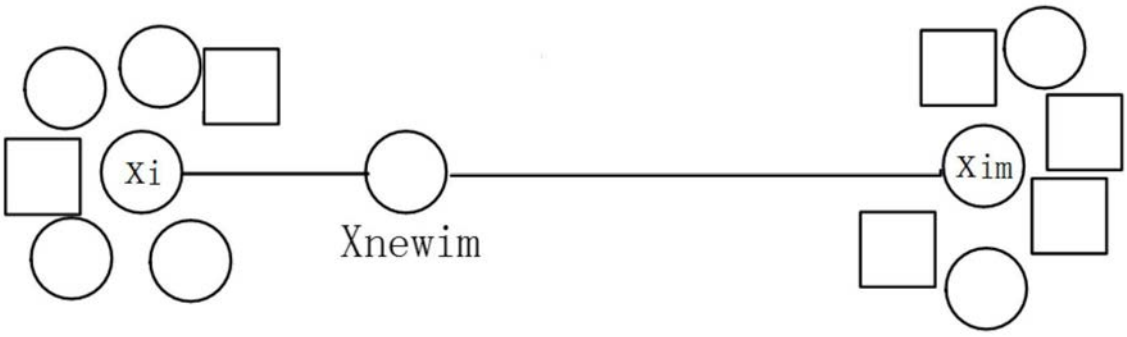

Xi
Xnewim
Xim

图3

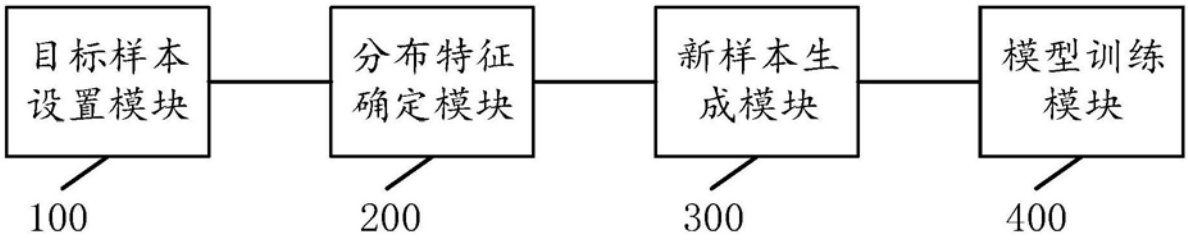


图4