(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 110708285 A
(43)申请公布日 2020.01.17

(21)申请号 201910818633.6

(22)申请日 2019.08.30

(71)申请人 中国平安人寿保险股份有限公司
地址 518000 广东省深圳市福田区益田路5033号平安金融中心14、15、16、41、44、45、46层

(72)发明人 高呈琳

(74)专利代理机构 深圳市隆天联鼎知识产权代理有限公司 44232
代理人 孙强

(51)Int.Cl.
H04L 29/06(2006.01)
G06N 3/08(2006.01)
G06N 3/02(2006.01)

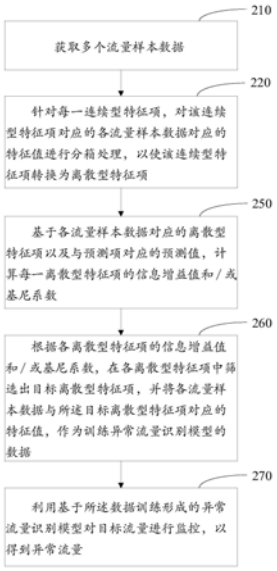
权利要求书3页 说明书13页 附图6页

(54)发明名称

流量监控方法、装置、介质及电子设备

(57)摘要

本发明涉及数据处理领域，揭示了一种基于异常流量识别模型的流量监控方法、装置、介质及电子设备。该方法包括：获取流量样本数据；针对每一连续型特征项，对该连续型特征项对应的各流量样本数据对应的特征值进行分箱，使该连续型特征项转换为离散型特征项；基于流量样本数据对应的离散型特征项和与预测项对应的预测值，计算离散型特征项的信息增益值和/或基尼系数；根据离散型特征项的信息增益值和/或基尼系数，获取目标离散型特征项，并将流量样本数据与目标离散型特征项对应的特征值，作为训练模型的数据；利用基于数据训练成的模型对流量进行监控。此方法下，提高了获取的数据的有效性，提高了模型的性能和用模型进行流量监控的精度。



CN 110708285 A

1.一种基于异常流量识别模型的流量监控方法,其特征在于,所述方法包括:

获取多个流量样本数据,每一所述流量样本数据包括与预测项对应的预测值以及与预设特征项集合中的至少一个特征项对应的特征值,所述预设特征项集合包括多个特征项,与每一特征项对应的特征值为离散值或连续值,其中,对应的特征值为离散值的特征项为离散型特征项,对应的特征值为连续值的特征项为连续型特征项,与预测项对应的预测值指示所述流量样本数据是否为异常流量;

针对每一连续型特征项,对该连续型特征项对应的各流量样本数据对应的特征值进行分箱处理,以使该连续型特征项转换为离散型特征项,其中,经分箱处理后得到的每一流量样本数据所属的箱为将所述连续型特征项转换为的离散型特征项对应的特征值;

基于各流量样本数据对应的离散型特征项以及与预测项对应的预测值,计算每一离散型特征项的信息增益值和/或基尼系数;

根据各离散型特征项的信息增益值和/或基尼系数,在各离散型特征项中筛选出目标离散型特征项,并将各流量样本数据与所述目标离散型特征项对应的特征值,作为训练异常流量识别模型的数据;

利用基于所述数据训练形成的异常流量识别模型对目标流量进行监控,以得到异常流量。

2.根据权利要求1所述的方法,其特征在于,所述获取多个流量样本数据,包括:

获取原始流量样本数据;

对所述原始流量样本数据进行数据清洗,以过滤掉原始流量样本数据中的异常数据,所述异常数据包括缺失值和/或离群值;

在经过数据清洗的原始流量样本数据中获取多个流量样本数据。

3.根据权利要求1所述的方法,其特征在于,所述针对每一连续型特征项,对该连续型特征项对应的各流量样本数据对应的特征值进行分箱处理,以使该连续型特征项转换为离散型特征项,包括:

针对每一连续型特征项,对各流量样本数据中与该连续型特征项对应的特征值进行聚类,以将与该连续型特征项对应的特征值划分为多个类;

针对每一连续型特征项,根据与该连续型特征项对应的特征值被划分为的多个类,对该连续型特征项对应的各流量样本数据对应的特征值进行分箱处理,以使该连续型特征项转换为离散型特征项。

4.根据权利要求1-3任意一项所述的方法,其特征在于,在基于各流量样本数据对应的离散型特征项以及与预测项对应的预测值,计算每一离散型特征项的信息增益值和/或基尼系数之前,所述方法还包括:

对各离散型特征项进行聚类,以将各离散型特征项划分为多个簇;

根据各离散型特征项被划分为的簇,在各离散型特征项中确定出目标离散型特征项;

所述基于各流量样本数据对应的离散型特征项以及与预测项对应的预测值,计算每一离散型特征项的信息增益值和/或基尼系数,包括:

基于各流量样本数据对应的目标离散型特征项以及与预测项对应的预测值,计算每一离散型特征项的信息增益值和/或基尼系数。

5.根据权利要求4所述的方法,其特征在于,所述对各离散型特征项进行聚类,以将各

离散型特征项划分为多个簇，包括：

确定各离散型特征项中每一对离散型特征项之间的皮尔逊相关系数；

利用所述皮尔逊相关系数，对各离散型特征项进行聚类，以将各离散型特征项划分为多个簇。

6.根据权利要求1所述的方法，其特征在于，所述根据各离散型特征项的信息增益值和/或基尼系数，在各离散型特征项中筛选出目标离散型特征项，包括：

根据各离散型特征项的信息增益值和/或基尼系数，在各离散型特征项中筛选出初始特征项；

重复执行初始特征项筛选步骤，得到初始特征项集合，直至重复次数达到第一预定数目，所述初始特征项筛选步骤包括：

利用由所述初始特征项构成的特征项集合建立随机森林，所述特征集合包括多个初始特征项，所述随机森林包括多个决策树，每一决策树包括多个初始特征项；

针对每一初始特征项，确定该初始特征项在每一决策树中的重要程度，并基于该初始特征项在每一决策树中的重要程度确定该初始特征项在所述随机森林中的重要程度；

对各初始特征项按照所述重要程度从高到低的顺序进行排列，获取排序在前第二预定数目的初始特征项，作为初始特征项集合；

将所有所述初始特征项集合的交集作为目标离散型特征项。

7.根据权利要求1所述的方法，其特征在于，在根据各离散型特征项的信息增益值和/或基尼系数，在各离散型特征项中筛选出目标离散型特征项，并将各流量样本数据与所述目标离散型特征项对应的特征值，作为训练异常流量识别模型的数据之后，所述方法还包括：

将获得的所述训练异常流量识别模型的数据输入至逻辑回归模型，以训练形成异常流量识别模型。

8.一种基于异常流量识别模型的流量监控装置，其特征在于，所述装置包括：

获取模块，被配置为获取多个流量样本数据，每一所述流量样本数据包括与预测项对应的预测值以及与预设特征项集合中的至少一个特征项对应的特征值，所述预设特征项集合包括多个特征项，与每一特征项对应的特征值为离散值或连续值，其中，对应的特征值为离散值的特征项为离散型特征项，对应的特征值为连续值的特征项为连续型特征项，与预测项对应的预测值指示所述流量样本数据是否为异常流量；

分箱模块，被配置为针对每一连续型特征项，对该连续型特征项对应的各流量样本数据对应的特征值进行分箱处理，以使该连续型特征项转换为离散型特征项，其中，经分箱处理后得到的每一流量样本数据所属的箱为将所述连续型特征项转换为的离散型特征项对应的特征值；

计算模块，被配置为基于各流量样本数据对应的离散型特征项以及与预测项对应的预测值，计算每一离散型特征项的信息增益值和/或基尼系数；

数据获取模块，被配置为根据各离散型特征项的信息增益值和/或基尼系数，在各离散型特征项中筛选出目标离散型特征项，并将各流量样本数据与所述目标离散型特征项对应的特征值，作为训练异常流量识别模型的数据；

监控模块，被配置为利用基于所述数据训练形成的异常流量识别模型对目标流量进行

监控,以得到异常流量。

9.一种计算机可读程序介质,其特征在于,其存储有计算机程序指令,当所述计算机程序指令被计算机执行时,使计算机执行根据权利要求1至7中任一项所述的方法。

10.一种电子设备,其特征在于,所述电子设备包括:

处理器;

存储器,所述存储器上存储有计算机可读指令,所述计算机可读指令被所述处理器执行时,实现如权利要求1至7任一项所述的方法。

流量监控方法、装置、介质及电子设备

技术领域

[0001] 本公开涉及数据处理领域，特别涉及一种基于异常流量识别模型的流量监控方法、装置、介质及电子设备。

背景技术

[0002] 随着机器学习和人工智能的发展，机器学习模型的应用越来越广泛，例如，在流量监控方面，机器学习模型也有用武之地，可以先通过训练机器学习模型，得到异常流量识别模型，然后利用异常流量识别模型进行异常流量的识别，从而达到流量监控的目的。然而，异常流量识别模型的训练严重依赖于数据，训练异常流量识别模型除了需要大量的流量样本数据之外，还需要对数据进行有针对性地获取，若获取用来训练模型数据不适合训练异常流量识别模型，则可能导致训练出的异常流量识别模型的性能较低，如何获取有效的数据来训练异常流量识别模型，提高训练出的异常流量识别模型的性能，进而提高利用模型监控流量的效果，是本领域亟需解决的问题。

发明内容

[0003] 为了解决相关技术中存在的上述技术问题，本公开提供了一种基于异常流量识别模型的流量监控方法、装置、介质及电子设备。

[0004] 根据本申请的一方面，提供了一种基于异常流量识别模型的流量监控方法，所述方法包括：

[0005] 获取多个流量样本数据，每一所述流量样本数据包括与预测项对应的预测值以及与预设特征项集合中的至少一个特征项对应的特征值，所述预设特征项集合包括多个特征项，与每一特征项对应的特征值为离散值或连续值，其中，对应的特征值为离散值的特征项为离散型特征项，对应的特征值为连续值的特征项为连续型特征项，与预测项对应的预测值指示所述流量样本数据是否为异常流量；

[0006] 针对每一连续型特征项，对该连续型特征项对应的各流量样本数据对应的特征值进行分箱处理，以使该连续型特征项转换为离散型特征项，其中，经分箱处理后得到的每一流量样本数据所属的箱为将所述连续型特征项转换为的离散型特征项对应的特征值；

[0007] 基于各流量样本数据对应的离散型特征项以及与预测项对应的预测值，计算每一离散型特征项的信息增益值和/或基尼系数；

[0008] 根据各离散型特征项的信息增益值和/或基尼系数，在各离散型特征项中筛选出目标离散型特征项，并将各流量样本数据与所述目标离散型特征项对应的特征值，作为训练异常流量识别模型的数据；

[0009] 利用基于所述数据训练形成的异常流量识别模型对目标流量进行监控，以得到异常流量。

[0010] 根据本申请的另一方面，提供了一种基于异常流量识别模型的流量监控装置，所述装置包括：

[0011] 获取模块，被配置为获取多个流量样本数据，每一所述流量样本数据包括与预测项对应的预测值以及与预设特征项集合中的至少一个特征项对应的特征值，所述预设特征项集合包括多个特征项，与每一特征项对应的特征值为离散值或连续值，其中，对应的特征值为离散值的特征项为离散型特征项，对应的特征值为连续值的特征项为连续型特征项，与预测项对应的预测值指示所述流量样本数据是否为异常流量；

[0012] 分箱模块，被配置为针对每一连续型特征项，对该连续型特征项对应的各流量样本数据对应的特征值进行分箱处理，以使该连续型特征项转换为离散型特征项，其中，经分箱处理后得到的每一流量样本数据所属的箱为将所述连续型特征项转换为的离散型特征项对应的特征值；

[0013] 计算模块，被配置为基于各流量样本数据对应的离散型特征项以及与预测项对应的预测值，计算每一离散型特征项的信息增益值和/或基尼系数；

[0014] 数据获取模块，被配置为根据各离散型特征项的信息增益值和/或基尼系数，在各离散型特征项中筛选出目标离散型特征项，并将各流量样本数据与所述目标离散型特征项对应的特征值，作为训练异常流量识别模型的数据；

[0015] 监控模块，被配置为利用基于所述数据训练形成的异常流量识别模型对目标流量进行监控，以得到异常流量。

[0016] 根据本申请的另一方面，提供了一种计算机可读程序介质，其存储有计算机程序指令，当所述计算机程序指令被计算机执行时，使计算机执行如前所述的方法。

[0017] 根据本申请的另一方面，提供了一种电子设备，所述电子设备包括：

[0018] 处理器；

[0019] 存储器，所述存储器上存储有计算机可读指令，所述计算机可读指令被所述处理器执行时，实现如前所述的方法。

[0020] 本发明的实施例提供的技术方案可以包括以下有益效果：对于本发明所提供的基于异常流量识别模型的流量监控方法，该方法包括：获取多个流量样本数据，每一所述流量样本数据包括与预测项对应的预测值以及与预设特征项集合中的至少一个特征项对应的特征值，所述预设特征项集合包括多个特征项，与每一特征项对应的特征值为离散值或连续值，其中，对应的特征值为离散值的特征项为离散型特征项，对应的特征值为连续值的特征项为连续型特征项，与预测项对应的预测值指示所述流量样本数据是否为异常流量；针对每一连续型特征项，对该连续型特征项对应的各流量样本数据对应的特征值进行分箱处理，以使该连续型特征项转换为离散型特征项，其中，经分箱处理后得到的每一流量样本数据所属的箱为将所述连续型特征项转换为的离散型特征项对应的特征值；基于各流量样本数据对应的离散型特征项以及与预测项对应的预测值，计算每一离散型特征项的信息增益值和/或基尼系数；根据各离散型特征项的信息增益值和/或基尼系数，在各离散型特征项中筛选出目标离散型特征项，并将各流量样本数据与所述目标离散型特征项对应的特征值，作为训练异常流量识别模型的数据；利用基于所述数据训练形成的异常流量识别模型对目标流量进行监控，以得到异常流量。

[0021] 此方法下，在获取到包含了离散型特征项和/或连续型特征项的流量样本数据后，通过先对连续型特征项进行分箱处理得到离散型特征项，然后根据离散型特征项的信息增益值和/或基尼系数获取目标离散型特征项，将与该目标离散型特征项对应的特征值作为

训练异常流量识别模型的数据，使得获取的目标离散型特征项更加适合用于训练异常流量识别模型，从而提高了获取的用来训练异常流量识别模型的流量数据的有效性，同时当获取的目标离散型特征项被用于训练异常流量识别模型时，能够避免过拟合，并可以提高训练出的异常流量识别模型的性能，当利用获取的数据训练出的异常流量识别模型用于流量监控时，可以提高监控异常流量的精度，进而提高异常流量监控效果。

[0022] 应当理解的是，以上的一般描述和后文的细节描述仅是示例性的，并不能限制本发明。

附图说明

[0023] 此处的附图被并入说明书中并构成本说明书的一部分，示出了符合本发明的实施例，并于说明书一起用于解释本发明的原理。

[0024] 图1是根据一示例性实施例示出的一种基于异常流量识别模型的流量监控方法的系统架构示意图；

[0025] 图2是根据一示例性实施例示出的一种基于异常流量识别模型的流量监控方法的流程图；

[0026] 图3是根据图2对应实施例示出的一实施例的步骤220的细节流程图；

[0027] 图4是根据图2对应实施例示出的一实施例的步骤250之前的步骤以及步骤250的细节流程图；

[0028] 图5是根据图4对应实施例示出的一实施例的步骤230的细节流程图；

[0029] 图6是根据一示例性实施例示出的一种基于异常流量识别模型的流量监控装置的框图；

[0030] 图7是根据一示例性实施例示出的一种用于实现上述基于异常流量识别模型的流量监控方法的电子设备示例框图；

[0031] 图8是根据一示例性实施例示出的一种用于实现上述基于异常流量识别模型的流量监控方法的计算机可读存储介质。

具体实施方式

[0032] 这里将详细地对示例性实施例执行说明，其示例表示在附图中。下面的描述涉及附图时，除非另有表示，不同附图中的相同数字表示相同或相似的要素。以下示例性实施例中所描述的实施方式并不代表与本发明相一致的所有实施方式。相反，它们仅是与如所附权利要求书中所详述的、本发明的一些方面相一致的装置和方法的例子。

[0033] 此外，附图仅为本公开的示意性图解，并非一定是按比例绘制。图中相同的附图标记表示相同或类似的部分，因而将省略对它们的重复描述。附图中所示的一些方框图是功能实体，不一定必须与物理或逻辑上独立的实体相对应。

[0034] 本公开首先提供了一种基于异常流量识别模型的流量监控方法。基于异常流量识别模型的流量监控方法是指本方法使用异常流量识别模型进行流量的监控。模型是指机器学习模型，可以包含各种类型，比如可以是决策树模型、逻辑回归模型、支持向量机等。训练机器学习模型需要先获取数据，数据是指按照一定形式组织的关于事物的信息，要建立异常流量识别模型，需要先获取用来训练异常流量识别模型的数据进行模型的训练，获取的

数据的类型一般为流量数据。训练异常流量识别模型是指将任意一种类型的模型训练成为能够处理异常流量识别任务的模型。训练异常流量识别模型是确定异常流量识别模型的参数的过程，而本公开提供的基于异常流量识别模型的流量监控方法可以先获取用于训练异常流量识别模型的数据，然后利用获取的数据训练异常流量识别模型，最终得到异常流量识别模型，从而利用异常流量识别模型实现对流量的有效监控。

[0035] 本公开的方法可以固定于各种终端，例如服务器、云计算的物理基础设施、智能手机、平板电脑、台式电脑、笔记本电脑、iPad、PDA(Personal Digital Assistant，简称掌上电脑)自助服务终端等任何具有运算处理功能和通信功能的设备或者设备集合。

[0036] 图1是根据一示例性实施例示出的一种基于异常流量识别模型的流量监控方法的系统架构示意图。如图1所示，包括服务器110、多个终端120以及数据库130，在图1示出的实施例中，本公开的实施终端为服务器110，多个终端120和数据库130都能够通过通信链路与服务器110进行通信，每一终端120在通过通信链路访问服务器110时会消耗流量，各终端120在访问服务器110时可能带来异常的大流量或非法流量服务器110在接收到终端120发来的访问请求后，可以将各终端访问时流量情况记录流量数据，然后将流量数据发送至与之相连的数据库130进行存储。出于特定的应用目的或者需求目的的需要，服务器110记录并存储至数据库130中的流量数据可能不是全都满足训练异常流量识别模型的需要，需要在这些流量数据中选择出适合训练异常流量识别模型的数据来进行异常流量识别模型的训练，而当本公开提供的基于异常流量识别模型的流量监控方法在本实施例中的实施终端——服务器110执行时，就能够先从数据库130中获取到适合异常流量识别模型的训练的流量数据，利用这些数据进行异常流量识别模型的训练可以使训练出的模型有更好的监控效果，从而当运行在服务器110上的训练好的异常流量识别模型对各终端的流量进行监控时，能够保证对异常流量的监控有较高的精度。

[0037] 值得一提的是，图1仅为本公开的一个实施例，虽然在图1示出的实施例中，流量来源于各终端，并由服务器将来自各终端的对应的流量数据记录并汇总至数据库中，并且训练好的异常流量识别模型是固定在服务器上，但在实际应用中，流量数据可以各种方式产生并存在，比如流量数据可以是其他终端转发来的数据包，一个数据包可以包括多个流量数据，获取到的流量数据也可以存储在本地或者其他终端，训练好的异常流量识别模型也可以固定在各种终端上。本公开对此不作任何限定，本公开的保护范围也不应因此而受到任何限制。

[0038] 图2是根据一示例性实施例示出的一种基于异常流量识别模型的流量监控方法的流程图。如图2所示，包括以下步骤：

[0039] 步骤210，获取多个流量样本数据。

[0040] 每一所述流量样本数据包括与预测项对应的预测值以及与预设特征项集合中的至少一个特征项对应的特征值，所述预设特征项集合包括多个特征项，与每一特征项对应的特征值为离散值或连续值，其中，对应的特征值为离散值的特征项为离散型特征项，对应的特征值为连续值的特征项为连续型特征项，与预测项对应的预测值指示所述流量样本数据是否为异常流量。

[0041] 流量样本数据可以是任意维度与产生访问流量的主体相关的数据，一个流量样本数据还包含与该流量样本数据是否为异常流量的标注结果的数据。预测项是要训练异常流

量识别模型所要预测的属性，即标注结果对应的属性，亦可以称之为标签或因变量；与预测项对应的预测值为该预测项的取值，即标注结果的值。特征项是表示流量样本数据一个维度的属性或者特征，特征值即为该属性或者特征的取值，为自变量。

[0042] 比如，流量样本数据可以是这样的，预测项为异常流量判断，与预测项对应的预测值即为一个流量样本数据是否为异常流量特征项的标签，特征项可以为IP地址、WI-FI名称、同一IP地址访问的次数、同一IP地址访问的账号的数目、预定时间段内请求访问的IP地址的数目、同一IP地址和WI-FI名称的组合的数目等，与每一特征项对应的特征值即为每一特征项的对应取值，其中，同一IP地址访问的次数、同一IP地址访问的账号的数目、预定时间段内请求访问的IP地址的数目、同一IP地址和WI-FI名称的组合的数目等特征项对应的取值是多样而连续的数值，同一IP地址访问的次数可以取值为1、2、3、4等数值中的任意一个，所以这些特征项对应的特征值为连续值，对应的特征项为连续型特征项；IP地址和WI-FI名称这些特征项的取值是离散的值，比如WI-FI名称取值可以为预设的字符串，这些特征项的取值没有明显的连续性，所以这些特征项对应的取值为离散值，这些特征项即为离散型特征项。

[0043] 在一个实施例中，每一所述流量样本数据包含的预测项对应的预测值为“0”或“1”，其中，“0”代表所述流量样本数据不为异常流量，而“1”代表所述流量样本数据为异常流量。

[0044] 在一个实施例中，每一所述流量样本数据包含的预测项对应的预测值为“是”或“否”，在训练模型时可以将“是”或“否”对应转换成“1”或“0”。

[0045] 每一所述流量样本数据包括的特征值所对应的特征项的数目与预设特征项集合包括的特征项的数目可以相同，也可以不同，各流量样本数据包括的特征值所对应的特征项的数目和种类可以相同，也可以不同。

[0046] 在一个实施例中，每一所述流量样本数据包括与预设特征项集合中的每一特征项对应的特征值。

[0047] 在一个实施例中，所述获取多个流量样本数据，包括：获取原始流量样本数据；对所述原始流量样本数据进行数据清洗，以过滤掉原始流量样本数据中的异常数据，所述异常数据包括缺失值和/或离群值；在经过数据清洗的原始流量样本数据中获取多个流量样本数据。

[0048] 缺失值是指流量样本数据的特征项对应的取值不存在，离群值是指流量样本数据的特征项对应的取值偏离正常水平，比如若一个流量样本数据的特征项为WI-FI名称，而该特征项的取值为空(NULL)，则该取值即为离群值。

[0049] 在一个实施例中，通过如下方式过滤掉原始流量样本数据中的离群值：

[0050] 针对每一连续型特征项，对各流量样本数据包含的与该连续型特征项对应的特征值从小到大进行排序；

[0051] 根据所述排序将包含的连续型特征项对应的特征值小于第2.5百分位或大于第97.5百分位的流量样本数据作为离群值过滤掉。

[0052] 步骤220，针对每一连续型特征项，对该连续型特征项对应的各流量样本数据对应的特征值进行分箱处理，以使该连续型特征项转换为离散型特征项，其中，经分箱处理后得到的每一流量样本数据所属的箱为将所述连续型特征项转换为的离散型特征项对应的特

征值。

[0053] 分箱是指一个连续型特征项对应的特征值进行分段处理的过程。

[0054] 在一个实施例中，利用分类与回归树(Classification and Regression Trees，CART)算法或者卡方分箱算法对该连续型特征项对应的各流量样本数据对应的特征值进行分箱处理。

[0055] 在一个实施例中，所述针对每一连续型特征项，对该连续型特征项对应的各流量样本数据对应的特征值进行分箱处理，以使该连续型特征项转换为离散型特征项，包括：

[0056] 针对每一连续型特征项，对各流量样本数据中与该连续型特征项对应的特征值从大到小进行排序；从排序在最前的与该连续型特征项对应的特征值开始，按照所述排序，选取预定数目个样本数据对应的特征值分为一箱，并将已被分箱的流量样本数据标记为已选取；从未被标记为已选取的排序在最前的流量样本数据中与该连续型特征项对应的特征值开始，按照所述排序，每次选取预定数目个与该连续型特征项对应的特征值分为一箱，并将已被分箱的与该连续型特征项对应的特征值标记为已选取，直至所有与该连续型特征项对应的特征值被标记为已选取，其中，当未被标记为已选取的与该连续型特征项对应的特征值的数目小于预定数目时，将所有未被标记为已选取的与该连续型特征项对应的特征值分为一箱。

[0057] 本实施例的好处在于，使被分至每一箱的与连续型特征项对应的特征值的数量大致相同，提高了分箱的均衡性。

[0058] 在一个实施例中，所述针对每一连续型特征项，对该连续型特征项对应的各流量样本数据对应的特征值进行分箱处理，以使该连续型特征项转换为离散型特征项，包括：

[0059] 针对每一连续型特征项，获取该连续型特征项的区间划分基准值和各流量样本数据中与该连续型特征项对应的特征值的最大值，所述区间划分基准值小于或等于各流量样本数据中与该连续型特征项对应的特征值的最小值；

[0060] 针对每一连续型特征项，将该连续型特征项的区间划分基准值与各流量样本数据中与该连续型特征项对应的特征值的最大值之间的区间平均划分为预定数目个特征值区间；

[0061] 对每一连续型特征项，针对每一流量样本数据，根据该流量样本数据中与该连续型特征项对应的特征值所属的特征值区间，对与该连续型特征项对应的流量样本数据中的特征值进行分箱，其中，与该连续型特征项对应的特征值所属的特征值区间为同一特征值区间的流量样本数据的特征值被分为一箱。

[0062] 本实施例的好处在于，使被分至同一箱的流量样本数据的特征值的差距不会太大，提高了分箱对各流量样本数据的特征值的差异的相关性。

[0063] 在一个实施例中，所述针对每一连续型特征项，对该连续型特征项对应的各流量样本数据对应的特征值进行分箱处理，以使该连续型特征项转换为离散型特征项，包括：

[0064] 针对每一连续型特征项，对各流量样本数据中与该连续型特征项对应的特征值从大到小进行排序；从排序在最前的与该连续型特征项对应的特征值开始，按照所述排序，选取预定数目个流量样本数据对应的特征值；当选取的流量样本数据对应的特征值中的最大值与最小值的差值大于预设的与该连续型特征项对应的差值阈值，将选取的流量样本数据对应的特征值中与所述最大值的差值小于或等于预设的与该连续型特征项对应的差值阈

值的流量样本数据对应的特征值分为一箱，并将已被分箱的流量样本数据标记为已选取；当选取的流量样本数据对应的特征值中的最大值与最小值的差值小于或等于预设的与该连续型特征项对应的差值阈值，将选取的所有流量样本数据对应的特征值分为一箱，并将已被分箱的流量样本数据标记为已选取；从未被标记为已选取的排序在最前的与该连续型特征项对应的特征值开始，按照所述排序，每次选取预定数目个与该连续型特征项对应的特征值，并根据选取的与该连续型特征项对应的特征值中最大值与最小值的差值对与该连续型特征项对应的特征值，其中，当选取的流量样本数据与该连续型特征项对应的特征值中的最大值与最小值的差值大于预设的与该连续型特征项对应的差值阈值时，将选取的流量样本数据对应的特征值中与所述最大值的差值小于或等于预设的与该连续型特征项对应的差值阈值的流量样本数据对应的特征值分为一箱，并将已被分箱的流量样本数据对应的特征值标记为已选取，当选取的流量样本数据对应的特征值中的最大值与最小值的差值小于或等于预设的与该连续型特征项对应的差值阈值时，将选取的所有流量样本数据对应的特征值分为一箱，并将已被分箱的流量样本数据对应的特征值标记为已选取，直至所有流量样本数据被标记为已选取，其中，当未被标记为已选取的流量样本数据的数目小于预定数目时，将所有未被标记为已选取的流量样本数据分为一箱。

[0065] 本实施例的好处在于，平衡和兼顾了分出的同一箱中与连续型特征项对应的特征值之间的差异以及分得的箱中连续型特征项对应的特征值的数量，使得分箱对特征值的划分更为合理。

[0066] 步骤250，基于各流量样本数据对应的离散型特征项以及与预测项对应的预测值，计算每一离散型特征项的信息增益值和/或基尼系数。

[0067] 信息增益和基尼系数都是衡量数据不纯度的重要指标。信息增益是选取按照某个自变量划分所需要的期望信息，该期望信息越小，划分的纯度越高。信息增益还可以定义为一个特征(变量)能够为分类带来多少信息，带来的信息越多，该变量越重要。基尼系数能够反映以离散型特征项对多个流量样本数据进行划分后划分的纯度。

[0068] 在一个实施例中，利用下列公式计算每一离散型特征项的信息增益值：

[0069] 利用如下公式计算获取的多个流量样本数据整体的信息熵：

[0070] $$\mathrm{H(D)} = -\sum_{k=1}^{m} p_k \log_2 p_k \quad ,$$

[0071] 其中，D为获取的多个流量样本数据，p_k为具有第k个预测项对应的预测值的流量样本数据的数目与获取的所有流量样本数据的数目的比值，k为获取的流量样本数据中所有预测项对应的预测值的种类数；

[0072] 基于获取的多个流量样本数据整体的信息熵利用如下公式计算离散型特征项b的信息增益：

[0073] $$Gain\left(D,\ b\right) = H(D) - \sum_{n=1}^{N} H\left(D_N\right),$$

[0074] 其中，D_N为所述多个流量样本数据中包含离散型特征项b的第N个取值的流量样本数据。

[0075] 在一个实施例中，利用下列公式计算离散型特征项b的基尼系数：

[0076] $$Gini(D) = 1 - \sum_{k=1}^{m} {p_k}^2 ,$$

[0077] $$Gini\left(D,\ b\right) = \sum_{n=1}^{N} Gini\left(D_N\right),$$

[0078] 其中，D为获取的多个流量样本数据，p_k为具有第k个预测项对应的预测值的流量样本数据的数目与获取的所有流量样本数据的数目的比值，k为获取的流量样本数据中所有预测项对应的预测值的种类数，D_N为所述多个流量样本数据中包含离散型特征项b的第N个取值的流量样本数据。

[0079] 通过计算每个离散型特征项的信息增益值和基尼系数，并以计算结果作为离散型特征项的筛选标准，实现了离散型特征项数量的精简，使得保留下来的特征项与预测项有更好的相关性，从而使最终获得的数据更加适合用于训练异常流量识别模型。

[0080] 步骤260，根据各离散型特征项的信息增益值和/或基尼系数，在各离散型特征项中筛选出目标离散型特征项，并将各流量样本数据与所述目标离散型特征项对应的特征值，作为训练异常流量识别模型的数据。

[0081] 在一个实施例中，根据各离散型特征项的信息增益值和/或基尼系数，在各离散型特征项中筛选出目标离散型特征项，包括：

[0082] 获取信息增益值大于预设信息增益值阈值或者基尼系数小于预设基尼系数阈值的离散型特征项，作为目标离散型特征项。

[0083] 在一个实施例中，所述根据各离散型特征项的信息增益值和/或基尼系数，在各离散型特征项中筛选出目标离散型特征项，包括：

[0084] 根据各离散型特征项的信息增益值和/或基尼系数，在各离散型特征项中筛选出初始特征项；

[0085] 重复执行初始特征项筛选步骤，得到初始特征项集合，直至重复次数达到第一预定数目，所述初始特征项筛选步骤包括：

[0086] 利用由所述初始特征项构成的特征项集合建立随机森林，所述特征集合包括多个初始特征项，所述随机森林包括多个决策树，每一决策树包括多个初始特征项；

[0087] 针对每一初始特征项，确定该初始特征项在每一决策树中的重要程度，并基于该初始特征项在每一决策树中的重要程度确定该初始特征项在所述随机森林中的重要程度；

[0088] 对各初始特征项按照所述重要程度从高到低的顺序进行排列，获取排序在前第二预定数目的初始特征项，作为初始特征项集合；

[0089] 将所有所述初始特征项集合的交集作为目标离散型特征项。

[0090] 在本实施例中，在根据各离散型特征项的信息增益值和/或基尼系数，获得初始特征项的基础上，再次利用随机森林获得目标离散型特征项，使得最终获得的目标离散型特征项更加适用于训练异常流量识别模型，从而可以使得训练出的异常流量识别模型能够更准确地进行异常流量的识别。

[0091] 在一个实施例中，所述针对每一初始特征项，确定该初始特征项在每一决策树中

的重要程度,包括:

[0092] 针对每一决策树,获取每一初始特征项对应的各节点在该决策树中分支前后的基尼指数变化量;

[0093] 针对每一决策树,针对每一初始特征项,获取针对该初始特征项的各节点获取的该决策树中分支前后的基尼指数变化量之和,作为该初始特征项在该决策树中的重要程度;

[0094] 所述基于该初始特征项在每一决策树中的重要程度确定该初始特征项在所述随机森林中的重要程度,包括:

[0095] 将该初始特征项在各决策树中的重要程度的平均值作为该初始特征项在所述随机森林中的重要程度。

[0096] 在一个实施例中,在根据各离散型特征项的信息增益值和/或基尼系数,在各离散型特征项中筛选出目标离散型特征项,并将各流量样本数据与所述目标离散型特征项对应的特征值,作为训练异常流量识别模型的数据之后,所述方法还包括:

[0097] 将获得的所述训练异常流量识别模型的数据输入至逻辑回归模型,以训练形成异常流量识别模型。

[0098] 步骤270,利用基于所述数据训练形成的异常流量识别模型对目标流量进行监控,以得到异常流量。

[0099] 当使用获取的数据训练好异常流量识别模型后,便可以利用异常流量识别模型进行流量的监控。

[0100] 综上所述,根据图2实施例示出的训练异常流量识别模型的数据的获取方法,能够使得获取的目标离散型特征项更加适合用于训练机器学习模型,从而提高了获取的用来训练异常流量识别模型的数据的有效性,当获取的数据用于训练异常流量识别模型时,可以提高训练出的异常流量识别模型的性能,从而提高训练出的异常流量识别模型的流量监控精度和流量监控效果。

[0101] 图3是根据图2对应实施例示出的一实施例的步骤220的细节流程图。如图3所示,步骤220包括以下步骤:

[0102] 步骤221,针对每一连续型特征项,对各流量样本数据中与该连续型特征项对应的特征值进行聚类,以将与该连续型特征项对应的特征值划分为多个类。

[0103] 在一个实施例中,利用DBSCAN(Density-Based Spatial Clustering of Applications with Noise,具有噪声的基于密度的聚类方法)算法对各流量样本数据中与连续型特征项对应的特征值进行聚类,以将与该连续型特征项对应的特征值划分为多个类。

[0104] 步骤222,针对每一连续型特征项,根据与该连续型特征项对应的特征值被划分为的多个类,对该连续型特征项对应的各流量样本数据对应的特征值进行分箱处理,以使该连续型特征项转换为离散型特征项。

[0105] 在一个实施例中,将被划分为一类的流量样本数据中与连续型特征项对应的特征值分为一个箱。

[0106] 在一个实施例中,所述针对每一连续型特征项,根据与该连续型特征项对应的特征值被划分为的多个类,对该连续型特征项对应的各流量样本数据对应的特征值进行分箱

处理，包括：

[0107] 针对每一连续型特征项，对与该连续型特征项对应的特征值被划分为的每一类按照类中该连续型特征项对应的特征值的平均值从小到大进行排序；

[0108] 重复执行分箱步骤，直至每一类都经过了分箱，所述分箱步骤包括：

[0109] 从排在最前的一个类开始，针对每一未经分箱的每一类，判断该类中包含的特征值的数目是否大于预设数目阈值；

[0110] 如果是，将该类分为一箱，并将该类标记为已分箱；

[0111] 如果否，从该类开始，按照所述排序，每次获取一个未被标记为已分箱的类，并判断已获取的所有未被标记为已分箱的类中包含的特征值的数目之和是否大于预设数目阈值，如果是，已获取的所有未被标记为已分箱的类中包含的特征值分为一箱。

[0112] 本实施例的好处在于，通过聚类的方式对各流量样本数据对应的特征值进行分箱处理，使得分箱结果能够对流量样本数据对应的特征值进行合理地划分，从而可以提高获取的用来训练异常流量识别模型的数据的有效性。

[0113] 图4是根据图2对应实施例示出的一实施例的步骤250之前的步骤以及步骤250的细节流程图。如图4所示，包括以下步骤：

[0114] 步骤230，对各离散型特征项进行聚类，以将各离散型特征项划分为多个簇。

[0115] 聚类是指对离散型特征项进行归类的过程。

[0116] 在一个实施例中，如图5所示，步骤230具体可以包括以下步骤：

[0117] 步骤231，确定各离散型特征项中每一对离散型特征项之间的皮尔逊相关系数。

[0118] 皮尔逊相关系数是用于衡量两个变量之间的相关程度的系数。

[0119] 在一个实施例中，利用如下公式计算每一对离散型特征项之间的皮尔逊相关系数：

[0120] $$P(x,y)=\frac{n\sum x_i y_i-\sum x_i\sum y_i}{\sqrt{n\sum x_i^2-(\sum x_i)^2}\sqrt{n\sum y_i^2-(\sum y_i)^2}},$$

[0121] 其中，x为第一离散型特征项，y为第二离散型特征项，x_i为第一离散型特征项的第i个特征值，y_i为第二离散型特征项的第i个特征值。

[0122] 步骤232，利用所述皮尔逊相关系数，对各离散型特征项进行聚类，以将各离散型特征项划分为多个簇。

[0123] 在一个实施例中，所述利用所述皮尔逊相关系数，对各离散型特征项进行聚类，以将各离散型特征项划分为多个簇，包括：

[0124] 重复执行簇内基准项获取步骤，直至不存在与该基准项的皮尔逊相关系数小于预设皮尔逊相关系数阈值的项，将获取的被标记过的基准项归为一簇，所述簇内基准项获取步骤，包括：

[0125] 获取一个未被标记过的离散型特征项，并将该离散型特征项标记为初始簇内基准项；

[0126] 获取与该基准项的皮尔逊相关系数小于预设皮尔逊相关系数阈值的项；

[0127] 取消该基准项的标记，并将获取的项标记为基准项，获取与该基准项的皮尔逊相关系数小于预设皮尔逊相关系数阈值的项并取消最近一次对基准项的标记并重新将获取

的项标记为基准项；

[0128] 再次获取一个未被标记过的离散型特征项，并将该离散型特征项标记为初始簇内基准项，重复执行簇内基准项获取步骤，直至不存在未被标记过的离散型特征项。

[0129] 步骤240，根据各离散型特征项被划分为的簇，在各离散型特征项中确定出目标离散型特征项；

[0130] 在一个实施例中，针对每一簇，任取一个离散型特征项，作为目标离散型特征项。

[0131] 步骤250'，基于各流量样本数据对应的目标离散型特征项以及与预测项对应的预测值，计算每一离散型特征项的信息增益值和/或基尼系数。

[0132] 综上所述，图4对应实施例的好处在于，通过先根据聚类获得将离散型特征项划归为多个簇，然后根据聚类得到的簇进行目标离散型特征项的选择，实现了对特征项的降维处理，在不显著降低信息量的情况下，提高了获取的离散型特征项的有效性，能够提高根据获取的离散型特征项而获取的训练异常流量识别模型的数据的有效性，进而能提高利用这些数据训练出的异常流量监控模型的性能，提高流量监控的精度和效果。

[0133] 本公开还提供了一种基于异常流量识别模型的流量监控装置，以下是本公开的装置实施例。

[0134] 图6是根据一示例性实施例示出的一种基于异常流量识别模型的流量监控装置的框图。如图6所示，该装置600包括：

[0135] 获取模块610，被配置为获取多个流量样本数据，每一所述流量样本数据包括与预测项对应的预测值以及与预设特征项集合中的至少一个特征项对应的特征值，所述预设特征项集合包括多个特征项，与每一特征项对应的特征值为离散值或连续值，其中，对应的特征值为离散值的特征项为离散型特征项，对应的特征值为连续值的特征项为连续型特征项，与预测项对应的预测值指示所述流量样本数据是否为异常流量；

[0136] 分箱模块620，被配置为针对每一连续型特征项，对该连续型特征项对应的各流量样本数据对应的特征值进行分箱处理，以使该连续型特征项转换为离散型特征项，其中，经分箱处理后得到的每一流量样本数据所属的箱为将所述连续型特征项转换为的离散型特征项对应的特征值；

[0137] 计算模块630，被配置为基于各流量样本数据对应的离散型特征项以及与预测项对应的预测值，计算每一离散型特征项的信息增益值和/或基尼系数；

[0138] 数据获取模块640，被配置为根据各离散型特征项的信息增益值和/或基尼系数，在各离散型特征项中筛选出目标离散型特征项，并将各流量样本数据与所述目标离散型特征项对应的特征值，作为训练异常流量识别模型的数据；

[0139] 监控模块650，利用基于所述数据训练形成的异常流量识别模型对目标流量进行监控，以得到异常流量。

[0140] 根据本公开的第三方面，还提供了一种能够实现上述基于异常流量识别模型的流量监控方法的电子设备。

[0141] 所属技术领域的技术人员能够理解，本发明的各个方面可以实现为系统、方法或程序产品。因此，本发明的各个方面可以具体实现为以下形式，即：完全的硬件实施方式、完全的软件实施方式(包括固件、微代码等)，或硬件和软件方面结合的实施方式，这里可以统称为“电路”、“模块”或“系统”。

[0142] 下面参照图7来描述根据本发明的这种实施方式的电子设备700。图7显示的电子设备700仅仅是一个示例，不应对本发明实施例的功能和使用范围带来任何限制。

[0143] 如图7所示，电子设备700以通用计算设备的形式表现。电子设备700的组件可以包括但不限于：上述至少一个处理单元710、上述至少一个存储单元720、连接不同系统组件(包括存储单元720和处理单元710)的总线730。

[0144] 其中，所述存储单元存储有程序代码，所述程序代码可以被所述处理单元710执行，使得所述处理单元710执行本说明书上述“实施例方法”部分中描述的根据本发明各种示例性实施方式的步骤。

[0145] 存储单元720可以包括易失性存储单元形式的可读介质，例如随机存取存储单元(RAM)721和/或高速缓存存储单元722，还可以进一步包括只读存储单元(ROM)723。

[0146] 存储单元720还可以包括具有一组(至少一个)程序模块725的程序/实用工具724，这样的程序模块725包括但不限于：操作系统、一个或者多个应用程序、其它程序模块以及程序数据，这些示例中的每一个或某种组合中可能包括网络环境的实现。

[0147] 总线730可以为表示几类总线结构中的一种或多种，包括存储单元总线或者存储单元控制器、外围总线、图形加速端口、处理单元或者使用多种总线结构中的任意总线结构的局域总线。

[0148] 电子设备700也可以与一个或多个外部设备900(例如键盘、指向设备、蓝牙设备等)通信，还可与一个或者多个使得用户能与该电子设备700交互的设备通信，和/或与使得该电子设备700能与一个或多个其它计算设备进行通信的任何设备(例如路由器、调制解调器等等)通信。这种通信可以通过输入/输出(I/O)接口750进行。并且，电子设备700还可以通过网络适配器760与一个或者多个网络(例如局域网(LAN)，广域网(WAN)和/或公共网络，例如因特网)通信。如图所示，网络适配器760通过总线730与电子设备700的其它模块通信。应当明白，尽管图中未示出，可以结合电子设备700使用其它硬件和/或软件模块，包括但不限于：微代码、设备驱动器、冗余处理单元、外部磁盘驱动阵列、RAID系统、磁带驱动器以及数据备份存储系统等。

[0149] 通过以上的实施方式的描述，本领域的技术人员易于理解，这里描述的示例实施方式可以通过软件实现，也可以通过软件结合必要的硬件的方式来实现。因此，根据本公开实施方式的技术方案可以以软件产品的形式体现出来，该软件产品可以存储在一个非易失性存储介质(可以是CD-ROM，U盘，移动硬盘等)中或网络上，包括若干指令以使得一台计算设备(可以是个人计算机、服务器、终端装置、或者网络设备等)执行根据本公开实施方式的方法。

[0150] 根据本公开的第四方面，还提供了一种计算机可读存储介质，其上存储有能够实现本说明书上述方法的程序产品。在一些可能的实施方式中，本发明的各个方面还可以实现为一种程序产品的形式，其包括程序代码，当所述程序产品在终端设备上运行时，所述程序代码用于使所述终端设备执行本说明书上述“示例性方法”部分中描述的根据本发明各种示例性实施方式的步骤。

[0151] 参考图8所示，描述了根据本发明的实施方式的用于实现上述方法的程序产品800，其可以采用便携式紧凑盘只读存储器(CD-ROM)并包括程序代码，并可以在终端设备，例如个人电脑上运行。然而，本发明的程序产品不限于此，在本文件中，可读存储介质可以

是任何包含或存储程序的有形介质，该程序可以被指令执行系统、装置或者器件使用或者与其结合使用。

[0152] 所述程序产品可以采用一个或多个可读介质的任意组合。可读介质可以是可读信号介质或者可读存储介质。可读存储介质例如可以为但不限于电、磁、光、电磁、红外线、或半导体的系统、装置或器件，或者任意以上的组合。可读存储介质的更具体的例子(非穷举的列表)包括：具有一个或多个导线的电连接、便携式盘、硬盘、随机存取存储器(RAM)、只读存储器(ROM)、可擦式可编程只读存储器(EPROM或闪存)、光纤、便携式紧凑盘只读存储器(CD-ROM)、光存储器件、磁存储器件、或者上述的任意合适的组合。

[0153] 计算机可读信号介质可以包括在基带中或者作为载波一部分传播的数据信号，其中承载了可读程序代码。这种传播的数据信号可以采用多种形式，包括但不限于电磁信号、光信号或上述的任意合适的组合。可读信号介质还可以是可读存储介质以外的任何可读介质，该可读介质可以发送、传播或者传输用于由指令执行系统、装置或者器件使用或者与其结合使用的程序。

[0154] 可读介质上包含的程序代码可以用任何适当的介质传输，包括但不限于无线、有线、光缆、RF等等，或者上述的任意合适的组合。

[0155] 可以以一种或多种程序设计语言的任意组合来编写用于执行本发明操作的程序代码，所述程序设计语言包括面向对象的程序设计语言—诸如Java、C++等，还包括常规的过程式程序设计语言—诸如“C”语言或类似的程序设计语言。程序代码可以完全地在用户计算设备上执行、部分地在用户设备上执行、作为一个独立的软件包执行、部分在用户计算设备上部分在远程计算设备上执行、或者完全在远程计算设备或服务器上执行。在涉及远程计算设备的情形中，远程计算设备可以通过任意种类的网络，包括局域网(LAN)或广域网(WAN)，连接到用户计算设备，或者，可以连接到外部计算设备(例如利用因特网服务提供商来通过因特网连接)。

[0156] 此外，上述附图仅是根据本发明示例性实施例的方法所包括的处理的示意性说明，而不是限制目的。易于理解，上述附图所示的处理并不表明或限制这些处理的时间顺序。另外，也易于理解，这些处理可以是例如在多个模块中同步或异步执行的。

[0157] 应当理解的是，本发明并不局限于上面已经描述并在附图中示出的精确结构，并且可以在不脱离其范围执行各种修改和改变。本发明的范围仅由所附的权利要求来限制。

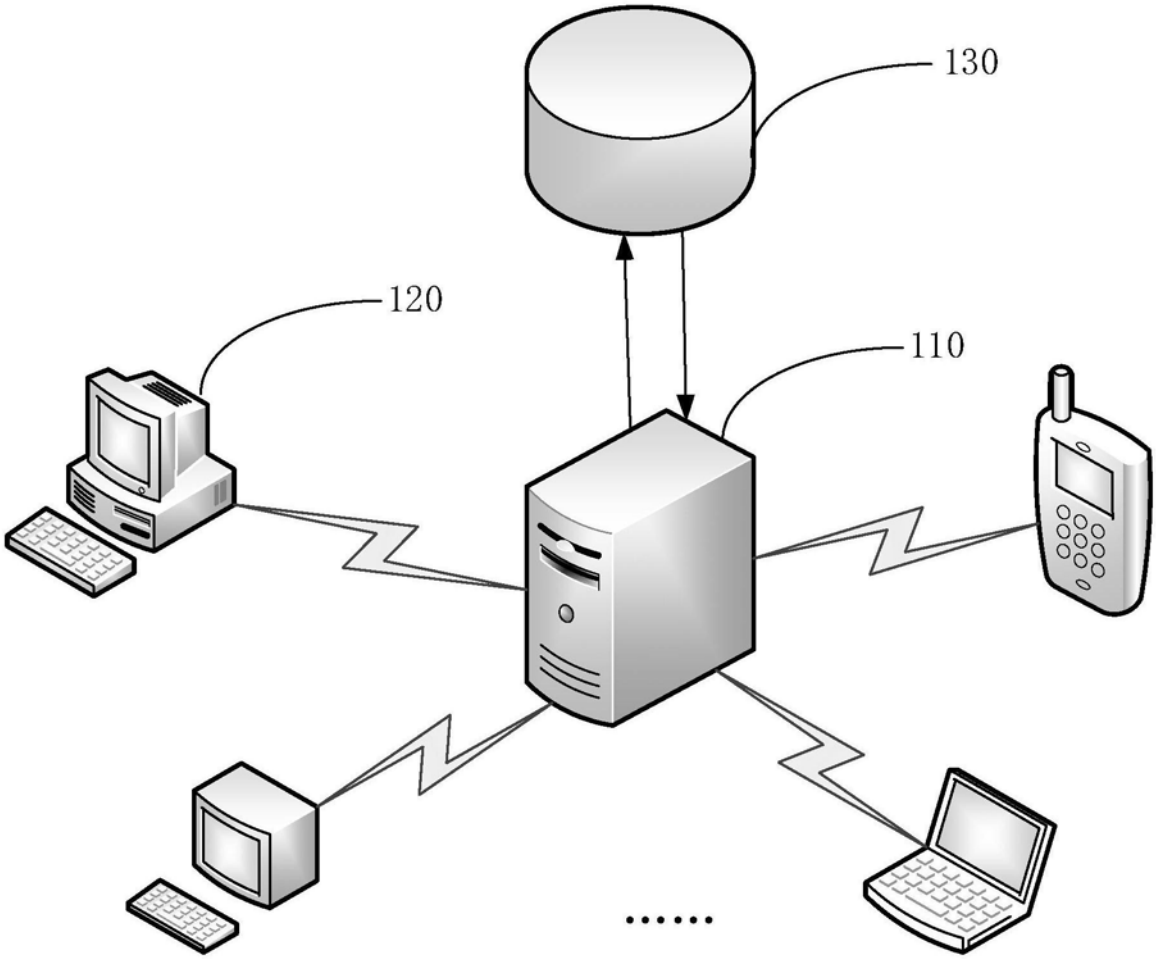


图1

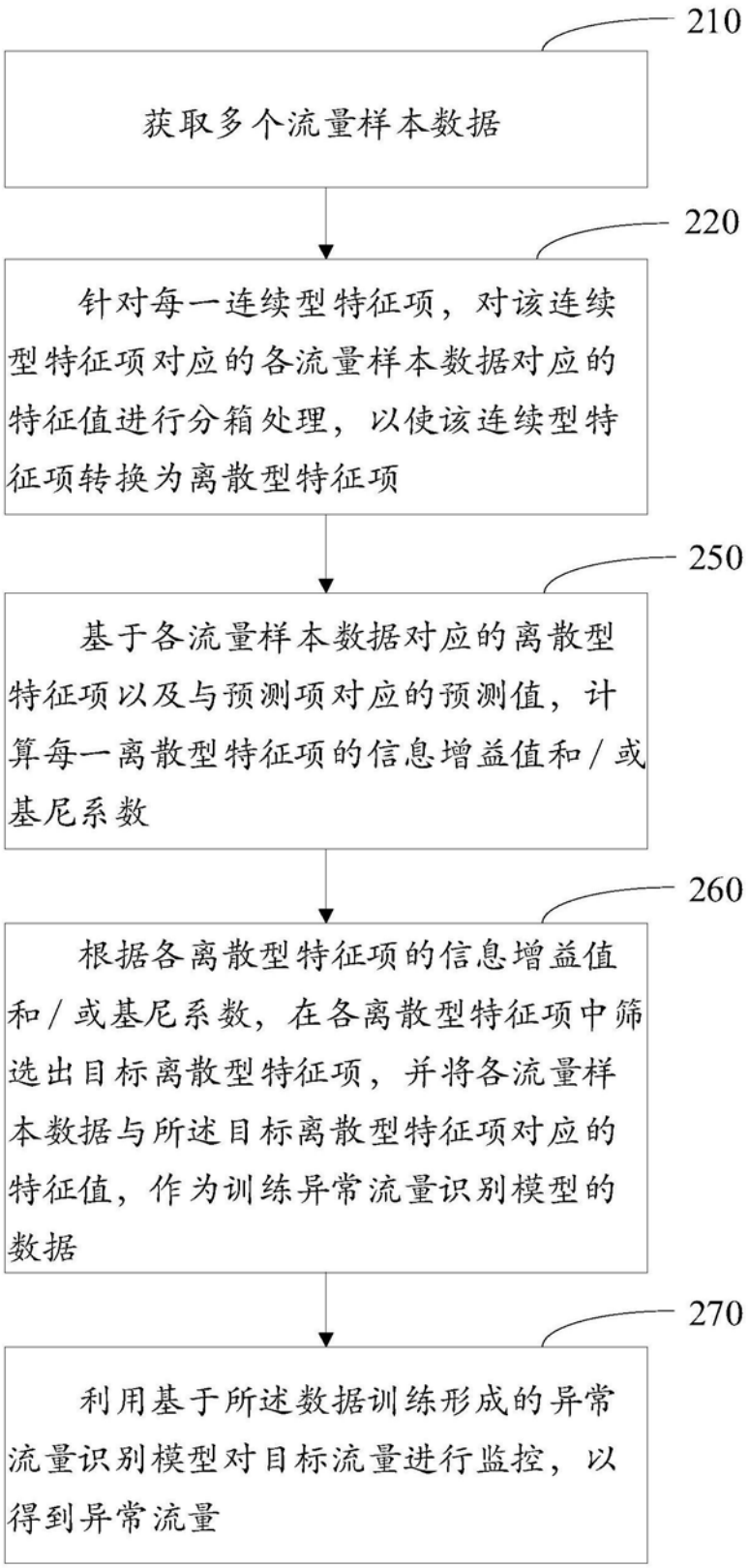

210
获取多个流量样本数据
220
针对每一连续型特征项，对该连续型特征项对应的各流量样本数据对应的特征值进行分箱处理，以使该连续型特征项转换为离散型特征项
250
基于各流量样本数据对应的离散型特征项以及与预测项对应的预测值，计算每一离散型特征项的信息增益值和/或基尼系数
260
根据各离散型特征项的信息增益值和/或基尼系数，在各离散型特征项中筛选出目标离散型特征项，并将各流量样本数据与所述目标离散型特征项对应的特征值，作为训练异常流量识别模型的数据
270
利用基于所述数据训练形成的异常流量识别模型对目标流量进行监控，以得到异常流量

图2


221
针对每一连续型特征项，对各流量样本数据中与该连续型特征项对应的特征值进行聚类，以将与该连续型特征项对应的特征值划分为多个类
222
针对每一连续型特征项，根据与该连续型特征项对应的特征值被划分为的多个类，对该连续型特征项对应的各流量样本数据对应的特征值进行分箱处理，以使该连续型特征项转换为离散型特征项

图3

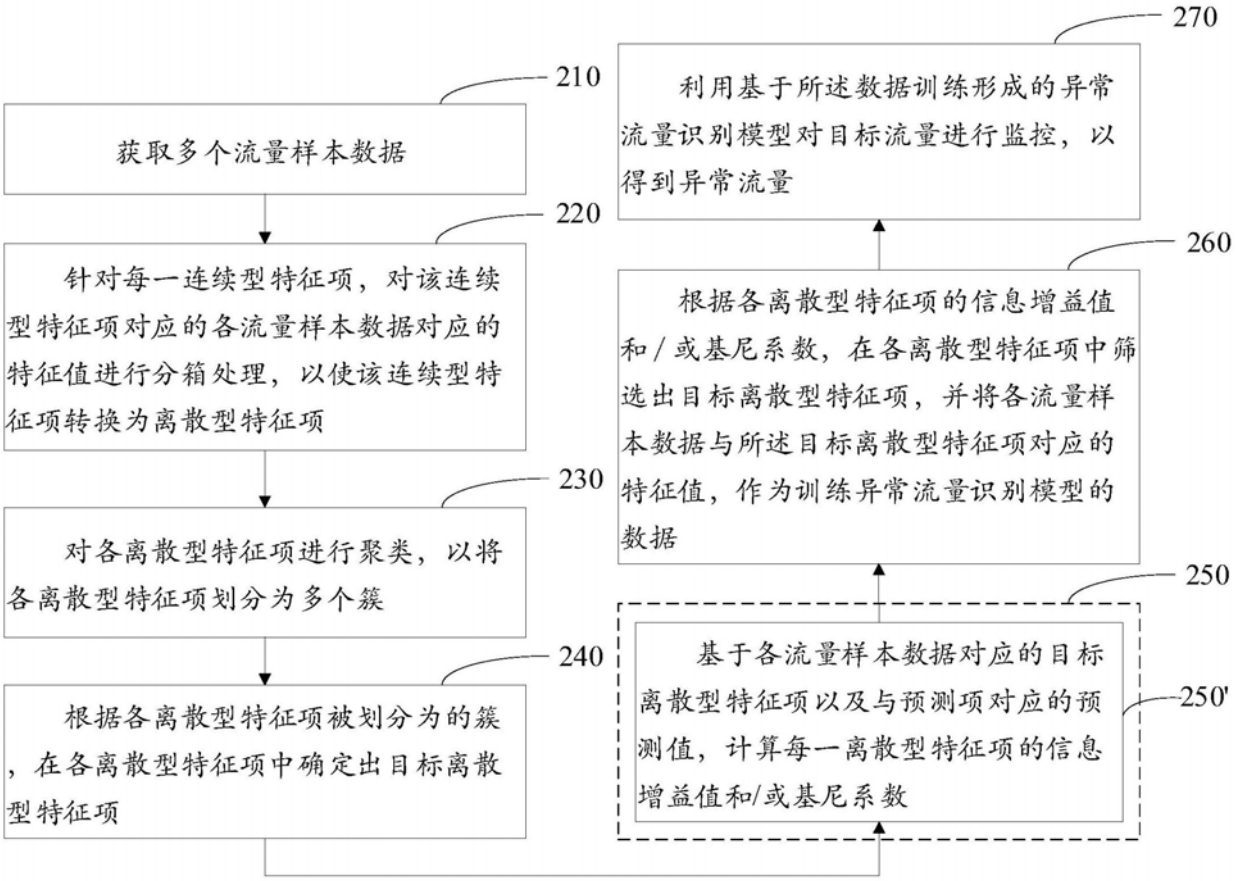

210
获取多个流量样本数据
220
针对每一连续型特征项，对该连续型特征项对应的各流量样本数据对应的特征值进行分箱处理，以使该连续型特征项转换为离散型特征项
230
对各离散型特征项进行聚类，以将各离散型特征项划分为多个簇
240
根据各离散型特征项被划分为的簇，在各离散型特征项中确定出目标离散型特征项
250
基于各流量样本数据对应的目标离散型特征项以及与预测项对应的预测值，计算每一离散型特征项的信息增益值和/或基尼系数
250'
260
根据各离散型特征项的信息增益值和/或基尼系数，在各离散型特征项中筛选出目标离散型特征项，并将各流量样本数据与所述目标离散型特征项对应的特征值，作为训练异常流量识别模型的数据
270
利用基于所述数据训练形成的异常流量识别模型对目标流量进行监控，以得到异常流量

图4

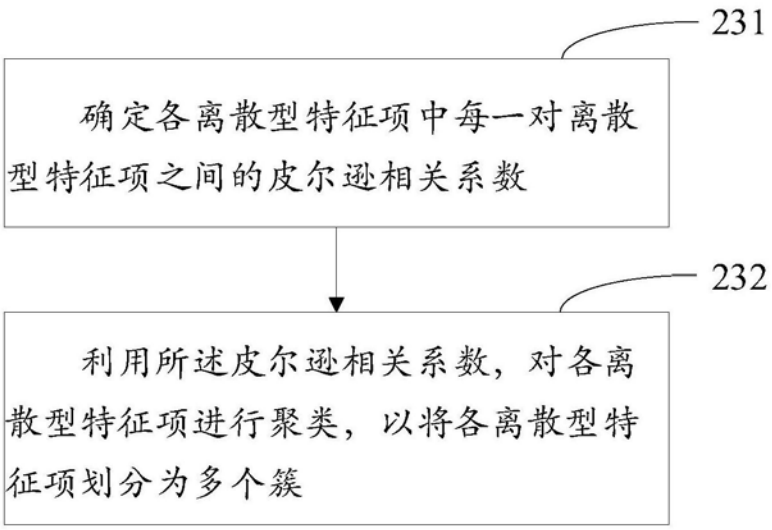

231
确定各离散型特征项中每一对离散型特征项之间的皮尔逊相关系数
232
利用所述皮尔逊相关系数，对各离散型特征项进行聚类，以将各离散型特征项划分为多个簇

图5

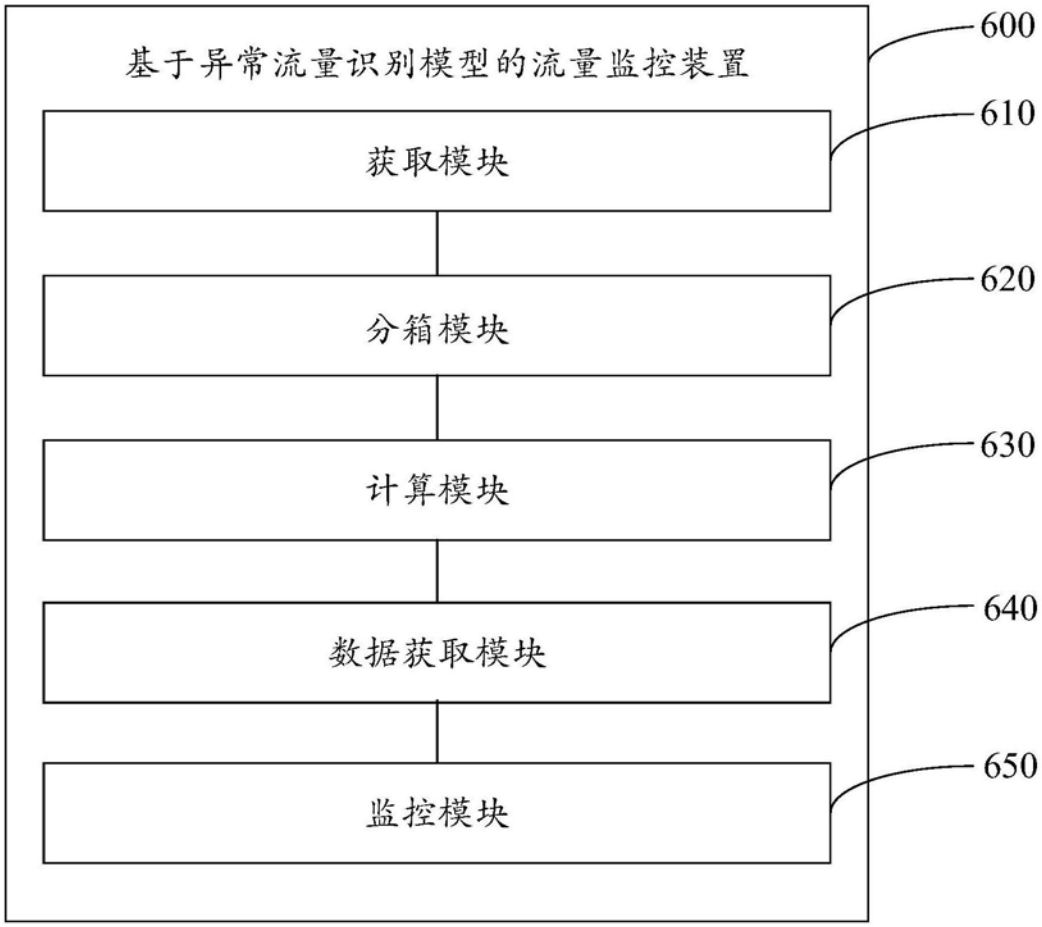

基于异常流量识别模型的流量监控装置
600
获取模块
610
分箱模块
620
计算模块
630
数据获取模块
640
监控模块
650

图6

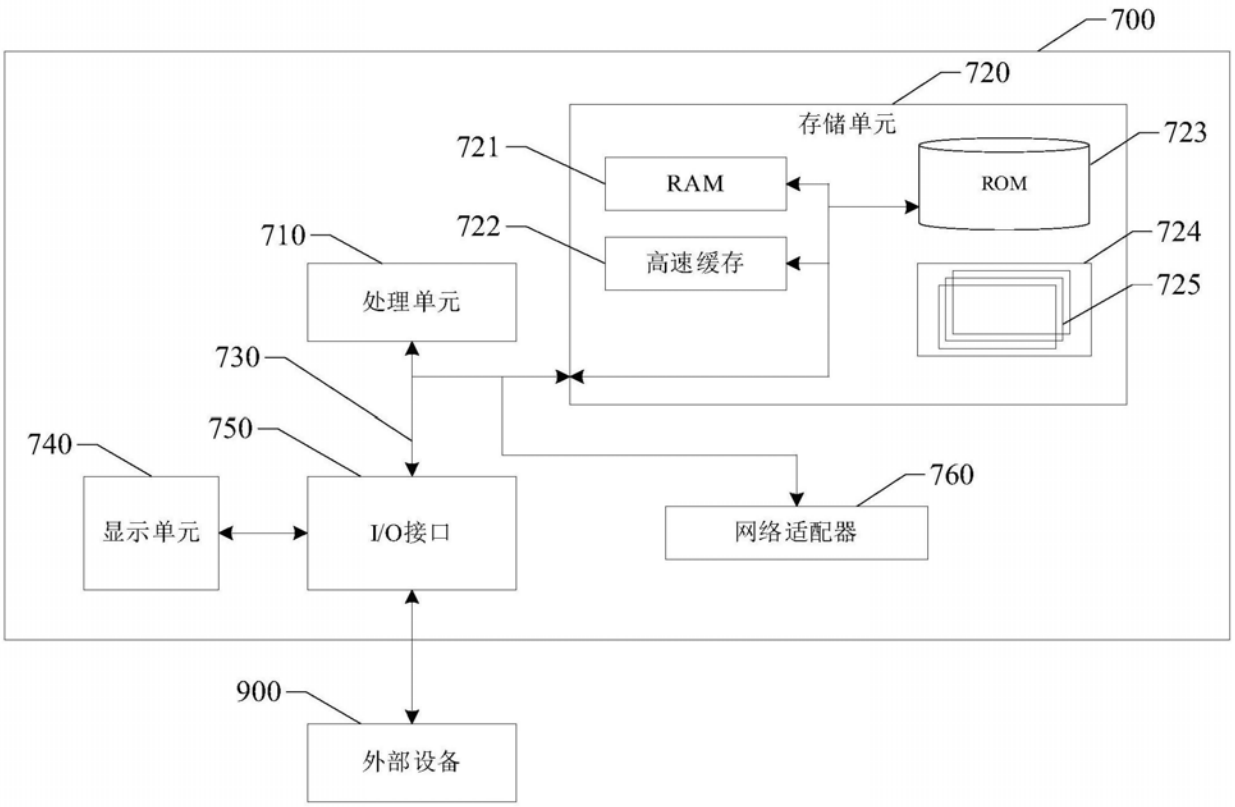

700
720
存储单元
721
RAM
723
ROM
722
高速缓存
724
725
710
处理单元
730
750
740
760
显示单元
I/O接口
网络适配器
900
外部设备

图7

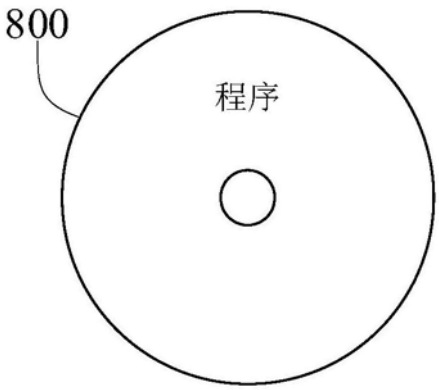

800
程序

图8