

(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

(43) 国际公布日
2020 年 4 月 2 日 (02.04.2020)



(10) 国际公布号
WO 2020/062689 A1

(51) 国际专利分类号:
G06F 16/2455 (2019.01)

(21) 国际申请号: PCT/CN2018/125246

(22) 国际申请日: 2018 年 12 月 29 日 (29.12.2018)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201811128269.2 2018 年 9 月 27 日 (27.09.2018) 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 518000 (CN)。

(72) 发明人: 孙家棣(SUN, Jiadi); 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 518000 (CN)。

马宁(MA, Ning); 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 518000 (CN)。

(74) 代理人: 深圳市隆天联鼎知识产权代理有限公司(SHENZHEN LUNGTIN LIANDING INTELLECTUAL PROPERTY AGENT LTD.); 中国广东省深圳市福田区南园路上田大厦 4A 刘抗美, Guangdong 518000 (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, 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,

(54) Title: CLUSTERING PROCESSING METHOD AND APPARATUS FOR TRAFFIC DATA, AND ELECTRONIC DEVICE

(54) 发明名称: 流量数据的聚类处理方法、装置及电子设备

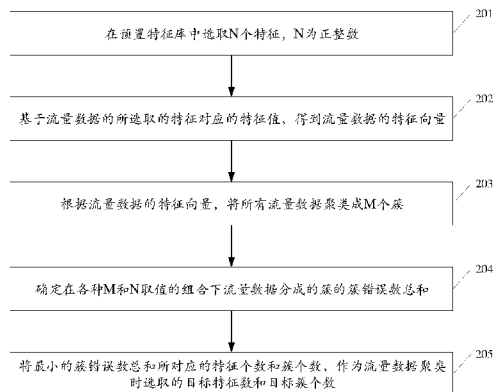


图 2

201 Select N features from a preset feature library, N being a positive integer
202 Obtain a feature vector of traffic data on the basis of a feature value corresponding to a selected feature of the traffic data
203 Cluster all traffic data into M clusters according to the feature vector of the traffic data
204 Determine the sum of cluster errors of the divided clusters of the traffic data under various combinations of M and N values
205 Select the number of the features and the number of the clusters corresponding to the minimum sum of the cluster errors as the number of target features and the number of target clusters selected when the traffic data is clustered

(57) Abstract: A clustering processing method and apparatus for traffic data, and an electronic device. The method comprises: selecting N features from a preset feature library (201); obtaining a feature vector of the traffic data on the basis of a feature value corresponding to a selected feature of the traffic data (202); clustering all traffic data into M clusters according to the feature vector of the traffic data (203); determining the sum of cluster errors of the divided clusters of the traffic data under various combinations of M and N values (204), the sum of the cluster errors being the result of addition of errors for each divided cluster, and the number of errors of each cluster referring to the number of the lesser in the number of white data and the number of black data in the cluster; and selecting the number of the features and the number of the clusters corresponding to the minimum sum of the cluster errors as the number of target features and the number of target clusters selected when the traffic data is clustered (205), so as to cluster a large amount of traffic data by using a clustering algorithm to meet the demands for overall analysis of the traffic data of the population.

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: 一种流量数据的聚类处理方法、装置及电子设备, 该方法包括: 在预置特征库中选取N个特征 (201); 基于流量数据的所选取的特征对应的特征值, 得到流量数据的特征向量 (202); 根据流量数据的特征向量, 将所有流量数据聚类成M个簇 (203); 确定在各种M和N取值的组合下流量数据分成的簇的簇错误数总和 (204), 簇错误数总和是分成的每个簇的错误数相加的结果, 每个簇的错误数是指该簇中白数据的数量和黑数据的数量中较少的一个数量; 将最小的簇错误数总和所对应的特征个数和簇个数, 作为流量数据聚类时选取的目标特征数和目标簇个数 (205), 从而利用聚类算法对大量流量数据进行聚类处理, 得以满足对于群体的流量数据整体分析的需求。

流量数据的聚类处理方法、装置及电子设备

技术领域

[0001] 相关申请的交叉引用：本申请基于并要求2018年9月27日递交、发明名称为“一种流量数据的聚类处理方法、装置及电子设备”的中国专利申请CN 201811128269.2的优先权，在此通过引用将其全部内容合并于此。

[0002] 本申请涉及大数据技术领域，特别涉及流量数据的聚类处理方法、装置及电子设备。

背景技术

[0003] 目前，随着互联网用户的日益增多，互联网领域正面临着大流量数据的挑战。大流量数据中难免会出现异常流量，这些异常流量会给互联网带来巨大的冲击与损失，例如，黑色产业形成的木马播种、流量交易和虚拟财产套现等诸多黑色产业链都会产生大量的异常流量。在现有技术的实现中，流量异常的识别一般是通过采集用户行为埋点和sdk数据来确定路径重复度、设备前后端登录埋点占比、ip访问账号数、ip访问次数、周期内手机号段用户登录均值和方差等特征，根据每一条流量数据的这些特征，确定该流量数据异常的概率。

[0004] 本申请的发明人意识到，现有技术的缺陷在于，黑色产业往往表现为群体的流量数据出现异常，而现有技术对于流量异常的识别是针对每一条流量数据孤立确定的，无法满足对于群体的流量数据整体分析的需求。

发明概述

技术问题

[0005] 为了解决相关技术中存在的无法满足对于群体的流量数据整体分析的需求的问题，本申请提供了一种流量数据的聚类处理方法、装置及电子设备。

问题的解决方案

技术解决方案

[0006] 根据本申请实施例的一方面，提供一种流量数据的聚类处理方法，所述流量数据包括白数据和黑数据，所述白数据是从确定为白用户的用户的数据流量中抽

取的流量数据，所述黑数据是从确定为黑用户的用户的数据流量中抽取的流量数据，所述白用户是确定为不会发出异常流量数据的用户，所述黑用户是确定为会发出异常流量数据的用户，所述方法包括：在预置特征库中选取 N 个特征， N 为正整数；基于流量数据的所选取的特征对应的特征值，得到所述流量数据的特征向量；所述特征向量包括所述流量数据的所述 N 个特征各自对应的特征值，其中，一个所述特征对应一个所述特征值；根据所述流量数据的特征向量，将所有所述流量数据聚类成 M 个簇， M 为大于等于2的正整数；确定在各种 M 和 N 取值的组合下所述流量数据分成的簇的簇错误数总和，所述簇错误数总和是分成的每个簇的错误数相加的结果，每个簇的错误数是指该簇中白数据的数量和黑数据的数量中较少的一个数量；将最小的簇错误数总和所对应的特征个数和簇个数，作为流量数据聚类时选取的目标特征数和目标簇个数。

[0007] 根据本申请实施例的另一方面，提供一种流量数据的聚类处理装置，所述流量数据包括白数据和黑数据，所述白数据是从确定为白用户的用户的数据流量中抽取的流量数据，所述黑数据是从确定为黑用户的用户的数据流量中抽取的流量数据，所述白用户是确定为不会发出异常流量数据的用户，所述黑用户是确定为会发出异常流量数据的用户，所述装置包括：选取单元，用于在预置特征库中选取 N 个特征， N 为正整数；获取单元，用于基于流量数据的所选取的特征对应的特征值，得到所述流量数据的特征向量；所述特征向量包括所述流量数据的所述 N 个特征各自对应的特征值；其中，一个所述特征对应一个所述特征值；聚类单元，用于根据所述流量数据的特征向量，将所有所述流量数据聚类成 M 个簇， M 为大于等于2的正整数；确定单元，用于确定在各种 M 和 N 取值的组合下所述流量数据分成的簇的簇错误数总和，所述簇错误数总和是分成的每个簇的错误数相加的结果，每个簇的错误数是指该簇中白数据的数量和黑数据的数量中较少的一个数量；设置单元，用于将最小的簇错误数总和所对应的特征个数和簇个数，作为流量数据聚类时选取的目标特征数和目标簇个数。

[0008] 根据本申请实施例的又一方面，提供一种计算机可读存储介质，其特征在于，其存储计算机程序，所述计算机程序在被计算机执行时使得计算机执行如前所述的方法。

[0009] 根据本申请实施例的又一方面，提供一种电子设备，包括：存储器，所述存储器上存储有计算机可读指令；处理器，其被配置为执行所述存储器上存储的计算机可读指令，其中所述计算机可读指令被所述处理器执行时，使得所述处理器执行如前所述的方法。

发明的有益效果

有益效果

[0010] 本申请的实施例提供的技术方案可以包括以下有益效果：

[0011] 本申请所提供的图像控制方法包括如下步骤，在预置特征库中选取 N 个特征， N 为正整数；基于流量数据的所选取的特征对应的特征值，得到所述流量数据的特征向量；所述特征向量包括所述流量数据的所述 N 个特征各自对应的特征值；其中，一个所述特征对应一个所述特征值；根据所述流量数据的特征向量，将所有所述流量数据聚类成 M 个簇， M 为大于等于2的正整数；确定在各种 M 和 N 取值的组合下所述流量数据分成的簇的簇错误数总和，所述簇错误数总和是分成的每个簇的错误数相加的结果，每个簇的错误数是指该簇中白数据的数量和黑数据的数量中较少的一个数量；将最小的簇错误数总和所对应的特征个数和簇个数，作为流量数据聚类时选取的目标特征数和目标簇个数。

[0012] 此方法下，对于流量异常的识别不是孤立确定的，而是将流量数据划分为若干个簇，结合若干个簇能够反映流量数据在一个群体、或在一个区域、或在一类人中呈现出的特点，有利于分析黑产业链的行为。综上，对于群体的流量数据整体分析的需求得以满足。

[0013] 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，并不能限制本申请。

对附图的简要说明

附图说明

[0014] 此处的附图被并入说明书中并构成本说明书的一部分，示出了符合本申请的实施例，并与说明书一起用于解释本申请的原理。

[0015] 图1是根据一示例性实施例示出的一种流量数据的聚类处理装置的示意图；

[0016] 图2是根据一示例性实施例示出的一种流量数据的聚类处理方法的流程图；

- [0017] 图3是根据图2对应实施例示出的步骤230的细节的流程图；
- [0018] 图4是根据另一示例性实施例示出的一种流量数据的聚类处理方法的流程图；
- [0019] 图5是根据一示例性实施例示出的一种流量数据的聚类处理装置的框图。

发明实施例

本发明的实施方式

- [0020] 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。
- [0021] 本发明的实施环境可以是便携移动设备，如智能手机、平板电脑、台式电脑。本发明所公开的流量数据的聚类处理方法可以适用于运行于便携移动设备上的任意应用程序。
- [0022] 图1是根据一示例性实施例示出的一种流量数据的聚类处理装置的示意图。装置100可以是上述便携移动设备。如图1所示，装置100可以包括以下一个或多个组件：处理组件102，存储器104，电源组件106，多媒体组件108，音频组件110，传感器组件114以及通信组件116。处理组件102通常控制装置100的整体操作，诸如与显示，电话呼叫，数据通信，相机操作以及记录操作相关联的操作等。处理组件102可以包括一个或多个处理器118来执行指令，以完成下述的方法的全部或部分步骤。此外，处理组件102可以包括一个或多个模块，用于便于处理组件102和其他组件之间的交互。例如，处理组件102可以包括多媒体模块，用于以方便多媒体组件108和处理组件102之间的交互。存储器104被配置为存储各种类型的数据以支持在装置100的操作。这些数据的示例包括用于在装置100上操作的任何应用程序或方法的指令。存储器104可以由任何类型的易失性或非易失性存储设备或者它们的组合实现，如静态随机存取存储器（Static Random Access Memory，简称SRAM），电可擦除可编程只读存储器（Electrically Erasable Programmable Read-Only Memory，简称EEPROM），可擦除可编程只读存储器（Erasable Programmable Read Only Memory，简称EPROM），可编程只

读存储器 (Programmable Red-Only Memory, 简称PROM), 只读存储器 (Read-Only Memory, 简称ROM), 磁存储器, 快闪存储器, 磁盘或光盘。存储器104中还存储有一个或多个模块, 用于该一个或多个模块被配置成由该一个或多个处理器118执行, 以完成如下所示方法中的全部或者部分步骤。电源组件106为装置100的各种组件提供电力。电源组件106可以包括电源管理系统, 一个或多个电源, 及其他与为装置100生成、管理和分配电力相关联的组件。多媒体组件108包括在所述装置100和用户之间的提供一个输出接口的屏幕。在一些实施例中, 屏幕可以包括液晶显示器 (Liquid Crystal Display, 简称LCD) 和触摸面板。如果屏幕包括触摸面板, 屏幕可以被实现为触摸屏, 以接收来自用户的输入信号。触摸面板包括一个或多个触摸传感器以感测触摸、滑动和触摸面板上的手势。所述触摸传感器可以不仅感测触摸或滑动动作的边界, 而且还检测与所述触摸或滑动操作相关的持续时间和压力。屏幕还可以包括有机电致发光显示器 (Organic Light Emitting Display, 简称OLED)。音频组件110被配置为输出和/或输入音频信号。例如, 音频组件110包括一个麦克风 (Microphone, 简称MIC), 当装置100处于操作模式, 如呼叫模式、记录模式和语音识别模式时, 麦克风被配置为接收外部音频信号。所接收的音频信号可以被进一步存储在存储器104或经由通信组件116发送。在一些实施例中, 音频组件110还包括一个扬声器, 用于输出音频信号。传感器组件114包括一个或多个传感器, 用于为装置100提供各个方面的状态评估。例如, 传感器组件114可以检测到装置100的打开/关闭状态, 组件的相对定位, 传感器组件114还可以检测装置100或装置100一个组件的位置改变以及装置100的温度变化。在一些实施例中, 该传感器组件114还可以包括磁传感器, 压力传感器或温度传感器。通信组件116被配置为便于装置100和其他设备之间有线或无线方式的通信。装置100可以接入基于通信标准的无线网络, 如WiFi (Wireless-Fidelity, 无线保真)。在一个示例性实施例中, 通信组件116经由广播信道接收来自外部广播管理系统的广播信号或广播相关信息。在一个示例性实施例中, 所述通信组件116还包括近场通信 (Near Field Communication, 简称NFC) 模块, 用于以促进短程通信。例如, 在NFC模块可基于射频识别 (Radio Frequency

Identification, 简称RFID) 技术, 红外数据协会 (Infrared Data Association, 简称IrDA) 技术, 超宽带 (Ultra Wideband, 简称UWB) 技术, 蓝牙技术和其他技术来实现。

[0023] 在示例性实施例中, 装置100可以被一个或多个应用专用集成电路 (Application Specific Integrated Circuit, 简称ASIC)、数字信号处理器、数字信号处理设备、可编程逻辑器件、现场可编程门阵列、控制器、微控制器、微处理器或其他电子元件实现, 用于执行下述方法。

[0024] 图2是根据一示例性实施例示出的一种流量数据的聚类处理方法的流程图。如图2所示, 此方法包括以下步骤。

[0025] 步骤210, 在预置特征库中选取N个特征, N为正整数。

[0026] 本申请实施例中, 每个用户的流量数据事先被规定若干个特征, 例如, 这些特征可以包括路径重复度、设备前后端登录埋点占比、ip访问账号数、ip访问次数、周期内手机号段用户登录均值和方差等, 预置数据库中包括但不限于上述若干个特征, 从预置特征库包括的若干个特征中选取N个特征, 其中, N可以为小于等于预置特征库中所有特征的数量正整数。其中, 特征可以由用户指定选取, 也可以随机选取, 也可以采用其他选取方式, 本申请实施例中不做限定。

[0027] 作为一种可选实施方式, 在预置特征库中选取N个特征可包括: 在预置特征库中选取卡方值从高到低前N个特征。本申请实施例中, 假设预置特征库中包含14个特征, 这时选取特征共有 种情况, 若聚成簇的类数在2-20之间取值, 有19种取值, 因此选取的特征和类数的组合有 种。如果每种组合都去遍历, 计算量非常大。此时可以按照每一特征对应的卡方值大小来选取目标特征。例如, 如果N为1, 则选取预置特征库中卡方值最高的特征作为目标特征, 如果N为2, 则选取预置特征库中卡方值最高和次高的特征作为目标特征, 由于卡方值越大, 该卡方值对应的特征对于良好的聚类越重要, 因此可以选取聚类效果最好的特征, 提升聚类效果。

[0028] 步骤220, 基于流量数据的所选取的特征对应的特征值, 得到流量数据的特征向量。

[0029] 本申请实施例中, 特征向量包括流量数据的N个特征各自对应的特征值; 其中

，一个特征对应一个特征值。例如， $a_1, a_2, \dots, a_n$ 分别是第1, 2, ..., N个特征的特征值，得到的流量数据的特征向量即为 $(a_1, a_2, \dots, a_n)$ 构成的集合。本申请实施例中，流量数据包括白数据和黑数据，白数据是从确定为白用户的用户的数据流量中抽取的流量数据，黑数据是从确定为黑用户的用户的数据流量中抽取的流量数据，白用户是确定为不会发出异常流量数据的用户，黑用户是确定为会发出异常流量数据的用户。可选的，流量数据中白数据和黑数据的比为预设比例且预设比例为1:1，预设比例也可以为其他比例，本申请实施例中不做限定。通过实施这种可选的实施方式，减少了因白数据与黑数据选取比例失衡导致局部最优的情况出现的概率。

[0030] 步骤230，根据流量数据的特征向量，将所有流量数据聚类成M个簇。

[0031] 本申请实施例中，M为大于等于2的正整数。

[0032] 步骤240，确定在各种M和N取值的组合下流量数据分成的簇的簇错误数总和。

[0033] 本申请实施例中，簇错误数总和是分成的每个簇的错误数相加的结果，每个簇的错误数是指该簇中白数据的数量和黑数据的数量中较少的一个数量。具体的，簇中如果只包括白数据或者只包括黑数据，认为此时聚类效果最佳。因此，将M个簇中白数据的数量大于黑数据的数量的簇确定为白簇，将M个簇中黑数据的数量大于白数据的数量的簇确定为黑簇，白簇的簇错误数为白簇中黑数据的数量，黑簇的簇错误数为黑簇中白流量数据的数量，M个簇的簇错误数总和即为所有白簇的簇错误数总和与所有黑簇的错误数总和累加得到的簇错误数总和。并且，各种M和N取值的组合为遍历N值取值范围的所有值与M值取值范围的所有值的组合。

[0034] 步骤250，将最小的簇错误数总和所对应的特征个数和簇个数，作为流量数据聚类时选取的目标特征数和目标簇个数。

[0035] 作为一种可选的实施方式，在将最小的簇错误数总和所对应的特征个数和簇个数，作为流量数据聚类时选取的目标特征数和目标簇个数之后，还可以执行以下步骤：确定在按照选取的目标特征数和目标簇个数聚类之后，聚成的每个簇的簇风险评分，该簇风险评分的公式如下：

$$\text{score} = \frac{N_1}{N_0 + N_1}$$

，其中，

N_0 、 N_1

分别表示该簇中白样本个数和黑样本个数， $score$ 为簇风险评分。其中，白样本个数即为该簇中白数据个数，黑样本个数即为该簇中黑数据个数。簇风险评分的取值范围为 $[0, 1]$ ，簇风险评分越大（簇风险评分越靠近1），表示该簇黑样本比例越大，该簇存在流量异常的风险也就越大。且将该簇的簇编号与该簇对应的簇风险评分对应存储，管理人员可以查看每一簇的簇风险评分，从而使得簇风险情况的呈现更加直观。可选的，当簇风险评分大于0.5时，确定该簇为流量异常的簇，其中，也可以在簇风险评分大于0.6或者0.7时确定该簇为流量异常的簇，具体的簇风险评分大于的某一数值本申请实施例中不做限定。

[0036] 作为另一种可选的实施方式，在执行完步骤250之后，还可以执行以下步骤：判断聚成的簇的数量是否大于预设数量；当判断出大于预设数量时，确定出每一聚成的簇的中心点；根据每一聚成的簇的中心点，将所有聚成的簇划分至预设簇中，其中，预设簇包括黑簇、白簇以及混合簇，黑簇为黑数据占主导的簇，白簇为白数据占主导的簇，混合簇为黑数据与白数据均不做主导的簇。通过实施这种可选实施方式，当聚成的簇的数量过多时，可进一步将聚成的簇划分得到三个簇，有利于后续根据划分得到的簇分析黑产业链的行为。

[0037] 上述方法下，对于流量异常的识别不是孤立确定的，而是依据目标特征数和目标簇个数将流量数据划分为若干个簇，结合若干个簇能够反映流量数据在一个群体、或在一个区域、或在一类人中呈现出的特点，有利于分析黑产业链的行为。综上，对于群体的流量数据整体分析的需求得以满足。

[0038] 图3是图2对应实施例示出的步骤230的细节的流程图。如图3所示，步骤230包括：

[0039] 步骤231，对流量数据的特征向量所包括的各特征值进行归一化处理，得到归一化特征向量。本申请实施例中，归一化处理是用流量数据的特征向量所包括的一个特征的特征值除以所有流量数据的特征向量所包括的该特征的最大特征值的结果。

[0040] 步骤232, 将归一化特征向量聚类成M个簇。

[0041] 图4是根据另一示例性实施例示出的一种流量数据的聚类处理方法的流程图。

如图4所示, 此方法包括以下步骤。

[0042] 步骤401, 在预置特征库中选取N个特征, N为正整数。作为一种可选实施方式, 在预置特征库中选取N个特征可包括: 在预置特征库中选取卡方值从高到低前N个特征。

[0043] 步骤402, 基于流量数据的所选取的特征对应的特征值, 得到流量数据的特征向量。本申请实施例中, 特征向量包括流量数据的N个特征各自对应的特征值; 其中一个特征对应一个特征值。本申请实施例中, 流量数据包括白数据和黑数据, 白数据是从确定为白用户的用户的数据流量中抽取的, 黑数据是从确定为黑用户的用户的数据流量中抽取的, 白用户是确定为不会发出异常流量数据的用户, 黑用户是确定为会发出异常流量数据的用户。可选的, 流量数据中白数据和黑数据的比为预设比例且预设比例为1:1。

[0044] 步骤403, 根据流量数据的特征向量, 将一部分流量数据聚类成M个簇, M为大于等于2的正整数。

[0045] 步骤404, 确定在各种M和N取值的组合下流量数据分成的簇的簇错误数总和, 簇错误数总和是分成的每个簇的错误数相加的结果。本申请实施例中, 每个簇的错误数是指该簇中白数据的数量和黑数据的数量中较少的一个数量。

[0046] 步骤405, 将从小到大前预定名次的簇错误数总和所对应的M和N的组合, 作为候选特征数M和N的组合。

[0047] 步骤406, 根据流量数据的特征向量, 将所有流量数据聚类成M个簇。本申请实施例中, M为大于等于2的正整数。

[0048] 步骤407, 确定在各种候选M和N取值的组合下流量数据分成的簇的簇错误数总和。

[0049] 本申请实施例中, 簇错误数总和是分成的每个簇的错误数相加的结果, 每个簇的错误数是指该簇中白数据的数量和黑数据的数量中较少的一个数量。

[0050] 步骤408, 将最小的簇错误数总和所对应的特征个数和簇个数, 作为流量数据聚类时选取的目标特征数和目标簇个数。

[0051] 作为一种可选的实施方式，在将最小的簇错误数总和所对应的特征个数和簇个数，作为流量数据聚类时选取的目标特征数和目标簇个数之后，还可以执行以下步骤：确定在按照选取的目标特征数和目标簇个数聚类之后，聚成的每个簇的簇风险评分，该簇风险评分的公式如下：

$$\text{score} = \frac{N1}{N0 + N1}$$

，其中，

N0、N1

分别表示该簇中白样本个数和黑样本个数，score为簇风险评分。上述方法下，能够在对预设数量的流量数据聚类成初始簇时，从中选取较优的候选M和N取值的组合，并在该选取较优的候选M和N取值的组合下针对所有流量数据进行聚类，从中选取簇错误数总和取值最小的簇错误数总和。这一过程的第一次聚类过程对部分流量数据聚类获取较优的候选M和N取值的组合，第二次聚类过程按照第一次聚类选取的较优的较优的候选M和N取值的组合，对全部流量数据聚类，可以同时兼顾处理效率与聚类的准确性。以下是本申请的装置实施例。

[0052] 图5是根据一示例性实施例示出的一种流量数据的聚类处理装置500的框图。如图5所示，该装置包括：

[0053] 选取单元501，用于在预置特征库中选取N个特征，N为正整数。作为一种可选的实施方式，选取单元501在预置特征库中选取N个特征，N为正整数可以包括：选取单元501在预置特征库中选取卡方值从高到低前N个特征。获取单元502，用于基于流量数据的所选取的特征对应的特征值，得到流量数据的特征向量；特征向量包括流量数据的N个特征各自对应的特征值；其中，一个特征对应一个特征值。本实施例中，流量数据中白数据和黑数据的比为预设比例，预设比例可以为1:1。第一聚类单元503，用于根据流量数据的特征向量，将所有流量数据聚类成M个簇，M为大于等于2的正整数。第一确定单元504，用于确定在各种M和N取值的组合下流量数据分成的簇的簇错误数总和，簇错误数总和是分成的每个簇的错误数相加的结果，每个簇的错误数是指该簇中白数据的数量和黑数据的

数量中较少的一个数量。第一设置单元505，用于将最小的簇错误数总和所对应的特征个数和簇个数，作为流量数据聚类时选取的目标特征数和目标簇个数。

[0054] 作为一种可选的实施方式，第一聚类单元503可以包括：归一化单元5031，其被配置为：对流量数据的特征向量所包括的各特征值进行归一化处理，得到归一化特征向量，其中归一化处理是用流量数据的特征向量所包括的一个特征的特征值除以所有流量数据的特征向量所包括的该特征的最大特征值的结果；聚类子单元5032，其被配置为：将归一化特征向量聚类成M个簇。

[0055] 作为一种可选的实施方式，如图5所示，该装置还可以包括：第二聚类单元506，其被配置为：在所述第一聚类单元503将所有所述流量数据聚类成M个簇之前，根据所述流量数据的特征向量，将一部分所述流量数据聚类成M个簇，M为大于等于2的正整数；第二确定单元507，其被配置为：确定在各种M和N取值的组合下所述一部分所述流量数据分成的簇的簇错误数总和，所述簇错误数总和是分成的每个簇的错误数相加的结果，每个簇的错误数是指该簇中白数据的数量和黑数据的数量中较少的一个数量；第二设置单元508，其被配置为：将所述第二确定单元507确定的从小到大前预定名次的簇错误数总和所对应的M和N的组合，作为候选特征数M和N的组合，其中，所述第一确定单元504被配置为通过执行如下步骤来实现所述确定在各种M和N取值的组合下所述流量数据分成的簇的簇错误数总和：确定在各种候选M和N取值的组合下所述流量数据分成的簇的簇错误数总和。作为一种可选的实施方式，该装置还可以包括风险确定单元509，其被配置为：在第一设置单元505将最小的簇错误数总和所对应的特征个数和簇个数，作为流量数据聚类时选取的目标特征数和目标簇个数之后，确定在按照选取的目标特征数和目标簇个数聚类之后，聚成的每个簇的簇风险评分，该簇风险评分的公式如下：

$$\text{score} = \frac{N1}{N0 + N1}$$

，其中，

N0、N1

分别表示该簇中白样本个数和黑样本个数，score为簇风险评分。

[0056] 作为另一种可选的实施方式，在第一设置单元505将最小的簇错误数总和所对应的特征个数和簇个数，作为流量数据聚类时选取的目标特征数和目标簇个数之后，第一设置单元505还可以被配置为：判断聚成的簇的数量是否大于预设数量；当判断出大于预设数量时，确定出每一聚成的簇的中心点；根据每一聚成的簇的中心点，将所有聚成的簇划分至预设簇中，其中，预设簇包括黑簇、白簇以及混合簇，黑簇为黑数据占主导的簇，白簇为白数据占主导的簇，混合簇为黑数据与白数据均不做主导的簇。通过实施这种可选的实施方式，当聚成的簇的数量过多时，可以进一步将聚成的簇划分得到三个簇，有利于后续根据划分得到的簇分析黑产业链的行为。

[0057] 上述装置中各个单元/模块的功能和作用的实现过程以及相关细节具体详见上述方法实施例中对应步骤的实现过程，在此不再赘述。以上各实施例中的装置实施例可以通过硬件、软件、固件或其组合的方式来实现，并且其可以被实现为一个单独的装置，也可以被实现为各组成单元/模块分散在一个或多个计算设备中并分别执行相应功能的逻辑集成系统。以上各实施例中组成装置的各单元/模块是根据逻辑功能而划分的，它们可以根据逻辑功能被重新划分，例如可以通过更多或更少的单元/模块来实现该装置。这些组成单元/模块分别可以通过硬件、软件、固件或其组合的方式来实现，它们可以是分别的独立部件，也可以是多个组件组合起来执行相应的逻辑功能的集成单元/模块。所述硬件、软件、固件或其组合的方式可以包括：分离的硬件组件，通过编程方式实现的功能模块、通过可编程逻辑器件实现的功能模块，等等，或者以上方式的组合。

[0058] 本申请还提供一种电子设备，该电子设备包括：存储器，该存储器上存储有计算机可读指令；处理器，其被配置为执行所述存储器上存储的计算机可读指令，该计算机可读指令被处理器执行时，使得计算机执行如前所述的流量数据的聚类处理方法。上面的实施例中所述的处理器可以指单个的处理单元，如中央处理单元CPU，也可以是包括多个分散的处理单元的分布式处理器系统。上面的实施例中所述的存储器可以包括一个或多个存储器，其可以是计算设备的内部存储器，例如暂态或非暂态的各种存储器，也可以是通过存储器接口连接到计

算设备的外部存储装置。

[0059] 该电子设备可以是图1所示的流量数据聚类处理装置100，其结构组成可以参考图1及上面对图1的描述。

[0060] 在一示例性实施例中，本申请还提供一种计算机可读存储介质，其上存储有计算机程序，该计算机程序被处理器执行时，使得处理器执行如前所述的流量数据的聚类处理方法。

[0061] 该存储介质可以是任何可以保持和存储可由指令执行设备使用的指令的有形设备。例如，其可以是——但不限于——电存储设备、磁存储设备、光存储设备、电磁存储设备、半导体存储设备或者上述的任意合适的组合。存储介质的更具体的例子（非穷举的列表）包括：便携式计算机盘、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）、静态随机存取存储器（SRAM）、便携式压缩盘只读存储器（CD-ROM）、数字多功能盘（DVD）、记忆棒、软盘、机械编码设备、例如其上存储有指令的打孔卡或凹槽内凸起结构、以及上述的任意合适的组合。

[0062] 这里所描述的计算机程序/计算机指令可以从计算机可读存储介质下载到各个计算/处理设备，或者通过网络、例如因特网、局域网、广域网和/或无线网下载到外部计算机或外部存储设备。网络可以包括铜传输电缆、光纤传输、无线传输、路由器、防火墙、交换机、网关计算机和/或边缘服务器。每个计算/处理设备中的网络适配卡或者网络接口从网络接收计算机可读程序指令，并转发该计算机可读程序指令，以供存储在各个计算/处理设备中的计算机可读存储介质中。

[0063] 本公开中所述的计算机程序指令可以是汇编指令、指令集架构（ISA）指令、机器指令、机器相关指令、微代码、固件指令、状态设置数据、或者以一种或多种编程语言的任意组合编写的源代码或目标代码，所述编程语言包括面向对象的编程语言—诸如Smalltalk、C++等，以及常规的过程式编程语言—诸如“C”语言或类似的编程语言。计算机可读程序指令可以完全地在用户计算机上执行、部分地在用户计算机上执行、作为一个独立的软件包执行、部分在用户计算机上部分在远程计算机上执行、或者完全在远程计算机或服务器上执行。在涉

及远程计算机的情形中，远程计算机可以通过任意种类的网络—包括局域网(LAN)或广域网(WAN)—连接到用户计算机，或者，可以连接到外部计算机（例如利用因特网服务提供商来通过因特网连接）。在一些实施例中，通过利用计算机可读程序指令的状态信息来个性化定制电子电路，例如可编程逻辑电路、现场可编程门阵列（FPGA）或可编程逻辑阵列（PLA），该电子电路可以执行计算机可读程序指令，从而实现本发明的各个方面。

[0064] 这里参照根据本发明实施例的方法、装置（系统）和计算机程序产品的流程图和/或框图描述了本发明的各个方面。应当理解，流程图和/或框图的每个方框以及流程图和/或框图中各方框的组合，都可以由计算机可读程序指令实现。

[0065] 这些计算机可读程序指令可以提供给通用计算机、专用计算机或其它可编程数据处理装置的处理器，从而生产出一种机器，使得这些指令在通过计算机或其它可编程数据处理装置的处理器执行时，产生了实现流程图和/或框图中的一个或多个方框中规定的功能/动作的装置。也可以把这些计算机可读程序指令存储在计算机可读存储介质中，这些指令使得计算机、可编程数据处理装置和/或其他设备以特定方式工作，从而，存储有指令的计算机可读介质则包括一个制造品，其包括实现流程图和/或框图中的一个或多个方框中规定的功能/动作的各个方面的指令。

[0066] 也可以把计算机可读程序指令加载到计算机、其它可编程数据处理装置、或其它设备上，使得在计算机、其它可编程数据处理装置或其它设备上执行一系列操作步骤，以产生计算机实现的过程，从而使得在计算机、其它可编程数据处理装置、或其它设备上执行的指令实现流程图和/或框图中的一个或多个方框中规定的功能/动作。

[0067] 应当理解的是，本申请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。本申请的范围仅由所附的权利要求来限制。

权利要求书

- [权利要求 1] 一种流量数据的聚类处理方法，其特征在于，所述流量数据包括白数据和黑数据，所述白数据是从确定为白用户的用户的数据流量中抽取的流量数据，所述黑数据是从确定为黑用户的用户的数据流量中抽取的流量数据，所述白用户是确定为不会发出异常流量数据的用户，所述黑用户是确定为会发出异常流量数据的用户，所述方法包括：
- 在预置特征库中选取 N 个特征， N 为正整数；
- 基于流量数据的所选取的特征对应的特征值，得到所述流量数据的特征向量；所述特征向量包括所述流量数据的所述 N 个特征各自对应的特征值，其中，一个所述特征对应一个所述特征值；
- 根据所述流量数据的特征向量，将所有所述流量数据聚类成 M 个簇， M 为大于等于2的正整数；
- 确定在各种 M 和 N 取值的组合下所述流量数据分成的簇的簇错误数总和，所述簇错误数总和是分成的每个簇的错误数相加的结果，每个簇的错误数是指该簇中白数据的数量和黑数据的数量中较少的一个数量；
- 将最小的簇错误数总和所对应的特征个数和簇个数，作为流量数据聚类时选取的目标特征数和目标簇个数。
- [权利要求 2] 根据权利要求1所述的方法，其特征在于，所述在预置特征库中选取 N 个特征包括：在预置特征库中选取卡方值从高到低前 N 个特征。
- [权利要求 3] 如权利要求1所述的方法，其特征在于，流量数据中白数据和黑数据的比为预设比例。
- [权利要求 4] 根据权利要求3所述的方法，其特征在于，所述预设比例为1:1。
- [权利要求 5] 根据权利要求1所述的方法，其特征在于，所述根据所述流量数据的特征向量，将所有所述流量数据聚类成 M 个簇，包括：
- 对所述流量数据的特征向量所包括的各特征值进行归一化处理，得到归一化特征向量，其中归一化处理是用所述流量数据的特征向量所包括的一个特征的特征值除以所有所述流量数据的特征向量所包括的该

特征的最大特征值的结果；

将所述归一化特征向量聚类成M个簇。

[权利要求 6]

根据权利要求1所述的方法，其特征在于，在根据所述流量数据的特征向量，将所有所述流量数据聚类成M个簇之前，所述方法还包括：根据流量数据的特征向量将一部分流量数据聚类成M个簇，M为大于等于2的正整数；

确定在各种M和N取值的组合下所述流量数据分成的簇的簇错误数总和，所述簇错误数总和是分成的每个簇的错误数相加的结果，每个簇的错误数是指该簇中白数据的数量和黑数据的数量中较少的一个数量；

将从小到大前预定名次的簇错误数总和所对应的M和N的组合，作为候选特征数M和N的组合，

且所述确定在各种M和N取值的组合下所述流量数据分成的簇的簇错误数总和包括：确定在各种候选M和N取值的组合下所述流量数据分成的簇的簇错误数总和。

[权利要求 7]

根据权利要求1所述的方法，其特征在于，在将最小的簇错误数总和所对应的特征个数和簇个数，作为流量数据聚类时选取的目标特征数和目标簇个数之后，所述方法还包括：确定在按照选取的目标特征数和目标簇个数聚类之后，聚成的每个簇的簇风险评分，该簇风险评分的公式如下：

$$\text{score} = \frac{N1}{N0 + N1}$$

，其中

N0、N1

分别表示该簇中白样本、黑样本个数，score为簇风险评分。

[权利要求 8]

一种流量数据的聚类处理装置，其特征在于，所述流量数据包括白数据和黑数据，所述白数据是从确定为白用户的用户的数据流量中抽取的流量数据，所述黑数据是从确定为黑用户的用户的数据流量中抽取

的流量数据，所述白用户是确定为不会发出异常流量数据的用户，所述黑用户是确定为会发出异常流量数据的用户，所述装置包括：

选取单元，其被配置为：在预置特征库中选取 N 个特征， N 为正整数；

获取单元，其被配置为：基于流量数据的所选取的特征对应的特征值，得到所述流量数据的特征向量，所述特征向量包括所述流量数据的所述 N 个特征各自对应的特征值；其中，一个所述特征对应一个所述特征值；

第一聚类单元，其被配置为：根据所述流量数据的特征向量，将所有所述流量数据聚类成 M 个簇， M 为大于等于2的正整数；

第一确定单元，其被配置为：确定在各种 M 和 N 取值的组合下所述流量数据分成的簇的簇错误数总和，所述簇错误数总和是分成的每个簇的错误数相加的结果，每个簇的错误数是指该簇中白数据的数量和黑数据的数量中较少的一个数量；

第一设置单元，其被配置为：将最小的簇错误数总和所对应的特征个数和簇个数，作为流量数据聚类时选取的目标特征数和目标簇个数。

[权利要求 9] 根据权利要求8所述的装置，其特征不在于，所述选取单元被配置为：在预置特征库中选取卡方值从高到低的前 N 个特征。

[权利要求 10] 如权利要求8所述的装置，其特征不在于，流量数据中白数据和黑数据的比为预设比例。

[权利要求 11] 根据权利要求10所述的装置，其特征不在于，所述预设比例为1:1。

[权利要求 12] 根据权利要求8所述的装置，其特征不在于，所述第一聚类单元包括：归一化单元，其被配置为：对所述流量数据的特征向量所包括的各特征值进行归一化处理，得到归一化特征向量，其中归一化处理是用所述流量数据的特征向量所包括的一个特征的特征值除以所有所述流量数据的特征向量所包括的该特征的最大特征值的结果；

聚类子单元，其被配置为：将所述归一化特征向量聚类成 M 个簇。

[权利要求 13] 根据权利要求8所述的装置，其特征不在于，还包括：

第二聚类单元，被配置为：在第一聚类单元将所有所述流量数据聚类成M个簇之前，根据流量数据的特征向量将一部分所述流量数据聚类成M个簇，M为大于等于2的正整数；

第二确定单元，其被配置为：确定在各种M和N取值的组合下所述一部分所述流量数据分成的簇的簇错误数总和，所述簇错误数总和是分成的每个簇的错误数相加的结果，每个簇的错误数是指该簇中白数据的数量和黑数据的数量中较少的一个数量；

第二设置单元，其被配置为：将所述第二确定单元确定的从小到大前预定名次的簇错误数总和和所对应的M和N的组合，作为候选特征数M和N的组合，

其中，所述第一确定单元被配置为通过执行如下步骤来实现所述确定在各种M和N取值的组合下所述流量数据分成的簇的簇错误数总和：

确定在各种候选M和N取值的组合下所述流量数据分成的簇的簇错误数总和。

[权利要求 14]

根据权利要求8所述的装置，其特征在于，还包括：

风险确定单元，其被配置为：确定在按照选取的目标特征数和目标簇个数聚类之后，聚成的每个簇的簇风险评分，该簇风险评分的公式如下：

$$\text{score} = \frac{N1}{N0+N1}$$

，其中

N0、N1

分别表示该簇中白样本、黑样本个数，score为簇风险评分。

[权利要求 15]

一种电子设备，其特征在于，所述电子设备包括：

存储器，所述存储器上存储有计算机可读指令；

处理器，其被配置为执行所述存储器上存储的计算机可读指令，其中，所述处理器通过执行所述计算机可读指令被配置为执行一种流量数据的聚类处理方法，其中，所述流量数据包括白数据和黑数据，所述

白数据是从确定为白用户的用户的数据流量中抽取的流量数据，所述黑数据是从确定为黑用户的用户的数据流量中抽取的流量数据，所述白用户是确定为不会发出异常流量数据的用户，所述黑用户是确定为会发出异常流量数据的用户，所述流量数据的聚类处理方法包括：

在预置特征库中选取 N 个特征， N 为正整数；

基于流量数据的所选取的特征对应的特征值得到所述流量数据的特征向量；所述特征向量包括流量数据的所述 N 个特征各自对应的特征值，其中一个所述特征对应一个所述特征值；

根据流量数据的特征向量将所有所述流量数据聚类成 M 个簇， M 为大于等于2的正整数；

确定在各种 M 和 N 取值的组合下所述流量数据分成的簇的簇错误数总和，所述簇错误数总和是分成的每个簇的错误数相加的结果，每个簇的错误数是指该簇中白数据的数量和黑数据的数量中较少的一个数量；

将最小的簇错误数总和所对应的特征个数和簇个数，作为流量数据聚类时选取的目标特征数和目标簇个数。

- [权利要求 16] 根据权利要求15所述的电子设备，其特征在于，所述在预置特征库中选取 N 个特征包括：在预置特征库中选取卡方值从高到低前 N 个特征。
- [权利要求 17] 根据权利要求15所述的电子设备，其特征在于，流量数据中白数据和黑数据的比为预设比例。
- [权利要求 18] 根据权利要求17所述的电子设备，其特征在于，所述预设比例为1:1。
- [权利要求 19] 根据权利要求15所述的电子设备，其特征在于，所述处理器通过执行所述计算机可读指令被配置为通过执行如下步骤来实现所述根据所述流量数据的特征向量，将所有所述流量数据聚类成 M 个簇：
- 对所述流量数据的特征向量所包括的各特征值进行归一化处理，得到归一化特征向量，其中归一化处理是用所述流量数据的特征向量所包

括的一个特征的特征值除以所有所述流量数据的特征向量所包括的该特征的最大特征值的结果；

将所述归一化特征向量聚类成M个簇。

[权利要求 20]

根据权利要求15所述的电子设备，其特征在于，所述处理器通过执行所述计算机可读指令被配置为：在根据所述流量数据的特征向量，将所有所述流量数据聚类成M个簇之前，执行如下步骤：

根据流量数据的特征向量将一部分流量数据聚类成M个簇，M为大于等于2的正整数；

确定在各种M和N取值的组合下所述流量数据分成的簇的簇错误数总和，所述簇错误数总和是分成的每个簇的错误数相加的结果，每个簇的错误数是指该簇中白数据的数量和黑数据的数量中较少的一个数量；

将从小到大前预定名次的簇错误数总和所对应的M和N的组合，作为候选特征数M和N的组合，

其中，所述处理器通过执行所述计算机可读指令被配置为通过执行如下步骤来实现所述确定在各种M和N取值的组合下所述流量数据分成的簇的簇错误数总和：确定在各种候选M和N取值的组合下所述流量数据分成的簇的簇错误数总和。

[权利要求 21]

根据权利要求15所述的电子设备，其特征在于，所述处理器通过执行所述计算机可读指令被配置为：在执行所述将最小的簇错误数总和所对应的特征个数和簇个数，作为流量数据聚类时选取的目标特征数和目标簇个数之后，执行步骤：确定在按照选取的目标特征数和目标簇个数聚类之后，聚成的每个簇的簇风险评分，该簇风险评分的公式如下：

$$\text{score} = \frac{N1}{N0+N1}$$

，其中

N0、N1

分别表示该簇中白样本、黑样本个数， score 为簇风险评分。

[权利要求 22]

一种计算机可读存储介质，其特征在于，其存储计算机程序，所述计算机程序在被计算机执行时使得所述计算机被配置为执行一种流量数据的聚类处理方法，其中，所述流量数据包括白数据和黑数据，所述白数据是从确定为白用户的用户的数据流量中抽取的流量数据，所述黑数据是从确定为黑用户的用户的数据流量中抽取的流量数据，所述白用户是确定为不会发出异常流量数据的用户，所述黑用户是确定为会发出异常流量数据的用户，所述流量数据的聚类处理方法包括：

在预置特征库中选取 N 个特征， N 为正整数；

基于流量数据的所选取的特征对应的特征值，得到所述流量数据的特征向量；所述特征向量包括所述流量数据的所述 N 个特征各自对应的特征值，其中，一个所述特征对应一个所述特征值；

根据所述流量数据的特征向量，将所有所述流量数据聚类成 M 个簇， M 为大于等于2的正整数；

确定在各种 M 和 N 取值的组合下所述流量数据分成的簇的簇错误数总和，所述簇错误数总和是分成的每个簇的错误数相加的结果，每个簇的错误数是指该簇中白数据的数量和黑数据的数量中较少的一个数量；

将最小的簇错误数总和所对应的特征个数和簇个数，作为流量数据聚类时选取的目标特征数和目标簇个数。

[权利要求 23]

根据权利要求22所述的计算机可读存储介质，其特征在于，所述在预置特征库中选取 N 个特征包括：在预置特征库中选取卡方值从高到低前 N 个特征。

[权利要求 24]

根据权利要求22所述的计算机可读存储介质，其特征在于，流量数据中白数据和黑数据的比为预设比例。

[权利要求 25]

根据权利要求24所述的计算机可读存储介质，其特征在于，所述预设比例为1:1。

[权利要求 26]

根据权利要求22所述的计算机可读存储介质，其特征在于，所述计算

机程序在被计算机执行时使得所述计算机被配置为通过执行如下步骤来实现所述根据所述流量数据的特征向量，将所有所述流量数据聚类成M个簇：

对所述流量数据的特征向量所包括的各特征值进行归一化处理，得到归一化特征向量，其中归一化处理是用所述流量数据的特征向量所包括的一个特征的特征值除以所有所述流量数据的特征向量所包括的该特征的最大特征值的结果；

将所述归一化特征向量聚类成M个簇。

[权利要求 27] 根据权利要求22所述的计算机可读存储介质，其特征在于，所述计算机程序在被计算机执行时使得所述计算机被配置为：在根据所述流量数据的特征向量，将所有所述流量数据聚类成M个簇之前，执行如下步骤：

根据所述流量数据的特征向量，将一部分所述流量数据聚类成M个簇，M为大于等于2的正整数；

确定在各种M和N取值的组合下所述流量数据分成的簇的簇错误数总和，所述簇错误数总和是分成的每个簇的错误数相加的结果，每个簇的错误数是指该簇中白数据的数量和黑数据的数量中较少的一个数量；

将从小到大前预定名次的簇错误数总和所对应的M和N的组合，作为候选特征数M和N的组合，其中，所述处理器通过执行所述计算机可读指令被配置为通过执行如下步骤来实现所述确定在各种M和N取值的组合下所述流量数据分成的簇的簇错误数总和：确定在各种候选M和N取值的组合下所述流量数据分成的簇的簇错误数总和。

[权利要求 28] 根据权利要求22所述的计算机可读存储介质，其特征在于，所述计算机程序在被计算机执行时使得所述计算机被配置为：在执行所述将最小的簇错误数总和所对应的特征个数和簇个数，作为流量数据聚类时选取的目标特征数和目标簇个数之后，执行步骤：确定在按照选取的目标特征数和目标簇个数聚类之后，聚成的每个簇的簇风险评分，该

簇风险评分的公式如下：

$$\text{score} = \frac{N1}{N0+N1}$$

，其中

$N0$ 、 $N1$ ：

分别表示该簇中白样本、黑样本个数， score 为簇风险评分。

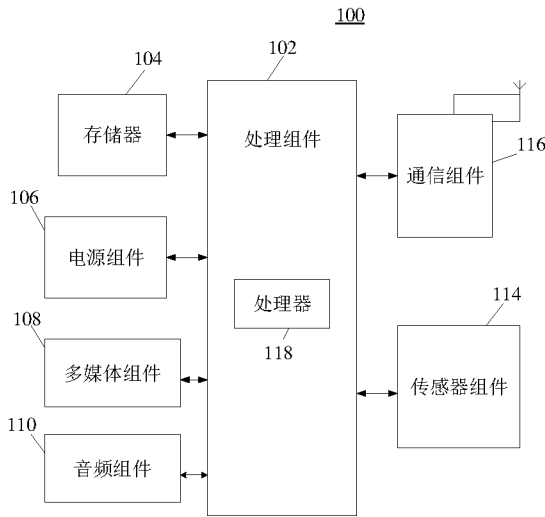


图 1

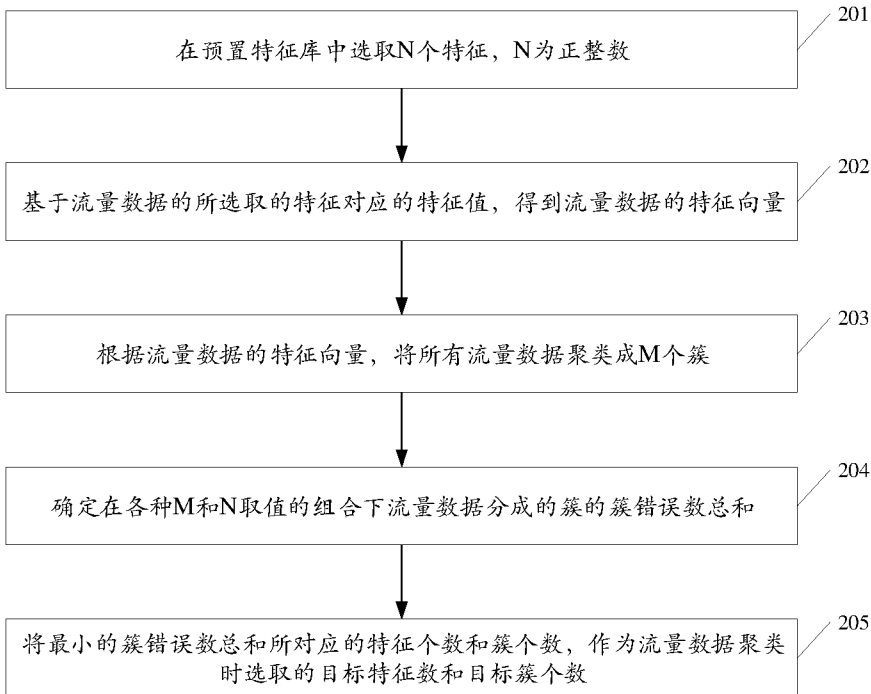


图 2

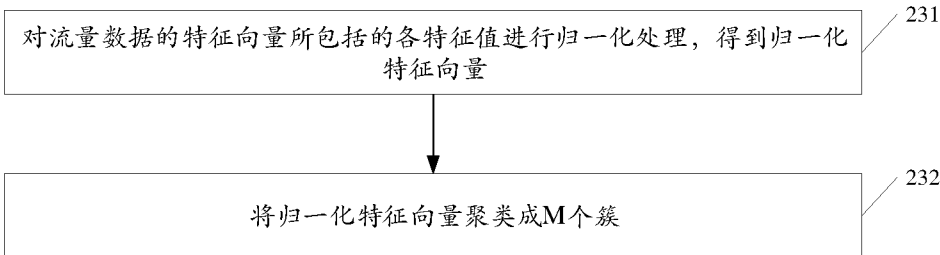


图 3

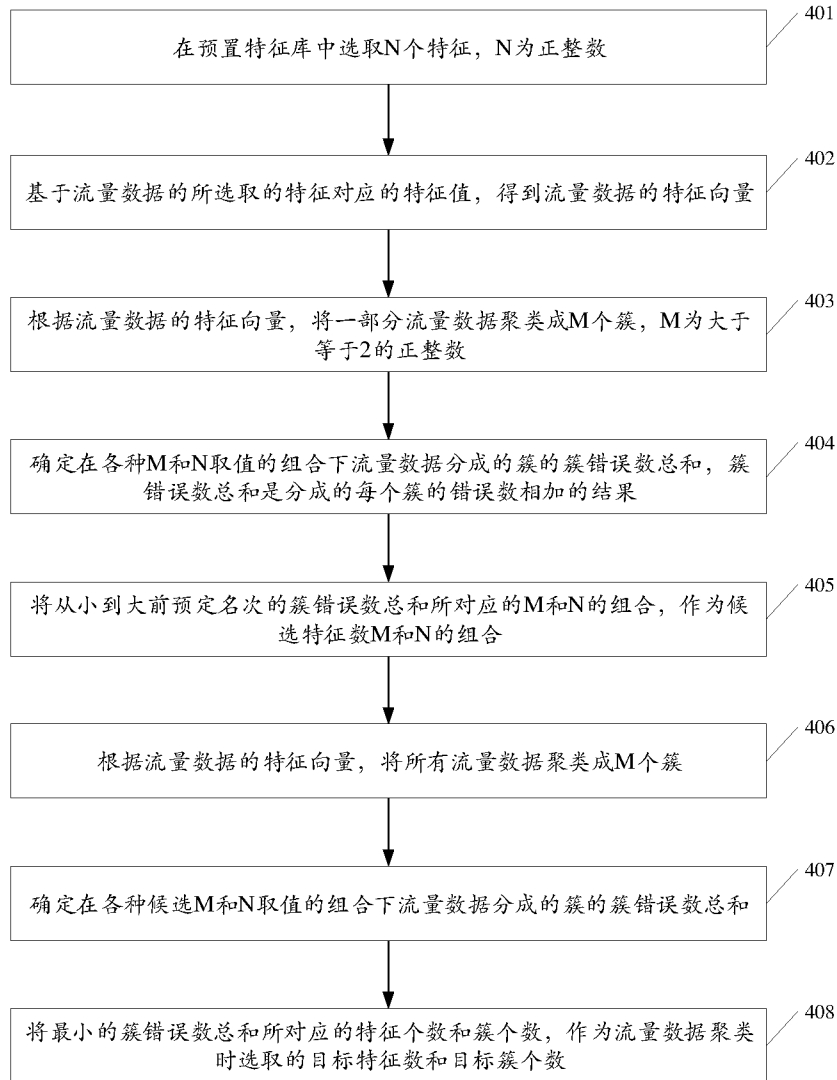


图 4

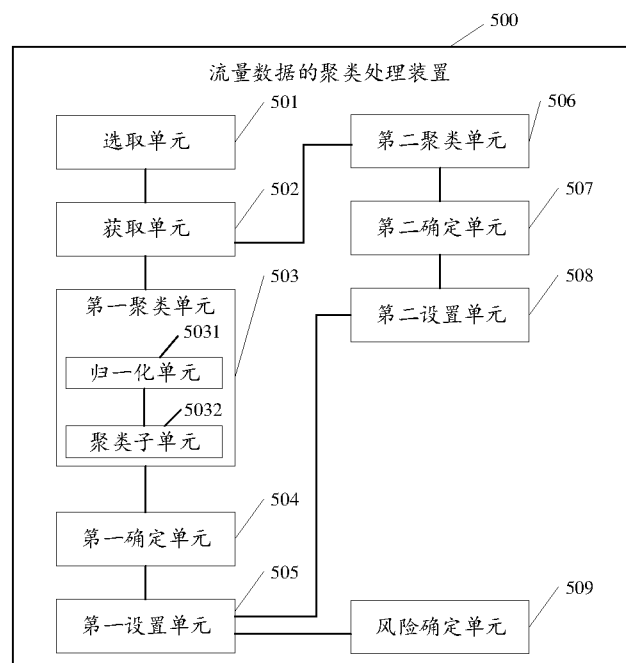


图 5

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2018/125246

A. CLASSIFICATION OF SUBJECT MATTER

G06F 16/2455(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)

CNPAT, CNKI, WPI, EPODOC, IEEE: 流量, 数据, 聚类, 白, 黑, 用户, 异常, 特征, 向量, 特征值, 簇, 错误, flow, traffic, data, cluster, white, black, user, abnormal, characteristic, vector, value, error

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 105141604 A (STATE GRID CORPORATION OF CHINA ET AL.) 09 December 2015 (2015-12-09) description, paragraphs [0032]-[0036]	1-28
A	CN 103001825 A (COMPUTER NETWORK INFORMATION CENTER, CHINESE ACADEMY OF SCIENCES) 27 March 2013 (2013-03-27) entire document	1-28
A	CN 107592323 A (JIANGSU IOT RESEARCH DEVELOPMENT CENTER) 16 January 2018 (2018-01-16) entire document	1-28
A	US 2017134401 A1 (RADWARE, LTD.) 11 May 2017 (2017-05-11) entire document	1-28

☐ 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

03 June 2019

Date of mailing of the international search report

03 July 2019

Name and mailing address of the ISA/CN

**China National Intellectual Property Administration (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/CN2018/125246

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN	105141604	A	09 December 2015	None	
CN	103001825	A	27 March 2013	None	
CN	107592323	A	16 January 2018	None	
US	2017134401	A1	11 May 2017	None	

国际检索报告

国际申请号

PCT/CN2018/125246

A. 主题的分类

G06F 16/2455 (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))

CNPAT, CNKI, WPI, EP0D0C, IEEE: 流量, 数据, 聚类, 白, 黑, 用户, 异常, 特征, 向量, 特征值, 簇, 错误, flow, traffic, data, cluster, white, black, user, abnormal, characteristic, vector, value, error

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 105141604 A (国家电网公司 等) 2015年 12月 9日 (2015 - 12 - 09) 说明书第[0032]-[0036]段	1-28
A	CN 103001825 A (中国科学院计算机网络信息中心) 2013年 3月 27日 (2013 - 03 - 27) 全文	1-28
A	CN 107592323 A (江苏物联网研究发展中心) 2018年 1月 16日 (2018 - 01 - 16) 全文	1-28
A	US 2017134401 A1 (RADWARE, LTD.) 2017年 5月 11日 (2017 - 05 - 11) 全文	1-28

☐ 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 6月 3日

国际检索报告邮寄日期

2019年 7月 3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

传真号 (86-10)62019451

受权官员

刘剑

电话号码 86-(10)-53961304

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2018/125246

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	105141604	A	2015年 12月 9日	无	
CN	103001825	A	2013年 3月 27日	无	
CN	107592323	A	2018年 1月 16日	无	
US	2017134401	A1	2017年 5月 11日	无	