(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 110365603 A
(43)申请公布日 2019.10.22

(21)申请号 201910579744.6

(22)申请日 2019.06.28

(71)申请人 西安交通大学
地址 710049 陕西省西安市咸宁西路28号

(72)发明人 曲桦 赵季红 都鹏飞 段喆琳 崔若星 徐阳

(74)专利代理机构 西安通大专利代理有限责任公司 61200
代理人 安彦彦

(51)Int.Cl.
H04L 12/851(2013.01)
H04L 12/26(2006.01)

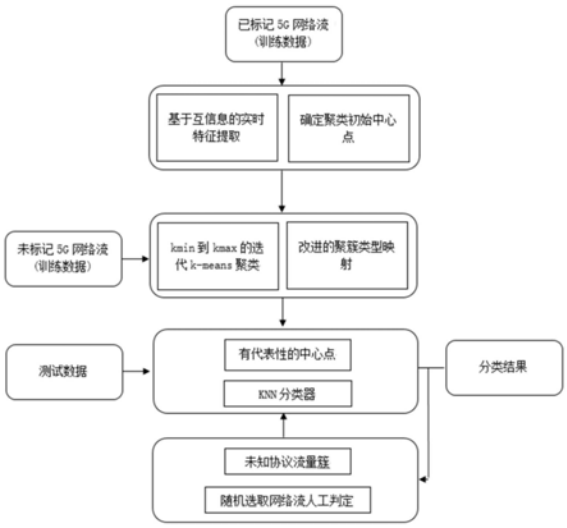
权利要求书3页 说明书8页 附图1页

(54)发明名称

一种基于5G网络能力开放的自适应网络流量分类方法

(57)摘要

本发明公开了一种基于5G网络能力开放的自适应网络流量分类方法,包括以下步骤:1)构建整体数据,再对整体数据的特征向量进行提取;2)将步骤1)提取的各特征向量作为数据样本,通过已标记数据样本的已知类别信息计算初始聚类中心以优化k-means算法,得若干k-means中心点,利用各k-means中心点构建初始中心点集合M;3)利用初始中心点集合M对网络流进行k-means聚类,利用k个簇及k个簇中心点根据评价函数得到聚簇结果;4)统计聚簇已标记网络流的数目,根据聚簇已标记网络流的数目进行网络流的分类,实现基于5G网络能力开放的自适应网络流量分类,该方法建模时间短,时空复杂度低,且应用范围广。



CN 110365603 A

1.一种基于5G网络能力开放的自适应网络流量分类方法,其特征在于,包括以下步骤:

1)采用自适应的滑动窗动态对网络增量数据进行处理,通过原始滑动窗口数据与增量滑动窗口数据构建整体数据,再对整体数据的特征向量进行提取;

2)将步骤1)提取的各特征向量作为数据样本,其中,一分部数据样本已标记网络流,另一部分数据样本未标记网络流,通过已标记数据样本的已知类别信息计算初始聚类中心以优化k-means算法,得若干k-means中心点,利用各k-means中心点构建初始中心点集合M;

3)利用初始中心点集合M对网络流进行k-means聚类,得k个簇和k个簇中心点,然后利用k个簇及k个簇中心点根据评价函数得到聚簇结果;

4)统计聚簇已标记网络流的数目,当聚簇中已标记网络流的数目小于预设网络流阈值时,则该聚簇为未知协议簇;当聚簇中已知网络流的数目大于等于预设网络流阈值时,则根据最大后验概率公式计算各类别的已标记网络流的后验概率,将最大后验概率值对应的类别作为该网络流的类别,实现基于5G网络能力开放的自适应网络流量分类。

2.根据权利要求1所述的基于5G网络能力开放的自适应网络流量分类方法,其特征在于,步骤1)的具体操作为:

采用自适应滑动窗口动态地对网络增量数据进行处理,将原始窗口数据与增量窗口数据分别用矩阵为$X_1=[x_1,x_2,\cdots,x_m]$及$X_2=[x_{m+1},x_{m+2},\cdots,x_{m+r}]$,所有数据样本可表示为$X=[X_1,X_2]$,设所有数据样本的互信息矩阵为S,原始窗口数据的互信息矩阵为S_1,新增窗口数据的互信息矩阵为S_2,则所有数据样本的互信息矩阵S为:

$$S=\frac{1}{m+r}\left(S_1+S_2\right) \quad (1)$$

利用S_1的特征分解将S_1对角化为单位阵,即

$$H_1^T S_1 H_1 = I \quad (2)$$

然后将S_2投影到由H_1张成的空间,则有

$$\overline{S_2}=H_1^T S_2 H_1 \quad (3)$$

将式(1)和式(2)相加,得:

$$H_1^T\left(S_1+S_2\right)H_1P_2=I+\overline{S} \quad (4)$$

求得$\overline{S_2}$的特征分解,即:

$$\overline{S_2}=P_2\Lambda_2P_2^T \quad (5)$$

将式(5)带入式(4),得:

$$P_2^T H_1^T\left(S_1+S_2\right)H_1P_2 = I+\Lambda_2 \quad (6)$$

由式(1)和式(6),得所有数据样本的互信息矩阵S的特征分解,由式(2)可知:

$$H_1=B_1\Lambda_1^{-\frac{1}{2}} \quad (7)$$

其中,$B_i\in R^{n\times k}$为原始数据的主成分决策矩阵,$\Lambda_1\in R^{m\times k}$为选取的前k个特征值组成的矩阵;

根据式(5)求出S_2的特征值$\Lambda_2=[\mu_1,\mu_2,\cdots,\mu_n]$、特征向量$P_2=[\beta_1,\beta_2,\cdots,\beta_n]$及其对应

的特征向量，根据所述k个特征值和特征向量求得S的特征值为：

$$\Lambda = \frac{1}{m+r}\left(I + \mu_i\right) \quad (8)$$

其中，m和r分别是历史数据和新增数据的样本数量；

S的特征向量：

$P = H_1\beta_i$ (9)。

3.根据权利要求1所述的基于5G网络能力开放的自适应网络流量分类方法，其特征在于，步骤2)的具体操作为：

将步骤1)得到的各特征向量作为数据样本，其中，一分部数据样本已标记网络流，另一部分数据样本未标记网络流；

通过已标记数据样本的已知类别信息计算初始聚类中心以优化k-means算法；并利用已标记网络流计算k-means中心点，其中，

$$m_i = \frac{1}{n_i}\sum_{f \in C_i} f \quad (10)$$

其中，每个k-means中心点m_i由属于类别C_i的已标记网络流f确定，n_i表示属于类别C_i的已标记网络流f的数目，利用各k-means中心点m_i构建初始中心点集合M。

4.根据权利要求1所述的基于5G网络能力开放的自适应网络流量分类方法，其特征在于，步骤3)的具体操作为：

31)利用初始中心点集合M对混合的网络流确定k-means聚类，得k个簇和k个簇中心点；

32)根据所述k个簇和k个簇中心点计算评价函数，得评价函数的值，同时利用所述k个簇中心点重置集合M，得新的集合M；

33)计算网络流特征向量集X中离所述新的集合M中所有中心点的距离和最大的k个向量点；

34)根据密度计算公式，确定在所述距离和最大的k个向量点中密度最大的向量点，并将所述密度最大的向量点添加到所述新的集合M中；

35)更新k值为k+1，转至步骤31)，直至k大于$\lfloor\sqrt{N}\rfloor$；

36)统计每次迭代时步骤32)中评价函数的值，从所有评价函数的值中选取最小值，获取最小评价函数的值对应的k值，再将该k值对应的聚簇结果输出。

5.根据权利要求1所述的基于5G网络能力开放的自适应网络流量分类方法，其特征在于，步骤4)的具体操作为：

对聚簇C_i，统计簇中已标记网络流的总数目n_{C_i}，当n_{C_i}小于预设网络流阈值γ_i时，将该簇C_i为未知协议簇；当n_{C_i}的取值大于等于预设网络流阈值γ_i时，则计算各类别的已标记网络流的后验概率，统计最大的后验概率值，并将该簇判定成最大后验概率值对应的网络流类型。

6.根据权利要求1所述的基于5G网络能力开放的自适应网络流量分类方法，其特征在于，聚簇中已标记网络流的最大后验概率为：

$$P\left(L = l_j\right) = \frac{n_{ij}}{n_i} \quad (14)$$

n_{ij}表示簇i中的已标记网络流中属于类型j的网络流数目，n_i表示簇i中已标记流的总数量。

一种基于5G网络能力开放的自适应网络流量分类方法

技术领域

[0001] 本发明属于网络信息领域,涉及一种基于5G网络能力开放的自适应网络流量分类方法。

背景技术

[0002] 近年来,能力开放市场潜力大,客户需求旺盛,通信能力开放与集成已成为运营商未来增长的热点和5G网络发展的重点。5G阶段是网络使能到业务使能的转变,对于网络能力的调用程度更广更深,能力开放的种类和范围更多。但是5G能力开放目前标准层面和业务需求层面仍处于研究探索阶段,从调研反馈情况来看,5G能力开放的产品研发工作还未深入展开,因此,非常有必要对标准层面和演进策略方面开展相关的研究工作,需要通过对5G网络能力开放开展深入的发展策略研究,进而推动业务能力平台的平滑演进。

[0003] 在当今复杂的网络环境下,为了实现对网络流量的有效监管和控制,对网络带宽资源进行合理分配并保证网络信息的安全可靠传输,网络流量分类技术的研究变得尤为重要。与此同时,与传统4G或3G的网络环境相比,在5G的网络中出现了大量的新型应用,这些应用所带来的未知协议流量使网络流量的构成变得更加复杂。据统计,在目前网络中属于新型应用的网络流量已占到未识别数据网络流的60%和未识别比特数的30%,因此,在进行网络流量分类时,若分类器不对这些新型未知协议网络流量进行处理,将严重影响到网络流量分类的整体准确率。大量出现的新型应用为传统的网络流量分类方法带来了一些技术难题,需要新的技术来改进原有的分类方案,适应如今复杂的网络环境。

[0004] 近年来,目前网络流量分类方法主要有4种:基于端口的流量分类技术、基于深度包检测(DeepPacket Inspection,DPI)的流量分类技术、基于流统计特征的机器学习(Machine Learning,ML)的流量分类方法和基于用户行为特征的流量分类技术。

[0005] 基于端口的流量分类技术

[0006] 在互联网发展的初期,网络中的流量种类和数目都相对较少,互联网数字分配机构IANA组织为一些常见的网络协议分配了固定的端口号,因此在早期对网络流量进行分类时,可以通过识别数据包的源端口号和目的端口号来判断该流量的所属的应用协议类型。

[0007] 基于深度包检测的流量分类技术

[0008] 数据包的载荷部分含有大量信息,DPI就是利用这些信息进行分类。基于DPI流量分类技术是依据特定协议或应用的特征码来实现的,通过对网络流量中的载荷数据进行特征码匹配,来获取流量的分类。

[0009] 基于DPI的分类技术虽然具有较高的准确率,但也存在一些缺点:消耗较多的计算资源,对数据加密分类能力较弱,各类应用特征码提取和更新比较困难,对载荷数据的分析会带来对用户隐私权的侵犯。

[0010] 基于流统计特征的机器学习方法

[0011] 近年来,随着人工智能技术的发展,越来越多的研究者开始利用机器学习算法来解决流量分类问题。利用机器学习解决流量分类问题,主要有两个部分:训练数据集和机器

学习算法。训练数据集的生成首先需要利用DPI工具、系统进程监控或人工的方法标注训练样本，得到样本标签，然后从网络流量中提取数据流的特征，最后利用训练集和机器学习算法得到分类器，即可用训练好的分类器对网络流量进行分类。

[0012] 基于用户行为特征的流量分类技术

[0013] 随着流特征加密技术的涌现，给基于流统计特征进行流量分类带来了一定的局限性。近年来，研究人员开始利用主机不同的通信行为模式进行网络流量分类，并提出了利用用户连接模式、连接图、网络连接直径等主机行为特征，对网络流量进行分析，开辟了分析网络流量分类的新方法。基于行为特征的流量分类技术主要通过分析网络协议和应用的连接特性和行为模式上的固有特性，达到对不同流量进行分类的目的。这种方法通常建模时间较长，时空复杂度高，应用有一定的局限性。

发明内容

[0014] 本发明的目的在于克服上述现有技术的缺点，提供了一种基于5G网络能力开放的自适应网络流量分类方法，该方法建模时间短，时空复杂度低，且应用范围广。

[0015] 为达到上述目的，本发明所述的基于5G网络能力开放的自适应网络流量分类方法包括以下步骤：

[0016] 1)采用自适应的滑动窗动态对网络增量数据进行处理，通过原始滑动窗口数据与增量滑动窗口数据构建整体数据，再对整体数据的特征向量进行提取；

[0017] 2)将步骤1)提取的各特征向量作为数据样本，其中，一分部数据样本已标记网络流，另一部分数据样本未标记网络流，通过已标记数据样本的已知类别信息计算初始聚类中心以优化k-means算法，得若干k-means中心点，利用各k-means中心点构建初始中心点集合M；

[0018] 3)利用初始中心点集合M对网络流进行k-means聚类，得k个簇和k个簇中心点，然后利用k个簇及k个簇中心点根据评价函数得到聚簇结果；

[0019] 4)统计聚簇已标记网络流的数目，当聚簇中已标记网络流的数目小于预设网络流阈值时，则该聚簇为未知协议簇；当聚簇中已知网络流的数目大于等于预设网络流阈值时，则根据最大后验概率公式计算各类别的已标记网络流的后验概率，将最大后验概率值对应的类别作为该网络流的类别，实现基于5G网络能力开放的自适应网络流量分类。

[0020] 步骤1)的具体操作为：

[0021] 采用自适应滑动窗口动态地对网络增量数据进行处理，将原始窗口数据与增量窗口数据分别用矩阵为$X_1=[x_1,x_2,\cdots,x_m]$及$X_2=[x_{m+1},x_{m+2},\cdots,x_{m+r}]$，所有数据样本可表示为$X=[X_1,X_2]$，设所有数据样本的互信息矩阵为S，原始窗口数据的互信息矩阵为S_1，新增窗口数据的互信息矩阵为S_2，则所有数据样本的互信息矩阵S为：

[0022] $$S=\frac{1}{m+r}\left(S_1+S_2\right) \quad (1)$$

[0023] 利用S_1的特征分解将S_1对角化为单位阵，即

[0024] $$H_1^T S_1 H_1 = I \quad (2)$$

[0025] 然后将S_2投影到由H_1张成的空间，则有

[0026] $\overline{S_2} = H_1^T S_2 H_1$ （3）

[0027] 将式(1)和式(2)相加，得：

[0028] $H_1^T \left(S_1 + S_2\right) H_1 P_2 = I + \overline{S}$ （4）

[0029] 求得$\overline{S_2}$的特征分解，即：

[0030] $\overline{S_2} = P_2 \Lambda_2 P_2^T$ （5）

[0031] 将式(5)带入式(4)，得：

[0032] $P_2^T H_1^T \left(S_1 + S_2\right) H_1 P_2 = I + \Lambda_2$ （6）

[0033] 由式(1)和式(6)，得所有数据样本的互信息矩阵S的特征分解，由式(2)可知：

[0034] $H_1 = B_1 \Lambda_1^{-\frac{1}{2}}$ （7）

[0035] 其中，$B_i \in R^{n \times k}$为原始数据的主成分决策矩阵，$\Lambda_1 \in R^{m \times k}$为选取的前k个特征值组成的矩阵；

[0036] 根据式(5)求出S_2的特征值$\Lambda_2 = [\mu_1, \mu_2, \cdots, \mu_n]$、特征向量$P_2 = [\beta_1, \beta_2, \cdots, \beta_n]$及其对应的特征向量，根据所述k个特征值和特征向量求得S的特征值为：

[0037] $$\Lambda = \frac{1}{m + r}\left(I + \mu_i\right) \quad (8)$$

[0038] 其中，m和r分别是历史数据和新增数据的样本数量；

[0039] S的特征向量：

[0040] $P = H_1 \beta_i$ (9)。

[0041] 步骤2)的具体操作为：

[0042] 将步骤1)得到的各特征向量作为数据样本，其中，一分部数据样本已标记网络流，另一部分数据样本未标记网络流；

[0043] 通过已标记数据样本的已知类别信息计算初始聚类中心以优化k-means算法；并利用已标记网络流计算k-means中心点，其中，

[0044] $$m_i = \frac{1}{n_i} \sum_{f \in C_i} f \quad (10)$$

[0045] 其中，每个k-means中心点m_i由属于类别C_i的已标记网络流f确定，n_i表示属于类别C_i的已标记网络流f的数目，利用各k-means中心点m_i构建初始中心点集合M。

[0046] 步骤3)的具体操作为：

[0047] 31)利用初始中心点集合M对混合的网络流确定k-means聚类，得k个簇和k个簇中心点；

[0048] 32)根据所述k个簇和k个簇中心点计算评价函数，得评价函数的值，同时利用所述k个簇中心点重置集合M，得新的集合M；

[0049] 33)计算网络流特征向量集X中离所述新的集合M中所有中心点的距离和最大的k个向量点；

[0050] 34)根据密度计算公式，确定在所述距离和最大的k个向量点中密度最大的向量

点，并将所述密度最大的向量点添加到所述新的集合M中；

[0051] 35)更新k值为k+1，转至步骤31)，直至k大于$\lfloor\sqrt{N}\rfloor$；

[0052] 36)统计每次迭代时步骤32)中评价函数的值，从所有评价函数的值中选取最小值，获取最小评价函数的值对应的k值，再将该k值对应的聚簇结果输出。

[0053] 步骤4)的具体操作为：

[0054] 对聚簇C_i，统计簇中已标记网络流的总数目n_{C_i}，当n_{C_i}小于预设网络流阈值γ_i时，将该簇C_i为未知协议簇；当n_{C_i}的取值大于等于预设网络流阈值γ_i时，则计算各类别的已标记网络流的后验概率，统计最大的后验概率值，并将该簇判定成最大后验概率值对应的网络流类型。

[0055] 聚簇中已标记网络流的最大后验概率为：

[0056] $$P\left(L = l_j\right) = \frac{n_{ij}}{n_i} \quad (14)$$

[0057] n_{ij}表示簇i中的已标记网络流中属于类型j的网络流数目，n_i表示簇i中已标记流的总数量。

[0058] 本发明具有以下有益效果：

[0059] 本发明所述的基于5G网络能力开放的自适应网络流量分类方法在具体操作时，先采用自适应的滑动窗动态对网络增量数据进行处理，得到整体数据的特征向量，再计算得到若干k-means中心点，然后利用得到的k-means中心点对网络流进行k-means聚簇，并利用评价函数得到聚簇结果，最后根据聚簇已标记网络流的数目利用后验概率公式获取网络流的类别，操作方便、简单，准确性较高，建模时间短，时空复杂度低，且应用范围广。

附图说明

[0060] 图1为本发明的流程图；

[0061] 图2为本发明中特征提取的示意图。

具体实施方式

[0062] 下面结合附图对本发明做进一步详细描述：

[0063] 参考图1及图2，本发明所述的基于5G网络能力开放的自适应网络流量分类方法包括以下步骤：

[0064] 1)采用自适应的滑动窗动态对网络增量数据进行处理，通过原始滑动窗口数据与增量滑动窗口数据构建整体数据，再对整体数据的特征向量进行提取；

[0065] 2)将步骤1)提取的各特征向量作为数据样本，其中，一分部数据样本已标记网络流，另一部分数据样本未标记网络流，通过已标记数据样本的已知类别信息计算初始聚类中心以优化k-means算法，得若干k-means中心点，利用各k-means中心点构建初始中心点集合M；

[0066] 3)利用初始中心点集合M对网络流进行k-means聚类，得k个簇和k个簇中心点，然后利用k个簇及k个簇中心点根据评价函数得到聚簇结果；

[0067] 4)统计聚簇已标记网络流的数目，当聚簇中已标记网络流的数目小于预设网络流

阈值时，则该聚簇为未知协议簇；当聚簇中已知网络流的数目大于等于预设网络流阈值时，则根据最大后验概率公式计算各类别的已标记网络流的后验概率，将最大后验概率值对应的类别作为该网络流的类别，实现基于5G网络能力开放的自适应网络流量分类。

[0068] 步骤1)的具体操作为：

[0069] 5G网络流中大部分数据并不是全局线性的，它们往往服从一定的形式的非线性分布规律，而传统的一些线性降维算法如主成分分析，Fisher判别分析对于非线性数据的降维效果很差，本发明提出的算法在保持高效的同时，对数据进行降维，保留了原始数据的大部分信息。

[0070] 此外，在未来实时的5G应用过程中，数据采集设备源源不断的从网络流中采集新的数据，传统的特征提取算法无法对增量数据进行快速处理，如果只是单纯地处理新增数据，不考虑历史数据对其影响，算法就无法从全局的角度进行特征提取，所提取的数据蕴含的信息也会大大降低。

[0071] 针对此问题，本发明将历史数据和新增数据相结合，采用自适应滑动窗口动态地对网络增量数据进行处理，将原始窗口数据与增量窗口数据分别用矩阵为$X_1=[x_1,x_2,\cdots,x_m]$及$X_2=[x_{m+1},x_{m+2},\cdots,x_{m+r}]$，所有数据样本可表示为$X=[X_1,X_2]$，设所有数据样本的互信息矩阵为S，原始窗口数据的互信息矩阵为S_1，新增窗口数据的互信息矩阵为S_2，则所有数据样本的互信息矩阵S为：

[0072] $$S=\frac{1}{m+r}\left(S_1+S_2\right) \quad (1)$$

[0073] 利用S_1的特征分解将S_1对角化为单位阵，即

[0074] $$H_1^T S_1 H_1 = I \quad (2)$$

[0075] 然后将S_2投影到由H_1张成的空间，则有

[0076] $$\overline{S_2}=H_1^T S_2 H_1 \quad (3)$$

[0077] 将式(1)和式(2)相加，得：

[0078] $$H_1^T\left(S_1+S_2\right)H_1P_2=I+\overline{S} \quad (4)$$

[0079] 求得$\overline{S_2}$的特征分解，即：

[0080] $$\overline{S_2}=P_2\Lambda_2P_2^T \quad (5)$$

[0081] 将式(5)带入式(4)，得：

[0082] $$P_2^T H_1^T\left(S_1+S_2\right)H_1P_2=I+\Lambda_2 \quad (6)$$

[0083] 由式(1)和式(6)，得所有数据样本的互信息矩阵S的特征分解，由式(2)可知：

[0084] $$H_1=B_1\Lambda_1^{-\frac{1}{2}} \quad (7)$$

[0085] 其中，$B_i\in R^{n\times k}$为原始数据的主成分决策矩阵，$\Lambda_1\in R^{m\times k}$为选取的前k个特征值组成的矩阵；

[0086] 根据式(5)求出S_2的特征值$\Lambda_2=[\mu_1,\mu_2,\cdots,\mu_n]$、特征向量$P_2=[\beta_1,\beta_2,\cdots,\beta_n]$及其对应的特征向量，根据所述k个特征值和特征向量求得S的特征值为：

[0087] $$\Lambda = \frac{1}{m+r}\left(I+\mu_i\right) \quad (8)$$

[0088] 其中,m和r分别是历史数据和新增数据的样本数量;

[0089] S的特征向量:

[0090] $P=H_1\beta_i$ (9)。

[0091] 组成主成分决策阵,将数据映射到主成分决策阵上,即实现了降维,后续的窗口重复此过程。

[0092] 针对传统特征提取算法不适用于非线性数据,且无法满足工业大数据实时性等问题,本发明提出了一种基于互信息的实时特征提取算法,其算法伪代码如下:

[0093] 输入:原始数据集

[0094] 输出:降维后的数据集

[0095] 1.将原始数据集按一定速率输入缓冲区buffer,当速率超过一定值时,动态增加缓冲区.

[0096] 2.滑动窗口从编号最小的缓冲区中读取数据;

[0097] 3.int id=1;/*缓冲区编号初始为1*/

[0098] 4.While(buffer[id]!=null)do

[0099] 5.Read Matrixi;

[0100] 6.if(Matrixi.Id==1)then

[0101] 7.MIandEigDesposition(Matrixi);

[0102] 8.unitedMatrix=UnitedMatrix(Matrixi);

[0103] 9.Output Matrixi*eigVecMatrix

[0104] 10.else

[0105] 11.compute MIMatrix;

[0106] 12.projection MIMatrix on unitedMatrix

[0107] 13.proMatrix=PorjectMatrix(MIMatrix);

[0108] 14.MIandEigDesposition(proMatrix);

[0109] 15.Output MIMatrix*eigVecMatrix

[0110] 16.end if

[0111] 17.id=id+1;

[0112] 18.end while

[0113] 其中,Matrix为窗口内的数据矩阵,用二维数组实现,UnitedMatrix()函数用来求解单位化矩阵,而ProjectMatrix()函数用来求投影后的矩阵;

[0114] 通过逐个扫描每个窗口,先判断当前窗口是否为第一个窗口,如果是,则求出当前窗口的互信息矩阵,然后进行特征分解,选出主成分决策矩阵,然后将原始矩阵映射到决策矩阵上,实现降维;否则,求出本窗口的互信息矩阵,然后将其投影在上个窗口的单位化矩阵上,然后根据式(3)和式(4)求出特征值及特征向量,并组成主成分决策阵,实现降维,整体流程图如图2所示。

[0115] 步骤2)的具体操作为:

[0116] 将步骤1)得到的各特征向量作为数据样本,其中,一分部数据样本已标记网络流,

另一部分数据样本未标记网络流；

[0117] 通过已标记数据样本的已知类别信息计算初始聚类中心以优化k-means算法，以降低k-means的收敛时间，并和下一阶段的迭代添加中心点的方法相结合，提高聚类的准确性。

[0118] 由于聚类算法的目的是将属于同种类别的数据聚集在一起，不同种类别的数据划分到不同的聚簇中去，因此利用已标记网络流的类别，可以计算出一组初始中心点，先大致确定一些聚类范围。利用已标记网络流计算k-means中心点，其中，

[0119] $$m_i = \frac{1}{n_i}\sum_{f \in C_i} f \quad (10)$$

[0120] 其中，每个k-means中心点m_i由属于类别C_i的已标记网络流f确定，n_i表示属于类别C_i的已标记网络流f的数目，利用各k-means中心点m_i构建初始中心点集合M。

[0121] 步骤3)的具体操作为：

[0122] 31)利用初始中心点集合M对混合的网络流确定k-means聚类，得k个簇和k个簇中心点；

[0123] 32)根据所述k个簇和k个簇中心点计算评价函数，得评价函数的值，同时利用所述k个簇中心点重置集合M，得新的集合M；

[0124] 33)计算网络流特征向量集X中离所述新的集合M中所有中心点的距离和最大的k个向量点；

[0125] 34)根据密度计算公式，确定在所述距离和最大的k个向量点中密度最大的向量点，并将所述密度最大的向量点添加到所述新的集合M中；

[0126] 35)更新k值为k+1，转至步骤31)，直至k大于$\lfloor\sqrt{N}\rfloor$；

[0127] 36)统计每次迭代时步骤32)中评价函数的值，从所有评价函数的值中选取最小值，获取最小评价函数的值对应的k值，再将该k值对应的聚簇结果输出。

[0128] 其中，改进的k-means算法中的距离度量为加权欧式距离，即

[0129] $$d(x_i, x_j) = \sqrt{w_1(x_{i1} - x_{j1}) + w_2(x_{i2} - x_{j2}) + \ldots + w_m(x_{im} - x_{jm})} \quad (11)$$

[0130] 其中，改进的k-means算法迭代过程中k值是由k_{min}＝p变化到$k_{max} = \sqrt{N}$，k_{min}的值是由输入的已标记流的类别数量决定的，k_{max}的值则是根据文献总结并验证出的k-means算法的经验最大值所决定的。

[0131] 其中，评价函数的值越小，说明每个簇中各样本点的距离越近，即聚类的效果越好。

[0132] 本发明考虑从距当前的中心点距离最远的k个中心点中选择密度最大的一个进行添加，因为距离最远的点可以有效地避免散发陷入局部最优，而密度最大则可以保证该点的代表性，密度的计算公式为：

[0133] $$\begin{cases} density(x) = \sum_{i=1}^{N} p(AvgDist - |x_i - x|) \\ p(y) = \begin{cases} 1 & y \geq 0 \\ 0 & 其它 \end{cases} \end{cases} \quad (12)$$

[0134] $$AvgDist = \frac{1}{C_N^2} \times \sum d\left(x_i, x_j\right) \quad (13)$$

[0135] 其中，d(xi,xj)表示向量点xi和向量点xj之间的加权欧式距离，C_N^2表示所有向量点两两组合时的计算次数，其中，N为所有向量点的数目。

[0136] 在多次迭代，添加具有代表性的中心点后，可以利用评价函数自动确定迭代过程中得到的最佳聚类结果以及其所对应的k值，不仅实现了参数的自适应，同时保证了输出聚类结果的高准确性。

[0137] 步骤4)的具体操作为：

[0138] 对聚簇Ci，统计簇中已标记网络流的总数目n_{C_i}，当n_{C_i}小于预设网络流阈值γ_i时，将该簇Ci为未知协议簇；当n_{C_i}的取值大于等于预设网络流阈值γ_i时，则计算各类别的已标记网络流的后验概率，统计最大的后验概率值，并将该簇判定成最大后验概率值对应的网络流类型。

[0139] 其中，聚簇中已标记网络流的最大后验概率为：

[0140] $$P\left(L = l_j\right) = \frac{n_{ij}}{n_i} \quad (14)$$

[0141] n_{ij}表示簇i中的已标记网络流中属于类型j的网络流数目，n_i表示簇i中已标记流的总数量。

[0142] 网络流阈值γ_i为：

[0143] $$\gamma_i = \frac{1}{2} r_i n_{C_i} \quad (15)$$

[0144] γ_i为混合输入的网络流中已标记网络流所占的比例，即若某个簇中所有类型的已标记网络流总数目相加，仍然小于该簇中所有网络流的数目乘以γ_i的1/2时，该簇将暂时判定为未知协议簇。考虑到已标记数据是随机选取的，对于属于非未知协议类别的网络流来说，他们所在的聚类簇中应该存在数目大于$r_i n_{C_i}$的已标记网络流。考虑到聚类结果的偶然性，可以认为当某一簇中的已标记网络流的数目小于(1/2)$r_i n_{C_i}$时，根据簇中已标记的网络流的类型进行簇的类型判定时的数据不充足，认为其判定结果不具有代表性，因此将这些簇暂时的划分为未知协议簇，需要在系统更新模块中在对其进行进一步的研究。

[0145] 通过改进的聚簇类别映射方法，可以使得在传统的半监督流量分类方法中会被错误划分到某已知协议类别中的未知协议类别簇也被识别和提取出来，利用这样的聚簇结果训练出的线上分类器，可以大大提高线上分类器的准确率，同时实现线上未知协议的提取。

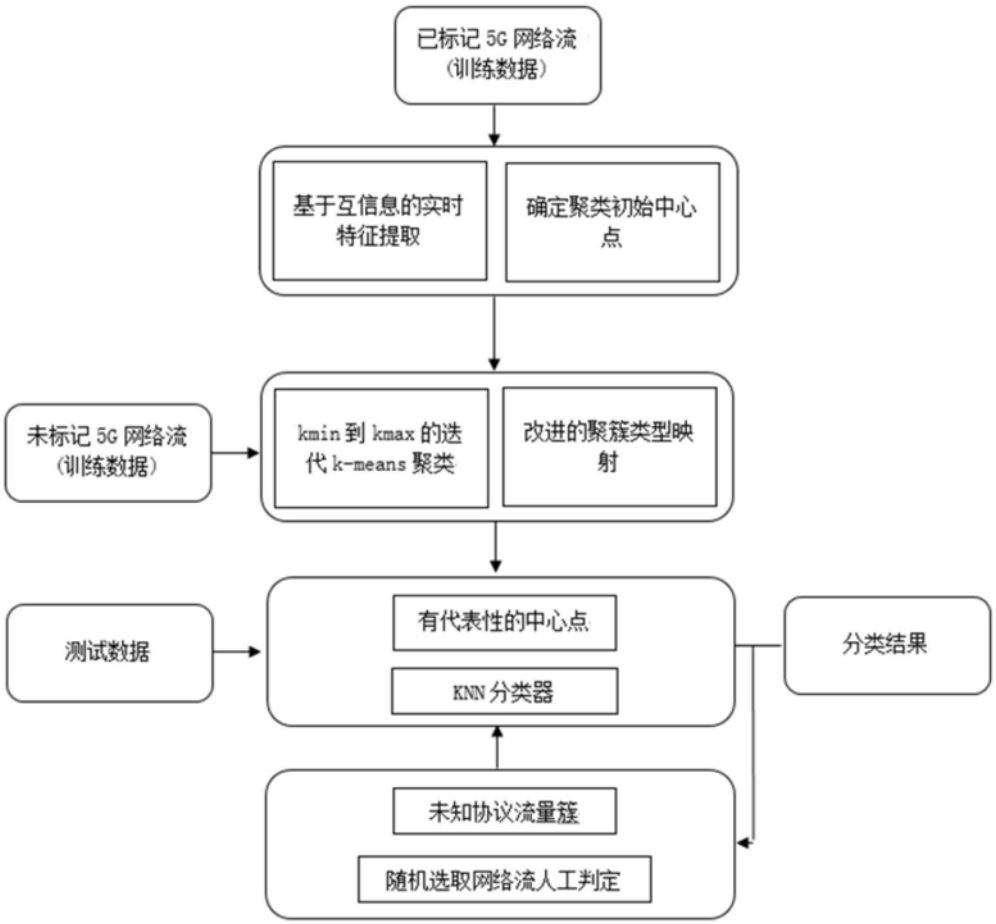

已标记 5G 网络流
(训练数据)
基于互信息的实时
特征提取
确定聚类初始中心
点
未标记 5G 网络流
(训练数据)
kmin 到 kmax 的迭
代 k-means 聚类
改进的聚簇类型映
射
测试数据
有代表性的中心点
KNN 分类器
分类结果
未知协议流量簇
随机选取网络流人工判定

图1

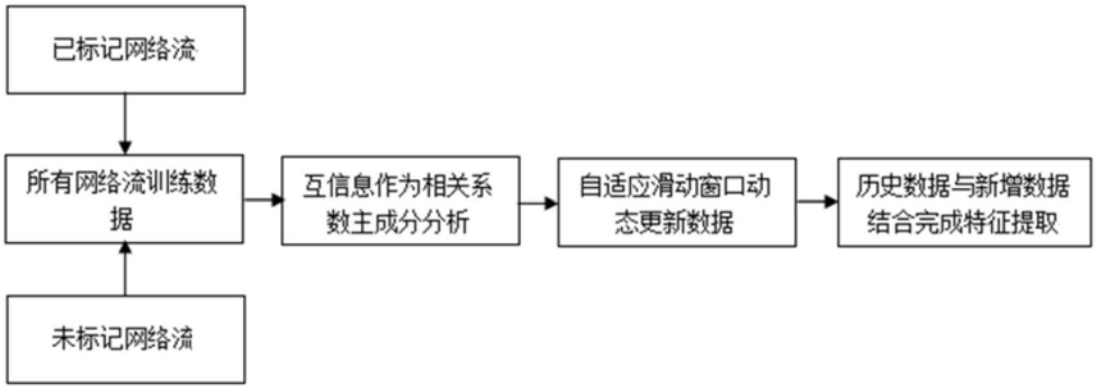

已标记网络流
所有网络流训练数
据
互信息作为相关系
数主成分分析
自适应滑动窗口动
态更新数据
历史数据与新增数据
结合完成特征提取
未标记网络流

图2