基于层次聚类的网络流识别算法研究

丁伟[1,2]，徐杰[1,2]，卓文辉[1,2]

（1.东南大学 计算机科学与工程学院，江苏 南京，211189；
2. 东南大学计算机网络与信息集成教育部重点实验室，江苏 南京，211189）

摘 要：利用核函数定理提出了一种改进的网络流识别算法。首先运用对称不确定性的概念选择出最相关的流测度，然后利用核函数定理对选择的网络流测度进行高维映射，以测度的高维空间距离作为度量各个类差别的标准，提高了聚类结果的准确性。采用光滑因子、轮廓系数和不确定熵来控制聚类过程。实验表明该算法的聚类结果更均匀，没有出现某个类占过大比重的情况且根据高维空间的类距离能够检测出网络流里的大部分流量。

关键字：流量识别；聚类分析；核函数映射；对称不确定分析

中图分类号：TP393 **文献标识码**：A

Net Traffic Identifier based on Hierarchical Clustering

DING Wei[1,2], XU Jie[1,2], ZHUO Weng-Hui[1,2]

(1. School of Computer Science and Engineering, Southeast University, Nanjing, 211189;
2. Key Laboratory of Computer Network and Information Integration, Ministry of Education, Nanjing, 211189)

Abstract: Proposed an improved net traffic identifier algorithm based on semi-supervised clustering. Symmetrical uncertainty was used to reduce the net flow attributes, and then kernel function was used to project the rest attributes to higher dimentional space. The train net flow was clustered in high dimentional space hierarchically. Smooth factor, sihouette coefficient and entropy controlled the cluster process to get a well result. Experiments show that the algorithm got flat clusters without any huge cluster and could identify most net flow even encrypted ones.

Key work: traffic identify; hierarchical cluster; kernel function; sihouette coefficient

1 引言

网络流量识别是网络安全和网络管理中的研究重点，在许多网络应用中有着重要作用如：入侵检测，QoS，网络计费等。传统的网络流量识别主要采用基于端口的

深度包检测(DPI)和深度流检测(DFI)技术是两种主要的流量识别技术它们各有优缺点。DPI 分析了数据包的内容其准确性相当高，但是它依赖于特征库，不能识别新出现的网络应用。DFI 依据网络流量属性对网络流进行分类，不依赖于具体的数据包内容，速度快但是准确率低。

本文提出了一种基于半监督聚类的网络流识别算法(Semi-supervised Clustering Identifier Algorithm，SCIA)，SCIA 利用 DPI 和 DFI 的特点采用半监督的机器学习方法对训练数据流进行聚类，聚类过程中利用轮廓系数、光滑因子和不确定熵来控制划分的类，使得聚类结果更均匀。

本文的研究结果将用于 NBOS 系统，代替原有的流量识别函数。NBOS 是为 211 三期工程开发的一个基于流记录的精细化网络管系统。流量的分类和统计是其功能的一部分。NBOS 系统现有的流量分类函数使用了一个简单的算法，由于篇幅的原因不在本文中介绍。

本文的第二部分介绍了相关的概念和有关的研究成果，第三部分论述了本文提出的网络流层次

收稿日期： 修回日期：

聚类分析算法；第四部分通过实验验证了算法的准确性；第五部分总结了本文的工作。

2 相关工作

2.1 基于聚类方法的网络流分类算法

聚类分析是机器学习和数据挖掘领域的重要组成部分，近年来许多学者将聚类分析的技术引入了网络流分类领域，并取得了较高的准确率。文献[1] [2][10]提出了一种在线层次聚类算法用于网络流量的异常检测。

文献[3] [4] [5]对流量属性进行半监督的聚类分析分别利用 K-means 算法、混合高斯密度和相似系数将已流量和未知流量进行区分。

文献[6]利用随机森林的邻近性测度代替欧氏距离来提高聚类结果的准确性 文献[7]综合多种方法提出了一种混合的网络流分类框架。文献[8]通过谱聚类的方法分析网络连接的行为模式；文献[9]分别利用主成分分析差分映射的方法对日志进行聚类分析找出异常流量的模式；文献[11]用调和平均值作为距离测度，对网络流进行聚类分析，将加密的数据流成功的划分出来。

上述方法主要通过改进距离测度，采用多种机器学习方法来提高聚类的准确性，但是有些流属性的对聚类分析没有帮助甚至回干扰聚类结果。由于网络流的复杂性仅仅通过对流属性的低维统计测度很难得到满意的结果，本文通过核函数的高维映射提高了聚类结果的准确性。

2.2 对称不确定性

熵是随机变量的不确定性度量，令 X 为一随机变量，其熵 H(X)定义为：

$$H(X) = -\sum_{i=1}^{|X|} p(x_i)\log_2(p(x_i)) \tag{1}$$

其中 P(xi) = P(X=xi)。H(X)越大，即 X 的不确定性越大，所携带的自信息量越大。在另一随机变量 Y 的值确定的情况下，变量 X 的条件熵 H(X|Y)定义为：

$$H(X|Y) = -\sum_{j=1}^{|Y|} p(y_j)\sum_{i}^{|X|} p(x_i|y_j)\log_2(p(x_i|y_j)) \tag{2}$$

H(X|Y)表示观察到随机变量 Y 的取值后，仍保留的关于变量 X 的不确定性，差值 H(X)－H(X|Y)表示由随机变量 Y 所提供的关于 X 的信息量， 在信息论中被称为 X 和 Y 之间的互信息量，表示为 I(X; Y)。

互信息量 I(X;Y)可以用来定量衡量两测度之间的相关关系。但是，由于 I(X; Y)的结果受变量取值和单位的影响，故进一步对其进行均一化，得到以下对称不确定性 SU(Symmetrical Uncertainty)[12]的定义：

$$SU(X;Y) = SU(Y;X) = 2\times\left[\frac{I(X;Y)}{H(X)+H(Y)}\right] \tag{3}$$

SU 取值范围在 [0, 1]之间，数值越大表示两变量间相关程度越强。由于 SU 具有较高的准确性和通用性，因此本文将其引入到网络流测度相关性分析领域中，作为定量衡量两流测度、流测度与流类别之间相关关系的标准。

2.3 高斯核函数

假设 $X = \{x_k \in R^N, k = 1,2,\cdots l\}$ 是一个非空的输入空间的样本集，被某种非线性映射Θ 映射到某一特征空间 H 得到$\Theta(x_1), \Theta(x_2), \cdots \Theta(x_l)$ 。如果函数$K: X \times X \to R$ 满足：

$$K(x_i, x_j) = (\Theta(x_i)\cdot\Theta(x_j)), \forall x_i, x_j \in X \tag{4}$$

则称 K 为核函数。本文采用的核函数为高斯核函数（见公式 5），因为高斯核函数的映射空间是无穷维的,在无穷维进行聚类可以更好的区分不同类。

$$K(x_i, x_j) = \exp(-\beta\|x - x\|^2), \beta > 0 \tag{6}$$

在输入空间样本 X 被映射到特征空间 H 后，特征空间的 Euclidean 距离可以表示为：

$$D_{H(x_i,x_j)} = \sqrt{\|x_i - x_j\|^2} = \sqrt{\Theta(x_i)\cdot\Theta(x_i) - 2\Theta(x_i)\cdot\Theta(x_j) + \Theta(x_j)\Theta(x_j)} \tag{6}$$

将公式（4）代入公式（6），即可得到公式（7）：

$$D_{H(x_i,x_j)} = \sqrt{K(x_i,x_i) - 2K(x_i,x_j) + K(x_j,x_j)} \tag{7}$$

再由公式（5），便可以方便的计算出特征空间中点的距离。

3 网络流识别

本文提出的 SCIA 算法共分为三个步骤：属性约简，数据流聚类，按类间距离标识流量。属性约简就是利用 SU 计算不同属性之间的关联性，去除无关测度。数据流聚类时需要对训练数据集进行计

算。NBOS 系统里流量识别模块的数据源为 NetFlow 流记录，无应用层负载。训练数据集采用的原始数据流为完整的网络数据包，包含了应用层的内容，通过 DPI 技术对其进行标识，然后再提取每个数据流的流属性就得到训练数据集。

3.1 流属性约简

NBOS 的数据源为 NetFlow 流（NT 流）记录，基于 NetFlow 的固有字段，可计算的流属性有源/宿端口、报文数、持续时间、平均报文到达间隔等共计 20 个。好的流测度组合应包含足够的类别信息且空间维数尽可能的低，因此测度选择的目的主要有两点，一是去除对类别属性无关的特征，二是去除冗余的特征。本文采用 SU 测度进行流属性选择,步骤如下：

1) 对 NT 流量按照 5 元组进行组流，计算出每条流 m(m=20)个测度的值，随机选取各种应用的流共计 n 条，构成测度集。测度集可以看作是一个 n×m 维的矩阵 X，矩阵中 n×1 维向量 X_j 表示第 j 个测度 n 条流的所有取值，元素 x_{ij} 表示第 i 条流第 j 个测度的取值。n 条流相对应的应用类型可描述为一个 n×1 的矩阵 C，矩阵中元素 c_i 表示第 i 条流所属的应用类别。如此，扩展矩阵 M=[X;C]便构成了我们的属性选择的样本集。

2) 计算向量 X_j 与向量 C 之间 SU 系数，若 SU 系数小于某阈值 δ_1，则认为测度 j 不能提供对分类有用的信息，属于无效测度，不选，即 X=X-{X_j}；反之不作处理。

3) 对 X 中剩余的测度，计算向量两两测度向量间的 SU 系数，若 SU 系数大于某阈值 δ_2，则认为测度 j 与测度 l 相互冗余，删除两者中与类别向量 C 的 SU 系数较小的测度；反之不作处理。

3.2 网络流分层聚类与识别

SCIA 通过对含有应用标号的网络数据流进行分层聚类，将相似的应用合并为一个类，计算出每个类的中心点。聚类的训练数据集为含有应用标识的 NetFlow 流，且流属性仅包含通过 SU 测度过滤后得到的最相关的属性。SCIA 通过光滑因子(Smooth Factor，sf)、轮廓系数(sihouette coefficient，sc)和信息熵来控制聚类的过程。sf 的定义如下：

$$sf = \frac{\sum_{i=1}^{k} \frac{\| c_i \|}{n} \log(\frac{\| c_i \|}{n})}{\log(\frac{1}{k})} \tag{8}$$

k 表示当前类的个数，c_i 表示第 i 类里流记录的个数，n 为整个训练集里流记录的个数。sf 的取值范围为(0,1]，当每个类里的元素都相等时 sf 为 1，此时的聚类结果最均衡，通过 sf 可以防止聚类结果里少数的类里的记录个数占整个训练集的比重过大，使得聚类结果相对均匀。

sc 定义如下：

$$sc = \frac{b - a}{\max(a, b)} \tag{9}$$

其中 a 表示某一样本点与其同簇中其他所有点之间的平均距离。b 表示某一样本点与其相邻簇中所有点之间的平均距离。由定义可知，该系数结合了凝聚度和分离度，因此，可以以此来判断聚类的优良性。轮廓系数的取值在区间[-1,1]间，值越大表示聚类效果越好。

在利用聚类结果进行分类时，可划分的类越多得到的有用信息越多，但是错误率也越高；当分的类越少时准确率会越高但是是以推广性为代价的，所得到的有用信息也越少。因此聚类结果即需要考虑不同类间的距离也需要考虑聚类结果的可用性。SCIA 通过定义最小光滑系数 min_sf，最小轮廓系数 min_sc 和最小熵 min_h 来控制聚类过程。

SCIA 采样自底向上的层次聚类算法，初始时所有具有相同应用标识的数据流为一个类，计算出每个类的中心，然后通过比较各个类之间中心距离的远近进行聚类。具体过程如下：

1) 计算每个初始类的中心；

2) 计算各个类的中心点间的距离；

3) 选取最近的两个中心点，将这两个中心点所在的类作为候选类；

4) 将这两个候选类合并，计算出新类的中心，计算分类的均衡性因子 sf，如果 sf > min_sf，则转第 5 步，否则将新合并类重新划分为原来的两个类，选取底下两个最近的类作为候选类转第 3 步；如果没有最近的两个类则聚类结束；

5) 计算新中心的轮廓系数 sc 和聚类的熵如果该点的轮廓系 h，如果 h<min_h 转第 6 步，如果 sc>min_sc 则转第 2 步，否则转第 6 步；

6) 将最近合并的一个类重新划分为原来的两类，聚类结束。

SCIA采用轮廓系数和熵来判断聚类何时停止。步骤 2 里对两个类间的距离采用两个类中心的高维

空间的欧式距离，距离的远近表示了两个类的接近程度。 通过选取不同的参数可以灵活的调整聚类结果，得到最佳的准确性和可推广性。SCIA 的聚类过程占用时间较多，其中最耗时的步骤为计算各个中心点之间的距离，其时间复杂度为：$O(n^2)$，n 为初始类的个数。

在识别未知流记录时通过计算该流记录与每个类的中心点之间的距离，然后选出距离最短的类，未知流记录即属于此类。虽然 SCIA 的聚类过程需要较多时间但是在分类流记录时其时间复杂度为 O(1)。

4 实验与分析

实验数据为江苏省网边界到 CERNET 国家主干路由之间时长 1 小时带报文负载的 IPTrace 数据。IPTrace 大小为 51GB，含有 4.86×10^8 个数据报文，通过组流后得到 5.67×10^7 个流记录,利用 nDPI 对 IPTrace 进行标识。nDPI[13]是在 OpenDPI 上扩充改进的 DPI 库，能识别多达 170 种应用，本文基于该开源软件，在融合端口识别、连接行为特征的基础上，对 IPtrace 中的流量进行了识别，得到了该数据源中 96%流量的应用分布情况，如表 1 所示。

表 1 IPTrace 流量分布

应用类别	所含协议	字节比重
Mail	pop,smtp,imap,etc	0.196
域名解析	Dns,mdns,whois-das,etc	11.25
DataBase	Postgresql,mssql,mysql,etc	0.498
P2P	cDonkey,BT,pplive,etc	12.67
Interactive	telnet,rdp,ssh,vnc,smb,elc	0.46
VoIP	Skype,teamspcak,viber,etc	4.52
WWW	http,http_proxy,elc	63.41
BULK	ftp	0.054
Service	Ntp,netbios,nfs,upnp,etc	0.65
其他	Remotescan,games,vpn	6.286

在进行属性约减时设 δ_1=0.45、δ_2=0.85 表格 1 给出了通过 SU 方法选择出来的测度结果：

选取 IPTrace 里前 80%的数据作为训练数据，后 20%的数据为测试数据，SCIA 的各参数设置为：min_sc=0.854，min_sf=0.21,min_h=12。最终得到 17 个大类，通过计算测试数据集里每条流记录与各个类的距离将测试数据集归入 17 个大类里，得到测试数据的分类标号。测试数据集中的每条流记录都含有应用标识，根据应用标识与聚类结果之间的关系可以判断每条流记录应该属于哪个大类，以此作为判断分类准确性的标准。测试结果显示分类准确性在 97%以上，接近于 DPI 的准确率。其中对一些加密应用如 hppts、Skype 等的分类准确率在 90%以上，这说明 SCIA 对加密流量的识别率较高，能够实时区此类应用。

表 2 测度描述

测度	描述	重要度
低位端口	NetFlow 字段	1
传输层协议	NetFlow 字段	0.621
双向字节数	前/后向字节数之和	0.582
平均报文长度	双向字节数/双向报文数	0.563
BPS	字节数/持续时间	0.538
平均到达间隔	持续时间/报文数	0.538
双向报文数	前/后向报文数之和	0.521
双向报文数比	前/后向报文数之比	0.519

5 总结

在复杂多变的网络环境下，网络流识别成为了研究的热点和难点。本文提出了一种半监督的网络流分类算法，该算法利用层次聚类的方法对待有应用标识的网络流数据进行聚类分析，将相似的数据流合并为一类。为了提高聚类结果的准确性首先通过对称不确定测度对流属性进行约减，然后利用高斯函数将流属性进行高维映射，在高维空间进行聚类。SCIA 的训练过程可以离线进行，在对流记录进行分类时算法的时间复杂性为常数，分类速度快。随着网络技术的发展网络带宽增长迅速，新的网络应用不断出现，进一步的研究工作是如何自动识别出新的应用、实时检测出不同类型的异常网络流量并将新的算法完全集成到 NBOS 系统。

参考文献:

[1] Truong, P., Guillemin, F.,"Dynamic Binary Tree for Hierarchical Clustering of IP Traffic"[C], Global Telecommunications Conference IEEE, Washington, DC, 2007:6-10.

[2] Xin Du ;Zhengzhou ; Yingjie Yang ,"Research of Applying Information Entropy and Clustering Technique on Network Traffic Analysis"[C],Computational Intelligence and Security, Suzhou, 2008:472-476.

[3] Shukla, D.B;Chandel, G.S;"An approach for classification of network

traffic on semi-supervised data using clustering techniques"[C],Nirma University International Conference on Engineering (NUiCONE), Ahmedabad,28-30 Nov. 2013,pp:1 - 6.

[4] Wang, Y.; Xiang,Y.; Zhang,J.; Zhou, W.,"Internet Traffic Classification Using Constrained Clustering"[C],Parallel and Distributed Systems, IEEE,2013.12,vol:99,pp:1-4

[5] Palnaty, R.P.; Rao, A. "JCADS: Semi-supervised clustering algorithm for network anomaly intrusion detection systems"[C],Advanced Computing Technologies (ICACT15th), Rajampet,21-22 Sept. 2013,:1-5.

[6] Yu Wang; Yang Xiang ; Jun Zhang;" Network traffic clustering using Random Forest proximities"[C],IEEE International Conference on Communications (ICC), Budapest, 9-13 June 2013:2058 - 2062

[7] Wei Lu; Ling Xue " A Heuristic-Based Co-clustering Algorithm for the Internet Traffic Classification"[C],28th International Conference on Advanced Information Networking and Applications Workshops (WAINA), Victoria, BC,13-16 May 2014，pp:49-54.

[8] Tao-Wei Chiou, Shi-Chun Tsai, Yi-Bing Lin "Network security management with traffic pattern clustering"[J],Soft Computing，vol.18,Jan 2014, pp 1757-1770

[9] Juvonen, A. Sipola, T;" Adaptive framework for network traffic classification using dimensionality reduction and clustering"[C],4th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT), St. Petersburg,3-5 Oct. 2012，pp:274-279.

[10] Choonho Son, Seok-Hyung Cho, Jae-Hyoung Yoo ," Volume Traffic Anomaly Detection Using Hierarchical Clustering"[J],Management Enabling the Future Internet for Changing Business and New Computing Services Lecture Notes in Computer Science, vol 5787, 2009, pp 291-300.

[11] Meng Zhang, Hongli Zhang, Bo Zhang ," Encrypted Traffic Classification Based on an Improved Clustering Algorithm"[J],Trustworthy Computing and Services Communications in Computer and Information Science，vol. 320, 2013, pp 124-131

[12] W H Press, S A Teukolsky, W T Vetterling, B P Flannery. Numerical Recipes in C [M]. London: Cambridge University Press, 1988.

[13] Karagiannis T, Papagiannaki K, Faloutsos M. "BLINC:multilevel traffic classification in the dark"[C],ACM SIGCOMM Computer Communication Review. ACM, Aug 2005, 35(4): 229-240.

作者简介：

丁伟（1962-），女，江苏南京人，东南大学教授、博士生导师，主要研究方向为网络测量和网络行为学。

徐杰（1989-），男，江苏泰州，博士研究生，主要研究方向：数据挖掘和网络测量。

卓文辉（1989-），女，江西，硕士，主要研究方向：网络测量。