文章编号：1001－9081(2010)10－2653－03

混合模式的网络流量分类方法

胡　婷[1]，王　勇[1]，陶晓玲[2]
(1. 桂林电子科技大学 计算机科学与工程学院，广西 桂林 541004；2. 桂林电子科技大学 网络中心，广西 桂林 541004)
(water.359@163.com)

摘　要：为了更好地满足用户对各类 Internet业务服务质量越来越精细的要求，流量分类是网络管理的重要环节之一。通过分析、对比基于端口号匹配、特征字段分析和流统计特征的机器学习分类方法的应用现状及其优缺点，针对单一分类方法存在的分类准确度不高、分类时间长等问题，提出一种混合模式的网络流量分类方案。此方案结合端口号匹配和机器学习分类方法，采用输出结果可视化的自组织映射网络算法实现网络流量在应用层的分类。实验表明，该方案能有效地实现对网络流量应用类型的分类，分类结果可视化效果好。

关键词：流量分类；统计特征；机器学习；自组织映射
中图分类号：TP393.08　　文献标志码：A

Network traffic classification based on hybrid model

HU Ting, WANG Yong, TAO Xiao-ling
(1. College of Computer Science and Engineering, Guilin University of Electronic Technology, Guilin Guangxi 541004, China;
2. Network Center, Guilin University of Electronic Technology, Guilin Guangxi 541004, China)

Abstract: In order to satisfy the requirements of users for more and more precise Internet service quality, the traffic classification is an important link in the network management process. Through analyzing and comparing the application situation and the advantages and disadvantages of each classification method by machine learning, which were separately based on port number matching, feature analysis and traffic characteristics, a hybrid model of network traffic classification method was proposed to solve the problems that rely on a single classification method, such as low accuracy, long classification time. This model combined the port number matching with machine learning, and applied Self-Organizing Map (SOM) of which the output result is visual. The experimental result shows that this method can effectively achieve the application type classification of network traffic, and obtain a good visual effect of classification result.

Key words: traffic classification; statistical characteristic; machine learning; Self-Organizing Map (SOM)

0　引言

随着网络规模的扩展，网络应用的复杂化，如何有效管理网络、提供较好的服务质量越来越引起人们的重视。网络流量是记录和反映网络及其用户活动的重要载体。通过对网络流量的各种应用进行分类，可以间接掌握网络的使用情况，可以按照用户需求对网络资源进行 QoS调度，并根据网络应用的发展趋势对现有网络进行扩容改进。此外，在流量计费、网络安全等方面，流量的有效分类也具有重大意义。

基于端口号匹配的分类根据国际互联网代理成员管理局(Internet Assigned Numbers Authority, IANA)建议的非强制端口号来区分不同的应用类型。随着 P2P和被动 FTP等新型网络应用的日益流行，数据传输中使用了大量的随机端口，使得这种方法不够准确[1]，但由于其简单和发展成熟，目前还未被完全淘汰。

基于特征字段分析的分类根据网络应用在传输过程中所具有的特征来区分不同的应用。它需要解析数据包并获得特征字段，准确性较高，但随着应用负载加密和新型应用的不断涌现，该方法的有效性逐步下降。目前 P2P的识别较多采用此方法。Cho等人[2]对未知端口的流量采用有效负载模式匹配的方法确定类型。此方法正确性较好，但在实时进程中受限。

这两种方法的比较见表 1，它们在目前网络中的实现较简单，但容易随着网络应用的改进而失效，并且分类结果易受网络环境变化的影响。

表 1　传统流分类方法比较

分类方法	优点	缺点
基于端口号匹配的分类方法	原理和实现较简单，可满足高速网络上的实时分类要求，可用硬件实现	容易受伪端口应用的干扰，只能识别已知端口的应用类别
基于特征字段分析的分类方法	对 P2P的识别准确性较高，可用于实时的流分类系统	需要深度解析包，开销大，涉及用户隐私，对载荷加密的流量不能分类

1　基于流统计特征的机器学习分类方法

由于传统的流量分类方法已不能很好地适应飞速发展的网络技术，许多研究人员根据网络流的一些属性的统计信息，将机器学习方法应用到流量分类领域，并取得了较好的结果。这种有效的流分类方法被称为基于流统计特征的机器学习分类方法。该方法的一般过程如图 1所示[3]，其关键步骤如下。

1)获取流量的特征属性最优组合集。流量属性集通过统计形成网络流的数据包的包头信息得到，包括包总数、字节总数、包长(平均值、标准差值)、时间间隔(平均值、标准差值)等。为了减少算法学习时所需要的数据量、缩短执行时间和提高分类正确率，需要从属性集合中选取对分类有效的

收稿日期：2010－04－21；修回日期：2010－07－01。　基金项目：国家自然科学基金资助项目(60872022)。
作者简介：胡婷(1986－)，女，陕西宝鸡人，硕士研究生，主要研究方向：网络安全；王勇(1964－)，男，四川阆中人，教授，博士，主要研究方向：计算机网络；陶晓玲(1977－)，女，浙江金华人，工程师，硕士，主要研究方向：网络信息安全、智能计算。

属性，作为最优属性集。常用的特征选择算法有 FCBF(Fast Correlation-Based Filter)、顺序前进法(Sequential Forward Selection, SFS)、相关性特征选择(Correlation Feature Selection, CFS)和遗传搜索算法等。

2)采用机器学习方法进行分类。已使用过的方法有K-近邻(K-Nearest Neighbor, K-NN)[4]、朴素贝叶斯方法(Naïve Bayes, NB)[5]、支持向量机(Support Vector Machine, SVM)[6]、C4.5[7]等。等。K-NN方法是最早引入网络流量分类中的机器学习方法，但该方法具有较大的计算开销，而且分类性能极易受到噪声数据的干扰。NB算法是 Moore等人对网络流量进行分类较早选用的机器学习，但此方法是一种基于概率的学习方法，过于依赖样本空间的分布，具有潜在的不稳定性。SVM和 C4.5可取得较高的分类准确率，但两种方法均是有监督的分类方法，必须事先标记流量的应用类型，因此不能适应完全意义上的实时分类。

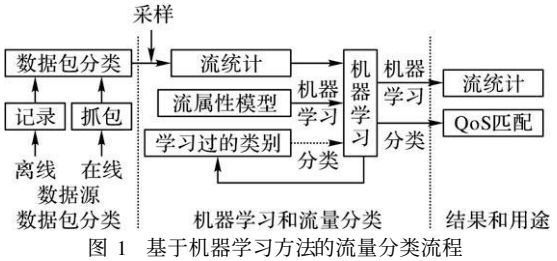


图 1 基于机器学习方法的流量分类流程

基于流统计特征的机器学习分类方法无需解析数据包，不受伪端口应用的影响；并且准确率高，能对多个应用类型分类，可识别新类型应用以及攻击流。但有些流量属性对网络动态变化较敏感，会对结果产生影响，分类算法可选范围广但实现较为复杂。

2 混合模式的流量分类方案

2.1 分类方案

在上述分析的基础上，本文采用简单易实现的基于端口号匹配方法，结合适应性强、准确度高的基于流特征的机器学习方法，提出一个混合模式的网络流量分类方案。图 2为方案流程。

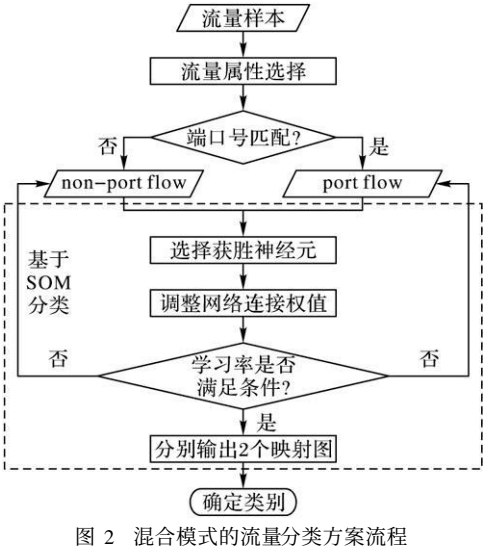


图 2 混合模式的流量分类方案流程

混合模式的流量分类方案具体实现过程如下。

1)对流量样本采用属性选择方法选出最优属性集，降低算法输入向量维数。

2)与常用协议的默认端口号(如 HTTP的默认端口号为 80, SMTP的默认端口号为 25)匹配，实现粗分。如果匹配成功，打“0”标签，归为常用端口流量样本集，记为 port flow；否则，打“1”标签，归为其他类别流量样本集，记为 non-port flow，该步减少了算法输入样本数，可提高算法处理速度。

3)分别采用基于自组织映射网络的分类方法进行细分。根据输出标签确定某一流量类别分布在 port flow映射图或 non-port flow映射图上。对分类结果采用手工处理方式，根据网格区域的颜色，结合训练样本，确定输出映射图中相应区域的流量类型。

2.2 自组织映射

神经网络对外界输入样本具有很强的识别与分类能力，可以很好地解决对非线性曲面的逼近，比传统的分类器具有更好的分类与识别能力，可适用于网络流量分类[8-10]。本文采用神经网络中的自组织映射[11](Self-Organizing Map, SOM)网络对流量按应用类型进行分类。

SOM网络是一种无人监督的竞争型神经网络。它通过对生物神经元的模拟，实现网络的自组织特性。它能够将高维的输入流量样本以拓扑有序的方式变换到二维的离散空间上，其输出分类结果可以直观的以棋盘状的二维平面阵显示。SOM网络可根据输入样本的内在联系，对样本进行自动聚类。因此，SOM可以通过对流量分类，识别新应用类型的流量样本。

设输入样本 $X=(x_1, x_2, \cdots, x_n)^T$，权向量为 $W_j=(w_1, w_2, \cdots, w_n)^T$ $(j=1, 2, \cdots, m)$。其中，n为输入样本维数，m为映射图神经元个数。X和 W全部进行归一化处理后为 $\hat{X}$和 $\hat{W}_j$。SOM网络学习过程主要有两个关键步骤。

1)选择获胜神经元。

$$d_{j^*}=\min_{j\in\{1,2,\cdots,m\}}\{\|\hat{X}-\hat{W}_j\|\} \quad (1)$$

2)调整获胜节点及邻域内的所有节点的网络权值。

$$W_{j^*}(t+1)=\hat{W}_{j^*}(t)+\eta(t)N(t)(\hat{X}-\hat{W}_{j^*}(t)) \quad (2)$$

其中：t为学习次数，$\eta(t)$为学习率，$N(t)$为优胜邻域。

采用 SOM方法对网络流量分类的过程为：将所有权值 w初始化为在$[-1, 1]$区间的随机数，随机选择一个流量样本 n，将其特征属性送入 SOM网的输入端 $x_1 \sim x_8$，设置初始邻域 $N(0)$和初始学习率 $\eta(0)$。输出层各神经元通过式(1)找出对应权向量与当前流量样本 n的属性最接近的优胜神经元 j^*。按式(2)，对 j^*及其邻域内的所有神经元向当前流量样本 n方向调整权值，然后缩小邻域 $N(t)$，减小学习率 $\eta(t)$，重新调整邻域内神经元的权值直到学习率衰减为 0。此时，若流量样本不空，继续随机选择剩余流量样本并重复以上迭代步骤。流量样本集为空则完成训练，此时输出一个流量类别映射图，根据样本激活神经元的位置可判断流量类别。

3 实验及结果分析

3.1 实验数据

本文所用数据集来源于 Moore等人在文献[5]中所用的实验数据集，简称 Moore_Set。此实验数据集是 Moore等人为进行流量分类实验，收集自某研究机构连接到 Internet的主机。此数据集中的每一条流按照流的五元组定义，且仅包括 TCP流，流特征信息来源于捕获到的包头部信息，共有 248个属性特征。表 2列出了数据集的统计信息。本文所用样本集是由 Moore_Set中按比例取出各种流量组成，共 3 780个样本。

3.2 特征选择

特征属性选择算法采用基于关联的快速过滤机制 FCBF算法，从所有流属性特征中选择最有代表性的属性特征形成

特征子集。FCBF算法从流的 248条属性中选出 8个特征属性,构成特征子集。

数据集中的流属性均是通过数据包的头部信息分析、计算得到,选中的特征属性描述为[13]:端口号、带 ACK位的包数(服务器端到客户端)、带 PUSH位的包数(服务器端到客户端)、MSS(客户端到服务器端)、提交到初始窗口的总字节数(服务器端到客户端)、丢失数据长度(客户端到服务器端)、重传最大个数和控制字节最小值。

表 2　数据集统计信息

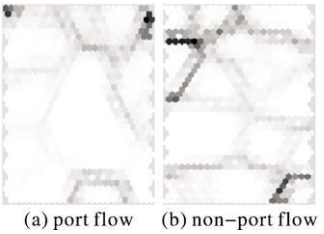
流类型	应用	流的数量	精度/%
WWW	HTTP, HTTPS	328 091	86.910
MAIL	IMAP, POP2/3, SMTP	28 567	7.567
BULK	FTP	11 539	3.056
DB	postgres, sqlnet, Oracle, ingres	2648	0.701
SERV	X11, DNS, IDENT, LDAP	2099	0.556
P2P	KaZaA, BitTorrent	2094	0.555
ATT	Internet worm and virus attacks	1793	0.475
MULT	WindowsMedia Player, Real	1152	0.305
总计	—	377 408	—

3.3　参数设置

经多次实验,SOM网络学习阶段,初始化参数的较优取值如下:1)输出映射图采用二维平面阵,神经元个数首选值设为 200,根据输入网络的样本数不同自适应调整网络结构;2)排序阶段,学习的最大次数为 1 200,学习率的初始值为 0.9,最终值为 0.05,邻域的初始值为 8,最终值为 1;3)收敛阶段,学习的最大次数为 2 000,学习率的初始值为 0.05,最终值为 0.01,邻域的初始值为 0,最终值为 0。实验所使用的数据分析工具为 Matlab 7.8.0。

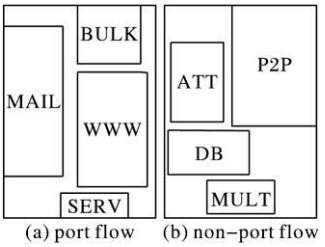
3.4　实验结果

混合模式的分类方法训练后输出映射图可通过 U矩阵直观反映,如图 3所示。在 U矩阵图中,两种类别间的分界面在 U矩阵中对应的邻节点距离较大,颜色较深。观察图 3可发现图(a)、(b)中均明显分为四个区域,即常用端口流量样本可进一步细分为四类,其他类别流量样本也可分为四类。再用手工确定映射图中各区域代表的流量类别,如图 4所示(图 3中流量区域是不规则的,在图 4中为了表示方便,近似表示)。手工确定方法为:获得流量类别映射图后,输入已知类型的流量样本,根据激活神经元的位置,可确定映射图中某一区域所代表的流量类别。



(a) port flow　(b) non-port flow

图 3　分类方法输出的类别映射图



(a) port flow　(b) non-port flow

图 4　二维 SOM网络表示的流量类别分布图

映射图中激活区域的大小和此类别流量样本数目及流量样本统计特征分布有关。WWW类和 MAIL类的样本数较大,因此激活的神经元范围较大;P2P流量发展迅速,不断出现新的应用类型,其统计特征标准偏差大;ATT类主要是攻击流,由于攻击对象不同,统计特征变化大,因此映射图中所占面积较大。此外,若分类前流量样本中类别数不确定,也可从输出的映射图中确定类别数,从而识别出未知流量。

由图 3、4可知,此分类方法可以实现对流量按照应用类型进行分类,并且可视化效果好。经多次实验证明,此方法的分类准确性高于 90%。并且由于该方法是无监督分类方法,可用于实时分类中。

4　结语

本文通过分析基于端口号匹配、特征字段分析和基于流统计特征的机器学习三种分类方法,提出了一个混合模式流量分类方案。此方法集合了不同算法的优点,既避免了端口号匹配方法准确率低的问题,又提高了对大样本、高维的流量采用机器学习算法的分类速度,因此适用于网络流量分类。

目前,基于流统计特征的机器学习的方法在网络流量分类中的应用还处在发展阶段,但由于该方法适应性强,可扩展性好,并且机器学习方法种类较多,因此具有一定的发展和应用。虽然传统的流量分类方法单独使用不能很好地满足实际应用,但其还有可适用的环境,可以在今后的研究中与基于机器学习的方法相组合,实现在高速网络中实时、有效的分类。

参考文献:

[1] MOORE A W, PAPAGIANNAKI K. Toward the accurate identification of network applications[C]// PAM 2005: Proceedings of the 6th International Workshop on Passive and Active Network Measurement. Berlin: Springer-Verlag, 2005: 41-54.

[2] CHOI K, CHOI K J. Pattern matching of packet payload for network traffic classification[C]// COIN-NGNCON 2006: The Joint International Conference on Optical Internet and Next Generation Network. Washington, DC: IEEE, 2006: 130-132.

[3] ZANDER S, NGUYEN T, ARMITAGE G. Automated traffic classification and application identification using machine learning[C]// LCN 2005: Proceedings of the IEEE Conference on Local Computer Networks, 30th Anniversary. Washington, DC: IEEE Computer Society, 2005: 250-257.

[4] ROUGHAN M, SEN S, SPATSCHECK O, et al. Class-of-service mapping for QoS: A statistical signature-based approach to IP traffic classification[C]// IMC'04: Proceedings of the ACM SIGCOMM Internet Measurement Conference. New York: ACM, 2004: 135-148.

[5] MOORE A W, ZUEV D. Internet traffic classification using Bayesian analysis techniques[C]// Proceedings of the 2005 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems. New York: ACM Press, 2005: 50-60.

[6] 徐鹏, 刘琼, 林森. 基于支持向量机的 Internet流量分类研究[J]. 计算机研究与发展, 2009, 46(3): 407-414.

[7] 徐鹏, 林森. 基于 C4.5决策树的流量分类方法[J]. 软件学报, 2009, 10(20): 2692-2704.

[8] TEUFL P, PAYER U, AMLINGER M, et al. InFoCT-Network traffic classification[C]// Proceedings of the seventh International Conference on Networking. Washington, DC: IEEE Computer Society, 2008: 439-444.

[9] KIZILOREN T, GERMEN E. Network traffic classification with self organizing maps[C]// ISCIS 2007: Proceedings of 22nd International Symposium on Computer and Information Sciences. Washington, DC: IEEE, 2007: 1-5.

[10] 王琳. 面向高速网络的智能化应用分类的研究[D]. 济南:济南大学, 2008.

[11] KOHONEN T. The self-organizing maps[J]. Proceedings of the IEEE, 1990, 78(9): 1464-148

[12] MOORE A W, ZUEV D. Internet traffic classification using Bayesian analysis techniques[C]// Proceedings of the 2005 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems. New York: ACM, 2005: 50-60.

[13] MOORE A W, ZUEV D, CROGAN M L. Discriminators for use in flow-based classification. RR-05-13[R/OL]. [2009-12-08]. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.101.7450&rep=rep1&type=pdf