

# 基于 Netflow 的流量分类方法研究

钱亚冠

(浙江科技学院 理学院,杭州 310023)

**摘 要:** 针对 Netflow 提供的流量信息有限的问题,在 Netflow 的基本信息基础上构建更丰富的特征空间,通过机器学习方法(决策树、朴素 Bayes 方法和 Bayes 网络)研究了 Netflow 用于流量分类的可行性。实验结果表明,决策树方法在 Netflow 数据上具有良好的分类效果;同时结合 Netflow 的广泛性,提出的方法具有良好的实用意义和推广价值。

**关键词:** Netflow;机器学习;流量分类

中图分类号: TN915.04

文献标志码: A

文章编号: 1671-8798(2014)05-0339-06

## Traffic classification based on netflow

QIAN Yaguan

(School of Sciences, Zhejiang University of Science and Technology, Hangzhou 310023, China)

**Abstract:** Due to the limited traffic information provided by Netflow, it is not considered as a suitable data set for traffic classification traditionally. We construct a richer feature space based on Netflow, and use machine learning methods (the decision tree, Navie Bayes and Bayes network) to explore the traffic classification. The experimental results show that the decision tree built on Netflow dataset has better precision than other two methods, and reinforce our suggestion that Netflow is fully appropriate for classification.

**Key words:** Netflow; machine learning; traffic classification

随着互联网应用的不断增多与传输带宽的持续增加,使得互联网变得更加复杂,于是对互联网管理提出了更高的要求。因此,需要更加有效的网络管理工具实现对应用流量的监控,而流量分类则是其中的核心技术。精确识别流量的应用类型,对实现分类计费、流量工程、容量规划等管理具有十分重要的意义。

基于 TCP 端口号的传统分类方法在 P2P 应用出现后受到了严峻的挑战。P2P 应用采用随机端口号的方法,甚至采用 http 协议的 80 端口躲避端口号的检测。而深度包检测(deep packet inspection, DPI)技术又遇到数据加密的难题。为了克服上述困难,近几年的研究工作开始转向流量的统计特征的研究。

---

收稿日期: 2014-05-09

基金项目: 浙江省网络媒体云处理与分析工程技术中心开放课题(2012E10023-14)

作者简介: 钱亚冠(1976—),男,浙江省嵊州人,副教授,博士,主要从事互联网流量建模、流量分类、流量异常检测、机器学习与大数据处理等研究。

究<sup>[1-2]</sup>,以期发现具体应用的特定流量模式<sup>[3-5]</sup>,从而确定应用类型。

目前,这类基于统计特征的方法通常需要很多的统计变量,有的甚至达到数百个<sup>[6]</sup>。对于实时性要求很高的网络管理任务来说,这类复杂的计算模型往往会严重影响管理效率。如何在保持较高的分类正确率的情况下获得精简的特征空间?这个问题启发人们研究是否可以利用 Netflow 信息进行流量分类<sup>[7]</sup>。笔者发现,思科的 Netflow 目前已得到广泛的部署,并已成为 IETF(internet engineering task force)的标准。Netflow 在数据流(flow)级别上实现了信息的汇集,包括源/目的 IP 地址、源/目的端口、字节总数、数据包总数等。由于 Netflow 中有关流量的信息有限,因此研究人员一直认为 Netflow 无法为分类提供足够的特征空间。而笔者的研究表明,利用 Netflow 进行流量分类具有 3 个优势:一是 Netflow 已被广泛部署在思科的路由器设备上,因此,采集数据变得非常方便,而不需要专门的流量采集设备;二是 Netflow 已经将数据包级的信息汇聚成了流级信息,可以免去大量的数据预处理工作;三是 Netflow 尽管提供的信息有限,但研究表明它完全可以支撑分类工作,并且可以满足实时性的要求。本研究正是基于上述认识,利用机器学习的方法展开对 Netflow 数据的分类研究。

## 1 相关工作

近几年,机器学习(machine learning, ML)方法开始被应用于流量分类领域,以便克服基于端口的方法及 DPI 方法的缺陷。机器学习是通过人工智能的学习理论,从大量的数据中获取知识,建立相应的分类模型,从而使模型具有对未知数据的预测(分类)能力。在流量分类中,利用已经获取的大量流量数据,通过机器学习,使得模型具有对未知流量的识别能力。目前,机器学习主要有基于监督的和无监督的学习方法 2 类。本研究采用基于监督的学习方法,即事先需要对训练数据进行分类标识,以便指导模型的建立。

目前,已有相关工作利用有监督的机器学习方法应用于互联网流量分类<sup>[8-15]</sup>,但这些工作均对数据包形式的流量进行处理,需要大量的模型训练时间,因此,很难真正部署到营运网络中。文献[6,16]等提出基于数据流(flow)的特征进行分类研究。数据流的特征包括流的持续时间、流的字节数、流的数据包数、流内的包到达间隔等。通过将数据包的信息进一步汇聚到数据流级别,可以显著减少数据量,从而有效地减少机器学习的模型训练时间。但是,目前数据流级别的分类方法采用的特征数仍然很多,文献[16]提出了 248 个可用的流特征,显著地增加了模型建立的复杂性。由此启发人们思考是否可以采用较少的特征来实现流级别的分类。最近研究发现 Netflow 具有流量特征空间简单,又与当前网络管理兼容的优点,非常适合营运网络的流量分类。据已有资料,目前还没有在 Netflow 上进行有效的工作。

## 2 基于 Netflow 的流量特征

Netflow 是思科公司为了收集网络流量信息而设计开发的一种网络协议,目前,它已成为 IETF 标准。Netflow 将具有相同五元组(源 IP 地址,目的 IP 地址,源端口,目的端口,协议号)的数据包归为同一数据流。Netflow 的基本工作原理是:利用标准的交换模式处理数据流的第一个 IP 包数据,生成 Netflow 缓存;随后,同样的数据基于缓存信息在同一个数据流中进行传输,不再匹配相关的访问控制等策略,Netflow 缓存收集随后数据流的统计信息。

支持 Netflow 协议的路由器或交换机可以收集自身所有端口的流量统计信息,并以 Netflow 记录的形式发送给服务器,用以分析处理。

Netflow 已经发展到第 10 版,但目前应用最广泛的是第 5 版,该版本被限制于 IPv4 的流量。考虑到目前流量仍然以 IPv4 为主,本研究采用第 5 版的 Netflow 数据,使用的 Netflow 信息见表 1。

从表 1 可以看出,Netflow 记录中可用于

表 1 Netflow 记录中可用于分类的信息

Table 1 Information of Netflow record used in classification

字节数	标记	解释
0~3	srcaddr	源 IP 地址
4~7	dstaddr	目的 IP 地址
16~19	dPkts	数据流中的数据包总数
20~23	dOctets	数据流中的 3 层数据包字节总数
24~27	first	数据流的起始时间
28~31	last	数据流的结束时间
32~33	srcport	TCP/UDP 源端口号
34~35	dstport	TCP/UDP 目的端口号

分类的信息非常少,正因为信息有限,所以没有引起研究者对 Netflow 在流量分类中的重视,甚至从根本上否定了它的意义。但笔者发现,从上述的基本信息中可以进一步推导出新的特征信息,如平均字节速率(B/s),平均数据包速率(Packets/s),平均数据包长度等,从而大大丰富了流量特征空间。

### 3 基于机器学习的流量分类方法

机器学习是研究计算机怎样模拟或实现人类的学习行为,以获取新的知识或技能,重新组织已有的知识结构使之不断改善自身的性能。因此,将机器学习中的有监督方法应用于流量分类中,可望获得良好的分类效果。有监督学习是指从给定的训练数据集中学习出一个函数,当新的数据到来时,可以根据这个函数预测结果。有监督学习的训练集需要事先标注好分类标签,用以指导机器学习。本研究采用朴素 Bayes 方法、Bayes 网络和决策树算法对 Netflow 流量数据进行分类研究。

#### 3.1 朴素 Bayes 方法

朴素 Bayes 方法源于概率论中的著名 Bayes 公式:

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \quad (1)$$

式(1)中: $H$ —假设; $X$ —证据; $P(H|X)$ —后验概率; $P(H)$ —先验概率。

朴素 Bayes 分类方法分类原理:

1)假设  $D$  是用于训练的 Netflow 流量数据集合, $X$  是训练集合的实例, $X = \{x_1, x_2, \dots, x_n\}$ ,也称为一个特征向量,其中  $x_n$  为分类标签。

2)又假设有  $m$  个流量分类,如 P2P,http 等,标记为  $C_1, C_2, \dots, C_m$ 。给定一个数据  $h$  流实例  $X$ ,预测具有最大后验概率的类,即预测  $X$  属于类  $C_i$  当且仅当

$$P(C_i|X) > P(C_j|X), \quad 1 \leq j \leq m, \quad j \neq i。$$

Bayes 公式中的先验概率容易从已获取的训练集合中估算得到: $P(C_i) = |C_{i,D}|/|D|$ ,其中  $|C_{i,D}|$  为  $D$  中  $C_i$  类的实例数。而计算  $P(X|C_i)$  的计算量巨大,需要枚举整个特征空间,因此引入实例特征之间关于类  $C_i$  条件独立的假设,使得后验概率的计算变为简单的乘积运算: $P(X|C_i) = \prod_{k=1}^n P(x_k|C_i)$ ,从而大大降低了计算复杂度。基于这种朴素的假设,此方法称为朴素 Bayes 方法。

#### 3.2 Bayes 网络

朴素 Bayes 方法假定特征之间可以有条件的独立,用于简化计算。当该假设成立时,朴素 Bayes 方法可获得很好的分类精度。但在实践中,特征之间往往可能存在依赖关系。Bayes 网络为克服这一不足,允许在特征子集之间定义条件独立性,并提供一种因果关系的图模型来进行学习(图 1)。

Bayes 网络由一个有向无环图和条件概率表构成。网络中的每个节点表示一个随机变量,可以是连续或离散值。每条有向弧表示一个概率依赖,连接的节点分别称为双亲和后代。每个变量关联着一个条件概率表, $P(Y|\text{parents}(Y))$ ,其中  $\text{parents}(Y)$  是  $Y$  的双亲。设变量  $X = \{x_1, x_2, \dots, x_n\}$ ,每个变量有

条件的独立于网络中的非后代,可得它的联合概率:

$$P(x_1, x_2, \dots, x_n) = \prod_{i=1}^n P(x_i | \text{parents}(Y_i)) \quad (2)$$

式(2)中: $P(x_1, x_2, \dots, x_n)$ — $X$  的某个特征组合的概率。

#### 3.3 决策树方法

决策树是一种基于判定的树结构,树中的每个分支节点表示在一个特征上的测试判定,而每个分支

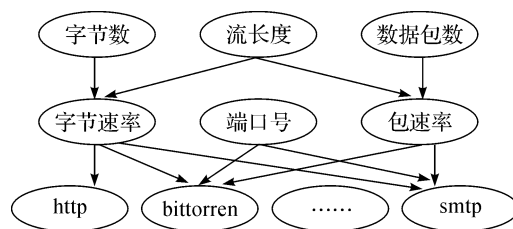


图 1 流量分类的 Bayes 网络

Fig. 1 Bayes networks applied in traffic classification

则表示一个测试判定的结果输出。每个叶节点则表示最终的输出,即分类标签。决策树从提出开始,已经产生了 3 种经典的算法:ID3,C4.5 和 CART,这些算法均采用贪心策略,自顶向下递归构造一棵决策树。

算法的核心思想是通过某种特征选择度量(如信息增益),选择“最佳”特征,将训练集合  $D$  分裂,每个特征值将产生一个分裂子集  $D_i$ 。递归地选择剩余候选特征中的“最佳”特征,继续将分裂子集  $D_i$  进行分裂,直到获得一个分类标号均相同(或占绝对优势)的子集。不同的决策树算法之间的差别在于创建树时的特征选择度量和剪枝策略。一旦一棵决策树从训练集合中构造成功,它就可以用来对未知实例进行预测分类。该过程非常直观和高效,从决策树的根节点出发,自顶向下沿着某个路径上的特征进行测试,直到到达叶节点(分类标签)。

#### 4 Netflow 数据集

从浙江大学校园网中心的某台路由器上获得了 Netflow 数据,共计 37 583 条数据流,并利用 DPI 工具 L7Filter 对数据流的应用类型进行了标识。共标识了 7 种应用类型:http,bittorrent,ssl,pop3,edonkey,skype 和 smtp。各种应用的数据流比例如表 2 所示。从表 2 可以看出,http 流量在字节总数上占绝对优势,这主要由于目前视频共享应用利用 http 协议传输短视频内容。基于 P2P 技术的 bittorrent 居第二大流量主体,尽管只有 4.99%,但每个数据流的平均字节总量却非常大,远超过 http 流量。

由表 3 可以明显发现,bittorrent 和 edonkey 这 2 种 P2P 应用每个流产生的字节流量最大,具有大象流(elephant flow)的特征。从网络管理的角度看,这种大象流对资源的占用很大,因此,识别该类流量具有十分重要的意义。

表 3 各种应用类型在数据集中的总字节数与数据流平均字节数的对比

Table 3 Comparison of total bytes and mean bytes of each application in traffic dataset

应用类型	总字节数/B	数据流中的平均字节数/B
http	943 808 458	31 152
bittorrent	159 925 669	85 248
ssl	43 097 597	11 401
pop3	1 745 128	3 317
edonkey	14 511 805	122 981
skype	961 572	22 362
smtp	491 956	2 827

#### 5 研究方法与实验结果

采用朴素 Bayes 方法、Bayes 网络和决策树算法对 Netflow 数据进行了实验研究,具体研究方案如下:

根据数据流数量的递增次序,分别设定 6 个训练数据集:数量从 3 000、5 000 递增到 21 000,集合内容上前者分别是后者的子集,呈包含关系,余下 16 000 个数据流作为测试集合。分别在 6 个训练集上用朴素 Bayes、Bayes 网络和决策树 C4.5 算法训练模型,并用同一测试集测试,分别获得图 2 中 3 种方法的分类精度比较结果。

表 2 各种应用类型在数据集中的比重(以字节计算)

Table 2 Percentage of each application in traffic dataset (in bytes)

应用类型	比例/%	应用类型	比例/%
http	80.61	edonkey	0.31
bittorrent	4.99	skype	0.11
ssl	10.06	smtp	0.46
pop3	1.40	others	2.06

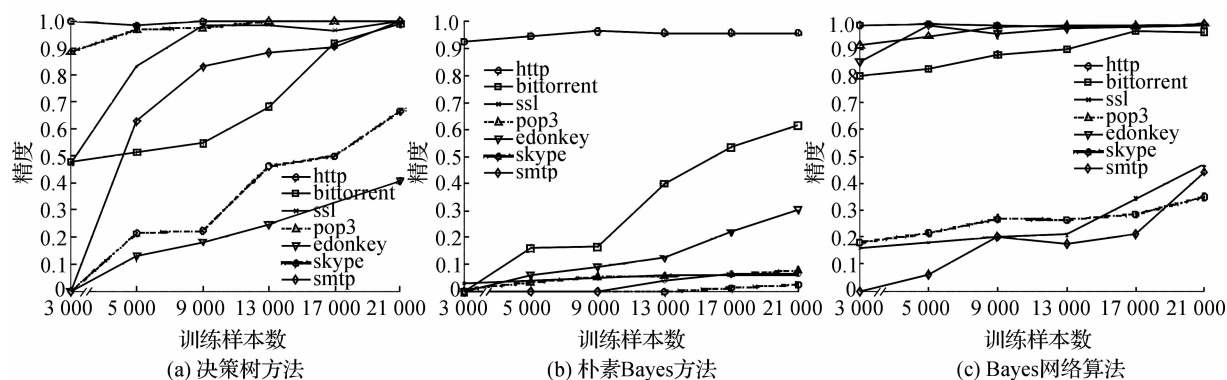


图2 3种不同机器学习方法的分类精度比较

Fig. 2 Comparison of precision among three machine learning methods

从图2(a)中可以发现,随着训练集合的增大,决策树方法的分类精度逐步提高。http, pop3 和 ssl 的分类精度在训练集超过 9 000 条记录后,提高不再明显,但均已超过 95% 的正确率。smtp 与 bittorrent 随着训练集的增大,分类精度提升迅速,在训练集合达到 21 000 条记录时已超过 98% 的准确率。skype 与 edonkey 虽然随着训练集的增大,精度也得到提高,但提高速度不大。在 21 000 条训练记录时,skype 接近 70%,而 edonkey 才达到 40% 的正确率。

图2(b)显示了朴素 Bayes 方法在不同训练集上的分类精度。从中可以看出,朴素 Bayes 方法对 http 应用的分类非常有效,只需 3 000 条 Netflow 记录就可以实现大于 90% 的正确率。但对于其余应用的分类效果明显不足,尤其对于 ssl, smtp 和 skype,其分类精度随着训练集的增大几乎没有提升。而 bittorrent 与 edonkey 虽有提升,但提升速度缓慢。

图2(c)显示的是 Bayes 网络的分类效果。可以明显发现,对于 http, bittorrent, pop3 和 edonkey 这 4 类应用, Bayes 网络可以在较小的训练集上达到大于 90% 的分类精度。与决策树相比,在训练集容量达到 21 000 条记录时, http, bittorrent, pop3 这 3 类应用的分类精度均可在 95% 以上,但 edonkey 在决策树下分类效率明显不及 Bayes 网络。可见, Bayes 网络对于 P2P 应用(bittorrent, edonkey)的区分能力优于决策树方法。在 ssl, smtp 和 skype 应用上, Bayes 网络的分类能力却不及决策树方法。

综上所述,决策树方法尽管在小的训练集下分类效率不及 Bayes 网络,但从图2(a)中可以发现随着训练集合容量的增大,各种应用的分类精度呈现不断上升的趋势。而 Bayes 网络在 skype, smtp 和 ssl 应用上的提升趋势却不是十分显著。可见,决策树方法在 3 种方法中具有较好的优势。

除了从分类精度上对上述 3 种方法进行了比较外,还从模型的训练时间上进行了对比(图3)。从图3中可以发现,朴素 Bayes 方法的模型训练时间是最短的,在训练集合增大到 21 000 条记录时,训练时间仍未超过 0.5 s。决策树方法与 Bayes 网络的训练时间基本接近,且与训练集合的容量成线性增长关系,即算法的时间复杂度为  $O(n)$ 。因此,从可计算理论的角度看,决策树和 Bayes 网络的算法复杂度是比较好的。

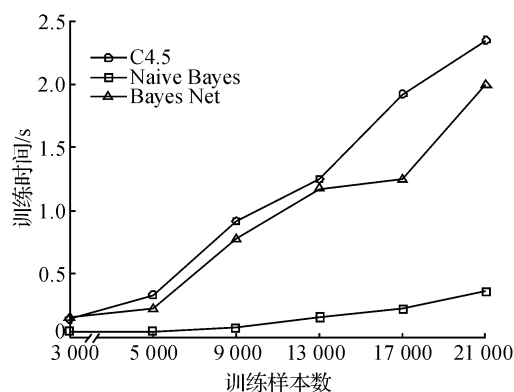


图3 3种机器学习方法在不同训练集下的模型建立时间

Fig. 3 Time taken to build models with three machine learning methods

## 6 结 语

从 Netflow 数据出发,利用朴素 Bayes 方法、Bayes 网络和决策树算法 3 种机器学习方法对 Netflow 数据中的应用类型进行了分类。实验仿真结果表明,这 3 种方法中决策树方法和 Bayes 网络具有较好的

分类性能。在有足够的训练实例下,各种应用在决策树方法中可达到理想的分类准确率。本研究的工作充分证明了 Netflow 数据应用于流量分类的可行性,从而改变了以往认为 Netflow 数据不适合流量分类的观点。在 Netflow 的基础上进行流量分类具有良好的实用性,与现有设备可保持良好的兼容性,因此,非常具有实际推广意义。

#### 参考文献:

- [1] Bernaille L, Teixeira R, Salamatian K. Early application identification[C]//Proceedings of the 2006 ACM CoNEXT conference. New York: ACM,2006:6.
- [2] Kim H, Claffy K C, Fomenkov M, et al. Internet traffic classification demystified: myths, caveats, and the best practices[C]//Proceedings of the 2008 ACM CoNEXT conference. New York: ACM,2008:11.
- [3] Iliofotou M, Kim H, Faloutsos M, et al. Graph-based P2P traffic classification at the internet backbone[C]. INFOCOM Workshops 2009, IEEE. Riode Janeiro: IEEE,2009:1-6.
- [4] Karagiannis T, Papagiannaki K, Faloutsos M. BLINC: multilevel traffic classification in the dark[J]. ACM SIGCOMM Computer Communication Review,2005,35(4):229-240.
- [5] Valenti S, Rossi D, Meo M, et al. Accurate, fine-grained classification of P2P-TV applications by simply counting packets[M]//Traffic Monitoring and Analysis. Papadopoulou M, Owezarski P, Pras A. Berlin: Springer,2009: 84-92.
- [6] Moore A W, Zuev D, Crogan M L. Discriminators for use in flow-based classification[EB/OL]. (2012-10-09)[2014-03-10]. <http://www.cl.cam.ac.uk/~awm22/publications/RR-05-13.pdf>.
- [7] Claise B. Cisco Systems NetFlow Services Export Version9; RFC 3954 (Informational)[EB/OL]. (2004-10-01)[2014-03-10]. <http://tools.ietf.org/html/rfc3954.html>.
- [8] Auld T, Moore A W, Gull S F. Bayesian neural networks for internet traffic classification[J]. IEEE Transactions on Neural Networks,2007,18(1):223-239.
- [9] Crotti M, Dusi M, Gringoli F, et al. Traffic classification through simple statistical fingerprinting[J]. ACM SIGCOMM Computer Communication Review,2007,37(1):5-16.
- [10] Haffner P, Sen S, Spatscheck O, et al. ACAS: automated construction of application signatures[C]//Proceedings of the 2005 ACM SIGCOMM workshop on mining network data. New York: ACM,2005:197-202.
- [11] Jiang H, Moore A W, Ge Z, et al. Lightweight application classification for network management[C]//Proceedings of the 2007 SIGCOMM workshop on Internet network management. New York: ACM,2007:299-304.
- [12] Moore A W, Zuev D. Internet traffic classification using bayesian analysis techniques[C]//ACM SIGMETRICS Performance Evaluation Review. New York: ACM,2005,33(1):50-60.
- [13] Roughan M, Sen S, Spatscheck O, et al. Class-of-service mapping for QoS: a statistical signature-based approach to IP traffic classification[C]//Proceedings of the 4th ACM SIGCOMM conference on Internet measurement. New York: ACM,2004:135-148.
- [14] Zuev D, Moore A W. Traffic classification using a statistical approach[M]//Passive and Active Network Measurement. Berlin: Springer,2005:321-324.
- [15] Szabó G, Szabó I, Orincsay D. Accurate traffic classification[C]//World of Wireless, Mobile and Multimedia Networks, 2007. Espoo: IEEE,2007:1-8.
- [16] Erman J, Mahanti A, Arlitt M, et al. Identifying and discriminating between web and peer-to-peer traffic in the network core[C]//Proceedings of the 16th international conference on World Wide Web. New York: ACM,2007: 883-892.