

基于支持向量机的流量分类方法^{*}

林 森^{a, b}, 徐 鹏^{a, b}, 刘 琼^{a, b}

(中国科学院 a. 软件研究所; b. 研究生院, 北京 100190)

摘 要: 针对现有流量分类方法存在的准确率低、应用范围受限、计算复杂度高等问题, 提出使用支持向量机方法来解决流量分类问题。使用公开的人工标注数据集作为训练集和测试集, 通过有监督学习构建支持向量机流量分类器。此外, 通过实验进一步分析了训练集大小、核函数、惩罚因子等因素对支持向量机分类性能的影响。实验结果表明支持向量机分类器可以达到 98% 以上的流分类准确率。

关键词: 流量分类; 支持向量机; 流量识别

中图分类号: TP393 文献标志码: A 文章编号: 1001-3695(2008)08-2488-03

Traffic classification based on support vector machine

LIN Sen^{a, b}, XU Peng^{a, b}, LIU Qiong^{a, b}

(a. Institute of Software, b. Graduate School, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: In order to solve the problems in current work, such as low accuracy, limited application region or high computation complexity, support vector machine (SVM) was applied to categorize traffic by application. The work capitalized on public hand-classified network dataset and used it to train and tested the supervised SVM traffic classifier. The improved accuracy of refined variants of this classifier was further illustrated, and the variants included the size of training dataset, kernel functions and penalty factors. The results indicate that it can achieve over 98% accuracy on per-flow classification with the SVM classifier.

Key words: traffic classification; support vector machine(SVM) ; traffic identification

1 相关研究

本文讨论的流量分类问题是指按应用类型将网络流量分类, 是网络管理的基础功能。准确的流量分类对于网络领域很多相关研究都很重要。例如入侵检测、流量调度、服务质量保证(QoS)、构建符合实际的流量模型、准确预测未来流量规模和需求等。目前, 已有的流量分类方法主要分为以下四类:

a) 基于固定端口号的分类方法。根据不同应用使用不同的传输层端口号来划分流量, 通常只有 50% ~70% 的准确率^[1, 2]。这种方法仅对使用传统固定端口的应用有效, 如 Web、DNS、mail 等。但是这类方法无法分类使用随机端口的应用, 或者错误分类使用传统固定端口的非传统应用。例如, 近几年来涌现出的 P2P 应用大量使用随机端口, 使基于固定端口流量分类方法失效。

b) 基于应用层内容的分类方法。根据不同应用具有不同的应用层特征字段来划分流量。该方法在加入人工干预的基础上能达到 99% 以上的准确率^[3]。其准确率虽然远远高于基于固定端口号的流量分类方法, 但是该方法需要获取完整的应用层负载内容, 其应用范围受到三个因素限制: (a) 涉嫌侵犯用户隐私; (b) 无法识别应用层加密的应用; (c) 无法识别特征字段未知的应用。

c) 基于传输层行为的分类方法。根据传输层不同级别的

行为特征来划分流量, 准确率约为 80% ~90%^[4]。该方法不依赖应用层内容, 不受基于应用层内容的分类方法三个限制因素的影响。但是传输层行为通常与网络环境密切相关, 相同应用在不同网络环境下的传输层行为很可能存在较大差异, 这种相关性限制了该方法的应用范围。

d) 基于统计学的分类方法。根据传输层的统计特征来划分流量。其中, 贝叶斯神经网络方法的准确率达到 99% 以上, 但该方法计算复杂度很高^[5]。目前关于统计学方法已有很多深入研究, 并且有许多成熟的算法实现。统计学方法被广泛应用于许多领域中的数据分析和处理, 取得很多成果。在网络流量分类领域, 这类方法的应用研究正在兴起。

基于统计学的分类方法可以有效克服前三种流量分类方法存在的问题, 因此成为当前流量分类的主要研究方向。特别是近两年来, 越来越多的机器学习方法被用于流量分类, 如贝叶斯算法^[6]、贝叶斯神经网络^[5]等, 都取得了很好的效果。尽管如此, 已有机器学习分类方法还是存在一些问题。例如贝叶斯算法^[6]依赖样本分布, 而且属性的相关性和冗余性会大幅度降低分类准确率; 贝叶斯神经网络^[5]同样依赖样本分布, 且存在过拟合、计算复杂度高等问题。SVM 方法有效地降低了样本分布、相关属性、冗余属性、过拟合以及计算复杂度高等因素的影响, 具有很好的泛化能力。本文提出利用支持向量机研究流量分类问题, 并在人工标注的公开数据集上进行模型训练和测试, 结果表明分类准确率能达到 98% 以上。

2 原理

2.1 简介

基于统计学习理论的支持向量机模型,自 20 世纪 90 年代中期提出以来,已经成为国际上机器学习领域的研究热点^[7]。支持向量机模型的基本思想主要有两个,即最大间隔原则和核技巧。最大间隔原则指通过寻求最优超平面,使产生的决策函数能够承受尽可能大的扰动;核技巧指通过核函数变换,将原输入空间需要用超平面划分的分类问题,转换为 Hilbert 空间中用超平面划分的分类问题,可以将非线性问题转换为线性问题,从而大大降低问题的复杂度。

支持向量机分类算法具有四个显著特点: a) 利用大间隔的思想降低分类器的 VC 维,实现结构风险最小化原则,控制分类器的推广能力; b) 利用 Mercer 核实现线性算法的非线性化; c) 稀疏性,即少量样本(支持向量)的系数不为零(就推广性而言,较少的支持向量数在统计意义上对应好的推广能力;从计算角度看,支持向量减少了核形式判别式的计算量); d) 算法设计成凸二次规划问题,避免了多解性^[8]。

2.2 支持向量机

假设训练样本集为

$$(x_1, y_1), (x_2, y_2), \dots, (x_l, y_l) \tag{1}$$

其中: $x_i \in R^n (i = 1, 2, \dots, l) (R \text{ 为实数域})$; 对于两类的分类问题, $y_i \in \{+1, -1\}$ 。

SVM 分类算法的原始形式可归结为下列二次规划问题:

$$\begin{aligned} \min & \frac{1}{2}(\omega \cdot \omega) + C \sum_{i=1}^l \xi_i \\ \text{s. t. } & y_i((\omega, x_i) + b) - 1 + \xi_i \geq 0 \end{aligned} \tag{2}$$

其中: $(\cdot, \cdot)$ 为两向量之间的内积; $\xi_i \geq 0$ 为松弛项,表示错分样本的惩罚程度; C 是惩罚因子,为常数,用于控制对错分样本惩罚的程度,实现在错分样本数与模型复杂性之间的折中; ω 和 b 为判决函数 $f(x) = (\omega \cdot x) + b$ 中的权向量和阈值。当无错分样本时,最小化目标函数的第一项等价于最大化两类间的间隔,可降低分类器的 VC 维,实现结构风险最小化原则。

上述二次规划的对偶形式为

$$\begin{aligned} \max & \sum_{i=1}^l \alpha_i - \frac{1}{2} \sum_{i,j=1}^l \alpha_i \alpha_j y_i y_j (x_i, x_j) \\ \text{s. t. } & \sum_{i=1}^l \alpha_i y_i = 0; 0 \leq \alpha_i \leq C; i = 1, 2, \dots, l \end{aligned} \tag{3}$$

其中: α_i 为 Lagrange 乘子。根据最优化理论中的 KKT 条件,只有少量样本(判决函数值等于 ± 1 的样本和错分样本)的 α_i 值不为零。Vapnik 等人称之为支持向量。

由于对偶式(3)中只出现两向量间的内积运算,Vapnik 等人用满足 Mercer 条件的核函数 $k(x_i, x_j)$ 来代替内积运算 (x_i, x_j) ,实现线性算法的非线性化。常用的核函数包括多项式核、径向基核以及二层神经网络。核形式的判别函数为

$$f(x) = \sum_{i=1}^l \alpha_i y_i k(x_i, x) + b \tag{4}$$

上面讨论是两类的分类问题,对于多类分类问题通常有如下几种方法:一类对余类、成对分类、纠错输出编码方法、确定多类目标函数方法等^[9]。

3 实验设置

3.1 数据集

本实验使用一套通过实际监测采集的数据集^[10]。该数据

集监测了一个名为 Genome Campus 的研究机构,大约有 1 000 名研究人员、管理人员和技术人员。监测点位于该机构的网络出口,监测持续 24 h 的双向流量。原始数据集较大,因此最终使用的数据集是对原始数据集的抽样。表 1 展示了该数据集中各类流的数目和百分比。

表 1 数据集统计信息		
类别	流数目	百分比 / %
WWW	328 091	86.91
MAIL	28 567	7.567
BULK	11 539	3.056
SERV	2 099	0.556
DB	2 648	0.701
INT	110	0.029
P2P	2 094	0.555
ATTACK	1 793	0.475
MMEDIA	1 152	0.305
GAMES	8	0.002
合计	377 526	100

3.2 分类对象、属性和类别

分类对象是完整的 TCP 双向流。每条流用源 IP、源端口、目的 IP、目的端口四元组惟一标志。每条流由 249 个属性表示,其中第 249 号属性表示类别。参考文献[11]列出了各属性的定义和完整说明。需要说明的是,这 249 个属性中前两个属性分别表示源端口号、目的端口号,为了避免对应用端口信息的依赖,在实验中并未使用这两个属性。表 2 列出了分类类别及其代表应用。

表 2 类别及其部分代表应用	
类别	应用举例
BULK	ftp
DB	postgres, sqlnet, oracle, ingres
INT	ssh, klogin, rlogin, telnet
MAIL	imap, pop2/3, smtp
SERV	X11, dns, ident, ldap, ntp
WWW	www
P2P	KaZaA, BitTorrent, GnuTella
ATTACK	Internet worm and virus attacks
GAMES	Half-Life
MMEDIA	Windows Media Player, Real

3.3 实验过程

3.3.1 数据集预处理

原始数据集中每个属性的取值范围不同。为了赋予每个属性同等的重要性,笔者通过变换将每个属性的取值范围映射到 $[-1, 1]$ 。本文采用线性变换映射,也可以尝试其他的映射方式。经过线性形映射后的数据集简称为 trace。

3.3.2 参数选择

支持向量机的参数选择不同取值,对最终模型的分类准确率有显著影响,所以在实验之前,先在较小的训练集上通过实验得到较好的参数取值,然后在较大的数据集上验证,通过验证的参数取值被用于后续实验。本文采用 LibSVM^[12]中的支持向量机分类算法 C-SVC。其中参数主要包括:

- a) C : 错分类惩罚因子,见式(2)(3),需要实验调整
- b) 核类型: 默认使用径向基核
 - (a) 线性核 $\mu \cdot v$
 - (b) 多项式核 $(\gamma^* (\mu \cdot v) + \text{coef})^{\text{degree}}$
 - (c) 径向基核 $e^{(-\gamma^* \|\mu - v\|^2)}$
- c) 核函数中的参数
 - (a) γ : 需要实验调整

- (b) coef: 默认值取 0
- (c) degree: 默认值取 3
- d) 停机条件, 默认值取 0.001

经过实验得, 当 $C = 512$, $\gamma = 0.031\ 25$ 时, 分类准确率较高; 后续实验中如未加特殊说明, 将 $C = 512$, $\gamma = 0.031\ 25$ 作为默认值使用。

3.3.3 实验

将数据集 trace 随机平均分为两个子集(各占 50%), 分别记为 trace1、trace2。Trace1 作为训练集, trace2 作为测试集。为了保证数据集中每类至少包含一个样例, 采用按类别比例随机抽样, 抽样后的数据集 trace1、trace2 与原数据集 trace 具有近似相同的类别比例。用训练集 trace1 训练分类器, 测试集 trace2 测试分类器性能。为了使实验结果更可靠, 重复以上抽样、训练、测试过程五次, 实验结果取平均值。

3.4 实验结果

本文使用评价标准定义如下:

定义 1 准确率表示分类的准确性。其中平均准确率定义为正确分类的流的个数与流的总个数之比; 某类的准确率定义为该类流中正确分类的个数与该类流的总个数之比。

表 3 表明支持向量机的平均准确率为 98.50%, 比改进的贝叶斯算法^[6] 74.80% 的准确率高, 比贝叶斯神经网络^[5] 99.26% 的准确率略低, 但是训练时间远小于贝叶斯神经网络。当训练集为 50% 的 trace 时, 即训练集大小为 188 763, 支持向量机的训练时间约为 2 h, 贝叶斯神经网络的训练时间约为 39 h。支持向量机方法平均准确率标准差为 0.01%, 远小于另外两种方法, 说明该方法有很好的稳定性。

表 3 不同算法分类准确率、训练时间比较		
算 法	准 确 率 / %	训 练 时 间
改进的贝叶斯算法	74.80 ±1.7	2 h 30 min
贝叶斯神经网络	99.26 ±0.4	38 h 48 min
支持向量机	98.50 ±0.01	2 h

4 讨论

本章将讨论训练集大小、核函数、惩罚因子等因素对支持向量机分类性能的影响。在本章的实验中如果没有特别说明, 实验的参数值使用 3.3.2 节中的默认值。

4.1 不同大小的训练集

支持向量机适合于处理小样本情况下的学习问题, 在训练集较小的情况下也能取得不错的性能。表 4 表明训练集大小为 191 时, 准确率超过 90%; 训练集大小为 1 894 时, 准确率接近 95%。说明支持向量机使用很小的训练集, 就能达到很高的准确率。增大训练集, 准确率会相应提高。使用较小训练集训练, 训练时间较短, 支持向量个数也较少, 可以减少计算量, 提高分类速度。

表 4 不同大小训练集的实验结果				
训练集大小	测试集大小	准 确 率 / %	训 练 时 间 / s	支持向量个数
191	188 157	90.33	< 1	91
1 894	188 157	94.91	4	331
18 936	188 157	96.85	106	1 884
94 684	188 157	97.44	1 800	7 153

4.2 不同的核函数

支持向量机的分类准确率很大程度上取决于属性空间通过核函数变换后线性可分的程度, 选取适当的核函数可以提

高准确率。另外核函数的复杂度会影响支持向量机的推广能力。分别采用线性核、多项式核和径向基核(核函数的形式见 3.3.2), 比较不同核函数条件下的准确率、训练时间、迭代次数和支持向量个数。在实验中使用 20% 的数据作为训练集, 80% 的数据作为测试集。从表 5 列出的实验结果可知在三种核函数中, 分类准确率都在 97% 以上, 其中多项式核略高, 径向基核训练时间最短, 训练过程中迭代次数远小于其他两种核。对于支持向量个数, 多项式核 < 径向基核 < 线性核, 与准确率的高低顺序相反, 说明较少的支持向量有较好的泛化能力。

表 5 不同核函数的实验结果				
核函数	准 确 率 / %	训 练 时 间 / s	迭 代 次 数 / M	支持向量个数
线 性	97.03	4 770	26	6 339
多 项 式	97.99	2 293	12	4 392
径 向 基	97.37	1 078	0.15	5 806

4.3 不同的惩罚因子

进一步的实验为每个类施加不同的惩罚因子, 对训练集中样例较少的类别使用较大的惩罚因子可以间接增大小类别的权重, 从而减轻训练数据的不均衡性带来的不利影响。使用 20% 的数据作为训练集, 80% 的数据作为测试集, 并且使用表 6 中列出的为不同类别选配的三组惩罚因子进行实验。表 7 列出了使用不同惩罚因子训练模型的分类准确率。 C_2 与 C_1 相比使用较大的惩罚因子, 准确率从 97.82% 提高到 98.16%。 C_2 与 C_3 相比, C_2 对不同类别使用不同的惩罚因子, 即训练样例较少的类别使用较大的惩罚因子, 变相地改变各类别训练样例的权重, 使训练集中各类别的样例数更均衡, 准确率从 97.37% 提高到 98.16%。

表 6 不同的惩罚因子					
C	WWW	MAIL	BULK	SERV	DB
	INT	P2P	ATT	MULT	GAME
C_1	100	200	300	500	500
	1 000	500	500	500	1 000
C_2	512	1 024	1 536	2 560	2 560
	5 120	2 560	2 560	2 560	5 120
C_3	512	512	512	512	512
	512	512	512	512	512

表 7 不同惩罚因子的准确率	
惩 罚 因 子	准 确 率 / %
C_1	97.82 ±0.07
C_2	98.16 ±0.06
C_3	97.37 ±0.02

5 结束语

本文提出使用支持向量机研究流量分类问题。通过在手工标注的公开数据集上训练与测试, 评估和分析了支持向量机模型进行流量分类的准确率等性能。实验结果表明支持向量机方法的分类准确率可以达到 98% 以上。本文进一步讨论了训练集大小、核函数、惩罚因子等因素对支持向量机分类性能的影响, 结果表明支持向量机在较小的训练集上也能达到较高的准确率, 并且通过调整不同类别的惩罚因子可以降低训练集不均衡带来的不利影响。此外, 使用线性核、多项式核、径向基核等核函数均能得到较好的分类性能。在以后的工作可以考虑从下述几个方面展开进一步的研究:

- a) 为了满足特定情况下分类的实时性要求, 研究在满足一定准确率的条件下, 如何减少支持向量个数, (下转第 2498 页)

的 RTGM 表, 获取翻译组播组信息; 退出翻译组播组。

(4) 接收到组播包。遍历组播组的组成员列表, 将组播包转发给每一个成员。

2) MAP 上发生的操作

(1) 接收到 TransReq。查询 RTGM 表, 如果有匹配项, 那么, 广播翻译组播组; 否则, 加入真实组播组, 分配并广播翻译组播组。

(2) 接收到组播包。在 RTGM 表中查询组播包所属组播组对应的翻译组播组; 通过翻译组播组将组播包发送出去。

5 结束语

本文提出的 DFHMM 协议适用于实际的层次型移动 IPv6 网络。DFHMM 不仅对外屏蔽了节点的域内切换, 而且还利用快速切换机制使区域组播代理提前加入组播组, 从而避免了域外切换时由于组播树重构引起的组播服务中断。此外, 移动锚点的动态发现机制保障了区域组播代理的配置灵活性, 其冗余的配备方式则为组播通信提供了较高的可靠性。

通过仿真量化 DFHMM 的性能是下一步的主要工作。

参考文献:

[1] ERKINS C. RFC 3344, IP mobility support for IPv4 [EB/OL] . (2002-08-01) . http: // www. ietf. org/ rfc/ rfc 3344. txt.

[2] JOHNSON D, PERKINS C. RFC 3775, Mobility Support in IPv6 [EB/OL] . (2004-06-08) . http: // www. ietf. org/ rfc/ rfc 3775. txt.

[3] DEERING S, CHERITON D. Multicast routing in datagram internet-works and extended LANs[J] . ACM Trans on Computer Sys-tems, 1990, 8(2) : 85-110.

[4] SOLIMAN H, CASTELLUCCIA C, MALKI K E L, *et al.* RFC 4140, Hierarchical mobile IPv6 mobility management (HMIPv6) [EB/OL] . (2005-08-01) . http: // www. ietf. org/ rfc/ rfc4140. txt.

[5] KOODLI R. RFC 4068, Fast handovers for mobile IPv6[EB/OL] .

(2005-07-01) . http: // www. ietf. org/ rfc/ rfc 4068. txt.

[6] 吴茜, 吴建平, 徐恪, 等. 移动 Internet 中的 IP 组播研究综述[J] . 软件学报, 2003, 14(7) : 1324-1337.

[7] ROMDHANI I, KELLIL M, LACH H Y. IP mobile multicast: chal-lenges and solutions[J] . IEEE Communication Surveys, 2004, 6(1) : 18-41.

[8] CHIKARMANE V, WILLIAMSON C L, BUNT R B, *et al.* Multicast support for mobile hosts using mobile IP: design issues and proposed architecture (MoM protocol) [J] . Mobile Networks and Applica-tions, 1998, 3(4) : 365-379.

[9] LIN C R, WANG K M. Mobile multicast support in IP networks [C] // Proc of IEEE INFOCOM. Tel Aviv: IEEE, 2000: 1664-1672.

[10] WANG Y, CHEN W. Supporting IP multicast for mobile hosts[J] . Mobile Networks and Applications, 2001, 6(1) : 57-66.

[11] TAN C L, PINK S. MobiCast: a multicast scheme for wireless net-works[J] . Mobile Networks and Applications, 2000, 5(4) : 259-271.

[12] SUH Y J, SHIN H S, KWON D H. An efficient multicast routing protocol in wireless mobile networks [J] . Wireless Networks, 2001, 7(5) : 443-453.

[13] JELGER C, NOEL T. Multicast for mobile hosts in IP networks: pro-gress and challenges[J] . IEEE Wireless Communications, 2002, 9(5) : 58-64.

[14] YE M H, YANG L Y, LIU Y, *et al.* The implementation of multicast in mobile IP[C] // Proc of IEEE Wireless Communications and Net-works. New Orleans: IEEE, 2003: 1796-1800.

[15] MCAULEY A, BOMMAIAH E, MISRA A, *et al.* Mobile multicast proxy[C] // Proc of IEEE MILCOM. Atlantic City: IEEE, 1999: 631-635.

[16] JUNG H Y, SOLIMAN H, KOH S J, *et al.* Internet draft fast hand-over for hierarchical MIPv6 (F-HMIPv6) [EB/OL] . (2005-10-01) . ht-tp: // draft-jung-mipshop-fhmipv6 -00. txt.

(上接第 2490 页) 以此减少计算量, 提高分类速度。

b) 深入研究不同属性对不同类别间的区分能力, 考虑对不同类别的区分使用不同的属性集。

c) 对目前新出现并且发展迅速的新应用的分类, 特别是 P2P 应用作更深入的研究。

参考文献:

[1] OORE D, KEYS K, KOGA R, *et al.* The CoralReef software suite as a tool for system and network administrators[C] // Proc of the 15th USENIX Conference on System Administration. California: USENIX Association, 2001: 133-144.

[2] LOGG C, COTTRELL L. Characterization of the traffic between SLAC and the Internet[EB/OL] . (2003-07-15) . http: // www. slac. stan-ford. edu/ comp/ net/ slac-netflow/html/ SLAC- netflow. html.

[3] MOORE A W, PAPAGIANNAKI D. Toward the accurate identifica-tion of network applications[C] // Proc of the 6th Passive and Active Measurement Workshop. Berlin: Springer, 2005: 41-54.

[4] KARAGIANNIS T, PAPAGIANNAKI K, FALOUTSOS M. BLINC: multilevel traffic classification in the dark[C] // Proc of ACM SIG-COMM Computer Communication. New York: ACM Press, 2005: 229-240.

[5] AULD T, MOORE A, GULL S. Bayesian neural networks for Internet traffic classification[J] . IEEE Trans on Neural Networks, 2007, 18(1) : 223-239.

[6] MOORE A W, ZUEV D. Internet traffic classification using Bayesian analysis techniques[C] // Proc of ACM SIGMETRICS. New York: ACM Press, 2005: 50-60.

[7] 张学工. 关于统计学习理论与支持向量机[J] . 自动化学报, 2000, 26(1) : 32-42.

[8] 许建华, 张学工, 李衍达. 支持向量机的新发展[J] . 控制与决策, 2004, 19(5) : 481-485.

[9] 邓乃扬, 田英杰. 数据挖掘中的新方法: 支持向量机[M] . 北京: 科学出版社, 2004.

[10] MOORE A W, HALL J, KREIBICH C, *et al.* Architecture of a net-work monitor[C] // Proc of the 4th Passive and Active Measurement Workshop. Berlin: Springer, 2003: 257-267.

[11] MOORE A W, ZUEV D. Discriminators for use in flow-based classifi-cation[R] . Cambridge: Intel Research, 2005.

[12] CHANG C C, LIN C J. LIBSVM: a library for support vector machines [EB/OL] . (2007-04-10) [2007-05-20] . http: // www. csie. ntu. edu. tw/-cjlin/libsvm.