doi:10.3772/j.issn.1002-0470.2016.02.002

基于特征加权的朴素贝叶斯流量分类方法研究①

张泽鑫②*** 李 俊* 常向青*

(*中国科学院计算机网络信息中心 北京 100190)

(**中国科学院大学 北京 100049)

摘 要 研究了被广泛应用于互联网流量分类的朴素贝叶斯分类方法的性能特点,针对此方法在给定类别下给出的所有流量特征同等重要并且是独立的假设在现实中难以满足,致使分类准确率不高的问题,提出一种基于特征加权的朴素贝叶斯流量分类算法。该算法基于 NetFlow 记录的特征信息,采用特征选择算法 ReliefF 和相关系数方法计算每个特征的权重值,然后将网络流量分配至后验概率最大的应用类别中。实验结果表明,这种基于特征加权的朴素贝叶斯算法具有超过 94% 的分类准确率,并且维持了朴素贝叶斯方法简单高效、分类稳定的特性,可以满足当前高带宽网络流量分类的需求。

关键词 流量分类(TC), ReliefF, 相关系数, 特征加权(AW), 朴素贝叶斯(NB), NetFlow

0 引 言

随着互联网的迅猛发展,越来越多的新型网络应用不断兴起,网络规模不断扩大,网络组成也越来越复杂。音频、视频等实时应用的兴起,更是从根本上改变了人们对网络的使用方式,网络复杂性上升的同时,网络的异构性也愈来愈强,各种新的应用和未知协议使网络日益复杂、多样化和难以管理。进行网络管理离不开互联网流量分类(traffic classification, TC),互联网流量分类是认识、管理和优化各种网络资源的重要依据。网络服务提供商(ISP)通过对网络流进行分类,获悉各类网络应用所占的比例,预测网络业务的发展趋势,对不同的网络业务实行差异化收费标准。流量分类在网络安全性检测方面也发挥着巨大的作用,可以实现更精确的异常检测,入侵检测等。因此,开展互联网流量分类研究具有重要实际意义和应用价值。

由于很多新的网络应用采用动态端口、数据载荷字段加密,导致传统的基于端口和基于有效载荷分析的流量分类方法变得越来越受限制,分类准确率下降。因此,基于流量行为特征,采用机器学习的方法处理流量分类问题逐渐成为国内外学者研究的热点。朴素贝叶斯(Naive Bayes, NB)分类方法因其实现简单、处理高效的特征被很多学者用于流量分类领域[1,2]。然而,朴素贝叶斯(NB)方法在估计类条件概率时,假设流量特征之间是同等重要且条件独立的,该假设在实际情况中很难满足,流量特征之间往往存在着相关性。解决该问题的一个方法是特征过滤,从流量特征集中删除冗余的特征,使用过滤后的特征子集进行分类模型训练。该方法提高了朴素贝叶斯进行流量分类的准确率,但是,并没有考虑不同流量特征对分类的重要性不同。因此,本文提出了一种特征加权(attribute weighting, AW)的朴素贝叶斯流量分类方法,该方法赋予每个特征一个权重,越重要的特征权重值越大,然后将流量分至后

① 973 计划(2012CB315803)和中国科学院计算机网络信息中心"一三五"计划(CNIC_PY-1401)资助项目。

② 女,1987 年生,博士生;研究方向:流量测量与特性分析,流量分类等;联系人,E-mail: zhangzexin@cstnet.cn

(收稿日期:2015-08-24)

验概率最大的应用类别中。特征加权是特征选择的一种扩展方法,赋予冗余度高的特征较小的权重,赋予对网络应用区分度大的特征较高的权重,既能解决流量特征之间存在冗余的问题,也考虑了不同流量特征对分类重要性不同。

1 相关研究

早期,网络应用通过周知的端口来进行分类,根据互联网编号分配机构(IANA)预定义和分配的端口映射表[3],每个端口号对应一个应用,比如众所周知的 web 应用的端口号是 80。然而,随着网络应用的层出不穷,大量的随机端口被用于数据通信,有些协议甚至封装后通过周知的端口进行通信。Moore 和 Papagiannaki[4] 通过实验研究发现使用 IANA 列表进行基于端口号的流量识别,准确率不超过 70%。基于端口的应用识别方法变得越来越受限[5-8]。

为了提高分类的准确率,很多学者开始关注网络流量的负载特征。剑桥大学的 Moore 等人[4]、AT&T 实验室的 Sen 等[9]在他们的论文中都提出了采用基于有效载荷特征匹配的方法来对互联网的业务流量进行分类。基于载荷的流量分类方法准确度非常高,但是需要检索每个数据包中的 Payload 字段,需要的计算资源非常大,并且有效载荷的分析侵犯了用户的隐私和安全性,其发展受到了很大的阻力。

随着研究的不断深入,基于行为特征的流量分类方法逐渐成为国内外研究的热点。该方法从不同的观测角度发现网络应用的不同行为特征,例如包大小、包时间间隔、字节数、持续时间等作为流量特征,应用机器学习方法对其建立相应的模型,然后应用于分类。基于流量的统计行为特征,各研究机构提出了基于机器学习的流量分类方法[10],通过统计流持续时间、分组到达间隔等流统计特性,采用有监督或者无监督的机器学习方法实现业务分类。McGregor 等[11] 采用基于 EM(expectation-maximization)算法无监督的学习方法对基于连接层统计特征的"流"进行分类。文献[12]提出将 P 个特征属性转换为 P 维向量,通过定义距离函数,采用聚类的方法对网络流进行分类。无监督的聚类方法分类精度不高,在类别数比较多的情况下,分类复杂度较大。同时有监督的机器学习方法也被用于流量分类中,Moore 等[1] 通过傅里叶变换构建了 248 个流特征,并采用 FCBF(fast correlation-based filter)进行特征选择,最后采用朴素贝叶斯方法来区分各应用。该方法的流量特征过多,增加了时间和空间复杂度,无法应用于流量的在线识别。朴素贝叶斯分类模型具有简单高效的特点,因此被很多研究者应用于流量分类中。文献[13]采用多种特征选择方法,结合不同的训练数据集预处理方法,评估朴素贝叶斯分类器在检测网络异常时的性能差异。文献[14]提出了一种新的特征选择方法对流量特征过滤,该方法能有效缓解多类不平衡的问题,最后采用朴素贝叶斯分类器对流量进行分类。Antonio 等[15] 为了改进朴素贝叶斯分类器的分类准确率,采用主成分分析法和关联特征法对网络流量特征进行了选择,结果显示这两种特征选择方法都提高了朴素贝叶斯的分类准确率。文献[16,17]也分别使用朴素贝叶斯分类算法对网络流量进行分类。

国内方面,徐鹏等人[18] 采用 C4.5 决策树方法对流量进行分类,实验证明决策树分类方法在处理动态变化的样本和大规模流量分类时具有较好的性能,然而采用的流量属性还无法实时获取,不适合用于网络流在线识别。陈亮等人[19] 首先采用简单相关系数方法进行流量特征选择,然后基于 Bayes 判别法进行流量分类。该方法虽然整体的分类准确率较高,但是有些应用的分类结果却不佳,例如 P2P、ATTACK 等。

2 特征加权的朴素贝叶斯流量分类方法

2.1 特征加权的朴素贝叶斯流量分类算法

朴素贝叶斯分类器具有稳定的分类效率,对缺失数据也不敏感,并且算法简单,因此,被广泛地应用在分类领域。朴素贝叶斯分类器利用贝叶斯定理计算待分类实例的最大后验概率,在估计类条件概率时假设属性之间条件独立,形式化描述为

$$P(A \mid y) = \prod_{k=1}^{d} P(A_k \mid Y = y) \quad (1)$$

其中 A 代表属性集 $A = \{A_1, A_2, \cdots, A_d\}$,包含 d 个属性。

在条件独立假设下,只需对给定的 Y,计算每一个 A_i 的条件概率。假设有 N 条网络流 $X = \{X_1, X_2, \cdots, X_n\}$,每条网络流 X_i 由 d 个属性值描述 $\{A_1, A_2, \cdots, A_d\}$,有 m 个网络应用类别,$Y = \{Y_1, Y_2, \cdots, Y_m\}$。对于网络流 X_i,属于类别 Y_j 的概率为:

$$P(Y_j \mid X_i) = \frac{P(X_i \mid Y_j)P(Y_j)}{P(X_i)} \quad (2)$$

其中,先验概率 $P(Y_j)$ 代表网络应用类别 Y_j 在整个网络流中占有的比例,$P(X_i \mid Y_j)$ 为类条件概率,表示在应用类别为 Y_j 时,X_i 出现的概率。朴素贝叶斯分类器的目标是找出使得 $P(X_i \mid Y_j)P(Y_j)$ $(j = 1, 2\cdots, m)$ 最大的类 Y_j,此时,流量 X_i 对应的类别即为 Y_j。根据式(1)、(2),朴素贝叶斯分类器的后验概率计算公式为

$$P(Y_j \mid X_i) = \frac{P(Y_j)\prod_{k=1}^{d} P(A_k \mid Y_j)}{P(X_i)} \quad (3)$$

朴素贝叶斯分类器对网络流量进行分类时,假定每条流的特征是相互独立的,并且每个特征对分类的贡献度都是一样的。然而,在真实的网络环境中,这些假设条件都是难以满足的,结果导致分类准确率降低。针对这个问题,文献[1]中采用 FCBF 方法进行流量特征选择,选择一个与类别相关性高,特征之间冗余度低的特征子集进行分类模型训练,削弱了流量特征之间的相关性,提高了分类的准确率。特征加权是另一种可以保留或删除特征的方法,特征越重要,赋予的权值越大,而不太重要的特征赋予较小的权值。特征加权是特征选择的普适化方法,特征选择是特征加权方法的一个特例,权值等于 0 或 1。特征加权的朴素贝叶斯分类器后验概率计算公式为

$$P(Y_j \mid X_i) = \frac{P(Y_j)\prod_{k=1}^{d} P(A_k \mid Y_j)^{\omega_k}}{P(X_i)} \quad (4)$$

其中 $\omega_k \in R^+$,代表特征的重要程度。

2.2 基于 ReliefF 和相关系数的流量特征加权算法

特征加权的朴素贝叶斯流量分类方法的关键是特征权值的计算,给每个流量特征分配合适的权重不仅能够区分特征之间的预测能力,也能降低违背特征独立性假设所带来的影响,提高朴素贝叶斯分类器的分类准确率。因此,本文从两个方面考虑流量特征权重的计算,首先,认为与类别相关性高的特征具有较高的权重,采用 ReliefF[20] 方法计算每个特征的权重值;其次,认为与其它特征冗余度高的特征具有较低的权重,采用相关系数方法修正每个特征的权重值。

ReliefF 是由 Kononenko 提出的一种多类别特征选择算法,其基本思想是给特征集中的每个特征赋予一个权重值,赋予和类别相关性高的特征较高的权重,最后根据这些权重值进行特征子集选择。本文则根据这些权重值对流量特征进行加权,权重向量为 $\omega = (\omega_1, \omega_2, \cdots, \omega_d)$,算法见表 1。Class$(X_i)$ 表示样本 X_i 所属的类别,函数 $diff(A, I_1, I_2)$ 计算样本 I_1 和样本 I_2 在特征 A 上的距离,$P(Y)$ 表示网络应用 Y 的先验概率。

表 1 基于 ReliefF 的流量特征权重计算算法

输入:网络流量训练集数据 X,样本抽样次数 m,最近邻样本个数 l

输出:流量特征集的权重向量 ω

1. 将权重向量置初值 $\omega = 0$
2. For i: = 1 to m do
3. 随机地选择一个网络流 X_i;
4. 从 X_i 的同类别流中,选出 l 条最邻近的流 H_s;
5. For 每个类 $Y \neq$ class(X_i) do
6. 从类 Y 中找出 l 条最邻近的流 M_s;
7. For K: = 1 to d do
8. $$\omega_k := \omega_k - \sum_{s=1}^{l} diff(A_k, X_i, H_s)/(m \cdot l) + \sum_{Y \neq class(X_i)} \left[\frac{P(Y)}{1 - P(class(X_i))} \sum_{s=1}^{l} diff(A_k, X_i, M_s)\right]/(m \cdot l)$$
9. End;

ReliefF 算法没有限定特征权值的取值范围,有可能为负值,所以,为了避免发生这种情况,根据 Han 等人[21] 提出的 min-max 方法对权值进行标准化操作。假设流量特征的权值向量为 $[\omega_1, \omega_2, \cdots, \omega_k]$,采用公式

$$\omega'_i = \frac{\omega_i - \min_\omega}{\max_\omega - \min_\omega}(new\_\max_\omega - new\_\min_\omega) + new\_\min_\omega \quad (5)$$

对权值标准化。其中，$i = 1,2,\cdots,k$，ω'_i 为特征 i 的权值 ω_i 变换后的新权重值，$\min_\omega$ 为原特征权值向量中的最小值，$\max_\omega$ 为原特征权值向量中的最大值，$new\_\max_\omega$ 为原特征权值向量标准化后的最大值，$new\_\min_\omega$ 为原特征权值向量标准化后的最小值。本文将标准化区间定义为[0,1]，即 $new\_\max_\omega = 1$，$new\_\min_\omega = 0$。

ReliefF 算法仅考虑了特征与类别之间的相关性程度，没有考虑特征之间的相关性，本文采用相关系数来度量特征之间的相关性程度，如下式所示：

$$\rho_{A_i,A_j} = \frac{\mathrm{cov}(A_i,A_j)}{\sqrt{\mathrm{var}(A_i)var(A_j)}} = \frac{\sum_{k=1}^{n}(A_{i,k} - \bar{A}_i)(A_{j,k} - \bar{A}_j)}{\sqrt{\sum_{k=1}^{n}(A_{i,k} - \bar{A}_i)^2(A_{j,k} - \bar{A}_j)^2}} \quad (6)$$

其中，A_i 和 A_j 为两个属性特征，$A_{i,k}$ 和 $A_{j,k}$ 为两个属性特征的观察值，$\bar{A}_i$ 和 $\bar{A}_j$ 为两个属性特征 n 个观察值的均值。$|\rho_{A_i,A_j}| \in [0,1]$，$\rho_{A_i,A_j}$ 值越大，表明属性特征之间的相关性越强，反之越弱。使用 ρ_{A_i,A_j} 值对特征权重 ω'_i、ω'_j 进行修正，如果 $\omega'_i > \omega'_j$，即特征 A_i 较特征 A_j 有较强的预测能力时，保持 ω'_i 不变，对 ω'_j 进行修正，$\omega'_j = \omega'^{1+P\cdot\rho_{A_i,A_j}}_j$，$\omega'_j \in [0,\omega'_j]$，$P$ 称为衰减系数，本文取 P 为 20。ρ_{A_i,A_j} 值小时，ω'_j 降低得少，ρ_{A_i,A_j} 值为 0 时，表明 A_i 和 A_j 之间不存在相关性，$\omega'_j = \omega'_j$。ρ_{A_i,A_j} 值大时，ω'_j 降低得多，ρ_{A_i,A_j} 值为 1 时，表明 A_i 和 A_j 之间的相关性最强，$\omega'_j \approx 0$。

3 实 验

为了验证特征加权的朴素贝叶斯流量分类算法的有效性，设置了两组对比实验。第一组采用标准的朴素贝叶斯分类方法进行流量分类；第二组首先采用文献[22]中提出的 FCBF 方法对流量特征过滤，然后使用过滤的训练数据集运行朴素贝叶斯分类器，构造分类模型。

3.1 实验数据

实验的数据集来自中国科技网(China Science and Technology Network，CSTNET)。为了验证不同时间段算法的性能，采集了 2014 年 12 月 16 日上午 10:00 ~ 11:00，下午 14:00 ~ 15:00，晚上 19:00 ~ 20:00 之间经过该出口的所有 Netflow 网络流量，分别为 CSTNET_1，CSTNET_2，CSTNET_3。经分析，TCP 流量仍为中国科技网的主要流量，所以本文仅对完整的 TCP 流进行分类。

为了避免用每种应用的训练样本过少而影响分类准确率，抽样时使得每种应用的流数基本保持一致，即采用均匀无放回抽样方法，从每个数据集中抽取约 3 万条数据，每种应用约 3000 条流数据作为样本集。经分析，表 2 中的 10 种应用为科技网 TCP 流量中的主要部分，所以本文选取这 10 种应用进行分类。每组数据集轮流作为训练集对分类器进行训练，剩余两组作为测试集。

表 2 数据集

应用名称	CSTNET_Set1 流数	CSTNET_Set2 流数	CSTNET_Set3 流数
Http	3218	2839	3198
Https	3025	2931	3141
Data-transfer	2851	3083	2532
SSH	2452	2497	2396
Jabber	3062	3093	2969
XunLei	3326	3068	3092
SMTP	2781	3004	2856
MSN	3031	2878	2851
IMAP	3173	3029	3016
BitTorrent	2391	3411	3067

文献[1]采用傅里叶变换得到 249 项网络流特征，在这些特征中存在着大量的冗余特征，对分类器进行训练时，不但增加了分类的时间复杂度，也降低了分类的准确率，并且不适合用于网络流的在线分类。为了降低分类系统采集报文、计算流量特征的开销，本文采用 NetFlow 的统计特征作为网络流的属性，见表 3。对于不同的网络应用，这些属性特征通常表现出较大的差异性，因此，可以利用 NetFlow 的流记录信息对网络应用进行识别。

表 3　网络流属性集合

属性	描述
bytes	流内字节数
packets	流内报文数
dest_port	目的端口
duration	流持续时间
IAT	流内平均报文间隔时间
packet_size	流内平均报文大小
pps	流内每秒平均传输包数
bps	流内每秒平均传输字节数

3.2　评估指标

评价一个网络流量分类器的好坏需要在一定指标下测试其分类结果,然后通过测试结果得出结论。分类器首先通过训练集进行模型的学习,拟合训练集数据中类标号和属性集之间的联系,建立分类模型。随后将该模型应用于测试集数据,分类模型的性能根据测试集数据的运行结果进行评估,运行结果计数存放在混淆矩阵中,如表 4。表中每项 f_{ij} 表示实际类标号为 i,但被预测为类 j 的记录数。

表 4　多类问题的混淆矩阵

		预测的类			
		类 = 1	类 = 2	…	类 = m
实际的类	类 = 1	f_{11}	f_{12}	…	f_{1m}
	类 = 2	f_{21}	f_{22}	…	f_{2m}
	⋮	⋮	⋮	…	⋮
	类 = m	f_{m1}	f_{m2}	…	f_{mm}

根据混淆矩阵,分类器性能评价指标计算公式如下:

(1) 准确率 $= \dfrac{\text{正确预测数}}{\text{预测总数}} = \dfrac{\sum_{i=1}^{m} f_{ii}}{\sum_{i=1}^{m}\sum_{j=1}^{m} f_{ij}}$,指分类器正确预测的样本占所有样本的比例。

(2) 类的精度 $= \dfrac{f_{ii}}{\sum_{j=1}^{m} f_{ji}}$,指被分类器分到某个类别,样本确实是这个类别所占的比例。

(3) 类的召回率 $= \dfrac{f_{ii}}{\sum_{j=1}^{m} f_{ij}}$,度量某个类别的样本被分类器正确分类的比例。

3.3　朴素贝叶斯网络流量分类

本节仅采用标准的朴素贝叶斯方法对网络流量进行分类,分别选取数据集中的一组数据作为训练集,另外两组数据作为测试集,每组实验重复 10 次,取每组实验的平均值作为分类结果。如表 5 所示,三组实验的平均分类准确率分别为80.7% ,81.0% ,81.2% 。从表 5 可以看出,朴素贝叶斯分类器的分类结果较差,整体的分类准确率较低,主要是由于朴素贝叶斯分类器在估计类条件概率时假设流量特征之间是条件独立的,流量特征之间的相关性违背了条件独立性假设,降低了朴素贝叶斯分类器的分类准确率。

表 5　基于朴素贝叶斯流量分类方法的整体分类准确率

训练集	测试集	分类准确率
CSTNET_1	CSTNET_2,CSTNET_3	80.7%
CSTNET_2	CSTNET_1,CSTNET_3	81.0%
CSTNET_3	CSTNET_1,CSTNET_2	81.2%

3.4　基于 FCBF 特征选择的朴素贝叶斯(FCBF_NB)网络流量分类

由于流量特征之间存在冗余,导致朴素贝叶斯分类器的分类准确率降低,所以采用文献[22]提出的 FCBF 方法对流量特征进行选择,选择与类别相关性高且特征之间冗余度低的流量特征用于构建朴素贝叶斯分类器模型。选出的特征集为 M = {dest_port, packet_size, duration, IAT, pps},然后在完成过滤的训练数据集上运行朴素贝叶斯分类器,构建分类模型,并使用数据集中的另外两组数据测试模型。每组实验重复 10 次,取每组实验的平均值为分类结果,如表 6 所示。采用 FCBF 方法对流量特征过滤之后,每组实验的整体分类准确率较特征过滤之前有了一定的提高,每组实验分别增长了3.6% ,3.7% ,3.2% ,但是整体的分类准确率仍然较低。主要是由于采用 FCBF 方法选出的特征子集,虽然使得特征之间的冗余度降低了,但是并不能完全消除,并且认为每个特征对分类是同等重要的,从而导致较低的分类准确率。

表 6 基于特征选择的朴素贝叶斯流量分类方法整体分类准确率

训练集	测试集	分类准确率
CSTNET_1	CSTNET_2,CSTNET_3	84.3%
CSTNET_2	CSTNET_1,CSTNET_3	84.7%
CSTNET_3	CSTNET_1,CSTNET_2	84.4%

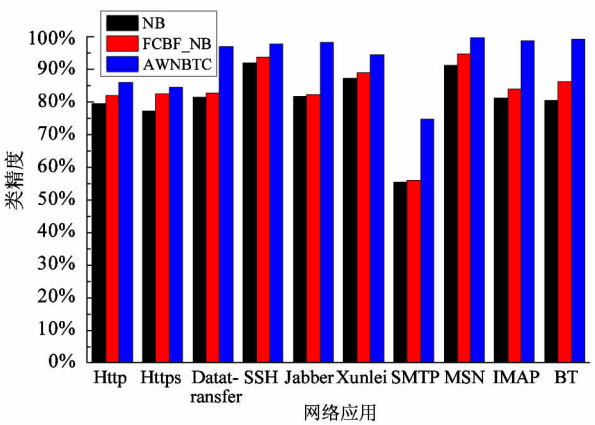
表 7 基于特征加权的朴素贝叶斯流量分类方法整体分类准确率

训练集	测试集	分类准确率
CSTNET_1	CSTNET_2,CSTNET_3	94.5%
CSTNET_2	CSTNET_1,CSTNET_3	94.2%
CSTNET_3	CSTNET_1,CSTNET_2	94.2%

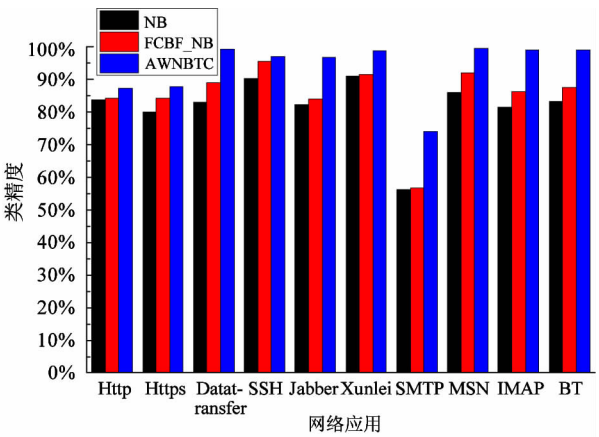
3.5 特征加权的朴素贝叶斯网络流量分类

考虑到不同的流量特征对分类的重要程度不同,越重要的特征赋予的权值越大,不太重要的特征赋予较小的权值。首先通过 ReliefF 和相关系数方法对流量特征计算权值。然后在训练数据集上运行特征加权的朴素贝叶斯分类器,构建分类模型,并使用数据集中的另外两组数据测试模型。每组实验重复 10 次,取每组实验的平均值为分类结果,如表 7 所示。较上述两种方法,特征加权(AW)的朴素贝叶斯(NB)流量分类(TC)算法(简称 AWNBTC 算法)的整体分类准确率有了显著提高,每组实验结果较朴素贝叶斯分类方法分别增长了13.8%,13.2%,13.0%,较基于 FCBT 特征选择的朴素贝叶斯(FCBF_NB)方法分别增长了10.2%,9.5%,9.8%。

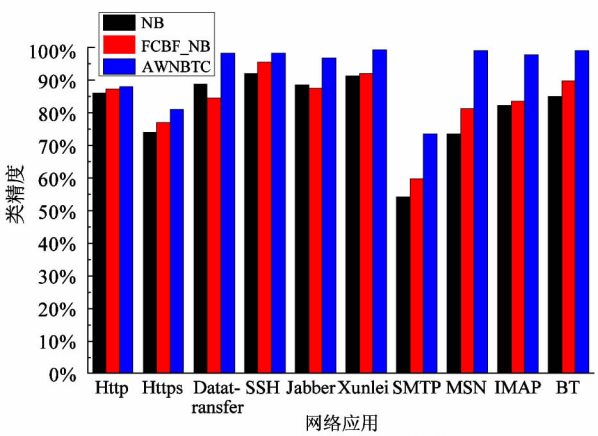
为进一步分析 AWNBTC 算法的分类准确性,通过图 1 和图 2 描述了每种网络应用的分类精度和召回率。从图中看出,较朴素贝叶斯(NB)方法和 FCBF_NB 方法,AWNBTC 算法每种应用的分类精度



(a) CSTNET_1 为训练集



(b) CSTNET_2 为训练集



(c) CSTNET_3 为训练集

图 1 网络应用类精度

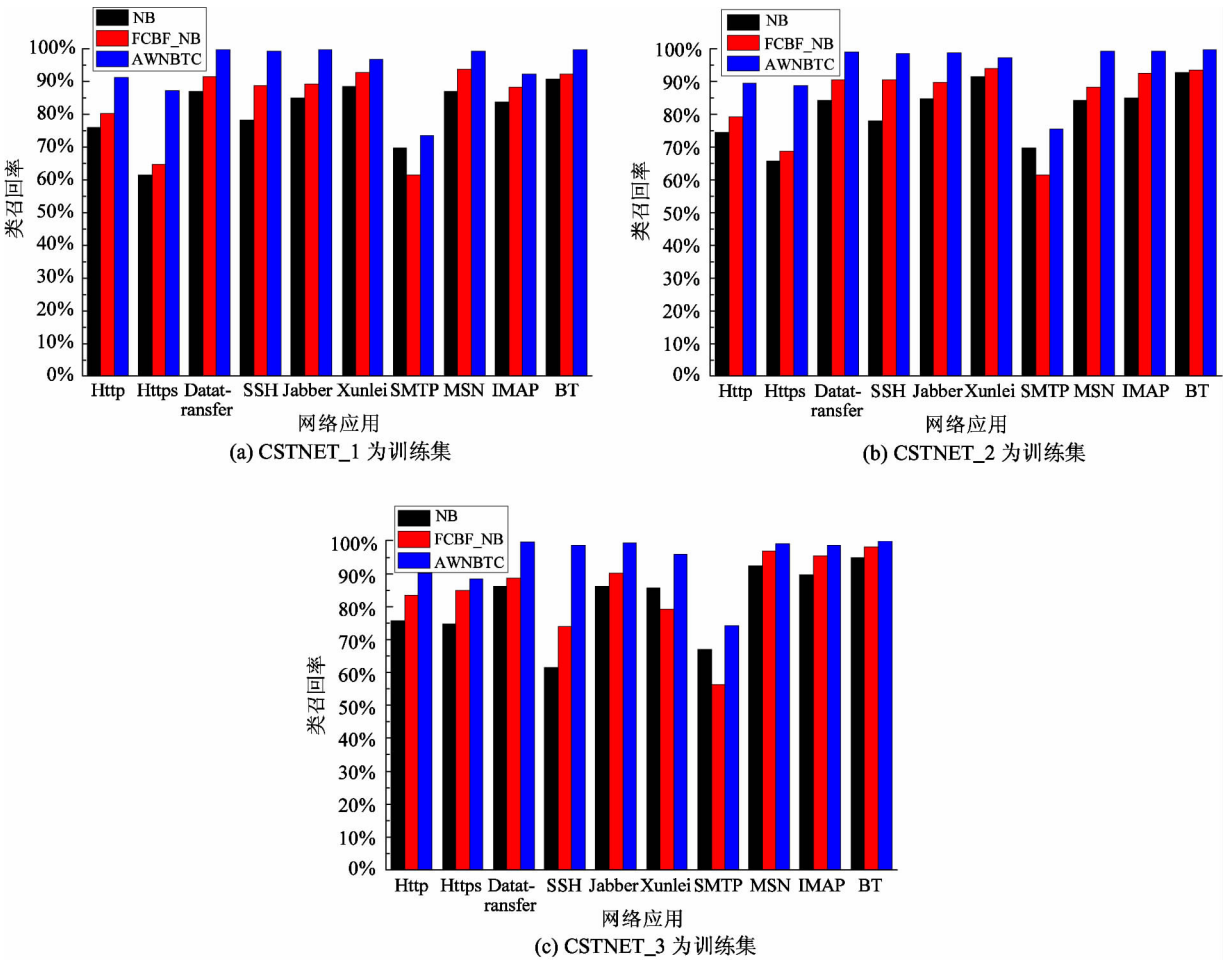


图 2　网络应用类召回率

和召回率都提高了,除了 Http、Https 和 SMTP,其他应用的分类精度和召回率超过了 96%,MSN 和 BT 的分类精度和召回率达到了 99%。SMTP 的分类精度改善得最多,较 NB 算法和 FCBF_NB 算法分别增长了18.7% 和16.6%,Datatransfer、Jabber、IMAP 和 BT 应用的分类精度较 NB 算法和 FCBF_NB 算法也平均提高了约14.9% 和12.6%。关于应用的召回率,Https 和 SSH 应用增长得最多,Https 分别提高了 20.8% 和 15.2%,SSH 则分别提高了 26.1% 和 14.3%。其他应用的召回率也有显著增长,Http、Datatransfer、Jabber 和 IMAP 较 NB 算法和 FCBF_NB 算法平均提高了约 13.2% 和 8.2%。由此可见,特征加权的朴素贝叶斯分类方法较 NB 和 FCBF_NB 方法,不仅具有较高的整体分类准确率,而且每类应用的分类准确性也很高。

4　算法评估

4.1　时效性

朴素贝叶斯分类器的实现主要包括模型训练和流量分类两部分,其中模型训练部分由流量特征离散化和朴素贝叶斯模型构建组成。首先采用基于等宽的离散化方法将各个流量特征的值域划分成具有相同宽度的区间,并用离散化后所在区间的标称值代替原来的值,其算法时间复杂度为 $O(N)$;模型构造主要为计算每个分类属性的类条件概率,时间复杂度为 $O(N)$,因此整个模型训练的时间复杂度为 $O(N)$,其中 N 为整个训练样本规模。流量分类部分首先对一条流的流量特征进行离散化,时间复杂度为 $O(1)$。然后通过朴素贝叶斯概率公式对每个

网络应用计算后验概率,并找出最大的后验概率,由于样本特征属性和协议类型是有限的,因此其时间复杂度也可以近似看成是常数 $O(1)$。所以,朴素贝叶斯分类器处理每一条流的时间复杂度都是常数,对于 n 个样本的分类处理,整个时间复杂度为 $O(N)$。

特征加权的朴素贝叶斯分类器仅在计算后验概率时为每个特征赋予一个权重,时间复杂度与朴素贝叶斯分类器一样,处理单条流的时间复杂度为 $O(1)$,n 个样本的分类处理时间复杂度也为 $O(N)$。为了测试 AWNBTC 方法的准确时间,使用三组数据分别对特征加权的朴素贝叶斯分类器进行训练,得到分类模型,然后,使用另外两组流量数据对分类模型进行测试,每组实验重复 10 次,得到的实验结果如表 8。从表中可以看出,AWNBTC 方法可每秒约处理 15000 条网络流,而中国科技网国际出口在一天当中高峰时刻的流约为 14000 条/s,AWNBTC 方法完全能够满足实时处理该出口的流量。

表 8 AWNBTC 方法的分类时间

训练集	CSTNET_1	CSTNET_2	CSTNET_3
训练模型	3.7s	3.6s	3.5s
测试模型	2.0s	2s	2s

4.2 时间稳定性

当前的网络应用可谓是百花齐放,层出不穷,时时刻刻都在发生着变化。为了测试分类器的时间稳定性,即旧的分类模型是否适用于新数据的分类,使用三组原数据分别进行模型训练,一个月后的流量数据(CSTNET_set4)进行模型测试,每组实验重复 10 次。实验结果表明整体的平均分类准确率为 94.2%,保持稳定。应用的分类精度和召回率如表 9所示,虽然每类应用的分类精度和召回率都有些波动,但总体仍保持较高的分类准确性。实验结果表明,AWNBTC 算法具有很强的动态适应性,不需经常更新样本和对模型进行重新训练,适合于流量的在线识别。

表 9 AWNBTC 算法时间稳定性测试结果

训练集	CSTNET_1		CSTNET_2		CSTNET_3	
	精度	召回率	精度	召回率	精度	召回率
Http	85.3%	92.5%	89.5%	91.5%	89.1%	91.3%
Https	83.4%	92.7%	83.6%	92.1%	84.4%	90.7%
Data-transfer	97.7%	99.4%	99.6%	99.3%	99.8%	99.4%
SSH	96.9%	98.8%	96.1%	98.0%	97.6%	97.7%
Jabber	98.6%	99.3%	98.2%	98.8%	98.6%	98.7%
XunLei	99.9%	99.6%	99.8%	98.6%	99.9%	99.7%
SMTP	79.8%	67.5%	77.9%	68.8%	76.3%	69.2%
MSN	99.6%	99.1%	97.9%	98.7%	98.7%	98.7%
IMAP	99.5%	99.0%	98.7%	98.6%	97.3%	97.9%
BitTorrent	98.6%	99.7%	99.5%	99.7%	98.3%	99.8%

5 结 论

实际网络中流量特征无法满足朴素贝叶斯分类方法的特征同等重要和条件独立性假设,从而导致分类准确率低,为此本文提出了一种特征加权的朴素贝叶斯流量分类算法。该算法有如下优势:(1)仅使用 NetFlow 记录的统计特征便得到很高的分类准确率,减少了基于数据报文计算流量特征的开销,提高了分类速度,节约了存储空间,更适合用于流量的在线识别。(2)具有朴素贝叶斯算法简单高效的分类特点,又不依赖于朴素贝叶斯的条件独立性假设,并且较朴素贝叶斯算法在整体的分类准确率、每类应用的分类精度和召回率上都有了很大的改进。

理论分析和实验结果表明,本文算法具有超过94%的分类准确率,并且能适应网络的动态变化,可以满足当前高带宽网络流量分类准确性、实时性和稳定性的需求。

参考文献

[1] Moore A W, Zuev D. Internet traffic classification using Bayesian analysis techniques. *ACM SIGMETRICS Performance Evaluation Review*,2005, 33(1):50-60

[2] Williams N,Zander S,Armitage G. A preliminaryperformance comparison of five machine learning algorithms forpractical IP traffic flowclassification. *ACM SIGCOMM Computer CommunicationReview*, 2006, 36(5):5-16

[3] Internet Assigned Numbers Authority. http://www.iana.org.

[4] Moore A W, Papagiannaki K. Toward the accurate identification of network applications. In: Proceedings of the 6th Passive and Active Measurement Workshop, Boston, USA, 2005. 41-54

[5] Zander S, Nguyen T, Armitage G. Automated trafficclassification and application identification using machine learning. In: Proceedings of the 30th IEEE Conference on Local Computer Networks, Sydney, Australia, 2005. 250-257

[6] Karagiannis T, Papagiannaki K, Faloutsos M. BLINC: multilevel traffic classification in the dark. *ACM SIGCOMMComputer Communication Review*, 2005, 35(4): 229-240

[7] Tavallaee M, Lu W, Ghorbani A. Online classification ofnetwork flows. In: Proceedings of the 7th Communication Networks and Services Research Conference, Moncton, Canada, 2009. 78-85

[8] Crotti M, Dusi M, Gringoli F, et al. Trafficclassification through simple statistical fingerprinting. *ACM SIGCOMMComputer Communication Review*, 2007, 37(1):5-16

[9] Sen S, Spatscheck O, Wang D. Accurate, scalable in network identification of P2P traffic using application signatures. In: Proceedings of the 13th International Conference on World Wide Web, New York, USA, 2004. 512-521

[10] Nguyen T T T, Armitage G. A survey of techniques for Internet traffic classification using machine learning. *IEEE Communications Surveys and Tutorials*, 2008, 10(4):56-76

[11] McGregor A, Hall M, Lorier P, et al. Flow clustering using machine learning techniques. In: Proceedings of the 5th Passive and Active Measurement Workshop, Antibes Juan-les-Pins, France, 2004. 205-214

[12] Bernaille L,Teixeira R, Akodkenou I, et al. Traffic classification on the fly. *ACM SIGCOMM Computer Communication Review*,2006, 36(2):23-26

[13] Katkar V D, Kulkarni S V. Experiments on detection of denial of service attacks using Naive Bayesian classifier. In: Proceedings of the 2013 IEEE International Conference on Green Computing, Communication and Conservation of Energy, Chennal, India, 2013. 725-730

[14] Zhen L, Qiong L. A new feature selection method for internet traffic classification using ml. *Physics Procedia*, 2012, 33: 1338-1345

[15] Antonio T, Paramita A S. Feature selection technique impact for Internet traffic classification using Naive Bayesian. *Jurnal Teknologi*, 2015, 72(5):141-145

[16] Raveendran R, Menon R. An efficient method for Internet traffic classification and identification using statistical features. *International Journal of Engineering Research and Technology*,2015, 4(7):297-303

[17] Ghofrani F, Jamshidi A, Keshavarz-Haddad A. Internet traffic classification using hidden naive Bayes model. In: Proceedings of the 23rd Iranian Conference on Electrical Engineering, Tehran, Iran, 2015. 235-240

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

[19] 陈亮,龚俭. 基于NetFlow记录的高速应用流量分类方法. 通信学报, 2012, 33(1):145-152

[20] Kononenko I. Estimating attributes: analysis and extensions of Relief. In: Proceedings of the 1994 European Conference on Machine Learning, Catania, Italy, 1994. 171-182

[21] Han J, Kamber M. Data Mining: Concepts and Techniques. San Francisco: Morgan Kaufmann, 2001

[22] Yu L, Liu H. Feature selection for high-dimensionaldata: A fast correlation-based filter solution. In: Proceedings of the 20th International Conference on Machine Learning, Washington D. C, USA, 2003. 856-863

Internet traffic classification using the attribute weighted naive Bayes algorithm

Zhang Zexin[*,**], Li Jun[*], Chang Xiangqing[*]

([*] Computer Network Information Center, Chinese Academy of Sciences, Beijing 100190)

([**] University of Chinese Academy of Sciences, Beijing 100049)

Abstract

Based on the analysis of the performance characteristics of the Naive Bayes (NB) method in wide use for network traffic classification, a novel attribute weighted Naïve Bayes classification algorithm was proposed to overcome the NB method's problem of low classification accuracy caused by its assumption that traffic attributes are of equal importance and independence is hard to satisfy in practice. The proposed algorithm uses the attribute selection algorithm of ReliefF and the correlation coefficient method to calculate the attribute weights based on the attribute information recorded by NetFlow. Then, it assigns a new instance to the most probable class, which has the largest posterior probability. The experiment showed that the classification accuracy of the proposed algorithm was over 94%, and the algorithm maintained simpleness, high efficiency and stability of the NB method. In short, this algorithm can fully meet the traffic classification demands of high-bandwidth networks.

Key words: traffic classification (TC), ReliefF, correlation coefficient, attribute weighting (AW), Naive Bayes (NB), NetFlow