基于 GA-CFS 和 AdaBoost 算法的网络流量分类

刺婷婷，师　军
（陕西师范大学 计算机科学学院，西安 710062）

摘　要：鉴于特征属性选择在网络流量分类中占据重要地位，为了确定最优特征子集，利用 CFS 作为适应度函数的改进遗传算法（GA-CFS），从网络流量的 249 个属性空间中提取主要属性并最终选定 18 个特征组合作为最优特征子集。通过 AdaBoost 算法把一系列的弱分类器提升为强分类器，对网络流量进行了深入的分类研究。实验结果表明，基于 GA-CFS 和 AdaBoost 的流量组合分类方法较弱分类器具有较高的分类准确率。

关键词：流量分类；相关性特征选择；适应度函数；AdaBoost 算法；弱分类器；权重
中图分类号：TP393　　**文献标志码**：A　　**文章编号**：1001-3695(2012)09-3411-04
doi:10.3969/j.issn.1001-3695.2012.09.056

Network traffic classification based on GA-CFS and AdaBoost algorithm

LA Ting-ting, SHI Jun
(*School of Computer Science, Shaanxi Normal University, Xi'an* 710062, *China*)

Abstract: The selection of feature attribute plays an important role in the network traffic classification. This paper applied a method considering the CFS algorithm as the fitness function of the improved genetic algorithm (GA-CFS) in order to extract the main flow statistical attributes in the space of 249 attributes and selected 18 attributes of a flow as the best feature subset. Finally it used the AdaBoost algorithm to enhance a series of weak classifiers to the strong classifiers. At the same time, it fulfilled the classification of the network traffic, and further studied the network traffic intensively. The experimental results indicate that GA-CFS and AdaBoost algorithm can achieve higher classification precision compared with the weak classifiers.

Key words: traffic classification; CFS; fitness function; AdaBoost algorithm; weak classifier; weight

基于应用类别的准确、高效的网络流量分类是入侵检测、实时网络流量模型的建立和网络规模预测的基础。近年来，一些诸如 P2P、文件共享和在线游戏等应用的不断流行，迫切需要找出一种快速的流量识别技术来划分这些多样的不断涌现的网络应用。

鉴于传统的网络流量分类方法已不能很好地适应飞速发展的网络技术，许多研究人员根据流属性的统计信息，将机器学习方法应用到流量分类领域中。使用机器学习方法进行网络流量分类的前提是将具有相同的元组信息（源 IP 地址、目的 IP 地址、源端口、目的端口、协议）的网络流作为一个流对象。首先，从一组属性向量中提取网络流量属性特征。从机器学习的角度来看，网络流量分类可以抽象为：已知流量属性特征向量 x_i，流量类别 y_i，$x_i=\{x_{i1},x_{i2},\cdots\}$ 表示流量属性特征集，y_i 表示流量类别集[1]。机器学习的最终目标是找到一个映射函数，来精确预测某个未知流量 x_i 的流量类别 y_i。基于流量统计特征的网络流量分类中，获得用于网络流量分类的最优特征子集是非常关键的一步。基于此，本文将遗传算法用于属性特征的选择，再结合一种新的分类方法——AdaBoost 算法，对网络流量进行分类。

1　流量特征属性选择

本文首先利用特征选择方法分别对每一类网络流量统计特征进行提取和降维，再利用 AdaBoost 算法进行网络流量的有监督分类。AdaBoost 网络流量分类模型框架如图 1 所示。

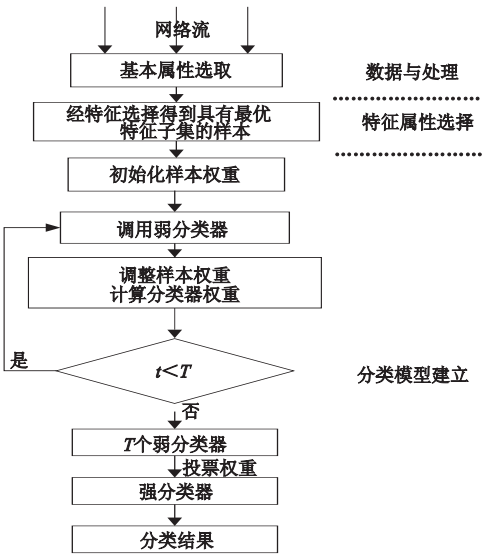

图1　AdaBoost网络流量分类模型框架

分类过程分为以下三步：

a）根据 Moore_set 的 249 个属性，选取各个样本流中含有的 37 个基本属性作为研究对象；

b）采用改进遗传算法进行特征选择，提取 37 个属性特征中对分类影响较大的主要属性；

c）根据提取的网络流主要特征，结合 AdaBoost 分类算法，

收稿日期：2011-12-26；**修回日期**：2012-02-13

作者简介：刺婷婷（1987-），女，山西介休人，硕士，主要研究方向为网络流量分类识别、网络信息系统安全（lakaiting@163.com）；师军（1957-），男，陕西西安人，副教授，主要研究方向为智能信息处理.

构建网络流量分类模型。

分类所使用的流量数据来自 Moore_set[2]，采集间隔约为 12 h，每一个数据集均包含数万条数据，这些数据集中的每一个样本都是从一条完整的 TCP 双向流抽象而来，包含 249 个属性[3]，其中最后一个属性即为分类目标属性，表明该样本的流量类型。参照 Moore_set 数据集，将网络流量的类别定义为 10 类[4]，分别为 WWW、MAIL、BULK、DATABASE、SERVICES、P2P、ATTACK、MULTMEDIA、INTERVATION、GAMES。

1.1 基于改进遗传算法的特征选择

为了使分类过程变得容易实现，希望依据最少的特征达到所要求的分类识别的正确率最高，因此，在初始获得网络流量的特征之后，再由这些原始特征产生出使得分类识别最有效、数目最少的特征子集，这就是网络流量分类中特征提取与选择的任务[5]。

在 Moore_set 数据集中，每一条网络样本流包含 249 个网络属性，其中有 100 多个属性是通过傅里叶变换技术得到的，存在众多的冗余属性和无关属性，这些属性的存在不仅会降低分类模型的准确率，而且会大大加重分类模型的计算负载[6]。

因此，为了形成有利于机器学习的流样本，避免对应用端口的依赖，在流的属性特征产生上本文考虑独立于协议、通信端口号的特征。先从 Moore_set 的 249 个属性中，选取数据包长度、间隔、到达时间、数目以及流持续时间等 37 个统计特征作为流的候选特征集。利用遗传算法将维度较高的网络流样本降维并简化。

遗传算法(genetic algorithm, GA)是由美国 Michigan 大学的 Holland 教授于 1975 年首先提出的。它是一种通过模拟自然进化搜索最优解的方法，首先将问题假设按照某种形式进行编码，通过编码组成初始群体，按照适者生存和优胜劣汰的原理，逐代演化产生出越来越好的近似解，在每一代，根据问题域中个体的适应度大小选择个体，并借助自然遗传学的遗传算子进行选择、交叉和变异，产生出代表新的解集的种群。末代种群中的最优个体经过解码，可以作为问题近似最优解[7]。由于简单遗传算法具有早熟收敛和局部搜索能力弱的缺陷，影响和制约它在特征选择算法中的高效应用[8]，因此本文采用 CFS 算法作为遗传算法的适应度函数来进行网络流属性选择。

1.2 GA-CFS 特征选择的具体过程

采用 CFS 算法作为遗传算法的适应度函数，对网络流样本数据进行迭代，最终得到用于网络流量分类的最优特征子集。基于 GA-CFS 的特征选择过程具体描述如下：

a)编码。初始输入 37 个网络流量属性，故染色体的长度为 37，个体的每个基因对应相应次序的特征，即当个体的某个基因为“1”时，表示对应的特征被选用；为“0”时，表示该特征未被选用。

b)设定初始群体和终止条件。本文初步设定每代群体个数为 3 000，利用随机函数产生 2 000 个染色体组成初始群体，进化代数设置为 200 代。

c)适应度函数的设定。特征选择是为了找到适合分类的最优特征子集，因此需要一个定量准则来衡量各个体对应的特征组合的分类能力。本文采用文献[9]中的 CFS 相关性特征选择作为适应度函数对每一代进行评估。

d)遗传算子的确定。通过轮盘赌方法、单点随机交叉的方法和基因位变异的方法，将交叉概率设定为 0.8，变异概率取 0.25 来确定遗传算子。

e)群体经过选择、交叉、变异运算之后得到下一代群体。

f)当循环结束时终止运算。最终得到最大适应度函数以及最优特征子集。

1.3 属性选择结果

依据以上过程，基于 MATLAB 7.7 平台，以 Moore_set 的 set01 和 set02 两个数据集作为特征选择的输入样本，各选取其中的 3 000 个网络流样本进行实验；以选取的 37 个候选特征作为初始特征，记为 $x_1, x_2, \cdots, x_{37}$，进行选择交叉变异的遗传操作，得到适应度最大的个体。此个体对应的网络流的 17 个特征属性，结合网络流量的类别属性，共同构成用于后续网络流量分类的最优特征子集。属性缩写及详细描述如表 1 所示。

表 1 遗传算法进行特征选择所得的最优特征子集

属性缩写	属性描述
total_packets_b a	下行流量的数据包总个数
actual_data_pkts_a b	上行流量的 TCP 数据包总个数
actual_data_bytes_b a	下行流量的总字节数
data_xmit_time_a b	一个完整数据包的总传送时间
min_retr_time_a b max_retr_time_a b	上行流量中任意两连接的最小、最大时间间隔
Avg_retr_time_a b Avg_retr_time_b a	上行、下行流量中任意两连接的平均时间间隔
var_data_ip_a b	上行 IP 数据包的可变字节数
Duration	连接的持续时间
means_data_ip_b a	下行 IP 数据包的平均字节数
Time_spent_idle	空闲时间
Effective_Bandwidth_a b	上行数据包的有效带宽
mean_data_wire_a b mean_data_wire_b a	上行、下行以太网数据包的平均字节数
var_data_wire_a b var_data_wire_b a	上行、下行以太网数据包的可变字节数
Class	流量类别

说明：a 表示客户端；b 表示服务器端；a b 表示上行流量；b a 表示下行流量。

2 AdaBoost 网络流量分类

由于真实网络环境中网络流量样本的分布是动态变化的，而朴素贝叶斯方法中使用特定条件下的采样值来估计实际网络中不断变化的动态值，因此将无法保证流量分类结果的稳定性[6]。众所周知，寻找一个精度比随机预测略高的弱学习算法比寻找高精度的强学习算法要容易得多。为此本文给出一种新的网络流量分类算法，即选用 C4.5 决策树算法作为弱分类器，采用 AdaBoost 算法将它们提升为强分类算法，进行更准确的流量分类。

2.1 AdaBoost 算法思想

AdaBoost(adaptive boosting)由 Freund 和 Schapire 提出。AdaBoost 最初用于处理两类问题，之后出现很多用于处理多类问题的扩展算法，如 AdaBoost M1、Adaboost M2、Adaboost MH、AdaBoost OC 以及 AdaBoost ECC 等[10,11]。

其基本思想是在整个训练样本上维护弱分类器权重。起初每个训练样本被赋予相同的权重，随后通过 AdaBoost 算法对赋予权重的训练集训练 T 轮，每一轮产生弱分类器假设，计算它的错误率，并运用错误率更新样本权重和弱分类器权重，对错误分类的样本分配赋予更大的权值，正确分类的样本赋予

更小的权值,这样下一轮学习时学习算法就会更集中于对错误分类的样本进行分类。T 轮学习后,会产生 T 个弱分类器,其中分类效果比较好的分类器的投票权重(不同于样本权重)比较大,最终的分类器采用一种有权重的投票方式产生,并且各个弱分类器之间的差异性越大,被集中的强分类器泛化能力越好[12]。

2.2　基于 AdaBoost 算法的网络流量分类模型的建立

网络流量多值分类算法在 AdaBoostM1 上的具体实现:

a)给定类别标定的训练样本集 $S=\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$,$y_i\in Y=\{1,2,\cdots,k\}$,$k$ 为类别数。

b)初始化。对于每一个样本$(x_i,y_i)\in S$,初始 $D_1(x_i,y_i)=1/n$。

c)开始循环 $t=1,2,\cdots,T$。

(a)在当前的分布 D_t 下,调用弱分类器(本文采用 C4.5 决策树分类算法)得到第 t 轮的分类规则:$h_t:X\rightarrow Y$;

(b)计算弱分类器 h_t 的分类误差:$\varepsilon_t=\sum_{h_t(x_i)\neq y_i}D_i(x_i,y_i)$;

(c)计算第 t 轮分类器的分类权重:$\beta_t=\varepsilon/(1-\varepsilon)$;

(d)更新样本权值:$D_{t+1}(x_i,y_i)=\dfrac{D_t(x_i,y_i)}{Z_j^t}\times\begin{cases}\beta_t & \text{if } h_t(x_i)=y_i\\ 1 & \text{otherwise}\end{cases}$,此处 Z_j^t 是标准化常量,使得 $\sum_{(x_i,y_i)\in s_j}D_t(x_i,y_i)=1$。

d)T 轮循环结束之后,输出最终的组合分类器:$h_f(x)=\arg\max_{y\in Y}\sum_{t:h_t(x_i=y)}\log 1/\beta_t$。

主要参数说明如下:

$D_i(x_i,y_i)$表示样本(x_i,y_i)在第 t 轮循环时的权重;

ε_t 表示第 t 轮弱分类器的训练错误率;

β_t 表示第 t 轮弱分类器的权重。

2.3　AdaBoost 算法的优点

在众多的分类学习算法中,AdaBoost 因其具有以下优点而被广泛应用:

a)可以集多种分类器之大成,具有减小训练样本分类错误的能力,广泛运用到各个领域。

b)AdaBoost 可以用来鉴别异常,重新赋权值后具有最高权重的样本很可能为异常,这有利于在网络流量测量领域进行异常流量数据的检测。

c)AdaBoost 不需要任何关于弱学习器性能的先验知识,因此可以非常容易地应用于实际问题中,并且容易处理弱学习算法得出的预测规则的正确性,理论上可以达到任何精度[13]。

3　实验建立和结果分析

本文的所有分类实验都使用 Weka 数据挖掘工具进行。Weka 全名为怀卡托智能分析环境(Waikato environment for analysis)[6]是由新西兰怀托大学开发,它是一个基于 Java、用于数据挖掘和知识发现的开源项目。

3.1　分类算法评价策略

为了有效评价分类算法,Weka 使用准确率(precision)、召回率(recall)和调和平均值(F-measure)三个评价参数来描述分类的结果。其中,准确率刻画的是分类器给出的类别为 C_i 的预测中正确结果的比例;召回率则代表了本应是 C_i 的实例,被正确预测的比例,即以多大的概率被正确召回到类;而 F-measure 是两者的几何平均值,F-measure 的值越接近 1,说明分类结果越精确[14]。这三个评价参数的计算式为

$$\text{precision}=\text{TP}/(\text{TP}+\text{FP})$$

$$\text{recall}=\text{TP}/(\text{TP}+\text{FN})$$

$$\text{F-measure}=(2\times\text{precision}\times\text{recall})/(\text{precision}+\text{recall})$$

其中:TP(truly positive)是指那些分类为 C_i 实际上也分到 C_i 类的网络流样本;FP(falsely positive)是指那些分类为 C_i 但实际上不是 C_i 类的网络流样本;FN(falsely negative)是指那些没有分到 C_i 类中,但是其类别实际上为 C_i 的网络流样本。

3.2　基于 AdaBoost 的准确率分析

本文对 Moore_set 的前五个数据集进行实验,采用十折交叉验证来测试 AdaBoost 算法的分类正确率。所谓十折交叉实验是一种常用的测试方法,即将数据集分成 10 份,轮流将其中 9 份作为训练数据,1 份作为测试数据来进行实验。每次实验都会得出相应的正确率,10 次结果正确率的平均值作为对算法精度的估计。分类结果如表 2 所示。

表 2　使用 AdaBoost + C4.5 的网络流量分类结果

数据集	precision/%	recall/%	F-measure/%	time/s
Set01	98.8	98.9	98.8	38.30
Set02	99.3	99.4	99.4	25.06
Set03	99.2	99.3	99.2	35.23
Set04	99.0	99.0	99.0	33.80
Set05	99.2	99.2	99.3	31.28

本次实验中,在考虑了特征选择的条件下,采用了最优特征子集的 18 个属性作为 AdaBoost 算法的输入。表 2 列出了 Moore_set 的前五个子集上的分类结果。分析可知,AdaBoost 算法在 Moore_set 上具有很高的分类精度,准确率最高为 99.4%,最低为 98.8%,波动幅度仅为 0.6%,这也说明 AdaBoost 用于网络流量分类具有很高的稳定性。

本文比较了以 C4.5 作为弱分类器的 AdaBoost 算法和 C4.5 分类方法处理网络流量分类问题的性能,比较结果如表 3 所示。从表 3 可知,两种方法都有较好的分类准确率和稳定性,这表明在使用了遗传算法进行特征选择之后,用少的属性特征可以达到较高的分类准确率;并在所选定的属性中去除了端口信息,可见基于机器学习的网络流量分类问题不依赖于端口信息,不依赖于应用层负载信息,是一种健壮性良好的网络流量分类方法。进一步分析表 3 的结果可知,基于 C4.5 的分类算法准确率为 97.8%,而用 AdaBoost 将 C4.5 分类器提升后,分类正确率高达 99.2%,可见 AdaBoost 分类算法可以有效地提高弱分类器的分类准确率。

表 3　AdaBoost 和 C4.5 分类结果比较

参数	C4.5	AdaBoost + C4.5
precision/%	97.8	99.2
recall/%	98.1	99.3
build time/s	3.03	35.23

3.3　特征选择对分类准确率的影响

为进一步研究特征选择算法对分类准确率的影响,本文对比了特征选择前网络流具有的 37 个属性和特征选择之后的 18 个属性的分类准确率,实验结果如表 4 所示。鉴于 GAMES 和 INTERVATION 这两类别样本数很少,故只列出另外八种类别。

表 4 特征选择前后各类别分类准确率对比

参数	AdaBoost + NB		NB	
	37 个	18 个	37 个	18 个
WWW/%	99.3	99.1	99.6	99.4
MAIL/%	99.1	98.1	99.5	99.1
FTP/%	89.7	80.0	80.3	82.2
ATTACK/%	98.4	98.4	12.6	23.8
P2P/%	92.3	92.2	44.2	52.6
DATABASE/%	78.8	91.6	82.6	82.6
MULTIMEDIA/%	65.1	64.3	32.4	30.4
SERVICES/%	98.6	99.0	91.1	78.2
平均/%	98.6	98.4	93.1	92.7
建模时间/s	61.14	18.20	3.08	1.53

从表 4 的分析可知,只需要 18 个属性,就能获得与 37 个属性相当甚至更好的分类准确率,大大节省了计算开销。

1)特征选择后,选择出的特征子集较好地保持了分类的结果。由表 4 可知,特征选择前,AdaBoost 算法建模时间为 61.14 s,而特征选择之后的 18 个属性特征的建模时间仅为 18.20 s,在保证平均分类准确率仅比特征选择之前少 0.2% 的情况下,时间复杂度明显降低。

2)运用 AdaBoost 分类算法将 Naïve Bayes 分类算法提升以后,分类准确率得到了很大提高,特别是针对现在迅速发展起来的 P2P 流量的分类,以及异常流量、多媒体流量都具有很好的识别。

4 结束语

在基于数据挖掘的网络流量分类技术的研究中,数据源的获取是研究的基础,数据的预处理直接影响网络流量分类的效果。本文采用改进遗传算法,对 Moore_set 中网络流的 249 个属性进行选择,得到的最优特征子集不仅降低了数据处理的复杂度,而且还保持了很好的分类准确率。随后在 Moore_set 的五个网络流数据集上应用 AdaBoost 算法进行分类实验,该算法获得了较好的优化性能,能够有效降低网络流量属性集合中冗余属性和无关属性对分类准确率的影响。这说明 AdaBoost 算法不但可以提高弱分类器的搜索效率,而且可以进一步改进分类器的整体性能。需要指出的是,本文虽然取得了一定成果,但未来还需在以下几个方面进行改进:

a)深入研究不同属性对不同类别的区分能力,考虑对不同类别的区分,使用不同的属性集;

b)对目前新出现并且发展迅速的新应用的分类,特别是 P2P 应用需进行更加深入的研究;

c)AdaBoost 与聚类方法结合,既能识别分类类型尚未定义的新型网络流量,即实现无监督的网络流量分类,又能使分类具有较高的效率;

d)传统的网络流量与新型的基于流统计特征的方法结合,也是今后可以探索的一个方面。

参考文献:

[1] WANG Jian-min, QIAN Cheng-lu, CHE Chun-hui, *et al*. Study on process of network traffic classification using machine learning[C]// Proc of the 5th Annual ChinaGrid Conference. 2005:262-266.

[2] MOORE A W, ZUEV D, CROGAN M. Discriminators for use in flow-based classification[M]. London: Queen Mary University of London, 2005.

[3] 徐鹏,林森,刘琼. 基于决策树的流量分类方法[J]. 计算机应用研究,2008,25(8):2484-2487.

[4] MOORE A W, PAPAGIANNAKI K. Toward the accurate identification of network application[C]//Proc of Passive & Active Measurement Workshop 2005. Boston: Springer-Verlag, 2005:41-54.

[5] 任江涛,黄焕宇,孙婧昊,等. 基于遗传算法及聚类的基因表达数据特征选择[J]. 计算机科学,2006,33(9): 155-156.

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

[7] 柳亚琴,石洪波. 基于 GA-CFS 属性选择的个体信用评估模型[J]. 计算机系统应用,2011,20(5):210-213.

[8] 张忠林,张军,米伟. 一种具有记忆功能的遗传算法属性约简方法[J]. 计算机应用研究,2010,27(1): 96-98.

[9] FREUND Y, SCHAPIRE R E. A decision-theoretic generalization of on-line learning and an application to boosting[J]. Journal of Computer and System Sciences, 1997,55(1):119-139.

[10] FRIEDMAN J, HASTIE T, TIBSHIRANI R. Additive logistic regression: a statistical view of boosting[J]. The Annals of Statistics, 2000,28(2):337-407.

[11] SCHAPIRE R E, SINGER Y. Improved boosting algorithms using confidence-rated predictions[J]. Machine Learning,1999,37(3): 297-336.

[12] SHAN Shi-guang, YANG Peng, CHEN Xi-lin, *et al*. AdaBoost gabor fisher classifier for face recognition[J]. Computer Science,2005, 32(3): 279-292.

[13] 张艳阳,顾明. 基于 AdaBoost 分类器的车牌字符识别算法研究[J]. 计算机应用研究,2006,23(5): 242-247.

[14] 许孟晋,张博峰. 基于机器学习的 Internet 流量分类[J]. 计算机应用,2010,30(1): 80-82.