基于GATS-C4.5的IP流分类

李文法[1,2] 陈 友[1,2] 段洣毅[1,3] 孙春来[3]
(中国科学院计算技术研究所 北京 100190)[1] (中国科学院研究生院 北京 100039)[2]
(北京交通大学计算技术研究所 北京 100029)[3]

摘 要 流分类技术在网络安全监控、QoS、入侵检测等应用领域起着重要的作用,是当前研究的热点。提出一种新的特征选择算法GATS-C4.5来构建轻量级的IP流分类器。该算法采用遗传算法与禁忌搜索相混合的搜索策略对特征子集空间进行随机搜索,然后利用提供的数据在C4.5上的分类正确率作为特征子集的评价标准来获取最优特征子集。在IP流数据集上进行了大量的实验,实验结果表明基于GATS-C4.5的流分类器在不影响检测准确度的情况下能够提高检测速度,并且基于GATS-C4.5的IP流分类器与NBK-FCBF(Naïve Bayes method with Kernel density estimation after Correlation-Based Filter)相比具有更小的计算复杂性与更高的检测率。

关键词 流分类,特征选择,遗传算法,禁忌搜索,决策树

IP Flow Classification Based on GATS-C4.5

LI Wen-fa[1,2] CHEN You[1,2] DUAN Mi-yi[1,3] SUN Chun-lai[3]
(Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China)[1]
(Graduate University of Chinese Academy of Sciences, Beijing 100039, China)[2]
(Institute of Computing Technology, Beijing Jiaotong University, Beijing 100029, China)[3]

Abstract Flow classification plays an important role in the research field of network security monitoring, Quality of Service, and intrusion detection. Recently, there has been much interest in them. We proposed a wrapper feature selection algorithm GATS-C4.5 aiming at modeling lightweight flow classifier by (1) using hybrid genetic-tabu approach as search strategy to specify candidate subsets for evaluation; (2) using C4.5 algorithm as wrapper approach to obtain the optimum feature subset. We examined the feasibility of our algorithm by conducting some experiments on flow datasets. The experimental results show that classifier with our approach can greatly improve computational performance without negative impact on classification accuracy. Further more, our approach is able not only to have smaller resource consumption, but also to have higher classification accuracy than Naïve Bayes method with Kernel density estimation after Fast Correlation-Based Filter (NBK-FCBF).

Keywords Flow classification, Feature selection, Genetic algorithm, Tabu search, Decision tree

1 引言

根据网络流准确实时地判断出其所属的应用类型是网络活动需要研究的重要内容。传统的基于端口与基于负载内容的分析技术已经显示出很多的弊端,现在越来越多的研究者把目光投向基于机器学习的流分类,这种流分类中的每一个流由一些特征组成,这些特征共同反映了这个流所属的应用类别。越来越多的特征被产生,用于流分类,但是这些特征中含有许多冗余与杂音特征,这使得对这些数据进行修剪与剔除变得尤为重要。特征选择针对IP流的高维特征空间存在大量的相关与冗余特征的特性,在此高维空间上应用搜索算法来寻找最优的特征子集,剔除那些相关与冗余特征。在得出的最优特征子集上建立的IP流分类器不仅可以降低分类器的时间复杂性,而且可以获得很好的检测效果。特征选择在流分类中扮演着十分重要的角色。

特征选择有filter和wrapper两种模型[1]。filter模型利用数据本身的特性作为特征子集的度量指标,而wrapper模型利用机器学习算法的分类正确率作为特征子集的度量指标。一般来说filter模型的效率高,效果差;wrapper模型的效率低,效果好。为了解决两种特征选择模型存在的问题,发挥它们的优势,很多学者提出了结合filter模型和wrapper模型的hybrid模型[1,2]。虽然hybrid模型在性能上有一定的提高,但是效果不理想。本文采用wrapper模型设计一种高效的特征选择算法,它不仅可以克服wrapper模型计算资源耗

到稿日期:2008-05-05 本文受国家"九七三"重点基础研究发展规划项目(2004CB318109),国家"八六三"高技术研究发展计划项目(2006AA01Z452),国家242信息安全计划项目(2005C39)资助。

李文法 博士,CCF会员,主要研究方向为流分类、网络攻防对抗、网络安全,E-mail:liwenfa@software.ict.ac.cn;陈 友 博士,主要研究方向为网络安全、数据挖掘;段洣毅 研究员,CCF理事,博士生导师,主要研究方向为系统工程、人工智能、网络通信;孙春来 高级工程师,CCF高级会员,主要研究方向为信息安全、通信网络。

用大的缺点，而且选择出的特征很大程度上提高了流分类器的检测率。

2 相关工作

轻量级的IP流分类包括流特征选择与流分类器两个部分，而特征选择包括搜索策略与评估函数。近年来，越来越多的研究者把目光聚集在基于机器学习的流分类上。在文献[3]中，作者采用爬山算法作为搜索策略、相关性(Correlation Feature Selection，CFS)与一致性评估函数的特征选择算法来选择最优的流特征子集。基于该最优特征子集建立的分类器在检测效果影响不大的情况下，降低了时间复杂性，提高了检测速度。文献[4]利用贝叶斯作为分类器来进行流分类，作者首先从网络流数据中产生248个流特征，然后利用CFS选择出最优特征子集，实验结果表明分类器只需要20个特征就能达到很好的分类性能。文献[5]提出了一种结合爬山算法与欧式距离的特征选择算法，实验结果表明不同的应用流可以被很好地划分开。当前很多研究都集中在基于特征选择算法的流分类器上，它们利用的特征选择算法是filter型的。虽然一定程度上提高了分类器的检测速度，但是分类性能却不理想。

本文提出一种wrapper型特征选择算法GATS-C4.5(Genetic Algorithm-Tabu Search as search strategy and C4.5 algorithm as evaluation function)来建立轻量级的流分类器。GATS-C4.5不仅可以提高特征选择的速度，而且基于它选出的特征建立的流分类器具有高检测速度与检测率。我们的贡献有：

(1)特征选择不仅降低了特征维数，而且帮助流分类器获取它们需要的特征子集；

(2)基于特征选择的流分类器不仅具有更小的时间复杂度，而且具有很快的检测速度与很高的检测率；

(3)GATS-C4.5不仅加快了特性选择的速度，而且获取了最优的特征子集。

3 特征选择算法的数学模型

给定一个特征子集 $F=\{f_1,f_2,\cdots,f_N\}$，N 是特征集的大小。一个特征子集可以用一个二进制向量表示：$S=(s_1,s_2,\cdots,s_N)$，$s_i\in\{0,1\}$，$i=1,2,\cdots,N$。$s_i=1$ 表示第 i 个特征 f_i 被选择，反之对第 i 个特征 f_i 不作选择。把C4.5在给定的特征子集 S 上所具有的性能 $G(S)$ 作为目标函数值，则特征选择问题转化为下列优化问题：

$$\max_S G(S) \tag{1}$$

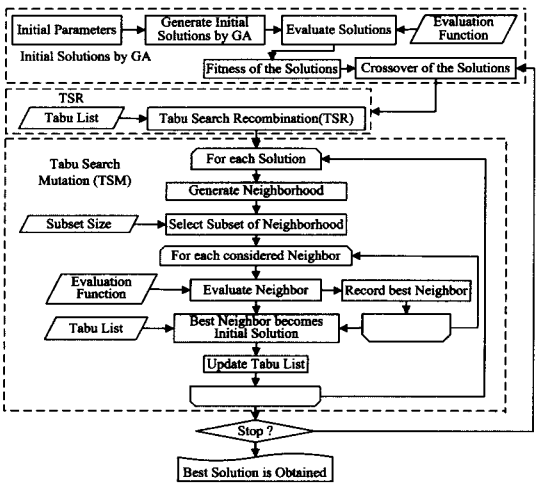
特征选择的求解优化问题 $\max_S G(S)$ 可以通过遗传算法和禁忌搜索的混合策略GATS来求解。

4 GATS-C4.5算法

遗传算法(Genetic algorithm，GA)是由美国学者Holland及其同事、学生提出的。该算法基于Darwin的进化论和Mendel的遗传学说，是一种进化搜索算法，已经成为求解任意函数优化问题的强有力工具[6]。其中重组和变异算子是GA的两个最重要组成部分。禁忌搜索(Tabu Search，TS)是另一个著名的启发式搜索算法，最早由Glover提出[7]，具有记忆功能是TS独创的特点之一。开发混合算法的目的是使原算法的优点被保持，弱点被克服或者被削弱，提高算法的力度。Mülenbein最早把记忆功能引入到GA[8]中，而TS的创始人Glover对混合GA与TS的必要性和可行性进行了理论上的分析和论述[9]，其被公认为是混合GA与TS的理论基础。在Glover理论基础上，本文提出一种GA与TS的混合策略GATS。把TS独有的记忆功能引入到GA进化搜索过程之中，构造了新的重组算子TSR。针对GA爬山能力差的缺陷，利用TS爬山能力强的优点，使用TS算法改进GA的爬山能力，即把TS作为GA的变异算子TSM。GATS综合了GA具有多出发点及TS具有记忆功能和爬山能力强的特点，克服了GA爬山能力差的弱点，并保持了GA具有多出发点的优势。C4.5[10]算法是一种决策树算法，在分类方面具有广泛的应用。GATS-C4.5算法由搜索方法GATS和评估函数C4.5组成。

4.1 GATS-C4.5算法流程

GATS-C4.5的算法流程如图1所示。特征选择算法由搜索策略与评估函数组成，GATS是搜索策略，C4.5是评估函数。流程分成3个阶段：首先初始化种群，得到初始个体集合，然后在每个个体上应用一定的算子进行操作，经过多次操作，最后形成最优的个体。图1主要由3部分组成：遗传算法初始化种群；禁忌搜索重组(Tabu Search Recombination)；禁忌搜索变异(Tabu Search Mutation)。下面主要从这3个方面来介绍GATS-C4.5算法，首先进行种群初始化，然后在初始化种群上应用TSR与TSM。



遗传禁忌作为搜索策略，C4.5决策树作为评估函数

图1 特征选择算法总体流程图

4.2 算法初始化

算法首先初始化种群规模、停止标准等参数，然后在初始化个体上进行评估，得出每个个体的适应值。对特征子集 $F=\{f_1,f_2,\cdots,f_N\}$ 中的每一个特征利用二进制进行编码，得到一个码长为 N 的二进制串：$h=h_1h_2\cdots h_N$，这一编码串表示对特征集所做的一次选择，其中 $h_i=1$ 表示第 i 特征被选择，所有选择的特征构成一个特征子集。$H=\{h_1h_2\cdots h_N\mid h_i\in\{0,1\},i=1,2,\cdots,N\}$ 为所有特征子集的集合，称为个体空间，个体空间的大小为 2^N。在评估函数上，我们利用C4.5在数据集上获取的分类错误率作为评估标准，针对每一个二进制串 h，其适应值的计算公式为

$$f(h)=\alpha \cdot P_{error}(h)+(1-\alpha) \cdot \frac{|h|}{N} \quad (2)$$

其中 $P_{error}(h)=\frac{1}{M}\sum_{i=1}^{M}\gamma_i$，$P_{error}(h)$是基于特征子集 h 的分类器的分类错误率的平均值，其中 M 表示类别数目，γ_i 是每一类的分类错误率，$|h|$是 h 中 1 的数目，N 是特征总数目。$f(h)$是分类错误率与特征子集 h 中 1 的数目占特征总数目的比例的组合。分类错误率与 h 中 1 的占有率之间通过 α 来权衡：α 越大，$f(h)$更强调错误分类率；α 越小，$f(h)$更强调 h 中 1 的数目。

4.3 禁忌搜索重组(Tabu Search Recombination, TSR)

TSR 算子作为重组算子，使用一个长度为 T 的禁忌表，表中记录染色体的适应值，渴望水平作为父代群体适应值的平均值。进行 TSR 操作时，首先把子代的适应值同渴望水平相比较。如果渴望水平好，则破禁，即这个染色体进入到下一代中。如果子代比渴望水平差，但不属于禁忌，也接受这个子代；若属于禁忌，则选择那个最好的父代进入到下一代中。TSR 的重组过程如图 2 所示。从 TSR 的重组过程可以看出，具有高适应值的子代进入到下一代的机会是很大的，但是并不是所有的高适应值的子代一定都进入到下一代。因为 TSR 使用了禁忌表，它可以限制适应值相同的子代出现的次数，因此可使群体中尽可能保持染色体结构的多样性，从而避免算法早熟。

```
Begin
  if fitness of x > average value of population
    then accept x;
  else
    if offspring x is not in tabu list
      accept x
    else
      choose the better of two parents to the next generation;
    update tabu list;
end
```

图 2 TSR 过程伪代码

4.4 禁忌搜索变异(Tabu Search Mutation, TSM)

TSM 与标准变异算子极为相似。首先，TSM 把一个染色体作为输入初始解，经过 TSM 作用，返回一个解作为输出。不同之处在于 TSM 是一个搜索过程，因此需要调用评价函数来确定移动值，并根据移动值和禁忌表 T 决定接受哪个移动输出。同样，由于 TSM 是一个 TS 搜索过程，在搜索过程中可以接受劣解，因此 TSM 具有强于其它(如到位和部分到位算子)的爬山能力。设 x 是一个染色体，则 TSM 的操作过程如图 3 所示。

```
Begin
  t=0; set the best solution x(0)=x; set T;
  while termination condition not satisfied do
    t=t+1;
    move x to x';
    update(x,x(0),tabu list);
end
```

图 3 TSM 过程伪代码

5 试验研究

为了验证特征选择算法对流分类计算复杂性与分类性能的影响，我们进行了大量的实验。首先用 GATS-C4.5 在给定的数据集上进行特征选择，然后比较结合 GATS-C4.5 的流分类器与没有使用特征选择的流分类器在系统建模时间、检测速度、检测率上的差异，最后比较基于 GATS-C4.5 的流分类器与 NBK-FCBF 在检测速度与检测率上的差异。

我们所有的实验都在文献[11]的数据集上完成，这个数据集由 Moore[4] 产生。这个数据集是在高性能网络监控器上获取的[12]，网络监控器所在的网络是一些与生物研究相关的研究机构，网络结点主要包括 3 个研究机构的 1000 多位研究者、管理者以及教员。每一个流集合都记录了 24h 双向流。关于流集合的详细内容将在下节介绍。

5.1 IP 流数据集与 IP 流特征

我们实验的数据集来自文献[11]，并且 Moore[4] 用贝叶斯方法在这个数据集上建立了流分类器 NBK-FCBF。在这个数据集中，每一个流是由一个元组定义的，元组包括许多特征，如连接流的两个端点主机的 IP 地址。协议类型如 ICMP，TCP，UDP；在 UDP，TCP 中的主机端口号，TCP 连接的持续时间等。每一个流有 248 个这样的独立特征，并且这些特征是流分类器的输入参数。表 1 列出的是部分特征，全部的特征见文献[13]。

表 1 作为分类器输入的流特征在流的两个方向上统计获取

Features
Flow Duration
TCP Port
Packet inter-arrival time
Payload size
Fourier Transform of packet inter-arrival time

每一个流除了 248 个特征之外，还有一个定义流的应用类型的类，如 WWW，P2P，MAIL，BULK 等。表 2 列出了数据集中所有的类，如 BULK 类是由 ftp 流组成的，这个流具有双向性，并且由 ftp 控制流与 ftp 数据流构成。关于流中类详细的描述见文献[14]。

表 2 网络流量的类型(每一个类型下含有多个应用实例)

Classification	Example Application
BULK	ftp
DATABASE	postgres, sqlnet, oracle, ingres
INTERACTIVE	ssh, klogin, rlogin, telnet
MAIL	imap, pop2/3, smtp
SERVICES	X11, dns, ident, ldap, ntp
WWW	www
P2P	KaZaA, BitTorrent, GnuTella
ATTACK	Internet worm and virus attacks
GAMEA	Half-Life
MULTIMEDIA	Windows Media Player, Real

数据集由 01 到 10 的 10 个小数据集组成，每个小数据集也是由一定数量的流构成的。10 个数据集的总体信息如表 3 所列。在整个数据集上如 WWW 的流数目是 328091，MAIL 的流数目是 28567，整个数据集所有流的数目为 377526。由表 3 可知类 INTTERACTIVE(INT)与类 GAMES 的流数量

很小，分别是 110 与 8。这些类由于流数量太小，没有提供足够的信息来进行分类，所以在流分类中将不考虑。

表 3　数据集的流数目统计结果

Total Flows	WWW	MAIL	BULK	SERV	DB
377526	328091	28567	11539	2099	2648
	INT	P2P	ATTACK	MMEDIA	GAMES
	110	2094	1793	1152	8

5.2　试验方案

在编号为 01 到 10 的 10 个数据集上我们进行了大量的实验，实验分为训练与测试两个部分。首先在 10 个数据集中的一个数据集上训练模型，然后把其它 9 个数据集作为测试集来测试在前面一个数据集上建立的分类器的性能。这样在每一个数据集上都建立模型，然后用其它 9 个数据集作为测试，就得到了 10 组数据，每一组数据对应一个分类器的性能。在我们的实验中，首先用 GATS-C4. 5 在训练集上选择特征子集，然后在此特征子集上建立分类器，最后用测试集测试分类器的性能，我们用到的分类器算法是 C4. 5[10]。分类器的性能主要由训练时间、测试速度、检测效果 3 个方面来评价，其中检测效果主要有两个指标：

准确率：被正确分类的流数目与流的总数目之比；

召回率：对每一类，被正确分类的流数目与这一类流的总数目之比。

从两个指标的定义可以看出，准确率是针对所有的类的衡量指标，而召回率是针对每一类的衡量指标。为了验证特征选择算法 GATS-C4. 5 的效果，我们对基于 GATS-C4. 5 的分类器在所有类，WWW，MAIL，P2P 上的性能进行了测试，并且把测试的结果与基于所有特征的分类器的性能进行了比较。为了从横向上比较 GATS-C4. 5 的性能，我们对基于 GATS-C4. 5 的分类器与 NBK-FCBF 在所有类，WWW，MAIL，P2P 上的检测率进行了对比。详细的实验结果将在 5. 4 和 5. 5 节介绍。所有的实验都在同一平台下完成，该平台的配置为：Intel processor 3. 0GHz，1. 00GB RAM，Windows 操作系统。

5.3　特征选择

GATS-C4. 5 作为特征选择算法，GATS 为搜索策略，C4. 5 为评估函数。先通过 GATS-C4. 5 对 01－10 的每一个数据集进行特征选择，然后在选择的特征上应用 C4. 5 算法建立分类器。表 4 是 10 个数据集特征选择的结果，其中 GATS-C4. 5 栏是 GATS-C4. 5 选择后的特征数目，FCBF 栏是 FCBF(Fast Correlation Based Filter)[4]选择后的特征数目。从表中

表 4　每一个训练集特征选择之后选择的特征数目

Training set	FCBF	GATS-C4. 5
01	4	7
02	3	6
03	7	7
04	7	9
05	11	6
06	5	4
07	15	9
08	16	8
09	28	12
10	49	8

可以看出，FCBF 与 GATS-C4. 5 对特征都有很大程度的削减，这样提高了分类器的测试速度。在 07－10 数据集上，GATS-C4. 5 选择的特征数目远远小于 FCBF 选择的特征数目。从这点可以看出，GATS-C4. 5 在测试速度与训练时间上可能要优于 FCBF。在后面的实验结果中我们也会对此进行论述。

5.4　特征选择在计算性能与分类能力上的效果

为了验证特征选择算法 GATS-C4. 5 的有效性，我们对基于所有特征的分类器与基于 GATS-C4. 5 的分类器在计算性能与分类能力上的效果进行了比较。在计算性能上，主要从建模时间与测试速度两个方面进行了对比；在分类能力上，主要从分类正确率与召回率上进行了对比。建模时间的对比如图 4 所示，图中记录了基于所有特征的分类器与基于 GATS-C4. 5 选择的特征的分类器在 10 个训练集上建立模型的时间。纵坐标的最高值 1 代表在我们测试平台下最长的建模时间 971s，其它的建模时间都相对于 971s 进行了标准化处理，所以纵坐标都小于 1。从图 4 可以看出，基于 GATS-C4. 5 的分类器建模时间相对于基于所有特征分类器的建模时间大大减少，这是因为特征的削减使得 C4. 5 分类器的结构简单化，这样建立这种结构的时间也就相应地减少了。

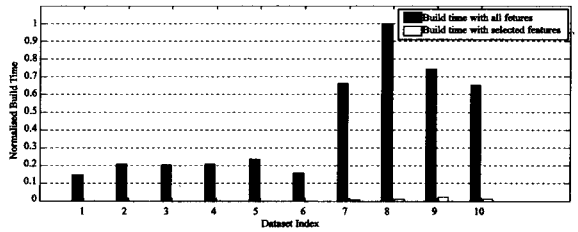


图 4　在 10 个训练集上基于所有特征的 C4. 5 分类器与基于选择特征的分类器训练时间对比图

图 5 提供了两种类型的分类器在测试速度上的区别。因有 10 个训练集，所以相应的两种类型的分类器都有 10 个。图中对每一个训练集上两种类型的分类器检测速度进行了对比。测试速度也经过标准化处理，最大的测试速度是每秒测试 50656 个流，这个速度标准化为 1，其它的测试速度相对于 50656 也进行了标准化，详细的比较如图 5 所示。基于 GATS-C4. 5 选择特征的分类器在测试速度上几乎是基于所有特征分类器的 1 倍。这是因为基于 GATS-C4. 5 选择特征的分类器在结构上比基于所有特征的分类器要精简，这样测试速度也就相应地跟上，这得力于特征选择的效应。

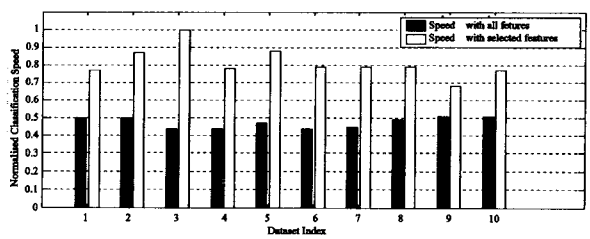


图 5　基于所有特征的 C4. 5 分类器与基于选择特征的分类器检测速度对比

上面详细比较了两种类型的分类器在计算性能上的差异。结果表明，基于 GATS-C4. 5 选择的特征的分类器在建模时间与测试速度上要远远优于基于所有特征的分类器。建

模时间与测试速度上的优势会不会导致分类器分类能力的下降呢？答案是否定的。图 6 到图 9 详细地展现了基于 GATS-C4.5 选择的特征的分类器与基于所有特征的分类器在分类能力上的区别。分类能力的比较主要从两方面进行：比较分类器在检测所有类别上的能力(图 6)；比较分类器在检测单个类别 WWW，MAIL，P2P 上的能力(图 7、图 8、图 9)。在图 6 中，基于 GATS-C4.5 选择特征的分类器在 10 组数据中，大部分点的数据要高于基于所有特征的分类器，并且它的平均准确率为 0.9899，要高于基于所有特征的分类器的平均准确率 0.9862，将近 0.4%的提高。这表明基于选择特征的分类器不仅在测试速度上有优势，同时在检测能力上也有一定的提高。可见特征选择能够剔除杂音与冗余特征，不仅精简了分类器的结构，同时提高了分类器的检测能力。

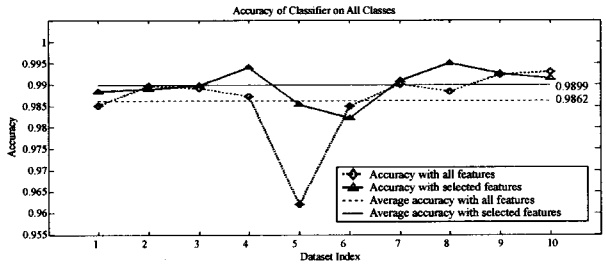


图 6　基于所有特征的 C4.5 分类器与基于 GATS-C4.5 选择特征的分类器在 WWW，MAIL，P2P 等所有类上召回率对比

图 7 到图 9 显示了基于 GATS-C4.5 选择特征的分类器与基于所有特征的分类器在 WWW，MAIL，P2P 上的召回率的比较。从这 3 幅图可以看出，基于特征选择的分类器在 WWW 上有更高的平均召回率(0.9982)，而在 MAIL(0.9966)，P2P(0.767)上的平均召回率要低于基于所有特征

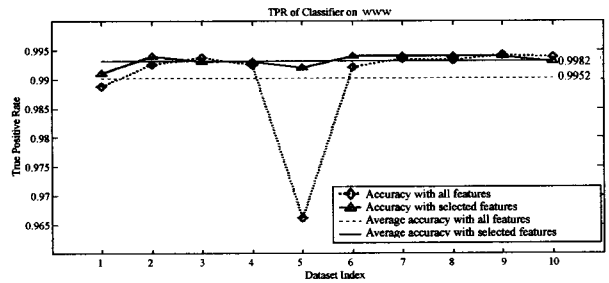


图 7　基于所有特征的 C4.5 分类器与基于 GATS-C4.5 选择特征的分类器在 WWW 上召回率对比

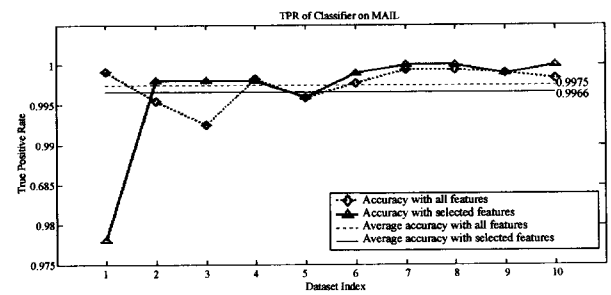


图 8　基于所有特征的 C4.5 分类器与基于 GATS-C4.5 选择特征的分类器在 MAIL 上召回率对比

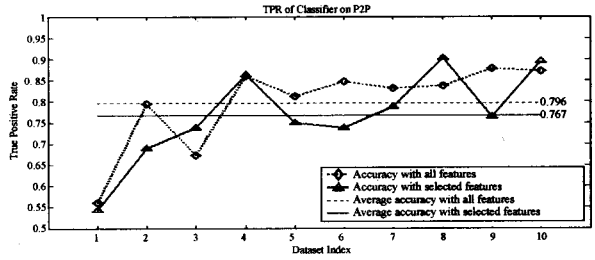


图 9　基于所有特征的 C4.5 分类器与基于 GATS-C4.5 选择特征的分类器在 P2P 上召回率对比

的分类器。WWW 上的平均召回率差不多增长 0.3%，在 MAIL，P2P 上的平均召回率下降将近 0.1%和 2.9%。特别是在 P2P 上下降比较多，但是在 P2P 上基于所有特征的分类器其召回率(79.2%)也不高，这说明数据集中的特征不能很好地反映出 P2P 这种应用。如果再在这样的基础上进行特征选择，效果可能会很差。虽然在平均召回率上有升有降，但是升降差别都不是很大，而在建模时间与测试速度上却有很大的提高，这说明特征选择在保证分类能力不减弱的情况下大大提高了分类器的时间性能。

5.5　GSTA-C4.5 与 NBK-FCBF 在计算性能与分类性能上的比较

5.4 节验证了特征选择算法的有效性，本节比较特征选择算法之间的性能。我们把基于 GATS-C4.5 选择特征的分类器与当前流行的基于特征选择的分类器 NBK-FCBF[4] 在计算性能与分类能力上进行了比较。Williams 在文献[3]中比较了分类器 C4.5 与分类器 NBK 的计算性能，结果表明 NBK 比 C4.5 有更短的建模时间，但是测试速度远小于C4.5。这是针对拥有相同数目特征的时候的结果对比，但是 GATS-C4.5 选择的特征数目比 FCBF 选择的特征数目要少，特别是在 07－10 四个数据集上，如表 4 所列。这表明在计算性能上 C4.5 要优于 NBK。在分类能力上我们也进行了比较，比较结果如图 10，图 11 所示。图 10 比较了基于 GATS-C4.5 的分类器与 NBK-FCBF 在所有类别，WWW，MAIL，P2P 上的分类能力，图 11 是基于 GATS-C4.5 选择特征的分类器在 ALL，WWW，MAIL，P2P 上超出 NBK-FCBF 的检测率。从图 10、图 11 可知，基于 GATS-C4.5 选择特征的分类器无论在 ALL，WWW，MAIL，P2P 上的检测率均高于 NBK-FCBF，并且平均高出将近 5%，特别是在 P2P 上高出 40%，这说明 GATS-C4.5 选择的特征子集效果好，特别是针对 P2P 其效果更好。

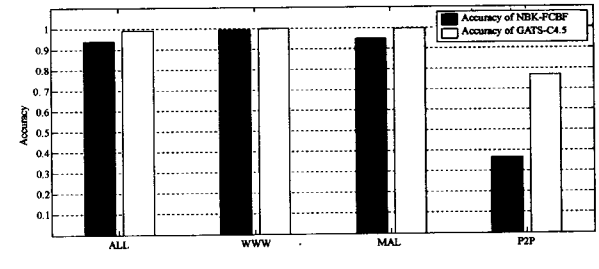


图 10　基于 GATS-C4.5 的 C4.5 分类器与 NBK-FCBF 在 ALL，WWW，MAIL，P2P 上的检测率对比

(下转第 96 页)

发现，当蠕虫爆发后的第 100 个时间片，蠕虫感染率 α 的估计值就已经稳定收敛到它的实际值 0.03，也就是说蠕虫检测算法在此时已经检测出了蠕虫的爆发。从预警时刻来看，此时间段处于蠕虫传播的第一阶段，即缓慢启动阶段。因此，使用该算法对蠕虫进行网络监测，根据监测结果及时预警，并采取相应的措施是有效并且可行的。

结束语 本文将卡尔曼滤波引入蠕虫检测，建立数学模型，给出相应的滤波方程。本算法具有实时递推性，可以对观测数据进行序贯处理。这种方法利用量测信息不断地修正估计值，而且无需保存过去的观测数据，大大减少了计算机的存储量和计算量，便于实时动态处理。仿真试验表明，采用卡尔曼滤波能够快速有效地检测出蠕虫的爆发。将来的改进工作包括抗差理论在卡尔曼滤波中的应用及对自适应参数设置的研究。

参 考 文 献

[1] Wen W P, Qing S H, Jiang J C, et al. Research and Development of Internet Worms [J]. Journal of Software, 2004, 15(8): 1208-1219

[2] Kruegel C, Vigna G. Anomaly Detection of Web-based Attacks [C] // Proc. of the 10th ACM Conference on Computer and Communications Security. Washington D C, USA, 2003: 251-261

[3] Zou C C, Gong W B, Towsley D, et al. The Monitoring and Early Detection of Internet Worms [J]. IEEE Transactions on Networking, 2005, 13(5): 961-974

[4] 陈博，方滨兴，云晓春. 基于最小二乘法的网络蠕虫检测方法 [J]. 哈尔滨工业大学学报，2007，39(3)：431-434

[5] 邓妍，戴冠中. 基于多监测点的蠕虫感染率的自回归估计[J]. 西北工业大学学报，2007，25(5)：657-661

[6] 秦永元. 卡尔曼滤波与组合导航原理[M]. 西安：西北工业大学出版社，2004

[7] David M, Colleen S N, Jeffery B. Code-Red: A Case Study on the Spread and Victims of an Internet Worm [EB/OL]. http://www.caida.org/publications/papers/2002/codered.pdf

(上接第 72 页)

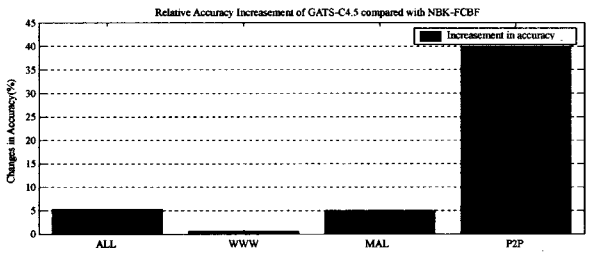


图 11 基于 GATS-C4.5 的 C4.5 分类器在检测准确率上相对于 NBK-FCBF 的百分比增长

结束语 IP 流分类在网络监控，QoS，入侵检测等领域应用广泛，为了能够快速地检测出系统的异常或者对某个特定结点进行实时监控，需要一种高效的 IP 流分类技术。网络的高速发展，使得网络中的数据量越来越大，并且这些数据中含有众多的相关与冗余信息，这使得对这些数据进行修剪与剔除变得尤为重要。流分类技术必须和数据预处理技术结合起来才能满足现代网络的需求，因此基于特征选择的流分类技术成为当前研究的热点。

针对传统特征选择算法的一些缺点，本文给出了一种 wrapper 型的特征选择算法 GATS-C4.5 来建立轻量级的流分类器，GATS-C4.5 不仅可以降低时间复杂度，而且可以获得最优的特征子集。基于 GATS-C4.5 的分类器具有更快的检测速度与更高的分类能力。我们在流数据集上进行了大量的实验，实验结果表明基于 GATS-C4.5 的流分类器在不降低分类能力的前提下可以大大降低分类器的计算复杂性，同时与 NBK-FCBF 相比，具有更快的检测速度与更高的分类能力。

在实验中我们发现 P2P 的召回率很低，在将来的研究中我们准备研究更好的特征选择算法与分类算法来解决这一问题。另外，将算法应用于入侵检测、网络监控等其它具体应用中也是我们进一步研究的内容。

参 考 文 献

[1] You Chen, Cheng Xue-qi, Li Yang, et al. Lightweight intrusion detection systems based on feature selection. Journal of Software, 2007, 18(7): 1639-1651

[2] Yu Lei, Liu Huan. Feature selection for high-dimensional data: A fast correlation-based filter solution // Proceedings of the Twentieth International Conference on Machine Learning (ICML 2003). 2003

[3] Williams N, Zander S, Armitage G. A Preliminary Performance Comparison of Five Machine Learning Algorithms for Practical IP Traffic Flow Classification. ACM SIGCOMM Computer Communication Review, 2006, 135(5)

[4] Moore A W, Zuev D. Internet Traffic Classification Using Bayesian Analysis Techniques // Proceedings of ACM SIGMETRICS. Banff, Canada, June 2005

[5] Zander S, Nguyen T T T, Armitage G. Automated Traffic Classification and Application Identification using Machine Learning // Proceedings of IEEE LCN. Australia, November 2005

[6] Holland J. Adaptation in natural and artificial systems. The University of Michigan Press, 1975

[7] Glover F. Future paths for integer programming and links to artificial intelligence. Computers and Operations Research, 1986, 13(4): 533-549

[8] Müblenbein H. Parallel generic algorithms in combinatorial optimization. Computer Science and Operations Research (Edited by Osman Balci), Oxford: Pergamon Press, 1995

[9] Glover F, Kelly J, Laguna M. Genetic algorithms and tabu search: hybrids for optimizations. Computers and Operations Research, 1995, 22(1): 111-134

[10] Quinlan J R. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, 1993

[11] The data-sets for flow classification. http://www.cl.cam.ac.uk/research/srg/netos/nprobe/data/papers/sigmetrics/index.html

[12] Moore A W, Hall J, Kreibich C, et al. Architecture of a Network Monitor // Passive and Active Measurement Workshop 2003 (PAM2003). La Jolla, CA, April 2003

[13] Moore A W, Zuev D. Discriminators for use in flow-based classification. Technical report. Intel Research, Cambridge, 2005

[14] Moore A W. Discrete content - based classification. a data set. Technical report. Intel Research, Cambridge, 2005