doi:10.3969/j.issn.1673-4785.2011.03.010

结合粗糙集和禁忌搜索的网络流量特征选择

顾成杰[1],张顺颐[1],杜安源[2]
(1.南京邮电大学 信息网络技术研究所,江苏 南京 210003;2.中国移动通信集团设计院有限公司 安徽分公司,安徽 合肥 230041)

摘 要:针对网络流量特征属性的优化选择问题,提出了一种结合粗糙集和禁忌搜索的网络流量特征选择方法(RS-TS).该方法通过粗糙集算法对网络流量特征属性进行约简,将所得到的特征子集作为禁忌搜索的初始解,并利用禁忌搜索得到最优特征子集.实验验证RS-TS方法优于基于GA的特征选择方法和基于IG的特征选择方法,能够有效地去除网络流量的冗余特征属性,提高网络流量分类精度.

关键词:粗糙集;禁忌搜索;特征选择;网络流量
中图分类号:TP391 **文献标识码**:A **文章编号**:1673-4785(2011)03-0254-07

Feature selection of network traffic using a rough set and tabu search

GU Chengjie[1], ZHANG Shunyi[1], DU Anyuan[2]
(1. Institute of Information Network Technology, Nanjing University of Posts and Telecommunications, Nanjing 210003, China; 2. Anhui Branch, China Mobile Group Design Institute, Hefei 230041, China)

Abstract: A feature selection of network traffic using a rough set and tabu search (RS-TS) was proposed for the purpose of optimization in the feature selection of traffic classification. This approach reduced the traffic feature attribute with a rough set and established the feature subset as the initial value of a tabu search, as well as the optimal feature subset on the basis of a tabu search. The optimal feature subset with a tabu search can be selected on the basis of a feature subset. In contrast with the traditional feature selection methods based on GA and IG, RS-TS was validated optimal by experimental results. It can diminish redundant feature attribution of network traffic effectively and greatly improve the classification accuracy.

Keywords: rough set; tabu search; feature selection; network traffic

网络流量的精确分类是网络流量建模、用户行为分析、网络流量控制等行为的重要前提[1].随着互联网流量日趋复杂和动态多变,越来越多的网络应用采用动态端口、伪装端口和应用层加密等规避手段,使得利用机器学习方法进行网络流量分类成为一个重要的研究方向[2].在使用机器学习方法处理网络流量分类问题时,研究对象是网络流量的流统计特征属性,一般情况下,特征属性越多,就越能精确地区分不同的网络流量[3];但网络流量中也存在很多冗余特征属性和弱特征属性,过多特征属性有时甚至会影响学习质量,导致分类精度下降,计算复杂度增加,计算时间延长[4].特征选择是在保持原有网络流量完整性的基础上,去除其中的冗余特征属性或弱特征属性,同时维持分类精度的过程[5].因此,寻找一种高效的网络流量特征选择方法以降低特征空间维数,提高网络流量分类的精度成为本文要解决的重要问题.

特征选择作为数据预处理的一类方法,是数据挖掘、机器学习和模式识别的重要方法[6].目前有基于信息熵、神经网络、粗糙集和遗传算法等各种特征选择方法.Kira和Rendell在文献[7]提出了一种基于特征相关性的Relief算法,能够有效地进行特征选择,但Relief算法只能使用在二值分类环境中,

收稿日期:2010-08-28.
基金项目:国家"863"计划资助项目(2006AA01Z232,2009AA01Z212,2009AA01Z202);江苏省自然科学基金资助项目(BK2007603);江苏省高技术研究计划资助项目(BG2007045);江苏省重大科技支撑计划资助项目(BE2008134);江苏省科技成果转化专项基金资助项目(BA2007012).
通信作者:顾成杰. E-mail:jackiee.gu@gmail.com.

而且不能识别出冗余特性. Li 等人利用遗传算法进行特征属性选择,具有搜索能力强等优点,适合于大规模复杂问题的求解,但容易过早收敛[8]. 文献[9]中将粒子群算法用于特征搜索,解决了遗传算法存在的计算量大的缺点;但算法中粒子被束缚在局部最优点及整个种群全局最优点附近,导致其搜索区域有限,易于陷入局部最优. 文献[10]中提出了一种基于免疫克隆算法的特征选择方法,该方法利用免疫克隆算法能够快速收敛于全局最优的特性,加快了搜索到最优特征子集的收敛速度,但存在求解过程中亲和度的计算工作量较大的缺点.

目前在网络流量特征选择研究方面,大多是采用已成熟的方法进行网络流量特征选择,但由于真实环境中的网络流量存在大量噪声,根据网络流量统计特征分析得到的特征属性数目庞大,同时也存在大量冗余特征属性,所以使用现有算法进行网络流量特征选择还存在一定局限性. 笔者在分析网络流量特性的基础上,提出一种结合粗糙集和禁忌搜索的网络流量特征选择方法,通过粗糙集算法对特征属性进行约简,将得到的特征子集作为禁忌搜索的初始解,利用禁忌搜索能够得到更广的搜索空间,从而寻找到全局最优解,得到最优特征子集,该方法能够较好地降低特征维度,提高分类性能.

1　结合粗糙集和禁忌搜索的网络流量特征选择方法

1.1　基于粗糙集的特征属性约简

粗糙集理论(rough set, RS)是波兰数学家 Paulat 提出的一种处理模糊和不确定知识的数学工具,其中特征约简是粗糙集理论的核心问题之一[11-12]. 特征约简的实质就是找到冗余特征属性. 在特征集中去掉冗余特征后的特征集称为最小特征子集,它具有与全部条件特征属性同样的区分决策能力[13].

给定一个四元组的信息系统 $S=(U,A,V,f)$,其中:U 是给定网络流量的数据样本集,为一个非空的有限集合;$A=C\cap D$,C 是网络流量样本中抽取的 n 个条件特征属性集合,$D=\{d\}$ 为决策属性集,即 A 表示网络流量样本的特征属性集合;$V=\cup V_r$ 是特征的取值范围构成的集合,其中 V_r 是特征 r 的值域;f 是 $U\times A\to V$ 的映射函数,它为每一个样本的每个特征属性赋予一个属性值,即 $\forall a\in A, x_i\in U$, $f(x_i,a)\in V_a$.

定义 1　最小特征子集 P　设 C、D 分别是信息系统 S 的条件特征属性集和决策特征属性集,特征属性集 $P(P\subseteq C)$ 是 C 的一个最小特征属性集,当且仅当 $\gamma(P,D)=\gamma(C,D)$,并且 $\forall P'\subset P,\gamma(P',D)\neq\gamma(C,D)$,则说明 P 是 C 的最小属性集,P 具有和 C 同样的区分决策能力.

定义 2　区分矩阵 G　信息系统 $S=(U,A,V,f)$ 可以用一个 $|U|\times|U|$ 的对称矩阵来表示,并对矩阵的每一项定义为:

$$C_{ij}=\begin{cases}\{a\in A\mid a(x_i)\neq a(x_j)\}, & D(x_i)\neq D(x_j);\\ \varnothing, & \text{其他.}\end{cases}$$

本文采用基于区分矩阵的启发式粗糙集约简算法,该算法利用特征属性在区分矩阵中出现的频率 $f(a_i)$ 作为启发规则,寻找最小特征子集. 特征属性的出现频率定义为

$$f(a_i)=f(a_i)+k\times\frac{|C|}{|B|}.$$

式中:$|C|$ 表示数据集中条件特征属性的个数;$|B|$ 表示区分矩阵 G 中每项所包含的特征属性个数;k 是权重系数,本文默认 $k=1$.

若某个特征属性在区分矩阵中出现的次数越多,则该特征属性的重要性就越大;若某个特征属性所出现的区分矩阵的项越短,则该特征属性的重要程度就越大,算法描述如下.

1)初始化算法参数,候选最小特征属性子集 $R=\varnothing$,每个条件特征属性在区分矩阵的出现频率 $f(a_i)=0$.

2)计算区分矩阵 G 的每项 m,同时计算各特征属性的加权频率 $f(a_i)$.

3)合并并按照特征属性的长度排序区分矩阵 G.

4)对区分矩阵 G 的每项 m 执行:

如果 $m\cap R==\varnothing$,选择 m 中 $f(a_i)$ 最大的特征属性 a_i,$R=R\cup\{a_i\}$.

5)返回得到的最小特征子集 R,算法结束.

1.2　禁忌搜索算法及相关参数设置

禁忌搜索(tabu search, TS)是 Glover 提出的一种全局搜索算法. 它模仿人类的记忆功能,在求解问题的过程中,采用禁忌技术对已经搜索过的局部最优解进行标记,并且在迭代中尽量避免重复的搜索,从而获得更广的搜索区间来寻找到全局最优解[14]. 但是它对初始解有较强的依赖性,好的初始解可使禁忌搜索在解空间中搜索到最优解,并且收敛速度快[15].

禁忌搜索算法首先确定一个初始可行解 x,并且对于每一个解 x 定义一个邻域 $N(x)$. 确定完初始可行解后,定义可行解 x 的邻域移动集 $s(x)$,然后从邻域移动集中挑选一个能改进当前解 x 的移动 $s(x)$,再从新解 x' 开始,重复搜索,如果邻域移动集中只接受比当前解 x 好的解,搜索就可能陷入循环的危险. 为避免陷入循环和局部最优,需要构造禁忌表(tabu list),禁忌表中存放刚刚进行过的 n 个邻域移动. 对于当前的移动,在以后的 n 次循环内是禁止的,以避免回到原先的解,n 次循环以后释放该移动. 即使引入了禁忌表,禁忌搜索算法仍有可能出现循环,因此还必须给定停止准则以避免算法出现循环,当在规定迭代次数内所发现的最优解无法改进或无法离开它时,则算法停止[16].

根据以上分析,为了利用禁忌搜索算法解决网络流量特征选择问题,需要解决解的表示、邻域结构、禁忌长度、禁忌表更新策略、终止条件等问题.

1)解的表示. 设解向量为 $(x_1, x_2, \cdots, x_i, \cdots, x_n)$,采用二进制编码形式表示,则解向量用 0/1 二进制位串来表示,1 表示第 i 个特征在最优特征子集中,0 表示第 i 个特征不在最优特征子集中.

2)初始解. 由于网络流样本存在很多冗余特征属性,为了解决禁忌搜索对初始解的依赖性,采用粗糙集算法来获得禁忌搜索所需要的初始解.

3)邻域结构. 用于实现邻域搜索,将初始解 x 的邻域结构定义为每次改变一个特征的状态,即每次加入 1 个特征属性或者减掉 1 个特征属性.

4)禁忌长度. 它表示被禁忌对象不允许被选取的迭代次数,体现了算法的短期记忆能力,禁忌长度越长,表示搜索范围越广泛,获得最优解的可能性越大. 本文设定禁忌长度 $n=10$.

5)禁忌表更新策略. 采用队列来记录禁忌表的操作,每次迭代将禁忌对象记录在队首,队列溢出则将排在队尾的禁忌对象删除.

6)特赦准则. 为了避免遗失全局最优解,如果某个解处于禁忌表中,但该解优于当前最优解,可以将该解重新选为新的解.

7)终止条件. 本文采用最大迭代次数 m_1 和最大无改进次数 m_2 作为终止条件,当禁忌搜索迭代到 m_1 次时,停止搜索;当通过 m_2 次搜索最优解没有改进时,停止搜索.

1.3 新的网络流量特征选择方法

网络流量特征选择目的是减少特征数以降低分类器的计算复杂度,同时维持分类精度. 网络流量的流统计特征属性反映了业务流的本质和度量,能以此来区分不同的业务流. 一般情况下,特征越多,越能精确地区分不同的业务;但获得特征数据不仅需要大量的测量时间,而且样本存储占用的空间,尤其当训练集中有很多流时,就会大大增加,不适于用在实时分类中. 另外过多、过细的网络流量特征虽然也能在一定程度上起到提高分类精度的作用,但这是以付出计算复杂度为代价,而且过多特征(或有些特征)甚至会影响学习质量,导致分类精度下降,计算复杂度增加,计算时间延长. 所以挖掘出最本质的信息,提取最能刻画样本数据的最少特征是非常重要的. 但最优特征子集选择是一个 NP 难问题,所以本文目的在于寻找一个较好的解决这个问题的近似方法. 首先基于粗糙集对网络流特征属性进行约简,剔除对分类起干扰作用的冗余特征属性,解决禁忌搜索对初始解依赖性强的问题. 然后将所得到的特征子集作为禁忌搜索的初始解,利用禁忌搜索得到最优特征子集,算法步骤如下.

1)初始化参数,禁忌长度 $n=10$,最大迭代次数 $m_1=500$,最大无改进次数 $m_2=50$.

2)使用粗糙集得到网络流量特征子集,作为禁忌搜索的初始解.

3)将网络流量特征子集用二进制编码形式表示,生成初始解,并且将禁忌表置空.

4)判断是否满足终止条件,若满足,转步骤 9);若不满足,转步骤 5).

5)利用初始解的邻域函数获得所有的邻域解,通过计算各个解的评价值,得到若干候选解.

6)判断是否满足特赦准则,若满足,转步骤 7);若不满足,转步骤 8).

7)将满足特赦准则的解作为当前解,其对应的对象替换最早进入禁忌表中的对象,更新最优解,转步骤 4).

8)计算候选解对应的各对象的禁忌属性,选择候选解中非禁忌对象的最优状态作为新的当前解,并用该对象替换最早进入禁忌表中的对象,转步骤 4).

9)结束算法,输出最优网络流量特征子集.

该方法将粗糙集和禁忌搜索相结合,通过粗糙集算法对特征属性进行约简,将得到的特征子集作为禁忌搜索的初始解,利用禁忌搜索能够在更广的搜索空间中寻找到全局最优解的特性,得到最优特征子集,降低了特征维度,加快了收敛速度,同时提高了后续分类性能. 其流程如图 1 所示.

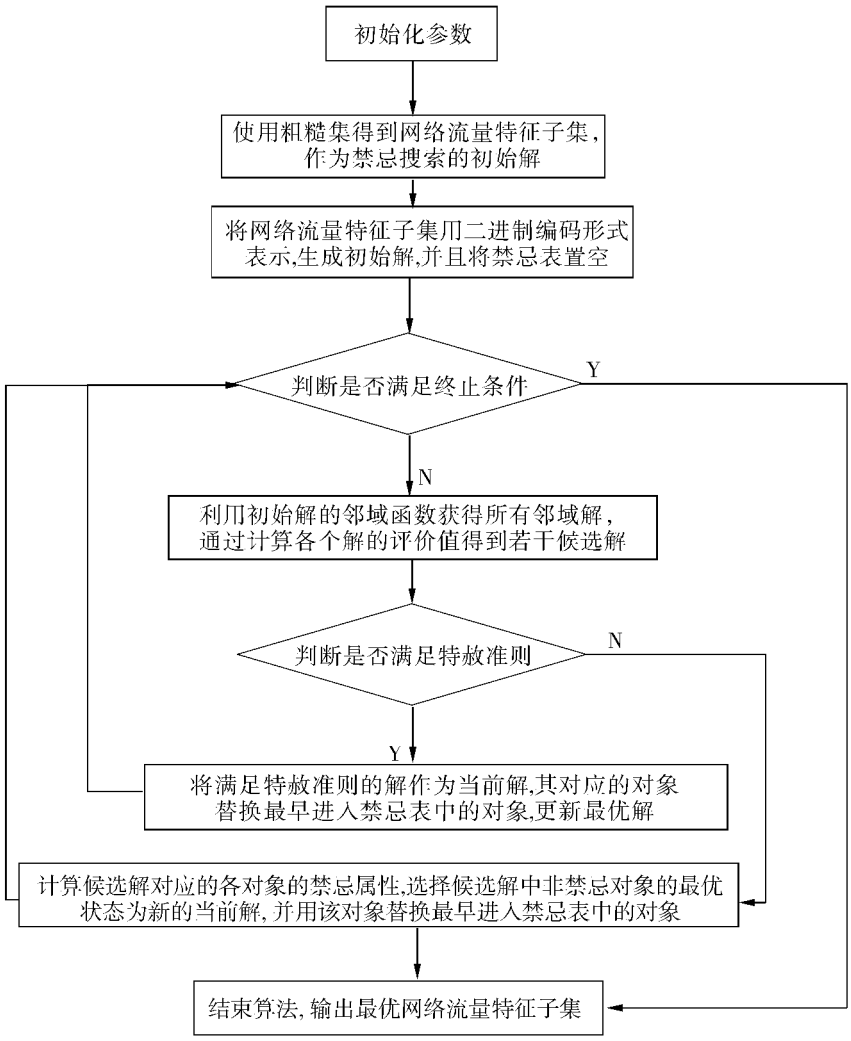


图 1　结合粗糙集和禁忌搜索的网络流量特征选择方法流程

Fig. 1　The flow chart of feature selection of network traffic using rough set and tabu search

2　实验验证

2.1　实验数据集

为了验证所提出的特征选择算法的有效性，通过实验来验证. 实验数据集的选择非常重要，经过分析和比较，采用剑桥大学 Moore 教授等人使用的网络流量实验数据集[17]，记为 Moore_Set. 该数据集采自 2003 年 8 月 20 日 0 时至 24 时流经某生物学研究所网络出口的双向网络流量，通过采样提取出 10 个平均抽样时间大约是1 680 s的子集，形成实验数据集. 该实验数据集中共包含了377 526个网络流样本，被分成 10 种类型，每种类型所包含的应用名称、每类网络流的数量和所占比例见表 1.

Moore_Set 中每条网络流样本都是从一条完整的 TCP 双向流抽象而来，包含 248 项属性，其中第 1 项属性和第 2 项属性分别是该流的源端口号和目的端口号. 为了避免端口信息对分类的影响，本文没有采用这 2 个属性. 使用粗糙集进行特征属性约简时，要求属性值用离散数据表示，因此还需要对各个属性值进行离散化.

表 1　实验数据集统计信息

Table 1　Statistics of Moore_Set

类　别	应用名称	流数目	所占比例/%
WWW	http, https	328 091	86.910
MAIL	Imap, pop3, smtp	28 567	7.567
BULK	ftp	11 539	3.056
DATABASE	oracle, mysql	2 648	0.701
SERVER	ident, ntp, x11, dns	2 099	0.556
P2P	kazaa, bittorrent	2 094	0.555
ATTACK	worm, virus	1 793	0.475
MEDIA	real, media player	1 152	0.305
INT	telnet, ssh, rlogin	110	0.029
GAME	half-life	8	0.002
总计	26 种	377 526	100

2.2 实验工具

实验所使用的数据挖掘工具是 Weka-3.5.8,该软件由新西兰 Waikato 大学所开发,可以在代码中调用,也可以直接使用. 本文采用 Matlab 7.0 用于数值计算,实验采用普通 PC 机,操作系统为 Windows XP Professional SP2, 其中 CPU 为 Intel Pentium 2.66 GHz,内存为 DDR-667 2GB.

2.3 评估策略与所采用的分类器

为了评估所提出方法的性能,同时便于与其他特征选择方法进行对比,采用反馈率(recall)和准确率(precious)作为衡量算法有效性的指标.

$$\text{反馈率} = \frac{\text{检索到的相关样本数量}}{\text{相关样本数量}} = \frac{N_{tp}}{N_{tp}+N_{fn}} \times 100\%,$$

$$\text{准确率} = \frac{\text{检索到的相关样本数量}}{\text{检索到的样本数量}} = \frac{N_{tp}}{N_{tp}+N_{fp}} \times 100\%.$$

式中:N_{tp}、N_{fn}、N_{fp} 含义定义如下:

1)N_{tp} (true positive):实际类型为 i 的样本被分类模型正确分类的样本数量;

2) N_{fn} (false negative):实际类型为 i 的样本被分类模型错误分类的样本数量;

3) N_{fp} (false positive):实际类型为非 i 的样本被分类模型分为 i 类的样本数量.

实验采用 C4.5 决策树分类器来对比各特征选择算法,因为 C4.5 决策树分类器在进行样本分类时,仅需要根据网络样本流特征属性自顶向下进行比较,处理相对简单,同时不依赖于网络流量样本的先验概率,能有效地避免网络流样本变化所带来的影响,具有较高的分类效率.

2.4 实验结果分析

为了验证结合粗糙集和禁忌搜索的特征选择方法(RS-TS)的正确性,在 Moore_Set 上运用 RS-TS 方法进行网络流量特征属性选择,得到相应的最优特征子集如表 2 所示. 从表中可以看出,网络流量样本具有大量特征属性,但同时也存在很多冗余特征属性和弱相关特征属性,经过 RS-TS 方法特征选择后,从 246 个特征属性中得到包含 23 个特征属性的最优特征子集,最优特征子集的特征个数仅为全部特征个数的9.35%. RS-TS 方法能够有效地获得对流量分类能够发挥更大权重的特征属性,同时也避免了分类精度受一些弱相关特征或冗余特征的影响.

表 2 运用 RS-TS 方法后得到的最优特征子集

Table 2 Optimal feature subset with RS-TS

序号	特征描述	缩写
1	Duration of the flow	duration
2	Number of packets in forward direction	fpkts
3	Number of packets in backward direction	bpkts
4	Number of bytes in forward direction	fbytes
5	Number of bytes in backward direction	bbytes
6	Minimum forward packet length	minfpktl
7	Mean forward packet length	meanfpktl
8	Maximum forward packet length	maxfpktl
9	Minimum backward packet length	minbpktl
10	Mean backward packet length	meanbpktl
11	Maximum backward packet length	maxbpktl
12	Minimum forward inter-arrival time	minfiat
13	Mean forward inter-arrival time	meanfiat
14	Maximum forward inter-arrival time	maxfiat
15	Minimum backward inter-arrival time	minbiat
16	Mean backward inter-arrival time	meanbiat
17	Maximum backward inter-arrival time	maxbiat
18	Minimum of active flow	minaf
19	Mean of active flow	meanaf
20	Maximum of active flow	maxaf
21	Minimum of idle flow	minif
22	Mean of idle flow	meanif
23	Maximum of idle flow	maxif

在相同的实验数据集上分别基于全部特征属性和最优特征属性训练 C4.5 决策树分类器,并利用得到的 C4.5 决策树分类器进行网络流量分类,表 3 是特征选择前后 C4.5 决策树分类器分类精度的对比结果.

表 3 全部特征属性与最优特征子集的分类精度对比

Table 3 Comparison of classification accuracy between full feature and optimal feature subset %

类 别	全部特征属性		最优特征子集	
	反馈率	准确率	反馈率	准确率
WWW	93.58	96.65	97.67	98.19
MAIL	92.69	93.28	96.82	94.07
BULK	85.23	85.83	88.96	86.85
DATABASE	90.72	90.18	91.07	90.37
SERVER	90.85	89.29	93.73	91.32
P2P	83.28	81.95	89.84	85.72
ATTACK	88.74	82.41	85.82	83.03
MEDIA	93.87	86.73	94.62	87.46
INT	85.36	81.59	87.18	83.92
GAME	82.03	79.65	88.22	83.07

从表 3 可以得出,对于 P2P 类别来说,使用最优特征子集比使用全部特征属性分类的准确率提高了 4.60%,从所有业务的平均值来看,分类准确率

提高了 1.89%. 由此可以得到,利用最优特征子集训练出来的分类器具有更高的分类精度. 实验结果表明,基于 RS-TS 的特征选择方法不仅降低了特征维数,而且提高了分类精度. 其原因在于,由于 Moore_Set 包含的 246 项网络流特征属性中,存在众多冗余属性和弱相关属性,而基于 RS-TS 的特征选择方法首先通过粗糙集进行特征属性的约减,在保证分类精度的同时极大地降低了特征空间的维度,并在此基础上使用禁忌搜索,得到网络流量最优特征子集,大大减弱了冗余特征对分类精度的影响.

为了说明结合粗糙集和禁忌搜索的特征选择方法的有效性,本文选择文献[8]所使用的基于 GA 的特征选择方法、文献[18]所使用的基于 IG 的特征选择方法与 RS-TS 方法对后续网络流量分类精度的影响进行对比试验. 在这 3 种特征选择方法后都采用 C4.5 分类器对网络流量进行分类,试验结果如表 4 所示.

表 4　不同特征选择方法对分类精度的影响

Table 4　Comparison of classification accuracy between different feature selection algorithms　%

类　别	RS-TS		CA		IG	
	反馈率	准确率	反馈率	准确率	反馈率	准确率
WWW	97.67	98.19	94.72	93.28	93.36	91.97
MAIL	96.82	94.07	93.34	92.92	90.12	89.25
BULK	88.96	86.85	87.09	83.34	79.89	81.29
DATABASE	91.07	90.37	89.63	88.03	84.47	83.85
SERVER	93.73	91.32	87.78	85.24	82.81	80.73
P2P	89.84	85.72	82.22	80.16	79.42	78.51
ATTACK	85.82	83.03	87.07	81.81	77.93	75.04
MEDIA	94.62	87.46	89.26	84.54	82.01	84.69
INT	87.18	83.92	83.19	79.93	80.36	77.73
GAME	88.22	83.07	79.27	76.78	74.17	69.94

表 4 可以得出,从所有类别的平均值来看,使用 RS-TS 方法比使用基于 GA 的特征选择方法的分类反馈率提高了 4.62%,比使用基于 IG 的特征选择方法的分类反馈率提高了 10.84%. 由此可以得出,RS-TS 较 GA、IG 方法可以得到较高的反馈率和准确率,说明了 RS-TS 方法的有效性. 其原因在于,相比于 RS-TS 方法,基于 GA 的特征选择方法容易过早收敛,很难在特征空间中获得全局最优解;基于 IG 的特征选择方法易受样本分布的影响,而实际网络环境中各种网络应用并不是均衡分布,所以在样本类别分布不均匀的网络环境下,基于 IG 的特征选择方法的分类精度严重受到影响. RS-TS 方法结合了粗糙集和禁忌搜索的优点,一方面通过粗糙集可以消除网络流量中大量的冗余特征属性,同时为搜索最优解提供较优的初始解,另一方面利用禁忌搜索避免陷入“局部最优”,保证能够在特征子集空间中找到全局最优解.

3　结束语

针对网络流量中存在大量冗余特征或弱特征,从而导致分类精度下降的问题. 根据粗糙集和禁忌搜索各自的优点,提出了一种结合粗糙集和禁忌搜索的网络流量特征选择方法(RS-TS). 实验表明,该方法在实际网络流量中可以取得优于基于 GA 的特征选择方法和基于 IG 的特征选择方法的效果. 该方法一方面利用粗糙集进行特征属性约简可以大大减少冗余特征属性的数量,解决禁忌搜索对初始解依赖性强的问题;并在此基础上使用禁忌搜索,可以加快收敛速度,较快地获得最优特征子集. 如何减小 RS-TS 的计算开销将是下一步研究的工作重点.

参考文献:

[1] NGUYEN T T, ARMITAGE G. A survey of techniques for internet traffic classification using machine learning[J]. IEEE Communications Surveys and Tutorials, 2008, 10(4): 56-76.

[2] CALLADO A, KAMIENSKI C. A survey on internet traffic identification[J]. IEEE Communications Surveys and Tutorials, 2009, 11(3): 37-52.

[3] ESTE A, GRINGOLI F, SALGARELLI L. Support vector machines for TCP traffic classification[J]. Computer Networks, 2009, 53(14): 2476-2490.

[4] POLAT K, GUNES S. A new feature selection method on classification of medical datasets: kernel F-score feature selection[J]. Expert Systems with Applications, 2009, 36(7): 10367-10373.

[5] PATRICIA E N, ENGELBRECHT A P. A decision rule-based method for feature selection in predictive data mining [J]. Expert Systems with Applications, 2010, 37(1):

602-609.

[6] BARALDI P,PEDRONI N,ZIO E. Application of a nicked Pareto genetic algorithm for selecting features for nuclear transients classification[J]. International Journal of Intelligent Systems,2009,24(2):118-151.

[7] KIRA K,RENDELL L A. Feature selection problem: traditional methods and a new algorithm[C]//Proceedings of the Tenth National Conference on Artificial Intelligence. San Jose,USA,1992:129-134.

[8] LI Yongming,ZHANG Sujuan,ZENG Xiaoping. Research of multi-population agent genetic algorithm for feature selection[J]. Expert Systems with Applications,2009,36(7):11570-11581.

[9] HUANG Chenglung,DUN Jianfan. A distributed PSO-SVM hybrid system with feature selection and parameter optimization[J]. Applied Soft Computing Journal,2008,8(4):1381-1391.

[10] ZHANG Li,MENG Xiangru,WU Weijia,et al. Network fault feature selection based on adaptive immune clone selection algorithm [C]//Proceedings of the 2009 International Joint Conference on Computational Sciences and Optimization. Sanya,China,2009:969-973.

[11] SWINIARSKI R W,SKOWRON A. Rough set methods in feature selection and recognition[J]. Pattern Recognition Letters,2003,24(6):833-849.

[12] 殷志伟,张健沛.基于浓缩布尔矩阵的属性约简算法[J].哈尔滨工程大学学报,2009,30(3):307-311.
YIN Zhiwei,ZHANG Jianpei.An attribute reduction algorithm based on a concentration Boolean matrix[J]. Journal of Harbin Engineering University,2009,30(3):307-311.

[13] THANGAVELK,PETHALAKSHMI A. Dimensionality reduction based on rough set theory: a review[J]. Applied Soft Computing Journal,2009,9(1):1-12.

[14] WANG Yong,LI Lin,N Jun,et al. Feature selection using tabu search with long-term memories and probabilistic neural networks[J]. Pattern Recognition Letters,2009,30(7):661-670.

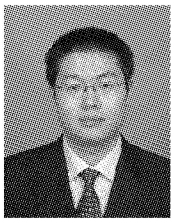
[15] HEDAR A R,WANG J,FUKUSHIMA M. Tabu search for attribute reduction in rough set theory[J]. Soft Computing,2008,12(9):909-918.

[16] CHIOU Chiewun,WU Muhcheng. A GA-tabu algorithm for scheduling in-line steppers in low-yield scenarios[J]. Expert Systems with Applications,2009,36(9):11925-11933.

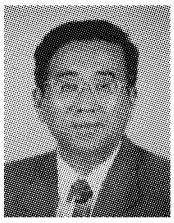
[17] MOORE A W,ZUEV D. Internet traffic classification using Bayesian analysis techniques[C]//International Conference on Measurement and Modeling of Computer Systems. New York,USA,2005:50-60.

[18] YANG Y M,PEDERSEN J O. A comparative study on feature selection in text categorization[C]//Proceedings of the Fourteenth International Conference on Machine Learning. Nashville,USA,1997:412-420.

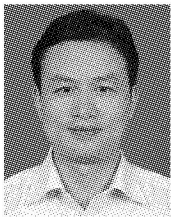
作者简介:



顾成杰,男,1985年生,博士研究生,主要研究方向为通信网与IP技术、P2P网络、网络流量识别与认知网络.



张顺颐,男,1944年生,教授,博士生导师,江苏省通信与网络工程技术研究中心主任,中国通信学会IP应用与增值电信业务专委会主任,中国电子学会通信学分会副主任.主要研究方向为计算机网络通信、下一代网络与IP技术、互联网络监测与管理.近年来先后主持完成国家"863"计划项目5项,国家科技支撑计划项目1项,江苏省重大科技成果产业化项目1项,江苏省高技术研究计划重点项目2项.获得省部级科技进步奖8项,完成专利申请30余项,发表学术论文100余篇.



杜安源,男,1979年生,工程师,主要研究方向为移动通信与无线资源管理、3G标准的应用性研究、移动通信技术设计与实施等.