doi: 10.3969/j.issn.1005-3026.2014.11.003

基于互信息和文化基因算法的网络流量特征选择

苗长胜[1], 原常青[1], 王兴伟[1], 常桂然[2]
(1. 东北大学 信息科学与工程学院, 辽宁 沈阳 110819; 2. 东北大学 计算中心, 辽宁 沈阳 110819)

摘 要: 利用文化基因框架的引导,提出一种结合了封装和过滤的混合型特征选择算法. 该算法在传统的遗传算法中采用了基于互信息的局部搜索算法,全局搜索以分类器精度为适应度函数,保证得到全局最优解;局部搜索以联合互信息为评价指标,加快了寻找最优特征子集的收敛速度. 实验表明,与现有算法相比,该算法在特征数量和计算复杂度上有显著改进,采用该算法的网络流量识别方法能以更少的特征获得更高的分类精度.

关 键 词: 互信息;文化基因算法;特征选择;流量识别
中图分类号: TP 393.07 **文献标志码**: A **文章编号**: 1005-3026(2014)11-1530-05

A Hybrid Feature Selection Algorithm Based on Mutual Information and Memetic Framework to Optimize Traffic Classification

MIAO Chang-sheng[1], *YUAN Chang-qing*[1], *WANG Xing-wei*[1], *CHANG Gui-ran*[2]
(1. School of Information Science & Engineering, Northeastern University, Shenyang 110819, China; 2. Computing Center, Northeastern University, Shenyang 110819, China. Corresponding author: MIAO Chang-sheng, E-mail: csmiao@126.com)

Abstract: Under the memetic framework, a new feature selection method combining filter and wrapper models was proposed. In the hybrid algorithm, classifier accruracy was used as fitness function to ensure global optimization, while joint mutual information was used as evaluation indicator to accelerate the process. The experimental results indicated that the proposed method outperformed the existing methods in computational efficeicecy and number of selected features. Applying this algorithm to traffic classification resulted in the improved accuracy with fewer features.

Key words: mutual information; memetic framework algorithm; feature selection; traffic classification

基于流统计特征的网络流量分类方法[1-3]根据流的一系列统计特征,如流持续时间、流平均包大小等,利用机器学习算法将流分类. 由于该方法具有保护用户隐私、识别加密和新型网络应用等优势,已成为当前网络流量识别领域的研究热点.

针对流量分类问题,Moore 等[4]给出了网络流量的248个特征. 流统计特征的合理选取对分类器效率和准确性有较大影响. 当前网络流量特征选择算法主要分为过滤型特征选择算法[5-6]和封装型特征选择算法两种. 特征选择是一个NP难组合优化问题[7],过滤型特征选择算法容易陷入局部最优,并不一定能够给出最优特征子集. 封装型特征选择算法[8-9]直接用分类器的分类性能作为特征子集的评估准则,虽然选择出的特征子集质量更高,但算法的复杂度太高,执行时间偏长,不适应高速网络环境下实时性的要求.

本文结合过滤型特征选择算法和封装型特征选择算法的优点,提出一种混合型特征选择算法,利用文化基因框架引导特征的选择,在传统的进化算法中结合局部提升算法.

收稿日期: 2013-09-13
基金项目: 国家自然科学基金资助项目(71071028, 70931001); 教育部高等学校博士学科点专项科研基金资助项目(20120042130003); 中央高校基本科研业务费专项资金资助项目(N120104001).
作者简介: 苗长胜(1981-),男,山东烟台人,东北大学博士研究生; 王兴伟(1968-),男,辽宁盖州人,东北大学教授,博士生导师; 常桂然(1946-),男,河北曲周人,东北大学教授,博士生导师.

1　相关技术

1.1　文化基因算法

Moscato 于 1989 年首次提出文化基因算法[10]的概念，由于群体算法搜索范围大，而局部搜索算法具有深度优势，因而这种算法进化效率比传统的遗传算法高得多.

针对网络流量分类的特征选择问题，采用克隆算法进行全局搜索，采用联合互信息的优化作为局部搜索策略.

1.2　互信息

互信息表示两个变量间共同拥有的信息含量，假定随机变量 X 和 Y 的边缘概率分布分别为 $p(x)$，$p(y)$，则二者之间的互信息 $I(X;Y)$ 定义为

$$I(X;Y) = \sum_{x\in\Phi}\sum_{y\in\Omega} p(x,y)\,\mathrm{lb}\,\frac{p(x,y)}{p(x)p(y)}. \tag{1}$$

联合互信息 $I(X_1,\cdots,X_n;Y)$ 表示多个变量 $X_1,\cdots,X_n$ 与 Y 之间的依赖程度，即：

$$I(X_1,\cdots,X_n;Y) = -\sum_{y\in\Omega}\sum_{x_n\in\Phi_n}\cdots\sum_{x_1\in\Phi_1} p(x_1,\cdots,x_n,y)\,\mathrm{lb}\,\frac{p(x_1,\cdots,x_n,y)}{p(x_1,\cdots,x_n)p(y)}. \tag{2}$$

由于联合互信息计算复杂，一般采取启发式算法，保证子集中特征变量与决策变量间互信息尽可能大，同时特征间的互信息尽可能小.

2　MIMF 特征选择算法

结合互信息和文化基因算法，提出一种 MIMF(combine mutual information with memetic framework)流量分类特征选择算法，算法伪代码见算法 1. 首先，随机初始化种群，每个染色体代表一个候选的特征子集. 然后，针对所有或者部分染色体，基于拉马克学习[11]进行局部提升. 最后，使用遗传算子产生下一代种群. 重复这一过程直到满足停止条件.

算法 1　MIMF 算法

1：Initialize：Randomly generate an initial population of feature subsets；
2：**While**(Stopping conditions are not satisfied)
3：Evaluate all feature subsets encoded in the population；
4：**For** each subset chosen to undergo the local improvement process；
5：Perform local search and replace it with locally improved solution；
6：**End For**
7：Perform evolutionary operators based on selection，crossover，and mutation；
8：**End While**

2.1　染色体编码和种群初始化

候选特征子集编码为一个染色体，染色体由一个长度等于特征数量的二进制字符串表示，每一比特位代表一个特征(见图 1). 在每个比特位上，“1”(“0”)表示对应特征被选择(排除). 设染色体长度为 n，每个染色体中“1”的最大数量为 M，即特征子集所包含的最多特征数量. 随机初始化种群，设种群规模为 P.

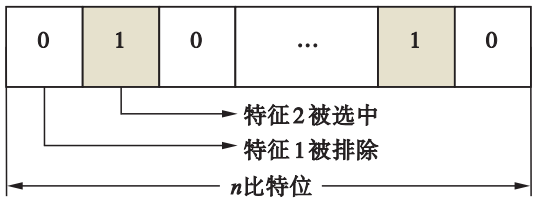

图 1　染色体的二进制位串表示法

Fig. 1　The binary string representation of chromosome

2.2　目标函数

将目标函数直接定义为分类精度：

$$\mathrm{Fintess}(c) = J(S_c). \tag{3}$$

其中：S_c 表示染色体 c 所对应的特征子集；目标函数 $J(S_c)$ 用来评价特征子集 S_c 的适应度. 当两个染色体具有相近适应度时，即当它们的适应度差距小于值 ε 时，给予特征数量较少的一个更高的存活几率.

2.3　局部搜索

采用过滤器的排序方法作为 meme 或者 LS 启发式算法. S 表示染色体 c 的特征子集，E 表示剩余特征的子集，满足 $F = S\cup E$. S 和 E 都采用信息增益进行排序. 下面定义 MIMF 算法局部搜索两个基本的 LS 算子：

1)“加”：采用线性排序选择，从 E 中选择一个特征 f_i，加入到 S 中，使得 $J(f_i)$ 取得最大值，优化目标见式(4). 并更新 $S = S\cup\{f_i\}$，$E = E\backslash\{f_i\}$.

$$\max_{f_i\in E}\left[I(f_i;c) - \frac{\beta}{|S|}\sum_{f_j\in S} I(f_i;f_j)\right]. \tag{4}$$

其中：$\beta = 1/(1+\mathrm{e}^{-\alpha g})$；$g$ 为进化代数；α 为修正因子.

2)“减”：采用线性排序选择，从 S 中选择一个特征，加入到 E 中，使得 $J(f_i)$ 取最小值，优化目标见式(5). 更新 $S = S\backslash\{f_i\}$，$E = E\cup\{f_i\}$.

$$\min_{f_i\in S}\left[I(f_i;c) - \frac{\beta}{|S|-1}\sum_{f_j\in S;f_j\neq f_i} I(f_i;f_j)\right]. \tag{5}$$

如图 2 所示,运用加算子,f_2 应该被加入到 S 中;运用减算子,f_3 应该被加入到 E 中.

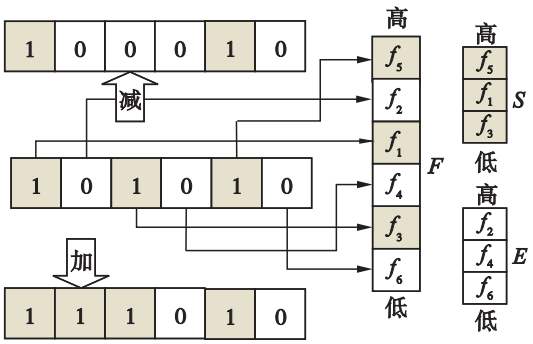

图 2 染色体的加算子和减算子

Fig. 2 The addition/subtraction operator of the chromosome

局部搜索的程度由搜索概率 w 和 LS 长度 l 确定. 搜索概率定义了每一代种群中进行局部搜索的精英染色体比例. LS 长度定义了局部搜索过程中加算子和减算子的最大数量. 改进优先局部搜索策略详见算法 2;贪心搜索策略伪代码与算法 2 类似,去掉算法 2 的第 10 行即可.

算法 2 改进优先局部搜索策略

1:Initialize l and w;
2:**For** each chromosome c among the $w \times P$ elitists
3:$c_{best} = c$;
4:**For** each of the l^2 combinations(k,d)
5:Repeat k times of Del operation;
6:Repeat d times of Add operation;
7:Calculate fitness of improved chromosome c' using
$F(c') = J(c')$;
8:**If**$((F(c') > F(c_{best}) > \varepsilon$ or$(|F(c') - F(c_{best})| < \varepsilon$ and $|c'| < |c_{best}|))$
9:$c_{best} = c'$;
10:break;
11:**End If**
12:**End For**
13:Replace the genotype c with the best improved c_{best};
14:**End For**

2.4 进化算子

在进化过程中,可以应用经典的遗传算子,例如选择、一致性交叉和变异算子. 可以限定每个染色体中 bit 位为 1 的数量的最大值为 M,由于标准的一致性交叉和变异算子可能违反这一限制,设计一种约束性的交叉和变异算子. 算法 3 概述了约束性交叉算子. 变异算子应用在双亲染色体的比特位为 1 的等位基因上. 约束性变异算子基于同样原理,详见算法 4.

算法 3 约束性交叉

1:Randomly select two parents p_1 and p_2;
2:Randomly generate a number r within$[0,1]$;
3:**If**$(r < c_p)$//c_p *denotes the crossover probability*
4:$k = \min(|p_1|, |p_2|)$;//$|p_1| \leqslant M, |p_2| \leqslant M$, ensure the number of bit '1' in the offspring does't exceed M;
5:**For**$(i = 1$ to $k)$
6:Locate the allele L_1 of the i^{th} bit '1' in p_1;
7:Locate the allele L_2 of the i^{th} bit '1' in p_2;
8:Crossover p_1 and p_2 in positins L_1 and L_2 with probability 0.5;
9:**End For**
10:**End If**

算法 4 约束性突变

1:**For**$(i = 1$ to $|c|)$
2:Locate the position L_1 of the i^{th} bit '1' in c;
3:Randomly select a bit '0' with position L_0;
4:Swap positions L_1 and L_0 with probability p;
//p denotes the mutation probability
5:**End For**
6:**For**$(i = 1$ **to** $M - |c|)$
7:Randomly flip a bit '0' with probability p;
8:**End For**

2.5 计算复杂度

由于 $I(f_i;c)$ 和 $I(f_i;f_j)$ 可以离线计算,每次局部搜索都可以重用其结果,其时间复杂度分别为 $O(|F| \cdot |C|)$ 和 $O(|F| \cdot |F-1|/2)$. 互信息的计算复杂度远远小于适应度函数(3),即训练分类器并分类测试集给出分类精度的计算代价.

设 GA 算法经过 G 次迭代收敛,则计算复杂度为 $O(PG)$,其中 P 是种群规模. 设 MIMF 算法经过 g 次迭代收敛,采用改进优先策略时计算复杂度是 $O(Pg + l^2wPg/2)$;采用贪心策略时,计算复杂度为 $O(Pg + l^2wPg)$. 通过实验可以发现,由于 $g \ll G$,MIMF 算法复杂度显著降低.

3 实验设计与分析

3.1 实验数据集

本文数据集于 2013 年 4 月第一周收集自 CERNET 东北地区中心节点. 实验中用 DPI 工具来产生大约包含 10^5 条网络流共 2.6 GB 字节的训练集,另外 10^5 条网络流约 2.8 GB 字节作为测

试集. 详细信息如表 1 所示.

表 1　数据集概况
Table 1　Survey of the dataset

数据集	训练集	测试集
Size/GB	2.6	2.8
Flows	10^5	10^5
TCP/%	63	54
TCP Bytes/%	90.0	91.9
Local IPs	1 384	1 516
Distant IPs	33.4	29.5

所有流量分为 8 类,详见表 2. 每个类型所占比例如表 3 所示,WEB 和 P2P 占据 tcp 和 udp 连接(flow)的绝大部分,而字节(size)主要由 P2P, STREAMING 和 WEB 组成,这主要是因为 STREAMING 类型的网络连接持续时间长,携带大量数据包.

表 2　分类策略
Table 2　Classification taxonomy

类型	程序/协议
WEB	HTTP,HTTPS
STREMING	PPLive,PPTV,HTTP Flv
P2P	Bittorrent,eMule,Xunlei
CHAT	MSN,QQ,HTTP Chat
MAIL	SMTP,POP3,HTTP Mail
FTP	Ftp－data,Ftp－control
GAMES	Counter Strike,WOW,Dota
OTHERS	NBS,Attacks

表 3　流量比例
Table 3　Proportion of class　　%

类型	流	字节
WEB	49.16	18.68
STREMING	5.71	32.39
P2P	36.4	36.85
CHAT	1.33	0.22
MAIL	1.29	1.10
FTP	0.13	0.71
GAMES	0.06	0.02
OTHERS	5.92	10.03

3.2　实验结果与分析

以文献[4]给出的 248 个流特征作为初始网络流量特征全集. 实验以 MIMF 作为特征选择算法,以 SVM 为分类器. 图 3 给出了分类精度随进化代数变化的曲线图,MIMF 算法在 20 代收敛,储慧琳等[12]提出的组合式算法在 45 代左右收敛. 同该组合式算法相比,MIMF 算法显著降低了特征选择算法的时间复杂度.

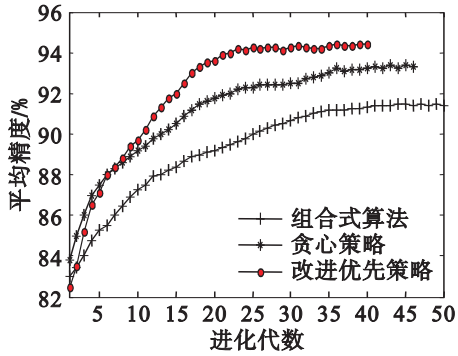

图 3　分类精度与进化代数的关系
Fig. 3　The ralationship between classification accuracy and evolution generation

流特征选择结果及分类平均准确率见表 4. 同文献[13]提出的 WSU_AUC 算法以及文献[12]提出的组合式算法相比,MIMF 算法提取的特征数量更少,分类精度更高.

表 4　实验结果
Table 4　Experimental results

特征选择算法	选取的流特征	特征个数	平均分类准确率/%
WSU_AUC	1,83,96,60,156,108,232,112,2,36,209,111	12	89.6
组合式算法	1,2,83,90,94,165,171	7	91.4
MIMF	1,2,83,60,94,165,112,171	8	94.4

4　结　　论

1）通过将全局搜索和局部搜索相结合,混合式特征选择算法 MIMF 获得了用于流量识别的最优流特征集合.

2）与传统 GA 算法以及 WSU_AUC 算法相比,该算法在计算复杂度上有显著改进,能在较少的特征数量上获得更高的分类精度.

参考文献:

[1] Du M, Chen X, Tan J. Online internet traffic identification algorithm based on multistage classifier [J]. *China Communications*, 2013, 10(2): 89－97.

[2] Camacho J, Padilla P, Teodoro P, *et al*. A generalizable dynamic flow pairing method for traffic classification [J]. *Computer Networks*, 2013, 57(14): 2718－2732.

[3] Nguyen T T, Armitage G, Branch P, *et al*. Timely and

continuous machine-learning-based classification for interactive IP traffic [J]. *IEEE/ACM Transactions on Networking (TON)*, 2012, 20(6): 1880 – 1894.

[4] Moore A, Zuev D, Crogan M, *et al*. Discriminators for use in flow-based classification [M]. London: Queen Mary and Westfield College Press, 2005: 1 – 14.

[5] Amiri F, Rezaei Y M, Lucas C, *et al*. Mutual information-based feature selection for intrusion detection systems [J]. *Journal of Network and Computer Applications*, 2011, 34(4): 1184 – 1199.

[6] Yang J, Ma J, Cheng G, *et al*. An empirical investigation of filter attribute selection techniques for high-speed network traffic flow classification [J]. *Wireless Personal Communications*, 2012, 66(3): 541 – 558.

[7] Jain A, Zongker D. Feature selection: evaluation, application, and small sample performance [J]. *Pattern Analysis and Machine Intelligence*, 1997, 19(2): 153 – 158.

[8] Shirazi H M, Namadchian A, Khalili T A. A combined anomaly base intrusion detection using memetic algorithm and bayesian networks [J]. *International Journal of Machine Learning and Computing*, 2012, 2(5): 706 – 710.

[9] Bacquet C, Zincir-Heywood A N, Heywood M I. Genetic optimization and hierarchical clustering applied to encrypted traffic identification [C]//Computational Intelligence in Cyber Security (CICS). Paris: IEEE, 2011: 194 – 201.

[10] Moscato P. On evolution, search, optimization, genetic algorithms and martial arts: towards memetic algorithms [M]. Pasadena: Caltech Concurrent Computation Program, 1989: 1 – 64.

[11] Ong Y S, Keane A J. Meta-lamarckian learning in memetic algorithms [J]. *Evolutionary Computation*, 2004, 8(2): 99 – 110.

[12] 储慧琳,张兴明. 一种组合式特征选择算法及其在网络流量识别中的应用[J]. 小型微型计算机系统, 2012, 33(2): 325 – 329.
(Chu Hui-lin, Zhang Xing-ming. Application of a hybrid feature selection algorithm in internet traffic identification [J]. *Journal of Chinese Computer Systems*, 2012, 33(2): 325 – 329.)

[13] Zhang H, Lu G, Qassrawi M T, *et al*. Feature selection for optimizing traffic classification [J]. *Computer Communications*, 2012, 35(12): 1457 – 1471.

(上接第 1524 页)

2) 针对含有界干扰的随机奇异 T – S 模糊系统,利用线性矩阵不等式方法给出该系统的有限时间鲁棒控制器存在的充分条件.

3) 通过数值算例验证了所提出方法的可行性和有效性.

参考文献:

[1] 严治国,张国山. 线性随机系统有限时间 H_∞ 控制[J]. 控制与决策, 2011, 26(8): 1224 – 1228.
(Yan Zhi-guo, Zhang Guo-shan. Finite-time H_∞ control for linear stochastic systems [J]. *Control and Decision*, 2011, 26(8): 1224 – 1228.)

[2] Zhang W H, Chen B S. State feedback control for a class of nonlinear stochastic systems [J]. *SIAM Journal on Control and Optimization*, 2006, 44(6): 1973 – 1991.

[3] 胡良剑,邵世煌,吴让泉. T – S 模糊随机系统的均方镇定[J]. 信息与控制, 2004, 33(5): 545 – 559.
(Hu Liang-jian, Shao Shi-huang, Wu Rang-quan. Mean square stabilization of T-S fuzzy stochastic systems [J]. *Information and Control*, 2004, 33(5): 545 – 559.)

[4] Hu L J, Zhao W G, Shao S H. Robust stochastic stabilization and robust H_∞ control for uncertain stochastic fuzzy systems [C]//Proceedings of the IEEE International Conference on Fuzzy System. Reno: IEEE, 2005: 254 – 259.

[5] Wang Z, Daniel W C H, Liu X H. A note on the robust stability of uncertain stochastic fuzzy systems with time-delay [J]. *IEEE Transactions on Systems Man and Cybernetics*, 2004, 34(4): 570 – 576.

[6] Dorato P. Short time stability in linear time-varying systems [C]//Procedure of IRE International Convention Record. New York, 1961: 83 – 87.

[7] Weiss I L E. Finite time stability under perturbing forces and on product spaces [J]. *IEEE Transaction on Automatic Control*, 1967, 12(1): 54 – 59.

[8] Amato F, Ariola M, Dorato P. Finite time control of linear system subject to parametric uncertainties and disturbances [J]. *Automatica*, 2001, 37(9): 1459 – 1463.

[9] Amato F, Ariola M, Dorato P. Finite time stabilization via dynamic output feedback [J]. *Automatica*, 2006, 42(2): 337 – 342.

[10] Oksendal B. Stochastic differential equations: an introduction with applications [M]. 5th ed. New York: Springer-Verlag, 2000.