

面向类不平衡网络流量的特征选择算法

唐 宏^{①②} 刘 丹^{*①②} 姚立霜^{①②} 王云锋^① 裴作飞^{①②}

^①(重庆邮电大学通信与信息工程学院 重庆 400065)

^②(移动通信技术重庆市重点实验室 重庆 400065)

摘 要: 针对网络流量分类过程中出现的类不平衡问题, 该文提出一种基于加权对称不确定性(WSU)和近似马尔科夫毯(AMB)的特征选择算法。首先, 根据类别分布信息, 定义了偏向于小类别的特征度量, 使得与小类别具有强相关性的特征更容易被选择出来; 其次, 充分考虑特征与类别间、特征与特征之间的相关性, 利用加权对称不确定性和近似马尔科夫毯删除不相关特征及冗余特征; 最后, 利用基于相关性度量的特征评估函数以及序列搜索算法进一步降低特征维数, 确定最优特征子集。实验表明, 在保证算法整体分类精确率的前提下, 算法能够有效提高小类别的分类性能。

关键词: 流量分类; 特征选择; 类不平衡; 加权对称不确定性; 近似马尔科夫毯

中图分类号: TN915; TP393

文献标识码: A

文章编号: 1009-5896(2021)04-0923-08

DOI: [10.11999/JEIT190992](https://doi.org/10.11999/JEIT190992)

Feature Selection Algorithm for Class Imbalanced Internet Traffic

TANG Hong^{①②} LIU Dan^{①②} YAO LiShuang^{①②}

WANG Yunfeng^① PEI Zuofei^{①②}

^①(School of Communication and Information Engineering, Chongqing University of
Posts and Communications, Chongqing 400065, China)

^②(Key Laboratory of Mobile Communications Technology, Chongqing University of
Posts and Communications, Chongqing 400065, China)

Abstract: Class imbalance always exists in the process of network traffic classification. Considering the problem, a new feature selection algorithm using Weighted Symmetric Uncertainty (WSU) and Approximate Markov Blanket (AMB) is proposed. Firstly, a feature metric is defined using category distribution information, which is biased to minority classes. This makes it easier pick out features which have strong correlation with minority classes. Then, considering the correlation between features and categories and between features and features, the weighted symmetry uncertainty and approximate Markov blanket are used to delete the unrelated features and redundant features. Finally, the feature dimension is further reduced to determine the optimal feature subset, by using feature evaluation functions based on correlation measures and sequence search algorithms. The experimental results demonstrate that the algorithm can effectively improve the classification performance of minority classes without sacrificing the accuracy of the overall classification.

Key words: Traffic classification; Feature selection; Class imbalance; Weighted Symmetric Uncertainty (WSU); Approximate Markov Blanket (AMB)

1 引言

网络流量分类技术广泛应用于网络管理和网络安全领域^[1]。基于端口号的流量分类方法^[2,3]在开放端口、伪装端口号等技术出现之后, 分类准确率大

大降低。基于特征字段的流量分类方法摆脱了对端口号的依赖, 可以取得较高的分类准确率, 但该方法只能对明文传输的数据包进行解析, 难以适用于加密流量的分类^[4-6]。基于传输层主机行为的流量分类方法不依赖于端口号, 也不需要解析报文, 但该方法对外界环境异常敏感, 多变的网络环境可能导致分类效果不够稳定。因此, 基于机器学习的网络流量分类方法得到了研究人员的青睐^[1,7]。

数据预处理是基于机器学习的网络流量分类过

收稿日期: 2019-12-11; 改回日期: 2021-02-22; 网络出版: 2021-03-04

*通信作者: 刘丹 s170101113@stu.cqupt.edu.cn

基金项目: 长江学者和创新团队发展计划(IRT_16R72)

Foundation Item: Changjiang Scholars and Innovative Research Team in University (IRT_16R72)

程中的重要步骤,该过程通常采用特征选择算法对特征进行降维。Moore等人^[8]使用快速相关滤波器去除冗余和无关的特征。Dai等人^[9]利用卡方和C4.5算法进行特征选择。Xu等人^[10]将二进制萤火虫算法与反向学习相结合进行特征选择。张震等人^[11]通过定义基于信息熵的“用户行为模式”来分析各个行为子簇背后表现出的业务特征,能有效降低算法的计算复杂度。Shafiq等人^[12]提出了一种基于机器学习的混合特征选择算法,该算法利用了加权互信息度量和受试者工作特征曲线(Receiver Operating Characteristic, ROC)下面积两个指标,使用该算法选择出的特征能够表现出很好的性能。Shi等人^[13]提出了一种新的基于深度学习和特征选择技术的特征优化方法,为流量分类提供具有强区分能力的特征。Wang等人^[14]提出一种基于多重过滤权重和多重特征权重的混合特征选择方法,能有效提高分类精度。王勇等人^[15]在特征提取过程中引入卷积神经网络,可以在避免复杂显式特征提取的同时达到提高分类精度的效果。Ren等人^[16]使用深层神经网络挑选数据的固有特征,利用树结构的递归神经网络处理大规模流量分类问题,可以在更少的训练时间内获得更高的性能。

现有的特征选择算法多数情况下选择出来的特征在多数类(大类)上表现良好,分类器对少数类(小类)的预测精度却很低,这就是类不平衡导致的问题。然而,在现实生活中,人们通常更关注小类的分类效果,错误地识别小类所带来的后果往往很严重。如在入侵检测中^[17-19],攻击类相对于正常流量就属于小类,错分攻击类可能会引起网络的瘫痪。

为了减轻类不平衡问题带来的不良影响,本文提出一种基于加权对称不确定性(Weighted Symmetric Uncertainty, WSU)和近似马尔科夫毯(Approximate Markov Blanket, AMB)的网络流量特征选择算法。该算法首先根据类别分布信息定义偏向于小类别的度量,使得识别小类别的特征更容易被选择出来,并基于加权信息熵计算特征与类别间的加权对称不确定性,利用特征排序算法删除不相关特征;然后,充分考虑特征与类别间、特征与特征间的相关性,利用近似马尔科夫毯条件删除冗余特征;最后,根据特征评估准则函数和序列搜索算法从候选特征集合中选出满足合适维数的特征子集。

2 特征选择

特征选择的流程图如图1所示,它主要包含4个基本环节:生成特征子集(搜索策略),评价准则,停止准则和结果验证^[20]。在原始特征集上运用搜索策略得到生成子集,利用评价准则对第1步选择出来的子集进行评估,最后根据停止准则结束搜索,并选择出来的最优特征子集进行检验,判断所选特征子集是否满足标准。

3 基于WSU_AMB的特征选择算法

基于加权对称不确定性和近似马尔科夫毯(WSU_AMB)的网络流量特征选择算法主要包含两个步骤。第1步,根据类别分布信息定义偏向于小类别的权值,基于加权信息熵计算出特征与类别间的加权对称不确定性,利用特征排序算法删除不相关特征,这一步可以过滤掉大多数特征;计算特征与特征之间的加权对称不确定性,利用近似马尔科夫毯删除冗余特征,确定候选特征集合。第2步,计算特征准则评估函数值,利用序列搜索算法选择出满足合适维度的特征子集。WSU_AMB算法的总体框架如图2所示。

3.1 加权对称不确定性

加权对称不确定性可以用来衡量特征与类别以及特征与特征之间的相关性,它是在加权信息熵的基础上计算出来的^[21]。加权对称不确定性可由式(1)进行计算

$$WSU(F, C) = 2 \left[\frac{IG_w(C|F)}{H_w(C) + H_w(F)} \right] \quad (1)$$

其中

$$IG_w(C|F) = H_w(C) - H_w(C|F) \quad (2)$$

$$H_w(C) = - \sum_i \sum_j w_i p(c_i, f_j) \log_2 p(c_i) \quad (3)$$

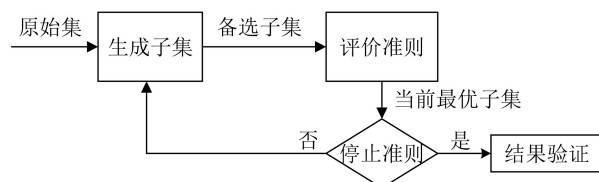


图1 特征选择流程图

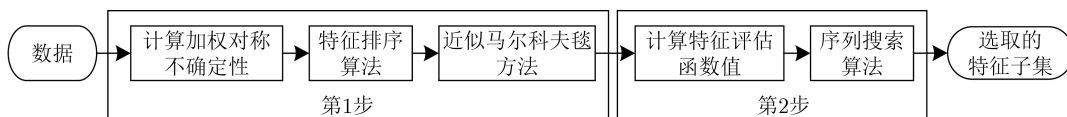


图2 WSU_AMB特征选择算法的总体框架

$$H_w(F) = - \sum_i \sum_j w_i p(c_i, f_j) \log_2 p(f_j) \quad (4)$$

$$H_w(C|F) = - \sum_i \sum_j p(f_j) w_i p(c_i|f_j) \log_2 p(c_i|f_j) \quad (5)$$

$$w_i = 1 - \frac{n_i}{N} \quad (6)$$

w_i 表示权值, $p(c_i, f_j)$ 表示类别 C 与特征 F 的联合概率, $p(c_i)$ 表示类别 C 的先验概率, $p(f_j)$ 表示特征 F 的先验概率, $p(c_i|f_j)$ 表示 F 发生的条件下 C 的后验概率, n_i 表示属于类别 c_i 的样本数, N 表示样本总量。

通过加权对称不确定性, 可对相关特征进行定义, 具体表述为: 计算每个特征与类别之间的加权对称不确定性 $WSU(f_i, C)$, 对该值进行降序排列, 排在最前面、值越大所对应的特征与类别的相关性越强。

3.2 近似马尔科夫毯

通常对特征与特征之间的相关性进行分析来判断某一特征是否冗余。根据马尔科夫毯思想可以形式化地给出冗余特征的定义, 但是马尔科夫毯的条件过于严格, 现实数据难以达到要求, 需要对该条件进行近似假设^[22], 基于此, 本节提出了近似马尔科夫毯来删除冗余特征。

所谓马尔科夫毯, 需满足以下条件。假设属性类别为 C , 特征集合为 F , 对于给定的特征 $f_i \in F$ 和特征子集 $M \subset F (f_i \notin M)$, 若有

$$f_i \perp \{F - M - \{f_i\}, C\} | M \quad (7)$$

则称能满足上述条件的特征子集 M 为 f_i 的马尔科夫毯。其中, 符号“ $\perp$ ”表示独立, “ $|M$ ”表示在给定 M 的条件下。

假设属性类别为 C , 特征集合为 F , 特征 f_i 为冗余特征的充要条件为当且仅当存在特征子集 M 为 f_i 的马尔科夫毯, 其中, $f_i \in F, M \subset F (f_i \notin M)$ 。在特征集合 F 中, 由于在特征 f_i 的马尔科夫毯 M 条件下, f_i 与其他非马尔科夫毯变量独立, 因此, 对于特征 f_i 而言, 所有非马尔科夫毯变量都是冗余的。

特征 f_i 是特征 f_j 的近似马尔科夫毯($i \neq j$), 需要满足式(8)的条件

$$\left. \begin{aligned} WSU(f_i, C) &> WSU(f_j, C) \\ WSU(f_j, C) &< WSU(f_i, f_j) \end{aligned} \right\} \quad (8)$$

特征与类别之间的 WSU 可由式(5)得到, 特征与特征之间的 WSU 的计算方法略有差别, 此时需要将其中一个特征看成类别属性。

3.3 相关性特征度量

在充分考虑特征的相关性的前提下, 有效减少特征维数, 提出一种特征准则评估函数

$$J(s) = \frac{n\bar{r}_{cf}}{\sqrt{n + n(n-1)\bar{r}_{ff}}} \quad (9)$$

其中, n 表示特征子集 s 中的特征个数, $\bar{r}_{cf}$ 表示特征子集 s 中各个特征与类别相关度的平均值, $\bar{r}_{ff}$ 表示特征子集 s 中特征之间相关度的平均值, r 为Pearson相关系数。两个变量之间的Pearson相关系数定义为两个变量之间的协方差和标准差的商

$$\begin{aligned} r_{X,Y} &= \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y} \\ &= \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}} \end{aligned} \quad (10)$$

3.4 算法流程描述

WSU_AMB 算法的实现过程如表1所示。第1阶段((1)~(9)行)利用加权对称不确定性和近似马尔科夫毯条件删除不相关特征和冗余特征, 得到候选特征集合; 第2阶段((10)~(20)行)利用特征评估准则函数和序列搜索算法找到最优特征子集。

4 实验验证

4.1 实验数据集

本实验使用Moore数据集^[23]来验证算法的性能, 表2展示了该数据集的统计信息。

4.2 性能评价指标

本实验采用整体精确率(accuracy), 小类别的准确率(precision)、召回率(recall)和F1值作为算法的性能评价指标。整体精确率可以反映多分类模型的综合预测能力, 准确率、召回率和F1值可以反映分类模型对单个应用的预测能力。

4.3 实验方案

利用Moore数据集的10个数据子集(DataSet1, DataSet2, ..., DataSet10)进行训练和预测, 每个子集中训练集的比例为70%。选用C4.5决策树作为分类器进行所提算法的最优参数选择。

在对网络流量进行分类时, 一般4~8个特征就能得到较好的分类效果。本文对10个数据子集进行实验, 可以发现相似的变化趋势, 只选取2个数据子集进行展示, 如图3, 经验证, 选取特征数 $L=6$ 。

阈值 δ 的设置对算法的性能有重要影响。因为 δ 值越高, 筛选特征的速度越快, 但会遗漏某些重要特征, 降低分类系统的性能; δ 值越低, 会把训

表 1 基于WSU_AMB的特征选择算法

输入: $D(f_1, f_2, \dots, f_N, C)$, WSU阈值 δ , $F = \{f_1, f_2, \dots, f_N\}$, 最优特征子集中特征数目 L	
输出: 最优特征子集 F_O	
第1阶段: 确定候选特征集合	
(1)	FOR $f_i \in F$
(2)	计算 $WSU(f_i, C)$
(3)	将特征按 $WSU(f_i, C)$ 值降序排列
(4)	IF $WSU(f_i, C) > \delta$
(5)	将特征 f_i 添加到特征子集 S^* 中
(6)	WHILE $S^* \neq \emptyset$
(7)	选择 S^* 中的第1个特征 f_i 作为显著特征, 将特征 f_i 加入特征子集 S , 从特征集合 S^* 中删除特征 f_i
(8)	查找以特征 f_i 为近似马尔科夫毯的特征子集 $\{f_j\}$
(9)	将特征子集 $\{f_j\}$ 从 S^* 中删除
第2阶段: 选择最优特征子集	
(10)	FOR $f_d \in S$
(11)	计算 $J(f_d)$
(12)	IF $J(f_a) = \max\{J(f_d)\}$
(13)	将特征 f_a 加入目标特征子集 F_O , 从候选特征集合 S 中删除特征 f_a
(14)	FOR $f_x \in S$
(15)	计算 $J(F_O \cup f_a)$
(16)	IF $J(f_1) = \max\{J(F_O \cup f_a)\}$
(17)	将特征 f_1 加入目标特征子集 F_O , 从候选特征集合 S 中删除特征
(18)	FOR Length(F_O) < L
(19)	重复(14)—(17)行
(20)	输出 F_O

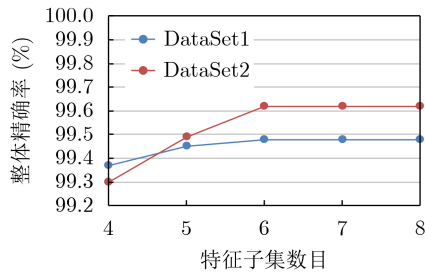
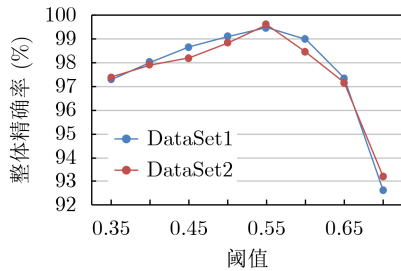
表 2 Moore数据集的统计信息

类别	应用实例	流量数	百分比(%)
WWW	www	328,092	86.905
MAIL	Imap, pop2/3, smtp	28,567	7.567
FTP-CONTROL	ftp-control	3,054	0.809
FTP-PASV	ftp-pasv	2,688	0.712
ATTACK	Internet worm, virus attacks	1,793	0.475
P2P	KaZaA, BitTorrent, GnuTella	2,094	0.555
DATABASE	Postgres, sqlnet oracle, ingres	2,648	0.702
FTP-DATA	ftp-data	5,797	1.536
MULTIMEDIA	Windows media player, Real	576	0.152
SERVICES	X11, dns, ident, Idap, ntp	2,099	0.556
INTERACTIVE	ssh, klogin, rlogin, telnet	110	0.029
GAMES	Half-Life	8	0.002
total	28	377,526	100

练样本自身的某些特点当成共性来学习, 会导致分类器泛化性能下降, 出现“过拟合”的现象。同样地, 利用10个数据子集进行阈值选择实验, 分类器的整体精确率变化趋势相似, 如图4所示, 本文选

取了2个数据子集进行实验展示。通过实验验证, 选取阈值 $\delta=0.55$ 。

在实验过程中, 利用4种分类器: 朴素贝叶斯(Naive Bayes, NB)、逻辑斯蒂回归模型(Logistic

图3 特征子集数目 L 对算法的影响图4 阈值 δ 对算法的影响

Regression, LR)、K近邻算法(K-Nearest Neighbour, KNN)和C4.5决策树(Decision Tree, DT), 将未进行特征选择的数据集(fullset)、基于相关的快速过滤器(Fast Correlation-Based Filter, FCBF)^[8]、卡方决策树算法(Chi-Square and Decision Tree, CSDT)^[9]、高效的特征优化方法(Efficient Feature Optimization Approach, EFOA)^[13]以及基于多重过滤权重和多重特征权重的混合特征选择方法(Hybrid Feature Selection based on Multi-Filter weights and Multi-Feature weights, MFHFS)^[14]与所提算法(WSU_AMB)进行对比。

4.4 实验结果与分析

4.4.1 所选特征集合

选用C4.5决策树作为分类器, 在不同的数据子集下, FCBF, CSDT, EFOA, MFHFS和WSU_AMB算法选择出来的特征数目如表3所示。选出的特征子集规模显示, 总体上WSU_AMB算法所选特征数较少, CSDT算法所选的特征数较多, 其他3种算法的降维能力相当。WSU_AMB算法的停止准则是基于搜索策略设定的, 一旦所选特征达到指定个数即停止搜索, 所以在不同的数据子集上, WSU_AMB算法选择的特征个数都相同。而其他4种算法的停止准则是基于评价准则设定的, 需要在评价价值达到最高时才停止搜索, 故每个数据子集所选特征个数会有所不同。

4.4.2 算法复杂度对比

表4对各算法的时间复杂度进行了定性分析, 其中, N 为特征总数, M 为数据集的实例数目, L 为候选特征集合中的特征数量, D 为隐层数, t 为

表3 不同特征选择方法所选特征数目

数据集	FCBF	CSDT	EFOA	MFHFS	WSU_AMB
DataSet1	8	9	9	8	6
DataSet2	7	5	6	7	6
DataSet3	5	10	7	6	6
DataSet4	6	14	5	5	6
DataSet5	7	5	8	7	6
DataSet6	6	9	5	6	6
DataSet7	7	11	6	5	6
DataSet8	8	9	8	7	6
DataSet9	6	14	7	7	6
DataSet10	7	6	7	6	6

表4 不同特征选择方法的时间复杂度分析

算法	时间复杂度
FCBF	$O(MN\log_2 N)$
CSDT	$O(N^2 + \log_2 L)$
EFOA	$O\left(\sum_{t=1}^D N_t \cdot C_t\right) + MN\log_2 N$
MFHFS	$O(N\log_2 N + L^3)$
WSU_AMB	$O(MN^2) + O(NK^2)$

当前所在的层, N_t 为该层的特征数, C_t 为该层的类别数, K 为选择的特征数。

为了进一步对算法的复杂度进行定量分析, 利用Moore数据集的10个特征子集对各算法的特征选择时间进行统计, 如图5所示。从图5中可看出算法运行时间与算法复杂度的定性分析基本保持一致, WSU_AMB算法的特征选择时间少于EFOA算法, MFHFS算法的特征选择时间最短; 因DataSet7~DataSet10的数据量较大, 是前6个数据子集的2倍多, 故EFOA算法的特征选择时间出现了明显的增长。

同时, 选用C4.5决策树作为分类器, 对各算法的分类时间进行了测试, 如表5所示。很明显, 使用基于CSDT算法和EFOA算法选择出来的特征进行C4.5分类器的训练, 分类时间较长, 而C4.5使用WSU_AMB算法选择出来的特征进行分类时, 分

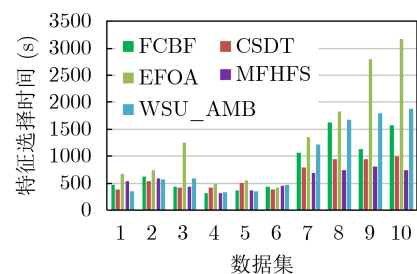


图5 不同算法在各数据集上的特征选择时间对比

类速度优于FCBF算法和MFHFS算法。虽然WSU_AMB算法在特征选择阶段的运行时间较长,但经过特征选择之后所选特征的特征值数目较小、数量较少,减少了分类过程中的计算开销,缩短了系统在分类阶段的耗时,提升了系统的分类效率。

4.4.3 分类性能对比

(1) 整体分类精确率。按照4.3节的实验方案进行实验,可以得到各特征选择算法在不同数据子集上的整体精确率。如图6所示,当NB, LR, 6NN和DT作为分类器时,WSU_AMB算法的平均整体精

确率均高于对比算法。使用6NN作为分类器时,各特征选择算法的分类性能更加稳定,当使用DT分类器时,5种特征选择算法均能取得较高的整体分类准确率。

(2) 小类准确率。选择ATTACK类和FTP-PASV类对5种特征选择方法在小类别上的性能进行分析。图7为基于不同数据子集,5种特征选择算法在DT分类器上的小类准确率对比。相比FTP-PASV类,ATTACK类在各个数据子集上的波动比较大。因为ATTACK类属于攻击类型,它通常会模仿其他流量的特征来躲避网络监测系统,故ATTACK类的分类结果呈现不稳定的特性。相比于其他4种对比算法,WSU_AMB算法对ATTACK类和FTP-PASV类的准确率提升明显,MFHFS算法的小类准确率性能仅次于WSU_AMB算法。

(3) 小类召回率。由于网络流量中的WWW类占绝大多数,所以在训练分类器时,往往对WWW类有利,容易把其他类别错分为WWW类,降低小类别的分类性能。而WSU_AMB算法为小类别选

表 5 分类时间的比较(ms)

算法	分类时间的均值
FCBF	153.6
CSDT	215.3
EFOA	234.6
MFHFS	146.4
WSU_AMB	120.7

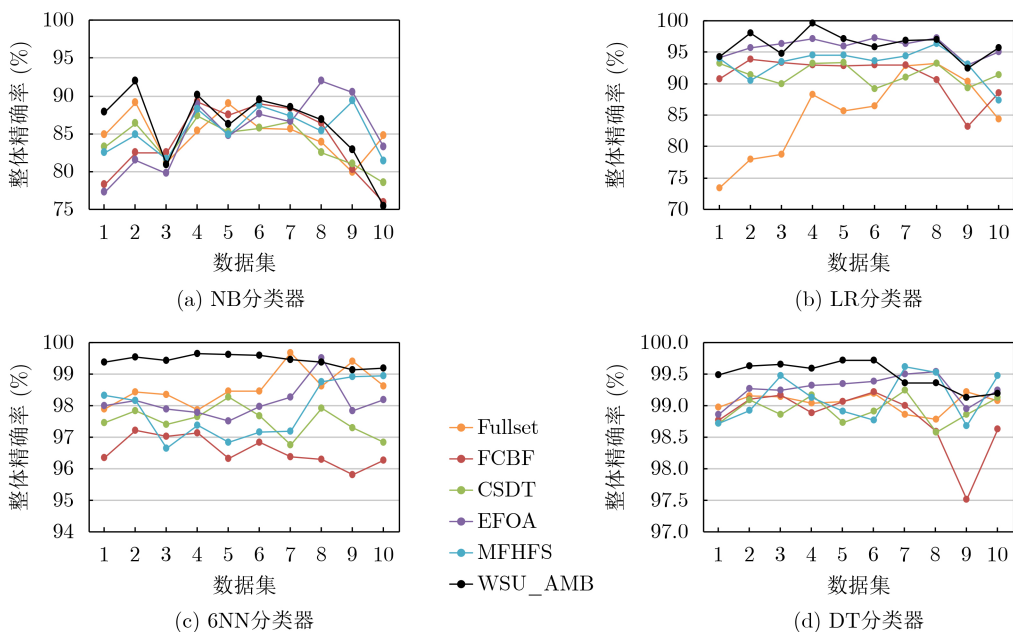


图 6 不同数据子集下各特征选择算法在4种分类器上的整体精确率对比

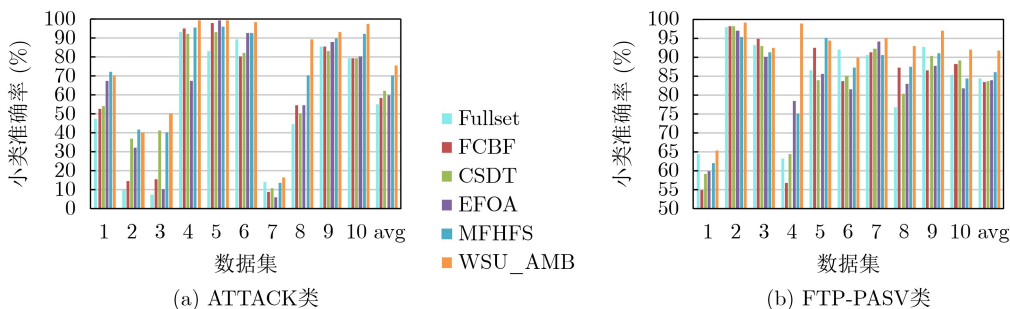


图 7 各特征选择算法的小类准确率对比

择出强相关性的特征，能够有效减少小类别被错分为WWW类的数量，提高小类别的召回率。图8为使用DT分类器，5种算法在各个子集上的小类召回率，可以看出，WSU_AMB算法的小类召回率好于对比算法。

(4) 小类F1值。图9为基于不同数据子集，5种

特征选择算法在DT分类器上的小类F1值对比。WSU_AMB算法在对每个特征进行度量时，识别小类别的特征权重更大；在对特征进行选择时，保证所选的特征具有较强的区分能力，即特征与类别之间高度相关，特征之间彼此不相关。因此，WSU_AMB算法的小类平均F1值能够高于其他4种方法。

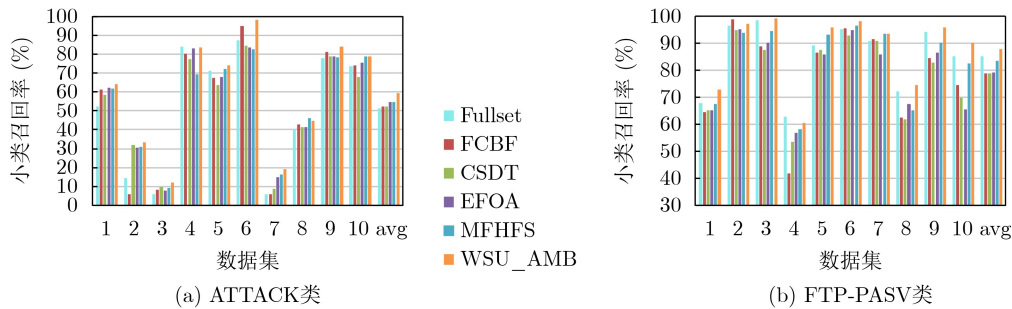


图8 各特征选择算法的小类召回率对比

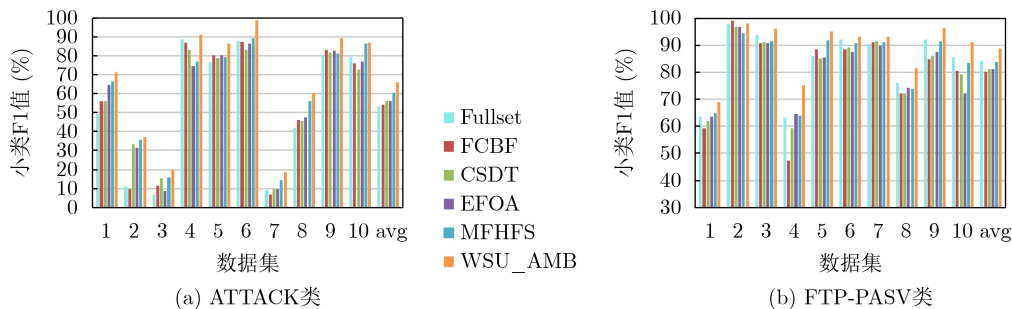


图9 各特征选择算法的小类F1值对比

5 结束语

针对网络流量分类技术存在的类不平衡问题，本文提出一种基于加权对称不确定性和近似马尔科夫毯的特征选择算法。充分考虑特征的相关性，利用偏向于小类别的加权对称不确定性和特征排序算法来滤除不相关特征，通过近似马尔科夫毯条件删除冗余特征；再根据特征评估准则函数和序列搜索算法进一步降低特征维数。实验表明，服务器的端口号和初始窗口中发送的字节总数是识别网络流量的两个重要特征，所提算法能够在不牺牲分类器整体精确率的前提下，有效提高小类别的准确率、召回率和F1值。进一步研究工作主要在以下几个方面：(1)减小数据漂移现象的影响；(2)有效降低算法在搜索最优特征子集时的时间消耗；(3)新应用的识别与分类。

参考文献

- [1] XUE Yibo, ZHANG Luoshi, and WANG Dawei. Traffic classification: Issues and challenges[J]. *Journal of Communications*, 2013, 8(4): 240–248. doi: [10.12720/jcm.2013.8.4.240-248](https://doi.org/10.12720/jcm.2013.8.4.240-248).
- [2] NGUYEN T T T and ARMITAGE G. A survey of techniques for internet traffic classification using machine learning[J]. *IEEE Communications Surveys & Tutorials*, 2008, 10(4): 56–76. doi: [10.1109/SURV.2008.080406](https://doi.org/10.1109/SURV.2008.080406).
- [3] DAINOTTI A, PESCAPE A, and CLAFFY K C. Issues and future directions in traffic classification[J]. *IEEE Network*, 2012, 26(1): 35–40. doi: [10.1109/mnet.2012.6135854](https://doi.org/10.1109/mnet.2012.6135854).
- [4] MOORE A W and PAPAGIANNAKI K. Toward the accurate identification of network applications[C]. The 6th International Workshop on Passive and Active Network Measurement, Boston, USA, 2005: 41–54. doi: [10.1007/978-3-540-31966-5_4](https://doi.org/10.1007/978-3-540-31966-5_4).
- [5] 叶春明, 王珍, 陈思, 等. 基于节点行为特征分析的网络流量分类方法[J]. *电子与信息学报*, 2014, 36(9): 2158–2165. doi: [10.3724/SP.J.1146.2013.01600](https://doi.org/10.3724/SP.J.1146.2013.01600).
- [6] YE Chunming, WANG Zhen, CHEN Si, et al. Internet Traffic classification based on hosts behavior analysis[J]. *Journal of Electronics & Information Technology*, 2014, 36(9): 2158–2165. doi: [10.3724/SP.J.1146.2013.01600](https://doi.org/10.3724/SP.J.1146.2013.01600).
- [6] DIAS K L, PONGELUPE M A, CAMINHAS W M, et al.

- An innovative approach for real-time network traffic classification[J]. *Computer Networks*, 2019, 158: 143–157. doi: [10.1016/j.comnet.2019.04.004](https://doi.org/10.1016/j.comnet.2019.04.004).
- [7] 鲁刚, 张宏莉, 叶麟. P2P流量识别[J]. 软件学报, 2011, 22(6): 1281–1298. doi: [10.3724/SP.J.1001.2011.03995](https://doi.org/10.3724/SP.J.1001.2011.03995).
LU Gang, ZHANG Hongli, and YE Lin. P2P traffic identification[J]. *Journal of Software*, 2011, 22(6): 1281–1298. doi: [10.3724/SP.J.1001.2011.03995](https://doi.org/10.3724/SP.J.1001.2011.03995).
- [8] MOORE A W and ZUZV D. Internet traffic classification using Bayesian analysis techniques[J]. *ACM SIGMETRICS Performance Evaluation Review*, 2005, 33(1): 50–60. doi: [10.1145/1071690.1064220](https://doi.org/10.1145/1071690.1064220).
- [9] DAI Lei, YUN Xiaochun, and XIAO Jun. Optimizing traffic classification using hybrid feature selection[C]. The 9th International Conference on Web-Age Information Management, Zhangjiajie, China, 2008: 520–525. doi: [10.1109/WAIM.2008.30](https://doi.org/10.1109/WAIM.2008.30).
- [10] XU Huali, YU Shuhao, CHEN Jiajun, *et al.* An improved firefly algorithm for feature selection in classification[J]. *Wireless Personal Communications*, 2018, 102(4): 2823–2834. doi: [10.1007/s11277-018-5309-1](https://doi.org/10.1007/s11277-018-5309-1).
- [11] 张震, 汪斌强, 陈鸿昶, 等. 互联网中基于用户连接图的流量分类机制[J]. 电子与信息学报, 2013, 35(4): 958–964. doi: [10.3724/SP.J.1146.2012.01040](https://doi.org/10.3724/SP.J.1146.2012.01040).
ZHANG Zhen, WANG Binqiang, CHEN Hongchang, *et al.* Internet traffic classification based on host connection graph[J]. *Journal of Electronics & Information Technology*, 2013, 35(4): 958–964. doi: [10.3724/SP.J.1146.2012.01040](https://doi.org/10.3724/SP.J.1146.2012.01040).
- [12] SHAFIQ M, YU Xiangzhan, BASHIR A K, *et al.* A machine learning approach for feature selection traffic classification using security analysis[J]. *The Journal of Supercomputing*, 2018, 74(10): 4867–4892. doi: [10.1007/s11227-018-2263-3](https://doi.org/10.1007/s11227-018-2263-3).
- [13] SHI Hongtao, LI Hongping, ZHANG Dan, *et al.* An efficient feature generation approach based on deep learning and feature selection techniques for traffic classification[J]. *Computer Networks*, 2018, 132: 81–89. doi: [10.1016/j.comnet.2018.01.007](https://doi.org/10.1016/j.comnet.2018.01.007).
- [14] WANG Youwei and FENG Lizhou. A new hybrid feature selection based on multi-filter weights and multi-feature weights[J]. *Applied Intelligence*, 2019, 49(12): 4033–4057. doi: [10.1007/s10489-019-01470-z](https://doi.org/10.1007/s10489-019-01470-z).
- [15] 王勇, 周慧怡, 俸皓, 等. 基于深度卷积神经网络的网络流量分类方法[J]. 通信学报, 2018, 39(1): 14–23. doi: [10.11959/j.issn.1000-436x.2018018](https://doi.org/10.11959/j.issn.1000-436x.2018018).
WANG Yong, ZHOU Huiyi, FENG Hao, *et al.* Network traffic classification method basing on CNN[J]. *Journal on Communications*, 2018, 39(1): 14–23. doi: [10.11959/j.issn.1000-436x.2018018](https://doi.org/10.11959/j.issn.1000-436x.2018018).
- [16] REN Xinming, GU Huaxi, and WEI Wenting. Tree-RNN: Tree structural recurrent neural network for network traffic classification[J]. *Expert Systems with Applications*, 2021, 167: 114363. doi: [10.1016/j.eswa.2020.114363](https://doi.org/10.1016/j.eswa.2020.114363).
- [17] LIN S Z, SHI Yong, and XUE Zhi. Character-level intrusion detection based on convolutional neural networks[C]. 2018 International Joint Conference on Neural Networks (IJCNN), Rio de Janeiro, Brazil, 2018: 1–8. doi: [10.1109/IJCNN.2018.8488987](https://doi.org/10.1109/IJCNN.2018.8488987).
- [18] 夏栋梁, 刘玉坤, 鲁书喜. 基于蚁群算法和改进SSO的混合网络入侵检测方法[J]. 重庆邮电大学学报: 自然科学版, 2016, 28(3): 406–413. doi: [10.3979/j.issn.1673-825X.2016.03.021](https://doi.org/10.3979/j.issn.1673-825X.2016.03.021).
XIA Dongliang, LIU Yukun, and LU Shuxi. Hybrid network intrusion detection method based on ant colony algorithm and improved simplified swarm optimization[J]. *Journal of Chongqing University of Posts and Telecommunications: Natural Science Edition*, 2016, 28(3): 406–413. doi: [10.3979/j.issn.1673-825X.2016.03.021](https://doi.org/10.3979/j.issn.1673-825X.2016.03.021).
- [19] LOPEZ-MARTIN M, CARRO B, SANCHEZ-ESGUEVILLAS A, *et al.* Shallow neural network with kernel approximation for prediction problems in highly demanding data networks[J]. *Expert Systems with Applications*, 2019, 124: 196–208. doi: [10.1016/j.eswa.2019.01.063](https://doi.org/10.1016/j.eswa.2019.01.063).
- [20] DASH M and LIU Huan. Consistency-based search in feature selection[J]. *Artificial Intelligence*, 2003, 151(1/2): 155–176. doi: [10.1016/s0004-3702\(03\)00079-1](https://doi.org/10.1016/s0004-3702(03)00079-1).
- [21] ZHANG Hongli, LU Gang, QASSRAWI M T, *et al.* Feature selection for optimizing traffic classification[J]. *Computer Communications*, 2012, 35(12): 1457–1471. doi: [10.1016/j.comcom.2012.04.012](https://doi.org/10.1016/j.comcom.2012.04.012).
- [22] 崔自峰, 徐宝文, 张卫丰, 等. 一种近似Markov Blanket最优特征选择算法[J]. 计算机学报, 2007, 30(12): 2074–2081. doi: [10.3321/j.issn:0254-4164.2007.12.002](https://doi.org/10.3321/j.issn:0254-4164.2007.12.002).
CUI Zifeng, XU Baowen, ZHANG Weifeng, *et al.* An approximate markov blanket feature selection algorithm[J]. *Chinese Journal of Computers*, 2007, 30(12): 2074–2081. doi: [10.3321/j.issn:0254-4164.2007.12.002](https://doi.org/10.3321/j.issn:0254-4164.2007.12.002).
- [23] MOORE A W. Dataset[EB/OL]. <https://www.cl.cam.ac.uk/research/srg/netos/nprobe/data/papers/sigmetrics/index.html>, 2005.
- 唐 宏: 男, 1967年生, 教授, 研究方向为计算机网络、移动通信。
刘 丹: 女, 1995年生, 硕士生, 研究方向为网络管理、机器学习。
姚立霜: 女, 1993年生, 硕士生, 研究方向为网络管理、机器学习。
王云锋: 男, 1992年生, 硕士生, 研究方向为机器学习、数据挖掘。
裴作飞: 男, 1994年生, 硕士生, 研究方向为机器学习、数据挖掘。

责任编辑: 余 蓉