https://doi.org/10.1051/jnwpu/20203820341

一种时效增强的机载网络流量识别方法

吕娜[1], 周家欣[1], 冯煊[2], 陈柯帆[1], 陈昨[3]

(1.空军工程大学 信息与导航学院, 陕西 西安 710077;
2.中国人民解放军 31006 部队, 北京 100000; 3.西北工业大学 网络空间安全学院, 陕西 西安 710072)

摘 要:机载网络拓扑动态性强,带宽受限等特点导致其难以为多样化的航空集群作战任务提供可靠的信息交互服务,因此需要对网络中的"大流量对象"进行实时识别,从而优化流量控制,提升网络性能。针对该问题,基于机器学习贝叶斯模型,提出一种时效增强的流量识别方法,首先通过对原始流量数据集进行预处理得到数据流训练子集,并基于贝叶斯网络模型构造子分类器,然后基于多窗口动态贝叶斯网络分类器模型实现大流量对象的早期识别。仿真结果表明,相较于现有的大流识别方法,所提方法可以在保证识别准确性的条件下有效提升识别时效性。

关 键 词:流量识别;机器学习;贝叶斯网络;航空集群;机载网络
中图分类号:TN915.851 **文献标志码**:A **文章编号**:1000-2758(2020)02-0341-10

随着航空电子技术的迅速发展以及战争理论的不断演进,航空平台间以集群方式连续执行观察-定位-决策-打击(observe-orient-decide-act, OODA)不同阶段作战任务已成为未来典型的空战模式,应用于航空集群作战的机载网络近年来也成为了研究热点[1-3]。航空集群多样化的作战任务导致机载网络业务流种类较多,不同种类业务流对传输带宽、时延需求差异性较大。为优化网络流量控制,有效提升机载网络性能,需要对网络流量识别方法展开研究。

统计与研究发现,各类通信网络中普遍存在"重尾分布"特性,即网络中大流量对象(本文中简称为"大流")数量较少,却占用了大量网络带宽,对网络整体性能影响较大[4]。基于此特性,实时、准确地识别网络中的大流对于优化网络传输带宽使用、提升网络性能具有重要意义,也使大流识别问题成为流量识别领域的焦点问题之一。相较于地面有线网络,航空集群机载网络存在网络拓扑动态变化、链路状态不稳定、传输带宽受限等特点,导致数据流对传输带宽更为敏感,因此大流的识别问题具有更高的研究价值。此外由于航空集群作战环境复杂,对抗性较强,作战任务具有高时间敏感性,一方面造成航空集群机载网络中的流量分布呈现高度不平稳的分布特征,另一方面也造成航空集群机载网络中的大流识别问题需要更加关注时效性。传统的大流识别方法,如基于抽样的方法[4-5]、基于 LRU(least recently used)队列管理[6]的方法普遍采用"后验统计"方式实现数据流中大流对象的提取,在数据分组到达分布极其不平稳的航空集群机载网络中难以获取理想的识别准确性,同时,由于该类方法需要通过统计的方式完成大流的判断,因此不具有大流预测能力,难以满足航空集群机载网络流量识别的时效性需求。

近年来,随着人工智能的兴起,机器学习在网络流量识别领域获得了越来越多的关注。基于机器学习的流量识别方法通过挖掘数据流多维特征与类别之间的关系,可以有效提升识别准确性,通过控制数据流特征的提取范围,也可以实现数据流类型的早期识别,从而为流量识别时效性的提升提供了新的思路[7]。目前,相关研究者已将贝叶斯分类器模型[8-9]、决策树模型[10-11]以及支持向量机模型[12-13]等经典机器学习算法引入流量识别应用,且均获得了较为理想的分类准确率。文献[7]提出了一种基于机器学习的大流量对象识别方法,该方法通过对

收稿日期:2019-06-21 **基金项目**:陕西省重点研发计划(2017GY-069)资助
作者简介:吕娜(1970—),女,空军工程大学教授,主要从事航空数据链、通信网络研究。

数据流的前部数据包特征进行提取,并通过多种机器学习算法实现了多种分类器模型构建。其实验结果显示该方法可以获得较高的大流识别准确性,表明通过机器学习可以实现数据流类型的早期判别,从而提升识别时效性。然而该方法仅对识别准确性与采集数据子流之间的关系进行了静态分析,而未提出动态的识别方法,因此实际应用价值有限。

为实现数据流类型的动态早期判断,需要分类器动态地根据采集到的数据子流特征为数据流类型预测提供边缘概率信息,这也使机器学习算法的选择成为关键问题之一。作为经典机器学习算法之一,贝叶斯分类器可以通过对某类对象特征先验分布的统计,利用贝叶斯公式计算得到该类对象所属类别的边缘分布概率,从而为数据流类型的动态识别提供条件。目前相关研究者已将贝叶斯分类器中的朴素贝叶斯算法引入流量识别,然而该模型要求样本特征间满足条件独立性假设,造成其无法客观地刻画各特征间的概率模型,因此其条件分布结果的准确性存在局限。

贝叶斯网络分类器(Bayesian network classifier, BNC)模型通过复杂的网络拓扑和各节点之间的条件概率关系,消除了朴素贝叶斯模型中特征间的条件独立性约束,可以获得更为客观的概率模型[14-15],常用于复杂系统内部的定性推理与数理分析。因此本文拟基于采用机器学习贝叶斯网络模型研究航空集群机载网络中的流量识别问题,提出一种时效增强型流量识别方法。该方法以数据流的前部子流段为对象,通过选取多维特征,构建一组贝叶斯网络分类器,从而实现流量分类;在此基础上,设计多窗口动态贝叶斯网络分类器模型,在保证识别准确性的前提下,有效提升识别时效性。

1 大流识别与贝叶斯网络相关理论

1.1 大流识别问题

数据流的“重尾分布”特性广泛存在于各类通信网络中,图 1 为某机载网络骨干节点在 30 min 内捕获到的数据流分布情况。随着数据流所包含数据包数量的增长,其所对应的概率密度迅速下降。进一步统计发现在该时间段内,包含数据包数量最多的前 1%的数据流所含数据包数量占该时间段内该节点捕获数据包总数的 58%。

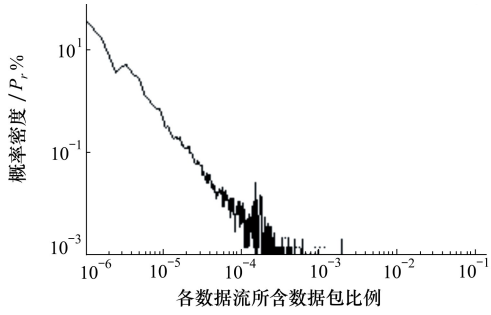


图 1 某机载网络中的数据流分布情况

基于大流对通信网络性能的显著影响,大流的识别具有重要意义,因此也得到了研究人员的广泛关注。目前学术界关于大流的定义通常根据其具体研究的网络环境实际情况出发,因此不存在统一的定义标准。常见的定义方式是一段时间内所含总数据分组数量超过某一固定阈值(如总链路容量的 0.1%)[4]的数据流。在此定义下,基于抽样、LRU 缓存队列、哈希散列函数的多种传统大流识别方法被提出。以上的传统大流识别方法通常关注传输带宽较大、数据传输速率较高、业务量大的地面有线网络,多从控制识别开销的角度出发,通过抽样、统计以及估计等手段实现大流的识别,并未充分考虑识别的时效性需求。与之相比,航空集群机载网络具有可用带宽受限、链路通断状态不稳定等特点,大流的出现对网络整体性能的影响更加明显,需要及时地发现大流并进行管控,因此,航空集群机载网络的大流识别问题更加关注于识别的时效性。

针对大流识别的时效性需求,文献[7]基于早期识别思想,通过对数据流前部固定数量的数据分组特征的提取,基于多种机器学习模型,构建了不同的分类器,从而在分析数据流前部几个数据分组的条件下实现大流的识别,该研究成果表明,机器学习方法为大流识别的时效性提升提供了条件。然而该文献仅对提取的数据流前部固定长度的数据分组特征进行了静态的分析,其选用的流量训练集与测试集较为理想,大流、小流样本的分布均衡,未充分考虑样本数量严重失衡对分类器训练的不利影响;此外,该方法也未充分考虑网络流量分布的随机性与流量样本尺度的差异性,无法根据数据流的实际分布情况灵活选取特征提取范围,未对其在流量分布呈现高随机性特点的实际网络中的应用问题展开探讨,因此无法直接应用于航空集群机载网络的流量识别。

1.2　贝叶斯网络相关理论

本文设计的时效增强性流量识别方法基于贝叶斯网络模型。作为贝叶斯分类器模型中的一种,贝叶斯网络模型由以下 2 个部分构成:

1) 结构部分　通过有向无环图 G 描述,无环图中的节点集由实例的特征集 $\boldsymbol{F}=\{f_1,f_2,\cdots,f_k\}$ 与目标类 $\boldsymbol{C}=\{c_1,c_2,\cdots,c_m\}$ 构成,用于定性地表示实例各特征及类别之间的依赖关系,其中每一个节点代表样本的一个类型或一项特征,节点之间的连接关系表示特征间或特征与类型间的依赖关系,具有直接连接关系的 2 个节点之间互为父子节点关系。

2) 参数部分　通过与各节点关联的条件概率表(condition probability table,CPT)Θ 描述,节点 N_i 的 CPT 定量地反映了当前节点关于其父节点集 $P_a(N_i)$ 的条件概率。

完整的 BNC 可通过以下公式计算得到特征与类别的联合概率 $P(U)$

$$P(U)=P(f_1,f_2,\cdots f_n,C)=\prod_{i=1}^{n+1}P(N_i\mid P_a(N_i)) \tag{1}$$

对于一组实例样本数据集 $\boldsymbol{X}=\{x_1,x_2,\cdots,x_n\}$,若该数据集中的全体实例分属于 m 个目标类 $\boldsymbol{C}=\{c_1,c_2,\cdots,c_m\}$,且选取的实例特征集为 $\boldsymbol{F}=\{f_1,f_2,\cdots f_k\}$,令 N_l 为 f_l 在贝叶斯网络有向无环图中对应的节点,当给定实例 x_i 的特征取值向量 $\boldsymbol{F}_i=(f_{i1},f_{i2},\cdots,f_{in})$ 与数据流目标类的先验概率 $P(c_j)$ 的条件下,基于 BNC 模型依据极大似然估计准则给出分类判决 q_i,即

$$\begin{aligned}q_i&=\arg\max_{c_j\in\boldsymbol{C}}P(c_j\mid\boldsymbol{F})=\\&\arg\max_{c_j\in\boldsymbol{C}}\frac{P(f_{i1},f_{i2},\cdots,f_{in},c_j)}{P(f_{i1},f_{i2},\cdots,f_{in})}=\\&\arg\max_{c_j\in\boldsymbol{C}}\frac{\prod_{l=1}^{k}P(N_l\mid P_a(N_l))\times P(c_j)}{P(f_{i1},f_{i2},\cdots,f_{in})}=\\&\arg\max_{c_j\in\boldsymbol{C}}\prod_{l=1}^{k}P(N_l\mid P_a(N_l))\times P(c_j)\end{aligned} \tag{2}$$

2　基于贝叶斯网络的流量识别方法

基于贝叶斯网络的流量识别方法的核心思想是通过前部子流窗口(initial sub-flow window,ISW)完成原始流量数据集的预处理,生成多个不同长度的数据流前部子流段(initial sub-flow field,ISF)样本训练子集,再利用贝叶斯网络模型对生成的多个训练子集进行学习,从而生成多个基于数据流前部子流的分类器实现数据流类型的早期识别。

2.1　基于前部子流的分类器模型构建

为避免分类器训练过程中出现的分类失衡现象,并实现大流的早期发现,本文通过前部子流窗口提取数据流的前部子流段特征,取代完整数据流特征作为训练及测试对象,实现一组前部子流段分类器(initial sub-flow field classifier,ISFC)的构建。

本文将前部子流窗口定义为在一定窗口生存期(window time to live,WTTL)内保持对独立数据流 ISF 记录的窗口,其中 WTTL 为 ISW 维持对单个数据流采集的时限;窗口值(ISW value,ISWV)表示 ISW 对单个数据流内 ISF 数据包采集数量。根据作用阶段,ISW 分为训练窗口(ISW for training,ISW-T)与捕获窗口(ISW for capture,ISW-C)。ISW-T 作用于 ISFC 线下训练阶段,实现训练数据流实例 ISF 特征的提取与训练集的过滤;ISW-C 作用于数据流线上识别阶段,实现对测试集中数据流 ISF 的捕获。

针对机载网络中基于 ISF 特征的大流识别,本文构造了多窗口动态贝叶斯网络分类模型(multi-window dynamic bayesian network classifier,MWD-BNC),如图 2 所示。

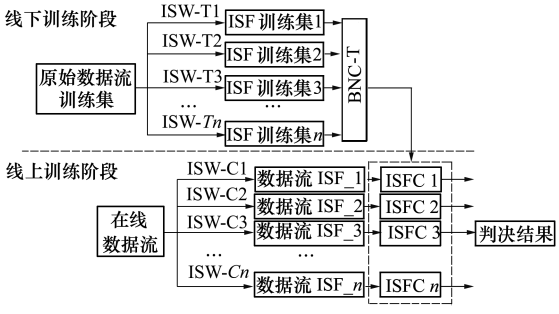


图 2　MWD-BNC 模型结构

在 ISFC 线下训练阶段,设置 n 个训练窗口值递增的 ISW-T,通过提取原始数据集流量在不同窗口值下的前部子流特征,从而将原始数据流训练集转换为 n 个数据流 ISF 子训练集,之后通过贝叶斯网络搜索算法生成 n 个相应的子 ISFC,最后按照窗口值升序,将 n 个子 ISFC 的依次组合,从而构成 MWD-BNC。在数据流线上识别阶段,相应设置 n 个窗口值与 ISW-T 相同的 ISW-C 实现线上对未知数据流 ISF 的捕获,提取 ISF 测试实例特征信息,然后

输入 MWD-BNC,动态分析识别准确性与时效性之间的平衡关系,并通过相应窗口值下的 BNSC 实现未知大流样本的识别。

2.2 基于 ISW-T 的训练集构建

受“重尾分布”特性的影响,原始流量数据集中存在的大量冗余负例(小流)样本将导致以此数据集构建的分类器趋向于牺牲正例(大流)的识别准确性以换取较高的整体识别准确性,即分类失衡现象[16]。然而相比于大量的小流,大流具有更高的识别价值,可以通过调整训练集的正负例比例构建偏向于大流的分类器。

本文以正负例数量的比例 $\eta = N_{\text{pos}}/N_{\text{neg}}$ 衡量训练集中正负例失衡程度。其中 N_{pos} 为训练集中大流数量,而 N_{neg} 则为小流数量。η 越小表明当前训练集中正负例比例越不平衡,而随着 η 的增大,基于该训练集构建的分类器逐渐偏向正例。以图 1 所示的机载网络流量数据集为例,设置大流判定阈值 θ_p = 0.01%(N = 121 个) 时,数据集 η = 0.009 5,此时以该原始数据集训练得到的 BNC 将出现严重的分类失衡现象。

基于原始网络流量数据集构建正负实例平衡的 ISF 训练集, ISW-T 对数据流 ISF 的提取流程设计如图 3 所示。

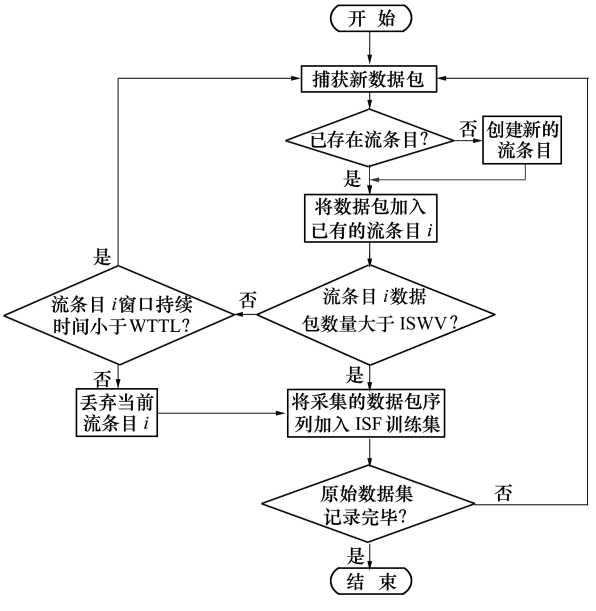


图 3 ISW 对数据流 ISF 的提取流程

在 MWD-BNC 模型线下训练阶段,根据 ISW-T 对数据流 ISF 的提取流程,所含数据包数量较多的数据流趋向于被 ISW-T 筛选进入 ISF 子训练集中;而对于原始数据集中大量存在的冗余小流实例而言,在 WTTL 内捕获到的数据包数量相对更难满足窗口值条件,因此趋向于被提前滤出 ISF 训练集。基于此现象,ISW-T 可在保留原始数据流训练集中大流 ISF 训练实例的前提下淘汰大量冗余小流 ISF 训练实例,从而使 ISF 训练集中正负例样本数趋于平衡。表 1 为在图 2 所示网络流量数据集中设置 ISW-T 后,不同 β 下获取的 η 取值,可以发现随着窗口值的增大,大量小流实例被预先淘汰,获得的 ISF 训练集的 η 逐渐向 1 逼近。

表 1 不同 β 下的 ISF 训练集正负例比值

β	0.1	0.15	0.2	0.25	0.3
η	0.193 8	0.282 3	0.390 0	0.525 1	0.671 4

2.3 基于 ISW-C 的数据流线上识别

根据 2.1 节中的 MWD-BNC 模型,ISW-C 实现线上数据流 ISF 的捕获,其窗口值为当前 ISFC 对特定数据流作出分类判决前所需要捕获的数据包数量。设当前机载网络环境中大流判别阈值为 N,定义窗口截取比 β 为捕获窗口值 I_C 与当前网络中大流判别阈值 N 的比,即

$$\beta = \frac{I_C}{N}(\beta < 1) \tag{3}$$

β 反映当前 ISFC 实现大流识别的时效性。随着 I_C 的增大,β 逐渐向 1 逼近,此时 ISFC 的识别时效性逐渐下降。

设数据流测试集为 $\boldsymbol{X} = \{x_1, x_2, \cdots, x_n\}$,其中 $x_1, x_2, \cdots, x_n$ 为分属于目标类 $\boldsymbol{C} = \{E, M\}$ 的独立数据流实例,$\boldsymbol{F} = \{f_1, f_2, \cdots f_k\}$ 为 ISF 特征集。$\boldsymbol{F}^{(i)} = (f_1^{(i)}, f_2^{(i)}, \cdots, f_k^{(i)})$ 表示线上捕获到数据流实例 x_i 中提取的 ISF 特征矢量,$c^{(i)}$ 表示数据流实例 x_i 所属类别,$c^{(i)} \in \boldsymbol{C}$,该 ISFC 的识别过程即实现特定数据流实例 ISF 特征矢量与流所属类别的映射 $\boldsymbol{F}^{(i)} \rightarrow c^{(i)}$。

根据公式(1) 可分别计算得到数据流实例 x_i 特征矢量与两目标类的联合概率,将正例联合概率 $P(c^{(i)} = E, \boldsymbol{F}^{(i)})$ 记作 $P_{\text{pos}}^{(i)}(U)$,负例联合概率 $P(c^{(i)} = M, \boldsymbol{F}^{(i)})$ 记作 $P_{\text{neg}}^{(i)}(U)$,并将数据流实例 x_i 的 BNC 判决参数 $\xi^{(i)}$ 表示为

$$\xi^{(i)} = \frac{P_{\text{pos}}^{(i)}(U)}{P_{\text{neg}}^{(i)}(U)} \tag{4}$$

由公式(2)给出的判定准则,当 $\xi^{(i)} < 1$ 时,分类器对实例 x_i 做出负例(小流) 判决,反之做出正例

(大流)判决。当$\xi^{(i)}=1$时,表明x_i位于正负例判决边界,此时 ISFC 做出随机判决。

对于实例x_i,若$P_{\text{pos}}^{(i)}(U)$与$P_{\text{neg}}^{(i)}(U)$之间相差越大,其距正负例判决边界越远,此时 ISFC 做出的判决更具可信性。为进一步量化 ISFC 对不同实例做出判决的可信性,本文对实例x_i的 ISFC 判决置信度参数$\alpha^{(i)}$定义为

$$\alpha^{(i)}=\frac{\max(\xi^{(i)},1)}{\min(\xi^{(i)},1)} \tag{5}$$

显然$\alpha^{(i)}\geqslant 1$且$\alpha^{(i)}$的取值越大表明当前 ISFC 对实例x_i作出的判决更具可信性。根据实际训练效果,预设一置信度阈值α_T,当 ISFC 对实例x_i作出判决后满足$\alpha^{(i)}\geqslant\alpha_T$时,则采纳该判决结果。

通过表 1 中对不同 ISWV 下构建的训练集的分析发现,随着β的增长,训练集的η向 1 逼近,训练所得的 ISFC 潜在的分类失衡现象也随之减轻;同时,线上分类过程中,随着 ISWV 的增大,大量小流在数据流捕获阶段被预先淘汰,ISFC 的预测空间也随之缩小。然而β的增长意味着线上识别过程中需采集更多的数据包才可作出判决,因此也造成大流识别时效性的下降。识别过程中,在保证准确性的前提下尽量提升时效性,MWD-BNC 模型对流i作出判决时的β应满足

$$\beta_i=\min_{\beta_i}\{\beta_i \mid \alpha^{(i)}\geqslant\alpha_T\} \tag{6}$$

基于以上分析,本文给出 MWD-BNC 模型的线上识别过程如下:

Input:　窗口生存期W_{TTL};
　　ISW-C 数量m;
　　ISW-C 窗口值数组w_s;
　　预设置信度阈值α_T;
1. Begin
2. for $k\leftarrow 1$ to ∞
3. if $X(k)$所属流缓存条目q不存在
4. 创建m,R_b^q
5. 初始化$i^q\leftarrow 1$,$t^q\leftarrow A_T^q(1)$,$j^q\leftarrow 1$
6. $R_b^q(i^q)=X(k)$
7. else
8. if $j^q<m$
9. if $t^q\leqslant W_{\text{TTL}}$
10. $R_b^q(i^q)=X(k)$
11. 记录累计采集数据包$i^q\leftarrow i^q+1$
12. 记录累计时间$t^q\leftarrow t^q+A_T(i^q)$
13. if $i^q>w_s(j^q)$
14. $Q^q\leftarrow Q^q\cup R_b^q(w_s(j^q-1):w_s(j^q))$
15. 提取并更新Q^q内的特征向量F^q
16. $c\leftarrow B^j(F^q)$
17. 根据分类结果计算α^q
18. if $\alpha^q>\alpha_T$
19. $D(q)\leftarrow c$, break
20. else
21. $j^q\leftarrow j^q+1$
22. end if
23. end if
24. end if
25. end if
26. if $D(q)==\phi$
27. $D(q)\leftarrow M$,break
28. end if
29. end if
30. end for
31. End
Output:　$D(X^{(q)})$

其中:$X(k)$表示线上捕获到的第k个数据包;R_b^q表示数据流q的数据包接收缓存数组;i^q表示捕获到的数据流q中的数据包序号;t^q表示数据流q的累积采集时间;$A_T(i^q)$表示数据流q中第i^q个数据包的到达时间间隔;j^q表示数据流q中当前的 ISW-C 序号;Q^q表示数据流q的 ISF 缓存条目数组;$\boldsymbol{F}^q$表示Q^q内提取到的 ISF 特征向量;c为数据流类型,其取值范围为$\{E,M\}$;$B^j(\boldsymbol{F}^q)$表示第j^q个 ISW-C 所对应的贝叶斯网络分类器,其根据提取到的 ISF 特征向量$\boldsymbol{F}^q$对所属流类型c进行判别。

3　实验及仿真分析

3.1　实验环境配置及性能评估指标

3.1.1　实验环境配置

本文中采用 Weka 3.8.3 作为机器学习平台,该软件基于 Java 环境,内置多种机器学习模型,可实现对数据的预处理、执行分类、回归、聚类等多种操作并对数据挖掘结果进行统计与分析。

为实现分类器的训练及验证,选取了某次飞行任务过程中机载网络中的实际流量数据作为原始流量数据集。该数据集保存了 30 min 内经过监测节点的全部数据包的报头信息,记为 ANset。根据 ANset 流量分布情况,将大流判定阈值设置为 100。参照 Moore 数据集中的流量特征选取方式,基于 ANset 中流量样本数据包到达分布统计特征以及数

据包头部字段特征,共同提取了 34 项 ISF 特征,并通过主成分分析(principal component analysis,PCA)特征选取算法降维得到 10 项 ISF 特征如表 2 所示。

表 2 数据流 ISF 选取特征及其描述

序号	特征名称	特征描述
1	protocol	传输层协议
2	duration_M	第 1 个数据包与第 M 个数据包之间持续时间
3	IAT_mean	聚合流中数据包平均到达时间
4	IAT_var	聚合流中数据包到达时间方差
5	IAT_ISW_mean	特定流中数据包平均到达时间
6	range_pktseq_ISW	特定流中数据包序号范围
7	IAN_mean	两连续数据包之间平均序号差
8	num_interval_100	数据包序号间隔在 100 ~ 1 000 之间的出现次数
9	num_interval_1 000	数据包序号间隔在 1 000 以上的出现次数
10	class	数据流类型(大流/小流)

3.1.2 准确性指标

在机载网络中,由于大流数量占比极低而其识别价值远高于小流,因此相较于分类器的整体识别准确性,大流的识别准确性更受关注。大流实例真实类别与分类类别的组合分为真阳性(true positive,TP),真阴性(true negative,TN),假阳性(false positive,FP)与假阴性(false negative,FN),分类结果如表 3 所示。

表 3 二类问题预测结果混淆矩阵

真实结果	预测结果	
	大流	小流
大流	T_P	F_N
小流	F_P	T_N

基于以上概念,本文采用正例查准率 P_{pos} 与正例查全率 R_{pos} 作为大流识别的准确性指标。P_{pos} 与 R_{pos} 分别表示如下

$$P_{\text{pos}} = \frac{T_P}{T_P + F_P} \tag{7}$$

$$R_{\text{pos}} = \frac{T_P}{T_P + F_N} \tag{8}$$

式中:P_{pos} 反映了分类器对大流识别的可信度;R_{pos} 反映了分类器对大流识别的覆盖度。

3.1.3 时效性指标

本文设计监测阈值比(monitor-threshold ratio,MTR)M 作为大流监测的时效性指标,将 MTR 定义为分类器对大流做出正确判决前,该流经过监测节点数据包数量占大流判定阈值的比值。设测试集中被分类器准确识别的大流数量为 m 条,大流判定阈值为 N,分类器对大流 i 作出准确判决时流 i 已有 n_i 数据包经过该节点,则该流的监测阈值比

$$M_i = \frac{n_i}{N} \tag{9}$$

该分类器基于此测试集的平均监测阈值比为

$$M_{\text{avrr}} = \frac{1}{m}\sum_{i=1}^{m}\frac{n_i}{N} \tag{10}$$

在本文提出的 MWD-BNC 模型中,任意一条大流 i 的 MTR 等于其被 BNC 准确识别时对应的 ISW-C 的窗口截取比 β_i,此时该模型基于测试集的 M_{avrr} 表示为

$$M_{\text{avrr}} = \frac{\sum_{i}^{m}\beta_i}{m} \tag{11}$$

对于一分类器,M_{avrr}越小,表明其具有更高的识别时效性。

3.2 识别性能评估

本文以 ANset 为原始数据集,分别选取:基于贝叶斯定理的周期性采样方法(periodically sampling based on bayesian theory,PS-BT)、基于 LRU 的方法(LRU_1)以及采样-保持方法(sample and hold,S&H)作为对照方法,验证所提出方法在机载网络下的大流识别性能。

3.2.1 模型训练正确性评估

首先对不同窗口值下的贝叶斯网络子分类器训练正确性进行评估。分别选取朴素贝叶斯算法(naïve bayesian,NB)、C4.5 决策树算法(C4.5 decision tree)以及贝叶斯网络分类器算法(bayesian network classifier,BNC),对不同窗口截取比下获得的训练子集训练所得 ISFC 识别准确性对比如图 4 所示。可以发现,NB 模型在 P_{pos} 方面表现最优,然而随着 β 的增长,相应 ISFC 的 R_{pos} 呈现整体下降趋势;而在 BNC 模型与 C4.5 模型下,随着 β 的增长,ISFC 的 P_{pos} 与 R_{pos} 均呈整体上升趋势。C4.5 与 BNC 在 P_{pos} 上表现相近,而在 R_{pos} 上 BNC 表现显著优于 C4.5。

3.2.2 准确性指标对比

为验证贝叶斯网络模型对大流识别准确性的提升效果,基于 3.1.1 节中选取的 10 项数据流特征,以 ANset 周期性采样数据集为训练集构建贝叶斯网络

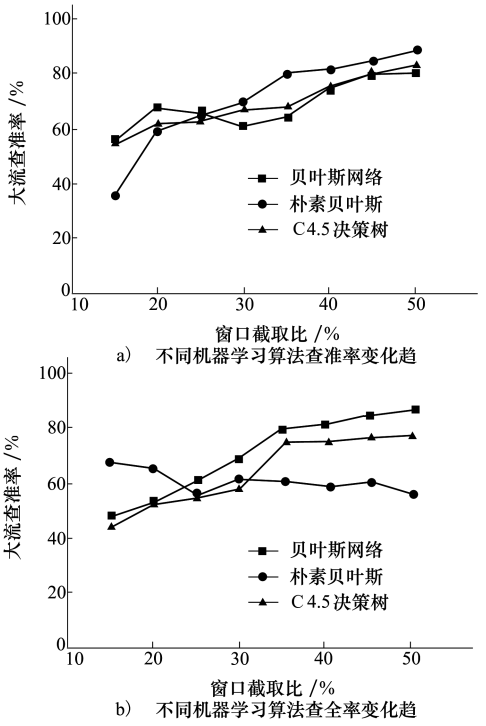


图 4　不同模型下大象流识别查准率与查全率

分类器,并对线上数据流进行周期性采样后所得的数据包组进行识别,将该方法记为 BNC-PS,流程如图 5 所示。

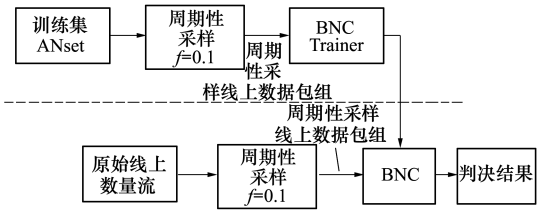


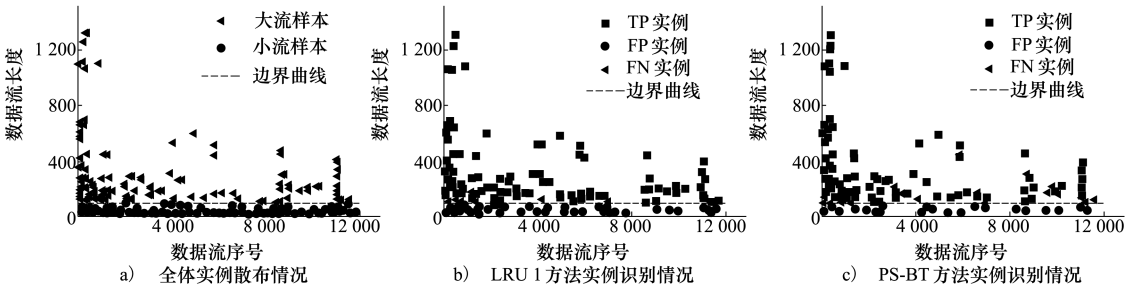
图 5　BNC-PS 方法流程图

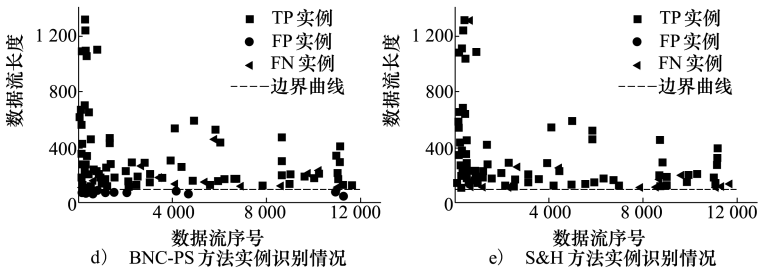
为控制变量,本文在 PS-BT、BNC-PS 与 S&H 方法中均设置了相同的采样率,并通过调节 LRU_1 方法中相关参数,使 3 种方法获得相近的 R_{pos},在此基础上以 P_{pos} 衡量各种方法的准确性,各方法参数设置如表 4 所示。

表 4　各识别方法参数配置情况

采用方法	参数配置
PS-BT	抽样率 $f=0.1$,测量间隔 $M_I=10$ min,采样识别阈值 $\hat{T}=6$
LRU_1	$C_q=70$,测量间隔 $M_I=10$ min
S&H	抽样率 $f=0.1$
BNC-PS	抽样率 $f=0.1$,测量间隔 $M_I=10$ min

图 5 所示为 ANset 在各方法下的识别情况,其中图 5a)为 ANset 中全体数据流实例分布情况,图 5b)~5e)分别为以 LRU_1 方法、PS-BT 方法、BNC-PS 方法以及 S&H 方法所获得的大流实例识别情况。4 种方法具有相近的 TP,但在 FP 方面,4 种方法的表现存在明显差异,其中,采用 LRU_1 方法时,FP 相对较高,识别结果可信度最低,这是因为本文中采用的 ANset 数据集中存在的大量突发性小流极易将大流记录条目替换出 LRU 缓存队列,该方法存在更高的误识别代价;S&H 方法以较大的计算开销为代价,保持了对各采样数据流的包计数,避免了误识别情况的发生,具有最高的可信度,而其对实例的覆盖度取决于抽样率的大小,因此,可认为采用 S&H 方法可以获取固定抽样率条件下准确性上限。PS-BT 方法通过设置采样识别阈值提升了大流识别门限,其 FP 相较于 LRU_1 方法大大降低,识别结果可信度有一定提升,另一方面,虽然设置了与 S&H 算法相同的抽样率,但抽样判决的过程引入了一定的误差,因此其识别覆盖度低于 S&H。





d) BNC-PS 方法实例识别情况 e) S&H 方法实例识别情况

图 6 不同方法下大流实例识别分布情况

采用 BNC-PS 方法识别可信度较 LRU_1 与 PS-BT 方法有了显著提升,在识别覆盖度方面,BNC-PS 方法的 TP 也有所上升,逼近了 S&H 方法所达到的特定抽样率下的准确性上限。各方法识别准确性指标如表 5 所示。

表 5 不同方法下大流识别准确性能指标

识别方法	TP 值	FP 值	P_{pos}	R_{pos}
LRU_1	119	72	0.856	0.613
PS-BT	118	36	0.849	0.766
BNC-PS	124	17	0.892	0.879
S&H	127	0	0.914	1

3.2.3 时效性指标对比

本文选取不同 α_T 下的 MWD-BNC 模型与 PS-BT 方法、LRU_1 方法以及 S&H 方法,验证不同的 MTR_{avrr} 下的识别准确性对比,其中,MWD-BNC 模型参照表 6 的配置方式,PS-BT、LRU_1 以及 S&H 方法的设置参照表 4。

表 6 MWD-BNC 参数配置情况

配置项	配置参数
WTTL(ISW-C)	15 min
ISW-C 数量	8 个
ISW-C 窗口值	(15,20,25,…,50)
置信度阈值设置	{2,5,10}

图 7 表示不同方法下 R_{pos} 随 MTR 的变化趋势,其中,LRU_1 方法在最低的 MTR(约 25%)下开始识别大流;而 PS-BT 方法的 R_{pos} 曲线相对滞后于 LRU_1 方法,2 种方法均在 $M=600\%$ 左右实现 R_{pos} 收敛;采用 S&H 方法时,由于保持了对采样数据流的包计数,全部大流在 $M=100\%$ 时开始被识别,R_{pos} 也在同时实现收敛。相比之下,MWD-BNC 模型的 R_{pos} 曲线远超前以上 3 种方法,本文设置 MWD-BNC 模型的最小 ISW-C 窗口值为 15,因此不同 α_T 下的 MWD-BNC 模型均在 $M=10\%$ 后开始识别大流,$\alpha_T=2$ 时 R_{pos} 在最低的 MTR 下实现收敛。

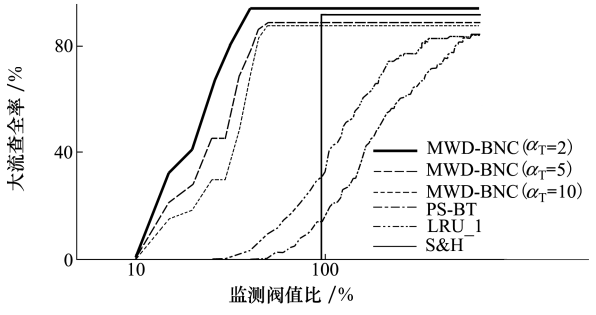


图 7 MWD-BNC 模型查全率收敛曲线

图 8 为不同方法下 FP 随 MTR 的变化,可以发现,S&H 方法通过保持计数的方式,可完全避免误识别情况的发生,因此 FP 始终保持为 0;采用 LRU_1 方法时,FP 的增长速度最快,最终的收敛值也最高;PS-BT 方法的 FP 曲线相对滞后于 LRU_1 方法,且最终的 FP 收敛值则更低,该结果表明,相比于 PS-BT 方法,LRU_1 方法虽具有较高的识别时效性,但其识别结果可信度较低。

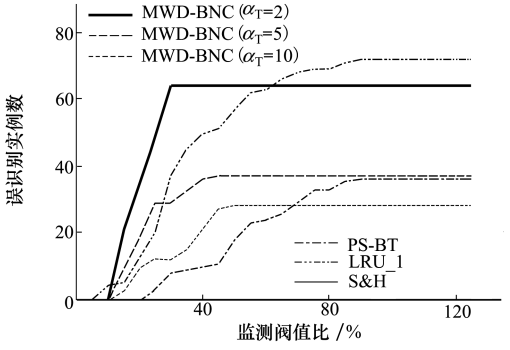


图 8 MWD-BNC 模型 FP 收敛曲线

与传统大流识别方法相比,MWD-BNC 模型的 FP 随各自 R_{pos} 同步增长,并几乎在相同的 MTR 值处实现收敛。其中,$\alpha_T=2$ 时的 FP 增长速度最快,最终 FP 收敛值也最高;而 $\alpha_T=10$ 的 FP 增长速度最

慢,并可获得最低的 FP 收敛值,表明在 MWD-BNC 模型中设置较高的 α_T 可以获得更高的识别可信度。各方法的时效性指标如表 7 所示。

表 7　不同方法下大流识别时效性指标　　%

识别方法	$M_{initial}$	M_{conv}	M_{avrr}
MWD-BNC($\alpha_T=2$)	15	40	24.08
MWD-BNC($\alpha_T=5$)	15	50	28.90
MWD-BNC($\alpha_T=10$)	15	50	33.24
LRU	30	600	214.23
PS-BT	45	585	212.58
S&H	100	100	100

4　结　论

本文基于机器学习方法中的贝叶斯网络分类器模型,结合航空集群机载网络的流量分布特点,提出了一种时效增强型流量识别方法,首先以数据流前部子流段训练窗口实现原始流量数据集的预处理,获取了一组基于数据流前部子流段的训练子集,然后基于贝叶斯网络模型对训练子集进行学习生成一组贝叶斯网络分类器实现大流量对象的识别。在此基础上构建了 MWD-BNC 模型,实现了时间代价敏感的机载网络大流识别。最后本文以实际网络流量数据集 ANset 进行了分类器的训练实验,并对本文所提方法的性能进行了分析,结果表明,本文所提方法可在保证识别准确性的前提下,有效提升识别时效性,实现大流的早期识别。

参考文献:

[1] 霍大军. 网络化集群作战研究[M]. 北京: 国防大学出版社,2013
HUO Dajun. Operation of Network Swarm[M]. Beijing: National Defence University Press, 2013 (in Chinese)

[2] 赵尚弘,陈柯帆,吕娜, 等. 软件定义航空集群机载战术网络[J]. 通信学报, 2017(8): 140-155
ZHAO Shanghong, CHEN Kefan, LYU Na, et al. Software Defined Airborne Tactical Network for Aeronautic Swarm[J]. Journal on Communications, 2017(8): 140-155 (in Chinese)

[3] 梁一鑫, 程光, 郭晓军,等. 机载网络体系结构及其协议栈研究进展[J]. 软件学报, 2016, 27(1): 96-111
LIANG Yixin, CHENG Guang, GUO Xiaojun, et al. Research Progress on Architecture and Protocol Stack of the Airborne Network[J]. Journal of Software, 2016, 27(1): 96-111 (in Chinese)

[4] ESTAN C, VARGHESE G. New Directions in Traffic Measurement and Accounting[J]. ACM Transactions on Computer Systems, 2003, 21(3): 270-313

[5] MORI T, UCHIDA M, KAWAHARA R, et al. Identifying Elephant Flows through Periodically Sampled Packets[C]// The Institute of Electronics, Information and Communication Engineers, 2004

[6] BAI Lei, TIAN Liqin, CHEN Chao. Elephant Flow Detection Algorithm for High Speed Networks Based on Flow Sampling and LRU[J]. Computer Applications and Software, 2016(4): 111-115

[7] HUANG Y H, SHIH W Y, et al. A Classification-Based Elephant Flow Detection Method Using Application Round on SDN Environment[C]// Asia-Pacific Nework Operation and Management Symposium, 2017

[8] MOORE A W, ZUEV D. Internet Traffic Classification Using Bayesian Analysis Techniques[J]. ACM Sigmetrics Performance Evaluation Review, 2005, 33(1):50

[9] WU K, KE J. A Scheme of Real-Time Traffic Classification in Secure Accsee of Power Enterprise Based on Improved Naïve Bayesian Classification Algorithm[C]// IEEE Internationnal Conference on Software Engineering & Service Science, 2017

[10] 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
XU Peng, LIN Sen. Internet Traffic Classification Using C4.5 Decision Tree[J]. Journal of Software, 2009, 20(10): 2692-2704 (in Chinese)

[11] TONG Da, QU Y R, PRASANNA V K. Accelerating Decision Tree Based Traffic Classification on FPGA and Multicore Platforms[J]. IEEE Trans on Parallel & Distributed Systems, 2017, 28(11): 3046-3059

[12] CAO Jie, FANG Zhiyi, QU Guannan, et al. An Accurate Traffic Classification Model Based on Support Vevtor Machines[J]. Networks, 2017, 27(1): 1962

[13] SUN Guanglu, CHEN Teng, SU Yangyang, et al. Internet Traffic Classification Based on Incremental Support Vector Machines

[J]. Mobile Networks & Applications. 2018, 23(14): 1-8

[14] DOGUC O, RAMIREZ, MARQUEZ J E. A Generic Method for Estimating System Reliability Using Bayesian Networks[J]. Reliability Engineering & System Safety, 2017, 94(2): 542-550

[15] FRIEDMAN Nir, GEIGER D, GOLDSZMIDT M. Bayesian Network Classifiers[J]. Machine Learning, 1997, 29(2/3): 131-163

[16] NG W W, HU J, YEUNG D S, et al. Diversified Sensitivity-Based Undersampling for Imbalance Classification Problems[J]. IEEE Trans on Cybernetics, 2017, 45(11): 2402-2412

A Timeliness-Enhanced Traffic Identification Method in Airborne Network

LYU Na[1], ZHOU Jiaxin[1], FENG Xuan[2], CHEN Kefan[1], CHEN Wu[3]

(1.School of Information and Navigation, Air Force Engineering University, Xi'an 710077, China;
2.PLA 31006 Troops, Beijing 100000, China;
3.School of Cybersecurity, Northwestern Polytechnical University, Xi'an 710072, China)

Abstract: High dynamic topology and limited bandwidth of the airborne network make it difficult to provide reliable information interaction services for diverse combat mission of aviation swarm operations. Therefore, it is necessary to identify the elephant flows in the network in real time to optimize the process of traffic control and improve the performance of airborne network. Aiming at this problem, a timeliness-enhanced traffic identification method based on machine learning Bayesian network model is proposed. Firstly, the data flow training subset is obtained by preprocessing the original traffic dataset, and the sub-classifier is constructed based on Bayesian network model. Then, the multi-window dynamic Bayesian network classifier model is designed to enable the early identification of elephant flow. The simulation results show that compared with the existing elephant flow identification method, the proposed method can effectively improve the timeliness of identification under the condition of ensuring the accuracy of identification.

Keywords: traffic classification; machine learning; Bayesian network; aeronautic swarm; airborne network; model; simulation; identification of elephant flow