ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 35,No. 2,Mar. 2020,pp. 331－343
DOI:10. 16337/j. 1004-9037. 2020. 02. 015

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

基于射频信号特征的Airmax设备指纹提取方法

季　澈[1]，彭林宁[1,2]，胡爱群[1,2]，王　栋[2]

(1. 东南大学网络空间安全学院,南京,211189;2. 网络通信与安全紫金山实验室,南京,211189)

摘　要：针对私有协议的Airmax设备,提出了一种新的射频指纹提取方法。首先,介绍了软硬件实验环境的搭建并简要介绍了Airmax技术,然后介绍了帧前导信号的提取方法,分为粗定位和精确定位,接着从理论分析和实验验证阐述了Airmax射频指纹的提取方法。提取的特征维数为14,其中频率偏移相关的特征有2个,幅度相关的特征有12个。最后,基于这14维特征使用K-means算法及决策树模型对设备特征数据集进行了训练和分类,计算了分类准确率,两个模型的准确率都达到了100%,对于4个设备的分类问题,K-means算法的准确率为92.4%,决策树模型的准确率为100%。

关键词：前导码提取;决策树;K-means算法;射频指纹;设备识别

中图分类号：TP312　　**文献标志码：**A

Fingerprint Extraction Method of Airmax Equipment Based on Radio Frequency Signal Characteristics

JI Che[1], PENG Linning[1,2], HU Aiqun[1,2], WANG Dong[2]

(1. Department of Cyberspace Security, Southeast Univercity, Nanjing, 211189, China; 2. Network Communication and Security Zijinshan Laboratory, Nanjing, 211189, China)

Abstract: A new radio frequency (RF) fingerprint extraction method is proposed for Airmax devices with proprietary protocols. Firstly, the construction of software and hardware experimental environment and the Airmax technology are introduced. Then the extraction method of the frame preamble signals is introduced, which is divided into two rough positioning and precise position. And the extraction method of Airmax RF fingerprint is expounded from theoretical analysis and experimental verification. A total of 14 dimensional features are extracted, in which 2 features are related to the frequency and 12 features are related to the amplitude. Finally, based on the 14-dimensional features, the K-means algorithm and the decision tree model are used to train and classify the data of features, and the classification precision is calculated. The precision of both models reach 100%. For the classification problem of four devices, the precisions of the K-means algorithm and the decision tree model are 92.4% and 100%, respectively.

Key words: preamble extraction; decision tree; K-means algorithm; radio frequency (RF) fingerprint; device identification

---

基金项目：国家自然科学基金(61941115)资助项目;江苏省重点研发计划(BE2019109)资助项目;网络通信与安全紫金山实验室资助项目。

收稿日期：2019-10-29;**修订日期：**2019-12-31

引　言

近年来,通讯设备,主要包括个人通讯设备和工业通讯设备的数量呈现飞速增长[1],其主要原因是无线通信理论及其应用的发展。通讯设备的广泛使用必然带来对设备安全隐患的研究和预防,其中无线设备的接入安全作为无线网络安全的重要一环,也受到学界重视并得到了深入研究。主要研究内容是无线接入点和终端用户对通信设备的识别和认证。

传统的认证模式,多以用户设备提供的认证信息为认证目标。例如,应用IPsec(IP security)协议的IPv6技术,通过ESP(Encapsulating security payloads)协议指定加解密算法和认证算法,通过IKE(Internet key exchange)协议来进行密钥的管理和交换,用户提供hash值等身份信息作为认证目标,AH协议可以使接收方认证发送方的身份信息,同时确保数据的完整性。

随着量子计算机的出现,计算机的计算性能大幅提升,密码学领域的现有研究亟待突破,此外,身份信息的泄露也是影响当前认证体系的问题之一。近年来,使用物理手段获取设备的身份信息这一方向得到了广泛研究,其中通过电磁波提取设备的射频指纹作为身份标识是一大热点[2-7]。

射频指纹的提取方法主要分为稳态响应的指纹提取和瞬态响应的指纹提取[8-10],其中基于瞬态响应的指纹提取主要是依据系统开启或者关闭的瞬间电磁波的包络和相位信息等对于不同的设备有差异性。瞬态响应受信噪比的影响较大,且捕捉难度较大,所以基于瞬态响应的指纹提取应用范围相对较窄。

基于稳态响应的指纹提取主要是对基带信号的频率偏移,I/Q偏移,幅度相位误差等射频特征的提取。由于基带信号相较瞬时信号稳定很多,所以基于稳态响应的指纹提取应用范围广于瞬态响应。

在基于稳态响应的指纹提取中,应用最多的是对频率偏移的提取,而在频率偏移的提取过程中,需要知道信号协议来进行帧前导的定位。不同协议的信号需要使用不同的方法进行提取,普适性较差,且遇到不同协议信号同时传输的情况难以处理,另外,若不知道信号具体的传输协议,则无法有效提取频偏特征。

本文试图找到一种对所有协议通用的射频指纹提取方法,主要是找到一种通用的前导码定位方法,因此本文研究Airmax信号这一基于私有协议的信号。本文以UBNT(Ubiquiti networks)公司生产的Rocket M2系列网桥为实验对象,用两台网桥和两台主机搭建一个局域网,通过USRP设备接收网桥发出的信号,使用信号处理的手段获取频率偏移等射频指纹特征,最后使用决策树,k近邻等机器学习模型对网桥设备进行分类识别和验证。由于Airmax信号的传输协议是一种私有协议,外部人员无法得知协议具体内容,以往的基于协议的提取方法都无法直接使用,本文没有破解信号协议,而是采用物理手段,即信号处理的一些方法,实现了频偏等射频特征的提取。

1　实验环境的搭建

本文所使用的硬件主要包括Rocket M2系列网桥以及USRP(Universal software radio peripheral)通用软件无线电外设平台和电脑主机。

软件无线电平台由Ettus公司生产的USRP N210主机和CBX射频子板构成,工作频段为1.2～6 GHz,最大采样率25 MS/s,上位机采用Linux系统接收和处理数据,实物图如图1所示。

Airmax设备的频点设定为2.4 GHz,USRP的采样率设定为20 MS/s,每次采集16 MS数据,采集的环境分别为LOS环境和NLOS环境,其中LOS环境共采集了10组数据,NLOS环境共采集了3组数据。

2　数据预处理

2.1　Airmax技术简介

Airmax是一种无线技术,由美国UBNT公司开发,用来提升无线传输速率和网络承载能力,目前只能兼容UBNT M2系列产品。

标准Wi-Fi协议基于RTS/CTS机制,共享带宽,当连接设备的传输速率水平不一时,整体效率会被速率较慢的设备拖累。Airmax设备采用TDMA技术,各台设备在不同的时隙工作,互不干扰,且系统可以智能地分配更好的资源给需求较大的设备。

但Airmax设备所用的是私有协议,在进行帧前导提取时无法沿用已有的方法。

图1　USRP实物图

Fig.1　USRP physical map

2.2　Airmax信号帧前导的捕获

图2是手动截取的一段帧前导,其中纵坐标表示经过归一化后的信号幅度,横坐标表示时间,单位为0.05 μs。从图2可以发现帧前导的波形与OFDM信号较为相似,都可看成是不同频率不同幅度的多个子载波相加得到。

通过大量实验观察手动截取的Airmax信号帧前导,发现前导码均是由10个符号组成。在20 MS/s采样率的前提下,每个符号共32点,因此每段前导码的长度确定为320点。

帧前导的捕获共分为两大步骤,分别是粗同步和精确同步,本文将依次介绍这两个步骤。

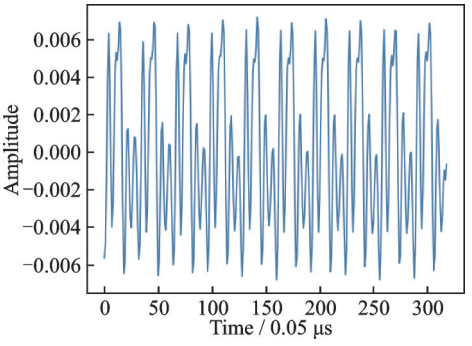

图2　手动截取的帧前导信号

Fig.2　Manually intercepted frame preamble

2.2.1　粗同步

对手动捕捉的大量前导码进行FFT,结果如图3所示。图3上半部分是手动抓取的前导码,下半部分是对该前导码进行了320点的FFT之后得到的频域图,其中横坐标是经过归一化后的频率。可以发现前导码只有在有限个点上有较大的值,其他点的值接近于0,这个特征和Wi-Fi信号类似。图4是802.11n信号前导码的时域和频域图。可以发现两者都是在个别频点有值,在其余频点受噪声影响有微弱能量。

记有明显能量分布的有限个频点的索引集合为I,所有频点的集合为U,傅里叶变换的长度为N,$x(i)$表示第i个点的信号,其中$i\in[1,N]$。记点数为一个FFT长度的信号块为一个搜索单元x_i,其中$x_i=[x(i),x(i+1),\cdots,x(i+N-1)]$。

定义粗同步判别阈值,帧前导的搜索范围为整个基带信号,使用帧前导搜索算法搜索出距当前索引位置最近的,判决项η_i大于阈值η的搜索单元。搜索算法如下。

算法1　步长帧前导搜索算法

输入:USRP采集的原始数据文件,当前索引为i,步长为T。

输出:η_i大于阈值η的第一个搜索单元的起始索引。

(1) 对于x_i,计算$\eta_i=10\lg\dfrac{\sum\limits_{i\in I}x(i)^2}{\sum\limits_{i\in X-I}x(i)^2}$。

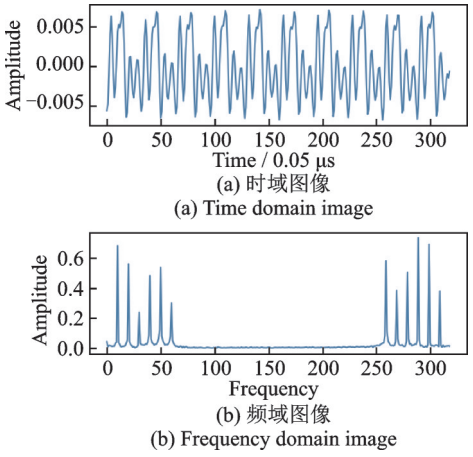

图3 Airmax前导码时域和频域图像

Fig.3 Time and frequency domain images of Airmax preamble

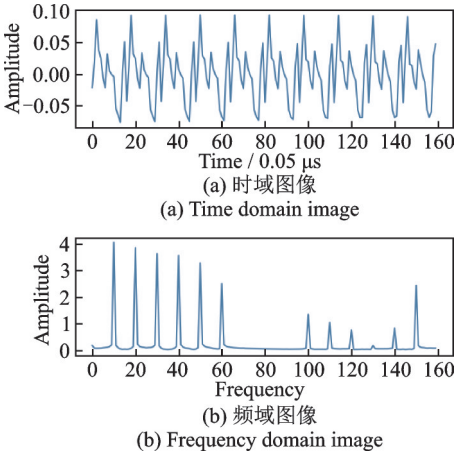

图4 Wi-Fi前导码时域和频域图像

Fig.4 Time and frequency domain images of Wi-Fi preamble

(2) 比较η_i与η,若$\eta_i \geqslant \eta$,返回i且算法结束,否则i加T,回到第1步。

阈值η的选取受信噪比的影响较大,记所有帧前导信号信噪比的最小值为$\mathrm{SNR}_{\min}$,其中

$$\mathrm{SNR}_{\min}=10\lg\frac{\sum\limits_{i\in I}x'(i)^2}{\sum\limits_{i\in X-I}x'(i)^2}$$

为了保证输出的搜索单元内至少有1个符号,需要设置$\eta \geqslant \eta'$,其中

$$\eta'=10\lg\frac{\sum\limits_{i\in I}\frac{1}{10}x'(i)^2}{\sum\limits_{i\in X-I}x'(i)^2}=\mathrm{SNR}_{\min}-10$$

若$\eta<\eta'$,可能会导致输出的搜索单元中不含有任何的帧前导符号;若η取值过大,可能会导致原本符合要求的帧前导未被搜索到。经过大量实验,发现η取4~9都符合要求,本文取η=7。

步长T用来控制搜索效率。T越小,搜索效率越低;T越大,搜索效率越高,但可能会出现帧前导未被搜索到的情况。本文取T=1 000。

2.2.2 精确同步

通过粗同步得到的索引只能确定帧前导在该索引附近,但无法准确确定帧前导的第一个索引位置,这是因为即使混入非帧前导信号,只要搜索单元中包含一个Airmax符号,经过FFT之后也可以显示出如图5所示的频谱特性,只是能量会相对较低。从图5可以明显看出,进行FFT的信号除了包含7个符号外,还包括了前面的一段噪声,因此为了准确确定帧前导的位置,还需要进行精确同步。

记标准帧前导的时域信号为$s(n)$,点数为FFT变换的长度N,则手动截取的帧前导信号$x(n)$可以表示为

$$x(n)=s(n)+\mu(n) \tag{1}$$

式中:$\mu(n)$为截取的帧前导信号中混杂的噪声信号,通过实验验证,该噪声可近似于白噪声。

记通过粗同步搜索到的信号为$x'(n)$,则有

$$x'(n)=s(n-L)+\mu'(n) \tag{2}$$

式中:$\mu'(n)$为搜索单元中混杂的噪声信号,L为帧前导信号在搜索单元中的偏移量。

若$L>$N,则表示$x'(n)$中没有任何符号,没有进行精确同步的意义。但粗同步可以保证输出的搜索单元内至少有1个符号,所以可以保证得到的偏移量$L\in[0,N]$。由下文的推导可知,当$L\in[0,N]$时,通过精确同步一定可以找到帧前导的准确起始位置。

因为帧前导信号后是一些长度不定的符号,不可以将其当作噪声处理,所以相关运算的起点只能在帧前导信号的左方。以图5为例,起点在帧前导信号前一段的噪声处,所以这里的L是正数,即搜索单元中的帧前导信号的起点应该在搜索单元起点的右侧。

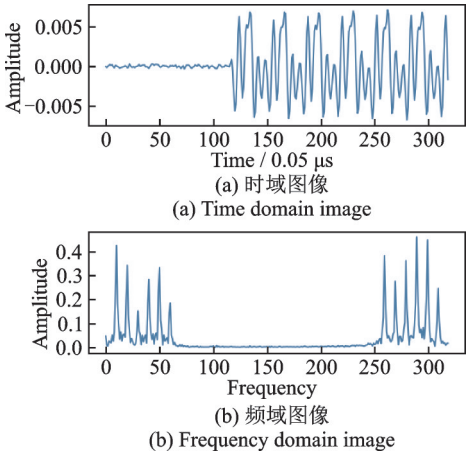

图5　包含噪声的Airmax前导码时域和频域图像

Fig.5　Time and frequency domain images of Airmax preamble with noise

对$x(n)$和$x'(n)$作相关运算,得

$$r_{xx'}(m)=\sum_{n=-\infty}^{+\infty}x^*(n)\,x'(n+m)=r_s(m-L)+r_{s\mu'}(m)+r_{\mu s}(m-L)+r_{\mu\mu'}(m)$$

$\mu(n)$,$\mu'(n)$可以看作白噪声,所以$r_{s\mu'}(m)$,$r_{\mu s}(m-L)$,$r_{\mu\mu'}(m)$均为0,最终得到

$$r_{xx'}(m)=r_s(m-L)\tag{3}$$

根据自相关函数的性质,$r_s(m-L)$的最大值在$m=L$时取得,因此函数$r_{xx'}(m)$的极大值点就是帧前导信号在搜索单元中的偏移量,也就是帧前导信号准确的起始位置。如图6所示,上半部分表示的是经过粗同步之后搜索到的搜索单元,下半部分是相关函数的图像,可以发现相关函数的极大值点与帧前导信号的起始点完全一致。

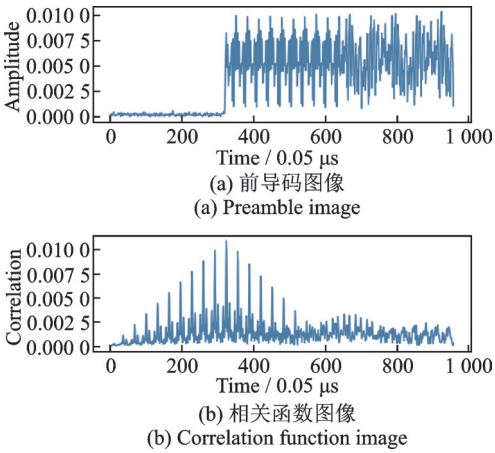

图6　前导码和相关函数图像

Fig.6　Preamble and correlation function image

2.2.3　时间复杂度分析

记n为数据文件内信号的总点数,m为搜索单元长度。粗同步过程对每一个搜索单元作多次基本运算(m次乘法,$m-2$次加法,1次选择)和一次对数运算。搜索整个数据文件的次数为$\frac{n}{m}$,因此粗同步的时间复杂度为$O\left(\frac{n}{m}*2m\right)=O(n)$。

精确同步在每次得到粗同步的结果后,进行m次乘法运算,所以精确同步的时间复杂度为$O(n)$。可以发现粗同步和精确同步的时间复杂度相当,但由于粗同步的基本运算次数是精确同步的两倍,所以粗同步的总时间会大于精确同步。

实验结果表明,提取单个帧前导,粗同步所花费的时间为0.017 4 s,而精确同步所花费的时间为0.007 82 s。粗同步花费的时间大约是精确同步的两倍,与理论相符。

3 Airmax射频指纹提取

通过粗同步和精确同步，可以找到原始数据文件中所有的帧前导，下面的工作就是从帧前导信号中提取射频指纹。稳态响应的射频指纹主要包括频率偏移，幅度误差和相位误差等，本文主要讨论频率偏移和幅度误差。

3.1 频率偏移的提取

设 $x_r(n)=R(n)\mathrm{e}^{-\mathrm{j}\left((\omega_b+\Delta\omega_r)n+\varphi\right)}$，其中 $x_r(n)$ 为接收机接收到的信号，$\Delta\omega_r$ 为接收机和发射机之间的频率偏移，ω_b 为基带信号频率。将一段帧前导信号分成长度相等的两段，每段长度为 N_1，对这两段信号进行共轭相乘，即对每一信号点 $x_r(n)$，计算 $x_r(n)\,x_r^*(n+N_1)$，可得

$$x_r(n)\,x_r^*(n+N_1)=R(n)\,R(n+N_1)\mathrm{e}^{\mathrm{j}\left(\omega_b+\Delta\omega_r\right)N_1} \tag{4}$$

基带信号两个符号对应位置的两个点的相位应该相同，即 $\omega_b N_1=2k\pi$，$k\in\mathbf{Z}$。对共轭相乘得到的复数取辐角，即可计算出 $\Delta\omega_r$

$$\Delta\omega_r=\frac{1}{N_1}\arg\left(x_r(n)\,x_r^*(n+N_1)\right) \tag{5}$$

由于信道噪声等影响，一些能量较弱的点会使得结果的精确度降低，因此需要对这些点进行筛选。定义筛选因子 α，信号的平均能量为 W，对于每段帧前导信号，计算 $R(n)\,R(n+N_1)$。

若 $R(n)\,R(n+N_1)\geqslant(1+\alpha)W$，则保留 $x_r(n)$，否则丢弃。α 可以根据系统需求自定义，但过大会导致保留的点数过少，一般取0.1～0.4之间，本文取 α=0.3。计算筛选过后的每个信号点的频偏，最后取平均值，就可以得到该前导码的估计频偏。

图7为200段前导码的频偏直方图，其中每项数据表示一段前导码的频偏，可以发现，在误差允许的范围内，频偏可以大致分为两类，而网桥设备确实有两台，与实际相符。

图8为频率分布直方图，通过图8可以更好地看出频偏的分布，从图8可以看出，频偏聚集在0.1和－0.5附近。

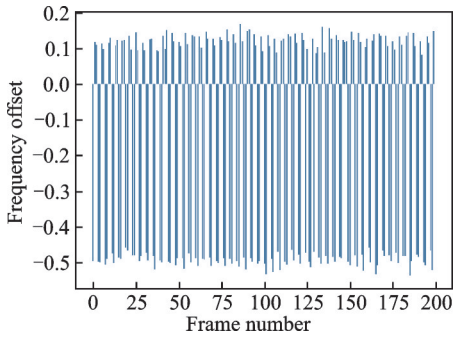

图7 200段前导码频偏直方图

Fig.7 Frequency offset histogram of 200 preambles

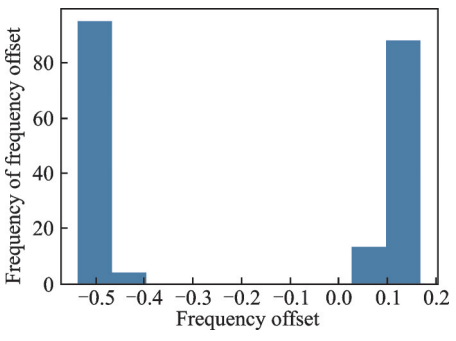

图8 频偏频率分布直方图

Fig.8 Frequency distribution histogram of frequency offset

3.2 频率偏移方差的提取

对于每一段帧前导信号，按照3.1节的方法，可以得到每个信号点频率偏移估计值。除了使用均值作为特征以外，还可以考虑使用这些频率偏移的方差，因为方差反映了频偏的稳定性，从而间接反映了

设备的稳定性。

200段前导码的频偏方差直方图如图9所示。可以发现虽然方差的稳定性不如频偏,但也能大致分为两类。图10是方差的频率分布直方图,虽然区分不如频偏的直方图明显,但从核密度估计曲线可以看出方差的分布围绕着两个中心。

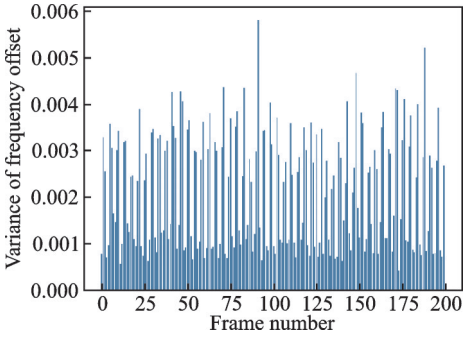

图9　频偏方差直方图

Fig.9　Histogram of frequency offset variance

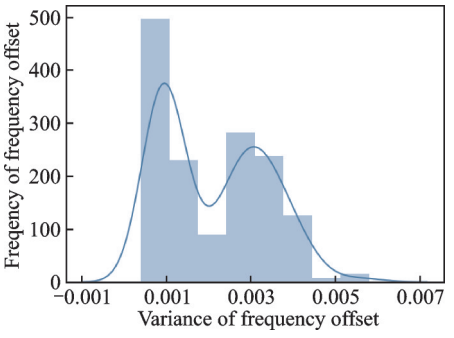

图10　频偏方差频率分布直方图

Fig.10　Frequency distribution histogram of frequency offset variance

3.3　幅度误差的提取

由于多径信道的影响,信号的幅频曲线不是一条直线,而会发生不同程度的衰落,即不同频点的信号能量会随机衰落。虽然不同时间的衰落曲线不一定相同,但本文假定相差极短时间的衰落曲线一致,即可以认为同一段帧前导信号的衰落曲线不随时间变化,在这个前提下,不同符号的相同频点应该有相同的衰落程度。

记Airmax帧前导频谱图中有值的12个频点的索引为$L_1,L_2,\cdots,L_{12}$,$X_b(L_i)$为基带信号在频点L_i的幅度;$X_r^n(L_i)$为接收机在频点L_i,时间n的信号幅度;$X_t^n(L_i)$为发射机在频点L_i,时间n的信号幅度;$X_c^n(L_i)$为信道信号在频点L_i,时间n的信号幅度;$g_r^n(L_i)=1+\eta_{L_i}^n$为接收机在频点$L_i$,时间$n$的增益,表征信道中的信号经过接收机后不同频点信号幅度的变化;$g_t^n(L_i)=1+\xi_{L_i}^n$为发射机在频点L_i,时间n的增益;$g_c^n(L_i)$为发射机在频点L_i,时间n的增益。

由上文的假设,同一帧信号不同符号的相同频点信道衰落相同,即$g_c^n(L_i)=g_c(L_i)$。$X_r^n(L_i)$可以通过式(6)计算得到。

$$\begin{aligned}X_r^n(L_i)&=g_r^n(L_i)g_c^n(L_i)g_t^n(L_i)X_b(L_i)=\\&g_c(L_i)(1+\eta_{L_i}^n)(1+\xi_{L_i}^n)X_b(L_i)\end{aligned}\tag{6}$$

将一段帧前导分为长度相等的两段,则每段都有相同数量的Airmax符号,且对应位置的时延为N_1,所以可以得到

$$X_r^{n+N_1}(L_i)=g_c(L_i)(1+\eta_{L_i}^{n+N_1})(1+\xi_{L_i}^{n+N_1})X_b(L_i)$$

由于存在$g_c(L_i)$这一衰落因子的影响,若单纯使用$X_r^n(L_i)$或者$X_r^{n+N_1}(L_i)$作为指纹特征,必然存在较大的不稳定性。为了消去多径信道的复杂影响,可以将对应频点的信号幅度相除,即

$$\rho(L_i)=\frac{X_r^{n+N_1}(L_i)}{X_r^n(L_i)}=\frac{(1+\eta_{L_i}^{n+N_1})(1+\xi_{L_i}^{n+N_1})}{(1+\eta_{L_i}^n)(1+\xi_{L_i}^n)}\tag{7}$$

可以将$\rho(L_i)$作为射频指纹的特征，因为共有12个频点，所以幅度相关的特征维数为12。

分别采集两台Airmax设备的信号，分别取100组帧前导信号计算每个设备的12个幅度相关的射频指纹特征，在每台设备的100组帧前导中任选一组，绘制成如图11所示的直方图。

图11横轴表示12个频点，纵轴表示两台设备在各个频点的幅度指纹特征$\rho(L_i)$，从图中可以看出两台设备的$\rho(L_1)$，$\rho(L_9)$，$\rho(L_{10})$等特征有着明显的区别，其他频点的特征区别不大。考虑到单个帧前导的结果存在偶然性，所以将每台设备的帧前导得到的指纹取平均值，得到的直方图如图12所示。

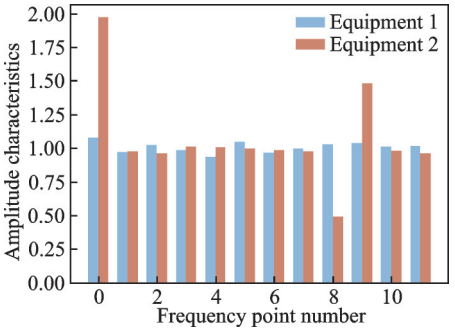

图11 幅度指纹直方图

Fig.11 Histogram of amplitude fingerprint

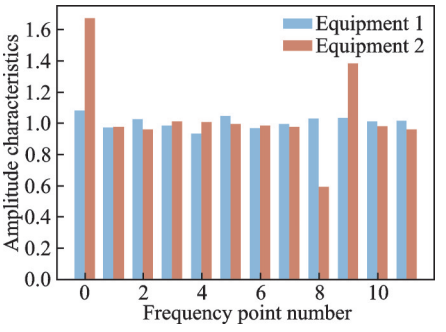

图12 平均幅度指纹直方图

Fig.12 Histogram of average amplitude fingerprint

从图12中可以发现两台设备的幅度指纹特征确实存在差异，在个别频点上存在明显差异，为了进一步说明这种差异，可以绘制频率分布直方图。

图13是设备1在频点L_1的特征$\rho(L_1)$的频率分布直方图，图14是设备2在频点L_1的特征$\rho(L_1)$的频率分布直方图。可以看出设备1的特征$\rho(L_1)$与设备2相比，集中分布在0.97附近，而设备2的特征集中分布在1.10附近。通过图15可以更好地看出这一点。

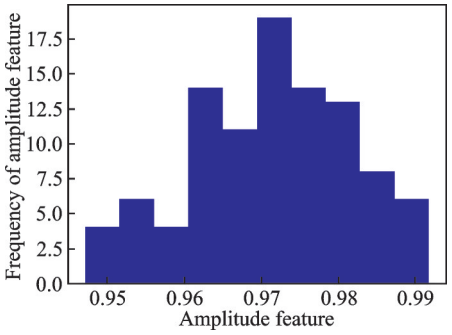

图13 设备1幅度特征直方图

Fig.13 Histogram of equipment 1 amplitude fingerprint

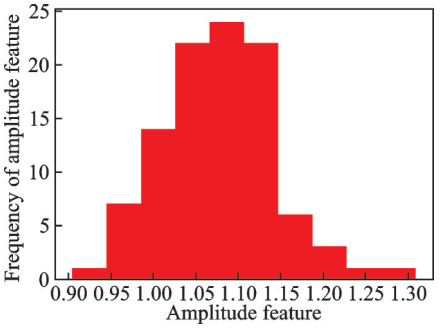

图14 设备2幅度特征直方图

Fig.4 Histogram of equipment 2 amplitude fingerprint

从图15可以看出$\rho(L_1)$的类间差距比类内差距更为明显，因此$\rho(L_1)$可以用作区分设备的特征之一。

其余11个频点的实验过程不再重复，实验结论是，尽管区分度小于频偏，幅度相关的特征也能作为

辅助特征使用。

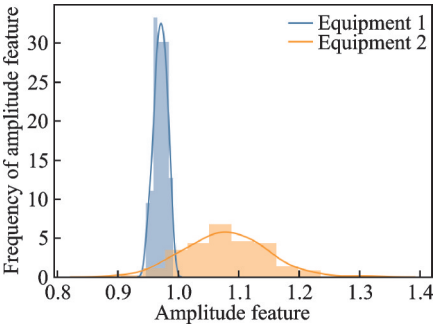

图15　两台设备的幅度特征频数分布直方图

Fig.15　Frequency distribution histogram of amplitude fingerprint

4　分类器的设计

本文使用K-means算法和决策树模型分别进行设备的分类训练,将分别阐述这两种方法的具体算法,实验过程与实验结果。

4.1　基于K-means算法的Airmax设备分类

受实验条件限制,可供实验的Airmax设备数量较少,且测试环境也是由两台Airmax组成的局域网,所以考虑使用简单的K-means算法进行设备分类。

4.1.1　K-means算法

算法2　K-means算法

输入:训练集D,聚类中心个数M。

输出:训练集D的分类结果。

(1) 对射频指纹特征进行标准化,将每类特征的数值缩放到-1～1之间,任选M个聚类中心。

(2) 将数据集D的每个样本按最小距离准则划分到M类中的某一类。

(3) 计算各类的均值向量作为重新分类后的聚类中心。

(4) 若聚类中心保持不变,算法结束,输出各个样本的类标,否则转第2步。

4.1.2　Airmax设备分类

由于每次实验都是两台Airmax设备同时工作,所以类标信息无法直接得到。本文采用了断电的方法,即在局域网工作状态中使一台设备断电,然后采集工作频率的信号,这样采集到的信号可以归为同一个设备,采集完成后再通电并使另一台设备断电,就可以采集到第二台设备的信号。使用算法3进行实验,得到的准确率为100%。以频偏为横轴,频偏方差为纵轴画出聚类图如图16所示。图中的两类点是聚类后的结果,可以看出在实验室环境下,K-means算法对这两台设备的分类较为准确。

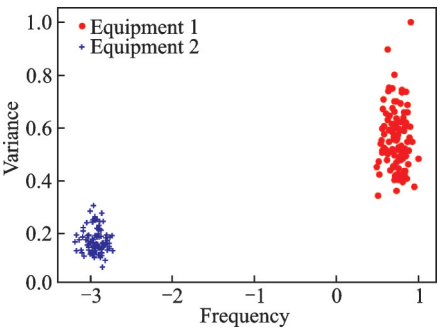

图16　设备特征聚类图

Fig.16　Device feature clustering image

K-means算法的优点是模型简单,运行效率高,但K-means算法属于无监督学习,样本的先验信息没有用到,且聚类中心个数的选取需要进行大量实验。

4.2　基于决策树模型的Airmax设备分类

决策树模型对于非线性分类往往有较好的效果,因此本文尝试用CART分类树算法进行设备识别与分类。优化目标选择基尼系数,基尼系数代表了模型的不纯度,基尼系数越小,则不纯度越低,特征越好[11-12]。假设有k个类别,第i个类别的概率为p_i,则基尼系数的表达式为

$$\mathrm{Gini}(p)=1-\sum_{i=1}^{k}p_i^2 \tag{8}$$

4.2.1 决策树生成算法

算法3 决策树生成算法

输入:指纹特征训练集,样本数量阈值,基尼系数阈值。

输出:决策树T。

(1) 判断样本个数,若小于阈值,返回决策子树。

(2) 计算并判断D指纹数据集的基尼系数,若小于阈值,返回决策子树。

(3) 计算指纹数据集关于每个特征的基尼系数。

(4) 对步骤3中的基尼系数排序,取出系数最小的特征,以该特征值作为切分依据,将指纹数据集分为两部分。

(5) 对两个子节点重复步骤(1)~(4),生成决策树。

4.2.2 Airmax设备分类

使用上述决策树模型进行分类。首先进行数据预处理,对样本进行数据标准化;再使用简单交叉验证法,选定一部分数据作为训练集,其余作为测试集;再通过训练集训练模型参数;最后在测试集上计算误差。

本文设定70%的数据作为训练集,30%的数据作为测试集。将样本的指纹特征及其类标信息存储到CSV文件中,作为指纹库,截取一部分如表1所示。使用决策树模型对指纹库进行分类,得到的准确率为100%,说明决策树模型在该二分类问题上的性能较好。但目前为止的实验都是针对二分类问题,对于多分类问题的准确性还有待确定。

为了验证该方法对于多分类问题的准确性,另取两台Airmax网桥设备入库并进行分类训练。实验结果表明使用决策树模型对4台设备的指纹库进行分类,得到的准确率为100%。

表1 指纹库部分样本的指纹特征

Table 1 Fingerprint characteristics of some samples in fingerprint database

Num	$\Delta\omega_r$	Variance	$\rho(L_1)$	…	$\rho(L_{12})$	Label
0	0.117 856	0.003 278	1.181 705	…	1.032 590	1
1	0.110 084	0.002 550	1.080 344	…	0.988 803	1
2	0.113 130	0.003 570	1.137 130	…	1.040 657	1
3	0.097 868	0.003 046	1.072 138	…	1.010 018	1
4	0.114 984	0.003 001	1.044 512	…	0.961 991	1
5	−0.496 020	0.000 775	0.973 811	…	0.794 775	2
6	−0.498 540	0.000 698	0.976 734	…	0.928 579	2
7	−0.500 250	0.000 957	0.967 109	…	0.960 016	2

4.3 基于SVM模型的Airmax设备分类

支持向量机(Support vector machine,SVM)也是常用的分类模型,并且核函数的使用使得SVM模型可以更好地应对非线性可分的数据集。本文选择高斯径向基核函数构造分类器。

4.3.1 SVM学习算法

算法4 基于核函数的SVM学习算法

输入:训练集$D=\{(\boldsymbol{x}_1,\boldsymbol{y}_1),\cdots,(\boldsymbol{x}_N,\boldsymbol{y}_N)\}$,其中$\boldsymbol{x}_i$为第$i$个样本的特征向量,$\boldsymbol{y}_i$为第$i$个样本的类标。

输出:分类决策函数。

(1)　选用高斯径向基函数K,求解最优化问题

$$\min \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j K(x_i,x_j)-\sum_{i=1}^{N}\alpha_i$$

$$\text{s.t.}\ \sum_{i=1}^{N}\alpha_i y_j=0$$

$$0\leqslant\alpha_i\leqslant 1,i=1,2,\cdots,N$$

(2) 选择步骤1中最优解的一个正分量α_j^*,计算

$$b^*=y_j-\alpha_j^* y_j K(x_i\cdot x_j)$$

(3) 构造决策函数

$$f(x)=\text{sign}\left(\sum_{i=1}^{N}\alpha_i{}^* y_j K(x_i\cdot x_j)+b^*\right)$$

4.3.2　Airmax设备分类

使用算法4生成决策函数,其中数据预处理的步骤与4.2.2节相同。由于类别总数为4,训练时采用间接法,依次把某个类别的样本归为一类,其他剩余的样本归为另一类,对这样4个数据集分别训练,测试时分别利用这4个训练模型测试,选取准确率最高的作为最后的结果。使用SVM模型对4台设备进行分类,得到的准确率为88.3%。

4.4　多种模型的分类结果比较

与K-means算法,决策树模型,SVM模型的流程类似,本文又选取了随机森林模型,逻辑斯蒂回归模型,神经网络模型等进行了Airmax设备分类。其中随机森林模型的子树数量为100,优化目标为基尼系数;逻辑斯蒂回归模型的学习方法为拟牛顿法,优化目标为似然函数;神经网络模型的学习方法为梯度下降法,损失函数定义为模型输出和观测结果之间的欧式距离,隐藏层数为1,隐藏层节点数为7,输入层节点数为14,输出层节点数为2。各模型的参数都经过实验调整至最优(准确率最高)。

本文对多种分类模型对于4台Airmax设备的分类结果(运行时间和测试集准确率)进行了比较,比较结果见表2。从表2可以发现,对于本文的4分类问题,K-means算法的运行时间最少,准确率相对于SVM算法和逻辑斯蒂回归模型较高,但相对于决策树,随机森林等模型较低。

表2　各模型的运行时间和准确率

Table 2　Run time and accuracy of each model

模型	运行时间/s	准确率/%
K-means	0.001 5	92.4
决策树	0.002 1	100.0
SVM	0.053 0	88.3
随机森林	0.013 0	100.0
逻辑斯蒂回归	0.023 0	82.5
神经网络	0.960 0	100.0

K-means算法和决策树模型由于模型简单,收敛速度较快,所以运行效率较高,同时由于该问题的特征维度和类别总数较少,所以分类准确率并不低。而神经网络等模型较为复杂,收敛速度相对较慢,因此运行效率低于K-means算法和决策树模型。

从准确率来看,对于线性不可分问题,逻辑斯谛回归模型的准确率要小于决策树和神经网络,神经

网络因为有了隐藏层,可以较好地处理线性不可分问题。

本文介绍了Airmax设备射频指纹的特征提取方法,包括粗同步,精确同步以及14维特征的提取,并使用K-means聚类模型,决策树模型分别对两台,4台设备进行分类和预测。系统的流程图如图17所示。首先通过粗同步和精确同步从基带信号中提取前导码;再从前导码中提取14维指纹特征作为指纹数据集;最后将数据集划分为训练集和测试集,并使用机器学习模型进行分类预测。

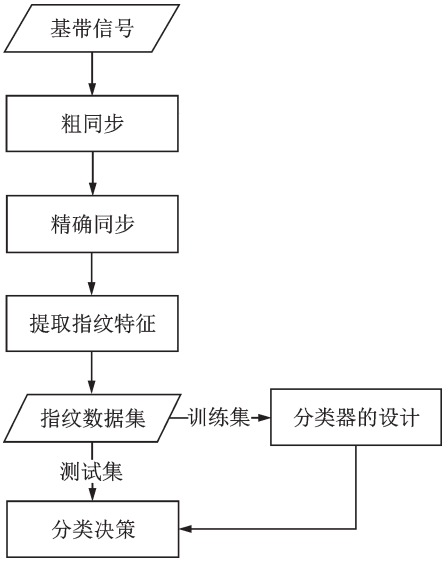

图17 系统流程图

Fig.17 Flow chart of system

5 结束语

K-means聚类模型较为简单,并且对两台设备的分类准确率达到100%,但是4台设备的分类准确率小于决策树模型,决策树模型对两台,4台设备的分类准确率均达到了100%。

本文提取射频指纹的过程中没有直接用到信号协议,在定位前导码的过程中采用了手动抓取一段样本前导码结合相关运算,傅里叶变换等信号处理方法,成功实现了不借助信号协议的前导码定位,而之前的射频指纹提取研究都借助了信号协议,例如Wi-Fi信号射频指纹的提取。在不知道信号协议的情况下,本文为射频指纹的提取提供了新的思路和方法。在分类识别过程中,由于实验条件的限制,目前还未能对更多的设备进行实验。未来在实验条件允许的情况下,可以增加设备的数量测试分类准确率,同时可以通过提高采样率来进一步提取瞬态响应,DCTF等特征。

参考文献:

[1] WANG Chiapin, TAI Tientsung. Achieving time-based fairness for VoIP applications in IEEE 802.11 WLAN using a cross-layer approach[C]//Proceedings of 2010 IEEE 21st International Symposium on, Personal Indoor and Mobile Radio Communications (PIMRC). [S.l.]: IEEE, 2010.

[2] DANEV B, ZANETTI D, CAPKUN S. On physical-layer identification of wireless devices[J]. ACM Computing Surveys, 2012, 45(1): 1-29.

[3] BRIK V, BANERJEE S, GRUTESER M, et al. Wireless device identification with radiometric signatures[C]//Proceedings of the 14th Annual International Conference on Mobile Computing and Networking, MOBICOM 2008. San Francisco, California, USA: ACM, 2008.

[4] YUAN H L, HU A Q. Fountainhead and uniqueness of RF fingerprint[J]. Journal of Southeast University (Natural Science Edition), 2009,39(2): 230-233.

[5] REISING D R, TEMPLE M A, JACKSON J A. Authorized and rogue device discrimination using dimensionally reduced RF-DNA fingerprints[J]. IEEE Transactions on Information Forensics and Security, 2015, 10(6): 1180-1192.

[6] FRÉDÉRIC D, ST-HILAIRE M. Radiometric identification of LTE transmitters[C]//Proceedings of 2013 IEEE Global Communications Conference (GLOBECOM). [S.l.]: IEEE, 2014.

[7] COGHILL C, REHMAN S U, SOWERBY K W. Radio-frequency fingerprinting for mitigating primary user emulation attack

in low-end cognitive radios[J]. IET Communications, 2014, 8(8): 1274-1284.

[8] WU Q, FERES C, KUZMENKO D, et al. Deep learning based RF fingerprinting for device identification and wireless security [J]. Electronics Letters, 2018, 54(24): 1405-1407.

[9] PENG L, HU A, JIANG Y, et al. A differential constellation trace figure based device identification method for ZigBee nodes [C]//Proceedings of International Conference on Wireless Communications & Signal Processing. [S.l.]: IEEE, 2016.

[10] YUAN H L. Research on physical-layer authentication of wireless network based on RF fingerprinting[D]. Nanjing: Southeast University, 2011.

[11] FAYYAD U M, IRANI K B. On the handling of continuous-valued attributes in decision tree generation[J]. Machine Learning, 1992, 8(1): 87-102.

[12] LINDEBERG T. Scale-space theory: A basic tool for analyzing structures at different scales[J]. Journal of Applied Statistics, 1994, 21(1/2): 225-270.

作者简介:

季澈(1996-),男,硕士研究生,研究方向:射频指纹识别,E-mail: 220184388@seu.edu.cn。

彭林宁(1984-),男,博士,副教授,研究方向:物理层安全、射频指纹、无线信道字符串生成、室内定位、软件无线电,E-mail:pengln@seu.edu.cn。

胡爱群(1964-),男,博士,教授,研究方向:无线通信安全、物理层安全,E-mail: aqhu@seu.edu.cn。

王栋(1986-),男,博士,研究方向:无线信道与射频指纹的分离技术、MIMO-OFDM射频指纹估计与识别技术、MIMO-OFDM射频指纹估计与识别技术、无线通信物理层密钥的生成技术,E-mail: wd@seu.edu,cn。

(编辑:夏道家)