基于集成学习的全双工中继系统安全中继选择方案研究

张 梦,郑建宏,刘香燕,何 云
(重庆邮电大学通信与信息工程学院,重庆 400065)

摘 要: 当处于无线信道状态信息(Channel State Information, CSI)快速变化或者多跳中继等应用场景时,利用集成学习算法解决安全中继选择问题能减少实时处理时延及计算复杂度.将合法信道和窃听信道的CSI作为训练模型输入,使系统安全容量到达最大的中继节点索引作为输出,把全双工中继系统安全中继选择问题转化为一个多类分类问题,并利用随机森林(Random Forest, RF)算法求解.安全中继选择方案的实现分为数据准备、模型建立和结果预测三个阶段.在数据准备阶段,由于RF算法要求训练模型输入为离散值,给出了均匀量化和非均匀量化两种特征提取法将CSI转化为离散值.最后,通过仿真实验验证方案性能.

关键词: 全双工; 中继选择; 物理层安全; 机器学习; 集成学习; 随机森林
中图分类号: TN918.91 **文献标识码:** A **文章编号:** 0372-2112(2021)09-1852-05
电子学报URL: http://www.ejournal.org.cn **DOI:** 10.12263/DZXB.20201340

Ensemble Learning-Based Relay Selection Scheme in Full-Duplex Relay System for Secure Transmission

ZHANG Meng, ZHENG Jian-hong, LIU Xiang-yan, HE Yun
(*School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing* 400065, *China*)

Abstract: Using the ensemble learning algorithm to solve the secure relay selection problem can reduce the real-time processing delay and computational complexity, when the wireless systems that with fast-changing channel state information (CSI) or multi-hop relay situation. The secure relay selection problem of a full-duplex relay system is modeled as a multi-class classification problem, which is solved by the random forest (RF) algorithm. The CSI of the legitimate channel and eavesdropping channel is taken as the input of the training model, and the index of the relay node which can maximize the system security capacity is taken as the output of the training model. The implementation of the proposed scheme is divided into three phases: training data preparation, model building and result prediction. In the training data preparation phase, as the input features demanded by the RF algorithm are discrete values, feature extraction methods based on both uniform and non-uniform quantization algorithms are proposed to transform the CSI into finite discrete values. Finally, simulation experiments are conducted to verify the performance of the proposed scheme.

Key words: full-duplex; relay selection; physical layer security; machine learning; ensemble learning; random forest

1 引言

由于无线信号的开放性,实现数据的安全传输是各类无线通信系统应用的重要前提.物理层安全技术通过有效地利用无线信道随机特性以及信号设计、处理技术替代传统密码体制中的共享密钥,实现数据安全传输[1].

文献[2]提出了基于不同优化目标的多个中继选择策略.主要有以降低系统中断概率为目标的连续中继选择策略;以优化系统安全容量为目标的最大最小中继选择策略;以增强干扰为目标的干扰中继选择策略.文献[3]研究了中继选择与功率分配联合优化提高系统安全性能的算法.文献[4]研究了通过安全天线选择来提高系统遍历安全容量的方法.目前大部分研究是基于点对点最优算法和精确的信道状态信息(Channel State Information, CSI),当CSI不易获得或者发生快速变化时,传统算法时效性较差、复杂度较高.

机器学习(Machine Learning, ML)算法可减少实时处理时延及计算复杂度,提高系统性能.文献[5]利用两个个体学习算法,k近邻算法和支持向量机(Support

收稿日期:2020-11-26;修回日期:2021-05-17;责任编辑:覃怀银
基金项目:国家科技重大专项(No. 2018ZX03001026-002); 重庆市教委科技项目(No. KJQN201800618, No. KJQN202002102)

Vector Machine, SVM)算法来解决半双工系统中安全天线选择问题. 文献[6]研究了在半双工中继系统中基于决策树(Decision Tree, DT)的安全中继选择问题.

目前,利用ML来解决系统安全性能优化问题的研究都是基于个体分类算法,而使用个体分类算法产生训练模型,通常会产生过拟合的情况,模型泛用性较差. 集成学习算法可将多个个体分类算法组合在一起,输出一个综合评价,有效减少训练模型的过度拟合. 随机森林(Random Forest, RF)是一种高效的集成学习算法[7],在生成训练模型时,利用装袋法对样本数据进行有放回的重采样让输入的训练样本更具多样性;利用随机子空间方法随机产生特征子集让训练样本更具随机性,以此生成多个不同的个体学习模型,比如多个不同的决策树模型,最后过组合和投票来确定最终的预测结果. 克服了个体分类算法容易产生过度拟合的问题,能提高预测准确率.

另外,上述研究构建的系统安全模型都是基于中继节点工作在半双工模式,全双工技术能在同一频段上同时接收和发送信号,其频谱效率是半双工传输的两倍,是下一代无线通信系统的关键技术[8~10]. 因此,研究给出一种基于RF算法的全双工中继系统安全中继选择方案. 安全中继选择问题就是找到使系统安全性能达到最优的中继节点,结合RF算法,将合法信道和窃听信道的CSI作为训练模型的输入,使系统安全容量到达最大的中继节点索引作为RF训练模型的输出标签,每一个分类标签代表了候选中继节点的索引,是一个多级标签. 因此,将全双工多中继系统安全中继选择问题转化为了基于RF算法的多类分类问题. 其具体实施方案分为三个阶段,即训练数据准备、模型建立和选择结果预测阶段. 在训练数据准备阶段,由于RF算法要求待分类对象的输入特征是离散值,研究提出了基于均匀量化和非均匀量化算法的特征提取方法,将合法信道和窃听信道CSI转化为有限的离散值. 并且由于采用了特征提取方法对CSI进行了离散化,相对于传统的点到点优化算法,节约了系统的反馈开销并且提高了系统对CSI的容错性.

2　系统模型

假设系统模型由一个源节点S、目的节点D、窃听节点E和K个解码转发全双工中继节点组成. 源节点S与中继节点$R_i(1 \leqslant i \leqslant K)$、源节点$S$与窃听节点$E$、中继节点$R_i$与目标节点$D$、中继节点$R_i$与窃听节点$E$之间存在直接链路,其信道系数分别表示为$h_{SR_i}$、$h_{SE}$、$h_{R_iD}$和$h_{R_iE}$. 包含所有CSI的集合为:$\Omega = \left\{h_{SR_1}, \cdots, h_{SR_K}, h_{R_1D}\cdots, h_{R_KD}, h_{R_1E}, \cdots, h_{R_KE}, h_{SE}\right\}$. 在节点$E$、$R_i$、$D$处的噪声定义为$n_E(t)$、$n_{R_i}(t)$、$n_D(t)$,且$n_E(t) \sim G(0,\delta_e^2)$、$n_{R_i}(t) \sim G(0,\delta_r^2)$、$n_D(t) \sim G(0,\delta_d^2)$,$G(0,\delta^2)$表示为均值为0,方差为$\delta^2$的高斯分布.

在时隙t,数据$x_s(t)$先由S传输到R_i,再由R_i解码转发到D,h_{C_i}为中继节点残余自干扰. 中继节点R_i在时隙t接收到的信息数据为:

$$y_{R_i}(t) = \sqrt{P_S}\, h_{SR_i} x_s(t) + \sqrt{P_{R_i}}\, h_{C_i} x_s(t-1) + n_{R_i}(t) \quad (1)$$

其中,$\sqrt{P_S}$和$\sqrt{P_{R_i}}$分别为源节点和中继节点的发射功率. 中继节点的可达容量C_{R_i}为:

$$C_{R_i} = \log_2\left(1 + \frac{P_S\left\|h_{SR_i}\right\|^2}{P_{R_i}\left\|h_{C_i}\right\|^2 + \delta_R^2}\right) \quad (2)$$

目的节点D在时隙t接收到的信息数据表示为:

$$y_D(t) = \sqrt{P_{R_i}}\, h_{R_iD} x_s(t-1) + n_D(t) + n_{R_i}(t) \quad (3)$$

其可达容量为:

$$C_D = \log_2\left(1 + \frac{P_{R_i}\left\|h_{R_iD}\right\|^2}{\delta_R^2 + \delta_D^2}\right) \quad (4)$$

窃听节点E在时隙$t-1$和t均能收到传输的信息,则:

$$y_e(t) = \sqrt{P_S}\, h_{SE} x_s(t) + \sqrt{P_{R_i}}\, h_{R_iE} x_s(t-1) + n_E(t) \quad (5)$$

假设一个传输块包含W个数据包,窃听节点将所接收的W个数据包堆叠在一起进行处理,式(5)可转化为矩阵形式:

$$\boldsymbol{Y}_E = \boldsymbol{H}\boldsymbol{X}_S + \boldsymbol{N}_E \quad (6)$$

其中,

$$\boldsymbol{Y}_E = \left[y_E(W+1), y_E(W), \cdots, y_E(1)\right]^{\mathrm{T}} \quad (7)$$

$$\boldsymbol{X}_S = \left[x_S(W), \cdots, x_S(1)\right]^{\mathrm{T}} \quad (8)$$

$$\boldsymbol{N}_E = \left[n_E(W+1), n_E(W), \cdots, n_E(1)\right]^{\mathrm{T}} \quad (9)$$

$$\boldsymbol{H} = \begin{bmatrix} \sqrt{P_{R_i}}\, h_{R_iE} & & \\ \sqrt{P_S}\, h_{SE} & \ddots & \\ & \ddots & \sqrt{P_{R_i}}\, h_{R_iE} \\ & & \sqrt{P_S}\, h_{SE} \end{bmatrix}_{(W+1)\times W} \quad (10)$$

则窃听节点E的窃听容量表示为:

$$C_E = \frac{1}{2}\log_2\left(\det\left(\boldsymbol{I} + \frac{\boldsymbol{H}^{\mathrm{H}}\boldsymbol{H}}{\delta_E^2}\right)^{\frac{1}{W}}\right) \quad (11)$$

其中,$\boldsymbol{I}$为单位矩阵. 对式(11)中矩阵$\boldsymbol{H}^{\mathrm{H}}\boldsymbol{H}$进行特征分解得到:

$$C_E = \frac{1}{2W\delta_E^2}\log_2\prod_{w=1}^{W}\left(1+\theta_w\right) \tag{12}$$

其中，θ_w为$\boldsymbol{H}^{\mathrm{H}}\boldsymbol{H}$的第$w$个特征值，可表示为[11]：

$$\theta_w = P_S\left\|h_{SE}\right\|^2 + P_{R_i}\left\|h_{R_iE}\right\|^2 + 2\sqrt{P_SP_{R_i}}\left\|h_{R_iE}h_{SE}\right\|\cos\frac{w\pi}{W+1} \tag{13}$$

将式(13)代入式(12)，窃听容量C_E可表示为：

$$C_E = \frac{1}{2}\log_2\left(1+\frac{P_S\left\|h_{SE}\right\|^2 + P_{R_i}\left\|h_{R_iE}\right\|^2}{\delta_E^2}\right) + \frac{1}{2W\delta_E^2}\sum_{w=1}^{W}\log_2\left(1+\frac{2\sqrt{P_SP_{R_i}}\left\|h_{R_iE}h_{SE}\right\|\cos\frac{w\pi}{W+1}}{1+P_S\left\|h_{SE}\right\|^2 + P_{R_i}\left\|h_{R_iE}\right\|^2}\right) \tag{14}$$

假设数据包个数W足够大，则C_E为：

$$C_E = \frac{1}{2}\log_2\left(1+\frac{P_S\left\|h_{SE}\right\|^2 + P_{R_i}\left\|h_{R_iE}\right\|^2}{\delta_E^2}\right) \tag{15}$$

系统的安全容量C_{S_i}定义为合法接收者的信道可达容量与窃听者的窃听容量之差. 源节点S和目的地节点D之间的链路包括两个部分，即源节点S到中继节点R_i的链路及中继节点R_i到目的节点D的链路，合法信道的最大可达容量为两者之间的较小值. 因此，C_{S_i}可表示为：

$$\begin{aligned} C_{S_i} &= \min\left(C_{R_i}, C_D\right) - C_E \\ &= \min\left(\log_2\left(1+\frac{P_S\left\|h_{SR_i}\right\|^2}{P_{R_i}\left\|h_{C_i}\right\|^2+\delta_R^2}\right), \log_2\left(1+\frac{P_{R_i}\left\|h_{R_iD}\right\|^2}{\delta_R^2+\delta_D^2}\right)\right) \\ &\quad - \frac{1}{2}\log_2\left(1+\frac{P_S\left\|h_{SE}\right\|^2 + P_{R_i}\left\|h_{R_iE}\right\|^2}{\delta_E^2}\right) \end{aligned} \tag{16}$$

最优安全中继选择是从多个中继节点中寻找出使系统安全容量最大的中继节点索引k^*. 该索引可表示为：

$$\begin{aligned} k^* &= \arg\max_{1\leqslant i\leqslant K} C_{S_i} \\ &= \arg\max_{1\leqslant i\leqslant K}\left\{\min\left(C_{R_i}, C_D\right) - C_E\right\} \end{aligned} \tag{17}$$

3 基于随机森林的安全中继选择方案

基于RF算法的安全中继选择方案分为三个阶段实现，即训练数据准备、模型建立和结果预测. 前两个阶段为模型的初始化，可在实际数据传输前离线进行，减少了实时决策时延及计算复杂度.

3.1 训练数据准备

RF训练样本集定义为：$Z = \left\{\left(H_1, Y_1\right), \left(H_2, Y_2\right), \cdots, \left(H_x, Y_x\right)\right\}$. $\left(H_x, Y_x\right)$表示训练样本x，其输入变量为H_x，输出变量为Y_x. 其中，H_x为包含合法信道和窃听信道CSI的样本数据；Y_x为分类标签，赋值为使系统安全容量最大的中继节点索引k^*. 由于RF算法要求输入特征是离散值，给出了基于均匀量化和非均匀量化的两种特征提取方法.

3.1.1 均匀量化法

均匀量化是指在离散化的过程中，各个量化区间间隔是相等的. 假设整个CSI分布区间为$[0, T]$，把它均匀地分成N个等长的间隔，即$S_1, \cdots, S_j, \cdots, S_N$. 其中，$S_1$的下边界为0，$S_{N-1}$的上边界为$T$. S_j的下边界和上边界分别定义为S_j^l和S_j^h. 其中，

$$S_j^l = \begin{cases} \dfrac{(j-1)T}{N}, & 1\leqslant j\leqslant N-1 \\ T, & j = N \end{cases} \tag{18}$$

$$S_j^h = \begin{cases} \dfrac{jT}{N}, & 1\leqslant j\leqslant N-1 \\ \infty, & j = N \end{cases} \tag{19}$$

假设h_x为初始CSI值，h_{uq}为量化后的特征值，定义为1到N之间的整数，则h_{uq}与h_x之间的映射关系可以表示为：

$$h_{uq} = \begin{cases} j, & S_j^l\leqslant h_x\leqslant S_j^h \\ N, & h_x > S_j^lN \end{cases} \tag{20}$$

3.1.2 非均匀量化法

非均匀量化方法的量化间隔是非等量的，根据量化区间数变化. 假设整个CSI分布区间为$[0, T]$，分成N个间隔，即$S_1, \cdots, S_j, \cdots, S_N$. S_1的下边界为0，S_{N-1}的上边界为T. S_j的下边界和上边界分别定义为S_j^{ul}和S_j^{uh}. 其中，

$$S_j^{ul} = \log_2(j), \quad 1\leqslant j\leqslant N \tag{21}$$

$$S_j^{uh} = \begin{cases} \log_2(j+1), & 1\leqslant j\leqslant N-1 \\ \infty, & j = N \end{cases} \tag{22}$$

h_{nq}为采用非均匀特征提取法量化后的特征值，定义为1到N之间的整数，则h_{nq}与h_x之间的映射关系可以表示为：

$$h_{nq} = \begin{cases} j, & S_j^{ul}\leqslant h_x\leqslant S_j^{uh} \\ N, & h_x > S_N^{ul} \end{cases} \tag{23}$$

3.2 模型建立

通过以下步骤训练生成RF模型：

(1) 利用装袋法对样本数据集Z进行有放回的重采样；

(2) 利用随机子空间方法从H_x中随机选取部分

特征;

(3) 根据满足最佳分割准则的特征将当前数据集分割成子集,分割后将此特征从特征子集中删除. 再根据剩余特征子集中的最佳特征递归分割分支子集,直到满足停止条件为止,叶子节点则代表了候选中继节点的索引;

(4) 重复步骤(1)~(3)M次,产生M棵决策树;

(5) 集成M棵决策树形成RF模型,并通过投票,得票数最多的分类结果为最终输出.

3.3 结果预测

对于每一次预测,首先中央控制器(Central Controller, CC)采集合法信道和窃听信道的瞬时CSI,然后进行量化处理,将离散CSI反馈到已训练好的RF模型中,最后判断得出最优中继节点索引. CC再将此结果广播给系统中的所有节点,被选中的中继节点准备工作. 由于预测时仅需要量化后的CSI,降低了系统对精确CSI的依赖性.

4 仿真结果及性能分析

本节对给出的方案性能进行仿真分析和验证. 首先,利用MATLAB的机器学习工具箱建立了RF模型,随机产生服从瑞利分布的10000个CSI数据,其中的70%形成训练集,30%形成测试集. 系统仿真的参数可以设定为 $P_S = P_{R_1} = \cdots = P_{R_K} = 30\text{dBm}$, $\delta_E^2 = \delta_D^2 = \delta_R^2 = 20\text{dBm}$, $h_{C_1} = \cdots = h_{C_K} = 0.01$. 将分类错误概率(Classification Error Probability, CEP)和测试集输出的平均安全容量作为衡量该案性能的评价指标. CEP为预测错误样本数与样本总数的比率.

图1对比了采用基于均匀量化和非均匀量化特征提取法时,分类错误概率(CEP)随中继节点个数K变化的情况. 可以看出,两种量化方法得到的CEP,都随着中继节点数量K的增加而增大,即随着中继节点数量的增加,训练模型的预测准确率都出现下降的情况. 总体看来,采用基于非均匀量化的特征提取法得到了相对更好的分类预测效果. 但采用均匀量化法,其预测性能衰减程度相对较小,说明若当网络拓扑结构突然发生变化时,即出现中继节点增加或者减少时,基于均匀量化特征提取法的RF安全中继选择方案鲁棒性更好.

图2分析了采用均匀量化方法时的量化间隔对方案性能的影响,量化间隔μ定义为$\mu = T/N$. 首先,CEP随着μ的增加而减小,这是由于刚开始量化间隔μ取值较太小,不能覆盖足够多的CSI样本,导致CEP较高;随着μ的逐渐增大,与CSI样本的数据特征匹配,CEP降到最低,最后随着μ的持续增大,过多CSI样本被划分到同一个区间,导致量化的精确度下降,CEP增加.

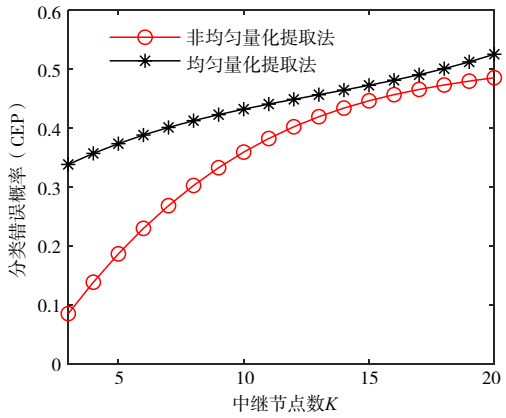

图1　采用两种特征提取方法得到分类错误概率对比情况

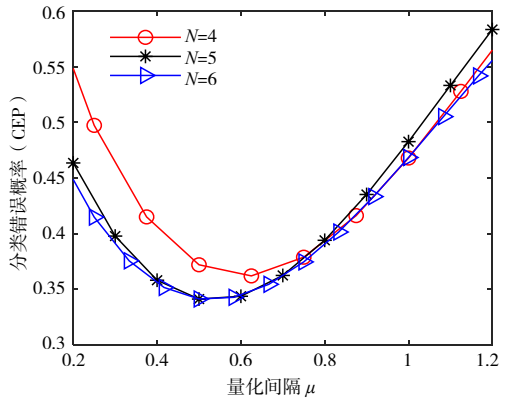

图2　分类错误概率随量化间隔 的变化情况

图3为采用非均匀量化方法时,量化区间个数N对方案性能的影响. N的增加意味着CSI离散化的度量过程更为精细,训练模型易出现过拟合的问题,因此,CEP随着N的增加而增加,但对整体预测性能影响较小.

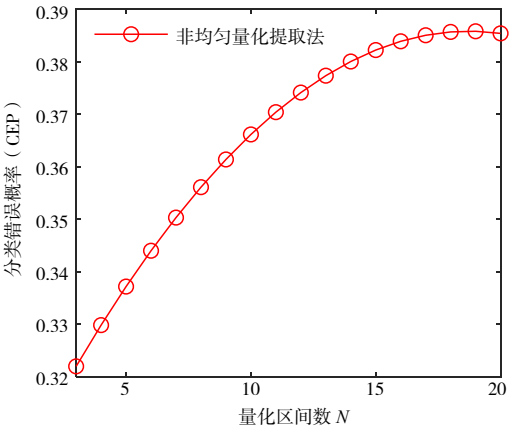

图3　分类错误概率随量化区间数N的变化情况

表1对比了分别采用RF算法与个体分类算法SVM、DT时系统的平均安全容量. 结果表明采用RF算法,系统可获得更高的安全性能.

表1 当中继节点数$K=4,6,8$时,采用三种优化算法系统平均安全容量对比情况

中继点数	系统平均安全容量(bit/s/Hz)		
	RF	SVM	DT
$K=4$	4.2	2.9	3.3
$K=6$	3.7	2.6	2.9
$K=8$	3.2	2.5	2.7

5 结论

研究了给一种基于RF算法的全双工中继系统安全中继选择方案.将合法信道和窃听信道CSI作为模型的输入,将使系统安全容量到达最大的中继节点索引作为模型的输出,把多中继选择问题建模为了一个基于RF算法的多类分类问题.该优化方案的实现主要分为三个阶段,即训练数据准备、模型建立和结果预测.在训练数据准备阶段,由于RF算法要求输入特征是离散值,提出了基于均匀和非均匀量化算法的特征提取方法,将合法信道和窃听信道的CSI转化为离散值.最后通过仿真实验验证了方案性能.

参考文献

[1] WYNER A D. The wire-tap channel[J]. Bell System Technical Journal, 1975, 8(54): 1335 – 1387.

[2] WANG W, TEH K C, LI K H. Relay selection for secure successive AF relaying networks with untrusted nodes[J]. IEEE Transactions on Information Forensics and Security, 2016, 11(11): 2466 – 2476.

[3] KUHESTANI A, MOHAMMADI A, MOHAMMADI M. Joint relay selection and power allocation in large-scale MIMO systems with untrusted relays and passive eavesdroppers[J]. IEEE Transactions on Information Forensics and Security, 2018, 13(2): 341 – 355.

[4] ZHOU Q, ZANG G Z, GAO Y Y, et al. Opportunistic relay selection for secure communication in AF multi-antenna relaying networks[A]. IEEE 2nd Advanced Information Technology, Electronic and Automation Control Conference[C]. Chongqing, China: IEEE, 2017. 212 – 217.

[5] DENG Z X, SANG Q, GAO Y, et al. Optimal relay selection for wireless relay channel with external eavesdropper: a NN-based approach[A]. 2018 IEEE/CIC International Conference on Communications in China[C]. Beijing, China: IEEE, 2019. 515 – 519.

[6] WANG X W. Decision-tree-based relay selection in dual-hop wireless communications[J]. IEEE Transactions on Vehicular Technology, 2019, 68(6): 5 – 23.

[7] BREIMAN L. Random forest[J]. Machine Learning, 2001, 5(1):5 – 32.

[8] WANG J, SHU F, HUANG X H, et al. Optimal coherent combining schemes for relay networks[J]. Wireless Personal Communications, 2016, 88(3): 575 – 585.

[9] SHU F, CHEN Y, YOU X H, et al. Low-complexity optimal spatial channel pairing for AF-based multi-pair two-way relay networks[J]. Science China Information Sciences, 2014, 57(10): 1 – 10.

[10] WANG J, YU H, WU Y P, et al. Pilot optimization and power allocation for OFDM-based full-duplex relay networks with IQ-imbalances[J]. IEEE Access, 2017, 5: 24344 – 24352.

[11] NOSHESE S, PASQUINI L, REICHEL L. Tridiagonal toeplitz matrices: properties and novel applications[J]. Numerical Linear Algebra, 2013, 20(2): 302 – 326.

作者简介

张 梦 女,1988年10月出生于重庆市.现为重庆邮电大学博士研究生.主要研究方向为物理层安全技术.
E-mail:mzhang@stu.cqupt.edu.cn

郑建宏 男,1961年8月出生于四川广安市.现为重庆邮电大学教授.主要研究方向为无线通信技术.
E-mail:zhengjh@cqupt.edu.cn

刘香燕 女,1992年2月出生于四川资阳市.现为重庆邮电大学博士研究生.主要研究方向为资源分配管理.
E-mail:xiangyan.leo@gmail.com

何 云 女,1979年11月出生于湖北武汉市.现为重庆邮电大学博士研究生.主要研究方向为新一代宽带无线通信系统.
E-mail:yunhe@cqupt.edu.cn