DOI:10.19651/j.cnki.emt.2416014

基于改进 KNN-RF 的信息补全算法*

张烈平[1,2] 陈 耀[1,2] 郑新鹏[1,2] 卢海钊[1,2] 张 翠[3]
(1.广西高校先进制造与自动化技术重点实验室(桂林理工大学) 桂林 541006;
2.广西特种工程装备与控制重点实验室(桂林航天工业学院) 桂林 541004; 3.南宁理工学院信息工程学院 桂林 541006)

摘 要: 针对室内指纹定位指纹库数据在实际环境中存在数据缺失导致定位误差大的问题,本文提出了一种改进距离公式的 K 近邻-随机森林的信息补全算法。首先,采用高斯滤波对收集的指纹数据进行预处理,去除干扰数据项,提高数据可靠性。其次,在将指纹数据划分为训练集和测试集的基础上,采用结合欧氏距离和曼哈顿距离的 KNN 算法获得近邻集合样本,随后用 RF 算法对近邻集合训练进行优化,再把各个决策树的预测结果取平均值,得到缺失数据的预测值。最后,将改进的补全算法与 KNN、改进的 KNN、RF 和 KNN-RF 补全算法进行对比。实验结果表明,本文的改进补全算法的预测准确率和精度均优于其他算法,预测的准确率达 91.3%。同时本文补全算法的指纹库平均定位误差为 1.82 m,相较于其他补全算法的指纹库定位误差降低了 1.6%~7.2%,定位性能更好。

关键词: 室内定位;KNN;RF;指纹数据库;信息补全
中图分类号: TP181;TN92 **文献标识码**: A **国家标准学科分类代码**: 520.2

Information completion algorithm based on improved KNN-RF

Zhang Lieping[1,2] Chen Yao[1,2] Zheng Xinpeng[1,2] Lu Haizhao[1,2] Zhang Cui[3]
(1. Key Laboratory of Advanced Manufacturing and Automation Technology(Guilin University of Technology), Education Department of Guangxi Zhuang Autonomous Region, Guilin 541006, China; 2. Guangxi Key Laboratory of Special Engineering Equipment and Control (Guilin University of Aerospace Technology), Guilin 541004, China; 3. School of Information Engineering, Nanning College of Technology, Guilin 541006, China)

Abstract: This paper proposes an information complementation algorithm of K nearest neighbor-random forest with an improved distance formula, aiming at the problem of indoor fingerprint localization fingerprint database data in the real environment with missing data leading to large positioning errors. First, the gathered fingerprint data is preprocessed using Gaussian filtering to eliminate interfering data points and enhance data dependability. Second, the nearest-neighbor set is sampled using the KNN algorithm, which combines Manhattan distance and Euclidean distance. The RF algorithm is then used to optimize the training of the nearest-neighbor set, and the prediction results of each individual decision tree are averaged to determine the predicted values of the missing data. This process is based on the division of the fingerprint data into training and testing sets. Finally, the improved complementary algorithm is compared with KNN, improved KNN, RF and KNN-RF complementary algorithms. The experimental results demonstrate that the modified complementary method in this study has superior prediction accuracy and precision than other algorithms, with a prediction accuracy of 91.3%. In the meantime, the fingerprint library of this paper's complimentary algorithm has an average positioning error of 1.82 m, which is 1.6%~7.2% less than that of other complementary algorithms, and the positioning performance is improved.

Keywords: indoor localization;KNN;RF;fingerprint database;information completion

0 引 言

近年来,无线技术的快速发展及其无处不在的应用,引起了人们对户外和室内定位跟踪实现的广泛关注。就户外定位服务而言,最著名且最受欢迎的技术是基于卫星的全球定位系统,它在旅游、导航、军事等领域提供无数服务。

收稿日期:2024-05-11
*基金项目:国家自然科学基金项目(61741303)、广西空间信息与测绘重点实验室基金项目(21-238-21-16)、梧州市 2022 年中央引导地方科技发展资金项目(202201001)资助

然而,由于全球定位系统(global positioning system,GPS)的信号来源为卫星且接收信号很薄弱,使该技术在城市、峡谷、高墙以及恶劣天气的情况下表现不佳。同时在大型多层室内场景中,卫星信号的传输方向可能会因建筑物的三维结构和材料而改变,从而影响定位的准确性,因此 GPS 在室内定位和地下环境中可能变得完全无用[1]。为了实现高精度的室内定位服务,近年来人们一直在寻找合适的室内定位技术,目前常用的技术有无线保真(wireless fidelity, WiFi)技术、射频识别(radio frequency identification,RFID)、超宽带(ultra wide band,UWB)和蓝牙技术[2]。由于 WiFi 技术具有部署成本低、覆盖范围广、传输速率快等优点,以及大多数移动设备都配备了内置的 WiFi 模块,从而避免了对额外硬件的需求。因此,基于 WiFi 的室内定位技术已成为室内定位领域内的研究热点。目前室内 WiFi 指纹定位主要是将接收信号强度(recevied signal strength indication,RSSI)作为指纹数据,与对应的待定位点进行匹配以实现定位。基于 WiFi 的指纹定位也因其复杂度低、易于实现且定位精度较高而被广泛使用[3]。

近年来,大多数室内定位研究使用经过训练的机器学习模型来取代传统技术,机器学习算法提供了从广泛的数据集中提取模式和关系的能力,从而提高了定位的精度[4]。商磊等[5]针对室内复杂环境下选取 K 近邻(K nearest neighbor,KNN)算法导致定位精度变差的问题,提出了一种基于 KNN 的改进定位算法,通过均值漂移(MeanShift)聚类选取自适应 KNN,利用几何位置对自适应 KNN 进行动态优选以提高定位精度和稳定性。Chen 等[6]针对无线定位系统中的 RSSI 值,提出一种改进的加权 KNN 指纹定位算法,使用 RSSI 对指纹数据进行分区,采用几何方法对相关度进行分析和定义,并对 KNN 进行重新加权,实现对该区域的精确定位。Ferreira 等[7]考虑了实际条件下 RSSI 值的变化,采用四分位数分析进行数据预处理,然后使用 KNN 进行定位。最后,在真实和仿真环境下进行实验,结果表明该方法能有效提高定位精度。牟平等[8]提出了一种改进的接入点(access point,AP)选择方法并融合 RF 分类算法进行实时室内定位,在离线阶段使用 AP 的接收信号强度(recevied signal strength,RSS)数据方差以及 AP 出现频率来衡量 AP 稳定性并选取前 m 个稳定的 AP,用拉普拉斯平滑处理避免方差为 0,并以此构建指纹数据库。但这种方法可能不足以应对所有实际情况。例如,当 AP 的 RSS 信号波动较大时,即使经过平滑处理,方差仍可能较大,导致误判 AP 的稳定性。Lee 等[9]提出了一个系统,该系统使用随机森林估计智能手表设备的室内位置,并使用基本服务集标识符(basic service set identifier,BSSID)和 RSS 来解决信号强度相似的问题,该方法依赖于 RSSI 的稳定性,而 RSSI 容易受到环境变化(如人员移动、家具摆放等)影响,从而导致定位误差增加。针对人为噪声下室内定位的预测精度严重下降的问题,Ko 等[10]提出了一种基于深度学习的基于随机森林(random forest,RF)滤波的室内定位方法,在人工噪声攻击的场景下仍然能提供良好的预测精度,但该方法对于其他类型的干扰源(如无线设备的干扰、物理障碍物等)影响定位精度的处理能力有限。

由以上研究结果可知,室内指纹定位系统利用机器学习算法取得了显著进展,通过收集信号强度等数据,系统就能够准确预测目标位置的坐标。但基于单一的机器学习算法的室内定位容易受到信号干扰,导致收集到的 RSSI 指纹库数据缺失率较高,从而影响定位精度和定位性能[11]。针对这一问题,本文提出了一种结合欧式距离和曼哈顿距离的 KNN-RF 信息补全算法(命名为改进的 KNN-RF 信息补全算法),对缺失的指纹数据进行预测填补,改善室内定位的精度。

1 KNN 算法和 RF 算法介绍

1.1 KNN 算法

KNN 算法,是一种基本的机器学习算法,也是一种分类和回归方法。KNN 算法的思想就是根据两点之间的距离的这一个物理量来衡量不同的点之间的相似程度,再根据相似程度进行分类和回归预测。假设在指纹定位的离线阶段,共布置了 L 个指纹点,即这 L 个指纹可被记为 $\{F_1, F_2, \cdots, F_l\}$,另外这 L 个指纹对应的坐标被记为 $\{L_1, L_2, \cdots, L_l\}$。指纹库建立完成后,把待测位置的指纹记为 S,为所有的 AP 多次测量后所取得的 RSSI 的平均值,即 $S=\{s_1, s_2, \cdots, s_n\}$。在存储的过程中,指纹都采取了 $F_i=\{r_1^i, r_2^i, \cdots, r_l^i\}$ 的形式,其中分别表示为第 i 个指纹中分别采取的第 $1-n$ 个 AP 点的值。通过采用如式(1)所示的欧氏距离或者如式(2)所示曼哈顿距离,可以得到待测位置的指纹 S 的坐标与离线指纹库中收集到的数据之间的距离差。

$$w_1=\sqrt{\sum_{i=1}^{n}(s_i-r_i)^2} \tag{1}$$

$$w_2=\sum_{i=1}^{n}|s_i-r_i| \tag{2}$$

其中,n 表示第 n 个 AP,s_i 为预测值,r_i 为实际值。

在完成上述步骤以后,可以根据距离的大小,选取其中的最小值来作为估算位置,求取其平均值来得到最终的结果,式(3)如下:

$$L=(\hat{x},\hat{y})=\frac{1}{k}\sum_{i=1}^{k}(x_i,y_i) \tag{3}$$

其中,k 表示取得 k 个近邻点,(x_i, y_i) 表示 k 个近邻点的每个点的标。

虽然 KNN 算法已经在许多研究中被广泛应用,但是 KNN 算法仍然面临初始训练阶段训练样本量过大、在更高维的特征数据中距离计算变模糊等问题,进而导致指纹数据库的精度降低[12]。而 RF 算法具有高准确性、鲁棒性、对高维数据的处理能力、自动特征选择等优势,完全可以用 RF 算法对 KNN 近邻集合进行优化。

1.2　RF 算法

RF 算法是 Breiman 在 2001 年提出的一种以多棵决策树为估计器的集成学习算法，它在构建决策树时从训练数据集中用有放回的方式抽取样本以保证随机抽样，这种抽样方法又称引导聚集(bootstrap aggregating，Bagging)[13]。

在构建随机森林时，首先假设数据集的样本个数为 N 个，以 Bagging 法从中随机取样创建训练集，其中特征数目为 M 个，再从 M 个不同的特征中随机选择 m 个特征作为决策树的分支节点，取 $m < M$，且保持取出的特征个数不变。然后对 m 个特征采用某种策略方法(信息增益、信息增益率、基尼指数)来确定一个分支节点的最佳属性。再重复以上步骤就可以构建大量的决策树，形成随机森林。当有指纹数据输入随机森林模型时，所有的决策树都会产生各自的预测结果，最后通过所有决策树投票来确定随机森林的输出结果。图 1 为随机森林算法的流程图。

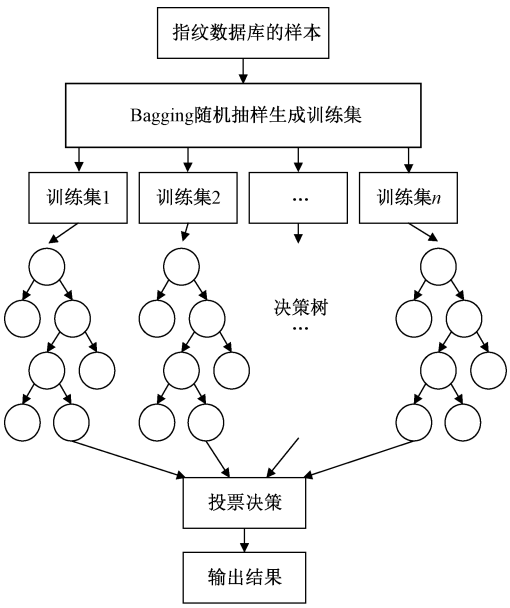

图 1　RF 算法流程图

Fig. 1　Flowchart of RF algorithm

RF 算法的随机特性体现在构建过程中数据样本的随机性和信息特征选择的随机性，能较好的应用在复杂相关性的指纹数据环境中[14]。但是 RF 算法也存在缺点，就是对于数据集的要求比较高，如果数据中存在大量离群冗余的数据时，数据之间的差异性就会偏大，不能体现指纹信息数据间的相关性，最终会导致预测性能降低。其次，当指纹数据中存在大量的缺失信息时，会增加另一部分信息的比重，导致在训练的时候损失过多的有效信息。

2　基于欧氏距离和曼哈顿距离的 KNN-RF 信息补全算法

2.1　算法设计

在 KNN 算法中 k 的取值很大程度上决定了信息补全的精度，对于缺失数据集数据本身的属性特征没有加以分析，会影响数据预测的精度。在 KNN 算法中，结果通常用欧式距离表示，但是在实际的室内定位指纹坐标中，该方法表现过于理想化，很容易忽略实际环境中的各种变量关系，与实际环境中的距离相比，欧氏距离的计算结果往往存在较大误差差。相反，曼哈顿距离更具稳定性，能够有效减少信号波动的影响范围。然而，该算法仅在指纹信息的特征值差异较小的情况下才适用，因为如果曼哈顿距离过大，距离的贡献值就会掩盖其他特征数据的近邻关系[15]。因此，本文提出了一种结合欧氏距离和曼哈顿距离，同时计算两种距离后求均值的方法来改进距离测量，再利用 RF 在缺失指纹特征补全的全局学习优点增强数据样本之间的关联性，结合改进后的 KNN 算法进行改善数据集，增强数据质量，从而完成算法的结合。基于欧氏距离和曼哈顿距离的 KNN-RF 信息补全算法流程图如图 2 所示。

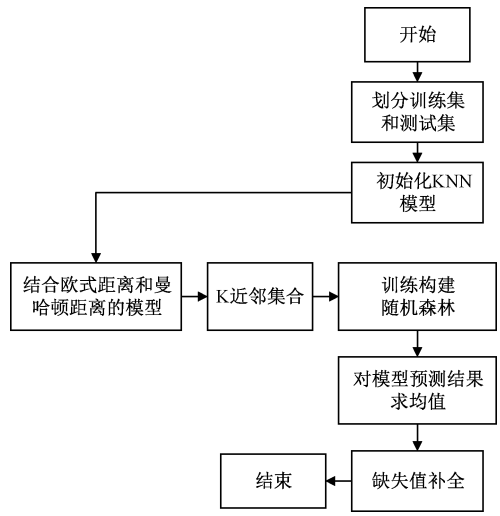

图 2　基于改进 KNN-RF 的信息补全算法流程图

Fig. 2　Flowchart of the information complementation algorithm based on improved KNN-RF

2.2　算法实现

设含有缺失指纹数据集为 X_{mis}，记为 $C_i = \{C_1, C_2, \cdots, C_n\}$，每个样本数据中包含 M 维属性，记为 $C_i = \{C_{i1}, C_{i2}, \cdots, C_{iM}\}$，$i = 1, 2, \cdots, n$。对原始数据集中的样本进行测试，计算缺失数据与其他所有样本 X_{els} 的欧氏距离与曼哈顿距离的均值，如式(4)～(6)所示。

$$D_1 = \sqrt{\sum_{i=1}^{n}(X_{mis,i} - X_{els,j})^2} \tag{4}$$

$$D_2 = \sum_{j=1}^{n} | X_{mis,i} - X_{els,j} | \tag{5}$$

$$D_{ij}(X_{mis}, X_{els}) = \frac{1}{2}(D_1 + D_2) \tag{6}$$

其中，$\sqrt{\sum_{i=1}^{n}(X_{mis,i} - X_{els,j})^2}$ 表示第 i 个缺失数据与第 j 其余样本数据之间的欧式距离；$\sum_{j=1}^{n} | X_{mis,j} - X_{els,j} |$ 表

示第 i 个缺失数据与第 j 其余样本数据之间的哈顿距离。

遍历 X_{mis} 中所有样本,计算得到每个缺失样本与 X_{mis} 的距离,得到距离集合 $D_{ij}(X_{mis},X_{els})$,再根据距离,就可以得到与 X_{mis} 最近的 K 的样本。

一般的,缺失数据的类别由其距离最近的几个共同元素决定,常用的方法是直接平均数或者是众数,但是这样就忽视了属性之间的相互关系性以及属性和样本之间的对应关系。因此,本文算法在通过 KNN 算法得到近邻集合后,利用 RF 进行下一优化。

首先通过改进距离后的 KNN 算法获取近邻集合,接着将含有缺失值的数据样本导入 RF 训练,得到训练集 $A_i=\{a_1,a_2,\cdots,a_n\}$。接着,从训练集 A_i 中有放回地抽取,每次抽样都是独立的不影响下一次的概率,形成训练子集 D_t。接着使用训练决策树,节点的切分是根据所有属性中随机选择 H 个特征($H<<S$),记 $H_i=H_1,H_2,\cdots$。最后,根据最优特征切分左右子树[16],如式(7)所示。

$$H=\frac{1}{2}\sqrt{S} \tag{7}$$

其中,S 表示训练集 A_i 中的样本数据中的维度特征。

信息熵(information entropy)是评价机器学习性能好坏的一种常用指标,用于衡量样本集合的"纯度",其数值越小则说明训练效果越好。假设训练子集 D_t 中的第 q 类样本所占比例记为 P_q,则训练子集 D_t 的信息熵为:

$$Ent(D_t)=-\sum_{q=1}^{y}P_q\log_2 P_q \tag{8}$$

其中,y 为 D_t 中的种类。$Ent(D_t)$ 值越小,则说明 D_t 的纯度越高,训练效果越好。对于 D_t 中的 H 维特征来说,不同的特征作节点分支所得到的效果也是不一样的。

假设特征集 H_i 中每一个属性 H' 都会有 h 多个可能的取值,采用 H' 对训练子集 D_t 进行划分操作,就会得到 h 个分支节点。当中的第 h' 个分支节点是包含了训练子集 D_t 中在特征 H_i 上取值 h' 的所有样本,把这些样本记为 $D_t^{h'}$,则所获得的信息增益可以表示为:

$$Gain(D_t,H')=Ent(D_t)-\sum_{h'=1}^{h}\frac{|D_t^{h'}|}{D_t}Ent(D_t^{h'}) \tag{9}$$

信息增益越大,则说明决策树的划分效果就很好,训练的效果也就更好。反复执行上面的流程,直到所有的特征都被用来训练或者标记为最终结果为止,这时决策树已经建立好了,预测的结果由这些树共同决定最终的补全结果[17]。如果是连续的数据,则缺失数据就是对 n 棵决策树的预测值取平均,如果是离散的数据,则缺失数据就是取决策树中类别数量较多的类型来作为补全结果[18]。

最后,KNN 算法取到的 K 个最近邻集合 $D_{ij}(X_{mis},X_{els})$ 输入到树模型中,得到多棵树的预测值,然后将多棵树的预测结果求均值就得到了最终的估算值,如式(10)所示。

$$\hat{L}_{mis}=\frac{1}{T}\sum_{t=1}^{T}Tree_t(D_{ij}) \tag{10}$$

其中,$\hat{L}_{mis}$ 是对缺失值的估算值;T 表示树的数量;$Tree_t(D_{ij})$ 表示第 t 棵树对指纹缺失数据的预测值。

3 实验与结果分析

3.1 实验参数及信息补全

实验采用的 AP 设备为 MERCURY- MW305R,一台安装自主开发的 RSSI 信息采集软件的手机作接收器,电脑 CPU 为 Intel Core i7-6850K,操作系统为 Windows 10 专业版。在实验区域内设置 5 个距离地面高度约为 0.9 m 的 AP 点作为信号源,位置如图 3 蓝色信号图标所示。在实验中采用网格切割的方法,在实验室区域内设置共 80 个参考点。在每个参考点分别采集 5 个 WiFi 热点的 RSSI 值,再对 RSSI 值高斯滤波处理后求均值,建立指纹数据库。

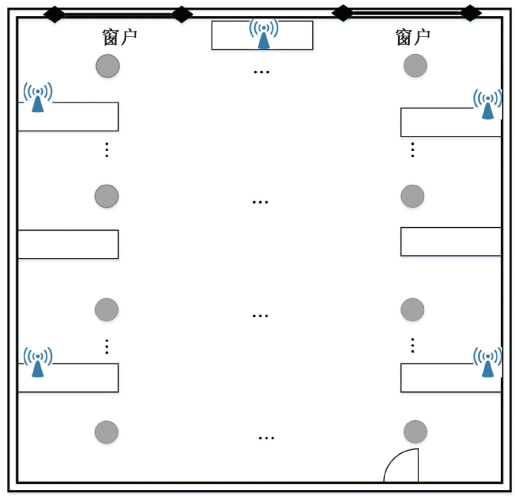

图 3 实验环境实景图

Fig. 3 Live view of the experimental environment

在经过高斯滤波预处理后的指纹库中发现,来自不同 AP 的多个参考点信息数据都出现了 RSSI 强度信息缺失情况,影响了指纹库数据的完整性和后期定位的精度。部分缺失的数据如表 1 所示,其中,信号单位为 dBm,Nan 表示缺失的指纹数据。

表 1 存在缺失值的部分数据

Table 1 Partial data with missing values

AP1	AP2	AP3	AP4	AP5	参考坐标
−37.19	−57.68	−68.17	Nan	−55.78	(1,1)
−37.06	Nan	−57.37	−57.55	−45.12	(2,1)
Nan	−56.04	−66.08	−52.28	−52.80	(5,1)
…	…	…	…	…	…
−40.57	−51.39	−60.16	Nan	Nan	(21,1)
−45.29	−51.32	Nan	−52.47	−52.09	(54,1)

首先初始化模型，在本次实验中将比例分别设置为80%和20%来划分为训练集和测试集。然后将这些数据输入改进距离公式后的KNN-RF算法模型中补全缺失值，从而得到完整的指纹数据库。在实验过程中，对多个参数进行调节，以确保模型的性能达到最佳状态，这些参数包括每棵决策树的最大树深(max_depth)、内部节点再划分的最小样本数(min_samples_split)、随机特征比例(Random_feature_ratio)和叶子节点包含的最小节点数(min_samples_leaf)[19]。这些参数使用默认的设置，对于参数 K 则使用 $Ent(D_t)$ 来衡量不同 K 值预测效果的好坏。max_depth、min_samples_split、Random_feature_ratio 和 min_samples_leaf 的参数分别是－1、6、0.8 和 4，参数 K 的设置结果如表 2 所示。

表 2　参数设置及实验结果

Table 2　Parameter settings and experimental results

K	最大树深	最小样本数	随机特征比例	最小节点数	Ent(D_t)
10	－1	6	0.8	4	0.851
20	－1	6	0.8	4	0.826
30	－1	6	0.8	4	0.763
50	－1	6	0.8	4	0.792
80	－1	6	0.8	4	0.810

根据表 2 中信息熵的大小情况，$Ent(D_t)$ 值越小说明预测性能越好，从而可以得到本次实验的参数 K 设置。因此，可以得出改进的 KNN-RF 信息补全算法模型各项的参数设置如表 3 所示。

表 3　最佳参数设置

Table 3　Parameter settings

参数	具体数值
K	30
最大树深	－1
最小样本数	6
随机特征比例	0.8
最小节点数	4

3.2　补全性能分析

将改进的 KNN-RF 信息补全算法的预测模型参数设置为表 3 中数值，对表 1 中的缺失数据进行缺失值补全，补全效果如表 4 所示。

为了进一步分析改进的 KNN-RF 信息补全算法的优劣性，使用上表 4 的实验数据，分别对比了基于 KNN 的信息补全法、基于改进 KNN 的信息补全法、基于 RF 的信息补全法、基于 KNN 和 RF 的信息补全法和基于改进的 KNN-RF 信息补全法的指纹信息预测准确率。为了验证对比，实验人为去除了指纹库的某段信息值，造成数据缺失，并利用以上几种算法的预测值和真实去除的值做比较分析，来验证算法的准确率，具体结果由图 4 所示。

表 4　基于改进的 KNN-RF 的补全数据表

Table 4　Complementary data table based on the improved KNN-RF

AP1	AP2	AP3	AP4	AP5	参考坐标
－37.19	－57.68	－68.17	－70.51	－55.78	(1,1)
－37.06	－53.56	－57.37	－57.55	－45.12	(2,1)
－33.56	－56.04	－66.08	－52.28	－52.80	(5,1)
…	…	…	…	…	…
－40.57	－51.39	－60.16	－54.59	－52.54	(21,1)
－45.29	－51.32	－64.39	－52.47	－52.09	(54,1)

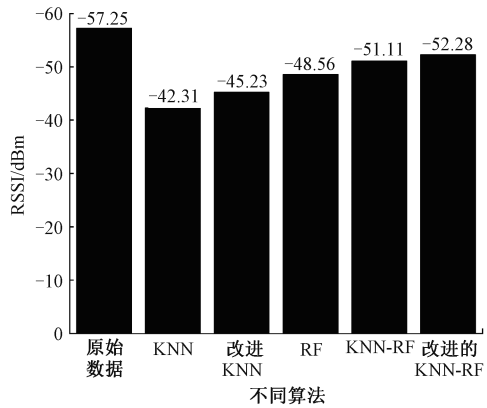

图 4　不同算法模型对比

Fig. 4　Comparison of different algorithmic models

图 4 中结果均为多次补全的平均值，从图 4 中可以看出，缺失的原始数据为－57.25 dBm，4 种算法都作出了各自的信息补全。在使用改进的 KNN-RF 补全算法对缺失值补全后得到的预测结果为－52.28 dBm，预测的准确率达91.3%。而 KNN 算法的预测结果为－42.31 dBm，准确率仅为 73.9%。改进 KNN 算法的预测结果为－45.23 dBm，准确率仅为 79%。RF 算法的预测结果为－48.56 dBm，准确率为 84.8%，KNN-RF 算法的预测结果为－51.11 dBm，准确率为 89.2%。改进的 KNN-RF 信息补全算法有更佳的预测效果，预测补全的平均值与真实值最为接近。

3.3　定位精度分析

为了探究补全算法对于定位精度的影响，分别使用基于 KNN 的信息补全法、基于改进 KNN 的信息补全法、基于 RF 的信息补全法、基于 KNN 和 RF 的信息补全法和改进的 KNN-RF 信息补全法处理同一缺失的指纹数据库，得到处理后的五个指纹数据库。使用同一参数的 KNN 定位算法来验证不同补全算法的性能。

实验结果如图 5 所示，其中 X 轴为实验测试的次数，Y 轴为定位误差。由图中可以看出，在原指纹库和经过 4 种

不同算法处理后的指纹库中,原指纹库在 KNN 算法下的定位误差范围在 1.99～1.95 m 之间波动。基于 KNN 补全的指纹库在 KNN 算法填的定位误差范围在 1.95～1.92 m 之间波动。基于改进 KNN 补全的指纹库在 KNN 算法填的定位误差范围在 1.91～1.87 m 之间波动。基于 RF 补全的指纹库在 KNN 算法填的定位误差范围在 1.90～1.87 m 之间波动。基于 KNN-RF 补全的指纹库在 KNN 算法填的定位误差范围在 1.86～1.83 m 之间波动。而基于改进的 KNN-RF 补全的指纹库在 KNN 算法填的定位误差范围在 1.84～1.81 m 之间波动。

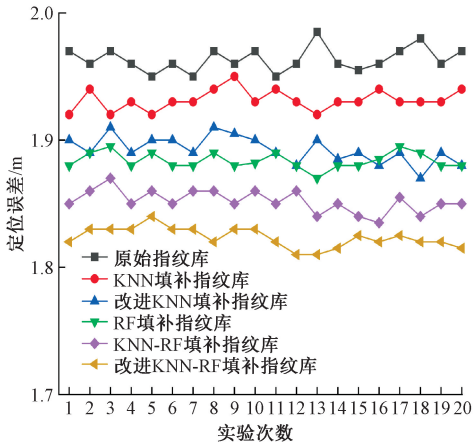

图 5 不同算法定位精度对比

Fig. 5 Comparison of positioning accuracy of different algorithms

实验的平均误差如表 5 所示,原指纹库的平均定位误差为 1.96 m,KNN 补全指纹库的平均误差为 1.93 m,改进 KNN 补全指纹库的平均误差为 1.89 m,RF 补全指纹库的平均定位误差为 1.88 m,KNN-RF 指纹库的平均定位误差为 1.85 m,而改进 KNN-RF 补全指纹库的定位误差为 1.82 m。相较与其他 4 种指纹库,本文提出的算法的定位误差波动范围和平均定位误差更小。总体可以看出,本文提出的改进 KNN-RF 信息补全算法在定位时的稳定性和精度都有所提高,在实际应用中会有更佳的表现。

表 5 不同模型指纹库平均误差

Table 5 Mean errors of different model fingerprint libraries

指纹库	平均定位误差/m
原始指纹库	1.96
KNN 补全指纹库	1.93
改进 KNN 补全指纹库	1.89
RF 补全指纹库	1.88
KNN-RF 补全指纹库	1.85
改进 KNN-RF 补全指纹库	1.82

4 结 论

基于 WiFi 的指纹定位算法仍是当前热门的研究课题,本文主要针对指纹库在实际环境中存在数据缺失的问题,提出一种解决办法。首先使用高斯滤波预处理指纹库数据,再对 KNN 算法的距离公式进行改进,提出一种基于欧氏距离和曼哈顿距离的 KNN-RF 指纹信息补全算法。然后对算法的参数选择进行实验分析,得到最佳的实验参数后开始训练预测模型。最后将存在缺失值的指纹数据输入模型进行预测补全。实验表明,在预测精度和应用在定位算法的定位精度上,本文提出的补全算法相较于 KNN、改进 KNN、RF 和 KNN-RF 补全算法都有明显提升,具有更好的实用性。在下一步的研究中将进一步改进该补全算法,使其兼容三维指纹数据库,实现空间上的指纹数据补全。

参考文献

[1] PAN L, ZHANG H, ZHANG L Y, et al. Indoor positioning fingerprint database construction basedon CSA-DBSCAN and RCVAE-GAN[J]. Physica Scripta, 2024, 99(5): 055002.

[2] XIANG L, XU Y, CUI J H, et al. GM(1,1)-based weighted K-Nearest neighbor algorithm for indoor localization[J]. Remote Sensing, 2023, 15(15): 3706.

[3] 王开亮,谢亚琴,宦海,等. 基于投票机制的室内 WiFi 指纹定位算法[J]. 电子测量技术, 2023, 46(12): 61-68.
WANG K L, XIE Y Q, HUAN H, et al. Indoor WiFi fingerprinting algorithm based on voting mechanism[J]. Electronic Measurement Technology, 2023, 46(12): 61-68.

[4] YADAV P, SHARMA S C. Unveiling the cutting edge: A comprehensive survey of localization techniques in WSN, leveraging optimization and machine learning approaches[J]. Wireless Personal Communications, 2023, 132(4): 2293-2362.

[5] 商磊,关维国,龚瑞雪. 基于聚类优选自适应 KNN 的改进定位算法[J]. 传感器与微系统, 2023, 42(3): 136-139.
SHANG L, GUAN W G, GONG R X. Improved localization algorithm based on clustering optimization and adaptive KNN[J]. Transducer and Microsystem Technologies, 2023, 42(3):136-139.

[6] CHEN B H, MA J, ZHANG L F, et al. An improved weighted KNN fingerprint positioning algorithm[J]. Wireless Networks, 2024, 30: 6011-6022.

[7] FERREIRA D, SOUZA R, CARVALHO C. QA-KNN: Indoor localization based on quartileanalysis and the KNN classifier for wireless networks[J]. Sensors, 2020, 20(17): 4714.

[8] 牟平,凌铭,胡锐. 基于改进 AP 选择的融合随机森林室内定位算法[J]. 全球定位系统, 2021, 46(5): 33-38.

MU P, LING M, HU R. Indoor location algorithm based on improved AP selection and random forest fusion[J]. GNSS World of China, 2021, 46(5): 33-38.

[9] LEE S, KIM J, MOON N. Random forest and WiFi fingerprint-based indoor location recognition system using smart watch[J]. Human-Centric Computing and Information Sciences, 2019, 9(1): 6.

[10] KO D, CHOI S H, AHN S, et al. Robust indoor localization methods using random forest-based filter against MAC spoofing attack[J]. Sensors, 2020, 20(23): 6756.

[11] 卢海钊,彭慧豪,唐滔,等. 基于 KNN 和 XGBoost 的室内指纹定位算法[J]. 电子测量技术, 2023, 46(2): 81-86.
LU H ZH, PENG H H, TANG T, et al. Indoor fingerprint localization algorithm based on KNN and XGBoost[J]. Electronic Measurement Technology, 2023, 46(2): 81-86.

[12] 欧锦添,乐燕芬,施伟斌. 基于密文 KNN 检索的室内定位隐私保护算法[J]. 数据采集与处理, 2024, 39(2): 456-470.
OU J T, LE Y F, SHI W B. Indoor location privacy protection algorithm based on ciphertext KNN retrieval[J]. Journal of Data Acquisition and Processing, 2024, 39(2): 456-470.

[13] 赵圣健,朱翠,王雅妮. 基于随机森林的 RFID 室内区域定位方法[J]. 北京信息科技大学学报(自然科学版), 2021, 36(3): 8-12.
ZHAO SH J, ZHU C, WANG Y N. RFID indoor area positioning method based on random forest[J]. Journal of Beijing Information Science & Technology University(Science and Technology Edition), 2021, 36(3): 8-12.

[14] LU M X, TAY L T, MOHAMAD-SALEH J. Landslide susceptibility analysis using random forest model with SMOTE-ENN resampling algorithm[J]. Geomatics, Natural Hazards and Risk, 2024, 15(1): 2314565.

[15] PÉREZ-NAVARRO A, MONTOLIU R, SANSANO-SANSANO E, et al. Accuracy of a single position estimate for kNN-based fingerprinting indoor positioning applying error propagation theory[J]. IEEE Sensors Journal, 2023, 23(16): 18765-18775.

[16] CAMANA M R, GARCIA C E, HWANG T, et al. A REM update methodology based on clustering and random forest[J]. Applied Sciences, 2023, 13(9): 5362.

[17] VARMA P S, ANAND V. Random forest learning based indoor localization as an IoT service for smart buildings[J]. Wireless Personal Communications, 2021, 117: 3209-3227.

[18] 曲佳,王旭东,吴楠,等. 基于随机森林算法的室内可见光指纹定位方法[J]. 光通信技术, 2023, 47(1): 1-7.
QU J, WANG X D, WU N, et al. Indoor visible light fingerprint positioning method based on random forest algorithm[J]. Optical Communication Technology, 2023, 47(1): 1-7.

[19] WANG M, YE X W, YING X H, et al. Data imputation of soil pressure on shield tunnel lining based on random forest model[J]. Sensors, 2024, 24(5): 1560.

作者简介

张烈平(通信作者),博士,教授,主要研究方向为传感器与智能信息处理技术。

E-mail:zlp@guat.edu.cn

陈耀,硕士研究生,主要研究方向为传感器与智能信息处理技术。

E-mail:2766997370@qq.com

郑新鹏,硕士研究生,主要研究方向为传感器与智能信息处理技术。

E-mail:1412819356@qq.com

卢海钊,硕士研究生,主要研究方向为传感器与智能信息处理技术。

E-mail:726997100@qq.com

张翠,硕士,副教授,主要研究方向为传感器与智能信息处理技术。

E-mail:361745092@qq.com