ISSN 1000-1239/CN 11-1777/TP
50(3): 568-577, 2013

一种基于 Wi-Fi 信号指纹的楼宇内定位算法

牛建伟　刘　洋　卢邦辉　宋文芳
(软件开发环境国家重点实验室(北京航空航天大学)　北京　100191)
(北京航空航天大学计算机学院　北京　100191)
(niujianwei@buaa.edu.cn)

An In-Building Localization Algorithm Based on Wi-Fi Signal Fingerprint

Niu Jianwei, Liu Yang, Lu Banghui, and Song Wenfang
(*State Key Laboratory of Software Development Environment* (*Beihang University*), *Beijing* 100191)
(*School of Computer Science and Engineering*, *Beihang University*, *Beijing* 100191)

Abstract　Since GPS cannot be used under in-building environment and current in-building localization approaches require pre-installed infrastructure, in-building localization becomes a problem demanding prompt solutions for location-based services. Therefore, this paper proposes a novel room-level in-building localization algorithm R-kNN (relativity k-nearest neighbor), which solves the localization problem by leveraging MAC address and RSSI (received signal strength indication) of Wi-Fi access points (APs) deployed in buildings. R-kNN falls into category of property-weighted k-nearest neighbor algorithm. By assigning the weight of each AP according to the relativity between AP pairs, R-kNN can reduce the negative effect of dimension redundancy. Moreover, since it makes no assumption on the physical distribution of rooms and APs, R-kNN can work well with existing APs without deploying any new infrastructure or modifying the existing ones. Experimental results demonstrate that when a large number of APs are available, the localization accuracy of R-kNN is bigger than those of the original kNN algorithm and naïve Bayes classifier, while its false positive ratio and false negative ratio is smaller than those of the original kNN algorithm and Naïve Bayes classifier in most cases.

Key words　in-building localization; Wi-Fi; received signal strength indication (RSSI); k-nearest neighbor; property-weighted k-nearest neighbor

摘　要　由于 GPS 无法在楼宇内使用,而目前的楼宇内定位技术一般都需要预先部署额外的设施,因此楼宇内无基础设施定位成为了一个热点研究问题.提出了一种利用 Wi-Fi 接入点的 MAC 地址和 RSSI (received signal strength indication)值,通过机器分类的方式实现楼宇内房间级定位的算法 R-kNN (relativity k-nearest neighbor). R-kNN 是一种属性加权 k 近邻算法,它通过将 AP 之间的相关性反应在权值的分配上,有效地降低了维度冗余对分类准确率的负面影响. R-kNN 没有对房间和 AP 的物理位置做出任何假设,只需要使用环境中现存的 AP 就可以取得较好的定位效果,无需部署任何额外设施或修改现有设施.实验结果表明,在 AP 数量较多的楼宇环境中,R-kNN 能够取得比 k 近邻算法和朴素贝叶斯分类器更好的定位效果.

关键词　楼宇内定位;Wi-Fi;RSSI;k 近邻算法;属性加权 k 近邻算法

中图法分类号　TP391

收稿日期:2011-08-22;**修回日期**:2011-11-16

基金项目:国家自然科学基金项目(61170296);新世纪优秀人才支持计划基金项目(NECT-09-0028);软件开发环境国家重点实验室基金项目(SKLSDE-2012ZX-17)

近年来,随着移动设备技术的快速发展和日益普及,如何充分利用设备的移动性为用户提供更加丰富和完善的服务引起了众多研究者的关注,而基于位置的服务(location based service, LBS)已经成为近年来移动计算研究领域的热点问题之一.提供基于位置服务的前提是移动设备需要知道自身所处的物理位置.当移动设备位于户外时,全球定位系统(global position system, GPS)可以为这类应用提供一种简单有效的解决方案.然而,GPS 无法在楼宇内工作,因此如何在楼宇内对移动设备进行定位依然是一个需要解决的科学问题.随着 Wi-Fi 技术应用的普及,WLAN(wireless local area network)接入点(access point, AP)在城市楼宇内已经广泛部署,这使得很多楼宇内环境(例如办公楼、咖啡厅等)中几乎每个角落都能够被 Wi-Fi 信号覆盖,因而基于 Wi-Fi 信号的楼宇内定位技术得到了迅速的发展.然而传统的基于到达角度定位法(angle of arrival, AOA)[1-2]、到达时间定位法(time of arrival, TOA)[3]和信号强度分析法[4-6]的 Wi-Fi 信号定位算法,在复杂的楼宇内环境应用时都难以取得较好的效果,而基于机器学习理论的定位算法由于能够规避信号强度测距等问题,因而可以获得较高的定位准确率.

本文提出了一种通过检测楼宇环境中 AP 的 MAC 地址和接收信号长度指示(received signal strength indication, RSSI)值,实现了一种用于楼宇内定位的机器学习算法 R-kNN(relativity k-nearest neighbor).R-kNN 是一种面向楼宇内定位改进的 k 近邻算法,能够充分利用环境中现有的 AP,实现较为精确的房间级楼宇内定位.作为一种 k 近邻的改进算法,R-kNN 能够在使用与原始 k 近邻算法和朴素贝叶斯分类器完全相同的训练数据的前提下,获得更高的定位准确率.而且在使用 R-kNN 算法进行楼宇内定位时,不需要考虑房间和 AP 的物理位置信息,也不需要部署任何额外设施,在移动设备上提供基于位置的服务时实现简洁,具有很好的可扩展性.

1 常见定位方法和研究现状

目前使用较广的基于 Wi-Fi 信号的定位方法主要有到达角度定位法、到达时间定位法、信号强度分析法和位置指纹分类法[6-12]等.

基于 AOA 的定位方法通过接收机天线阵列测出 AP 发射无线信号的入射角,然后根据多个 AP 的角度方位线的交叉点来估计移动设备的位置,该方法适用于视距(line of sight)传播的情况,设备复杂度较高.基于 TOA 的定位方法根据信号的传播时间来估计移动设备与 AP 的距离,该方法要求 AP 有非常精准的时钟,且收发信号的双方能够实现精确时钟同步.基于 AOA 和 TOA 的定位算法一般都需要一定数量预先知道确切位置的 AP 才能工作,且需要复杂的仪器设备或者修改 IEEE802.11 的物理层协议.虽然 Llombart 等人[3]提出了一种不需要修改 AP 软硬件的到达时间定位方法,但这种方法的定位精度过于依赖 AP 和移动设备的运行状况,且仍然需要预先知道 AP 的确切位置.在楼宇内 WLAN 环境下,单个 AP 的覆盖范围有限,计算无线信号的传输时延的误差较大.另外,由于楼宇内环境复杂,无线信号入射角度的测量也很难十分准确.因此,基于 AOA 或 TOA 的方法并不太适用于楼宇内环境.

信号强度分析法的基础是电磁波的衰减特性.电磁波在介质中传播时接收到的信号强度与传输的距离平方成反比,因此可以利用 RSSI 的观测值来估计该点到 AP 的距离.这类定位方法中效果较好的是 Bahl 等提出的 RADAR 系统[5].RADAR 利用基站测得的 RSSI 推测用户相对基站的距离和方向实现定位,同时 RADAR 还通过综合多个基站的测量结果来改善定位精度.然而,楼宇内空间中存在着大量的障碍物,如墙壁、房门、桌椅、箱柜等等,当它们处于无线信号的传播路径上时会对信号的衰减程度产生难以计算的影响,因此在楼宇内环境下利用 RSSI 值的测距精度不高.而且信号强度分析法与 AOA 和 TOA 一样,需要预先知道一定数量已知位置的 AP 才能够工作.

位置指纹分类法是一类基于机器学习的定位方法,其基本思想是记录特定位置的信号指纹(一系列属性的观测值,一般是 AP 的 RSSI 值),通过比较测试样本与位置指纹的相似程度,判定测试点是否在指定的位置,实质上是将定位问题转换为一个分类问题.虽然 Castro 等人[9]提出使用信噪比作为 IEEE802.11 网络中用于定位的位置特征,但是,由于无线信号在楼宇内环境的传播受到多种因素的干扰,信噪比的变化更加不稳定.因此,目前在该领域的研究工作基本上都是基于信号强度 RSSI 来进行的.与 AOA,TOA 和信号强度分析法相比,基于 RSSI 的位置指纹分类法不依赖于角度或距离等几何量,而是将 RSSI 的观测值本身作为标定位置的依据,这样就规避了接收信号强度测距等问题.在位

置指纹分类法中使用比较广泛的分类算法有 k 近邻算法[6-7]和朴素贝叶斯分类器[8-11]. 在使用朴素贝叶斯分类器时,往往需要假设各 AP 的 RSSI 的观测值之间相互独立,然而在实际环境中各 AP 的 RSSI 的分布并不独立,而是具有一定的相关性. 相比之下,k 近邻算法是无参数的分类方法,对位置和属性之间的对应关系不需要做出任何假设,所以更加具有普适性.

2 R-kNN 算法

原始 k 近邻算法虽然具有较高的普适性,但由于考虑的因素过于简单,使得原始 k 近邻算法在不同的数据集上的分类准确率波动较大. 针对一些原始 k 近邻算法无法取得较高分类准确率的应用场景,学者们提出了很多针对 k 近邻算法的改进策略[13-15],其中比较典型的有距离加权方法和属性加权方法.

按距离(本文后续部分中如无特殊说明,"距离"一词均指代样本空间中的欧氏距离)加权的 k 近邻算法最早由 Dudani[14]提出,其基本思想是:与测试样本距离越近的训练样本与测试样本的关系越紧密. 距离加权方法按训练样本和测试样本的距离给不同的训练样本赋予不同的权值,且在算法统计特定类别的训练样本数量时不再做单纯的计数,而是将属于该类别的训练样本的权值进行累加. 距离加权方法使得测试样本更容易被分类为离它较近的训练样本所属的类别. 当训练样本的分布特别稀疏时,距离加权方法能够有效地降低过大的 k 值对分类准确率产生的不良影响. 然而在 Dudani 之后,Bailey 等人[15]证明了当训练样本数量较少,距离加权方法并不一定能够取得比原始 k 近邻算法更好的分类效果. 另外,当训练样本规模趋近于无穷时,原始 k 近邻算法总能够取得比任何一种距离加权方案更好的分类效果.

按属性加权的 k 近邻算法的使用更为广泛,其实质是对欧式空间中的坐标轴进行一定程度的缩放. 属性加权 k 近邻算法的基本思想是:当属性较多,即样本空间的维度较高时,往往只有有限的几个属性对分类产生决定性的影响,而其它大部分属性对分类的贡献不大. 由于大部分的属性都对分类没有太大的贡献,如果平等地对待所有属性,则计算欧式距离时关键属性的影响可能被大量的无贡献属性淹没,使得分类结果出现较大的偏差(这一现象被称为维度灾难),所以有必要使用加权的方法对不同的属性区别对待. 当每个属性都被赋予了较为合理的权值时,属性加权方法能够有效地降低维度冗余对分类准确率的影响. 然而,如何为属性分配权值仍然是一个需要针对具体应用场景分析研究的问题.

R-kNN 算法是一种面向楼宇内定位的分类算法,它的设计目标是利用环境中现存的 AP 准确地进行房间级的定位而不需要部署任意额外的基础设施,这种设计适用于办公楼等 AP 分布较集中的场合. R-kNN 算法是一种基于属性加权的 k 近邻改进算法,它能够在同时检测到大量 AP 的情况下,降低维度冗余对分类准确率的影响. 由于 R-kNN 算法中属性的权值是从各 AP 之间的相关系数计算得出,所以将之命名为 R-kNN(relativity k-nearest neighbor).

2.1 AP 的相对位置关系对定位的影响

根据三角定位方法,理论上只需要检测到 3 个不共线的 AP 的 RSSI 值便可以确定一个点的位置. 然而在楼宇内,特别是在办公环境中,通常能够检测到 AP 的数量远大于 5,这就意味着如果将每一个 AP 的 RSSI 值都当做一个独立维度,则必然存在维度冗余,如果直接计算距离将会造成较大的偏差.

事实上,对于楼宇内定位问题来说,这种维度冗余主要产生于 AP 之间的相对位置关系. 当两个 AP 的空间距离较近(如图 1(a)所示)时,对这两个 AP 的 RSSI 的观测值将具有很高的正相关性. 考虑两个 AP 完全重合的极端情况,则从任意位置测量这两个 AP 的 RSSI 值都将相等,即它们的 RSSI 值是完全正相关的. 当两个 AP 处于测试区域的两侧(如图 1(b)所示)时,对这两个 AP 的 RSSI 的观测值将具有很高的负相关性. 如果两个 AP 处于相反方向的无穷远处,则这两个 AP 的 RSSI 的观测值将是完全负相关的.

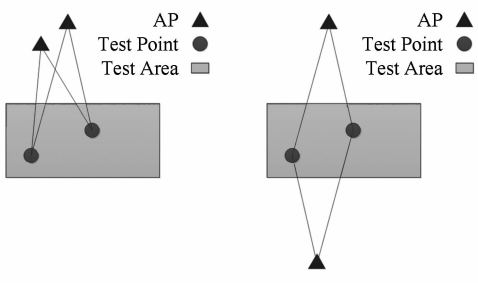


Fig. 1 Schematic diagram of APs' relative location.

图 1 AP 之间相对位置关系示意图

不同 AP 的 RSSI 观测值之间的相关性将对距

离的计算产生显著的影响,进而反映到 k 近邻算法的分类准确率上.如图 2 所示,尽管点 a 到点 b 和点 c 的实际空间距离是相等的,由于在竖直方向上聚集了较多数量的 AP(5 个),而在水平方向上的 AP 数量较少(2 个),所以如果忽略障碍物所造成的影响,认为测得的各个 AP 的 RSSI 可以准确地反映出测试点到 AP 的距离,则计算得出的距离 $D_{ab}^2 \approx 5x^2$ 和 $D_{ac}^2 \approx 2x^2$.虽然基于 k 近邻的定位算法并不要求利用 RSSI 的差来估计实际空间中的距离,然而这种距离计算的误差将可能加大在实际空间中位于同一房间的两个样本在样本空间中的距离,从而造成分类的错误.

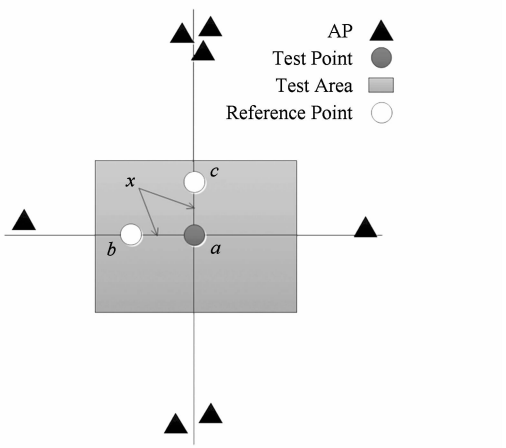


Fig. 2 Dimension redundancy's impact on distance computation.

图 2 维度冗余对距离计算的影响

2.2 权值分配算法

R-kNN 定位算法的核心就是 AP 的权值分配算法.设测试区域内能够检测到的全部 AP 的集合为 V,样本 a 将使用样本属性向量 $\boldsymbol{s}_a=(s_{a,1},s_{a,2},\cdots,s_{a,n})$ 来描述,其中每一个分量 $s_{a,i}$ 表示该样本中第 i 个属性的值(即第 i 个 AP 的 RSSI 的观测值),n 是全部 AP 的数量.如果在一个样本中没有检测到特定的 AP,则将其 RSSI 值设为 -129,即比 RSSI 的下限值低 1.

在给每个 AP 分配权值之前,首先需要计算在测试区域内能够检测到的所有 AP 之间的相关系数.记两个 AP 之间相关系数为 r_{ij},其计算公式如式(1),

$$r_{ij}=\frac{Cov(Rssi_i,Rssi_j)}{\sqrt{D(Rssi_i)\times D(Rssi_j)}}. \tag{1}$$

其中 $i,j\in V$,$Rssi_i$ 和 $Rssi_j$ 分别表示第 i,j 个 AP 的 RSSI 的观测值,$D(a)$ 表示随机变量 a 的方差,$Cov(a,b)$ 表示随机变量 a 和 b 的协方差.全部相关系数可以组成相关系数矩阵 $\boldsymbol{R}$.相关系数矩阵的计算过程使用全部的训练数据,并不区分训练样本所属的类,所以最后得出的相关系数是每对 AP 在全部测试区域内的相关系数.另外,计算一对 AP 的相关系数时,只考虑这两个 AP 能够被同时检测到的数据.

由式(1)可以看出,在计算相关系数矩阵时,可能遇到如下情况使得相关系数无定义:1)两个 AP 的可检测范围没有交集,即它们从来没有被同时检测到过;2)其中至少一个 AP 的 RSSI 的观测值的方差为 0.为了方便计算,当出现相关系数无定义时,定义该 AP 对的相关系数为 0.另外,对 $\forall i\in V$,定义 $r_{ii}=1$.

得到相关系数矩阵之后,就可以用式(2)计算出每一个 AP 的权值:

$$w_i=\frac{1}{\boldsymbol{R}_i\cdot\boldsymbol{R}_i^{\mathrm{T}}},i\in V. \tag{2}$$

式(2)中,w_i 是分配给第 i 个 AP 的权值,$\boldsymbol{R}_i=(r_{i1},r_{i2},\cdots,r_{in})$ 是相关系数矩阵 $\boldsymbol{R}$ 中的第 i 个行向量.简单来说,每个 AP 的权值是它与所有 V 中的 AP(包括它自己)的相关系数的平方和的倒数,所以一个 AP 与测试区域内能够检测到的所有 AP 的相关性(相关系数的绝对值)越高则它的权值越低.这是因为该 AP 与其它 AP 的相关性越高,就表示它所提供的信息与其它 AP 所提供的信息有更多的重复;反之,一个 AP 与其它 AP 越独立则它的权值越大,且对 $\forall i\in V$ 有 $w_i\in(0,1]$.

为了便于分析训练数据的特征,定义所有 AP 的权值之和为训练数据集的维度 D,如式(3)所示:

$$D=\boldsymbol{W}\cdot\boldsymbol{N}^{\mathrm{T}},\boldsymbol{N}=(1,1,\cdots,1). \tag{3}$$

其中 $\boldsymbol{W}=(w_1,w_2,\cdots,w_n)$ 为训练数据集的权值向量,$\boldsymbol{N}$ 是 n 维行向量且所有分量的值都为 1.显然当一个训练数据集中包含的 AP 之间相关性越高,则这个训练数据集的维度越小;反之,则训练数据集的维度越大,且对任意的训练数据集有维度 $D\in[1,n]$.以下是一些典型和极端场景中的 AP 权值分布和训练数据集的维度.

场景 1. 任意两个 AP 都完全相关(正相关或者负相关),此时对 $\forall i\in V$ 有 $w_i=1/n$ 和 $D=1$.说明当所有 AP 都完全相关时,训练样本空间只相当于一个一维的向量空间,同时所有 AP 都是平等的,所以权值都相等.

场景 2. 任意两个 AP 都完全独立,此时对 $\forall i\in V$ 有 $w_i=1$ 和 $D=n$.说明当所有 AP 之间都独立

时,每一个 AP 的 RSSI 观测值都可以被看做一个维度,而训练样本空间是一个 n 维的向量空间,同时所有 AP 都是平等的,所以权值都相等.

场景 3. 训练数据集中一共出现了 6 个 AP,其中 1 和 2 完全相关,3,4 和 5 两两完全相关,其它的 AP 对都完全无关. 则明显有 $w_1=w_2=1/2$, $w_3=w_4=w_5=1/3$, $w_6=1$ 和 $D=3$. 当 AP 中有很多紧密相关的集团时,训练样本空间的维度和集团的个数相同,且每个集团内部的 AP 平均分配这个维度的权值.

由场景 1 和场景 2 可以看出,当 AP 之间具有最强的相关性(两两完全相关)和最低的相关性(两两相互独立)时,R-kNN 算法都会退化为原始的 k 近邻算法. 这是因为,在这些场景下 AP 之间并没有统计上的区别,也就没有必要对它们的权值加以区分.

在得到每个 AP 的权值后,就可以简单地使用式(4)代替欧氏距离并使用 k 近邻算法对测试样本进行分类,如式(4)所示:

$$d(a,b)=\sqrt{\sum_{i\in V}w_i(s_{a,i}-s_{b,i})^2}, \tag{4}$$

其中,$d(a,b)$表示 a,b 两个样本之间的距离.

算法 1 为 R-kNN 算法的伪代码,该过程接受训练样本集、权值向量和待定位的测试样本作为参数,返回测试样本的预测位置.

算法 1. R-kNN.

输入:训练样本集 T,权值向量 $\boldsymbol{W}$,测试样本 a;

输出:分类结果 r.

符号定义:

k:最近邻居个数.

nil:空集.

$d(a,b)$:样本 a 和样本 b 之间的距离,即 $d(a,b)=\sqrt{\sum_{i\in V}w_i(s_{a,i}-s_{b,i})^2}$.

C:样本 a 可能属于的所有类的集合.

$class[a]$:样本 a 所在的类.

① $Q\leftarrow\varnothing$;
② $m\leftarrow$nil;
③ for all training sample $b\in T$ do
④ $d\leftarrow d(a,b)$;
⑤ if $|Q|<k$ then
⑥ $Q\leftarrow Q\cup\{b\}$;
⑦ $m\leftarrow\arg\max_{i\in Q}\{d(a,i)\}$;
⑧ else if $d<d(a,m)$ then
⑨ $Q\leftarrow(Q-\{m\})\cup\{b\}$;
⑩ $m\leftarrow\arg\max_{i\in Q}\{d(a,i)\}$;
⑪ end if
⑫ end for
⑬ $R\leftarrow\varnothing$;
⑭ for all $c\in C$ do
⑮ $w\leftarrow 0$;
⑯ for all $q\in Q$ do
⑰ if $class[q]=c$ then
⑱ $w\leftarrow w+1$;
⑲ end if
⑳ end for
㉑ $R\leftarrow R\cup\{<c,w>\}$;
㉒ end for
㉓ $r\leftarrow\arg\max_{i\in C}\{w|<i,w>\in R\}$;
㉔ return r.

3 实验分析和评价

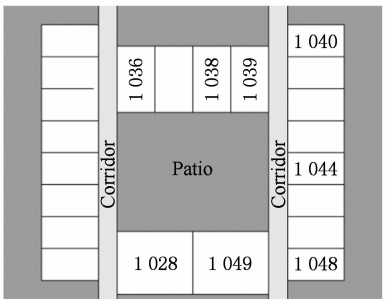
3.1 实验环境

实验在北航新主楼 G 座 10 层 8 个房间(位置分布如图 3 所示)与 11 层、9 层和 8 层的走廊上进行. 在实验环境中,总共检测到 165 个 AP,每个房间可以同时检测到的 AP 数量在 45~75 之间(具体为 1028 房间 74 个,1036 房间 61 个,1038 房间 69 个,1039 房间 67 个,1040 房间 48 个,1044 房间 53 个,1048 房间 47 个,1049 房间 75 个). 所有 AP 都是原先已经部署的,实验过程中未添加任何额外设备也未对任何 AP 的位置、软件等进行修改. 实验数据(包括训练数据和测试数据)均使用笔记本电脑 Thinkpad R400 的无线网卡测得. 在本实验中,对上下楼层的走廊采集数据的范围覆盖了对应 10 层全部实验房间的位置.

3.2 实验结果及分析

作者先后在该楼中进行了 30 d 的实地 AP 信号采集和测试(每天早中晚各 1 次,每次 2 h),总共在每个房间采集了 15 万~20 万组样本数据(包括所有可检测到的 AP 的 MAC 地址和 RSSI 值). 实验结果分别见图 4~7 所示.

图 4 所示为 k 近邻算法、R-kNN 算法和朴素贝叶斯分类器在不同房间和楼层中的定位准确率,即



The white rectangular blocks indicate rooms.
The number inside is room number.

Fig. 3 Room layout of the 10th floor.

图 3 10 层房间分布示意图

在特定位置上采集的测试样本被正确定位的比例.从总体上看,3 种定位算法都能够以较大概率(平均大于 80%)正确判定测试样本所处的房间,表明了基于 Wi-Fi RSSI 的位置指纹分类法用于房间级楼宇内定位具有较好的效果.从图 4 可以看出,贝叶斯分类器的分类效果在不同房间和楼层之间波动非常大,在很多位置的分类效果不是很理想,因此平均分类准确率不高.k 近邻算法也存在同样的问题.相比而言,R-kNN 算法在不同房间和楼层之间定位准确率比较稳定,能够稳定地保持最高或者接近最高的定位准确率.

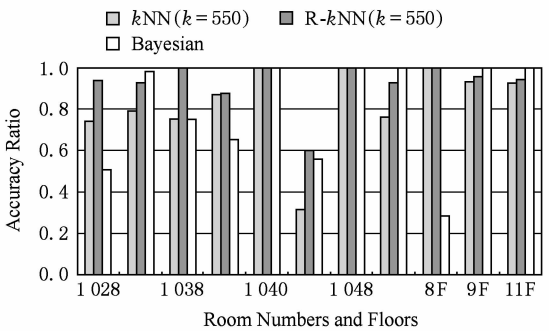


Fig. 4 Localization accuracies of different algorithms.

图 4 不同定位算法在不同房间和楼层的定位准确率

图 5 给出了 k 近邻算法、R-kNN 算法和朴素贝叶斯分类器的定位准确率(被正确定位的测试样本数与所有测试样本数的比值)相对于 k 的变化曲线.由于朴素贝叶斯分类器的定位准确率与 k 值无关,所以在图 5 以及图 6 和图 7 中表现为一条水平的直线.在测量样本点的过程中,周围环境的变化会使测量得到的样本出现一些随机性的扰动,训练样本中的扰动将使不同类的训练样本的交叉更加严重,进而使得基于 k 近邻的分类算法在 k 值较小时的分类准确率波动较大.从图 5 可以看出,当 k 值较小时两种算法的定位准确率没有明显的差距,因为 R-kNN 的加权策略并不是为了抵御随机扰动而设计的,所以此时与 k 近邻算法的区别并不明显.当 k 值继续增大时,相对于 k 近邻算法定位准确率迅速下降,R-kNN 算法的曲线更加平稳,这表明了 R-kNN 算法通过为 AP 分配合理的权值,使得同属于一类的样本更加紧密的聚集在一起,测试样本附近属于同一类的训练样本的密度更大,对 k 值的变化具有更强的鲁棒性.当 k 值大于某一阈值(700)时,这两种算法的定位准确率均稳定在某一水平上(R-kNN 定位准确率稳定在 91%左右,k 近邻定位准确率稳定在 84%左右).从整体上看,随着 k 值的变化 k 近邻算法的定位准确率在朴素贝叶斯分类器的两侧呈现波动状,而 R-kNN 算法可以保持高于前两种算法的定位准确率.

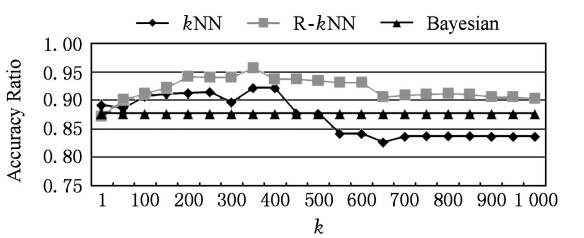


Fig. 5 Localization accuracies with k variation of kNN, R-kNN and Bayesian.

图 5 k 近邻算法、R-kNN 算法和朴素贝叶斯分类器定位准确率随 k 取值的变化

图 6 所示为 k 近邻算法、R-kNN 算法和朴素贝叶斯分类器的假阳性(false positive)比率(被误判为该类的测试样本占所有不属于该类的测试样本的比例)变化曲线,其中图 6(e),(k)中 k 近邻算法和 R-kNN 算法的曲线完全重合,图 6(f)中 R-kNN 算法和朴素贝叶斯分类器的曲线完全重合.各房间的假阴性比率(被误判为不属于该类的样本占所有实际属于该类的样本的比例)变化曲线如图 7 所示,其中图 7(d)中 k 近邻算法和 R-kNN 算法的曲线完全重合,图 7(e),(g)中 3 条曲线完全重合.

从整体上看假阴性比率较低意味着判定准确率较高,即很少有测试样本遭到误判,所以每个类的假阳性比率也应该较低.但是对某些分类算法而言,存在一种特殊的现象:当实际属于某个类的测试样本容易被正确判定为该类时,其它类的测试样本往往也很容易被误判为该类,使得低假阴性比率和高假阳性比率并存.这种现象被称为“分类黑洞”,因为这种同时拥有低假阴性比率和高假阳性比率的类会像黑洞一样吸引大量的测试样本,即诱使分类算法将

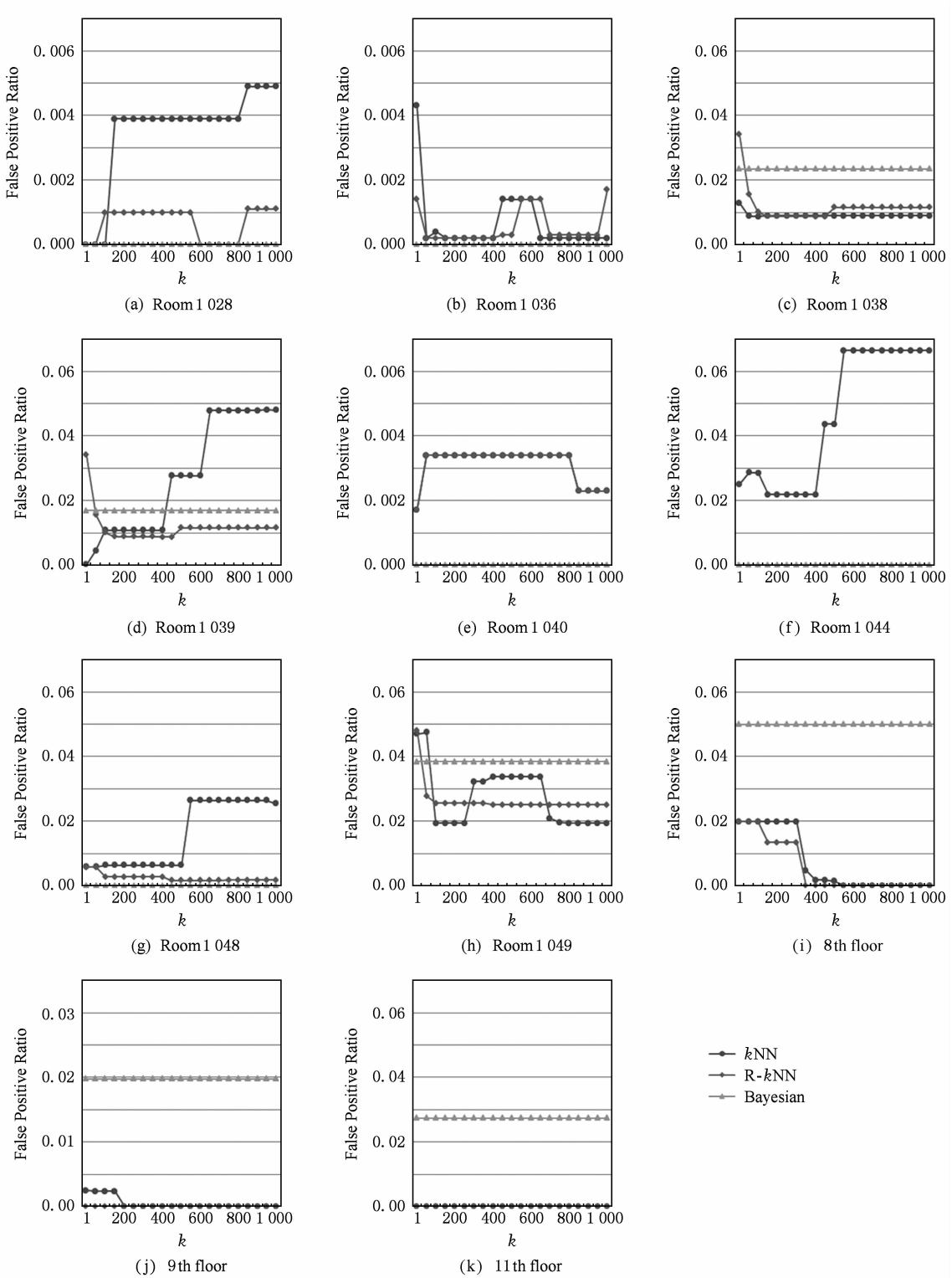


Fig. 6 False positive ratios with k variation of kNN, R-kNN and Bayesian in different rooms.

图 6 各房间中 k 近邻算法、R-kNN 算法和朴素贝叶斯分类器的假阳性比率随 k 取值的变化

大量样本判定为该类.

从图 6 和图 7 中的曲线可以看出,在除 1049 以外的房间和楼层中,3 种算法的定位效果有明显的区别. 虽然朴素贝叶斯分类器在很多位置上都具有较低的假阴性比率(较高的定位准确率),然而其在 9 层~11 层的假阳性比率却相对较高. 相比之下,k 近邻算法和 R-kNN 算法在这两个楼层中的假阴性比率和假阳性比率都较低,说明与朴素贝叶斯分类器相比,k 近邻算法和 R-kNN 算法在区分处于不同楼层的样本上具有更好的效果. 另一方面,在 10 层

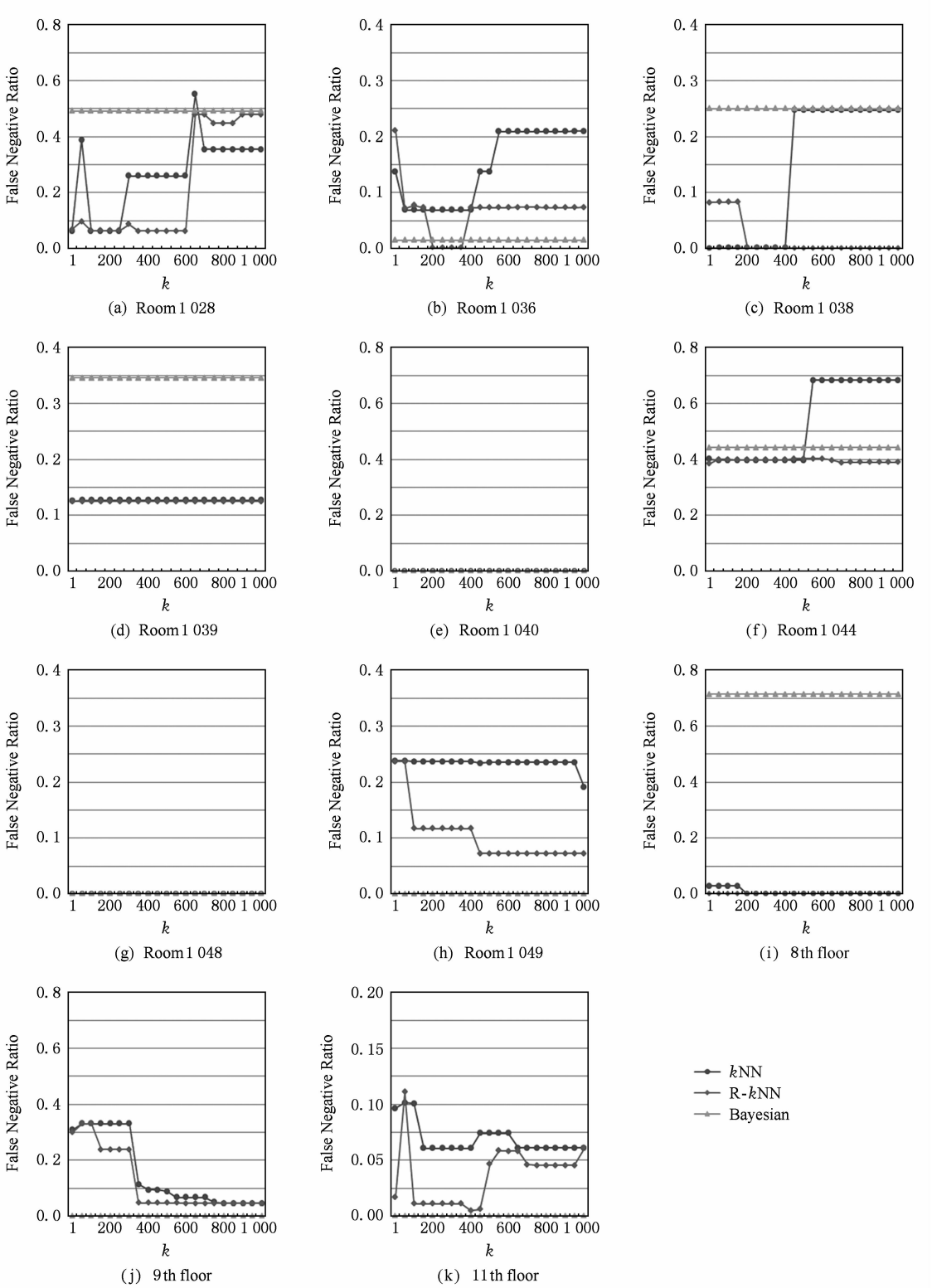


Fig. 7 False negative ratios with k variation of kNN, R-kNN and Bayesian in different rooms.

图 7 各房间中 k 近邻算法、R-kNN 算法和朴素贝叶斯分类器的假阴性比率随 k 取值的变化

的测试房间中,k 近邻算法在 1039 房间和 1048 房间中都出现了一定程度的“分类黑洞”现象,而 R-kNN 算法不但在这两个房间中的假阴性比率很低,而且其假阳性比率也同时较低.

从整体上看,即使是在 R-kNN 算法假阴性比率很低的几个位置(1028,1036,1038,1040,1048,8层,9 层,11 层等)中,R-kNN 算法的假阳性比率都能够被控制在较低的水平,而不会出现“分类黑洞”

现象.这说明即使定位准确率相同,R-kNN 算法的定位效果也比原始 k 近邻算法和朴素贝叶斯分类器更稳定.另外,随着 k 值的不断增大,R-kNN 算法相对于各类的假阳性比率呈现下降的趋势;相反,原始 k 近邻算法的假阳性比率却逐渐上升.图 6 中的曲线也同样说明了 R-kNN 算法能够增大特定样本周围与它同属一类的训练样本的密度,从而当 k 值较大时具有更好的抗随机扰动能力,能够比 k 近邻算法更加有效地排除随机扰动造成的误差.

为了分析训练数据集的规模与定位准确率之间的关系,本文分别使用不同规模的训练数据进行了定位实验,并得到了如图 8 所示的结果.

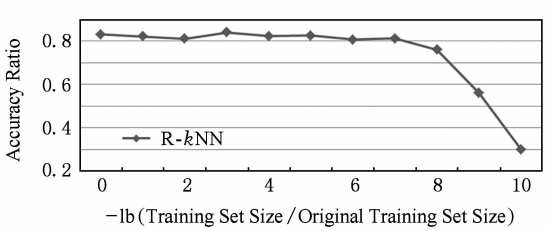


Fig. 8 Training set size's impact on localization accuracy of R-kNN.

图 8 训练数据集的规模对 R-kNN 算法定位准确率的影响

在该实验中,不同规模的训练数据集通过从原始数据集中随机挑选一定比例的训练样本构成,原始数据集中包含的样本数为 53 万,每个房间和楼层的训练样本数量基本相同.图 8 中,横坐标是训练样本比例对 2 的对数的相反数,同时图中每一个数据点为多组相同规模训练数据集实验结果的平均值.从图中可以看出,当样本比例大于 $2^{-8}\approx 0.4\%$,即训练样本数在 2 000 以上时,定位准确率对训练数据规模的变化并不敏感,当训练数据集规模继续减小时,定位准确率才有了明显的下降.该实验结果表明,R-kNN 算法在训练数据集规模相对较小时同样能够取得较高的定位准确率,但要求训练数据的采集过程只受到随机误差的影响,即数据采集点的空间分布要均匀,不同位置采集的样本数量要均衡.

4 结 论

本文提出了一种利用 Wi-Fi 接入点的 MAC 地址和 RSSI 值实现楼宇内房间级定位的算法 R-kNN. R-kNN 是一种属性加权 k 近邻算法,它通过将各个 AP 之间的相关性反应在权值的分配上,有效地降低了维度冗余对分类精度的负面影响.由于 R-kNN 没有对房间和 AP 的物理位置做出任何假设,所以只需要使用环境中现存的 AP 就可以取得较好的定位效果,而不需要部署任何额外设施或修改现有设施.实验结果表明,与原始 k 近邻算法和朴素贝叶斯分类器相比,R-kNN 算法可以在 AP 密度高的环境下取得较高且相对稳定的定位准确率;对各房间和楼层的假阳性比率和假阴性比率的分析结果表明,即使在平均定位准确率相近的情况下,R-kNN 算法的定位效果也比原始 k 近邻算法和朴素贝叶斯分类器工作更稳定.我们下一步工作将进一步研究 AP 之间相关性的计算方法,期望进一步提高 R-kNN 算法的定位准确率.

参 考 文 献

[1] Niculescu D, Nath B. Ad hoc positioning system (APS) using AOA [C] //Proc of the 22nd Annual Joint Conf of the IEEE Computer and Communications Societies. Piscataway, NJ: IEEE, 2003: 1734-1743

[2] Zhang Y W, Brown A K, Malik W Q, et al. High resolution 3-D angle of arrival determination for indoor UWB multipath propagation [J]. IEEE Trans on Wireless Communications, 2008, 7(8): 3047-3055

[3] Llombart M, Ciurana M, Barceló-Arroyo F. On the scalability of a novel WLAN positioning system based on time of arrival measurements [C] //Proc of the 5th Workshop on Positioning, Navigation and Communication. Piscataway, NJ: IEEE, 2008: 15-21

[4] Han D, Andersen D, Kaminsky M, et al. Access point localization using local signal strength gradient [C] //Passive and Active Network Measurement. Berlin: Springer, 2009: 99-108

[5] Bahl P, Padmanabhan V N. RADAR: An in-building RF-based user location and tracking system [C] //Proc of the 19th Annual Joint Conf of the IEEE Computer and Communications Societies (INFOCOM 2000). Piscataway, NJ: IEEE, 2000: 775-784

[6] Gümüskaya H, Hakkoymaz H. WiPoD wireless positioning system based on 802.11 WLAN infrastructure [J]. Enformatika, 2005, 8(9): 126-130

[7] Patel S, Truong K, Abowd G. PowerLine positioning: A practical sub-room-level indoor location system for domestic use [C] //Proc of UbiComp 2006. Berlin: Springer, 2006: 441-458

[8] Park J G, Charrow B, Curtis D, et al. Growing an organic indoor location system [C] //Proc of MobiSys'10. New York: ACM, 2010: 271-283

[9] Castro P, Chiu P, Kremenek T, et al. A probabilistic room location service for wireless networked environments [C] //Proc of UbiComp 2001. Berlin: Springer, 2001: 18-34

[10] Castro L A, Favela J. Continuous tracking of user location in WLANs using recurrent neural networks [C] //Proc of the 6th Mexican Int Conf on Computer Science. Piscataway, NJ: IEEE, 2005: 174-181

[11] Ito S, Kawaguchi N. Bayesian based location estimation system using wireless LAN [C] //Proc of the 3rd IEEE Int Conf on Pervasive Computing and Communications Workshops (PERCOMW'05). Piscataway, NJ: IEEE, 2005: 273-278

[12] Xiao Ling, Li Renfa, Luo Juan. A sensor localization algorithm in wireless sensor networks based on nonmetric multidimensional scaling [J]. Journal of Computer Research and Development, 2007, 44(3): 399-405 (in Chinese)
(肖玲, 李仁发, 罗娟. 基于非度量多维标度的无线传感器网络节点定位算法[J]. 计算机研究与发展, 2007, 44(3): 399-405)

[13] Hao Xiulan, Tao Xiaopeng, Xu Hexiang, et al. A strategy to class imbalance problem for *k*NN text classifier [J]. Journal of Computer Research and Development, 2009, 46(1): 52-61 (in Chinese)
(郝秀兰, 陶晓鹏, 徐和祥, 等. *k*NN 文本分类器类偏斜问题的一种处理对策[J]. 计算机研究与发展, 2009, 46(1): 52-61)

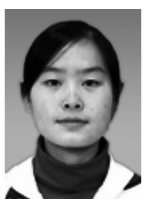
[14] Dudani S A. The distance-weighted *k*-nearest-neighbor rule [J]. IEEE Trans on Systems, Man and Cybernetics, 1976, 6(4): 325-327

[15] Bailey T, Jain A K. A note on distance-weighted *k*-nearest neighbor rules [J]. IEEE Trans on Systems, Man and Cybernetics, 1978, 8(4): 311-313

Niu Jianwei, born in 1969. PhD, professor of Beihang University. Senior member of China Computer Federation. His current research interests include embedded and mobile computing.

Liu Yang, born in 1987. PhD candidate in computer science. His current research interests include mobile and pervasive computing.

Lu Banghui, born in 1985. PhD candidate in computer science. His current research interests include opportunistic networks and wireless sensor networks.



Song Wenfang, born in 1988. Master candidate in computer science. Her current research interests include mobile and ubiquitous computing.