基于 WiFi 指纹的室内定位系统

王旭强，万亚平，李洪飞，耿家兴
（南华大学计算机学院，湖南 衡阳 421001）

摘　要：室内定位系统旨在以无线方式定位建筑物内的物体，并为室内定位移动应用带来发展。为了探索这种不成熟的系统设计，选择阿里第三方数据集，使用 PCA 进行特征选择，并分别建立基于 Gradient boosting、kNN 和 SVM 的预测模型。实验结果表明，kNN 和 Gradient Boosting 的组合为室内定位提供了高精度的预测。kNN 对于样本量大于 1000 的大量数据集表现出良好的性能，并且 Gradient Boosting 在小数据量上交叉验证错误很小。

关键词：室内定位；PCA 特征；Gradient boosting 方法；kNN 方法；SVM 方法
DOI:10.16184/j.cnki.comprg.2019.05.050

1　概述

随着互联网移动支付的迅速普及，人们享受到越来越多的生活便利。如走入商场的某家餐厅时，手机会自动弹出该餐厅的优惠券；当走入商场服装店时，手机可以自动推荐这家店里喜欢的衣服；在路过商场一家珠宝店时，手机可以自动提示了一款钻戒已经有货了；离开商场停车场时，手机在许可下可以自动交停车费。这些服务都离不开背后大数据挖掘和机器学习的支持。因此，在正确的时间、正确的地点给用户最有效的服务是当下室内定位的一个热点。

室内定位系统（IPS）旨在用磁传感器网络或其他数据源来定位建筑物内的物体或人。随着移动设备变得无处不在，应用程序的附近感知已成为开发人员优先考虑的方向。然而，大多数应用程序目前依赖于 GPS，并且在室内定位功能较差。到目前为止，IPS 系统设计还没有标准[1]。由于无线局域网（WLAN）和移动设备的激增，基于 WiFi 的 IPS 已经成为 IPS [2][3]的一种实用且有效的方法，不需要额外的设备成本。

2　相关工作

Mike Y. Chen，Timothy Sohn 等人探讨了数据大小和预测算法对位置预测精度的影响，并提出利用质心算法，有限大小的数据集可以提供高度可靠的结果[4]。Sunkywu Woo，Seongsu Jeong 等人为 WiFi 定位系统选择了指纹方法[5]。通过采用比较算法和 RFID 设备作为接收机，实现了 5m 内的定位精度。William Ching，Rue Jing Teh 等人使用 T-mobile G-1 手机进行了类似的结果，并建议通过用户贡献，换句话说，通过不断增加数据大小来提高预测准确度[6]。

从已有的研究成果看，虽然有大量对 WiFi 室内定位的研究文献，但这些文献所采用的通常为传统的单模型，如 KNN、朴素贝叶斯 、神经网络等经典方法。由于 WiFi 定位具有定位不准，会产生漂移，导致定位结果的不准确。因此， 将新的机器学习方法应用到 WiFi 定位具有十分重要的意义。

通过机器学习方法使用 WiFi 指纹定位移动设备的楼层，并探索数据大小、特征维度、模型组合和参数选择，以维持预测准确性，适用于不同的测试环境。

3　数据预处理

使用阿里的开源数据。 它由使用 16 种不同的手机型号检测到的 520 个 RSSI 指纹组成，有 18 位用户。这个数据集提供了数以千计的 RSSI 样本。因此，在给定的指纹路线图作为训练时产生在建筑物内的不同位置设置，可以使用机器学习生成模型技术，可以用它来预测未知移动设备持有人的楼层号码某个建筑物。

高维特征空间会损害计算效率。 因此，使用主成分分析（PCA）提取主要特征。 如图 1 所示，找出前 200 个主要特征。发现前三名特征很重要，并且超过 200 个特征以后的每个特征的重要性水平都小于 1。 从图 2 中可以看出，在特征大小和预测准确度之间存在

基金项目：湖南省研究生科研创新项目（CX2018B604）、南华大学研究生科学基金项目（2018KYZ015）。
作者简介：王旭强（1993-），男，硕士，研究方向：机器学习；万亚平（1973-），男，教授，CCF 会员；耿家兴，男，硕士；李洪飞，男，硕士。
收稿日期：2019-02-27

一定的关系。根据关于每个学习算法的特点，为合适的算法选择 5,50 和 200 个特征比较和探索最佳预测准确性。

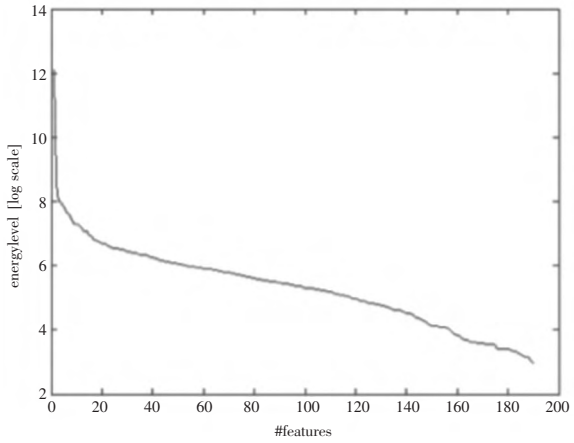

图 1　经过{}PCA 处理的特征重要性排序

#Feature\#Sample	500	5000
5	0.1193	0.0281
50	0.034	0.0017
200	0.02	0.0012

图 2　特征大小 VSKNN 的错误率

原始数据集有超过 5000 个样本，每栋建筑都需要付出很多努力为了收集数据在建立模型之前。因此，随机选择要探索的原始样本空间的子集数据大小对模型准确性的影响。在各种实验中探索了数据大小由 100，200，500，1000，2000 和 5000 个样本。

4　模型选择

在机器中实现了 3 种分类方法学习，包括 k–最近邻（kNN），梯度提升和支持向量机（SVM）。所有这 3 种方法都使用 10 倍交叉验证，以避免过拟合。

4.1　KNN

KNN 似乎是分类的好选择。这是因为 kNN 试图制造通过计算之间的距离进行分类功能，而各种 RSSI 信号的强度取决于关于 WiFi 源与网络之间的物理距离手机。在这种情况下，特征空间的接近度是物理空间密切的一个很好的迹象。

4.2　SVM

应用多类 SVM 来确定决策类之间的界限。但是，结果比决策树和 KNN 更差。一个潜在的原因 SVM 失败的原因是与无关的变量有关高维数据集。预测准确度高即使减少特征维度也很难实现从 520 到 200.要解决这个问题，进一步探索降维的努力 PCA。

4.3　梯度提升

在大幅度保持树木的大部分优势的同时提高准确性，包装算法是一个不错的选择。这里选择梯度增强提高准确性。另外，为了处理缺失的数据，使用代理来分发实例。最佳人数迭代（树的数量）使用交叉识别验证和每个简单树的深度设置为 4 岁。这个参数可以进一步研究得到一个更准确的。

5　实验与结果分析

分别使用 3 栋建筑的数据测试了模型。在每个建筑物中，在特征空间上执行 PCA 以减小其尺寸，并随机选择样本以在 4 个分类模型上执行 10 倍交叉验证。对于每个模型，样本空间中的减少维数和样本数都是可调的，以实现最小的错误。每组参数执行 5 轮，并且在这些轮次中平均误差以减少噪声。

5.1　KNN

首先探讨需要考虑的最近邻居的数量，以便对测试集进行分类。如图 3 所示，分类误差随 k 的增加而增加。然后研究 kNN（k = 1）分类中样本数量和主要特征数量的影响。随着样本量的增加，误差会急剧减少，但是通过添加更多主要特征可以看出没有太大的改进。

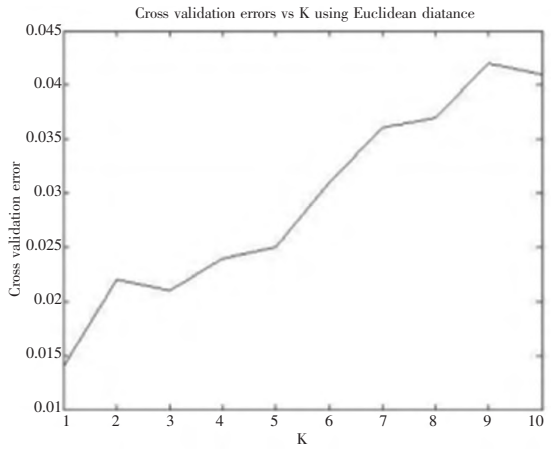

图 3　使用欧几里德距离的 kNN 方法的误差与 K.的关系

5.2　SVM

SVM 在此问题上表现不佳。在这里探索线性内核和三阶多项式内核，并决定使用多项式内核以获得更好的准确性。从图 4 中可以看出随着样本量的增加 SVM 误差的下降趋势，具有 150 个主要特征的数据比具有 50 个主要特征的数据表现更好。

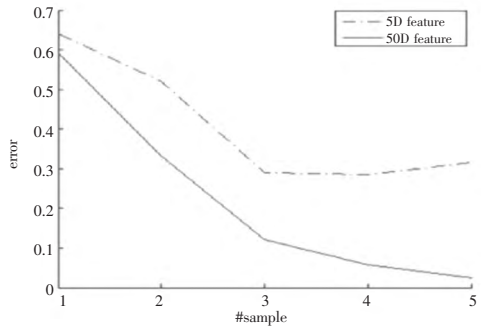

图 4　SVM 错误与样本大小。虚线表示使用 50 个特征的结果;实线表示使用 500 个特征的结果

没有用 5000 个样本测试实验，因为在给定足够的边际数据的情况下，SVM 往往对小数据集表现更好。因此，实验大的特征和样本空间将增加优化问题的收敛难度。

5.3　Gradient Boosting

基于训练数据拟合梯度增强（GB）模型。通过交叉验证，最佳迭代次数确定为 189.图 5 显示了错误分类错误风险与迭代次数的关系。对于测试集，错误分类错误计算为 0.05。图 6 显示了每个变量的相对重要性。图 7 显示了部分依赖性最重要的变量 V3。从这个数字可以看出楼层号码是强烈依赖在变量 V3 上，表示 V3 来自强大的信号源。图 8 显示了每层楼的整体错误率。发现 GB 错误率非常低，楼层 1 为 0，楼层 2 为 0.08 楼层 3 为 0.063 楼，楼层 4 为 0。

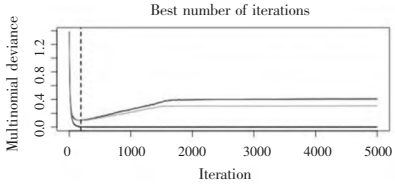

图 5　错误分类错误风险与梯度增强方法的迭代次数

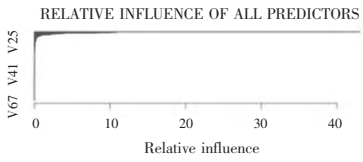

图 6　所有预测变量的相对影响

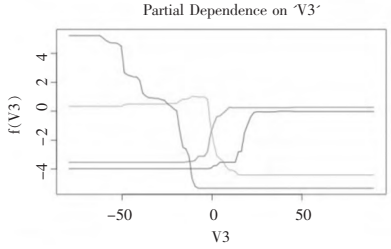

图 7　部分依赖于 V3

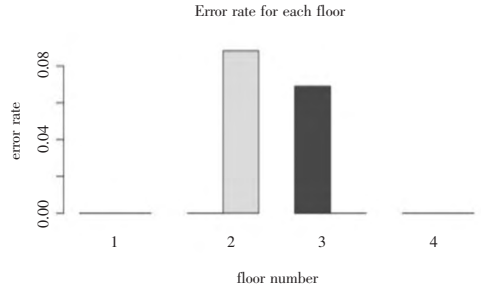

图 8　每层楼的错误率

6　结语

最简单的 KNN 模型在给定相对较小的特征空间和合理的大数据空间的情况下提供了良好的准确性。但是，SVM 在此分类算法上表现不佳。通过梯度增强对多个树进行装袋可以大大提高预测精度。为了获得高精度，同时保持预测小型和大型数据集的能力，建议将 KNN 和梯度增强结合用于室内定位系统。

参考文献

[1] Zhou, Junyi Shi, Jing. RFID localization algorithms and applications: a review. Journal of Intelligent Manufacturing, 20:, 695–707, 2009.

[2] Ferris, Brian Fox, Dieter Lawrence, Neil D. WiFiSLAM Using Gaussian Process Latent Variable Models. IJCAI, 7:, 2480–2485, 2007.

[3] Ma rques, Nelson Meneses, Filipe Moreira, Adria no. Combining similarity functions and majority rules for multi –building, multi –floor, WiFi positioning. IEEE Xplore, 2012.

[4] Chen, Mike Y Sohn, Timothy Chmelev, Dmitri Haehnel, Dirk Hightower, Jeffrey Hughes, Jeff LaMarca, Anthony Potter, Fred Smith, Ian Varshavsky, Alex/ Practical metropolitan –scale positioning for gsm phones. UbiComp 2006: Ubiquitous Computing, 225 – 242, 2006.

[5] Woo, Sunkyu Jeong, Seongsu Mok, Esmond Xia, Linyuan Choi, Chang su Pyeon, Muwook Heo, Joon. Application of WiFi–based indoor positioning system for labor tracking at construction sites: A case study in Guangzhou MTR. Automation in Construction, 20:, 3 – 13, 2011.

[6] Ching, William Teh, Rue Jing Li, Binghao Rizos, Chris. Uniwide WiFi based positioning system. Technology and Society (ISTAS), 2010 IEEE International Symposium on, 180–189, 2010.