湖北理工学院

毕 业 设 计（论文）
外 文 文 献 翻 译

题　　目： Indoor Positioning System Using Wifi Fingerprint

学　　院：计算机学院

专业名称：计算机科学与技术

学　　号：201440410230

学生姓名：陈自民

指导教师：刘军

2018 年 3 月 30 日

室内 Wifi 指纹定位系统

Dan Li, Le Wang, Shiqi Wu

Stanford University

室内定位系统旨在使用无线信号在建筑物内部定位物体，并且对于室内位置感知具有巨大的益处移动应用。为了探索这个不成熟的系统设计，我们选择 UJIndoorLoc 数据库作为我们的数据集，使用 PCA 特征选择以及基于决策树，梯度提升，KNN 和 SVM 建立预测模型。我们的实验结果表明，KNN 和 Gradient Boosting 的结合为室内提供了较高的预测精度定位。对于样本量大于 1000 的大量数据集和梯度提升，KNN 表现出良好的性能对于小数据量具有小的交叉验证错误，并且对丢失数据很有效。

Mike Y. Chen，Timothy Sohn 等人已经探索过数据大小和预测算法的影响对位置预测的准确性，并提出了建议用质心算法，有限的数据量集可以提供一个高度可靠的结果。SunkywuWoo，SeongsuJeong 等人选择了指纹 Wifi 定位系统的方法。通过采用比较算法，并采用 RFID 器件作为接收器，实现了定位精度在 5 米以内。WilliamChing，Rue Jing Teh 等人使用 T-mobile G-1 电话进行了类似的结果，并建议预测的准确性会是随着用户的贡献而改善，换句话说，通过不断增加数据量。华金 Torres-Sospedra，Raul Mntoliu 等人提出 UJIndoorLoc 数据库，用于公共公共数据库用于基于 WLAN 指纹的室内定位。受以前工作的启发，我们计划使用指纹的 Web 接入点（WAP）作为预测功能移动设备支架的位置。指纹的我们使用的 WAP 是接收信号强度指示器(RSSI)。在这个项目中，我们定位了楼层通过机器使用 Wifi 指纹的移动设备学习方法，并探索数据大小，功能尺寸，模型组合和参数选择如果不能提高预测准确度，针对不同的测试环境。

1 介绍

室内定位系统（IPS）以无线方式为目标在建筑物内找到物体或人物磁性传感器网络或其他数据源。室内定位的主要用户是在室内做一些具有收益的工作，如增强现实，商店导航，等等。随着移动设备变得无处不在，使用移动设备成为目前。目前大多数应用都依赖于 GPS，然而，GPS 在室内的定位功能功能不佳。至目前为止，IPS 系统设计没有事实上的标准。由于无线局域网的扩张网络和移动设备，基于 WiFi IPS 已经成为 IPS 的一种实用且有效的方法不需要额外的设施成本。然而，

基于 Wifi 的位置系统精度取决于指纹数据库中采集的位置指纹的数量。在定位过程中信号波动发生可能会增加路径中的错误和用户定位的不准确性。

2 方法

2.1 数据处理

我们在这个项目中使用 UJIndoorLoc 数据库。它由使用 16 个不同的检测到的 520 个 RSSI 指纹组成手机型号由 18 个用户从 4 个到 5 个不同 3 座建筑物的楼层，因此，以给定的指纹路线图作为训练设置，我们可以使用机器学习生成一个模型技术，我们可以预测未知移动设备持有人的楼层号码某个建筑物。

2.2 降维处理

然而，高维特征空间具有冗余功能会损害计算效率因此，我们使用主成分分析（PCA）提取主要特征。 能量显示了前 200 个主要组件的级别在图 2 中。 我们发现前三名校长组件包含大部分的能量，并为此超过前 200 个元件，每个元件的能量级别小于 1。 从图 3 可以看出，在预测准确度和准确度之间有一个折衷需要的功能数量，表示计算复杂性，为学习任务。根据关于每个学习算法的特点，我们为适合的算法选择 5,50 和 200 个顶级特征比较和探索最佳预测准确性

2.3 决策树

然后使用决策树进行实现，决策树有能够快速高效的去除许多不相关的变量，但是决策树也有一些缺点，例如相比于 KNN 算法而言决策树存在一些不精确的比较。多种的标准被应用在决策树的实施上面，通过不同的标准来使决策树中的数据属性进行分离，以此来达到建造决策树的目的。

3 结果和讨论

我们使用三座建筑物的数据测试了模型。如前所述，在每个建筑物，我们在特征空间上执行 PCA 以减少其特征空间维度和随机选择的样本来执行对四种分类进行十倍交叉验证样本数据。 减少室内的尺寸和指纹点数量样本，空间中的样本数量是可调的为每个模型实现最小的误差的数量。 将每次测量的数量取平均值得到最终的最小误差样本的数量

3.1 KNN

我们首先探索我们最近的邻居的数量需要考虑对测试集进行分类，分类的误差随之增加 k 的增加。然后我们看看来自的影响样本数量和最高主体数量特征在 kNN 的 k=1 分类中。错误随着样本量的增加而显着减少，但是增加更多样本数据并没有

多大改善特征。

3.2 SVN

我们应用多类 SVM 来确定决策类之间的边界。但是，结果是比决策树或 KNN 更糟。一个潜在的原因 SVM 失败的原因是因为无关变量高维数据集。预测准确度高即使我们减少特征维度也难以实现从 520 到 200.为了解决这个问题，我们进一步探索降维的 PCA。

3.3 决策树

决策树在后面被实现，决策树在快速处理数据方面有着较大的优势，它是很容易的去排除相关的变量，但是决策树也有计算结果不准确的特点，在对比使用 KNN 的算法过程中，在对决策树的处理方面，会由于对决策树的一些操作导致在一些数据集中数据的缺少。因此这种算法鼓励变量在高的数据集中。

3.4 降维处理

梯度提升模型基于训练数据。通过交叉验证，最佳数量迭代确定为 189.图 9 显示了错误分类错误风险与迭代次数。错误分类误差计算为 0.05 测试集。图 10 显示了每个的相对重要性变量和图 11 显示了部分依赖关系最重要的变量 V3。从这个数字我们可以看到底线数量强烈依赖在变量 V3 上,这表明 V3 来自 a 强大的信号源。图 12 显示了总体错误每层的价格。发现 GB 非常低错误率，1 楼为 0，2 楼为 0.08，0.063 楼和 0 楼 4 楼。

4.结论

正如本文所展示的那样，最简单的 kNN 模型给出了很好的精度，给出了相对较小的特征空间和合理的大数据空间。 然而，SVM 在这种分类算法上表现不佳。虽然一个决策树没有给出令人满意的结果，通过梯度提升对多棵树进行剪枝可以大大提高预测的准确性。我们建议如果想获得较高的精确度就要保持样本容量，使用小型和大型数据集 kNN 和梯度提升的组合算法来用于室内定位系统。

5.未来展望

由于梯度增强是缺失值的稳健性，因此当前 kNN 模型缺失值的影响是调查。此外，超越目前的模式，追踪移动用户，电话类型和最小号码的 Wifi 资源需要准确的定位在未来的工作中探索。

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] Marques, Nelson Meneses, Filipe Moreira, Adriano. 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, Changsu 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.

[7] Torres-Sospedra, Joaquın Montoliu,Raúl Martınez-Usó, Adolfo Avariento, Joan P Arnau,Tomás J Benedito-Bordonau, Mauri Huerta,Joaquın. UJIIndoorLoc: A New Multi-building and Multi-floor Database for WLAN FingerprintbasedIndoor Localization Problems.

Indoor Positioning System Using Wifi Fingerprint

Indoor Positioning System aims at locating objects inside buildings wirelessly, and have huge benefit for indoor location-aware mobile application. To explore this immature system design, we choose UJIndoorLoc database as our data set, use PCA for feature selection, and build prediction models based on decision tree, gradient boosting, kNN and SVM, respectively. Our experiment results indicate that combination of kNN and Gradient Boosting provides high prediction accuracy for Indoor Positioning. kNN shows good performance for large volume of data set with sample size greater 1000, and Gradient Boosting has small cross validation error for small data volume and is robust to missing data.

1.Introduction

Indoor Positioning System (IPS) aims at wirelessly locating objects or people inside buildings based on magnetic sensor network, or other source of data. The major consumer benefit of indoor positioning is the expansion of location-aware mobile computing indoors, such as augmented reality, store navigation, etc. As mobile devices become ubiquitous, contextual awareness for applications has become a priority for developers. Most applications currently rely on GPS, however, and function poorly indoors. Up till now, there is no de facto standards for IPS system design [1]. Due to the proliferation of both wireless local area networks (WLANs) and mobile devices, WiFi-based IPS has become a practical and valid approach for IPS [2][3] that does not require extra facility cost. However, Wifi-based position system as (WPS) accuracy depends on the number of positions that have been entered into the database. The possible signal fluctuations that may occur can increase errors and inaccuracies in the path of the user.

Mike Y. Chen, Timothy Sohn, et al have explored the influence of data size and prediction algorithm on location predicting accuracy, and has proposed that with centroid algorithm, a limited size of data set can provide provide a highly reliable result[4]. Sunkywu Woo, Seongsu Jeong, et al

have chosen fingerprint methods for Wifi positioning system[5]. By adapting comparison algorithm and using RFID device as receiver, they achieved locating accuracy of within 5m. William Ching, Rue Jing Teh, et al have conducted similar result using T-mobile G-1 phone, and suggested that the predicting accuracy would be improved with the user contribution, in other words, by constantly increasing the data size[6]. Joaquin Torres-Sospedra, Raul Mntoliu, et al, have proposed UJIndoorLoc database for a common public database for WLAN fingerprint-based indoor localization[7].

Inspired by previous work, we plan to use fingerprint of Web Access Points(WAPs) as features to predict the position of mobile device holder. The fingerprint of WAP we use is the Received Signal Strength Indicator(RSSI). In this project, we locate the floor level of a mobile device using Wifi fingerprint via machine learning methods, and explore the data size, feature dimension, model combination and parameter selection to maintain, if not improve, prediction accuracy, for different test environment.

2. Methods

2.1 Data preprocessing

We use UJIndoorLoc database [7] for this project. It consists of 520 RSSI fingerprints detected using 16 different phone models by 18 users from 4 to 5 different floors of 3 buildings, as shown in figure 1. For each building, this dataset gives us thousands of RSSI samples generated at various locations inside the building. 1 Thus, with given roadmap of fingerprints as training set, we could generate a model using machine learning techniques, with which we would be able to predict the floor number of unknown mobile device holder in a certain building.

2.2 dimensionality reduction

However, the high dimensional feature space with redundant features would hurt computational effi- ciency. Therefore, we used Principal Component

Analysis(PCA) to extract principal features. The energy levels of the first 200 principal components are shown in figure 2 . We found that the top three principal components contain most of the energy, and for the components beyond the first 200, each of their energy levels is less than one. As we can see from figure 3, there is a tradeoff between prediction accuracy and number of features required, which indicates the computation complexity, for the learning task. Depending on the characteristics of each learning algorithm, we choose 5, 50 and 200 top features for the suitable algorithms to compare and explore for the best prediction accuracy

2.3 Decision Tree

Decision tree is then implemented, which has the advantages of fast training process, easy interpretation and resistance to many irrelevant variables. But decision tree has the disadvantage of inaccuracy compared with kNN, even cross validation is used. In decision tree, two criteria are applied to prune the tree. One is cross validation and the other is one stand error. Surrogate splits are used in construction of the optimal tree as a missing value strategy, which encourages variables within highly correlated sets.

3 Results & Discussion

We tested our models using data from three buildings separately. As mentioned before, at each building, we performed a PCA on the feature space to reduce its dimension, and randomly selected samples to perform a ten-fold cross validation on the four classification models. Both of the number of reduce dimension and the number of samples in the sample space are tunable for each model to achieve the least error. Each set of parameters was performed for five rounds and the error is averaged among those rounds to reduce noise.

3.1 KNN

We first explore the number of nearest neighbors we need to consider

for classifying a testing set. As shown in figure 4, the error of classification increases with increasing of k. Then we look into the influence from the number of samples and the number of top principal features in kNN (k=1) classification. The error reduces dramatically with increase of sample size, but not much improvement is seen from adding more principal features.

3.2 SVM

SVM does not perform well for this problem. Here we explore both linear kernel and third order polynomial kernel, and decided to use polynomial kernel for better accuracy. From figure 5, we can see the descending trends of SVM error with increasing of sample size, and data with 150 principal features perform better than data with 50 principal features.

3.3 Decision Tree

Decision tree is then implemented, which has the advantages of fast training process, easy interpretation and resistance to many irrelevant variables. But decision tree has the disadvantage of inaccuracy compared with kNN, even cross validation is used. In decision tree, two criteria are applied to prune the tree. One is cross validation and the other is one stand error. Surrogate splits are used in construction of the optimal tree as a missing value strategy, which encourages variables within highly correlated sets.

3.4 Gradient Boosting

To maintain most advantages of trees while dramatically improve accuracy, bagging algorithm could be a good choice. Here gradient boosting is chosen to improve the accuracy. Also, to handle missing data, we use surrogates to distribute instances. Best number of iterations (number of trees) are identified using cross validation and the depth for each simple tree is set to be four. This parameter could be further studied get a better accuracy.

4. Conclusion

As demonstrated in this paper, the simplest kNN model gives good accuracy, given a relative small feature space and reasonable large data space. However, SVM performs poorly on this classification algorithm. Although one decision tree does not give satisfying result, bagging of multiple trees through gradient boosting could highly increase the prediction accuracy. To acquire high accuracy, while maintaining the capacity for predicting both small and large data set, we suggest the combination of kNN and gradient boosting for the indoor positioning system.

5. Future Work

Since Gradient boosting is robust of missing value, the effect of missing value for current kNN model is to be investigated. Also, beyond the current models, tracking of moving user, type of phones and minimum number of Wifi sources required for accurate positioning will be explored in the future work.

References

[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] Marques, Nelson Meneses, Filipe Moreira, Adriano. 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, Changsu 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.

[7] Torres-Sospedra, Joaquın Montoliu, Raúl Martınez-Usó, Adolfo Avariento, Joan P Arnau, Tomás J Benedito-Bordonau, Mauri Huerta, Joaquın. UJIIndoorLoc: A New Multi-building and Multi-floor Database for WLAN Fingerprintbased Indoor Localization Problems.