Centro de Informática – UFPE
Aluno: Douglas Tavares Ribeiro Paulino Silva - dtrps@cin.ufpe.br
Orientador: Daniel Carvalho da Cunha - dcunha@cin.ufpe.br

Localização de Estações móveis: Técnica de localização *Fingerprinting* Utilizando *Machine Learning* LightGBM

Áreas: Comunicações Móveis e Aprendizagem de máquina

1.Resumo

Ao longo dos últimos anos tem observado-se um grande número de aplicações IoT e serviços para dispositivos móveis que utilizam o sistema de localização tanto *outdoor* quanto *indoor*, porém o que tem ganhado grande destaque são os serviços de localização em ambientes *indoor* que apresentam inovação para o marketing, segurança, saúde pública entre outros. Uma das tecnologias bastante conhecida é o *Global Positioning System* (GPS), porém apresenta uma precisão baixa e acurácia média não sendo ideal para localizar o usuário em ambientes *indoor* e nem para aplicações IoT pelo alto consumo de bateria. Neste trabalho é proposto uma nova solução que utiliza uma técnica de localização conhecida como *fingerprinting* que irá utilizar *Machine Learning* (ML) e compara os resultados com algum artigo da literatura. O algoritmo de ML proposto é o LightGBM que na literatura propõe um tempo de treinamento e consumo de memória menor em relação aos outros algoritmos de ML mantendo a mesma acurácia, além disso este algoritmo é ideal para treinar com um grande número de instâncias de dados e um grande número de *features*.

Palavras-chave: Sistemas de Localização *Outdoor* e *Indoor*, *Fingerprinting* utilizando *Machine Learning*, Redes móveis, LightGBM.

2. Introdução

Ao longo dos anos, a localização de usuários em redes móveis e sem fio tornou-se muito importante para a inclusão digital. Alguns aplicativos de *smartphones* precisam da localização do usuário para melhor atendimento e qualidade do serviço prestado. Governos dos países também utilizam a localização dos *smartphones* dos seus cidadãos para informá-los sobre tragédias ou conscientizá-los com alertas em carros de som. Foi o caso durante a pandemia do novo coronavírus (Sars-CoV-2), em que o Governo do Brasil, fez uma parceria com as operadoras de telefonia para monitorar como estavam as medidas de isolamento no país [1]. Porém, esse monitoramento requer uma alta precisão e acurácia para identificar se as pessoas estavam realmente dentro de casa, o que não se torna possível saber porque o erro médio dos dados de geolocalização das operadoras assume valores no intervalo de 1.3 Km [2] a 32 Km para regiões mais afastadas das Estações Rádio Bases (ERBs), de acordo com a *Federal Communications Commission* (FCC) [3]. Assim, distâncias menores que 1.3 Km não seriam detectados e apenas deslocamentos de raio de 32 Km seriam altamente confiáveis.

O serviço de localização mais conhecido atualmente é o *Global Positioning System* (GPS), muito utilizado por aplicativos como *Google Maps* [4], *Waze* [5] e *Foursquare* [6]. Porém, o GPS é mais recomendado para ambientes *outdoor*, pois a acurácia é de 30 à 60 metros segundo os principais provedores de sistemas operacionais para *smartphones* [7] [8]. Já no ambiente *indoor* o GPS não é adequado, pois apresenta uma precisão baixa, acurácia média e frequência baixa. Além disso existem interferências no sinal de materiais como: tijolos, madeiras, vidros e prédios próximos. A interferência causa ruídos no sinal de rádio emitido e recibo pela rede de celular, o que dificulta a localização do usuário do ponto exato de onde ele está.

Outro ponto observado é sobre as redes IoT que são caracterizadas por um grande número de objetos. Uma vez que aplicações IoT exigem uma coleta de dados com alta periodicidade, o consumo de bateria do GPS aumentaria consideravelmente, inviabilizando o seu uso para aplicações IoT tanto *outdoor* quanto *indoor* [9].

Visando melhorar a acurácia para ambientes *indoor*, algumas tecnologias estão sendo exploradas como A-GPS, sensores de *smartphones*, Bluetooth, Ultra-wideband, ZigBee, GSM, *high sensitivity* GNSS e até mesmo Wi-Fi. Os especialistas buscam a melhor combinação dessas tecnologias para entregar boas acurácias com margem de erro de poucos metros [10] e identificar o andar onde encontra-se o usuário em caso de prédios.

Enquanto alguns especialistas investem em tecnologias para rastrear os usuários em ambientes *indoor*, outros investem em técnicas de radiolocalização tradicionais. A radiolocalização é uma técnica de localização que faz uso de parâmetros presentes nos sinais de rádio para prover a localização do usuário. Esta técnica pode utilizar diferentes parâmetros, tais como: *Time of arrival* (ToA), *Angle of Arrival* (AoA), *Received Signal Strength Indicator* (RSSI) e *Time difference of Arrival* (TDoA). Várias propostas podem ser encontradas na literatura, mas essas técnicas tendem a ser imprecisas, principalmente devido à propagação de *multipath* e *non-line-of-sight* (NLoS) [11].

Um problema que pode ser encontrado na radiolocalização é que o valor absoluto do RSSI pode variar de acordo com o fabricante do *smartphone* e do modelo de medição. Essa variação pode causar problemas na acurácia das técnicas de localização baseadas em *fingerprinting* [10].

Algumas técnicas de radiolocalização que destaca-se são: o *fingerprinting* que é um vetor com várias *features* como: ToA, AoA, RSSI e TDoA que são usadas para correspondência de padrões para predizer a posição do usuário [11]. E a trilateração que estima a distância aproximada da estação móvel para cada uma das 3 ERBs mais próximas. As posições das ERBs são os pontos de referência, então é possível calcular a interseção dos raios das ERBs com a estação móvel. Em seguida, montar um sistema de equações e determinar a posição geográfica aproximada do usuário [12].

Como temos várias *features* disponíveis na radiolocalização, algumas propostas encontradas na literatura modelam o problema de Localização dos dispositivos móveis dos usuários como um problema de *Machine Learning* (ML) [13]. ML é um subconjunto da Artifical Intelligence (AI), a principal definição é que ML é orientada a dados, os algoritmos são modelos matemáticos baseados nestes dados. A amostra de dados é conhecida como dados de treinamento que são utilizados para treinar o algoritmo e predizer os resultados esperados. Neste trabalho será utilizado a aprendizagem supervisionada, ou seja, o algoritmo precisa dos dados de entrada e dos dados de saída na fase de treinamento para que possa aprender e predizer novas entradas.

Ao trabalhar com radiolocalização utilizando ML, um problema surge para coletar as amostras das ondas de rádio emitidas por dispositivos móveis em ambientes *indoor*. Nesses ambientes o acesso para a coleta de dados na maioria das vezes é privado, pois tratam-se de casas, fazendas e empresas que restringem o acesso impossibilitando a medição com o *drive test*. Apesar das restrições de acesso para os técnicos realizarem as medições, os usuários daquele ambiente ainda querem os serviços de localização com qualidade, ou seja, com precisão e acurácia boas. Por isto neste trabalho é proposto o treinamento com dados *outdoor* para prever a localização do usuário tanto em ambientes *outdoor* quanto *indoor*.

Neste trabalho, é proposto utilizar um algoritmo de ML conhecido como LightGBM que será responsável por predizer o *path loss* nos ambientes *outdoor* e *indoor* e com esta predição será possível localizar no *grid* do *Fingerprinting* a posição do usuário. O algoritmo LightGBM é um *Gradient Boosting Decision Tree* (GBDT) altamente eficiente que utiliza duas novas técnicas: *Gradient-based One-Side Sampling* (GOSS) e *Exclusive Feature Bundling* (EFB). As duas técnicas GOSS e EFB lidam com um grande número de instâncias de dados e com um grande número de *features* respectivamente. O artigo do LightGBM mostra que o algoritmo consegue superar outras técnicas de ML em termos de velocidade computacional e consumo de memória, mantendo a mesma acurácia das outras técnicas de ML para o mesmo conjunto de dados [14].

2. Objetivo

O objetivo geral deste trabalho é comparar o *baseline* que será a técnica de localização *Fingerprinting* (FP) utilizando *Machine Learning* (ML) como em [11] ou [16] com o *Fingerprinting* utilizando o LightGBM. O objetivo é treinar as duas técnicas apresentadas acima com apenas dados *outdoor* para prever a localização do usuário tanto em ambientes *outdoor* e *indoor*. Ao final será comparado o tempo de treinamento, consumo de memória e verificar se a acurácia da técnica de localização proposta continua igual ou melhor do que o *baseline*.

Sendo assim para alcançar esse objetivo é necessário alguns objetivos específicos para ambos algoritmos como a divisão dos dados em duas fases: uma de treinamento (*off-line)* e outra de predição (*on-line)*. A coleta das amostras de dados consiste na utilização de uma base de dados já construída no artigo [16] para que possam ser realizados os experimentos. Na fase de predição (*on-line*) serão analisados 3 cenários: apenas dados *outdoor*, dados *outdoor-indoor* e apenas dados *indoor*.

O segundo e o terceiro cenários devem ser analisados com mais cuidado, pois podem apresentar ruídos devido às interferências de árvores e prédios próximos. O objetivo específico é analisar como o treinamento apenas com dados *outdoor* afetou a acurácia em cada um desses cenários e qual cenário apresentou os melhores resultados. Outro caso é se o LightGBM consiga manter a acurácia, diminuir o tempo de treinamento e consumo de memória em relação ao *baseline*. Será analisado quais os benefícios que o *fingerprinting* utilizando o LightGBM traz para as aplicações que utilizam a técnica de localização.

3. Metodologia

Será realizado um estudo inicial para entender todos conceitos relacionados a este trabalho, tais como: as tecnologias que ajudam na localização do usuário, as técnicas de localização existentes na literatura das mais simples às que tratam este problema como um problema de *Machine Learning* (ML) e verificar as aplicações que dependem da localização com eficiência como aplicações IoT. Após a leitura dos artigos da literatura sobre localização de dispositivos móveis e as principais técnicas de localização utilizadas, o algoritmo LightGBM será estudado e implementado, o que permitirá o início dos experimentos.

Para realizar os experimentos deste trabalho, simulações computacionais serão realizadas por meio da linguagem Python 3. A linguagem de programação Python 3 tende a ser melhor para trabalhar com problemas de ML devido à disponibilidade de várias bibliotecas que facilitam tanto a análise dos dados, quanto a implementação dos algoritmos. Serão utilizadas as bibliotecas de Python para *Machine Learning* como Scikit, Pandas, Numpy e Seaborn, além de uma biblioteca auxiliar de cálculo de distâncias geográficas: PyRadioLoc [15].

A primeira etapa da experimentação consiste na utilização de uma base de dados já construída no artigo [16]. No artigo citado, com um *drive-test* é coletado uma amostra que compõe um coordenada geográfica (latitude e longitude), um vetor de RSSI e um conjunto de medições de *propagation delay* (PD) [16]. Essa base de dados contém dados *outdoor* e *indoor* que servirá para realizar o treinamento dos algoritmos e os experimentos futuros.

Com estas amostras adquiridas usaremos apenas dos dados *outdoor* que serviram para o treinamento (*off-line*) para o *baseline* e a nova técnica proposta. Na etapa seguinte é a fase de predição (*on-line)* para medir a acurácia das duas técnicas. Serão realizados testes com 3 cenários diferentes: o primeiro cenário apenas com dados *outdoor*, o segundo cenário com dados *outdoor* e *indoor* e por último um cenário apenas com dados *indoor*. O objetivo desta divisão de 3 cenários é analisar qual foi o impacto na acurácia de treinar ambos os algoritmos com dados *outdoor* e compará-los. No segundo e terceiro cenário, poderá ser observado um comportamento diferente na técnica de localização que seriam os *outliers*. Caso observe-se que a acurácia no segundo e terceiro

cenário foi bastante afetada negativamente em relação ao primeiro cenário pode ser verificado formas de amenizar esse efeito causado pelos *outliers*.

Referências

[1] Governo vai usar dados de operadoras para monitorar aglomeração na pandemia, https://www1.folha.uol.com.br/mercado/2020/04/governo-vai-usar-dados-de-operadoras-para-monitorar-deslocamentos-na-pandemia.shtml, Acesso em: 02 de abril de 2020.

[2] Accurate Location Detection, https://transition.fcc.gov/pshs/911/Apps%20Wrkshp%20215/911_Help_SMS_WhitePaper0515.pdf Acesso em: 02 de abril de 2020.

[3] Cell tower data may be misleading, https://www.heraldcourier.com/opinion/cell-tower-data-may-be-misleading/article_f5e63928-7c76-5df2-8d38-b36388154a83.html, Acesso em: 02 de abril de 2020.

[4] Google Maps, https://www.google.com.br/maps, Acesso em: 29 de março de 2020.

[5] Waze, https://www.waze.com/pt-BR/, Acesso em: 29 de março de 2020.

[6] FourSquare, https://pt.foursquare.com/, Acesso em: 29 de março de 2020.

[7] Precisão, acurácia, frequência e recência, https://www.meioemensagem.com.br/home/opiniao/2018/11/22/precisao-acuracia-frequencia-e-recencia.html, Acesso: 04/04/2020

[8] GPS Accuracy, https://www.gps.gov/systems/gps/performance/accuracy/, Acesso: 04/04/2020

[9] GOLDONI, E, et al. Experimental data set analysis of RSSI-based indoor and outdoor localization in LoRa networks. Internet Technology Letters, [S.1], v.2, n.1, p.e 75, 2019 DOI: https://doi.org/10.1002/itl2.75

[10] LUI, G. et al. Differences in RSSI readings made by different Wi-Fi chipsets: a limitation of wlan localization. In: INT. CONFERENCE ON LOCALIZATION AND GNSS (ICL - GNSS 2011). Proceedings… [S.1.: s.n.], 2011. p.53-57.

[11] TIMOTEO, Robson D.A.. SILVA, Lizandro N.. CUNHA, Daniel C.. CAVALCANTI, George D. C.. An approach using support vector regression for mobile location in cellular networks. Computer Networks 95, 2016.

[12] Savvides, A., Han, C. C., and Srivastava, M. B. (2001). Dynamic fine grained localization in ad-hoc sensor networks. In Proceedings of ACM Mobile Communications (MobiCom).

[13] OLIVEIRA, L. L.. OLIVEIRA JR, L. A.. SILVA, G. W. A.. TIMOTEO, R. D. A.. CUNHA, D. C.. An RSS-based regression model for user equipment location in cellular networks using machine learning. Springer Science+Business Media, LLC, part of Springer Nature 2018. https://DOI.org/10.1007/s11276-018-1774-4

[14] KE, GUOLIN et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. 31st Conference on Neural Information Processing Systems (NIPS 2017) Pages 3149–3157, Long Beach, CA, USA.

[15] Biblioteca para o cálculo de distâncias geográficas: PyRadioLoc https://github.com/timotrob /pyRadioLoc, Acesso: 09/04/2020

[16] TIMOTEO, R. D.A.; CUNHA, D. C., A scalable fingerprint-based angle-of-arrival machine learning approach for cellular mobile radio localization, Computer Communications 157 (2020) 92–101. https://doi.org/10.1016/j.comcom.2020.04.014