CLÁUDIO GUSTAVO SANTOS CAPANEMA

DETECÇÃO DE PONTOS DE INTERESSE E PREDIÇÃO DE PRÓXIMO LOCAL DE VISITA DE USUÁRIOS MÓVEIS COM BASE EM DADOS ESPARSOS

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Ciência da Computação, para obtenção do título de *Magister Scientiae*.

Orientador: Fabrício Aguiar Silva

VIÇOSA - MINAS GERAIS
2020

Ficha catalográfica preparada pela Biblioteca Central da Universidade Federal de Viçosa - Câmpus Viçosa

T
C236d
2020

Capanema, Cláudio Gustavo Santos, 1994-
Detecção de pontos de interesse e predição de próximo local de visita de usuários móveis com base em dados esparsos / Cláudio Gustavo Santos Capanema. – Viçosa, MG, 2020.
48 f. : il. (algumas color.) ; 29 cm.

Orientador: Fabrício Aguiar Silva.
Dissertação (mestrado) - Universidade Federal de Viçosa.
Referências bibliográficas: f. 46-48.

1. Computação ubíqua. 2. Redes neurais (Computação). 3. Mineração de dados (Computação). I. Universidade Federal de Viçosa. Departamento de Informática. Programa de Pós-Graduação em Ciência da Computação. II. Título.

CDD 22. ed. 004

CLÁUDIO GUSTAVO SANTOS CAPANEMA

DETECÇÃO DE PONTOS DE INTERESSE E PREDIÇÃO DE PRÓXIMO LOCAL DE VISITA DE USUÁRIOS MÓVEIS COM BASE EM DADOS ESPARSOS

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Ciência da Computação, para obtenção do título de *Magister Scientiae*.

APROVADA: 20 de março de 2020.

Assentimento:

Cláudio Gustavo Santos Capanema
Cláudio Gustavo Santos Capanema
Autor

Fabrício Aguiar Silva
Fabrício Aguiar Silva
Orientador

A Deus, pela graça da vida, pelas bênçãos dadas e pelos ensinamentos. A minha família pelo apoio aos meus estudos, e em especial a minha mãe Alessandra da Fonseca Santos Capanema pelas sábias orientações e por estar ao lado nos momentos mais difíceis.

Agradecimentos

A Jesus pela graça dada à minha vida e carreira acadêmica.

À minha família pelo apoio e compreensão, e em especial a minha mãe Alessandra da Fonseca Santos Capanema.

Ao orientador Fabrício Aguia Silva pelo conhecimento compartilhado, atenção, e sábios conselhos durante todo o curso.

Ao apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

Às amizades construídas durante o curso.

A todas as pessoas que de alguma forma me ajudaram nessa caminhada.

E por fim, a todos os colaboradores da Universidade Federal de Viçosa, em especial aos professores do Departamento de Informática.

Pois tu és a minha esperança,
ó Soberano Senhor,
em ti está a minha confiança desde a juventude.
(Salmos 71:5)

Resumo

CAPANEMA, Cláudio Gustavo Santos, M.Sc., Universidade Federal de Viçosa, março de 2020. **Detecção de Pontos de Interesse e Predição de Próximo Local de Visita de Usuários Móveis com Base em Dados Esparsos**. Orientador: Fabrício Aguiar Silva.

Dados de localização provenientes de dispositivos móveis são importantes para o estudo da mobilidade humana. Ao se conhecer melhor seus usuários, provedores de serviços móveis têm o interesse em aprimorar os seus produtos e aumentar o engajamento de clientes. A maioria das soluções presentes na literatura foram desenvolvidas com base em dados de localização que foram coletados intensivamente, o que leva a uma alta de demanda por recursos energéticos, de armazenamento, processamento e de rede nos dispositivos móveis. Por outro lado, os dados esparsos, mesmo que mais limitados, podem ser gerados por um grande número de usuários sem afetar a autonomia energética de dispositivos móveis. Assim, explorar as suas possibilidades é objeto de estudo com demanda real e crescente. Neste sentido, surgem dois problemas a serem pesquisados na área: detecção de pontos de interesse (PoI) e previsão de próximo local de visita. Neste trabalho, são propostas soluções para esses dois problemas, considerando dados esparsos. O método proposto para a detecção de PoIs se destaca pela capacidade de definir o tipo do local de interesse em Casa ou Trabalho mesmo se a rotina de um determinado usuário é menos comum, como ir ao trabalho durante a noite e permanecer em casa durante o dia. Já a rede neural *MFA-RNN*, proposta para a predição de próximo local de visita, engloba convenientemente as mais recentes técnicas existentes na literatura, como utilização de múltiplos fatores de entrada (localização, tempo, identificação do usuário e tipo do dia), e aplicação do mecanismo *MHSA* (*Multi-Head Self-Attention*). Assim, diferentes aspectos podem ser aprendidos e correlacionados pela rede neural. Além disso, é descrito um método para o preenchimento de dados esparsos, que visa contribuir com o treinamento do modelo *MFA-RNN*. Os resultados obtidos demonstram que ambas as soluções desenvolvidas são eficazes para dados esparsos, e neste sentido, superam os principais métodos da literatura.

Palavras-chave: Pontos de interesse. Predição de próximo local. Dados esparsos

Abstract

CAPANEMA, Cláudio Gustavo Santos, M.Sc., Universidade Federal de Viçosa, March, 2020. **Points of Interest Detection and Next Place of Visit Prediction of Mobile Users Based on Sparse Data**. Advisor: Fabrício Aguiar Silva.

Location data from mobile devices is important for the study of human mobility. By better knowing their users, mobile service providers are interested in improving their products and increasing the customers engagement. The majority of the existing solutions in the literature were developed based on location data that was collected intensively, which leads to a high demand for resources such as energy, storing, processing and network on mobile devices. On the other hand, sparse data, even more limited, can be generated by a large number of users without affecting the power autonomy of their mobile devices. Thus, exploring its possibilities is an object of study with real and growing demand. In this sense, two problems arise to be researched in the area: points of interest (PoI) detection and next place prediction. In this work, solutions are proposed for these two problems, considering sparse data. The proposed method for PoI detection stands out for its ability to define the type of the location of interest as Home or Work even if the routine of a particular user is less common, such as going to work during the night and staying at home during the day. The MFA-RNN neural network, proposed for the next place prediction, conveniently includes the most recent techniques used in the literature, such as the use of multiple input features (location, time, user's ID and type of the day), and the application of the MHSA (Multi-Head Self-Attention) mechanism. Thus, different aspects can be learned and correlated by the neural network. In addition, it is described a method for sparse data filling, which aims to contribute to the training of the MFA-RNN model. The obtained results show that both developed solutions are effective for sparse data, and in this sense, they outperform the main methods of the literature.

Keywords: Points of interest. Next place prediction. Sparse data

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução Geral

Diversos estudos sobre a mobilidade humana têm sido conduzidos com a utilização em massa dos dispositivos móveis, como *smartphones* e *tablets*. A possibilidade de se inferir a localização geográfica de milhares de usuários móveis tem tornado mais eficiente o dia a dia humano em diversas áreas, como previsão de tráfego rodoviário (Gao et al., 2016), melhoria da interação entre pessoas (Yao et al., 2016), planejamento urbano (Rathore et al., 2016), dentre outros. Em comum, essas contribuições estão relacionadas ao conceito de Cidades Inteligentes, onde a tecnologia atua como agente de melhorias constantes.

Grande parte das soluções atuais relacionadas ao estudo da mobilidade humana utiliza dados de GPS coletados de dispositivos móveis para avaliar os seus métodos. O problema dessas soluções é que, em geral, os dados de GPS foram coletados de modo intensivo, o que leva a um alto custo em um cenário real de larga-escala, em que os aparelhos seriam principalmente afetados pelo alto consumo energético. Dados esparsos, no entanto, são mais apropriados a cenários reais, uma vez que são coletados entre intervalos maiores de tempo e por isso demandam menos recursos dos dispositivos móveis.

Dentre os diferentes problemas em aberto que podem ser explorados estão: a detecção de pontos de interesse (PoI) e a previsão de próximo local de visita. Ambos problemas são complementares, e juntos são capazes de descrever padrões de mobilidade de indivíduos e diferentes grupos de pessoas.

A tarefa de detecção de PoI busca identificar quais são os locais de maior interesse de um indivíduo bem como detalhar o seu tipo. Neste trabalho, são classificados dois tipos de PoI individuais com base em dados esparsos: Casa e Trabalho. Diferentemente de outras abordagens, o algoritmo proposto é capaz de reconhecer os diferentes tipos de PoI mesmo que um determinado usuário possua uma rotina adversa, como ir ao Trabalho à noite e permanecer em Casa durante o dia. Comparando-se com os principais métodos da literatura, os resultados obtidos apresentam melhorias na identificação de PoI em pelo menos 13%, e de no mínimo 10% e 4% para a classificação dos PoI dos tipos Casa e Outro, respectivamente.

Provedores de serviços móveis estão interessados em aprimorar os seus produtos e aumentar o engajamento dos seus clientes não somente através da detecção de pontos de interesse, mas do reconhecimento do padrão de mobilidade dos seus usuários como um todo. Neste sentido, prever o próximo local de visita de um usuário de dispositivo móvel é uma tarefa importante para aplicações que necessitam de respostas sobre o futuro próximo. Dessa forma, a segunda contribuição deste trabalho é o modelo *MFA-RNN* (*Multi-Factor Attention Recurrent Neural Network*), uma rede neural recorrente para predição de próximo local de visita que engloba mútiplos fatores dos traços de localização coletados (localização, horário, identificação do usuário e tipo do dia), e o mecanismo do estado da arte *Multi-Head Self-Attention*, o qual é capaz

de extrair correlações sobre diferentes partes da entrada da rede neural. Além disso, é proposto um novo método para o preenchimento de dados esparsos, tendo como objetivo contribuir com o reconhecimento de padrões de mobilidade pela rede neural. Comparando-se com as principais abordagens da literatura, a solução atual é a primeira a ser desenvolvida para dados esparsos, e os resultados obtidos apresentam desempenho superior em diferentes cenários de testes, principalmente quando o próximo local de visita é um PoI do tipo Casa ou Outro.

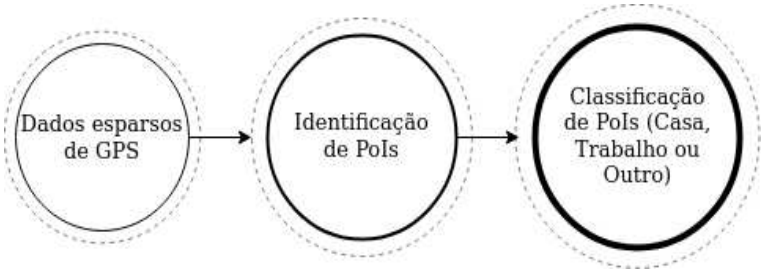

Figura 1.1: Fluxo de detecção de PoI.

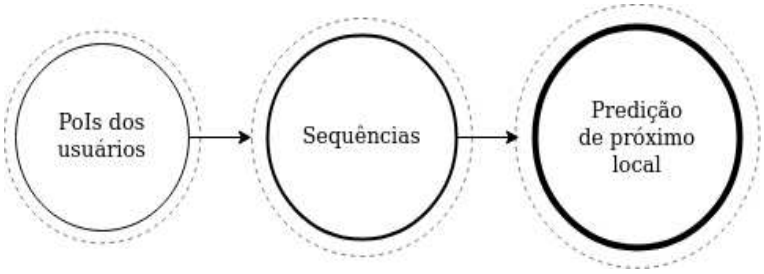

Figura 1.2: Fluxo de predição de próximo local de visita.

Por último, os fluxos de execução das tarefas das duas contribuições da pesquisa estão sumarizados a seguir:

- Com base nos dados de GPS esparsos de diferentes usuários móveis, são identificados os pontos de interesse de cada pessoa. Dentre esses locais, o método proposto classifica aqueles que correspondem à Casa e ao Trabalho. Quando um PoI não pertence a uma das classes anteriores ele é considerado do tipo Outro (veja Figura 1.1).
- A partir dos pontos de interesse detectados, é possível estabelecer o histórico de movimentação de cada pessoa em termos de PoIs visitados. Com base nesse histórico, são formadas sequências de locais visitados que servem de entrada para treinar o modelo de predição de próximo local de visita. Dessa forma, o método proposto é capaz de indicar o próximo PoI a ser visitado dada uma sequência de PoIs percorridos anteriormente por uma pessoa (veja Figura 1.2).

1.1 Justificativa

A detecção de pontos de interesse, bem como a previsão do próximo PoI de visita, são utilizados por serviços baseados em localização para diversas aplicações, como: sistemas de recomendação e mineração do padrão de mobilidade individual (Montoliu

et al., 2013), planejamento urbano (Hoteit et al., 2016; Rathore et al., 2016), melhoria na interação humana (Yao et al., 2016), previsões de tráfego (Gao et al., 2016) e de utilização de aplicativos (Yu et al., 2018), *marketing* baseado em localização (Schreckenberger et al., 2018), dentre outros.

Tabela 1.1: Métricas para cada tipo de dado.

	Denso	Esparso
Duração	✓	✗
Frequência	✓	✓
Trajetória	✓	✗
Informação temporal	✓	✓
Contatos	✓	✗

Diversos estudos já foram realizados na área, e a crescente utilização de dispositivos móveis continua demandando novas soluções capazes de minerar os dados gerados. No entanto, um dos aspectos relevantes para o desenvolvimento de novos métodos, é o conjunto de dados utilizado para avaliar as soluções, onde o volume e a qualidade dos mesmos são determinantes. Neste sentido, é importante que os dados de teste possam refletir a realidade.

Tabela 1.2: Consumo de recursos para cada tipo de dado.

	Denso	Esparso
Energia	alto	baixo
Processamento	alto	baixo
Armazenamento	alto	baixo
Rede	alto	baixo

A maioria dos trabalhos recentes da literatura se baseiam em dados de GPS densos, ou seja, coletados entre pequenos intervalos de tempo pelos dispositivos móveis. Com dados densos é possível se obter com precisão um maior número de métricas, como pode ser observado na Tabela 1.1. Nesse sentido, esse tipo de dado se destaca pelos seguintes aspectos: possibilidade de se conhecer o horário de chegada e saída de um local visitado (e consequentemente a duração de permanência), obtenção da trajetória percorrida e dos contatos estabelecidos entre usuários que visitaram um mesmo local. No entanto, a coleta intensiva de dados requer maior capacidade de armazenamento, tráfego de rede e processamento, além de afetar consideravelmente a autonomia energética dos aparelhos (veja Tabela 1.2), o que em alguns casos, pode dificultar a sua implantação em ambiente de produção. Assim, é preciso avaliar soluções sob um grande volume de informações que tenham sido geradas sem a necessidade de alta demanda de recursos pelos dispositivos móveis. Os dados de GPS

esparsos, onde podem existir longos períodos entre a geração de informações, são mais apropriados para cenários reais, com a coleta de dados de milhares de usuários não afetando significativamente o consumo de recursos dos dispositivos. Ao mesmo tempo, a redução na quantidade de informações geradas torna mais desafiadora a análise dos dados, surgindo uma grande demanda por novas soluções.

1.2 Objetivos

O objetivo geral desta pesquisa é propor novos métodos para a compreensão da mobilidade humana considerando-se dados esparsos de localização coletados de dispositivos móveis. Os objetivos específicos estão descritos a seguir:

1. Identificar pontos de interesse (PoI).
2. Classificar o tipo dos pontos de interesse em Casa, Trabalho e Outro.
3. Propor uma solução que englobe as principais técnicas do estado da arte para a tarefa de predição de próximo local de visita.
4. Apresentar um método para preenchimento de dados esparsos, com o objetivo de tornar mais efetivo o reconhecimento de padrões de mobilidade.
5. Avaliar como a rotina humana tende a variar ao longo da semana e aproveitar as informações obtidas para treinar a rede neural proposta.

1.3 Contribuições

- Modelo de identificação e classificação de pontos de interesse com base em dados esparsos que avança o estado da arte ao conseguir melhores resultados do que os principais trabalhos da literatura.
- Método de preenchimento de dados esparsos juntamente com uma rede neural para a previsão de próximo local de visita que, através dos resultados obtidos, representam uma nova abordagem no estado da arte.
- Implantação do método proposto em Capanema et al. (2019) em ambiente de produção, na empresa que forneceu os dados utilizados na pesquisa.

1.4 Organização da Dissertação

A organização desta dissertação está em conformidade com o padrão estabelecido pela Comissão do Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Viçosa (UFV). São apresentados dois artigos em formato de coletânea, sendo o primeiro publicado em uma conferência e o segundo submetido para o mesmo evento.

O Capítulo 2 apresenta o artigo Capanema et al. (2019), publicado nos anais do XXXVII Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos (SBRC 2019). O trabalho descreve uma abordagem eficaz para a detecção e classificação de pontos de interesse com base em dados esparsos.

O Capítulo 3 é composto pelo artigo “MFA-RNN: Uma Rede Neural Recorrente para Predição de Próximo Local de Visita com Base em Dados Esparsos”, que foi submetido e aceito para o XXXVIII Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos (SBRC 2020). Este trabalho aproveita as contribuições do artigo anterior para detecção de PoI com o objetivo de descrever a movimentação de cada usuário móvel em termos da sequência de pontos de interesse visitados. Com esses dados, é realizado o treinamento da rede neural recorrente proposta, a fim de prever o próximo local de visita de cada usuário. Além disso, é apresentado um novo método para preenchimento de dados esparsos que contribui com o desempenho do modelo *MFA-RNN*.

No Capítulo 4, as conclusões e trabalhos futuros são discutidos com base nos dois artigos produzidos.

Capítulo 2

Identificação e Classificação de Pontos de Interesse Individuais com Base em Dados Esparsos

CAPANEMA, Cláudio Gustavo Santos; SILVA, Fabrício Aguiar; SILVA, Thais Regina M. B. Identificação e Classificação de Pontos de Interesse Individuais com Base em Dados Esparsos. Em: Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos (SBRC). **Anais do XXXVII Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos - SBRC 2019**. Gramado: Brasil, 2019. p. 16-29.

Abstract

Geo-spatial data are important sources to understand mobile users profile, helping providers to offer better services. With this type of data, it is possible to identify relevant visiting points of a user, and even to classify these points as home and work locations. With this knowledge, mobile service providers can increase the engagement and the retention of these users. However, identifying and classifying point of interest (PoI) are not trivial tasks, and the majority of existing works assume that the data have to be collected with a high frequency, making the process harder and more expensive. In this work, we propose approaches to identify and classify PoIs based on sparse data that were collected during long time intervals. The results, when compared with literature solutions, show precision improvements of at least 13% on the identification of PoIs, and 10% and 4% in classification of home and work, respectively.

Resumo

Dados de localização de dispositivos móveis são fontes importantes para entender o perfil de usuários, ajudando os provedores a oferecerem melhores serviços. Com esse tipo de dado, é possível identificar os pontos relevantes de um usuário, e até mesmo classificar esses pontos como locais de casa ou trabalho. Com esse conhecimento, provedores de serviços móveis podem aumentar o engajamento e a retenção de seus clientes. No entanto, identificar e classificar pontos de interesse (PoI) não são tarefas triviais, e a maioria dos trabalhos existentes assumem que os dados devem ser coletados com uma frequência alta, dificultando e encarecendo o processo. Neste trabalho, são propostas abordagens para identificar e classificar PoIs com base em dados esparsos, ou seja, que foram coletados em intervalos longos de tempo. Os re-

sultados, quando comparados com soluções da literatura, mostram melhorias de pelo menos 13% na precisão para a identificação dos PoIs, e de 10% e 4% na classificação de pontos de casa e de trabalho, respectivamente.

2.1 Introdução

O advento da utilização em massa de dispositivos móveis, como *smartphones* e *tablets*, trouxe consigo a geração de grandes volumes de dados de localização de usuários. Diversas aplicações fazem o uso do sensor de GPS para fornecer serviços de ofertas baseadas em localização, auxílio na mobilidade, buscas na Web e entrega de conteúdo digital orientados à localização. Além disso, empresas de telefonia coletam dados de registros de acessos, chamados CDR (*Call Detail Records*), que também representam uma fonte importante de dados de localização (Naboulsi et al., 2016).

Além de facilitar a oferta de serviços e conteúdos baseados em localização, dados georreferenciados têm sido utilizados para o entendimento de padrões de mobilidade tanto de indivíduos quanto de grupos de pessoas (Pavan et al., 2015; Naboulsi et al., 2016). Essa fonte de dados está intimamente relacionada com o conceito de Cidades Inteligentes, uma vez que auxilia, por exemplo, no planejamento urbano (Rathore et al., 2016) e na previsão de volume de tráfego rodoviário (Castro et al., 2012). Além disso, empresas de diversos ramos utilizam dados georreferenciados para conhecer melhor os seus clientes para, assim, oferecem serviços mais personalizados.

Um aspecto importante da mobilidade urbana e do perfil de usuários móveis refere-se à identificação e classificação de pontos de interesse (PoIs) dos usuários. Esses pontos correspondem a locais que uma pessoa visita com certa frequência, podendo representar locais de residência, trabalho, lazer, escolas, locais em que se costuma fazer compras e se alimentar, dentre outros. Neste contexto, a identificação de PoIs refere-se a encontrar esses locais (i.e., definir as coordenadas aproximadas), enquanto a classificação visa categorizar um PoI pelo seu tipo (i.e., casa, trabalho, lazer, dentre outros).

A maioria dos trabalhos que visam identificar ou classificar PoIs utilizando informações geradas por GPS baseiam-se em dados densos, ou seja, com alta frequência de coleta (i.e., na ordem de poucos segundos). Com dados densos, é possível observar vários aspectos, como horário de chegada e partida de um local (e consequentemente o tempo de permanência), o trajeto feito de um local a outro e o tempo de deslocamento. No entanto, a coleta intensiva de dados georreferenciados leva a um alto consumo energético dos aparelhos móveis devido à utilização do sensor de GPS, um alto consumo de rede para a transmissão desses dados e uma necessidade maior de capacidade de armazenamento e processamento no servidor. Por isso, geralmente dados densos não são coletados, ou são coletados para amostras pequenas de usuários voluntários. Por outro lado, a coleta de dados georreferenciados de forma esparsa é mais fácil e menos custosa de ser alcançada, sendo uma alternativa viável tanto para os dispositivos móveis quanto para o servidor de armazenamento e processamento desses dados.

Dadas as informações acima, surgem as seguintes perguntas de pesquisa:

- É possível identificar com precisão PoIs de usuários com base em dados esparsos?

- É possível classificar com precisão os PoIs de usuários em *Casa* e *Trabalho* com base em dados esparsos?

Para responder a essas perguntas, o objetivo deste trabalho é propor algoritmos para identificação e classificação de PoIs individuais com base em dados esparsos. O algoritmo de identificação visa inferir os PoIs de um usuário com base em seus locais visitados. Já o algoritmo de classificação visa classificar quais pontos de interesse representam o local de *Casa* e *Trabalho*. Os algoritmos propostos foram comparados com soluções bem conhecidas da literatura utilizando a mesma base de dados esparsos. Observou-se que a proposta atual supera Montoliu et al. (2013) e Cuttone et al. (2014) na precisão para a identificação de PoIs em pelo menos 13%. Comparando com Hoteit et al. (2016) e Kung et al. (2014), as melhorias encontradas para a classificação de PoIs foram de pelo menos 10% para *Casa* e de 4% para *Trabalho*.

Este trabalho está organizado da seguinte forma. Inicialmente, na Seção 3.2 é apresentada uma revisão da literatura contendo os principais estudos relacionados à área de identificação e classificação de PoIs. Em seguida, na Seção 2.3 são descritas as características da base de dados utilizada no trabalho. Na Seção 2.4 o algoritmo proposto para a identificação de PoIs é apresentado e comparado com dois estudos da literatura. Em seguida, a Seção 2.5 contém a descrição e avaliação do algoritmo proposto para a classificação de PoIs em *Casa* ou *Trabalho*. Por último, na Seção 3.7 a conclusão e os trabalhos futuros são apresentados.

2.2 Trabalhos Relacionados

A identificação e classificação de PoIs é um assunto que faz parte de estudos sobre a mobilidade humana, e que vem crescendo nos últimos anos. Nesta seção, são apresentados os principais trabalhos da área. Inicialmente são apresentadas as soluções para a identificação de PoIs. Em seguida, são citadas propostas de classificação de PoIs em *Casa* e *Trabalho*. A Tabela 2.1 sumariza os principais trabalhos da literatura.

Para a identificação de PoIs, é comum utilizar-se algoritmos de agrupamento para processar as localizações. O trabalho apresentado em Csáji et al. (2013) utiliza CDRs e diagrama de Voronoi para associar as torres de telefonia mais frequentemente utilizadas com as suas regiões de cobertura, e em seguida as agrupa utilizando um método de triangulação. O trabalho de Frias-Martinez et al. (2010) também utiliza diagrama de Voronoi com o mesmo propósito do artigo citado anteriormente. Os autores de Isaacman et al. (2011); Ranjan et al. (2012) também processam dados de CDRs, e utilizam o algoritmo *Leader* para agrupar as torres de celular próximas. Além disso, Isaacman et al. (2011) recorre à regressão logística para se obter a relevância de cada local. Já o trabalho Lee et al. (2015) recorre a uma extensão do algoritmo DBScan, na qual é feito o agrupamento de traços de GPS com base em restrições de distância e velocidade calculadas entre dois pontos para se descobrir locais relevantes. Os autores de Cuttone et al. (2014) apresentam uma solução baseada no algoritmo *GMM*, e outra que utiliza o DBScan juntamente com uma restrição de distância máxima entre pares de coordenadas de GPS para identificar PoIs. Em Pavan et al. (2015), são consideradas restrições de tempo, distância e velocidade máximas para filtrar pares de registros consecutivos. Por último, o trabalho Montoliu et al. (2013) apresenta soluções de agrupamento que utilizam restrições de tempo e distância entre registros consecutivos.

Tabela 2.1: Comparação entre soluções.

Solução	Fonte(s) de dado(s)	Técnica(s) para identificação	Técnica(s) para classificação
(Csáji et al., 2013)	CDR	Diagrama de Voronoi e triangulação	*K-means*
(Trestian et al., 2009)	CDR	Tempo de permanência	Tempo de permanência
(Kung et al., 2014)	CDR e GPS	Tempo de permanência	Tempo de permanência
(Isaacman et al., 2011)	CDR	Leader e regressão logística	Quantidade de registros
(Hoteit et al., 2016)	GPS	Quantidade de registros	Quantidade de registros
(Järv et al., 2014)	CDR	*Multiple Linkage Analysis*	*Não classifica
(Schneider et al., 2013)	CDR	Tempo de permanência	Tempo de permanência
(Cuttone et al., 2014)	GPS, Wi-Fi e GSM	*Gaussian Mixture Method* e DBScan	*Não classifica
(Lee et al., 2015)	GPS	DBScan	*Não classifica PoI
(Ranjan et al., 2012)	CDR	*Leader*	Quantidade de registros
Frias-Martinez et al. (2010)	CDR	Diagrama de Voronoi	Algoritmo genético e quantidade de registros
(Montoliu et al., 2013)	GPS, GSM, Wi-Fi, Bluetooth e acelerômetro	Agrupamento por Grade e DBScan, e restrições de distância e tempo entre registros	*Não classifica
(Pavan et al., 2015)	GPS	Restrições de distância, tempo e velocidade entre registros	*Não classifica
Proposta Atual	**GPS**	**DBScan, e restrições de dias e horas diferentes**	**Quantidade de registros em horários específicos para cada usuário**

Dentre esses trabalhos, Csáji et al. (2013), Trestian et al. (2009), Kung et al. (2014), Isaacman et al. (2011), Järv et al. (2014), Schneider et al. (2013), Ranjan et al. (2012) e Frias-Martinez et al. (2010) utilizam CDRs, que são dados geralmente disponibilizados por operadoras de telefonia e que possuem apenas a localização das torres, e

não dos aparelhos. Já Lee et al. (2015); Pavan et al. (2015); Montoliu et al. (2013) assumem que os dados de GPS são densos, para se conhecer detalhes de deslocamento e permanência dos usuários em cada local. O estudo de Cuttone et al. (2014) é um dos poucos que considera dados de GPS esparsos e está mais diretamente relacionado ao presente trabalho, sendo utilizado como base de comparação na avaliação dos resultados. Considerando a relevância e a qualidade do trabalho apresentado em Montoliu et al. (2013), o mesmo também é utilizado como base de comparação, sendo que seus parâmetros foram ajustados para que ele seja melhor adaptado para dados esparsos, que é o foco do trabalho atual.

Outras técnicas além de agrupamentos também são utilizadas para a identificação de PoIs. Os autores de Järv et al. (2014) utilizaram *Multiple Linkage Analysis* para se obter as localizações com maior número de chamadas em cada mês de um usuário, que são consideradas suas localizações de interesse. Por outro lado, o tempo em que um usuário permaneceu em cada local visitado é considerado um fator relevante para a identificação de PoIs pelas propostas de Trestian et al. (2009), Kung et al. (2014) e Schneider et al. (2013). No entanto, esses trabalhos utilizam dados esparsos de CDRs, o que dificulta a estimativa precisa de tempo de permanência em um local.

Para a classificação de PoIs em *Casa* e *Trabalho*, o tempo de permanência em intervalos de horários pré-definidos é uma métrica comumente utilizada. De acordo com Trestian et al. (2009), o local de maior tempo de permanência entre 22:00h e 6:00h do dia seguinte é classificado como *Casa*; por outro lado, o local de *Trabalho* corresponde aos períodos de 10:00h às 12:00h e de 14:00h às 17:00h. Em Kung et al. (2014), o local no qual o usuário permaneceu por mais tempo de 8:00h às 20:00h corresponde ao *Trabalho*, e de 20:00h às 8:00h à *Casa*. O artigo Schneider et al. (2013) considera como *Casa* o local de maior tempo de permanência de meia noite às 06:00h horas da manhã.

A ideia de se classificar os locais como *Casa* e *Trabalho* com base no tempo de permanência em determinadas faixas de horário faz sentido, pois em geral as pessoas passam boa parte do dia em seu local de *Trabalho*, e da noite em *Casa*. Porém, medidas de tempo de permanência não são precisas o suficiente em dados esparsos, uma vez que o usuário pode visitar diversos locais sem gerar registros. Além disso, intervalos de tempo fixos não necessariamente representam todos os usuários, que possuem rotinas diferentes. Neste trabalho, é proposto um algoritmo que define intervalos de horário específicos para cada usuário para a classificação dos locais de *Casa* e *Trabalho*.

Para Csáji et al. (2013) é possível discernir claramente os locais de *Casa* e *Trabalho* após a aplicação do algoritmo *k-means* sobre os PoIs. Além disso, trabalhos como Ranjan et al. (2012); Isaacman et al. (2011); Frias-Martinez et al. (2010); Hoteit et al. (2016) contabilizam a quantidade de registros em determinadas faixas de horários para classificar os locais em *Casa* e *Trabalho*. As soluções Hoteit et al. (2016); Kung et al. (2014) foram utilizadas neste trabalho como bases de comparação com o nosso algoritmo de classificação de PoI por se tratarem de trabalhos recentes e relevantes da literatura.

Um dos grandes desafios do problema de identificar e classificar PoIs é a validação, já que dados rotulados com essas informações não são facilmente obtidos. Considerando CDRs, empresas de telefonia podem utilizar dados dos contratos de serviços para validar a região da residência e trabalho de seus usuários. No entanto, a precisão obtida ao se utilizar esse tipo de dado é baixa, já que uma antena cobre de centenas de metros a quilômetros. Por outro lado, trabalhos que utilizam dados de GPS necessitam da colaboração de usuários para prover suas informações, e esses

dados são em baixa escala, raros e privados. Neste quesito, os trabalhos Hoteit et al. (2016), Cuttone et al. (2014), Lee et al. (2015), Montoliu et al. (2013) e Pavan et al. (2015) processaram dados de, respectivamente, 32, 6, 46, 8 e 182 usuários. Já no presente trabalho, foi possível analisar dados reais de 194 usuários, superando as abordagens que utilizaram dados de GPS. A Tabela 2.1 apresenta os principais aspectos de cada um dos trabalhos relacionados, destacando os tipos de dados, e as técnicas utilizadas por cada um. Sensores como WiFi, *bluetooth* e acelerômetro, além da rede de telefonia (GSM) também podem ser associados às fontes de dados CDR e GPS para se inferir a localização de usuários.

2.3 O Conjunto de Dados

Os dados utilizados neste trabalho foram disponibilizados por uma empresa provedora de serviços móveis sob um acordo de confidencialidade. Foram fornecidos dados de localização de 194 usuários voluntários de dispositivos móveis, durante um período de até 62 dias. Os dados foram coletados mediante a permissão prévia dos usuários.

Para esse conjunto de usuários, foi gerado um registro de localização sempre que o usuário se deslocar para um novo local, distante pelo menos 100 metros do local anterior, e desbloquear a tela do seu *smartphone*. Ou seja, os dados são esparsos pois um único registro de localização é gerado quando há um deslocamento do usuário e um acesso ao *smartphone* no novo local. Um registro de localização contém o identificador do usuário, data e horário, latitude e longitude obtidas pelo sensor de GPS. O identificador do usuário é um número aleatório e único para cada usuário, que não permite a identificação do mesmo de forma alguma.

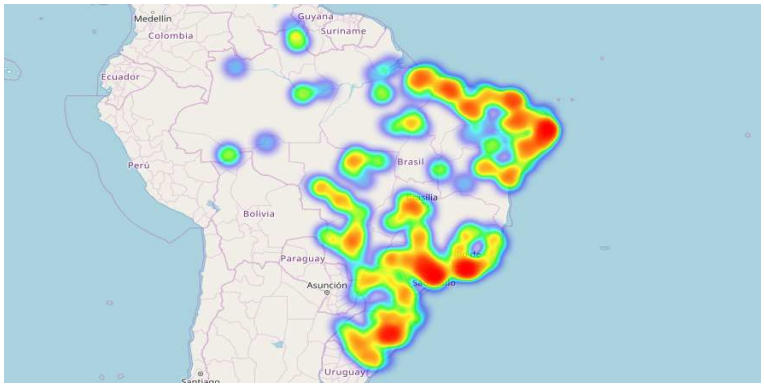

Figura 2.1: Mapa de calor dos registros coletados dos 194 usuários.

Os dados abrangem o território brasileiro, e como pode ser observado na Figura 2.1, há uma maior concentração de registros nas regiões nordeste, sudeste e sul. A Figura 2.2 exibe a quantidade total de registros gerados em cada hora do dia, sendo possível observar que os usuários são mais ativos entre 8:00h e 10:00h da manhã, e entre 13:00h e 16:00h da tarde. Além disso, a partir da Figura 2.3 percebe-se que a maioria dos usuários gerou entre 500 e 600 registros durante todo o período de coleta. É importante ressaltar que esse período de coleta varia de usuário para usuário, e os valores de média e mediana correspondentes são ambos de 34 dias. Isso indica que, em geral, os dados dos usuários foram coletados em uma mesma quantidade de dias.

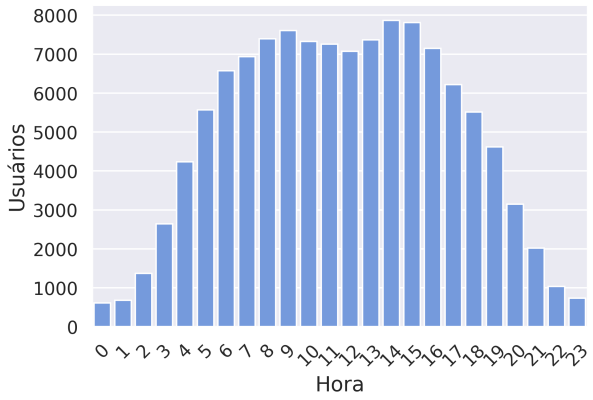

Figura 2.2: Registros por hora.

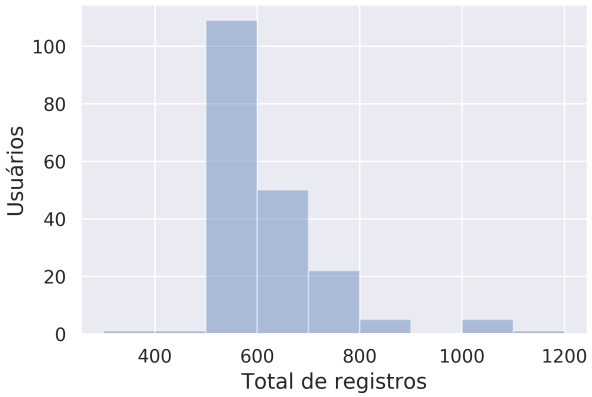

Figura 2.3: Total de registros por usuário.

Com o objetivo de demonstrar o quão esparsos são os dados, foram gerados gráficos de CDF (Função de Distribuição Acumulada) para a distância (Figura 3.1) e para o intervalo de tempo (Figura 3.2) entre pares de registros consecutivos dos usuários. Para as distâncias, a mediana é de 874 metros, a média é de 2.469 metros, e o desvio padrão é de 4.723 metros. Optamos por representar distâncias de até 47.000 metros, para criar um gráfico visualmente mais intuitivo, sendo que 1,31% das distâncias são superiores a 47.000 metros. Já para os intervalos de tempo, a mediana é de 26 minutos, a média corresponde a 66 minutos, e o desvio padrão é de 116 minutos. Além disso, optamos por representar tempos de até 700 minutos, também com o objetivo de criar um gráfico visualmente mais intuitivo, sendo que 1,75% dos tempos são maiores do 700 minutos.

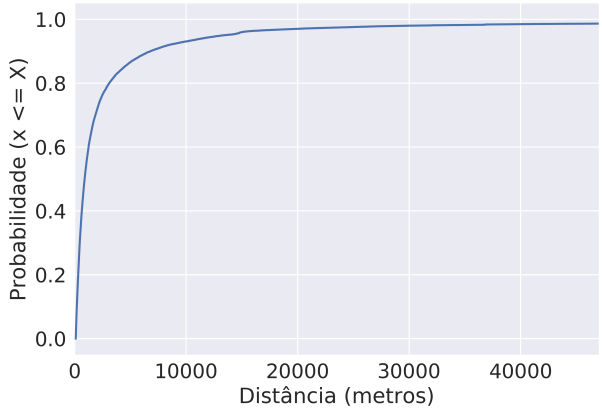

Figura 2.4: CDF de distância entre cada par de registros consecutivos.

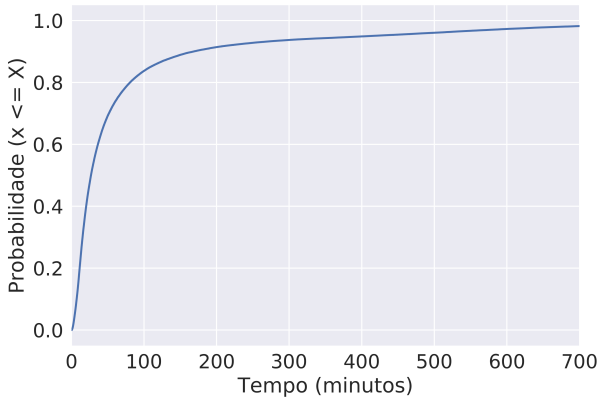

Figura 2.5: CDF do intervalo de tempo entre cada par de registros consecutivos.

2.4 Identificação de PoI

A identificação de pontos de interesse é uma tarefa essencial para o estudo da mobilidade humana. Com a informação dos PoIs individuais, é possível estudar padrões de mobilidade, perfil de usuários, demanda por infraestrutura computacional e viária, dentre outros. Porém, quando os dados são esparsos, os desafios são maiores pois não

é possível ter certeza do horário de chegada em um local, do tempo de permanência, e nem da intensidade de uso de um dispositivo móvel em cada local.

2.4.1 Solução

Neste trabalho, é proposto e validado um algoritmo para identificar pontos de interesse que seja adequado à característica esparsa dos dados. O algoritmo possui duas etapas: inicialmente, as localizações dos registros são agrupadas utilizando-se um algoritmo de agrupamento por densidade (DBScan), e em seguida os grupos encontrados são filtrados com base em duas métricas possíveis de serem conhecidas em dados esparsos, que são o total de dias visitados e a diversidade de horas de visita. Dessa forma, é possível se obter um conjunto de PoIs mais relevantes, melhorando assim o desempenho do algoritmo.

Para o agrupamento inicial, o DBScan foi escolhido por ser um algoritmo eficiente baseado em densidade, que não exige o número fixo de grupos como entrada e desconsidera pontos que não atendem às restrições estabelecidas por seus parâmetros. Neste caso, o parâmetro de número mínimo de amostras do DBScan foi definido de forma que cada grupo tenha mais de um registro por semana. Em avaliações empíricas, chegou-se a um valor de 18% do total de dias de coleta dos dados de cada usuário, o que corresponde a uma média de 1,26 registros por semana. Além disso, o raio da vizinhança de um ponto *core* é considerado 10 metros (parâmetro *eps* do *DBScan*). Esses parâmetros foram definidos empiricamente, e garantem que locais visitados muito esporadicamente, e que não estejam próximos a outros locais, não sejam considerados como candidatos a PoIs.

Ao final dessa primeira etapa, cada grupo gerado corresponde a um PoI candidato, e possui um conjunto de coordenadas e seus respectivos horários e datas. Dentre os PoIs candidatos gerados, são selecionados aqueles que obedecem a duas restrições. A primeira seleciona os PoIs que foram visitados em pelo menos 15% dos dias dado todo o período de amostragem do usuário. Isso significa que, por exemplo, para um período de 28 dias de coleta de dados, um PoI deve ter sido visitado pelo usuário pelo menos 4 dias diferentes no período, ou o equivalente a uma vez por semana. É importante ressaltar que o parâmetro de número mínimo de amostras permite que locais visitados em um pequeno número de dias diferentes sejam considerados como candidatos a PoI, enquanto que a restrição descrita impede isso. Por outro lado, a segunda restrição objetiva responder à seguinte pergunta: é possível estabelecer um número mínimo de horas diferentes do dia em que os registros pertencentes a um PoI foram gerados, de modo que se melhore o desempenho do algoritmo para a identificação de pontos de interesse? Após a realização de vários testes empíricos, observou-se que a seleção de PoIs que possuem pontos coletados em pelo menos 7 horas diferentes melhora o desempenho do algoritmo. Essas regras fazem com que apenas locais mais relevantes sejam considerados como PoIs.

O Algoritmo 1 representa a solução para a identificação de PoIs, que recebe como entrada a lista de registros de um usuário e retorna a lista de PoIs desse usuário. Na linha 2, os registros recebidos como entrada são agrupados pelo algoritmo DBScan. Em seguida, a partir da linha 3 até a linha 13, cada um dos grupos gerados são processados para se obter a quantidade de dias e horas diferentes que os seus registros representam. Na linha 14, é obtido o valor, em dias, do período entre o primeiro e o último registros gerados daquele grupo. Se a quantidade de dias diferentes for maior

ou igual a 15% do período que o usuário visitou aquele local, e se a quantidade de horas diferentes for maior ou igual a 7, então este grupo é considerado um PoI do usuário, como descrito no algoritmo nas linhas 15 e 16.

Algorithm 1 Identificação de PoIs de usuário

Require: $D, r = (r_1, ..., r_n)$: {Período de dias de coleta, e Lista de registros}
Ensure: $PoIs$ {Lista de pontos de interesse}
1: $PoIs \leftarrow \emptyset$
2: $Grupos = DBScan(data \leftarrow r, eps \leftarrow 10m, min\_samples \leftarrow |D| * 0.18)$
3: **for** $grupo \in Grupos$ **do**
4: $dias\_diferentes \leftarrow \emptyset$ {Lista com dias diferentes no grupo}
5: $horas\_diferentes \leftarrow \emptyset$ {Lista com horas do dia diferentes no grupo}
6: **for** $ponto \in grupo$ **do**
7: **if** $ponto.dia \notin dias\_diferentes$ **then**
8: $dias\_diferentes \leftarrow dias\_diferentes \cup \{ponto.dia\}$
9: **end if**
10: **if** $ponto.hora \notin horas\_diferentes$ **then**
11: $horas\_diferentes \leftarrow horas\_diferentes \cup \{ponto.hora\}$
12: **end if**
13: **end for**
14: $periodo \leftarrow Periodo(grupo)$
15: **if** $|dias\_diferentes| \geq periodo * 0.15$ **and** $|horas\_diferentes| \geq 7$ **then**
16: $PoIs \leftarrow PoIs \cup \{grupo\}$
17: **end if**
18: **end for**

2.4.2 Soluções Base

Com o objetivo de comparar o desempenho do algoritmo proposto com outras soluções da literatura, os algoritmos apresentados por Cuttone et al. (2014) e Montoliu et al. (2013) foram implementados. Cuttone et al. (2014) é um dos poucos trabalhos que consideram dados georreferenciados esparsos para a identificação de PoIs. Foi apresentada uma solução baseada no algoritmo *GMM (Gaussian Mixture Method)*, e uma baseada no algoritmo DBSCan, a qual foi escolhida devido ao melhor desempenho.

Montoliu et al. (2013) apresentou um algoritmo de agrupamento baseado em grade, e outro baseado no DBSCan, sendo que esse último obteve melhor desempenho sobre a nossa base de dados, e portanto foi escolhido como uma das soluções base. Esse algoritmo primeiramente agrupa os registros em relação ao tempo de permanência nos locais, para em seguida agrupar utilizando o DBScan. Apesar de se basear em dados densos, essa solução foi escolhida por ser uma referência na literatura, e para verificar se a mesma pode também ser utilizada com dados esparsos, o que ainda não foi feito por outros estudos. Para isso, os parâmetros da solução foram ajustados empiricamente para que sejam os mais apropriados possível a dados esparsos.

2.4.3 Resultados

Seja PoI_u o conjunto de PoIs do usuário u rotulados e conhecidos, $PoI_{u,c}$ o conjunto de PoIs do usuário u identificados corretamente e $PoI_{u,i}$ o conjunto de PoIs do usuário u identificados incorretamente por um algoritmo. A precisão ($p = \frac{|PoI_{u,c}|}{|PoI_{u,c}|+|PoI_{u,i}|}$) indica, dentre os pontos identificados, quantos estão corretos. A revocação ($r = \frac{|PoI_{u,c}|}{|PoI_u|}$) indica, dentre todos os PoIs reais de um usuário, quantos foram identificados corretamente pelo algoritmo. Por fim, o *f-score* ($f = 2 * \frac{p*r}{p+r}$) é derivado dessas duas métrica. Para as três métricas, quanto mais próximo de 1, melhor.

Primeiramente, foi avaliado o quão distantes os PoIs identificados pelos algoritmos estão dos PoIs reais dos usuários. Para isso, a Figura 2.6 ilustra a CDF (Função de Distribuição Acumulada) das distâncias entre todo PoI identificado pelos algoritmos e o PoI real mais próximo. Essa figura mostra o eixo-x com valor máximo de 1.000 metros por questões de visualização, sendo que a proposta atual tem 4,48% das distâncias acima desse valor, enquanto as soluções Cuttone et al. (2014) e Montoliu et al. (2013) têm 5,6% e 9,4%, respectivamente. Com base nesse gráfico, é possível perceber que a proposta deste artigo leva à identificação de PoIs mais próximos dos reais.

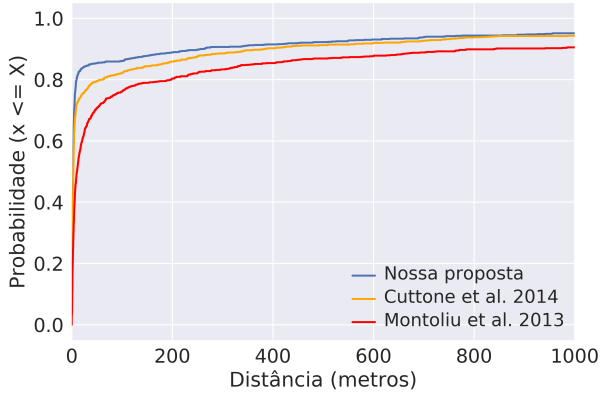

Figura 2.6: CDF de distâncias entre PoIs identificados e reais

Para contabilizar os PoIs identificados corretamente, é definida uma distância máxima para a margem de erro, já que dificilmente um algoritmo irá identificar exatamente as mesmas latitudes e longitudes de um PoI real. Neste trabalho, essa distância de margem de erro foi variada entre 10 e 100 metros. Assim, um PoI identificado é considerado correto se ele está distante a no máximo a distância de margem de erro de um PoI real.

As Figuras 2.7, 2.8 e 2.9 apresentam os resultados médios e o intervalo de confiança de 95% da proposta atual e das soluções base Cuttone et al. (2014) e Montoliu et al. (2013). Os desempenhos variam de acordo com a distância da margem de erro (eixo-x) entre o PoI encontrado por cada solução e o PoI real mais próximo. Pode-se perceber um desempenho superior da proposta atual em relação às soluções base nas métricas precisão, revocação e *f-score*. A precisão na nossa proposta varia de 70% a 74%, enquanto para Cuttone et al. (2014) o intervalo é de 56% a 61%, e para Montoliu et al. (2013) a precisão varia de 38% a 52%. Para a métrica revocação, os desempenhos das abordagens citadas anteriormente variam respectivamente, de 80% a 84%, de 73% a 80%, e entre 58% e 79%. Os desempenhos da proposta deste trabalho, de Cuttone

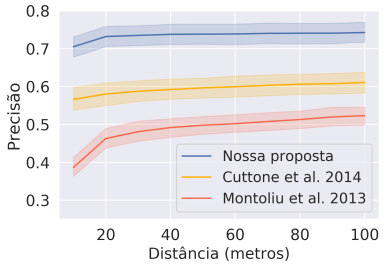

Figura 2.7: Precisão.

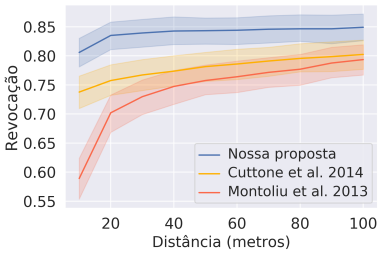

Figura 2.8: Revocação.

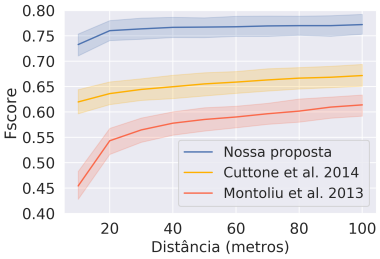

Figura 2.9: F-score.

et al. (2014) e Montoliu et al. (2013) para a métrica *f-score* variaram respectivamente entre 73% e 77%, 61% e 67%, e 45% e 61%.

Esses resultados mostram que, apesar de ser uma solução bem elaborada e eficiente para dados densos, a proposta de Montoliu et al. (2013) não é apropriada para dados esparsos. A solução proposta por Cuttone et al. (2014), por ser focada em dados esparsos, consegue alcançar resultados melhores que Montoliu et al. (2013). A proposta deste trabalho atual consegue melhores resultados ao considerar o número de horas e dias distintos de visitas aos locais, desconsiderando assim locais visitados um número significativo de vezes (o que faz com que um grupo seja considerado pelo DBScan), mas em poucos dias e horários diferentes (o que faz com que as restrições do algoritmo o desconsidere).

2.5 Classificação de PoIs

A classificação dos tipos de locais que cada usuário frequenta também corresponde a um passo relevante no processo de análise da mobilidade humana. Ao se conhecer os locais de casa e trabalho, e os padrões de deslocamento entre esses locais, provedores de serviços podem personalizar seus serviços de acordo com o perfil do usuário. Neste trabalho, são classificados os locais de *Casa* e *Trabalho* de cada usuário. Tanto a proposta apresentada quanto as soluções base selecionadas processam apenas pontos de GPS coletados durante dias de semana. Isso porque, em geral, a rotina humana é melhor definida durante esses dias, o que torna mais eficaz a classificação desses tipos de PoIs.

2.5.1 Solução

A *Casa* e o *Trabalho* são definidos como os locais onde mais registros foram gerados em horários específicos. O grande diferencial desta proposta é que esses horários variam de usuário para usuário, e não são fixos como nos outros trabalhos.

Inicialmente, identifica-se o maior período de inatividade do usuário, ou seja, o maior intervalo de horário de 00:00h até as 23:00h no qual não se geraram registros, considerando todos os dias de amostragem. O horário para a identificação da *Casa* corresponde ao intervalo de 2 horas anteriores até as 2 horas posteriores ao intervalo de inatividade encontrado na etapa anterior. A justificativa para isso é que, em geral, as pessoas começam e terminam os seus dias no local de sua moradia, e portanto é mais provável que sejam gerados registros em casa no intervalo próximo ao período de inatividade. Caso não seja encontrado um período de inatividade, consideramos

intervalos fixos iguais aos da solução base Hoteit et al. (2016), que será descrita posteriormente, uma vez que empiricamente esses foram os intervalos fixos que trouxeram os melhores resultados. O intervalo de horário para a classificação do local de *Trabalho* corresponde ao período do dia oposto ao intervalo de horário definido para a classificação da *Casa*. Dessa forma, supomos que o *Trabalho* é o local onde uma pessoa gera mais registros quando está fora de casa.

Em outras palavras, são definidos dois intervalos de horários para contabilizar registros em cada local. O local que possui mais registros no período definido para classificar a casa é classificado como *Casa*, e o mesmo processo é utilizado para classificar o local que representa o *Trabalho* do usuário. Portanto, caso o intervalo de horário de inatividade de um dado usuário seja das 22:00h até as 05:00h do dia seguinte, por exemplo, então o período para a classificação em *Casa* é das 20:00 (22:00h - 2h) às 07:00h (05:00h + 2h). Dessa forma, o período para a classificação do PoI em *Trabalho* é das 08:00h às 19:00h.

O Algoritmo 2 apresenta a solução proposta para a classificação de PoIs para cada usuário. A partir da linha 1 até a linha 9, são obtidas as horas que os registros de todos os PoIs foram gerados. Com essas informações, é possível obter o intervalo de inatividade através da função utilizada na linha 10. Como mencionado anteriormente, caso não exista um intervalo de inatividade para o usuário, considera-se o mesmo intervalo utilizado pela solução de Hoteit et al. (2016). Nas linhas 11 e 12, as funções utilizadas retornam o PoI que representa o local de *Casa*, e o PoI que representa o local de *Trabalho*. Para todo usuário, o algoritmo sempre define os locais de casa e trabalho.

Algorithm 2 Classificação de PoIs de usuário

Require: $PoIs = (p_1, ..., p_n)$: {Lista de pontos de interesse}
Ensure: $casa \in PoIs, trabalho \in PoIs$ {Pontos de interesse classificados}
1: $horas \leftarrow \emptyset$
2: **for** $poi \in PoIs$ **do**
3: **for** $ponto \in poi$ **do**
4: $hora \leftarrow ponto.hora$
5: **if** $hora \notin horas$ **then**
6: $horas \leftarrow horas \cup hora$
7: **end if**
8: **end for**
9: **end for**
10: $intervalo \leftarrow Intervalo\_Inativo(horas)$
11: $casa \leftarrow Classifica\_Casa(PoIs, intervalo)$
12: $trabalho \leftarrow Classifica\_Trabalho(PoIs, intervalo)$

2.5.2 Resultados

Foram utilizadas as propostas de Kung et al. (2014) e Hoteit et al. (2016) como soluções base. A primeira foi adaptada para contabilizar a quantidade de registros de cada PoI entre intervalos de horários, e não o tempo de permanência. Esta solução considera o intervalo de horário de 20:00h às 08:00h para determinar o local de *Casa*, e 08:00h às 20:00h para determinar o local de *Trabalho*. A segunda proposta classifica a *Casa* como o local onde o usuário gerou mais registros no período de tempo de 22:00h

às 7:00h do dia seguinte, e o *Trabalho* como o local de maior número de registros de 9:00h às 17:00h. Nesta avaliação, foi considerada uma distância de 100m de margem de erro. Em outras palavras, se o local de *Casa* classificado pelo algoritmo estiver a menos de 100m do local real da casa do usuário, então considera-se um acerto. Vale destacar que os PoIs considerados como reais são os originais dos dados rotulados, e não os obtidos pelo algoritmo proposto neste trabalho. Isto permite que a validação seja feita de modo independente, assumindo que a identificação de PoIs já foi realizada. Além disso, somente foram considerados na avaliação os PoIs rotulados como *Casa* ou *Trabalho* nos dados originais, visando assim avaliar a eficiência das soluções em distinguir entre um tipo de local e outro. A precisão é dada pela divisão do número de acertos pelo número de usuários, pois foram classificados os locais de todos os usuários.

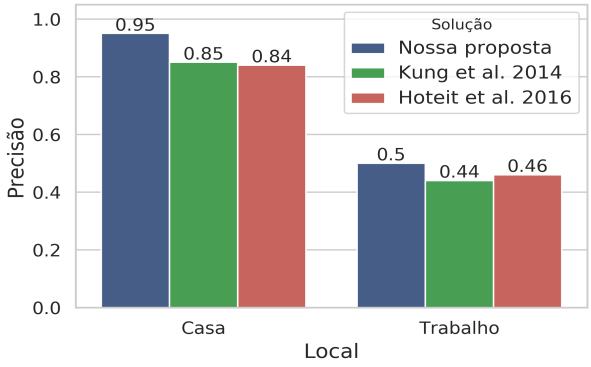

Figura 2.10: Precisão de classificação de PoIs.

A Figura 2.10 apresenta o desempenho do algoritmo proposto em comparação com as soluções base Kung et al. (2014) e Hoteit et al. (2016), para classificar PoIs em *Casa* e *Trabalho*. A utilização de intervalos variáveis de horário levou a um desempenho melhor, de 95% para a classificação do local de *Casa* e de 50% para a classificação do local de *Trabalho*. Os respectivos desempenhos alcançados por Kung et al. (2014) foram de 85% e 44%. Por outro lado, a solução Hoteit et al. (2016) obteve valores de 84% para classificação de *Casa* e 46% para *Trabalho*. Além disso, foi possível utilizar intervalos variáveis de horário para 112 dos 194 usuários, demonstrando que a proposta atual pode ser aplicada de maneira abrangente. A partir dos resultados, também é possível notar que a tarefa de encontrar o PoI Trabalho é mais difícil do que encontrar o local de Casa. A razão para isso é que, cada pessoa pode ter um tipo de trabalho diferente, o que requer padrões de mobilidade específicos.

2.6 Conclusões e Trabalhos Futuros

O presente trabalho apresentou duas contribuições para a área de análise de mobilidade, tendo como diferencial a utilização de dados esparsos. A primeira contribuição corresponde à identificação de pontos de interesse (PoIs) de usuários de dispositivos móveis, e a segunda corresponde à classificação de PoIs em casa e trabalho. Os algoritmos apresentados foram validados comparando-se os resultados gerados com soluções conhecidas da literatura. As avaliações comparativas mostraram que foi possível alcançar melhores resultados, demonstrando que as hipóteses assumidas em termos

da diversidade de visitas (para identificação) e da personalização dos intervalos de horários (para classificação) foram eficientes.

Como trabalhos futuros, pretende-se melhorar a classificação de locais de trabalho, além de classificar outros tipos de PoIs, como locais de refeição e lazer. Além disso, também é relevante analisar aspectos de deslocamento entre os PoIs, para se prever possíveis demandas por recursos computacionais e viários.

2.7 Agradecimento

Este trabalho contou com o apoio da Fapemig, CNPq e CAPES.

Capítulo 3

MFA-RNN: Uma Rede Neural Recorrente para Predição de Próximo Local de Visita com Base em Dados Esparsos

Abstract

Predicting the user's mobility is an important task to enhance the effectiveness of mobile applications. In this work, we present the MFA-RNN (Multi-Factor Attention Recurrent Neural Network), a neural network that uses the Multi-Head Self-Attention technique to extract correlations under several features of the sequence of the visited places. The proposed model is able to predict the next place of visit considering multiple factors (user, location, time and type of the day) of each record of the sequence. Moreover, we propose a method to fill sparse data to enhance the performance of the solution. The obtained results indicate the effectiveness of the MFA-RNN model in relation to four known solutions of the literature.

Resumo

Prever a mobilidade de um usuário é uma tarefa importante para se elevar a efetividade de aplicações móveis. Neste trabalho, é apresentada a *MFA-RNN* (*Multi-Factor Attention Recurrent Neural Network*), uma rede neural recorrente que utiliza a técnica *Multi-Head Self-Attention* para extrair correlações sob diversos aspectos da sequência de locais visitados. O modelo é capaz de prever o próximo local de visita considerando múltiplos fatores (usuário, localização, tempo e tipo do dia) de cada registro da sequência. Além disso, é proposto um método para o preenchimento de dados esparsos para melhorar o desempenho da solução. Os resultados obtidos indicam a eficácia do modelo *MFA-RNN* em relação a quatro soluções conhecidas na literatura.

3.1 Introdução

A popularização dos dispositivos móveis trouxe consigo o conceito de *Location-Based Services (LBS)*, que contempla um conjunto de serviços destinados a tornar mais eficiente o dia a dia de usuários móveis (Al-Molegi and Martínez-Ballesté, 2018). Esses serviços têm como principal recurso a utilização de informação georreferenciada gerada pelos sensores dos dispositivos móveis. Considerando que o comportamento humano tende a seguir padrões de mobilidade de pessoa para pessoa (Banovic et al., 2016), é importante que as rotinas de milhares de usuários sejam estudadas.

Neste contexto, surge o problema de predição de próximo local de visita considerando o histórico de movimentação de um usuário. Essa tarefa é especialmente importante para *LBS* que necessitam de informação sobre o futuro próximo. Dentre as aplicações estão sistemas de recomendação (Wei et al., 2012), previsão de tráfego (Gao et al., 2016), melhoria na interação humana (Yao et al., 2016), dentre outras. Assim, ao se prever o próximo local de visita, provedores de serviços podem aprimorar os seus produtos, e aumentar o engajamento e a satisfação de seus clientes.

Diversos métodos têm sido apresentados para se prever o próximo local de visita. As redes neurais recorrentes têm se destacado por tornar possível identificar padrões de mobilidade de diferentes agentes sob vários fatores. Recentemente, estudos têm sido conduzidos para avaliar quais tipos de informações trazem maior impacto na predição do próximo local de visita. No entanto, um aspecto pouco abordado, mas muito relevante, diz respeito ao método de coleta e o volume dos dados utilizados para avaliar as soluções desenvolvidas atualmente. Dados gerados por meio de *check-ins* de localização em redes sociais não são capazes de representar por completo a rotina humana, uma vez que é necessário que o usuário sempre informe a sua posição de forma ativa (i.e., coleta ativa). Como alternativa, alguns trabalhos recorrem a dados de *GPS* coletados entre pequenos intervalos de tempo sem a necessidade de interferência humana (i.e., coleta passiva), a fim de avaliar melhor as suas soluções. Esse método de coleta de informações, no entanto, dificilmente pode ser aplicado em larga escala, onde a elevada carga de utilização do sensor de *GPS* pode comprometer a autonomia energética dos dispositivos móveis. Assim, soluções desenvolvidas sob dados de *GPS* esparsos, onde podem existir longos períodos entre a coleta de informações, tendem a ser mais atrativas para serem aplicadas na prática.

Neste trabalho, é proposta a *MFA-RNN* (*Multi-Factor Attention Recurrent Neural Network*), uma rede neural recorrente para a predição de próximo local de visita baseada em dados esparsos. Comparando-se com estudos anteriores, a *MFA-RNN* engloba convenientemente o mecanismo de *Multi-Head Self-Attention* (Vaswani et al., 2017) com diferentes tipos de informações que podem ser obtidas dos traços de localização, e que não foram utilizadas simultaneamente em outros trabalhos. Além disso, é proposto um novo método de preenchimento de dados de *GPS* esparsos, a fim de reduzir os efeitos da falta de dados e elevar o desempenho da rede neural. Dessa forma, as contribuições deste trabalho são resumidas em:

- Proposta de uma nova arquitetura capaz de aplicar a técnica de *Multi-Head Self-Attention* no contexto de predição do próximo local de visita ao mesmo tempo que engloba diferentes fatores.
- Utilização de múltiplos fatores de entrada para o modelo: localização, horário, ID do usuário e o tipo do dia (dia de semana ou final de semana).
- Considerando o trabalho de (Capanema et al., 2019), é possível inferir se uma pessoa está em casa em horários específicos mesmo que não tenham sido gerados registros. Essa contribuição é estendida neste trabalho para preencher a base de dados esparsos, e assim, facilitar o reconhecimento de padrões pela rede neural.
- Diferentemente de outros trabalhos, é apresentada uma análise que indica como a rotina humana varia entre dias de semana e finais de semana. Juntamente com

os resultados obtidos, a utilização da informação de tipo do dia do registro é, dessa forma, fundamentada tanto do ponto de vista teórico quanto prático.

- Validação da proposta *MFA-RNN* considerando-se dados de 5.272 usuários reais, o que representa a maior base de dados coletados de forma passiva entre os trabalhos existentes encontrados na literatura.

O restante do trabalho está organizado da seguinte forma. Inicialmente, a Seção 3.2 apresenta os principais trabalhos relacionados. A Seção 3.3 descreve a base de dados e o método utilizado de preenchimento de dados esparsos. Na Seção 3.4, é apresentada a análise sobre a rotina dos usuários. Posteriormente, a arquitetura *MFA-RNN* é apresentada na Seção 3.5. Por último, os resultados dos experimentos, e a conclusão e trabalhos futuros estão organizados nas Seções 3.6 e 3.7, respectivamente.

3.2 Trabalhos Relacionados

Diversas técnicas já foram desenvolvidas para a predição de próximo local de visita. As redes neurais recorrentes são uma variação de rede neural que têm ganhado destaque sobre métodos tradicionais, como cadeias de *Markov* (Zeng et al., 2019), para solucionar o problema. A evolução das soluções tem ocorrido em dois principais aspectos: utilização de diferentes tipos de informações que compõem os traços de localização, e aprimoramento da arquitetura da rede neural com a implementação de novas camadas.

Os modelos *STF-RNN* (Al-Molegi et al., 2016) e *MAP* (Al-Molegi et al., 2018) consideram como entradas para seus modelos apenas informações espaço-temporais (localização e horário) dos eventos gerados pelos dispositivos móveis. Em geral, cada localização corresponde ao identificador do ponto de interesse (PoI) ou região de interesse (RoI) visitado pelo usuário. Já a informação temporal é comumente representada pela hora do dia em que o evento foi gerado. Em busca de aprimoramentos, (Al-Molegi and Martínez-Ballesté, 2018) concluíram, através do modelo chamado *STW-RNN + WKD/WKE*, que o tipo do dia (dia de semana ou final de semana) do evento é uma informação relevante para a rede neural. No entanto, não foram conduzidas análises do quanto a rotina humana pode variar ao longo da semana. Por outro lado, os trabalhos (Feng et al., 2018; Yao et al., 2017; Zeng et al., 2019) avançaram na área ao incorporarem à rede neural o identificador do usuário que gerou cada evento, tornando possível que o modelo aprenda a rotina específica de cada pessoa. Além disso, na proposta de (Yao et al., 2017), de nome *SERM*, foi introduzida a utilização de informações textuais de dados do *Twitter* e *Foursquare* através de uma camada *Embedding* associada com a técnica *GloVe* (Pennington et al., 2014).

Com relação à arquitetura de redes neurais, os autores de (Vaswani et al., 2017) introduziram uma nova camada de nome *Multi-Head Self-Attention* (*MHSA*), originalmente desenvolvida para o problema de Tradução de Máquina. O mecanismo é capaz de extrair correlações sob diferentes partes de uma sequência, e assim como demonstrado em (Zeng et al., 2019), pode ser aplicado no problema de predição de próximo local de visita. A solução de (Zeng et al., 2019) é referida como *MHSA+PE* no presente trabalho, onde *PE* (*Positional Enconding*) é uma camada utilizada para propagar a noção de ordem dos eventos de uma sequência. Já em (Al-Molegi et al., 2018), um mecanismo de atenção simplificado foi utilizado no modelo *MAP*. Dentre as camadas

recorrentes, as soluções de (Al-Molegi et al., 2016, 2018) utilizaram a *Simple RNN*, enquanto em (Yao et al., 2017) a *Long Short-Term Memory* (*LSTM*) foi aplicada. Os autores de (Zeng et al., 2019; Feng et al., 2018) recorreram a uma camada *Gated Recurrent Unit* (*GRU*) que, assim como a *LSTM*, tem o diferencial de aprender melhor com longas sequências.

Tabela 3.1: Comparação com trabalhos da literatura.

Solução	**Camada recorrente**	**Dados de entrada**	**Mecanismo de atenção**	**Usuários avaliados**
MAP	*Simple RNN*	Localização e tempo	Simples	182 (coleta passiva), 319.063 (coleta ativa)
STF-RNN	*Simple RNN*	Localização e tempo	Não possui	182 (coleta passiva)
SERM	*LSTM*	Localização, tempo, ID do usuário e informação contextual	Não possui	7.826 (coleta ativa)
DeepMove	*GRU*	Localização, tempo e ID do usuário	Simples	15.639 (coleta ativa), 1.075 (coleta passiva)
STW-RNN+WKD/WKE	*Simple RNN*	Localização, tempo e tipo do dia	Não possui	182 (coleta passiva), 319.063 (coleta ativa)
MHSA+PE	*GRU*	Localização, tempo e ID do usuário	*MHSA*	1.083 (coleta ativa)
MFA-RNN (Proposta Atual)	***GRU***	**Localização, tempo, ID do usuário e tipo do dia**	***MHSA***	**5.272 (coleta passiva)**

Neste trabalho, a solução proposta *MFA-RNN* engloba as principais técnicas descritas na literatura em termos de redes neurais (camadas *MHSA* e *GRU*, e múltiplas entradas), para a predição do próximo local de visita. Além disso, diferentemente das soluções existentes, a proposta também considera que os dados são esparsos e portanto, realiza um preenchimento dos mesmos. A Tabela 3.1 sumariza as principais

soluções que envolvem redes neurais, e as compara com o presente trabalho. Em geral, as soluções da literatura são avaliadas sob um grande número de usuários quando o método de coleta de dados é ativo (e.g., *check-ins*). Porém, quando o método de coleta é passivo, poucos usuários de teste são considerados. Para aferir a qualidade da proposta *MFA-RNN*, as soluções *MAP* (Al-Molegi et al., 2018), *STF-RNN* (Al-Molegi et al., 2016), *SERM* (Yao et al., 2017) e *MHSA+PE* (Zeng et al., 2019) são utilizadas como base de comparação neste trabalho.

3.3 Descrição dos Dados

3.3.1 Características

O conjunto de dados utilizado foi fornecido sob confidencialidade por um provedor de serviços privado. Foram coletados, voluntariamente, dados de 5.272 usuários de dispositivos móveis durante um período de 62 dias. Cada evento de localização contendo a coordenada geográfica, o horário e o identificador do usuário era gerado quando o usuário se deslocava em pelo menos 100 metros, e desbloqueava a tela do seu aparelho. Apesar da necessidade de o usuário desbloquear a tela para que a coleta fosse realizada, diferentemente de *check-ins* utilizados na coleta ativa, o desbloqueio da tela é uma ação usual feita constantemente pelos usuários, principalmente em seus pontos de interesse. Esse método de coleta leva a uma base de dados esparsa, em que longos períodos podem existir entre a coleta de um evento e outro. As Figuras 3.1 e 3.2 exibem gráficos de CDF (*Função de distribuição acumulada*), obtidos pela distância e o tempo entre eventos consecutivos. Existe 50% de probabilidade de que a distância entre eventos consecutivos seja maior do que 1.049 metros e 50% de probabilidade para o tempo seja maior do que 35 minutos.

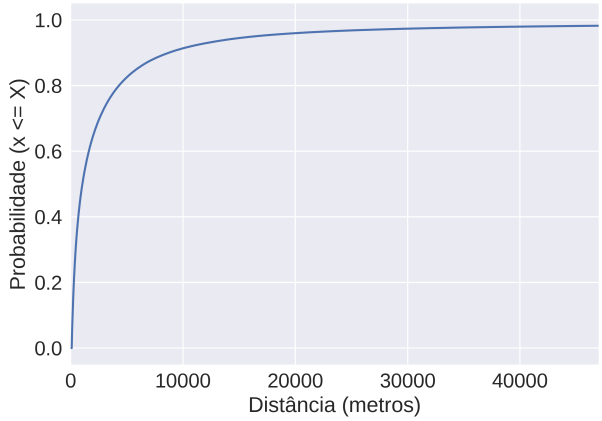

Figura 3.1: CDF de distância entre cada par de registros consecutivos.

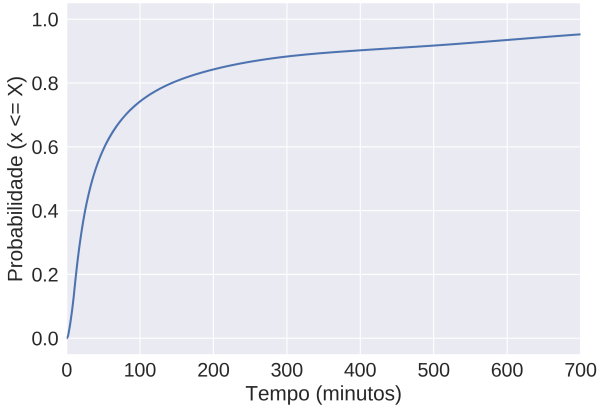

Figura 3.2: CDF do intervalo de tempo entre cada par de registros consecutivos.

3.3.2 Preenchimento

Para cada usuário, é necessário primeiro definir a sua mobilidade em termos de pontos de interesse visitados. Para isso, foi utilizada a solução proposta em (Capanema et al., 2019), desenvolvida sob dados esparsos. Foram identificados e classificados

dois tipos de pontos de interesses pessoais: *Casa* e *Outro*. A detecção do PoI *Trabalho* possui uma assertividade baixa, e dessa forma, optou-se considerá-lo como *Outro* neste trabalho. Assim, para cada usuário é possível identificar o seu local de moradia e PoIs diversos que são definidos como *Outro*. Além disso, um terceiro estado de localização é considerado quando o usuário não está em nenhum PoI: *Deslocamento*. Essas informações são utilizadas para construir o conjunto de sequências de locais visitados para cada usuário. Cada sequência $S(u)$ de um usuário u é composta por N eventos $e = (l, t, id, td)$ gerados durante a coleta de dados. Os elementos de um evento correspondem, respectivamente, ao tipo da localização (Casa, Outro ou Deslocamento), hora do dia, usuário e tipo do dia (dia de semana e final de semana). O problema de aplicar puramente essa abordagem, no entanto, é a existência de longos períodos entre um evento e outro, o que reduz a chance de real correlação entre a ordem dos locais visitados por uma pessoa.

Considerando o método de coleta descrito anteriormente, o trabalho de (Capanema et al., 2019) mostrou que o usuário tende a estar em casa durante o maior período de tempo onde não foram gerados eventos por um usuário. Isso ocorre porque é necessária a ação do usuário com o desbloqueio do aparelho e uma movimentação mínima para se gerar dados de localização. Porém, no trabalho anterior, essa informação havia sido utilizada apenas para detectar PoIs do tipo Casa, e não para preencher sequências. Dessa forma, neste trabalho, esse conhecimento é estendido e aplicado para preencher a sequência de localizações do usuário com o PoI *Casa* durante o período de inatividade do usuário.

3.4 Análise de Rotina

A mobilidade humana pode variar de acordo com diversos fatores, dentre eles os tipos dos dias de uma semana (dia de semana ou final de semana). Nesta seção, é apresentado um breve estudo que objetiva analisar o quanto a rotina humana pode variar ao longo da semana, e consequentemente, justificar o aproveitamento dessa informação na proposta *MFA-RNN*.

Após a identificação e classificação dos PoIs de um usuário, é possível caracterizar a sua rotina em termos da sua localização para cada hora do dia. Para isso, são contabilizadas as quantidades de eventos gerados em cada tipo de localização (*Casa*, *Outro* e *Deslocamento*) de acordo com cada hora do dia. Dessa forma, é possível estabelecer as probabilidades de se estar em *Casa* às 20:00 horas ou em *Deslocamento* às 11:00 horas, por exemplo.

Tabela 3.2: Exemplo da rotina de um usuário.

Horário	**00:00-01:00**	**01:00-02:00**	**...**	**22:00-23:00**	**23:00-00:00**
Casa	0.33	0.33	...	0.7	0.8
Outro	0.33	0.33	...	0.1	0.1
Deslocamento	0.33	0.33	...	0.2	0.1
Entropia	1.10	1.10	...	0.8	0.64

A Tabela 3.2 exemplifica a distribuição de probabilidades para um usuário, em que para cada hora do dia é calculada a entropia de *Shannon* (Shannon, 1948), com

base e, para medir o nível de incerteza do respectivo conjunto de probabilidades. Assim, para cada hora se obtém o quão previsível é a rotina do usuário. Como o trabalho é avaliado sob dados esparsos, é comum que em algumas horas do dia nenhum registro tenha sido gerado por um determinado usuário. Nesse caso, as probabilidades de se estar em Casa, Outro e Deslocamento ser tornam iguais, e a entropia correspondente tem o valor máximo de 1.10, indicando elevada incerteza sobre a rotina naquele horário. A média das entropias obtidas em cada horário indica o quão previsível pode ser a rotina de uma determinada pessoa considerando o seu padrão de mobilidade. Intuitivamente, a rotina humana tende a variar ao longo da semana. Dessa forma, a caracterização de rotina é feita separadamente para dias de semana e finais de semana.

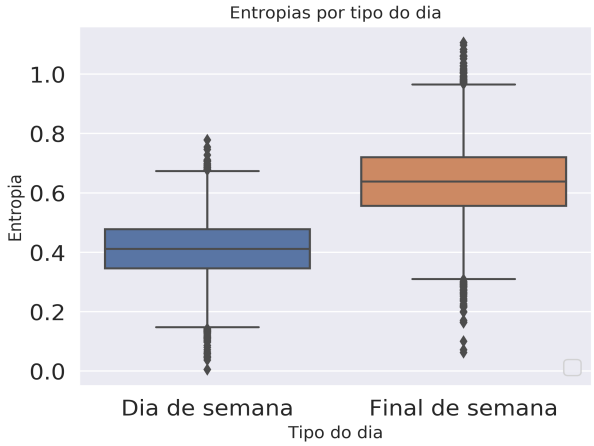

Figura 3.3: Distribuição das entropias médias de todos os usuários. Quanto menor o valor, mais previsível é a rotina.

A Figura 3.3 apresenta as distribuições das entropias de todos os usuários separadamente para dias de semana e finais de semana. Pode-se perceber que existe uma maior previsibilidade da rotina dos usuários analisados durante os dias de semana, uma vez que o valor da entropia é menor. Por outro lado, o valor maior de entropia durante finais de semana indica que as rotinas tendem a variar muito nesses dias.

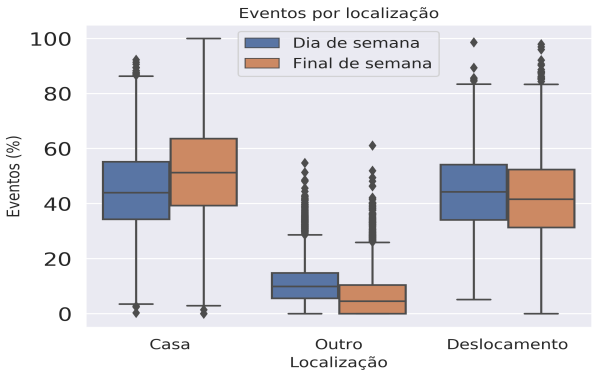

Figura 3.4: Comparação entre a porcentagem de eventos de cada localização gerados em dia de semana e final de semana.

Outra informação importante é que a proporção de eventos gerados em cada localização varia de acordo com o tipo de dia. A Figura 3.4 indica que existem pro-

porcionalmente mais eventos em *Casa* nos finais de semana do que nos demais dias. Por outro lado, os pontos de interesse do tipo *Outro* têm um menor peso no número de visitas nos sábados e nos domingos do que nos dias de semana. Similarmente, os usuários têm uma tendência menor de estar em *Deslocamento* durante os finais de semana, em relação ao demais dias.

Portanto, indicar o dia da semana em que um evento foi gerado tende a ser uma informação relevante para uma rede neural prever o próximo ponto de visita de um usuário. A variação da entropia, e a troca de proporção de eventos entre *Casa* e *Outro* durante a semana, sugerem as seguintes hipóteses:

1. Existe uma probabilidade maior para que o próximo local de visita seja *Casa* nos finais de semana do que nos demais dias. Por outro lado, de segunda-feira à sexta-feira os usuários tendem a estar em *Deslocamento* e visitar mais o PoI *Outro*.

2. Em geral, durante dias de semana, menos localizações diferentes são visitadas em um mesmo horário, o que reduz a entropia. Assim, a tendência é que cada horário esteja associado a uma localização predominante apenas.

Assim, a utilização pela rede neural da informação de tipo de dia da semana em que eventos foram gerados pode ser fundamentada do ponto de vista teórico.

3.5 Arquitetura MFA-RNN

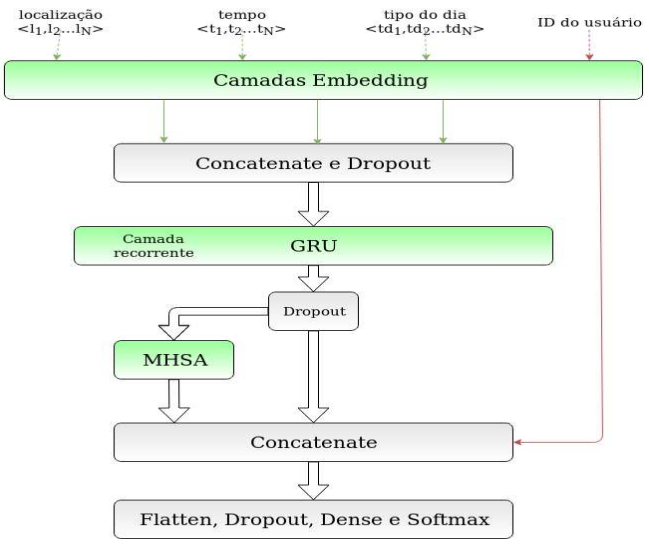

Figura 3.5: Arquitetura MFA-RNN.

Redes neurais recorrentes são tradicionalmente utilizadas para processar sequências de informações textuais, onde se busca compreender, por exemplo, as correlações entre a ordem de ocorrência de palavras de uma frase. Em uma interpretação mais ampla, as palavras podem ser compreendidas como variáveis categóricas, que no contexto deste trabalho correspondem aos elementos que compõem cada evento (ver Seção 3.3.2).

A rede neural proposta MFA-RNN, ilustrada na Figura 3.5, é capaz de aprender sob múltiplos fatores (localização, tempo, usuário, tipo do dia) para prever o próximo local de visita. Diversas camadas (*Embedding*, *Concatenate*, *GRU*, *MHSA* e *Dropout*) estão convenientemente dispostas, a fim de se alcançar um modelo eficiente e generalista.

A seguir, a arquitetura da rede neural é descrita destacando-se as características das principais camadas utilizadas.

3.5.1 Camadas *Embedding*

A entrada da rede neural pode ser compreendida como uma composição de quatro matrizes de localização, tempo, tipo do dia e ID do usuário, que representam os eventos de uma sequência codificados em inteiros positivos. A função das camadas *Embeddings* é representar cada um desses inteiros na forma de vetores numéricos de tamanho fixo. Assim, os valores de cada representação vetorial são atualizados durante o treinamento a fim de que palavras que tendem a ocorrer próximas nas sequências tenham representações vetoriais parecidas. Além disso, a utilização de vetores densos torna possível trabalhar eficientemente com grandes vocabulários de palavras, o que não ocorre com a técnica de *one-hot-encoding*, onde os vetores tendem a ser de grande dimensão e esparsos (Feng et al., 2018).

Após a camada *Embedding*, os vetores de saída de localização, tempo e tipo do dia são concatenados, e a técnica *Dropout* é aplicada com o intuito de evitar sobre-ajuste de treinamento e assim obter um modelo mais generalista.

3.5.2 Camada recorrente: *Gated Recurrent Unit (GRU)*

A intuição por trás de uma camada recorrente é que existe uma ordem lógica na disposição dos elementos em uma sequência, que portanto deve ser aprendida. No contexto deste trabalho, a ordem representa a sequência de locais visitados pelo usuário.

Diferentemente da *Simple RNN*, as camadas *LSTM* (*Long Short-Term Memory*) e *GRU* (*Gated Recurrent Unit*) são variantes de redes neurais recorrentes capazes de extrair correlações entre eventos de longas sequências (Fu et al., 2016). A *GRU* engloba informações de passos anteriores e da entrada corrente, de onde são selecionadas as informações a serem processadas. Em termos práticos, a vantagem da *GRU* sobre a *LSTM* está no seu custo computacional (Fu et al., 2016), e por este motivo foi utilizada no modelo proposto *MFA-RNN*. Por fim, a técnica *Dropout* é aplicada para evitar sobre-ajuste.

3.5.3 *Multi-Head Self-Attention* (*MHSA*)

O mecanismo *Self-Attention* foi inicialmente desenvolvido para o problema de Tradução de Máquina, onde se buscava compreender a correlação entre as palavras de uma sequência em um determinado idioma, com palavras de uma segunda sequência em outro idioma. No contexto deste trabalho, ao invés de se comparar sequências diferentes, as palavras da própria sequência são correlacionadas, e por isso o nome dado *Self-Attention*. A técnica *Multi-Head Self-Attention* introduzida em (Vaswani et al., 2017), aplica o *Self-Attention* sobre diferentes partes do vetor de entrada da camada,

sendo cada parte nomeada de *head*. A quantidade de *heads* que trouxe melhores resultados foi 4, indicando que existem 4 fatores que devem ser processados separadamente. Portanto, a partir da camada *MHSA* é possível “prestar atenção"sobre diferentes subespaços e fatores, correlacionando elementos de um mesmo vetor.

3.5.4 Englobando resultados

O mecanismo descrito na Seção 3.5.3 não consegue extrair correlações com base na ordem das sequências por si só (Vaswani et al., 2017). Por esse motivo, a noção de ordem das sequências, obtida pela camada recorrente (*GRU*), é propagada diretamente. Dessa forma, as saídas das camadas *MHSA*, *GRU*/*Dropout* e *Embedding* para o ID do usuário, são concatenadas em um tensor resultante. Assim, a rede neural pode aprender sobre o perfil de mobilidade de um determinado usuário sob a perspectiva de duas técnicas simultâneas, *MHSA* e *GRU*. Em seguida, é realizada a redução de dimensão dos dados com *Flatten*, e o *Dropout* é aplicado para evitar sobre-ajuste. Por fim, uma camada *Dense* é utilizada juntamente com a função de ativação *softmax* para prever os três tipos de localização possíveis para o próximo local de visita.

3.6 Resultados e Análises

3.6.1 Configuração

Foram conduzidos experimentos com o objetivo de comparar o desempenho da rede neural proposta MFA-RNN, com trabalhos bem conhecidos da literatura. Como a base de dados é desbalanceada, ou seja, não possui uma quantidade semelhante de eventos das três classes de localizações possíveis (Casa, Outro e Deslocamento), a utilização da métrica de acurácia não é a mais adequada. Dessa forma, o objetivo de cada abordagem se torna obter o melhor desempenho para prever o próximo local de visita mediante as métricas Precisão, Revocação e F1-score. Cada modelo foi testado sob o método de validação cruzada, dividindo-se a trajetória de cada usuário em 5 partes.

As abordagens *MHSA+PE* ((Zeng et al., 2019), *STF* ((Al-Molegi et al., 2016)), *SERM* ((Yao et al., 2017)) e *MAP* ((Al-Molegi et al., 2018)) são recentes e utilizaram uma grande variedade de técnicas em seus respectivos modelos. Portanto, elas foram selecionadas como bases de comparação com a solução proposta *MFA-RNN*.

Neste trabalho, foram conduzidos dois experimentos. No primeiro cenário, tanto o método proposto quanto as soluções base utilizam a base de dados esparsos. Para verificar o impacto positivo que o preenchimento dos dados traz, no segundo experimento a técnica de preenchimento (Ver Seção 3.3.2) é adotada por todas as soluções.

3.6.2 Avaliação Geral

O objetivo do primeiro experimento é avaliar o desempenho do modelo *MFA-RNN* e das *baselines* sob a base de dados original, ou seja, sem a aplicação do método proposto de preenchimento de dados. As Figuras 3.6, 3.7 e 3.8 exibem a Precisão, Revocação e *F1-score* de cada método, para os três tipos de localização existentes.

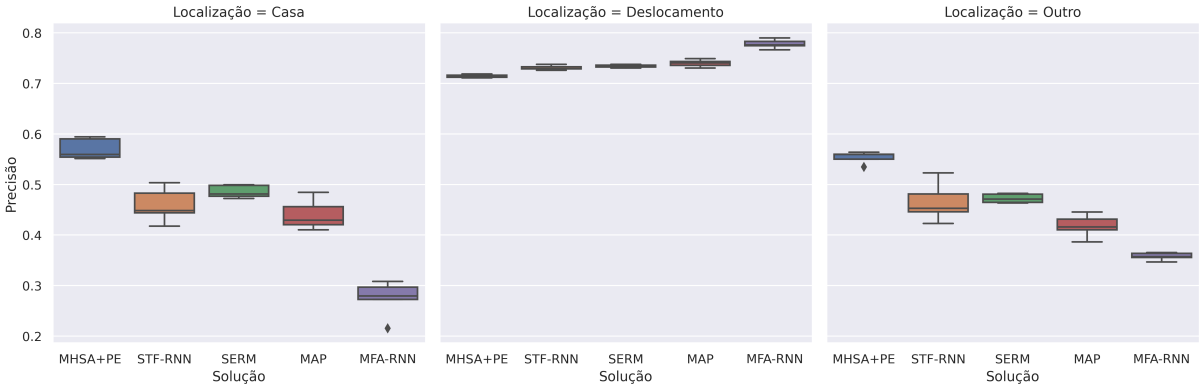

Figura 3.6: Precisão por localização (dados originais).

Apesar de o desempenho para se prever quando um usuário estará em *Deslocamento* ser ligeiramente inferior ao das soluções base, o *F1-score* da proposta *MFA-RNN* se manteve acima de 70%. Com relação à previsão dos PoIs *Casa* e *Outro*, as *baselines*, em geral, apresentam alta Precisão e baixa Revocação, o que afeta negativamente no valor final do *F1-score*.

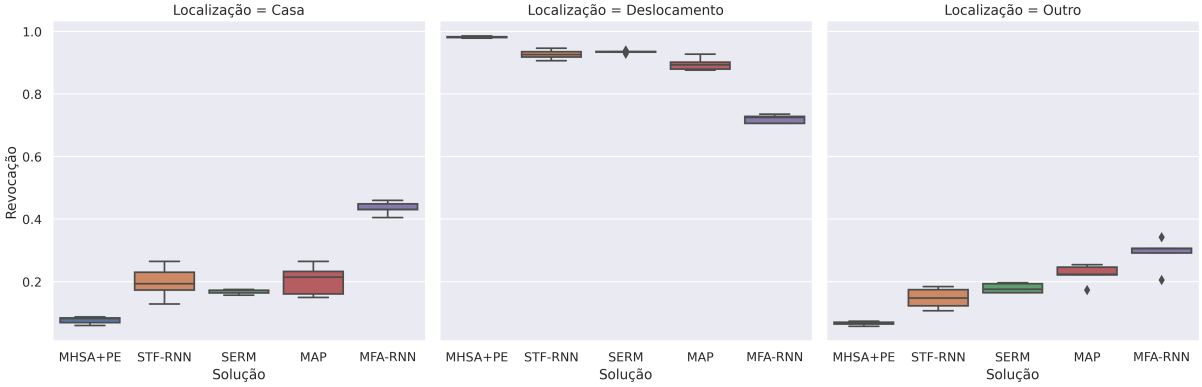

Figura 3.7: Revocação por localização (dados originais).

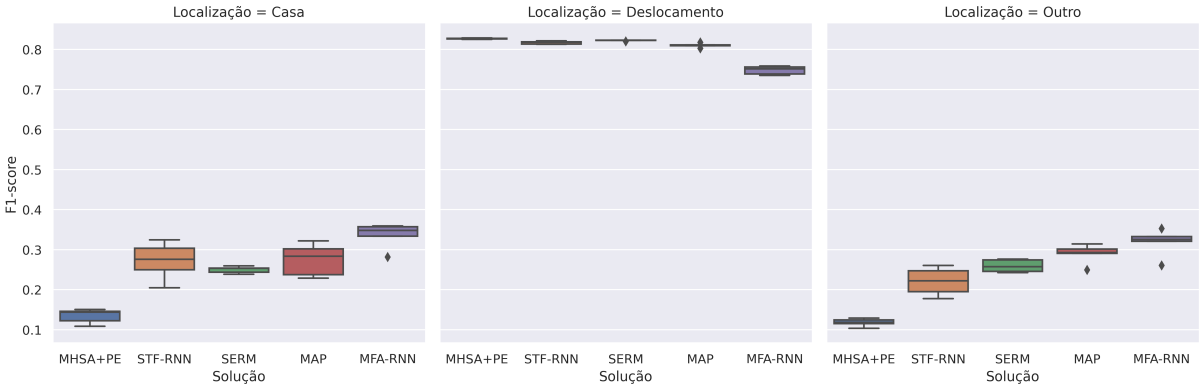

Figura 3.8: F1-score por localização (dados originais).

Ao mesmo tempo, os resultados indicam que o modelo *MFA-RNN* apresentou melhorias justamente onde as soluções base são deficitárias, ou seja, no *F1-score* de

predição dos locais *Casa* e *Outro*. Dessa forma, é possível observar o ganho obtido com a arquitetura proposta, destacando-se a utilização da camada *MHSA*, e de múltiplos fatores como entrada pela rede neural, em especial através da *feature* “tipo do dia".

3.6.3 Impacto do Preenchimento dos Dados

Neste experimento, todos os métodos são comparados sob a mesma base de dados preenchida. Assim, é possível avaliar o impacto do preenchimento de dados no desempenho das soluções. As Figuras 3.9, 3.10 e 3.11 exibem os desempenhos de cada método de acordo com as métricas Precisão, Revocação e *F1-score*, respectivamente.

O preenchimento de dados possibilitou que os desempenhos das soluções base alcançassem valores acima de 70% de *F1-score* para prever o PoI Casa. Juntamente com a melhoria na predição de Casa, as soluções base se aproximaram de 30% de *F1-score* para prever a localização do tipo Outro (exceto a solução *MHSA+PE*). Por outro lado, o desempenho para Deslocamento foi levemente reduzido entre essas abordagens, e, neste caso, o valor final de *F1-score* se aproximou da proposta *MFA-RNN*. Apesar de a solução *MHSA+PE* possuir a camada *MHSA*, a arquitetura apresentada não foi capaz de gerar bons resultados em ambos os cenários avaliados. A camada implementada *PE* (*Positional Encoding*) não contribuiu com a previsão do PoI Outro. A abordagem similar *MAP* utilizou um mecanismo simples de *Attention*, e optou por propagar a noção de ordem dos elementos da sequência diretamente pela camada recorrente em detrimento da utilização de *PE*, o que trouxe melhores resultados. Essa intuição é aproveitada no presente trabalho, mas com o diferencial de se utilizar a técnica mais recente de *Multi-Head Self-Attention*.

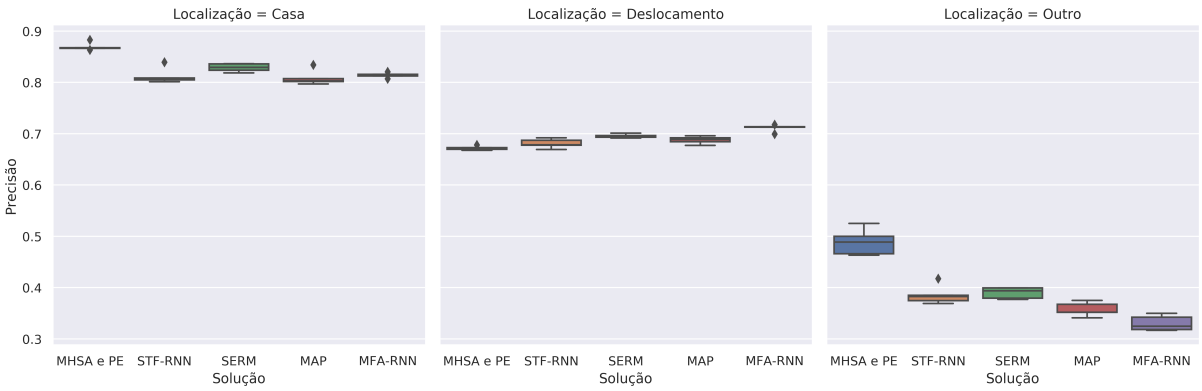

Figura 3.9: Precisão por localização (dados preenchidos).

Além disso, neste experimento o modelo *MFA-RNN* também se destaca quando o próximo local de visita é do tipo Outro, ao alcançar cerca de 35% de *F1-score*. A eficácia do método para Deslocamento é similar às soluções base, e portanto, mesmo aplicando preenchimento de dados em todas as abordagens, a rede neural proposta *MFA-RNN* apresenta um desempenho geral superior.

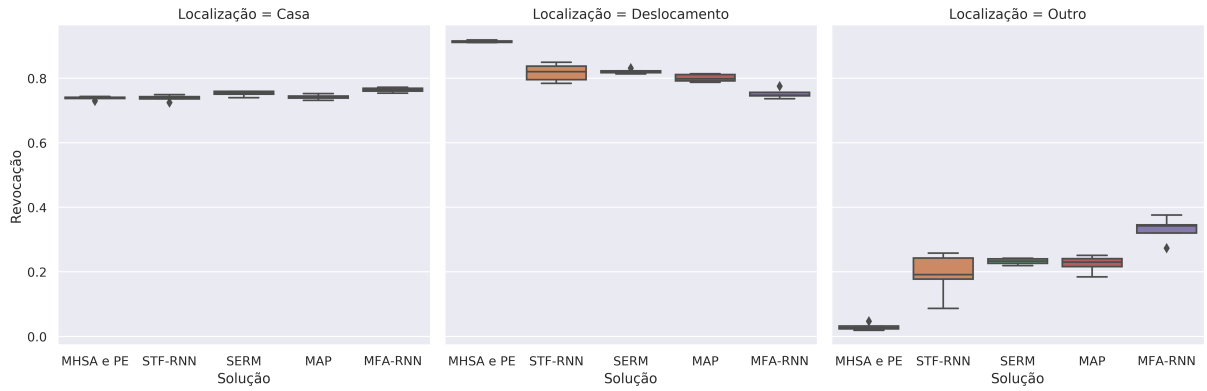

Figura 3.10: Revocação por localização (dados preenchidos).

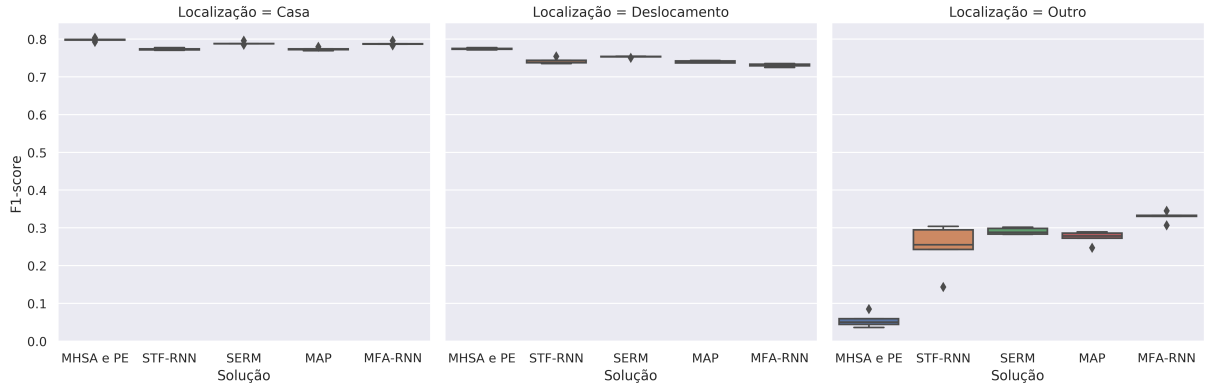

Figura 3.11: F1-score por localização (dados preenchidos).

Tanto no experimento anterior como no atual, a utilização de múltiplos dados de entrada gerou resultados mais consistentes (menor dispersão de valores de *F1-score*) para as soluções *MFA-RNN* e *SERM*. Comparando-se *SERM* com *STF-RNN* (cuja dispersão de *F1-score* é alta), as principais diferenças entre essas abordagens são que em *SERM* foi utilizada como entrada o ID do usuário, e o horário foi dividido em 48 partes (24 horas para dia de semana e 24 horas para finais de semana), funcionando analogamente à entrada de tipo do dia.

Por último, conclui-se que o método de preenchimento de dados pode contribuir com o desempenho tanto das soluções base quanto da rede neural apresentada. Em especial, ocorreram melhorias na predição dos PoIs Outro e, principalmente, Casa.

3.7 Conclusões e Trabalhos Futuros

Neste trabalho, foi apresentada a rede neural *MFA-RNN* para a predição de próximo local de visita de usuários móveis. O modelo se destaca por utilizar, ao mesmo tempo, múltiplos fatores de entrada da rede neural e o mecanismo de estado da arte *Multi-Head Self-Attention*. Com o objetivo de melhorar o desempenho do método desenvolvido, também foi apresentada uma abordagem para o preenchimento de dados esparsos, que é capaz de inferir quando um usuário esteve no PoI Casa. O modelo *MFA-RNN* obteve melhores resultados que as soluções base principalmente quando o

próximo local de visita for do tipo Casa ou Outro. Os experimentos também demonstraram que o método de preenchimento de dados também pode ser utilizado para melhorar o desempenho de outras soluções da literatura. Além disso, nossos testes foram realizados a partir de uma base de dados contendo 5.272 usuários, superando os principais trabalhos da literatura que utilizam bases de dados esparsas, contendo dados georreferenciados de alta precisão e que foram coletados passivamente.

Como trabalhos futuros, poderão ser previstas mais categorias de PoI como Lazer, Hospedagem e Saúde, por exemplo, a fim de enriquecer os resultados. Dessa forma, o modelo *MFA-RNN* poderá prever próximos locais de visita de tipos mais específicos, o que ampliará ainda mais a sua efetividade.

Agradecimentos

Este trabalho contou com o apoio da CAPES, CNPq e Fapemig.

Capítulo 4

Conclusões Gerais e Trabalhos Futuros

A pesquisa desenvolvida apresentou duas contribuições científicas para a literatura na área de análise da mobilidade humana. As soluções apresentadas são focadas em dados esparsos, e por isso são mais apropriadas para serem utilizadas em cenários nos quais os desempenhos dos dispositivos móveis devem ser preservados. A primeira solução corresponde a um método de identificação e classificação de pontos de interesse individuais de usuários móveis, e a segunda corresponde a uma abordagem para prever o próximo PoI que um usuário possivelmente irá visitar dada uma sequência de localizações prévias.

O artigo contido no Capítulo 2 apresentou os métodos de identificação de classificação de pontos de interesse (Casa ou Trabalho) utilizando dados esparsos de 194 usuários, a maior quantidade entre as soluções avaliadas. Comparando-se com os principais métodos da literatura, a solução apresentada alcançou melhorias de pelo menos 13% no *f-score* para a identificação de pontos de interesse. Para a classificação do tipo do PoI em Casa e Trabalho, contabilizou-se a quantidade de eventos gerados em intervalos específicos de horário com base no intervalo de inatividade. Esse conceito se mostrou eficaz diante de soluções bem conhecidas da literatura, alcançando melhorias de pelo menos 10% e 4% para a classificação dos PoI Casa e Trabalho, respectivamente.

O trabalho presente no Capítulo 3 corresponde a uma solução para prever a mobilidade humana juntamente com uma abordagem para o preenchimento de dados esparsos. A rede neural desenvolvida engloba as principais técnicas do estado da arte utilizadas para o problema de predição de próximo local de visita. Múltiplos fatores de entrada para o modelo são considerados (localização, tempo, identificador do usuário e tipo do dia) ao mesmo tempo em que se utiliza o mecanismo *MHSA* (*Multi-Head Self-Attention*) para se extrair correlações entre fatores. Os resultados obtidos demonstram um significativo avanço perante os mais recentes métodos de predição de próximo local de visita, especialmente onde essas soluções têm um desempenho deficitário, como para prever quando o próximo local de visita é do tipo Casa ou Outro. Além disso, demonstrou-se que o método de preenchimento de dados não é apenas capaz de contribuir com o desempenho geral da rede neural *MFA-RNN*, mas também com as soluções base avaliadas. Os testes foram conduzidos sobre uma base de dados de 5.272 usuários, sendo o maior valor entre as abordagens atuais considerando-se o método de coleta passivo, o que torna ainda mais consistentes os resultados.

Como trabalhos futuros, pretende-se principalmente avaliar o modelo *MFA-RNN* a partir de melhorias na eficácia do método de detecção de pontos de interesse. As possíveis contribuições estão descritas a seguir:

- Melhorar a acurácia da classificação do PoI Trabalho.

- Detalhar melhor a semântica dos pontos de interesse em outros tipos como *shopping*, restaurante, academia, aeroporto, dentre outros.
- A partir da proposta de contribuição descrita no item anterior, é possível avaliar o modelo *MFA-RNN* para a previsão de uma maior variedade de próximos locais de visita.
- Avaliar os métodos apresentados utilizando outras bases de dados esparsos.

Referências Bibliográficas

Al-Molegi, A., Jabreel, M., and Ghaleb, B. (2016). Stf-rnn: Space time features-based recurrent neural network for predicting people next location. In *2016 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1–7. IEEE.

Al-Molegi, A., Jabreel, M., and Martínez-Ballesté, A. (2018). Move, attend and predict: An attention-based neural model for people's movement prediction. *Pattern Recognition Letters*, 112:34–40.

Al-Molegi, A. and Martínez-Ballesté, A. (2018). The effect of space-time representation learning in predicting people's next location. In *CCIA*, pages 64–73.

Banovic, N., Buzali, T., Chevalier, F., Mankoff, J., and Dey, A. K. (2016). Modeling and understanding human routine behavior. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 248–260. ACM.

Capanema, C. G. S., Silva, F. A., and Silva, T. R. M. B. (2019). Identificação e classificação de pontos de interesse individuais com base em dados esparsos. *Anais do XXXVII Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos - SBRC 2019*, pages 16–29.

Castro, P. S., Zhang, D., and Li, S. (2012). Urban traffic modelling and prediction using large scale taxi gps traces. In *International Conference on Pervasive Computing*, pages 57–72. Springer.

Csáji, B. C., Browet, A., Traag, V. A., Delvenne, J.-C., Huens, E., Van Dooren, P., Smoreda, Z., and Blondel, V. D. (2013). Exploring the mobility of mobile phone users. *Physica A: statistical mechanics and its applications*, 392(6):1459–1473.

Cuttone, A., Lehmann, S., and Larsen, J. E. (2014). Inferring human mobility from sparse low accuracy mobile sensing data. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, pages 995–1004. ACM.

Feng, J., Li, Y., Zhang, C., Sun, F., Meng, F., Guo, A., and Jin, D. (2018). Deepmove: Predicting human mobility with attentional recurrent networks. In *Proceedings of the 2018 World Wide Web Conference*, pages 1459–1468. International World Wide Web Conferences Steering Committee.

Frias-Martinez, V., Virseda, J., Rubio, A., and Frias-Martinez, E. (2010). Towards large scale technology impact analyses: Automatic residential localization from mobile phone-call data. In *Proceedings of the 4th ACM/IEEE international conference on information and communication technologies and development*, page 11. ACM.

Fu, R., Zhang, Z., and Li, L. (2016). Using lstm and gru neural network methods for traffic flow prediction. In *2016 31st Youth Academic Annual Conference of Chinese Association of Automation (YAC)*, pages 324–328. IEEE.

Gao, J., Sun, Y., Liu, W., and Yang, S. (2016). Predicting traffic congestions with global signatures discovered by frequent pattern mining. In *2016 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)*, pages 554–560. IEEE.

Hoteit, S., Chen, G., Viana, A., and Fiore, M. (2016). Filling the gaps: On the completion of sparse call detail records for mobility analysis. In *Proceedings of the Eleventh ACM Workshop on Challenged Networks*, pages 45–50. ACM.

Isaacman, S., Becker, R., Cáceres, R., Kobourov, S., Martonosi, M., Rowland, J., and Varshavsky, A. (2011). Identifying important places in people's lives from cellular network data. In *International Conference on Pervasive Computing*, pages 133–151. Springer.

Järv, O., Ahas, R., and Witlox, F. (2014). Understanding monthly variability in human activity spaces: A twelve-month study using mobile phone call detail records. *Transportation Research Part C: Emerging Technologies*, 38:122–135.

Kung, K. S., Greco, K., Sobolevsky, S., and Ratti, C. (2014). Exploring universal patterns in human home-work commuting from mobile phone data. *PloS one*, 9(6):e96180.

Lee, S., Choi, Y., Lim, S., and Park, J. (2015). A spatio-temporal distance based clustering approach for discovering significant places from trajectory data.

Montoliu, R., Blom, J., and Gatica-Perez, D. (2013). Discovering places of interest in everyday life from smartphone data. *Multimedia tools and applications*, 62(1):179–207.

Naboulsi, D., Fiore, M., Ribot, S., and Stanica, R. (2016). Large-scale mobile traffic analysis: a survey. *IEEE Communications Surveys & Tutorials*, 18(1):124–161.

Pavan, M., Mizzaro, S., Scagnetto, I., and Beggiato, A. (2015). Finding important locations: A feature-based approach. In *Mobile Data Management (MDM), 2015 16th IEEE International Conference on*, volume 1, pages 110–115. IEEE.

Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Ranjan, G., Zang, H., Zhang, Z.-L., and Bolot, J. (2012). Are call detail records biased for sampling human mobility? *ACM SIGMOBILE Mobile Computing and Communications Review*, 16(3):33–44.

Rathore, M. M., Ahmad, A., Paul, A., and Rho, S. (2016). Urban planning and building smart cities based on the internet of things using big data analytics. *Computer Networks*, 101:63–80.

Schneider, C. M., Belik, V., Couronné, T., Smoreda, Z., and González, M. C. (2013). Unravelling daily human mobility motifs. *Journal of The Royal Society Interface*, 10(84):20130246.

Schreckenberger, C., Beckmann, S., and Bartelt, C. (2018). Next place prediction: A systematic literature review. In *Proceedings of the 2nd ACM SIGSPATIAL Workshop on Prediction of Human Mobility*, pages 37–45.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell system technical journal*, 27(3):379–423.

Trestian, I., Ranjan, S., Kuzmanovic, A., and Nucci, A. (2009). Measuring serendipity: connecting people, locations and interests in a mobile 3g network. In *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement*, pages 267–279. Acm.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Wei, L.-Y., Zheng, Y., and Peng, W.-C. (2012). Constructing popular routes from uncertain trajectories. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 195–203. ACM.

Yao, D., Zhang, C., Huang, J., and Bi, J. (2017). Serm: A recurrent model for next location prediction in semantic trajectories. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 2411–2414. ACM.

Yao, Z., Fu, Y., Liu, B., Liu, Y., and Xiong, H. (2016). Poi recommendation: A temporal matching between poi popularity and user regularity. In *2016 IEEE 16th international conference on data mining (ICDM)*, pages 549–558. IEEE.

Yu, D., Li, Y., Xu, F., Zhang, P., and Kostakos, V. (2018). Smartphone app usage prediction using points of interest. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(4):1–21.

Zeng, J., He, X., Tang, H., and Wen, J. (2019). A next location predicting approach based on a recurrent neural network and self-attention. In *International Conference on Collaborative Computing: Networking, Applications and Worksharing*, pages 309–322. Springer.