Universidade Federal
de Campina Grande

UNIVERSIDADE FEDERAL DE CAMPINA GRANDE

Centro de Engenharia Elétrica e Informática

Programa de Pós-Graduação em Ciência da Computação

Luiz Henrique de Andrade

TrajTax: Uma Taxonomia para o Domínio de Trajetórias

Campina Grande – PB

2020

Universidade Federal de Campina Grande
Centro de Engenharia Elétrica e Informática
Coordenação de Pós-Graduação em Ciência da Computação

TrajTax: Uma Taxonomia para o Domínio de Trajetórias

Luiz Henrique de Andrade

Dissertação submetida à Coordenação do Curso de Pós-Graduação em Ciência da Computação da Universidade Federal de Campina Grande - Campus I como parte dos requisitos necessários para obtenção do grau de Mestre em Ciência da Computação.

Área de Concentração: Ciência da Computação
Linha de Pesquisa: Metodologia e Técnicas da Computação

Cláudio de Souza Baptista, Ph.D.
(Orientador)

Campina Grande, Paraíba, Brasil

A553t Andrade, Luiz Henrique de.
TrajTax: uma taxonomia para o domínio de uma trajetórias / Luiz Henrique de Andrade. - Campina Grande, 2020.
92f. : il. Color.

Dissertação (Mestrado em Ciência da Computação) - Universidade Federal de Campina Grande, Centro de Engenharia Elétrica e Informática, 2020.
"Orientação: Prof. Dr. Cláudio de Souza Baptista".
Referências.

1. Taxonomia de Atributos. 2. Trajetórias. 3. Interpretabilidade. 4. Engenharia de Atributos. 5. Aprendizagem de Máquina. I. Baptista, Cláudio de Souza. II. Título.

CDU 004.89(043)

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECÁRIA ITAPUANA SOARES DIAS CRB-15/93

TRAJTAX: UMA TAXONOMIA PARA O DOMÍNIO DE TRAJETÓRIAS

LUIZ HENRIQUE DE ANDRADE

DISSERTAÇÃO APROVADA EM 17/08/2020

CLÁUDIO DE SOUZA BAPTISTA, PhD., UFCG
Orientador(a)

CARLOS EDUARDO SANTOS PIRES, Dr., UFCG
Examinador(a)

AMILCAR SOARES JUNIOR, Dr.
Examinador(a)

CAMPINA GRANDE - PB

Resumo

O aumento na quantidade de dados disponíveis relacionados a trajetórias viabiliza inúmeros trabalhos em diferentes domínios: trajetórias de pedestres, animais, navios, aviões, dentre outros. Com isso, diversas aplicações de aprendizagem de máquina têm sido propostas com base nos dados. Para o sucesso destas aplicações, é importante o entendimento do domínio do problema, das técnicas disponíveis e dos atributos que são utilizados como entrada. A combinação destes pilares é a base para que as pesquisas possam evoluir e chegar a bons resultados. Muito se é discutido sobre as técnicas disponíveis e diferentes domínios, todavia, pouca atenção é dada aos atributos utilizados.

Outro ponto a ser considerado no contexto de trajetórias é que a interpretabilidade dos resultados gerados é importante. A necessidade de um modelo interpretável gera uma restrição quanto à complexidade dos atributos utilizados no processo de aprendizagem de máquina. A opção de geração automática de atributos torna-se inviável para diferentes domínios devido à complexidade dos atributos resultantes. Desta forma, o processo de engenharia de atributos depende mais de atributos projetados por especialistas que são baseados na teoria atrelada ao contexto.

Esta pesquisa propõe uma taxonomia, denominada TrajTax, que classifica atributos relacionados à trajetórias, utilizados no estado da arte. A taxonomia tem como propósito auxiliar no processo de engenharia de atributos, servindo como base para especialistas discutirem sobre atributos que são utilizados atualmente e para proporem novos atributos de acordo com a necessidade do projeto. Para o desenvolvimento da taxonomia, foi feito o levantamento dos atributos utilizados em trabalhos baseados em dados de trajetórias com diferentes domínios e utilizou-se uma metodologia de construção de taxonomias. Os atributos são definidos com base em suas definições na literatura. A taxonomia proposta consiste de atributos de compreensão simplificada, para que possam auxiliar na interpretabilidade do modelo. Assim, como contribuições deste trabalho, tem-se o levantamento dos atributos utilizados em trabalhos na área de trajetória em diferentes domínios e a criação da taxonomia TrajTax.

Palavras chave: Trajetórias, Taxonomia, Interpretabilidade, Engenharia de atributos, Aprendizagem de máquina

Abstract

The large volume of data concerning trajectories enables research on different domains including trajectories of pedestrians, animals, ships, and airplanes. As a result, several machine learning applications have been proposed based on these data. To achieve the success of these applications, it is important to understand the application domain, the available techniques, and the various input features. This comprehension is fundamental for the research to evolve and obtain good results. Much has been discussed on the available techniques and their different domains; however, research on features has received little to none of the attention.

Another point to be considered regarding trajectories is that in some domains the interpretability of results must be rated as most important. The need for an interpretable model creates a constraint regarding the complexity of the features used along the machine learning process. The option of automatic features generation is impracticable for different domains. This is due to the complexity of the resulting features. Thus, the feature engineering process depends more on features designed by specialists. These features are based on theories regarding their various contexts.

We propose a taxonomy called TrajTax that classifies trajectory features used in the state-of-the-art. The taxonomy is intended to help researchers along the feature engineering process; this will serve as a basis for specialists to discuss the features which are currently used and to suggest new ones. For the development of the taxonomy, we conducted a survey on works based on trajectory features along different domains, and used a methodology for taxonomy construction. The features were outlined based on their definitions proposed in the state-of-the-art. The Trajtax consists of a simplified comprehension concerning the features so it can help with the model interpretability. Consequently, the main contributions of the present work encompass a survey of the features used in trajectory machine learning applications on different domains, and the creation of the TrajTax taxonomy.

Keywords: Trajectories, Taxonomy, Interpretability, Feature Engineering, Machine Learning

Agradecimentos

Primeiramente, agradeço a Deus por tudo. Sem Ele nada disso teria sentido. Que o Senhor continue abençoando meu trilhar e me guiando nos caminhos do bem. Obrigado por me dar o que nunca imaginei e por ouvir pacientemente minhas reclamações no caminho.

Ao professor Cláudio, que tem um papel muito importante na minha vida acadêmica. Muitíssimo obrigado professor por todas as oportunidades e ensinamentos durante essa jornada que começou na graduação. Durante esse tempo, houveram muitos "Desculpa aí, professor", dos quais em cada um deles ficava um ensinamento para mim. Até agora estou na dívida quanto à crase, mas eu chego lá. Eu sou extremamente grato por tudo.

Aos professores, membros da banca, Carlos Eduardo e Amílcar Soares, pelos comentários e pela contribuição realizada por meio de sugestões e conhecimentos. À todos da Dalhousie University, em especial ao professor Luis Torgo, que me orientou durante o período que passei no Canadá. Aproveitando, gostaria de agradecer a todos os professores que contribuíram para a minha formação. Gostaria de agradecer também aos funcionários da COPIN que me auxiliaram com todas as dúvidas - que não foram poucas - durante esse processo.

Aos meus pais, Marta Maria e José Ribamar, por todo o apoio e educação. Com certeza, apesar de ter tido uma educação formal muito boa, o que aprendi de melhor na vida foi através do exemplo dentro de casa pelos atos de amor, carinho e atenção. Em nome deles, quero agradecer a toda minha família que é a minha base e me ajudou de todas as formas possíveis para que eu chegasse até aqui.

A meu amor Pollyanne, muito obrigado por me acalmar e acreditar em mim. Por me apoiar e me ajudar quando tudo o que eu queria era que acabasse. Você é demais. Eu te amo muito. Muito obrigado por toda paciência durante esta fase e por me ajudar a superar tudo isso. A todos do Laboratório de Sistemas de informação. Durante muito tempo estamos juntos nas alegrias e nos perrengues, mas sempre que dá, com salgadinhos na mesa. Aprendi demais com todos vocês. Gostaria de agradecer em especial a Anderson pela ajuda com o português. A todos os meus amigos, desculpem a ausência algumas vezes. Estamos sempre juntos.

À CAPES pelo incentivo e suporte financeiro.

Conteúdo

Lista de Abreviaturas e Siglas

Acc - *Aceleração*

AIS - *Automatic Identification Systems*

CB-Smot - *Clustering-Based Stops and Moves of Trajectories*

cm - *Centímetros*

COG - *Course over ground*

Convex3 - *Convexity*

COSAspct - *Cosine of aspect*

CSV - *Comma-Separated Values*

DBSCAN - *Density-Based Spatial Clustering of Applications with Noise*

DistCLSD - *Distance to the nearest closed road*

DistCWat - *Distance to the nearest water source from within a cattle pasture*

DistEdge - *Distance to nearest edge*

DistEFnc - *Distance to the nearest ungulate-proof fence*

DistEWat - *Distance to the nearest water source from within an ungulate-proof pasture*

DistOPEN - *Distance to the nearest open road*

DistRSTR - *Distance to the nearest restricted access road*

Elev - *Elevation*

ForgProd - *Forage production*

GPS - *Global Positioning System*

Hc - *Head Change*

m - *Metros*

m^2 - *Metros Quadrados*

m/s - *Metros por Segundos*

m/s^2 - *Metros por Segundos ao Quadrado*

MDS - *MultiDimensional Scaling*

Obswt - *Observation weight*

PerSlope - *Percent Slope*

POI - *Point of Interest*

Rad/s - *Radiano por segundo*

RadNum - *Radio collar number*

ROI - *Region of Iterest*

ROT - *Rate of Turn*

RPM - *Rotações por minuto*

SINAspct - *Sine of aspect*

SOG - *Speed Over Ground*

SoilDpth - *Soil depth*

SVM - *Support Vector Machine*

USDA - *United States Department of Agriculture*

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Iniciaremos o nosso trabalho explicando que uma trajetória consiste na representação do movimento de um objeto e que, apesar da natureza contínua das trajetórias, esse movimento é representado em aplicações e pesquisas de forma discreta por um conjunto de pontos espaço-temporais [53]. Neste sentido, de acordo com Zheng [79], as trajetórias podem estar associadas à mobilidade de pessoas, animais, veículos de transporte e fenômenos naturais. Todavia, em cada um desses casos, o objeto móvel é diferente e tem características de movimento distintas. Assim, o estudo de dados de trajetórias possui diversas vertentes de pesquisas e aplicações.

Com o desenvolvimento tecnológico, a aquisição de dados de trajetória transformou-se em um processo mais simples e de baixo custo. Dessa forma, há uma grande quantidade de dados de trajetórias disponíveis; o que viabiliza a utilização de técnicas de aprendizagem de máquina para uma melhor análise descritiva e preditiva do material em questão. Como consequência dos avanços de tecnologia, surgiu o desenvolvimento de pesquisas em diferentes domínios, por exemplo: predição de rota e de destino [41], detecção de trajetória incomum em uma região de interesse (permanecer ao redor, fuga ou retorno) [4], detecção de rotas maiores que a rota padrão para táxis [69], detecção de anomalias no tráfego a longo prazo [31], cálculo dos riscos de acidente de trânsito [46], detecção do tipo de transporte utilizado durante a trajetória [78], detecção de rota de pesca que auxilia na prevenção dessa atividade em lugares irregulares [13], predição de rotas de embarcações que auxiliam no processo de evitar colisão [48], detecção de anomalias em rotas de embarcações que ajudam na fiscalização para verificar se elas estão de acordo com as normas estabelecidas por tipo de

embarcação [42], entre outros.

As trajetórias são objeto de interesse em aplicações de aprendizagem de máquina de diferentes domínios. Dentre os exemplos citados, fica clara a importância dessas aplicações em problemas reais, como a gestão de tráfego facilitando tomadas de decisão. Um exemplo útil que podemos citar é: se for possível prever congestionamentos, baseado em trajetórias anteriores, o órgão gestor pode preparar alternativas para determinada via, antes de existir um problema crônico.

A combinação dos dados de trajetórias com outras fontes (e.g. informações climáticas nos pontos relativos à trajetória) geram uma grande gama de atributos que podem ser utilizados. Todavia, nem sempre um alto número de atributos gera resultados melhores, pois, em alguns casos, a utilização deles leva a uma piora nos resultados em comparação com modelos treinados com um subconjunto desses atributos. Desta forma, é necessário adequar os dados para que possam ser utilizados da melhor forma com relação ao problema representado pelo modelo. Esse processo de adequação dos dados para obtenção de melhores resultados é conhecido como engenharia de atributos. Ozdemir e Sursala [45] definem o termo como o processo de transformação dos dados em atributos mais representativos do problema para a obtenção de melhores resultados em processos de aprendizagem de máquina. Tal processo consiste na aplicação de um conjunto de técnicas para a obtenção do melhor conjunto de atributos a ser utilizado nas atividades de aprendizagem de máquina.

Apesar do desenvolvimento e da grande quantidade de dados disponíveis na área, a falta de interpretabilidade dos modelos em alguns domínios ainda é um desafio em aberto [7]. Na prática, quando se lida com trajetórias, muitas das vezes busca-se detectar um problema e dar soluções para ele, por exemplo: se uma embarcação está em velocidade não permitida naquela determinada região, um modelo de detecção de anomalia cria um alerta. Todavia, para que essa solução seja de fato eficaz, o alerta deve ser específico quanto à causa da anomalia, pois o responsável pela fiscalização não pode apenas afirmar ao comandante que tem algo errado com a trajetória da embarcação. Em casos como esse, é necessário que o modelo seja suficientemente interpretável, a fim de que sejam obtidas conclusões claras.

Um primeiro passo para diminuição do problema de interpretabilidade do modelo é a compreensão dos dados que são passados como entrada para os modelos. Os atributos associados às trajetórias variam de acordo com o domínio (e.g. atributos utilizados para trajetó-

rias de animais, como proximidade a uma fonte de água, não são utilizados para trajetórias de táxis); contudo, mesmo trajetórias de domínios diferentes contêm traços em comum. Assim, um ponto importante a ser levado em consideração para lidar com o problema de interpretabilidade é o entendimento dos atributos utilizados. Para se ter uma ideia geral sobre eles, é importante levar em consideração os traços comuns entre esses atributos e como representam as trajetórias.

A forma como os atributos são gerados influencia na interpretabilidade deles, e existem diferentes maneiras para lidar com sua geração, dentre elas, está o uso de algumas técnicas de geração automática de atributos, como aplicação de transformações dos dados ou treinamento dos modelos de aprendizagem profunda e a utilização dos valores da penúltima camada como entrada para os algoritmos. No contexto de trajetórias, essa abordagem torna-se inviável, porque o entendimento acerca do atributo é um ponto importante e os que forem gerados através de transformações automáticas são, de forma geral, valores de difícil interpretação. Por conta disso, a criação manual de atributos é a opção mais viável, porém, essa técnica demanda recursos humanos e especialistas de domínio, levando em consideração a situação circunstancial específica do processo de criação do atributo.

Nesse cenário apresentado, um problema, entretanto, é que a criação manual de atributos demanda o conhecimento aprofundado sobre os dados e os padrões de comportamento relacionados ao objeto em movimento e esse problema não é presente apenas no contexto de trajetórias. Fromm, Wambsganss e Söllner [20] afirmam que muitos dos atributos utilizados no caso da mineração de texto são gerados manualmente e dependem do conhecimento do projetista para que gere bons resultados. Os autores propõem, então, a criação de uma taxonomia. Com isso, pesquisadores podem ter uma base sobre o que deve ser levado em consideração durante o processo de geração de atributos para seja utilizado neste contexto. Desta forma, facilita-se a proposição de novos atributos, aumentando as chances de sucesso deles.

Levando estes pontos em consideração, o presente trabalho apresenta os principais atributos relacionados a trajetórias, conforme encontrados na literatura, além de trazer uma taxonomia criada com bases nos mesmos, denominada TrajTax. O processo até a criação da taxonomia para o domínio de trajetórias é descrito em detalhes e aborda pontos importantes associados aos atributos, servindo como referência para futuras pesquisas na área.

De forma simples, taxonomias podem ser definidas como formas de organização classificatórias, considerando princípios associados aos objetos classificados [59]. A taxonomia TrajTax aqui proposta é realizada com base na metodologia orientada por Nickerson, Varshney e Muntermann [43]. O processo de criação da taxonomia passa por diversas etapas, sendo que uma etapa de grande importância é o levantamento dos atributos presentes na literatura. A taxonomia resultante abrirá espaço para discussões e facilitará o entendimento de novos pesquisadores na área em relação aos atributos que já foram utilizados e o que deve ser levado em conta durante o processo de geração de novos atributos.

Este trabalho tem como propósito, portanto, o auxílio para pesquisas que envolvem aplicação de métodos de aprendizagem de máquina em dados de trajetória para os quais a interpretabilidade do modelo resultante é uma condição necessária. Como contribuições desta pesquisa tem-se o levantamento dos atributos utilizados em trabalhos na área de trajetória em diferentes domínios, e a criação da taxonomia TrajTax. Os atributos levantados nessa dissertação são apresentados e definidos, e a taxonomia representa as características deles. A taxonomia TrajTax foi desenvolvida para ser utilizada tanto no processo de seleção de atributos, tanto quando se busca por eles de maneira que se adéquam ao contexto do trabalho, quanto na proposição de novos atributos, destacando os pontos que devem ser observados.

1.1 Objetivos

Nesta seção, será discutido o objetivo geral desta pesquisa. Posteriormente, serão expostos os objetivos específicos deste trabalho.

1.1.1 Objetivo Geral

Este trabalho tem como objetivo o desenvolvimento de uma taxonomia para classificar os atributos interpretáveis associados aos dados de trajetórias. A taxonomia resultante representa as principais características destes atributos e pode auxiliar no processo de engenharia de atributos, dando uma visão geral dos que estão disponíveis, com suas características.

1.1.2 Objetivos específicos

Visando atingir o objetivo geral deste trabalho, é necessária a divisão do mesmo em objetivos específicos, que são definidos como:

- Realizar um levantamento dos atributos interpretáveis que são usados em diferentes domínios de trajetórias, o que é possível através do estudo dos trabalhos relacionados;
- Selecionar bases de dados de trajetórias a serem utilizadas durante o processo;
- Gerar atributos calculados a partir dos pontos espaço-temporais;
- Avaliar a taxonomia com relação aos critérios propostos na metodologia utilizada.

1.2 Relevância

Trajetórias são compostas por múltiplos pontos espaço-temporais. Isso implica em uma gama de possíveis atributos que podem ser utilizados em processos de aprendizagem de máquina. De acordo com Domingos [15], a diferença entre o sucesso e o fracasso de um modelo de aprendizagem de máquina está na escolha de atributos. Assim sendo, é importante que o pesquisador conheça as opções de atributos existentes para que, aplicando técnicas de engenharia de atributos, possa chegar a um subconjunto de atributos que melhor representa os dados para o determinado propósito.

Outro ponto a ser ponderado é que, para determinados domínios, é relevante que os resultados gerados por modelos sejam interpretáveis, e esse ainda é um desafio em aberto, importante em alguns domínios no contexto de trajetórias [7]. Além da escolha do algoritmo que facilite a interpretação, é necessário que os atributos passados como entrada do algoritmo sejam claros com relação aos seus significados. Em casos como esse, onde é necessário que os resultados sejam compreensíveis, é ainda mais indispensável o conhecimento do pesquisador sobre os atributos utilizados em seus modelos. Sobre isso, ressaltamos que muitos dos atributos utilizados na literatura são relacionados a métricas que foram propostas e que são baseadas em teorias que podem ser explicadas a outros pesquisadores.

Nos trabalhos associados a trajetórias presentes na literatura, muito se discute sobre os modelos aplicáveis aos dados de trajetória [79; 18]. Além disso, também é possível encontrar

definições sobre diferentes domínios associados a trajetórias. Todavia, pouca atenção é dada aos atributos que são utilizados em processos de aprendizagem de máquina. Os trabalhos na área dão pouco destaque aos atributos que são usados como entrada, enfatizando mais quais técnicas foram empregadas individualmente ou combinadas. Isso auxilia os pesquisadores quanto a factibilidade da aplicação que está sendo explorada, todavia, não eleva a discussão sobre os possíveis dados de entrada.

Visando assistir as atividades de aprendizagem de máquina com dados de trajetória, este trabalho propõe uma taxonomia baseada nos atributos relacionados à trajetória, além de auxiliar no desenvolvimento de novos atributos. Nesse sentido, o uso de taxonomia para esse propósito já foi abordado em outras áreas do conhecimento e pode ajudar também no processo de engenharia de atributos na área de trajetórias.

Com o levantamento dos atributos utilizados em trabalhos do estado da arte, o processo de escolha deles no processo de aprendizagem de máquina será facilitado. Com o foco em atributos interpretáveis, cada um deles é descrito de acordo com sua definição apresentada nos respectivos trabalhos científicos. Uma lista com atributos utilizados, com referências para os artigos que reportaram a utilização, e suas respectivas descrições, dá uma visão geral sobre as possibilidades existentes. Também é plausível afirmar que a lista que possa servir como base para a criação de novos atributos de acordo com a necessidade.

Outro ponto que requer destaque é a expansão da biblioteca Trajlib [17] para aproveitamento em distintos domínios. Essa biblioteca facilita na geração de atributos, para serem utilizados tanto em trabalhos que consideram todos os pontos, como nos que levam em consideração segmentos inteiros. A ferramenta pode ser usada ainda para segmentar trajetórias de acordo com o propósito do estudo. Também é possível encontrar dados de exemplo que são de livre utilização como a base de dados de trajetórias de animais empregadas neste trabalho.

A criação de uma taxonomia auxilia no entendimento dos atributos existentes. Ademais, abre espaço para discussões sobre as possibilidades de novos atributos e facilita a avaliação de outros que tenham estrutura similar aos que já são conhecidamente efetivos para o domínio em estudo. As informações estruturais podem ser de extrema importância na decisão de quais atributos escolher durante o processo de engenharia de atributos para que os melhores resultados sejam obtidos, pois é possível direcionar os esforços nas escolhas de atributos que se encaixam melhor no escopo do domínio.

É relevante que as informações associadas aos atributos possam ser facilmente compreendidas para que os resultados obtidos na área possam melhorar. Com isso, a geração de uma taxonomia, junto com o levantamento dos atributos presentes na literatura, auxilia na aceleração do desenvolvimento de pesquisas no domínio de trajetórias.

1.3 Organização Estrutural

Os próximos capítulos desta dissertação são organizados da seguinte forma: no capítulo 2, são introduzidos os conceitos essenciais utilizados nesta pesquisa, como: definições dos dados de trajetória, formas de representação de trajetórias, engenharia de atributos e aprendizagem de máquina, matriz de correlação e taxonomia. No capítulo 3, são apresentados os trabalhos relacionados a este estudo, dos quais foram retirados os atributos utilizados no processo de criação da taxonomia. No capítulo 4, a taxonomia Trajtax é exposta e com ela todos os passos no processo de levantamento dos atributos, bases de dados utilizadas, expansão da biblioteca Trajlib, geração de atributos e desenvolvimento da taxonomia. O capítulo 5, é, por fim, composto pelas considerações finais, com proposições para trabalhos futuros.

Capítulo 2

Fundamentação teórica

Neste capítulo são apresentados os principais conceitos que fundamentam o desenvolvimento deste trabalho. Para isso, o capítulo foi dividido em seções relacionadas às áreas de conhecimento. Na seção 2.1, são explicados os dados de trajetórias de forma geral e as suas possíveis representações. Na seção 2.2, o processo de aprendizagem de máquina é descrito. Na seção 2.3, o processo de engenharia de atributos é discutido, assim como sua aplicabilidade aos dados de trajetória. Na seção 2.4, é apresentado o conceito de matriz de correlação. Na seção 2.5, os conceitos associados a taxonomias são expostos, e na seção 2.6, é apresentada a metodologia de desenvolvimento de uma taxonomia utilizada nesta pesquisa. Por fim, na seção 2.7 são feitas as considerações finais.

2.1 Dados de trajetória

As atividades de aprendizagem de máquina com dados de trajetórias podem ser aplicadas em diferentes domínios (e.g. mobilidade urbana e mobilidade de animais). Apesar de existirem diferenças entre os domínios, os conceitos básicos de trajetória são aplicados de forma geral a todos. Nesta seção, são introduzidas as definições de alguns conceitos gerais associados a trajetórias que são importantes para o entendimento das atividades na área.

Para que uma trajetória seja originada, é necessário que haja um objeto em movimento que gere um rastro. Este objeto é tratado na literatura como um objeto móvel, do inglês *moving object*, que pode ser definido como um objeto que muda sua localização com o tempo, criando assim um rastro em forma de conjunto de dados espaço-temporais [53].O

tipo do objeto pode variar de acordo com o domínio dos dados (e.g. embarcações [48], animais [29], fenômenos naturais [32]).

Zheng [79] define uma trajetória (τ) como um rastro gerado por um objeto em movimento (O_{id}) que é representado como uma sequência ordenada cronologicamente de pontos espaço-temporais $\tau = [P_1, P_2, ..., P_n]$. Esta definição foi utilizada como base durante o desenvolvimento deste trabalho, pois descreve a trajetória com relação a todos os seus pontos, facilitando a generalização para diferentes contextos. Renso, Spaccapietra e Zimányi [53], em sua definição de rastreio de movimento, argumentam que não existem dois pontos coletados ao mesmo tempo e para cada ponto há um possível conjunto de atributos (X_i) que podem ser obtidos através de dispositivos. Na Figura 2.1 tem-se uma representação de uma trajetória $\tau = [P_1, P_2, ..., P_{10}]$ realizada pelo objeto móvel representado pelo id Oid.

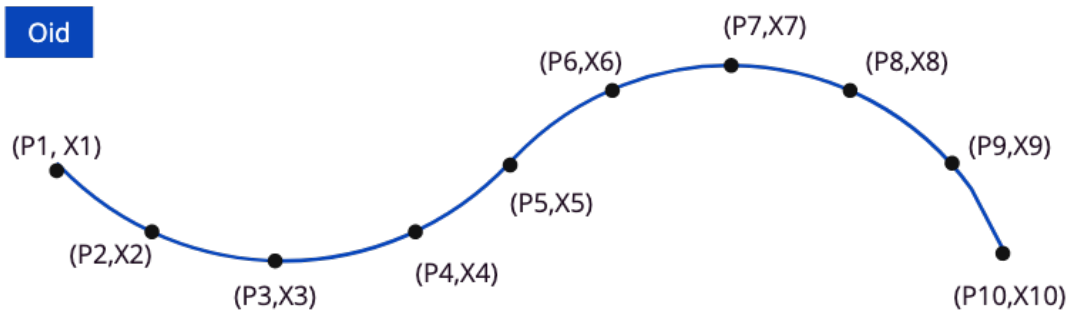

Figura 2.1: Exemplo de uma trajetória.

Como vista na definição de trajetória, a representação do movimento tem por base pontos. Um ponto espaço-temporal (P) é formado por uma marca temporal (t), uma localização geográfica (L) que tem a forma ($L = (latitude, longitude)$). Logo, o ponto P pode ser representado como: $P = (latitude, longitude, t)$ [79].

Outra questão que pode ser destacada na definição de trajetória é que podem existir atributos associados aos pontos espaço-temporais. De forma geral, um atributo é, como estabelecido por Chandrashekar e Sahin [5], uma propriedade individual de um objeto que está sendo observado que pode ser mensurada. Como visto na concepção de Renso, Spaccapietra e Zimányi [53], cada ponto da trajetória pode ou não ter atributos associados. Todavia, atributos não estão restritos apenas ao contexto de pontos quando tratamos de trajetórias; é possível que eles estejam associados a um objeto mais geral como a uma traje-

tória completa ou um segmento [78; 75; 74], ou mesmo com múltiplas trajetórias [12; 46; 57].

Trajetórias representam o movimento de um determinado objeto, todavia, existem casos em que ela contém diferentes comportamentos (e.g. uma pessoa pode sair caminhando e após um tempo pegar um ônibus). Desta forma, como ponderaram Renso, Spaccapietra e Zimányi [53] é viável a divisão das trajetórias em sub-trajetórias que estão associadas com um mesmo valor, de modo que esse processo é chamado de segmentação.

Assim, de acordo com o problema, pode ser feito o uso de trajetórias ou segmentos como entrada nas atividades de aprendizagem de máquina. Dessa maneira, um segmento ou sub-trajetória s é definido, de acordo com Lee, Han e Whang [32], como um subconjunto dos pontos espaço-temporais $< P_i, P_{i+1}, \ldots, P_j >$ de uma trajetória $\tau =< P_1, P_2, \ldots, P_n >$, de modo que $1 \leq i \leq j \leq n$.

2.1.1 Formas de representação de trajetórias

A forma pela qual os dados são estruturados na base de dados tem impacto no processo de aprendizagem de máquina, pois cada domínio relacionado com os dados de trajetórias, usualmente, tem restrições associadas e requer uma organização particular para esses elementos. A seguir, portanto, são apresentadas formas de estruturar os dados de trajetórias e comentados pontos importantes sobre essas abordagens.

Uma forma de estruturar informações de trajetórias em uma base de dados é salvando-as sobre todos os pontos espaço-temporais, que podem ser agrupados, ou não, em sub-trajetórias de acordo com o contexto. De Souza et al. [13] utilizam essa abordagem de usar os dados de todos os pontos para a criação dos modelos, separando-os em sub-trajetórias. Furtado et al. [22] lidam com os dados a partir de todos os pontos da trajetória para a geração de elipses que são utilizadas no cálculo de incerteza associado a uma trajetória. Apesar da completude dos dados, esse tipo de organização demanda um poder computacional maior devido ao grande número de pontos que são operacionalizados.

Outra maneira de trabalhar com os dados associados a pontos da trajetória é comprimindo a trajetória e usando apenas alguns pontos importantes delas. Essa abordagem é utilizada por Li et al. [34], onde os autores lidam com a redução de escala multidimensional, do inglês *multidimensional scaling* (MDS), para a geração de uma representação dos pontos reduzida.

Zheng [79] apresenta três tipos de compressão que podem ser aplicadas em trajetórias, que são: *online*, *offline* e contextual. No caso da compressão online, decide-se durante a coleta dos dados manter ou não o ponto. A compressão *offline* é feita depois que todos os dados já foram coletados. Por fim, no caso da compressão contextual, busca-se manter os pontos que atribuem significado à trajetória.

Além da utilização de todas as informações, ainda existe a opção de uso dos valores agrupados que representam a trajetória ou segmento. Esses valores são normalmente vetores de atributos, obtidos por meio de métricas estatísticas aplicadas aos valores originais dos atributos [14; 17; 29; 70]. Outra abordagem utilizada para resumir os dados é através do algoritmo de decomposição do perfil, que inclui as características locais do movimento nos atributos e não apenas um resumo geral [14; 70]. Esta forma de estruturar os dados diminui consideravelmente o custo computacional das atividades de aprendizagem de máquina.

De acordo com Zheng [79], além da utilização da forma original dos dados de trajetórias, é possível transformá-las em outras estruturas de dados como matrizes, grafos e tensores, aos quais é possível também fazer a aplicação de técnicas de aprendizagem de máquina. Deste modo, é importante o conhecimento do domínio, pois existem diferentes formas possíveis de representação e cada uma delas tem atributos associados distintos.

2.2 Aprendizagem de máquina

Alpaydin [3] define aprendizagem de máquina como uma forma de programar computadores para otimizar uma métrica de acordo com dados passados. Baseado em um conjunto de modelos, na aprendizagem, busca-se encontrar a melhor parametrização do modelo adequado para o conjunto de dados que são passados como entrada. Os modelos gerados podem ser prescritivos, preditivos, descritivos ou preditivos e descritivos.

Os modelos de aprendizagem de máquina podem ser de três tipos: não-supervisionado, semi-supervisionado e supervisionado. De acordo com Zhu e Goldberg [81], os modelos não-supervisionados trabalham em um conjunto de dados com n instâncias que não contêm nenhuma forma de guia sobre como as instâncias devem ser organizadas, e apontam como principais atividades: agregação, detecção de valores extremos (*outliers*) e redução de dimensionalidade.

Já no caso de modelos supervisionados, cada um dos objetos utilizados no processo de treinamento de um modelo tem uma classe associada. Zhu e Goldberg [81] definem modelos supervisionados como o treinamento, a partir de uma amostra de treino, de uma função F com o objetivo mapear o conjunto de atributos do objeto X para o conjunto de classes Y.O processo consiste em buscar uma função que, quando um novo objeto precise ser classificado, o seu resultado seja o mais próximo do valor real da classe daquele objeto. Os autores ainda destacam que, de acordo com o domínio das classes, o processo pode ser classificado como: classificação ou regressão.

Os modelos semi-supervisionados são uma combinação dos dois modelos anteriores. Zhu e Goldberg [81] afirmam que a maioria das atividades de máquina semi-supervisionada consiste em estender os modelos supervisionados ou não-supervisionados, de modo a adicionar informações complementares do outro paradigma.

2.3 Engenharia de atributos

De acordo com VanderPlas [66], o processo de engenharia de atributos consiste em transformar dados brutos em informações que podem servir de entrada para um algoritmo de aprendizagem de máquina. A transformação da informação para servir como entrada de um algoritmo de aprendizagem de máquina é de fato importante, todavia, não é o único passo executado no processo de engenharia de atributos.

Como Zheng e Casari [73] descrevem, o processo de engenharia de atributos consiste em formular os atributos mais apropriados para um determinado conjunto de dados, um dado modelo e uma determinada atividade. Os autores ainda complementam informando que a quantidade de atributos é importante pois, se forem poucos, pode não ser possível extrair informações relevantes, e se forem muitos, pode ser muito custoso para treinar o modelo. Em ambos casos, podem ocorrer erros no treinamento, afetando o desempenho.

A qualidade dos atributos utilizados está diretamente associada ao resultado gerado pelo modelo de aprendizado de máquina. De acordo com Domingos [15], o que faz a diferença entre o sucesso e o fracasso de um projeto de aprendizagem de máquina são os atributos que são utilizados como entrada. Desta forma, a utilização das técnicas de engenharia de atributos é necessária para adequar os atributos ao problema estudado.

Como levantado por Fromm, Wambsganss e Söllner [20], parte do processo de engenharia de atributos ainda depende do conhecimento de especialistas para a criação desses atributos. Assim, uma organização dos atributos existentes pode auxiliar no processo de criação deles.

No contexto de trajetórias, a interpretabilidade de um modelo é importante e ainda é um desafio [7]. Parte da interpretabilidade de um modelo depende dos atributos que são utilizados como entrada dos algoritmos, o que requer que eles sejam simples e de fácil compreensão.

Os atributos discutidos neste trabalho são utilizados em diferentes pesquisas com variados propósitos. Como visto, o sucesso das atividades de aprendizagem de máquina está na escolha correta dos atributos utilizados como entrada. O conhecimento em relação ao objeto de estudo é uma das partes mais importantes do processo. Assim, é essencial conhecer os atributos associados às trajetórias.

2.4 Matriz de correlação

A matriz de correlação de Pearson contém informações sobre a correlação para cada par de variáveis presente em um conjunto de dados. A correlação mede a força e a direção da relação linear entre duas variáveis quantitativas [39]. A associação linear entre duas variáveis implica que, quando uma das variáveis cresce em uma unidade, a outra cresce ou decresce em um valor fixo.

O coeficiente de correlação, também conhecido como coeficiente de correlação de Pearson, pode variar de -1 até 1. Quando o valor de correlação é zero, significa a não existência de relação linear entre as variáveis. No caso em que a correlação é próxima de um, significa que as duas variáveis crescem linearmente juntas. Já no caso em que a correlação é próxima de menos um, enquanto uma variável cresce, a outra decresce [63].

Ainda de acordo com Swinscow et al.[63], o coeficiente de correlação de Pearson, r, pode ser calculado a partir da fórmula na Equação 2.1:

$$r = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{[\sum(x - \bar{x})^2(y - \bar{y})^2]}} \tag{2.1}$$

Onde x e y representam os valores das variáveis quantitativas e $\bar{x}$ e $\bar{y}$ representam as médias aritméticas destas variáveis.

Uma matriz de correlação é uma matriz m por m onde cada um dos elementos da matriz representa o valor da correlação entres os elementos par a par [64]. A matriz de correlação é uma matriz simétrica e os valores da diagonal principal tem valor um. Uma matriz de correlação é exemplificada na Figura 2.2.

$$corr = \begin{bmatrix} 1 & \frac{\sum(x_1-\bar{x_1})(x_2-\bar{x_2})}{\sqrt{[\sum(x_1-\bar{x_1})^2(x_2-\bar{x_2})^2]}} & \cdots & \frac{\sum(x_1-\bar{x_1})(x_n-\bar{x_n})}{\sqrt{[\sum(x_1-\bar{x_1})^2(x_n-\bar{x_n})^2]}} \\ \frac{\sum(x_2-\bar{x_2})(x_1-\bar{x_1})}{\sqrt{[\sum(x_2-\bar{x_2})^2(x_1-\bar{x_1})^2]}} & 1 & \cdots & \frac{\sum(x_2-\bar{x_2})(x_n-\bar{x_n})}{\sqrt{[\sum(x_2-\bar{x_2})^2(x_n-\bar{x_n})^2]}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\sum(x_n-\bar{x_n})(x_1-\bar{x_1})}{\sqrt{[\sum(x_n-\bar{x_n})^2(x_1-\bar{x_1})^2]}} & \frac{\sum(x_n-\bar{x_n})(x_2-\bar{x_2})}{\sqrt{[\sum(x_n-\bar{x_n})^2(x_2-\bar{x_2})^2]}} & \cdots & 1 \end{bmatrix}$$

Figura 2.2: Exemplo de matriz de correlação de Pearson.

2.5 Taxonomia

Existem diferentes formas de classificar objetos, dentre as quais está o ato de criar uma taxonomia. As taxonomias são comumente utilizadas nos estudos em biologia para classificação dos seres. De acordo com Sokal e Sneath [60], diferentes termos são utilizados com relação às taxonomias e clarificar todos eles implicaria em um livro só para isto. Os autores propõem, então, a utilização da definição de taxonomia proposta por Simpson [59] como referência, onde, a taxonomia é definida como o estudo teórico da classificação, incluindo suas bases, princípios, processos e regras. Os autores ainda destacam que o termo taxonomia é atribuído ao resultado do processo. O termo pode ser usado em ambos os casos de acordo com o contexto.

Taxonomias são importantes em pesquisas e gestão, pois a classificação auxilia pesquisadores a entender e analisar problemas complexos [43]. De acordo com Glass e Vessey [23], o uso de taxonomia estrutura ou organiza o conhecimento em uma determinada área de estudo, trazendo vantagens potencialmente que levam ao avanço na área. Miller e Ruth [38] afirmam que as taxonomias fornecem descrições sobre os objetos em estudo que são úteis para discussão, pesquisa e pedagogia. Além disto, o uso de taxonomia auxilia no processo de entendimento das divergências em estudos prévios sobre o objeto de estudo [56].

O processo de criação de taxonomia tem mudado através do tempo. De acordo com Sokal [61], no princípio da ciência moderna, os seres eram agrupados por apenas uma caracterís-

tica, por exemplo: unicelular ou multicelular; de forma que os objetos de uma classe, obrigatoriamente, compartilhavam essa determinada característica e a decisão de qual é utilizada podia ser arbitrária. Em contrapartida, taxonomias baseadas em múltiplas características não requerem que esses atributos sejam universais, de modo que um objeto pode ser de um grupo sem conter todas as características dele. Os autores ainda destacam que com a utilização de múltiplas características, o processo de desenvolvimento de uma taxonomia torna-se mais complexo e requer utilização de técnicas apropriadas.

Como defendido por Nickerson, Varshney e Muntermann [43], o desenvolvimento de uma taxonomia é um processo complexo e não tem sido aplicado de forma adequada na área de sistemas de informação. Os autores estudaram as taxonomias que foram propostas na área de sistemas de informação e chegaram à conclusão de que em muitos casos, as taxonomias são feitas de forma intuitiva sem uma base metodológica. Por fim, eles apresentam uma metodologia para a criação de taxonomia que é aplicada neste trabalho.

Outro ponto importante destacado por Nickerson, Varshney e Muntermann [43] é que taxonomias podem ser classificadas como: indutiva, dedutiva ou intuitiva. De acordo com os autores, uma taxonomia é considerada indutiva se o processo de determinação das dimensões e características é baseado na observação empírica dos casos seguida de uma análise destes. Já no caso das taxonomias dedutivas, a derivação dá-se a partir das teorias e conceitos associados. Por fim, os autores expõem a taxonomia intuitiva na qual o pesquisador cria a taxonomia de acordo com o que ele considera fazer mais sentido.

2.6 Metodologia de geração de taxonomias

Ao longo da história, diferentes formas de criação de taxonomias foram propostas e aplicadas em diversos domínios do conhecimento. Com o avanço tecnológico, o poder de processamento dos dados aumentou. Assim, o que anteriormente era feito manualmente passou a ser desenvolvido em computadores, e a avaliação de um número maior de características passou a ser possível no processo de geração de taxonomia. O método de criação de taxonomia utilizado neste trabalho foi proposto por Nickerson, Varshney e Muntermann [43] em um trabalho que avaliou as taxonomias publicadas em periódicos de sistemas de informação. A partir do estudo das taxonomias existentes, os autores chegaram à conclusão de que mui-

tos dos trabalhos na área não seguiam uma metodologia. Com isso, os autores propuseram uma metodologia para criação de taxonomias na área de sistemas de informação. De acordo com eles, esse processo de desenvolvimento foi criado com base na literatura associada à taxonomia em outras áreas.

2.6.1 Definição e características de uma taxonomia

Uma taxonomia pode ser definida como um conjunto de n dimensões, onde cada uma delas consiste de k características que são mutuamente excludentes e coletivamente exaustivas. Uma dimensão pode ser entendida como uma variável associada aos objetos descritos pela taxonomia. Já as características são os possíveis valores que estas variáveis podem assumir. A taxonomia gerada utilizando a metodologia proposta deve estar de acordo com essa definição.

O fato das características serem mutuamente excludentes implica que nenhum objeto pode ser classificado em dois atributos na mesma dimensão. Já o fato de ser coletivamente exaustiva significa que, para cada uma das dimensões, o objeto deve ter uma das características daquela dimensão. Assim, para cada dimensão, o objeto tem que obrigatoriamente estar associado a um aspecto.

Uma característica associada às taxonomias é a dinâmica. Isso significa dizer que é possível a adequação da taxonomia caso exista mudança no comportamento do objeto de estudo. Considerando que para ser útil durante muito tempo é preciso que haja mudança na taxonomia, o desenvolvimento de uma taxonomia perfeita é inviável. Logo, no processo de desenvolvimento, busca-se por uma que seja útil. A seguir são descritos os pontos argumentados por Nickerson, Varshney e Muntermann [43] para definir uma taxonomia útil:

- **É concisa:** a capacidade cognitiva de um humano é limitada. Assim, para que uma taxonomia seja útil é necessário que ela tenha um número limitado de dimensões e de características, caso contrário, a compreensão da mesma pode ser comprometida com um grande número de informações.

- **É robusta:** uma taxonomia muito simples pode não representar os dados da melhor forma possível. De acordo com os autores, uma taxonomia com uma dimensão e duas características normalmente não é muito útil. Como esta característica está em conflito

com a necessidade de ser concisa, cabe ao pesquisador chegar a um balanço que possa ser representativo e compreensível ao mesmo tempo.

- **É abrangente:** o fato de uma taxonomia ser abrangente pode ser interpretada de duas formas. A primeira está atrelada ao fato de que a taxonomia pode classificar todos os objetos de um domínio estudado. A segunda está associada à inclusão de todas as dimensões dos objetos de interesse.

- **É extensível:** em uma taxonomia, é necessário que seja possível adicionar novas dimensões e novas características em uma dimensão quando novos objetos são avaliados. Os autores destacam que uma taxonomia estática tem chances de se tornar obsoleta.

- **É explicativa:** uma taxonomia útil não contém todos os detalhes de um objeto, mas possui explicações importantes sobre a natureza dos objetos estudados e futuros objetos que possam ser estudados. Assim, se alguém precisar entender um objeto, não necessita saber todas as informações sobre ele, apenas as que o localizam na taxonomia.

Essas características são necessárias para que uma taxonomia seja útil, todavia não são necessariamente as únicas que devem ser consideradas. Elas representam a fundação para a avaliação da taxonomia e guiam pesquisadores no processo de criação. A forma mais completa de avaliar uma taxonomia é observar sua utilização através do tempo. Uma forma de avaliar a taxonomia, nos casos em que é a avaliação da utilização por outros pesquisadores não é possível, é a verificação da utilidade da mesma perante o seu propósito.

2.6.2 Processo de criação de uma taxonomia

A metodologia proposta por Nickerson, Varshney e Muntermann [43] consiste em um conjunto de passos para o desenvolvimento de uma taxonomia. Nesta subseção estes passos propostos pelos autores são detalhados. A Figura 2.3 contém um fluxograma com cada um dos passos da metodologia. Estes passos são genéricos e podem ser aplicados a diferentes áreas de sistemas de informação.

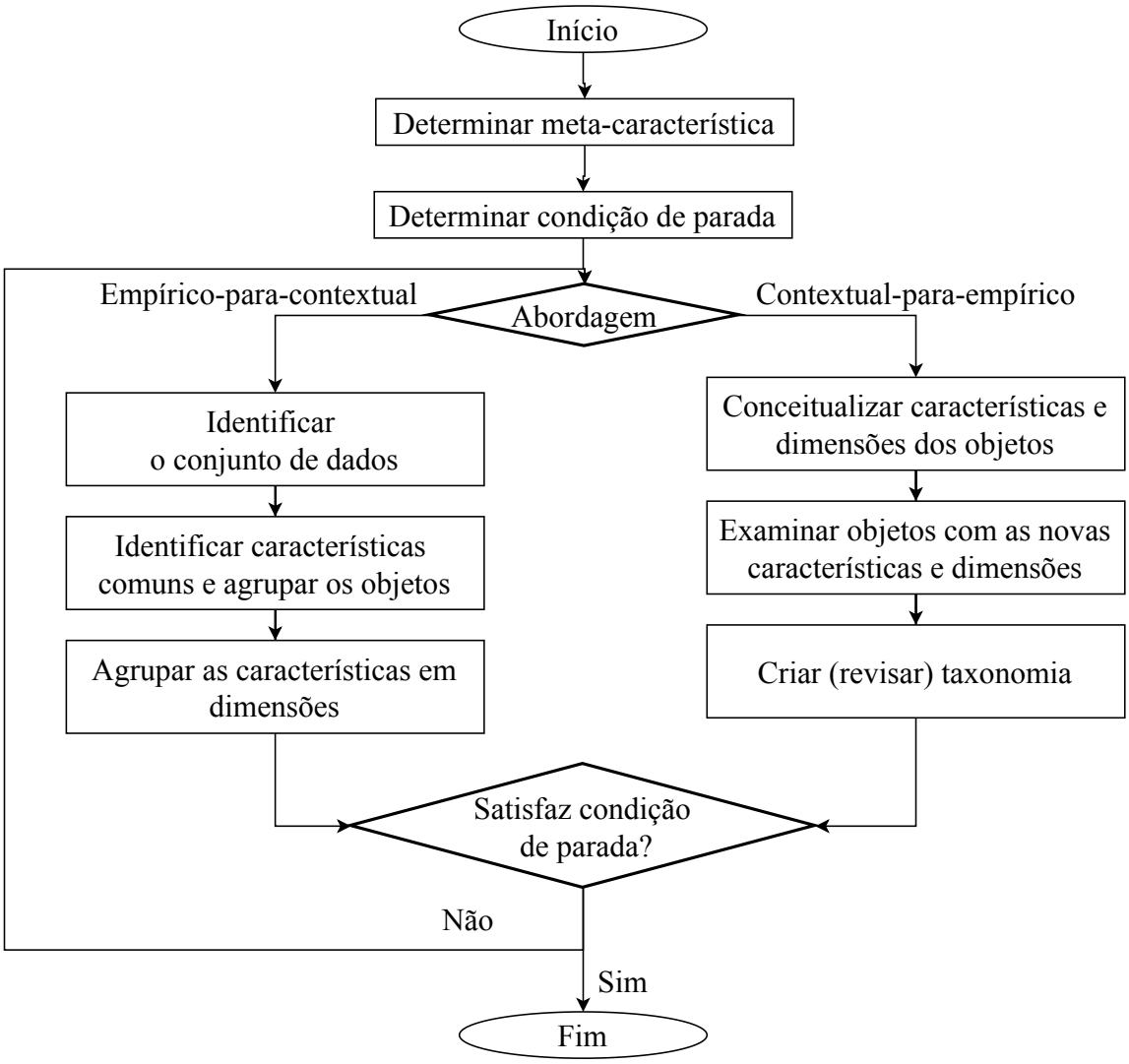

Figura 2.3: Fluxograma do método de desenvolvimento de taxonomia proposto por Nickerson, Varshney e Muntermann [43]

O primeiro conceito apresentado pelos autores é o de meta-característica. A seleção das características de uma taxonomia é uma parte importante no processo de desenvolvimento de uma taxonomia. A meta-característica é utilizada para evitar que vários atributos não relacionados sejam propostas na busca de encontrar um padrão. A meta-característica é a característica mais compreensível de todas, e a partir dela, as outras são derivadas. A escolha da meta-característica deve estar diretamente associada ao propósito da taxonomia.

Para exemplificar, os autores utilizam o exemplo de criar uma taxonomia para classificar uma plataforma de acesso on-line. Caso a taxonomia seja para classificar o poder de processamento, a meta-característica deve ser as características de *hardware* e *software* da

plataforma e uma possível característica seria memória disponível. Caso a classificação seja em relação à utilização por usuário, a meta-característica deve ser a capacidade da plataforma interagir com o usuário e um possível atributo seria o número máximo de usuários utilizando a plataforma simultaneamente.

A decisão de qual meta-característica será aplicada é importante pois impacta no resultado da taxonomia, uma vez que está diretamente associada ao propósito da mesma. Os autores destacam que, apesar de influenciar no resultado, a escolha da meta-característica não é clara até que a metade do processo seja atingido.

Outra parte relevante no processo de criação de uma taxonomia é a condição de parada de processo. As condições de parada são tanto objetivas como subjetivas. Uma condição de parada destacada como fundamental pelos autores é que a definição de taxonomia seja respeitada e cada dimensão tenha suas características mutuamente excludentes e elas sejam coletivamente exaustivas.

Além de estar de acordo com a definição de taxonomia, os teóricos sugerem uma lista de possíveis condições objetivas que podem ser usadas para a determinação do fim do processo de desenvolvimento da taxonomia. A seguir são apresentadas as possíveis condições que podem ser utilizadas:

- Todos os objetos de estudo ou uma amostra representativa dos objetos foram examinados;
- Nenhum objeto foi combinado com o objeto similar ou separado em múltiplos objetos na iteração final;
- Pelo menos um objeto é classificado como cada característica em cada dimensão;
- Nenhuma dimensão ou característica foi adicionada na última iteração;
- Cada dimensão é única e não se repete;
- Cada característica é única na sua dimensão;
- Cada combinação de dimensão e característica é única e não se repete.

As condições subjetivas propostas pelo trabalho estão associadas à utilidade da taxonomia de acordo com os requisitos propostos pelos autores. Assim, para que o processo de

desenvolvimento esteja completo, é necessário que a taxonomia resultante seja concisa, robusta, abrangente, extensível e explicativa. Desta forma, unindo as condições objetivas com as subjetivas, aumentam as chances da taxonomia ser construída de forma correta, diminuindo a influência do pesquisador no resultado. e que a mesma seja útil para o propósito que foi planejada.

Uma vez que o desenvolvimento das taxonomias, utilizando este método, resulta em taxonomias que associam o objeto pelas similaridades, é possível que essas similaridades sejam obtidas tanto a partir de conceitos associados aos objetos, assim como de informações obtidas através da avaliação do conjunto de objetos avaliados. A metodologia escolhida compreende a combinação das duas abordagens para que o pesquisador chegue a uma taxonomia útil da melhor maneira possível. Assim, durante o processo iterativo, é possível o uso da abordagem conceitual para a empírica, ou o inverso. As duas abordagens são descritas a seguir.

A primeira consiste em partir dos conceitos existentes sobre os objetos, e criar possíveis dimensões sem levar em consideração os objetos que estão sendo avaliados de fato. Assim, a partir dos conceitos existentes, o pesquisador cria as dimensões associadas aos objetos e avalia se as características destas dimensões podem ser derivadas da meta-característica utilizada como base para o desenvolvimento. Caso as características não possam ser derivadas da meta-característica, elas são eliminadas. Uma vez que as características são validadas quanto à meta-característica, elas são avaliadas quanto à relação de modo que devem ser mutuamente excludentes e coletivamente exaustivas. Os autores nomeiam este processo como conceitual para empírico.

A segunda abordagem, por sua vez, consiste, primeiramente, na identificação dos dados que serão classificados. A partir desse conjunto de dados, o pesquisador encontra as características comuns entre estes objetos que devem ser deriváveis da meta-característica. As características escolhidas devem ser discriminativas entre os objetos, pois não é útil uma característica que é comum a todos os objetos. Uma vez que o conjunto de características é criado, as mesmas podem ser agrupadas com alguma técnica estatística para a criação das dimensões. Uma vez que os grupos são criados, é necessário a criação de um rótulo conceitual. Os autores nomearam esta abordagem de empírico-para-contextual.

Com as fases do processo definidas, é possível descrever o fluxo do mesmo. Primei-

ramente, é decidida qual meta-característica será utilizada para a criação da taxonomia de acordo os critérios discutidos previamente. A condição de parada escolhida é avaliada ao final de cada iteração. Uma vez que a base é definida, começa o processo iterativo. A primeira decisão é qual abordagem será utilizada: conceitual-para-empírica ou empírica para conceitual. Ao fim da criação das novas dimensões, são feitas as mudanças necessárias com relação às dimensões repetidas ou que precisam ser divididas. Uma vez finalizado o processo escolhido, a taxonomia é avaliada através da condição de parada. Caso não satisfaça os critérios de parada deve-se voltar no processo para a fase de seleção de abordagem e o processo de repete.

2.7 Considerações finais

Neste capítulo foram apresentados assuntos que são o fundamento deste trabalho. Foram introduzidas definições sobre os dados de trajetórias e as formas de representação de uma trajetória. Também foram detalhados os conceitos sobre aprendizagem de máquina e engenharia de atributos. Os conceitos relacionados à matriz de correlação também foram discutidos. Por fim, o conceito de taxonomia foi definido junto com suas utilidades e possíveis processos de criação. A metodologia utilizada no desenvolvimento da taxonomia proposta neste trabalho foi exposta detalhando os pontos principais a serem considerados no processo.

No próximo capítulo são apresentados alguns trabalhos relacionados a esta pesquisa.

Capítulo 3

Trabalhos relacionados

Com o avanço da tecnologia em diversas áreas, cada vez mais a aplicação de algoritmos de aprendizagem de máquina aos dados gerados tem sido necessária em todas as áreas de conhecimento, para a obtenção de informações relevantes. Com relação às trajetórias, diversas frentes têm sido exploradas com relação aos dados, dependendo do domínio do problema. Para cada domínio, diversas técnicas vêm sendo estudadas, visando a obtenção de bons resultados.

Neste capítulo, serão abordados os trabalhos concernentes às trajetórias de embarcações, de mobilidade humana e independentes de domínio. As seções foram apresentadas com essa estrutura para ficar de acordo com as bases de dados utilizadas nesta pesquisa. Os trabalhos com proposta independente de domínio utilizam múltiplas bases de dados para avaliação do algoritmo proposto, buscando generalizar o resultado, e incluem dados de trajetórias de animais e fenômenos naturais. Por fim, são apresentados estudos que utilizam taxonomias relacionadas às trajetórias, assim como relacionadas ao processo de aprendizagem de máquina.

3.1 Trajetórias de embarcações

Trajetórias de embarcações podem assumir diferentes formas por não terem nenhuma delimitação de caminho obrigatório a ser seguido, como ocorre no caso das trajetórias em rodovias, por exemplo. Existem também regras associadas aos movimentos em algumas regiões específicas, como no entorno de portos. Além disso, embarcações têm diferentes padrões de

deslocamento, tipo, por exemplo, uma embarcação cargueira movimentar-se de uma determinada forma, enquanto um barco em rota de pesca se comporta de maneira distinta. Nesta seção, são, portanto, apresentados trabalhos que abrangem diferentes domínios relacionados às trajetórias de embarcações.

Prever onde um objeto vai estar, de acordo com o seu histórico, pode ser útil em diversas aplicações, tal como a prevenção de acidentes. Em seu trabalho, Pallotta et al. [48] recorreram aos dados históricos pré-processados de sistemas automáticos de identificação. Os teóricos utilizaram o processo estocástico de Ornstein-Uhlenbeck [65] para a predição da posição de um navio que esteja seguindo alguma das rotas históricas. A partir dos dados históricos, foram então derivadas informações contextuais, que resultaram em pontos de notificação (*waypoints*) como entradas, saídas e áreas estacionárias. Também foram derivadas rotas representativas entre os pontos de notificação que são usadas como entrada para o modelo. Estatísticas, a tipo de exemplo a velocidade sobre a terra e curso sobre a terra, tiveram uso.

Mazzarella, Arguedas e Vespe [37] trabalharam com dados de AIS (Automatic Identification Systems) de sistemas de navegação marítimos. Os autores apresentaram um método desenvolvido para melhorar a previsão de localização de embarcações, oferecendo melhorias no conhecimento da situação marítima. No estudo é realizada a aplicação de um algoritmo de previsão Bayesiano, baseado em um filtro de partículas a dados históricos de AIS. Para avaliar o sistema proposto, são utilizados dados reais de uma área específica entre o estreito de Gilbraltar e o estreito de Dover. Os teóricos utilizaram informações relativas às localizações geográfica dos pontos (latitude e longitude), uma discretização dos valores de tempo, velocidade sobre a terra e curso sobre a terra.

Nguyen et al. [42] também recorreram aos dados de AIS com o objetivo de monitorar navios. O foco da ferramenta desenvolvida foi lidar com as dificuldades presentes na análise de dados de AIS: volume de dados gigante, ruído e irregularidade na amostragem. A ferramenta apresentada realiza reconstrução de trajetórias, detecção de anomalias e identificação de tipo de embarcação. Os estudiosos focaram no fato de lidar com grandes quantidades de dados e na utilização de aprendizagem profunda. O modelo de aprendizado utilizado neste trabalho foi rede neural recorrente. Como entrada o algoritmo, os autores utilizaram uma forma de *hot encoding* dos atributos: latitude, longitude, velocidade sobre a terra e curso

sobre a terra.

Li et al. [34] explanaram uma combinação de técnicas de agregação e mapeamento de trajetórias, com o objetivo de explorar a grande quantidade de dados gerada pelos dispositivos de sistemas automáticos de identificação (AIS). De acordo com os autores, esses sistemas são comumente utilizados como complementares a radares em navios com o objetivo de melhorar a segurança na navegação. A aplicação das técnicas de mapeamento espaço-temporal das trajetórias permite um melhor desempenho na clusterização. Para isso, os autores propõem a utilização de *merge distance* para calcular a similaridade entre as trajetórias, *multidimensional scaling* para criar uma expressão espacial das similaridades entre as trajetórias e uma versão melhorada do *DBSCAN* para trajetórias como algoritmo de clusterização. Os atributos levados em consideração no processo de clusterização neste trabalho foram as localizações geográficas e o tempo associado à coleta dos mesmos.

Varlamis et al. [68] trabalharam com dados de AIS para extrair informações de trajetórias de embarcações. Eles propuseram uma abordagem baseada em redes semânticas, que podem ser genéricas e conter informações importantes, e podem ser processadas com análises de rede ou outras técnicas de mineração de dados - até mesmo não-supervisionadas - para encontrar pontos fora da curva. Os atributos deste estudo foram distância coberta, velocidade, aceleração, direção e taxa de mudança de direção.

3.2 Trajetórias de mobilidade humana

O estudo de trajetórias geradas por humanos ajuda a entender o comportamento de um grupo de pessoas. Muitas informações importantes podem ser derivadas de dados de trajetórias que têm aplicação em áreas como propaganda direcionada, que se utiliza da informação da localidade do usuário para a oferta de um determinado produto, quando disponível em uma loja próxima, ou mesmo para encontrar padrões de comportamentos de mobilidade associados a uma região de interesse. Nesse contexto, trajetórias podem ser geradas das mais variadas formas. O usuário pode estar em uma rota caminhando, ou em um transporte público, ou em seu próprio veículo, e em cada um desses contextos a rota da trajetória é restringida por caminhos distintos. A habilidade de extrair informação relevante a partir de trajetórias auxilia no planejamento tanto individual quanto no coletivo. Nesta seção são

expostos trabalhos relacionados a conjuntos de trajetórias associados à mobilidade humana.

Quando se trata de movimento, diferentes formas de anomalias podem estar associadas. Por exemplo, uma anomalia pode estar associada ao movimento de um indivíduo, assim como também pode estar associada ao movimento de um grupo, ou até mesmo aos movimentos ocorridos em uma determinada região. Neste sentido, Barragana, Alvares e Bogorny [4] apresentam um algoritmo de detecção de comportamentos não usuais próximos a pontos de interesses. Segundo eles, para cada objeto que se move, são avaliados três tipos de comportamentos não usuais: permanecer ao redor, fuga e retorno. Um conjunto de dados reais é então utilizado para avaliar o desempenho da solução proposta e os indivíduos desse conjunto são classificados de acordo com o grau de comportamento não usual para um determinado conjunto de pontos de interesse. Os autores apresentam o trabalho como uma promissora aplicação em sistemas de segurança, visto que os pontos de interesse podem ser câmeras de segurança, edifícios comerciais, entre outros. Os atributos usados por eles são: tempo, velocidade, localização dos pontos (latitude e longitude), ângulos entre duas sub-trajetórias e região de interesse.

Wang et al. [69] discorrem sobre um método de criação de grupos desenvolvido para detectar trajetórias anômalas de táxis. Nele, trajetórias com mesma origem e destinos são consideradas e um algoritmo de *edit distance* modificado é aplicado para medir a similaridade entre as trajetórias. As trajetórias anômalas são classificadas a partir da hierarquização dos grupos. Os atributos aos quais os pesquisadores recorreram foram as localizações geográficas dos pontos da trajetória e o tempo associado a eles.

Outra aplicação de detecção de anomalias quando se trata de movimento humano é o acompanhamento do tráfego de veículos. Em seu trabalho, Kong et al. [31] operam dados de trajetórias para detectar anomalias no tráfego de veículos a longo prazo em cidades. A ideia presente no artigo é encontrar as áreas com anomalias no tráfego a partir do conjunto de dados coletados de rotas de ônibus, para sugerir um melhor planejamento nessas regiões. Para a obtenção dos resultados foram utilizados no experimento os atributos relacionados à velocidade do ônibus, o tempo de parada por ponto de ônibus e as localizações geográficas dos pontos de pontos das trajetórias e dos pontos de ônibus.

Também é possível trabalhar com os dados de trajetórias humanas para calcular os riscos de acidente no trânsito. Sobre esse assunto, Paefgen et al. [46] derivaram informações

relativas ao risco de acidente para motoristas. Eles propuseram a criação de atributos derivados dos dados originais para o domínio de acidentes. Os atributos utilizados por eles foram: frequência de viagens do veículo por mês, distância percorrida pelo veículo por mês, número de horas dirigidas pelo motorista, tipo de rodovia.

No processo de agregação, entidades são comparadas e agrupadas por densidade ou por um número definido de grupos. Uma parte importante na atividade de agregação é exatamente a definição de uma medida de distância entres as entidades. No caso de trajetória, uma medida de distância deve abranger tanto as características espaciais quanto as temporais. O uso dessas duas características é o que foi proposto no trabalho de Hong, Chen e Mahmassani [25]. Eles expuseram um método de clusterização para trajetórias que leva em consideração a matriz viária que os transportes transitam. Com as informações dos caminhos disponíveis, os autores utilizam uma abordagem de menor caminho associada ao tempo no processo de criação dos grupos. Os atributos considerados aqui são: tempo, velocidade e localização geográfica dos pontos nas trajetórias.

Scherrer et al. [57] propõem um método para identificar diferentes tipos de usuários: moradores do local ou visitantes. Os dados utilizados foram coletados por uma aplicação de celular de sistemas de navegação. Os pesquisadores levam em consideração que pessoas residentes em um determinado local movimentam-se de forma diferente dos turistas. Os atributos extraídos pelos autores são: número de dias nos quais o usuário está usando o aplicativo de localização, número de dias consecutivos que o usuário recorre ao aplicativo de localização, número de dias da semana que o usuário utilizou o aplicativo, número total de dias que o usuário precisou do aplicativo, tempo total de uso, distância, distância média entre os pontos visitados, região percorrida, porcentagem da caminho que se repete, velocidade, número de paradas, tempo de paradas.

Zheng et al. [74] tem como objetivo a classificação do meio de transporte que está sendo usufruído pelo usuário em um determinado ponto. Nessa perspectiva, se comparam o desempenho de quatro classificadores, dentre os quais a árvore de decisão obteve melhor resultado. Os autores também introduziram ao algoritmo de particionamento à ideia de pontos de mudança entre um segmento caminhado e um segmento percorrido, usando outro tipo de meio de transporte. Como entrada para os classificadores, os autores utilizaram: comprimento do segmento, velocidade e aceleração.

Zheng et al. [75] fizeram uso de um modelo baseado em árvore de decisão seguido de um pós-processamento dos resultados. O trabalho introduz três atributos que são deriváveis dos dados originais e que melhoraram os resultados. Aqui, foi introduzida a ideia de taxa de variação de direção, taxa de mudança de velocidade e taxa de paradas. Além dos atributos introduzidos, os autores ainda utilizavam informações relativas à velocidade, aceleração e comprimento do segmento percorrido. Em um trabalho posterior, Zheng et al. [78] recorreram à mesma abordagem citada, melhorando o algoritmo de particionamento, considerando que para a mudança de meio de transporte sempre existe um segmento caminhado. Eles obtiveram melhores resultados conseguindo separar as trajetórias nos diferentes meios de transporte sem a ajuda de informações externas, como pontos de interesse ou mapas associados às trajetórias.

Através da presença de acelerômetro e GPS em aparelho celulares, Reddy et al. [52] usaram uma árvore de decisão junto com um modelo oculto de Markov. A abordagem identificava se o usuário estava parado, andando, correndo, em uma bicicleta ou em um veículo motorizado. Os autores usaram a magnitude do vetor de força gerado pelo sensor de acelerômetro e a velocidade como base para os atributos utilizados.

Etemad, Soares Júnior e Matwin [17] criaram um *framework* para detecção de meios de transporte utilizados pelos usuários em uma trajetória. O *framework* segue cinco passos: Os pontos são agrupados em trajetórias, os atributos associados aos pontos são calculados, atributos relacionados a trajetória completa são calculados, os valores com ruídos são removidos, e os valores dos atributos ficam normalizados. Os atributos dessa abordagem foram: aceleração, arranque, direção, taxa de mudança de direção, taxa de mudança na taxa de mudança de direção, velocidade, distância do ponto anterior.

3.3 Trabalhos trajetórias independentes de domínio

É relevante destacar que algumas propriedades de trajetórias são comuns a todos os domínios. Diante disso, muitos trabalhos propõem algoritmos gerais que possam ser aplicados a diferentes conjuntos de dados. Os atributos utilizados neste tipo de abordagem normalmente são derivados da localização geográfica e do tempo de coleta dos pontos da trajetória. Um destaque é que como o movimento de animais e de fenômenos naturais são livres, os

trabalhos que lidam com esses dados são genéricos e se encaixam nessa categoria. Nesta seção, são discutidos os trabalhos que focam em abordagens mais gerais, que independem do domínio das trajetórias, bem como as características associadas aos mesmos.

Muitas vezes, as bases de dados necessitam de uma rotulação manual para serem empregadas em aplicações de aprendizagem de máquina supervisionadas. Porém, este trabalho de anotação é muito custoso e demanda conhecimento de especialista. Para a obtenção de bons resultados, é necessário a rotulação das partes mais críticas que podem diferenciar as trajetórias de acordo com suas classes. Para auxiliar neste processo de anotação, Soares Júnior et al. [29] utilizaram *active learning* para seleção de sub-trajetórias que geram melhores resultados no decorrer da aprendizagem de máquina, visando a diminuir o esforço no trabalho de anotação. Os autores testaram sua abordagem com classificação utilizando três bases de dados relacionadas a diferentes contextos, explicados a seguir.

A primeira base de dados em Soares Júnior et al. [29] é relativa à trajetória de animais, onde os atributos utilizados foram: velocidade em m/s, a variação de direção em graus, a distância do ponto anterior, profundidade do solo, distância para o ponto de água mais próximo, a distância para o ponto de água mais próximo dentre de uma área de pasto, altitude, fechamento da copa de todas as árvores próximas ao animal e o percentual de inclinação com relação a área ao redor. . A segunda base de dados é relativa à navegação de navios na costa nordeste do Brasil, com o intuito de descobrir se o barco está ou não em atividade de pesca. Os atributos aqui empregados foram: a velocidade em m/s, a variação de direção em graus e a distância percorrida desde o último ponto. A terceira base utilizada no experimento foi a Geolife, que é relativa à mobilidade urbana com o alvo de classificar o tipo de meio de transporte associado a uma trajetória. Os atributos da Geolife foram: velocidade em m/s e aceleração em m/s^2. Para cada trajetória, com os valores dos atributos associados aos pontos, foram calculados o mínimo, o máximo e a média, e esses valores foram utilizados nos classificadores.

Furtado et al. [22] propõem uma medida de similaridade para trajetória que visa diminuir a incerteza associada com o cálculo. A ideia dos autores é a utilização de uma elipse que tem seu tamanho variável de acordo com taxa de coleta dos pontos da trajetória: quanto mais próximos os pontos, menor será a elipse gerada. Com isso, as trajetórias podem ser comparadas com suas áreas mesmo que elas não tenham a mesma taxa de coleta dos pontos.

Os atributos utilizados nessa abordagem são as localizações geográficas dos pontos.

Com o aumento do número de dados de trajetórias, a comparação entre elas para encontrar similaridade tem um grande impacto na escalabilidade do processo. Visando a diminuir o número de comparações realizadas durante esse processo que, na maioria dos casos, é de ordem quadrática, Furtado, Pilla e Borgorny [21] argumentaram a favor de uma estratégia que usa as propriedades de distância em espaços euclidianos para diminuir o número de comparações realizadas com a utilização de um limiar. As localizações dos pontos são aplicadas na comparação.

Como visto, é possível a criação de grupos de diferentes formas dependendo da medida de similaridade ou distância entre as trajetórias. Nanni e Pedreschi [40] recomendam uma adaptação para trajetória baseada em densidade, que utiliza apenas pares de trajetórias que são contemporâneos para o cálculo de distâncias. Ou seja, para cada par de trajetórias, é calculada a distância euclidiana entres os pontos que foram coletados no mesmo tempo. Como as trajetórias são contemporâneas, não há a necessidade de nenhuma transformação nelas para o cálculo da distância. Os atributos nesse caso são a localização geográfica e os dados temporais do momento no qual o ponto da trajetória foi coletado.

Os algoritmos empregados para a criação de grupos são baseados na utilização da trajetória completa. Lee, Han e Whang [32] partem do princípio de que agrupar trajetórias completas pode levar à não identificação de sub-trajetórias similares. Assim, os autores sugerem outra abordagem para a criação de grupos, que é baseada na ideia de sub-trajetórias comuns. Ao invés de calcular as distâncias entre as trajetórias como um todo, as trajetórias são particionadas em pontos importantes das trajetórias onde ocorrem mudanças no movimento. As sub-trajetórias geradas são agrupadas através de métricas de distância, que levam em consideração a forma da trajetória e desconsideram os tempos associados a ela. Tanto na separação das sub-trajetórias quanto no agrupamento das mesmas, o atributo utilizado é a posição geográfica dos pontos da trajetória. Os autores validam sua abordagem com o emprego de duas bases de dados: a primeira contém trajetórias relativas ao movimento de animais e a segunda é relacionada aos movimentos de furacões.

Yao et al. [72] discutem uma metodologia para extração de atributos das trajetórias que são independentes de tempo e da localização, para que trajetórias com comportamento semelhante, mesmo que não tenham acontecido no mesmo lugar, possam ser agrupadas de forma

correta. Os autores utilizam uma janela deslizante para que os atributos sejam independentes de tempo e, ainda assim, representativos da trajetória. A metodologia proposta é testada com trajetórias artificiais e reais, com resultados superiores aos métodos existentes para ambos. Os atributos extraídos pelos autores são: distância percorrida, intervalo de tempo, variação da velocidade e a taxa de giro.

Outra abordagem existente na literatura é proposta por Ferrero et al. [19], que apresenta uma nova metodologia para encontrar sub-trajetórias relevantes denominada Movelets. O algoritmo é inspirado no conceito de séries temporais *Shapelets*. Eles sugerem medidas de distâncias que podem comparar uma trajetória com sub-trajetórias. No algoritmo proposto, são aplicados como atributos: as distâncias entre as trajetórias avaliadas e as sub-trajetórias importantes encontradas pelo algoritmo.

Lee et al. [33] trabalham um *framework* para classificação de trajetórias. Eles aplicam a técnica de clusterização descrita em [32] e, com os resultados, criam dois atributos que podem ser empregados em qualquer classificador. Os atributos gerados são derivados de dois tipos de grupos: baseado em região *(region-based)* e baseado em trajetória *(trajectory-based)*. Os autores validam seus atributos através das bases de dados de movimentos de animais e movimentos de furacões, com o classificador Máquina de Vetores de Suporte *(Support vector machine - SVM)*

3.4 Taxonomias

Taxonomias podem ser utilizadas para representação de problemas complexos. Assim como em tantas outras áreas do conhecimento, diversas taxonomias foram propostas destacando aspectos relacionados à ciência da computação. Existem diferentes formas de geração de taxonomias, e por isso as taxonomias existentes em uma área variam de acordo com a metodologia utilizada em cada contexto. Nesta seção, portanto, serão discutidos trabalhos relevantes relacionados às taxonomias no contexto de trajetória e de engenharia de atributos.

Na área de engenharia de atributos, Fromm, Wambsganss e Söllner [20] propuseram uma taxonomia para representação das características dos atributos associados às atividades de processamento de linguagem natural. Os autores destacam que muitos atributos associados a esta área são de difícil compreensão e que muitos dos atributos utilizados costumam ser

propostos por especialistas na área.

Oudah e Henschel [44] utilizam uma taxonomia que representa microbiomas no processo de engenharia de atributos para a classificação dos mesmos. Os autores conseguiram derivar informações importantes sobre os microbiomas reduzindo o conjunto de atributos para ser o mais sucinto e informativo possível. Com a aplicação da metodologia, os autores conseguiram melhores resultados do que alguns trabalhos que utilizaram processos de engenharia de atributo mais simples.

Na área de trajetórias, Alamri, Taniar e Safar [1] propuseram uma taxonomia para representar as possibilidades de consultas relacionadas aos objetos em movimento. Os autores descrevem as diferentes consultas possíveis em cinco categorias: localização, movimento, objeto, temporal e padrões. Esses são aspectos que podem ser considerados quando uma consulta é criada. Os autores focam nos índices e estruturas associadas para a geração de consultas eficazes.

Ainda no domínio de trajetórias, entretanto visando o problema de visualização, Allain, Turkay e Dykes [2] sugerem uma taxonomia para destacar algumas características relacionas às trajetórias. Eles propuseram a criação de três categorias: a capacidade de decisão do objeto em movimento com relação à trajetória, seu nível de restrições associadas e dados contextuais. Nesse estudo, foram avaliados os trabalhos relacionados à visualização de trajetória com relação as categorias propostas.

Apesar do termo taxonomia não ser muito dito em trabalhos relacionados às trajetórias, existem pesquisas que, a partir de uma revisão da literatura, propuseram uma organização das informações com relação a aplicações e modelos utilizados. Nesse sentido, Zheng [79] discute estudos relacionados à mineração de dados de trajetórias e as organiza em grupos, de acordo com sua aplicabilidade. Seguindo a mesma abordagem, Mazimpaka e Timpf [36] organizaram conceitos relacionados aos métodos e aplicações de mineração de dados relacionados às trajetórias.

Um ponto importante que deve ser considerado por pesquisadores na área são os conceitos associados às trajetórias. Sobre isso, Renso, Spaccapietra e Zimányi [53] trazem definições e conceitos relativos às trajetórias e aplicações de aprendizagem de máquina. Essas definições podem servir de guia para o entendimento desta área de conhecimento.

Dentre os trabalhos presentes na literatura, não foi encontrado nenhum que direciona

atenção diretamente aos atributos de trajetória e suas características. De forma geral, o estado da arte está voltado para possíveis aplicações e modelos. Entretanto, alguns dos conceitos de trajetória apresentados na literatura estão diretamente associados aos atributos.

Por exemplo: a utilização de múltiplas trajetórias em projeto de aprendizagem de máquina, destacado por Renso, Spaccapietra e Zimányi [53], está diretamente relacionada à granularidade dos atributos neste contexto. Assim, nesta dissertação foram direcionados esforços para a organização dos atributos e derivação de informações importantes a partir deles.

3.5 Considerações Finais

Neste capítulo, discutimos os artigos relevantes na área de aprendizagem de máquina que são focados em dados de trajetórias considerados nesta pesquisa. Devido à importância da interpretabilidade dos resultados, é essencial o entendimento dos atributos utilizados nas diferentes tarefas da aprendizagem de máquina. Porém, percebe-se que pouco destaque é dado para os atributos - que são parte importante do processo. Outro ponto a destacar é que alguns dos atributos usados mudam de acordo com o domínio da aplicação, o que abre a possibilidade de encontrar padrões ou grupos existentes. Diferentes áreas da computação já abordaram a criação de taxonomias. Na área de trajetórias, alguns aspectos gerais foram cobertos, porém não foram encontrados trabalhos de taxonomia que modelam atributos com um maior detalhamento.

No próximo capítulo, os atributos aos quais recorremos em nosso estudo são apresentados e descritos de acordo com suas definições, e o processo de geração da Trajtax é descrito.

Capítulo 4

Trajtax: uma taxonomia para atributos de trajetória

Neste capítulo, o processo de criação de taxonomia TrajTax é descrito. Nesse sentido, atributos associados às trajetórias representam suas características espaço-temporais e a informação contida em um desses atributos pode estar associada a um dos diferentes aspectos como, por exemplo, localização ou forma do movimento. Assim, é importante ter a compreensão dos atributos de trajetórias a serem utilizados em modelos de aprendizagem de máquina. E exatamente para auxiliar nesse entendimento é que a taxonomia Trajtax foi proposta, cujos resultados obtidos durante a execução desta pesquisa são discutidos neste tópico.

Destacamos que os passos propostos na metodologia, presentes na Seção 2.6, , são aplicados aos atributos de trajetórias e os seus resultados são discutidos. Este capítulo é, portanto, dividido em seis seções. A primeira, Seção 4.1, apresenta o levantamento dos atributos relacionados às trajetórias. A segunda, Seção 4.2, aborda as bases de dados utilizadas na criação da taxonomia. A terceira, Seção 4.3, contém informações sobre a expansão da biblioteca Trajlib, usada para adicionar atributos às bases de dados. A quarta, Seção 4.4, tem informações sobre a geração de atributos complementares às bases de dados. A quinta, Seção 4.5, traz os passos realizados durante o desenvolvimento da taxonomia. A sexta, Seção 4.6, mostra a aplicação da taxonomia gerada e discute sobre sua utilidade, e, por último, a Seção 4.7 é o espaço no qual tratamos algumas considerações finais do capítulo.

4.1 Levantamento de atributos no contexto de trajetórias

Os atributos são definidos, de forma geral, por Chandrashekar e Sahin [5] como propriedades mensuráveis de um objeto que está sendo observado. Quando se trata de trajetórias, o objeto analisado pode variar de acordo com o propósito do estudo. Assim, ele tanto pode ser cada ponto da trajetória, quanto a trajetória completa, ou mesmo múltiplas trajetórias com comportamento em comum. Por esse fator, o propósito do trabalho influencia diretamente nos atributos que podem ser usados.

Soares Júnior et al. [29]propõem que os atributos sejam categorizados enquanto dois tipos: do ponto e de segmento. Segundo os autores, o atributo de ponto, *point feature*, do inglês, é associado a um ponto na trajetória. Já o atributo de segmento,*segment feature*, associa um atributo a um segmento completo de uma trajetória. O termo *características globais* globais também tem sido usado para atributos que resumem uma trajetória ou um segmento inteiro [70].

Todavia, é possível que essas definições sejam restritivas quando se é levado em consideração a forma de descrever um segmento, que pode ser formado por apenas dois pontos. Dessa maneira, qualquer atributo que recorra às informações de dois ou mais pontos para ser calculado pode ser considerado um atributo de segmento. Pode-se pegar o atributo velocidade como exemplo para verificar que o valor da velocidade em um ponto, e quando esse valor é calculado, depende de informações do ponto predecessor. Além disso, existem casos em que os valores dos atributos estão associados a múltiplos segmentos ou múltiplas trajetórias.

Levando isso em conta, uma outra possibilidade de trabalhar os atributos relacionados a trajetórias é considerar a frequência de cálculo do atributo e quais dados são necessários para seu cálculo. Primeiramente, devem ser levados em consideração quais dados são necessários para criação de um atributo. Por exemplo, para o cálculo da distância percorrida em relação ao ponto anterior, precisa-se das posições do ponto atual e do seu predecessor. O segundo aspecto considerado diz respeito à qual frequência em que esse atributo é calculado. No caso da distância, a frequência pode ser calculada para todos os pontos que tiverem um predecessor na trajetória. Assim, é possível aplicar o conceito a atributos de diferentes níveis, incluindo os derivados a partir de múltiplas trajetórias.

Em alguns domínios associados aos dados de trajetórias, a interpretabilidade do modelo de aprendizagem de máquina é muito importante e ainda é um desafio [7]. Uma parte fundamental no entendimento de um modelo é a clareza nos dados que alimentam o classificador. Assim sendo, nesta pesquisa foram levantados os atributos utilizados em trabalhos com distintos domínios no contexto de trajetória, junto com as suas descrições, para que seja possível a compreensão sobre eles. Neste ponto, convém destacar que, os atributos que são transformações não entraram no escopo desta pesquisa, em vez disso, foram coletados os atributos primitivos nos quais podem ser aplicadas transformações.

Nesta seção, os atributos revisados são nomeados e descritos com detalhes para que possam servir como base para outros trabalhos na área, podendo, assim, auxiliar outros pesquisadores que busquem um guia do que já foi utilizado. Os atributos descritos nesta seção foram levados em consideração no processo de desenvolvimento da taxonomia.

Os atributos levantados são descritos neste trabalho na seguinte estrutura:

Nome: descrição, número de pontos aos quais o atributo está associado, trabalhos nos quais os atributos foram utilizados.

- **Distância entre pontos:** distância entre dois pontos presentes na trajetória utilizando uma métrica de distância que melhor se ajusta ao cenário estudado. Pontos associados ao atributo: quaisquer dois pontos em uma trajetória [11; 16].

- **Velocidade:** distância percorrida entre dois pontos, usando uma métrica de distância que melhor se ajusta ao cenário estudado, dividida pela duração deste movimento. Pontos associados ao atributo: quaisquer dois pontos de uma trajetória [29; 58; 33; 72; 51; 31; 26; 80; 17; 28; 78; 11; 16; 74; 75; 10; 67; 55].

- **Aceleração:** mudança na velocidade de um objeto em movimento, dividida pelo tempo decorrido no curso dessa mudança. Pontos associados ao atributo: quaisquer dois pontos de uma trajetória [29; 58; 17; 78; 11; 16; 74; 75; 10; 67; 6].

- **Auto-interseção:** um ponto de auto-interseção é um ponto no qual a trajetória passa pelo menos duas vezes por ele. Pontos associados ao atributo: grupo de pontos que pode ser um segmento ou uma trajetória [27; 28].

- **Arranque:** alteração na aceleração de um objeto em movimento, dividida pelo tempo

decorrido durante essa alteração. Pontos associados ao atributo: quaisquer dois pontos de uma trajetória [11; 10].

- **Taxa de giro:** é o arco-tangente entre a localização geográfica de um ponto e a localização geográfica do ponto antecessor. Pontos associados ao atributo: quaisquer dois pontos consecutivos de uma trajetória [72].

- **Direção:** mede o ângulo entre a linha que conecta dois pontos sucessivos e uma linha que liga o primeiro ponto e o ponto de referência (por exemplo, o norte magnético ou verdadeiro). Pontos associados ao atributo: quaisquer dois pontos consecutivos de uma trajetória [10; 67].

- **Taxa de mudança de direção:** diferença absoluta entre dois valores de direção. Pontos associados ao atributo: quaisquer dois pontos consecutivos de uma trajetória [10; 67].

- **Taxa de mudança na taxa de mudança de direção:** diferença entre os valores de duas taxas de mudança de direção consecutivas divididas pela diferença de tempo. Pontos associados ao atributo: quaisquer dois pontos consecutivos de uma trajetória [17].

- **Ponto de permanência:** um ponto de permanência s representa uma região geográfica em que um objeto em movimento permaneceu por um determinado intervalo de tempo. A extração de um ponto de permanência depende de dois parâmetros de escala, um limite de tempo (T_{threh}) e um limite de distância (D_{threh}). Pontos associados ao atributo: um grupo de pontos que pode estar associado a uma trajetória ou segmento[76].

- **Taxa de parada:** número de pontos em que a velocidade está abaixo de um limite em um(a) segmento/trajetória dividido pelo comprimento do(a) segmento/trajetória. Pontos associados ao atributo: um grupo de pontos que pode estar associado a uma trajetória ou segmento [75; 78].

- **Taxa de mudança de velocidade:** número de pontos cuja mudança percentual na velocidade em relação ao ponto anterior excede um determinado limite, dividido pelo

comprimento do segmento. Pontos associados ao atributo: um grupo de pontos que pode estar associado a uma trajetória ou segmento [75; 78].

- **Taxa de mudança de direção:** número de pontos nos quais um objeto altera sua direção por um certo limite (Hc) dividido pelo comprimento do segmento. Pontos associados ao atributo: um grupo de pontos que pode estar associado a uma trajetória ou segmento [78; 75; 16].

- **Localização geográfica dos pontos inicial e final** localização geográfica do ponto inicial da trajetória e do ponto final. Pontos associados ao atributo: um grupo de pontos que pode estar associado a uma trajetória ou segmento [47].

- **Tamanho do segmento:** soma de todas as distâncias entre os pontos de um(a) segmento/trajetória. Pontos associados ao atributo: um grupo de pontos que pode estar associado a uma trajetória ou segmento [74; 55].

- **Semana, final de semana/feriado:** indica se a trajetória foi coletada em um dia da semana ou durante o final de semana ou feriado. Pontos associados ao atributo: um ponto presente na trajetória [80].

- **Dia da semana:** atributo categórico associado ao dia da semana (por exemplo, segunda-feira) no qual a trajetória aconteceu. Pontos associados ao atributo: um ponto presente na trajetória [80; 47; 58].

- **Intervalo de tempo:** o tempo é dividido em intervalos de acordo com o que melhor se adequa ao problema (por exemplo, intervalos de uma hora). Pontos associados ao atributo: um ponto presente na trajetória [47; 33; 13].

- **Tempo no início do(a) segmento/trajetória:** instante em que o(a) segmento/trajetória começou. Pontos associados ao atributo: um grupo de pontos que pode estar associado a uma trajetória ou segmento [26].

- **Duração do movimento:** diferença de tempo entre o ponto inicial e o ponto final de um(a) segmento/trajetória. Pontos associados ao atributo: um grupo de pontos que pode estar associado a uma trajetória ou segmento [47; 51; 71].

- **Taxa irregular de roteamento:** diferença entre uma trajetória e trajetória histórica com a mesma configuração de origem-destino. Pontos associados ao atributo: múltiplas trajetórias [62].

- **Localização típica do objeto:** locais em que objetos em movimento ficam a maior parte do tempo. Esse atributo é definido por um grupo de n localizações geográficas nas quais um objeto passa a maior parte do tempo ao longo do dia. Pontos associados ao atributo: múltiplas trajetórias [50].

- **Quilômetro por mês percorrido pelo objeto em movimento:** número de quilômetros percorridos pelo objeto em movimento no mês. Pontos associados ao atributo: múltiplas trajetórias [24].

- **Dias consecutivos:** número de dias consecutivos do objeto em movimento gerando trajetórias. Pontos associados ao atributo: múltiplas trajetórias [57].

- **Período de geração de trajetórias:** número de dias entre a primeira e a última trajetória. Pontos associados ao atributo: múltiplas trajetórias [57].

- **Tempo total das trajetórias:** soma da duração do tempo de todas as trajetórias. Pontos associados ao atributo: múltiplas trajetórias [57].

- **Distância total:** soma de toda a distância percorrida pelo objeto em movimento em todas as suas trajetórias. Pontos associados ao atributo: múltiplas trajetórias [57].

- **Distância máxima:** distância entre os dois pontos mais distantes visitados por um objeto em movimento presente nas trajetórias. Pontos associados ao atributo: múltiplas trajetórias [57].

- **Distância diária:** distância média percorrida pelo objeto em movimento por dia. Pontos associados ao atributo: múltiplas trajetórias [57].

- ***Distância entre os centroides diários:*** distância média de dois centroides diários consecutivos, onde o centroide diário é o centroide do polígono que rodeia os pontos relativos ao movimento diário de um objeto. Pontos associados ao atributo: múltiplas trajetórias [57].

- ***Tamanho médio do segmento:*** distância média percorrida em segmentos de movimentação. Um segmento de movimento descreve um segmento de movimento contínuo sem um segmento de parada. Pontos associados ao atributo: múltiplas trajetórias [57].

- **Área:** área total em que o objeto em movimento cobriu em m^2 calculada com base no polígono que rodeia todos os pontos visitados pelo objeto em movimento. Pontos associados ao atributo: múltiplas trajetórias [57].

- **Área diária:** área média em que o objeto em movimento cobre por dia em m^2. Pontos associados ao atributo: múltiplas trajetórias [57].

- **Sobreposição:** porcentagem média de sobreposição de duas áreas diárias consecutivas cobertas. Pontos associados ao atributo: múltiplas trajetórias [57].

- **Número de movimentos:** número absoluto de segmentos de movimento nas trajetórias do objeto em movimento. Um segmento de movimento descreve um segmento de movimento contínuo sem um segmento de parada. Pontos associados ao atributo: múltiplas trajetórias [57].

- **Número de paradas:** número total de paradas nas trajetórias do objeto em movimento. Pontos associados ao atributo: múltiplas trajetórias [57].

- **Duração de parada:** duração média das paradas nas trajetórias do objeto em movimento. Pontos associados ao atributo: múltiplas trajetórias [57].

- **Tempo de uso do veículo:** tempo de uso de um veículo em um mês. Pontos associados ao atributo: múltiplas trajetórias [47].

- **Acelerômetro:** uso do sensor do acelerômetro para adicionar os dados das forças de aceleração ao movimento que dá um sentido no qual o objeto está se movendo. Pontos associados ao atributo: um ponto presente na trajetória [52].

- **Profundidade do solo:** profundidade do solo até a camada que impede o movimento de água e ar através do solo. Pontos associados ao atributo: um ponto presente na trajetória [29].

- **Tipo de rodovia:** valor categórico que distingue o tipo de estrada (por exemplo, instalações urbanas, arteriais e rodoviárias). Pontos associados ao atributo: um ponto presente na trajetória [47].

- **Largura da estrada:** largura da estrada pela qual o objeto em movimento está percorrendo. Pontos associados ao atributo: um ponto presente na trajetória [71].

- **Limite de velocidade da estrada:** velocidade máxima aceita na estrada pela qual o objeto em movimento está percorrendo. Pontos associados ao atributo: um ponto presente na trajetória [62].

- **Velocidade do vento:** velocidade do vento associada a uma localização de ponto e tempo em uma trajetória. Pontos associados ao atributo: um ponto presente na trajetória [30].

- ***Ponto de interesse (POI):*** é uma área frequentemente visitada que é importante no contexto localizado. Pontos associados ao atributo: um ponto presente na trajetória [4].

- ***Região de interesse (ROI):*** é uma região em torno de um POI com um raio r. Pontos associados ao atributo: um grupo de pontos relacionados a um(a) segmento/trajetória [4].

- **Porcentagem de inclinação:** porcentagem de inclinação da superfície pela qual o objeto está se movendo. Pontos associados ao atributo: um ponto presente na trajetória [29].

- **Velocidade do motor - Rotações por minuto (RPM):** mecanismo de velocidade do transporte do objeto em movimento. Pontos associados ao atributo: um ponto presente na trajetória [24].

- **Velocidade do veículo alerta:** velocidade superior ao limite de velocidade da estrada depois de corresponder aos tipos de estrada por GPS. Pontos associados ao atributo: um ponto presente na trajetória [24].

- **Valor acima de um limiar:** valor de uma característica do movimento está acima de um limiar. Pontos associados ao atributo: um ponto presente na trajetória [24].

- **Velocidade angular do veículo:** mudança na orientação de um objeto, em radianos, a cada segundo (Rad / s). Pontos associados ao atributo: um ponto presente na trajetória [24].

- **Tempo de parada de ônibus:** tempo em que o objeto em movimento fica em uma estação de ônibus. Pontos associados ao atributo: um grupo de pontos que pode estar associado a uma trajetória ou segmento [31].

- **Distância para uma sub-trajetória importante** sub-trajetórias importantes são derivadas do conjunto de dados e as distâncias da trajetória para cada sub-trajetória são usadas como atributos. Pontos associados ao atributo: múltiplas trajetórias [19].

- ***Distância para atributos dependentes de contextos:*** distância dos pontos associados a um determinado contexto (por exemplo, o local mais próximo com água ao lidar com o movimento de animais). Pontos associados ao atributo: um ponto presente na trajetória [29; 58; 78].

- **Velocidade sobre a terra:** é a velocidade de uma embarcação em relação à superfície da terra, do inglês *Speed over ground (SOG)*. Pontos associados ao atributo: um ponto presente na trajetória [37; 6; 13].

- **Curso sobre a terra:** curso sobre a terra, do inglês *Course over ground (COG)*, é a direção real do progresso de uma embarcação, entre dois pontos, em relação à superfície da Terra. Pontos associados ao atributo: um ponto presente na trajetória [37; 6].

As tabelas 4.1, 4.2 e 4.3 associam os atributos apresentados nesta seção aos diferentes domínios de trajetórias. É possível ainda destacar alguns pontos importantes a partir da análise da tabela. O primeiro ponto é que os atributos relacionados às características do movimento são muito trabalhados em estudos que têm como foco os veículos. Outro ponto, é o fato de que pode-se verificar a presença de um grande número de atributos contextuais relacionados à trajetórias, que representam a mobilidade urbana.

Tabela 4.1: Utilização dos atributos associados a pontos nos diferentes domínios

	Mobilidade humana	Animais	Veículos	Fenômenos naturais
Curso sobre a Terra			X	
Distância para atributos dependentes de contexto		X		
Largura da estrada	X			
Limite de velocidade da estrada	X			
Ponto de interesse	X			
Porcentagem de inclinação		X		
Profundidade do solo		X		
Tipo de rodovia	X			
Auto-interseção	X	X	X	X
Direção			X	
Distância entre pontos			X	
Taxa de giro	X	X	X	X
Taxa de mudança de direção			X	
Acelerômetro			X	
Velocidade do vento				X
Velocidade sobre a Terra			X	
Aceleração	X	X	X	X
Arranque			X	
Taxa de mudança na taxa de mudança de direção			X	
Velocidade	X	X	X	X
Crescimento anormal da velocidade do motor	X			
Velocidade do motor	X			
Dia da semana	X			
Intervalo de tempo	X	X	X	X
Semana ou fim de semana/feriado	X			

Tabela 4.2: Utilização dos atributos associados a trajetória completa ou segmento nos diferentes domínios

	Mobilidade humana	Animais	Veículos	Fenômenos naturais
Distância para uma sub-trajetória importante	X	X	X	X
Região de interesse	X			
Localização geográfica dos pontos inicial e final	X			
Tamanho da(o) trajetória/segmento	X	X	X	X
Tempo de parada de ônibus	X			
Velocidade angular do veículo	X			
Ponto de permanência	X			
Taxa de mudança de velocidade			X	
Taxa de parada			X	
Valor comparado a um limiar	X			
Duração do movimento	X			
Tempo no início do movimento	X			

Tabela 4.3: Utilização dos atributos associados a múltiplas trajetórias nos diferentes domínios

	Mobilidade humana	**Animais**	**Veículos**	**Fenômenos naturais**
Quilômetros percorridos por mês	X			
Taxa irregular de roteamento	X			
Área	X			
Distância máxima	X			
Distância total	X			
Sobreposição	X			
Localização típica do objeto	X			
Área diária	X			
Distância diária	X			
Distância entre os centroides diários	X			
Duração da parada	X			
Número de movimentos	X			
Número de paradas	X			
Tamanho médio do segmento	X			
Tempo de uso do veículo	X			
Periodo de geração de trajetórias	X			
Dias consecutivos	X			

4.2 Bases de dados utilizadas

Uma parte importante do processo de desenvolvimento da taxonomia é a seleção dos dados utilizados como base no processo. Eles são importantes tanto para a criação das dimensões empíricas, quanto para a avaliação das dimensões conceituais. Desta forma, a escolha de bases de dados derivadas de contextos diferentes foi realizada visando a comparar domínios distintos relacionados à trajetórias, para que os resultados não estejam atrelados a um domínio. As três bases de dados escolhidas para o experimento deste trabalho são dos seguintes domínios: mobilidade urbana com diferentes meios de transporte, trajetórias de barcos e tra-

jetórias de animais. Os dados apresentados nesta seção são aplicados para o desenvolvimento da taxonomia proposta nesta dissertação.

4.2.1 Mobilidade urbana com meios de transportes

Geolife é uma base de dados resultante do *Microsoft Research GeoLife Project* que contém informações sobre a mobilidade de usuários em Pequim, China [77]. Os dados contêm anotações sobre o tipo de meios de transporte que o usuário fez uso em cada um dos pontos das trajetórias. Além da informação do tipo de meios de transporte, a base de dados contém detalhes sobre a localização geográfica de cada ponto (latitude e longitude), o dia e hora que o ponto foi coletado e a altitude daquele ponto.

Os dados das trajetórias de 182 usuários foram coletados durante o período de abril de 2007 a agosto de 2012. Neles, estão contidas informações sobre 17.621 trajetórias com cerca de 1,2 milhões de quilômetros e duração total de mais de 48.000 horas. Os dados foram coletados por diferentes celulares e dispositivos registradores GPS em frequências distintas. É importante ressaltar que 91 por cento dos dados foram coletados no intervalo temporal entre 1 e 5 segundos e no intervalo espacial entre 5 e 10 metros. Mais informações sobre isso, assim como a opção para baixar, podem ser encontradas no site[1] da Microsoft.

Devido ao grande número de trajetórias e ao fato de ser uma base de dados pública, a Geolife tem sido utilizada com variados propósitos como: encontrar a trajetória que passa mais próximo a uma sequência de pontos [8], achar correlação entre localizações geográficas [76], dentre outras aplicações. No entanto, a base citada é mais utilizada na detecção de meios de transporte a partir de trajetórias [17; 35; 74].

Considerando que pode ser custoso executar algoritmos de aprendizagem de máquina quando o conjunto de dados tem um número grande de instâncias, apenas uma amostra da base de dados original foi selecionada para os experimentos neste trabalho. Visando a manter a representatividade das classes existentes, para a amostra, foram selecionados os usuários que utilizaram o maior número de meios de transporte em suas trajetórias. Com essa abordagem, a base de dados foi reduzida a 987.518 pontos, usando, então, 5% da base original.

As classes de meio de transporte presentes na amostra selecionada são: caminhada, carro,

[1] https://www.microsoft.com/en-us/download/details.aspx?id=52367

ônibus, bicicleta, táxi, metrô, trem, avião, barco, corrida e motocicleta. Sobre isso, existe um desbalanceamento quanto aos meios de transporte utilizados nas trajetórias e a porcentagem de cada uma das classes presentes na amostra selecionada pode ser visualizada na Figura 4.1.

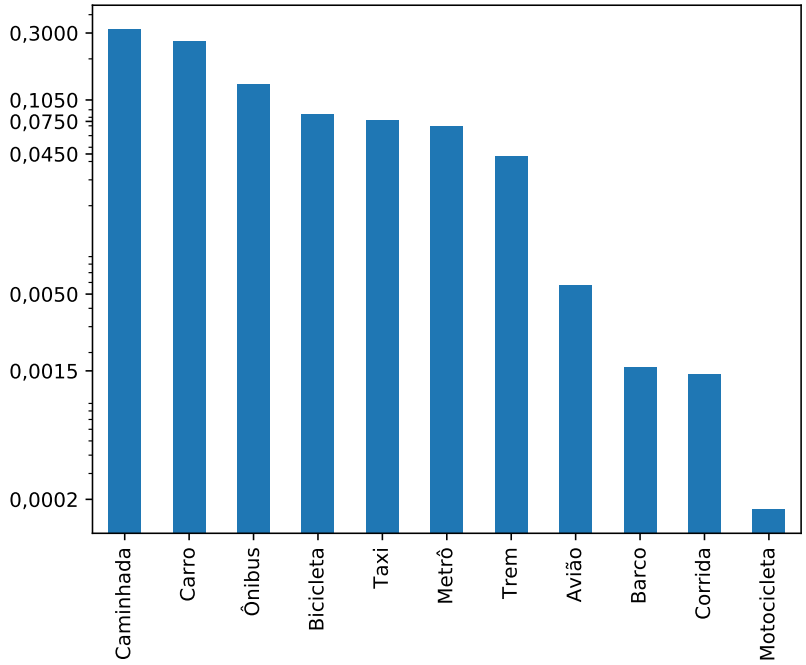

Figura 4.1: Porcentagem de dados relativos a cada meio de transporte na base de dados Geolife.

Como pode ser visto na Figura 4.1, cerca de 30% dos pontos da amostra foram coletados por pessoas que estavam caminhando. O segundo meio de transporte mais frequente é o carro, com cerca de 26% de ocorrência na amostra. Como meio de transportes menos frequentes, estão: motocicleta (0,01%), corrida (0,14%) e barco (0,15%).

4.2.2 Trajetórias de barcos

A segunda base de dados contém informações sobre trajetórias de barcos. Esta base é um subconjunto da base de dados utilizada por de Souza et al. [13]. A classe relacionada a cada ponto é referente ao tipo da atividade do barco naquele instante, mais especificamente se o barco está em rota de pesca ou não. As informações sobre a atividade do barco foram geradas por especialistas que anotaram as trajetórias de acordo com as características do movimento do barco.

A base de dados contém dados referentes às trajetórias de dezesseis navios. A base de dados contém informações sobre a localização geográfica (latitude e longitude), a data e a hora e velocidade sobre a terra, do inglês *speed over ground (SOG)*, e a variável alvo relativa ao comportamento do barco naquele ponto (se a atividade é pesca ou não). Existem 512.749 pontos associados às trajetórias dos barcos. Os dados são relativos ao período de junho de 2012 a dezembro de 2013. A mediana da frequência de coleta dos dados é de 19 segundos. As trajetórias somadas têm cerca de 929.836 km percorridos.

Os comportamentos possíveis das sub-trajetórias são: pesca e não pesca. A base é formada por um número maior de pontos relacionados ao comportamento de pesca. Na Figura 4.2 pode ser vista a porcentagem de cada uma das classes presentes, na qual pode-se constatar que cerca de 70% dos pontos são referentes à atividade pesqueira.

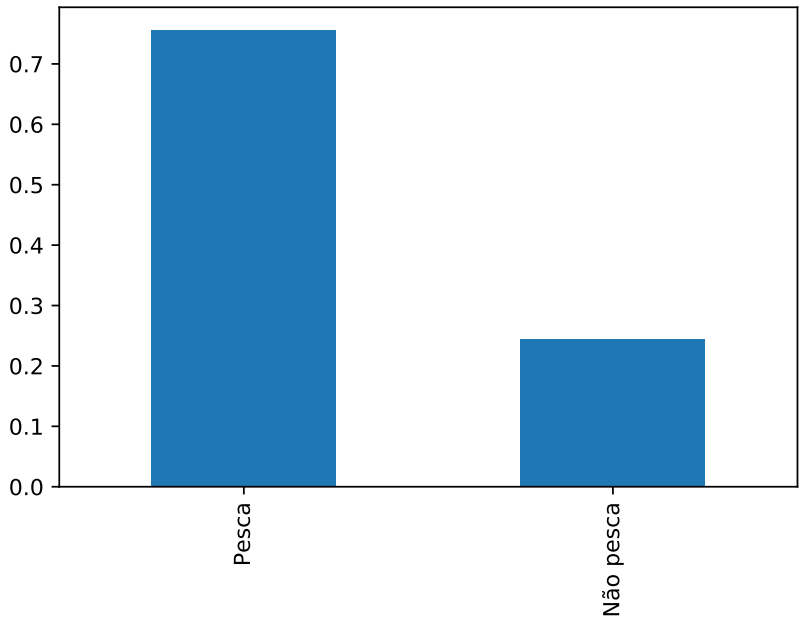

Figura 4.2: Porcentagem de dados relativos ao tipo de comportamento dos barcos.

4.2.3 Trajetórias de animais

Os dados das trajetórias dos animais são resultados de um projeto chamado *The Starkey Project* [2], que estuda a vida de animais selvagens. O estudo foi conduzido em uma parceria entre *Oregon Department of Fish*, *Wildlife* e *the USDA Forest Service*. Os animais acompanhados

[2]https://www.fs.fed.us/pnw/starkey/

estavam na área chamada *the Starkey Experimental Forest and Range* que é localizada no estado de Oregon nos Estados Unidos da América. A coleta dos dados teve início no ano de 1989 e duração de 10 anos [54].

Os dados contêm trajetórias de 259 animais que tinham seus movimentos rastreados por sensores. A base de dados é composta por 287.092 instâncias relativas aos pontos que os animais passaram. Os dados coletados são esparsos, tendo como valor mediano de tempo entre as coletas dos dados 5944,50 segundos. Como a área é controlada, os dados das trajetórias dos animais foram enriquecidos com um grande número de atributos obtidos através do mapeamento do local. A base de dados é composta pelos seguintes atributos: Segundos desde o início da coleta (StarkeyTime), Identificador do rastreador (RadNum), Peso atribuído a cada observação (Obswt), Profundidade do solo (SoilDpth), Porcentagem de inclinação (PerSlope), Seno da direção da inclinação (SINAspct), Cosseno da direção da inclinação (COSAspct), Convexidade (Convex3), Distância a fonte de água mais próxima pelo pasto do gado(DistCWat), Fechamento da copa (Canopy), Elevação do terreno (Elev), Distância para uma fonte de água mais próxima em pastos a prova de ungulados (DistEWat), Distância para estrada pública mais próxima (DistOPEN), Distância para estrada administrativa mais próxima (DistRSTR), Distância para estrada mais próxima (DistCLSD), Distância para a cerca mais próxima (DistEFnc), Produção de forragem (ForgProd), Distância para borda mais próxima (DistEdge).

Os animais que tiveram suas trajetórias coletadas e salvas na base de dados são: alces, veados e gado. Existe dentre os pontos uma desproporção em relação aos animais, pois os alces aparecem mais vezes que os outros dois grupos. Na Figura 4.3 pode ser vista a porcentagem de cada uma das classes presentes. Pode-se observar que os dados de alces representam mais de 50% dos dados, enquanto veados e gado representam cerca de 25% e 20% das trajetórias na base, respectivamente.

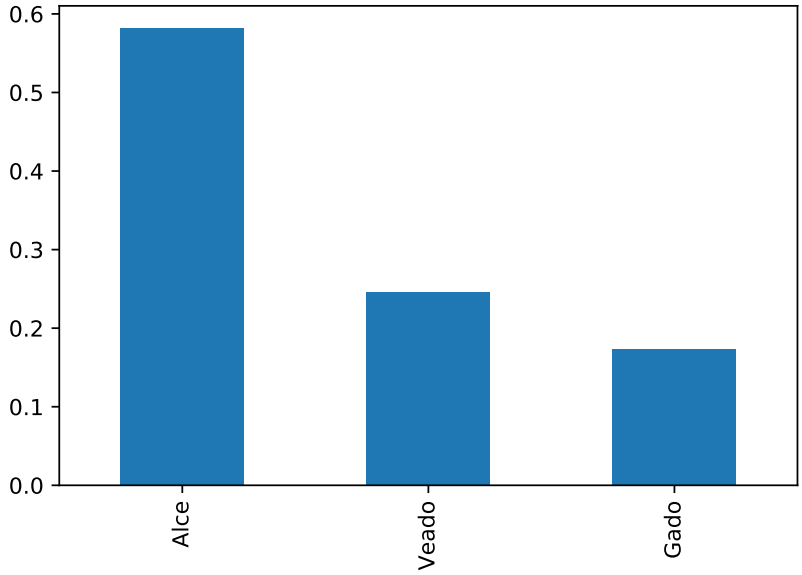

Figura 4.3: Porcentagem de dados relativos a cada tipo de animal.

4.3 Biblioteca Trajlib

Em muitos casos, os atributos associados a uma trajetória podem ser calculados a partir do conjunto original de pontos espaço-temporais. Dependendo do atributo, pode-se levar em consideração as informações sobre o espaço, o tempo, o tipo de segmentação da trajetória, dentre outras especificidades dela. Para facilitar este processo de geração de atributos, calculados a partir das informações de espaço e tempo relacionados à trajetória é que foi desenvolvida a biblioteca *Trajlib*.

A biblioteca foi criada pelo grupo de pesquisa do *Institute for Big Data Analytics* na *Dalhousie University*. Ela foi estendida nesta pesquisa por dois aspectos: a adição de um conjunto maior de características calculáveis a partir dos dados primitivos e a implementação do algoritmo de segmentação CB-Smot. A biblioteca foi projetada para a realização de experimentos de predição do tipo de meio de transporte utilizado durante uma trajetória [17]. O estudo demonstrou bons resultados, considerando a utilização dos atributos propostos junto com a redução de ruídos presentes nos dados. Apesar do propósito inicial da biblioteca ser lidar com dados relacionados ao domínio de predição de tipo de meio de transportes, os atributos gerados são gerais e podem ser usados em diferentes domínios.

Como discutido por Christ et al. [9], o processo de engenharia de atributos em uma série

cronológica demanda tempo para a extração de atributos significativos. Os autores propõem uma biblioteca *Python*, que calcula um conjunto de atributos que podem ser trabalhados em conjunto com outras bibliotecas *Python*, como *Pandas* e *Scikit-learn*, para análises exploratórias e aplicações de ciência de dados. Trajetórias também são séries temporais, todavia, com informações extras sobre o espaço que não seriam consideradas na utilização de bibliotecas como *Tsfresh*, que é focada em geração de atributos para séries temporais. Assim, a *Trajlib*, tal qual a *Tsfresh*, tem o propósito de auxiliar no processo de engenharia de atributos e as especificidades das trajetórias são levadas em conta ao longo da geração de atributos.

A biblioteca *Trajlib* é construída através da linguagem de programação *Python* na versão 3.7. Ela tem como dependência os pacotes *Numpy* e *Pandas*, possui o código aberto e encontra-se disponível no GitHub[3]. A biblioteca recebe como entrada um arquivo CSV, que deve ter no mínimo as colunas latitude, longitude e instante de tempo. O arquivo pode conter outras colunas associadas ao domínio da trajetória e não está restrito às colunas espaço temporais. Cada linha presente no arquivo é relativa a um ponto espaço-temporal. A biblioteca tem um exemplo de uso[4] com dados de trajetórias de animais.

A *Trajlib* é composta por três partes principais. A primeira consiste em uma classe representativa da trajetória, a qual recebe os dados que são carregados a partir do CSV e métodos úteis associados a pré-processamentos, exportação e visualização. A segunda parte é relativa à segmentação das trajetórias e influencia diretamente na criação de atributos relativos a segmentos. A forma de segmentar uma trajetória está diretamente relacionada ao propósito do estudo. A terceira parte é o foco principal da biblioteca e consiste na geração dos atributos.

Os atributos são calculados para cada um dos pontos e então, caso necessário, são resumidos para representar um segmento ou uma trajetória completa, dependendo da forma de segmentação escolhida. Os métodos estatísticos utilizados como sumários dos atributos dos pontos são: valor mínimo, valor máximo, média aritmética, mediana, desvio padrão, 10º percentil, 25º percentil, 50º percentil, 75º percentil, 90º percentil. Para cada atributo, esses valores são calculados levando em consideração os valores associados a cada um dos pontos espaço-temporais.

Uma das extensões dadas à biblioteca *Trajlib* foi a adição de um conjunto maior de ca-

[3]https://github.com/metemaad/TrajLib

[4]https://github.com/metemaad/TrajLib/blob/master/TrajLib_Usage_Example.ipynb

racterísticas calculáveis a partir dos dados originais. Para tanto, foram adicionados atributos relacionados ao tempo, como dia da semana, e se é dia útil ou fim de semana. Também foram adicionados atributos relacionados às trajetórias completas ou segmentos, tais quais: duração da(o) trajetória(segmento) em segundos, distância do ponto inicial ao ponto final, distância percorrida por dia, tempo de total de paradas, número de auto-interseções e distância total percorrida.

Além disso, outra extensão feita nesta pesquisa está relacionada à implementação do algoritmo de segmentação CB-Smot [49], que tem a divisão das trajetórias com base em paradas e movimentos efetuados pelo objeto. A detecção de parada e movimento é realizada através da criação de grupos de pontos com comportamentos comuns. Este tipo de segmentação é importante em casos como a classificação de veículos, pois quando há uma mudança de veículo em uma mesma trajetória, ela acontece durante uma parada. Assim, a segmentação baseada em paradas auxilia no processo de classificação de modo que os segmentos têm apenas um tipo de veículo utilizado.

A biblioteca *Trajlib* auxilia na geração de atributos durante a fase de engenharia de atributos. Além da possibilidade de uso dos atributos já implementados, a biblioteca pode ser estendida de acordo com a necessidade do projeto para novos atributos, uma vez que é expansível. Portanto, a biblioteca é genérica, podendo gerar atributos de forma independente do domínio associados com os dados das trajetórias. Desta forma, novos atributos e uma nova forma de segmentação foram adicionados à *Trajlib* durante esta pesquisa, deixando-a ainda mais robusta e contribuindo para que a geração atributos em outros projetos possa ser facilitada.

4.4 Geração de atributos

Os dados originais disponíveis nas bases de dados consistem na localização geográfica, no tempo e alguns atributos que podem ser associados ao contexto. Todavia, muitos dos atributos utilizados na literatura podem ser derivados a partir dos dados de localização e tempo e podem ser utilizados em diferentes domínios. A combinação dos dados disponíveis para derivar informações é parte do processo de engenharia de atributos. Assim, é relevante para o entendimento dos atributos que as bases de dados utilizadas na avaliação tenham o maior

número de atributos possíveis.

Os atributos que foram gerados e acrescidos às bases de dados durante esta pesquisa são aqueles que aparecem comumente na literatura. As definições dos atributos gerados foram retiradas dos artigos de destaque na área [10; 11; 80]. Os atributos adicionados aos conjuntos de dados foram calculados a partir das informações dos pontos espaço-temporais das trajetórias. Os atributos são atribuídos a cada um dos pontos e são, em uma maioria, calculados com base nos pontos e seus predecessores.

Para todas as bases de dados utilizadas neste trabalho foram adicionados os atributos descritos na Tabela 4.4 através da biblioteca *Trajlib*. Os atributos apresentados são genéricos e podem ser aplicados a todos os domínios, possibilitando a adição deles a todas as bases de dados utilizadas neste trabalho. Os dados associados aos domínios das bases foram mantidos e os valores calculados foram acrescidos, agregando maiores informações às trajetórias.

Tabela 4.4: Atributos adicionados às bases de dados gerados pela biblioteca *Trajlib*

Atributo	Descrição
Distância	Distância entre um ponto e o seu predecessor.
Velocidade	Velocidade do ponto, calculada utilizando a informação do tempo e distância percorrida entre o ponto e o seu predecessor.
Aceleração	Aceleração do ponto, calculada utilizando a informação do tempo e velocidade entre o ponto e o seu predecessor.
Bearing	Ângulo entre a linha formada entre o ponto e um ponto de referência (ex. norte verdadeiro) e a linha entre o ponto e o seu sucessor.
Bearing rate	Diferença absoluta entre o bearing do ponto e o bearing do seu sucessor.
Rate of bearing rate	Diferença absoluta entre o bearing rate do ponto e o bearing rate do seu sucessor.
Rate of turn	O arco tangente da diferença entre as longitudes do ponto e seu predecessor dividido pela diferença entre as suas latitudes
jerk	A diferença entre a aceleração de um ponto e a aceleração do seu predecessor, dividido pela diferença do tempo de coleta no ponto e o tempo de coleta do seu predecessor
is_weekday	Valor binário que indica se o dia de coleta do ponto é na semana ou no fim de semana.
hour	Hora no formato 24h do momento em que o ponto foi coletado.
minute	Minuto do momento que o ponto foi coletado.
weekday	Dia da semana que o ponto foi coletado.

4.5 Desenvolvimento da taxonomia

Nesta seção o processo de desenvolvimento da taxonomia Trajtax é apresentado. O primeiro passo consiste na seleção da meta-característica e condição de uma parada. Este passo é descrito na subseção 4.5.1, apresentada a seguir. Posteriormente, nas sub-seções 4.5.2 e 4.5.3, são apresentados os processos aplicados ao contexto de atributos de trajetórias, seguindo as abordagens: empírica para conceitual e conceitual para empírica.

4.5.1 Passo 1: Seleção da meta-característica e condição de parada

Antes da definição da meta-característica, faz-se necessária a apresentação do propósito da taxonomia. Isso porque a meta-característica serve como base no processo de criação da taxonomia e deve ser associada ao propósito dela. Nesse sentido, o propósito da taxonomia deste trabalho é facilitar o entendimento dos atributos utilizados em atividades de aprendizagem de máquina com dados de trajetórias. Esse entendimento dos atributos é importante pois um dos desafios associados aos modelos de trajetória é a interpretabilidade dos mesmos, por isso, o foco desta pesquisa são atributos definidos na literatura que não advêm de transformação, que impossibilitam a interpretação dos mesmos.

Tendo em vista o propósito da taxonomia, é possível a definição da meta-característica. A meta-característica utilizada aqui pode ser definida como a avaliação das especificidades dos atributos interpretáveis no domínio de trajetórias, porque, as características propostas na taxonomia são propriedades que explicam os tipos de atributos existentes. A partir da meta-característica, é possível a derivação das características e dimensões da taxonomia. As dimensões, por sua vez, são variáveis consideradas importantes associadas ao propósito da taxonomia. As características são os possíveis valores que estas variáveis assumem quando consideradas no contexto proposto.

As condições de parada são classificadas em dois grupos: objetivas e subjetivas. Com relação às objetivas, a condição de parada fundamental é que, na taxonomia resultante, todas as características devem ser mutuamente excludentes e coletivamente exaustivas deixando a taxonomia em conformidade com a definição. Além desta, a outra condição de parada objetiva escolhida é que cada dimensão seja única e que cada característica seja única na sua dimensão. Essa condição evita que a mesma informação seja repetida de diferentes formas, garantindo que a taxonomia seja sucinta.

Com relação às condições subjetivas, optou-se o seguinte: para que a taxonomia Trajtax seja considerada completa, ela deve ser concisa, robusta, abrangente, extensível e explanatória. Diferentemente das condições objetivas, as condições subjetivas não têm métricas de avaliação definidas e são, em alguns casos, difíceis de serem avaliadas. Essas características estão associadas com a utilidade da taxonomia e destacam pontos que são essenciais para que a taxonomia seja usada.

Uma vez definidos os pontos iniciais que guiam o processo de desenvolvimento de uma

taxonomia, o objetivo, a meta-característica e as condições de parada, são apresentadas a seguir:

- **Objetivo da taxonomia:** facilitar o entendimento dos atributos utilizados em atividades de aprendizagem de máquina com dados de trajetórias.
- **Meta-característica:** a avaliação das especificidades dos atributos interpretáveis no domínio de trajetórias.
- **Condições de parada objetivas:**
  - Todas as características devem ser mutuamente excludentes e coletivamente exaustivas.
  - Cada dimensão única e cada característica seja única na sua dimensão.
- **Condições de parada subjetivas:**
  - Concisa;
  - Robusta;
  - Abrangente;
  - Extensível e
  - Explanatória.

A partir da obtenção das definições anteriormente citadas, definiu-se qual das abordagens seria seguida na primeira iteração. É possível a aplicação da abordagem empírica para conceitual ou conceitual para empírica. O processo empírico para conceitual foi escolhido como primeira fase deste processo, conforme veremos a seguir.

4.5.2 Passo 2: Abordagem empírica para conceitual

Com as bases de dados de três domínios diferentes foi possível a derivação de informações mais gerais, levando em consideração semelhanças e diferenças elas. Assim, a primeira parte do processo foi a seleção dos dados a serem usados. As bases de dados utilizadas nesta

pesquisa são descritas na seção 4.2 e consistem em trajetórias de animais, barcos e mobilidade urbana. As escolhas destas bases de dados estão associadas à variedade de domínio e à disponibilidade.

A próxima etapa foi encontrar características comuns que estão diretamente associadas à meta-característica. Para tal, nesta fase recorremos à teoria de taxonomia numérica, definida por Sokal e Sneath [60] como uma avaliação numérica da similaridade utilizando a distância entre as unidades taxonômicas e organizando-as em taxas de acordo com suas similaridades. Portanto, é possível avaliar os valores disponíveis de acordo com suas similaridades nos diferentes domínios. Com o resultado, é possível encontrar características em comum.

No caso de atributos, busca-se agrupar os que têm valores semelhantes. Dessa forma, para o cálculo de distância dos atributos, foi usado o coeficiente de correlação de Pearson, que avalia duas variáveis e indica a existência, ou não, de um relacionamento linear entre elas. Com os valores das correlações entre pares, montou-se uma matriz de correlação e os valores foram organizados em grupos de acordo com a proximidade entre os atributos. Nesta fase, outras aplicações para agrupar os atributos eram possíveis, entretanto, a matriz de correlação de Pearson foi a escolhida, pois ela dá uma ideia de distância entre as instâncias avaliadas de maneira não restrita apenas ao agrupamento dos atributos.

Para esta fase do processo, os dados originais foram acrescidos dos dados gerados pela biblioteca *Trajlib*, descrita na seção 4.3. . A partir dos conjuntos de dados, foram geradas matrizes de correlação. Como resultado, os atributos são ordenados de acordo com os valores resultantes das matrizes de correlação. Para a geração das matrizes e a ordenação dos valores, foi utilizada a linguagem de programação *Python* com as bibliotecas *Pandas*, *Seaborn* e *Scipy*.

Matrizes de correlação

As matrizes de correlação contêm os valores do coeficiente de correlação de Pearson para cada um dos pares de variáveis. Esses valores variam entre menos um e um. O valor de coeficiente zero implica que as variáveis não são linearmente dependentes e, quando o coeficiente está próximo aos extremos, indica que são linearmente dependentes. Para facilitar o entendimento, cores foram estabelecidas para representar os valores. Em ocasiões nas quais for obtido algo próximo de zero, os valores assumem a cor cinza, e quanto mais próximo for

dos extremos, assumem as cores azul (um) e vermelho (menos um).

O primeiro conjunto de dados avaliado é relativo à base de dados relacionada ao movimento de animais. O número de atributos presentes nessa base é maior que as outras duas trabalhadas nesta pesquisa. O número de atributos presentes está associado ao fato de que os dados derivam de um experimento controlado, e assim diferentes informações foram associadas às localizações em que as trajetórias acontecem. Como primeira parte do processo, os dados foram organizados através do resultado do algoritmo de agrupamento hierárquico, com a biblioteca *Scipy*. O resultado pode ser visto na Figura 4.4.

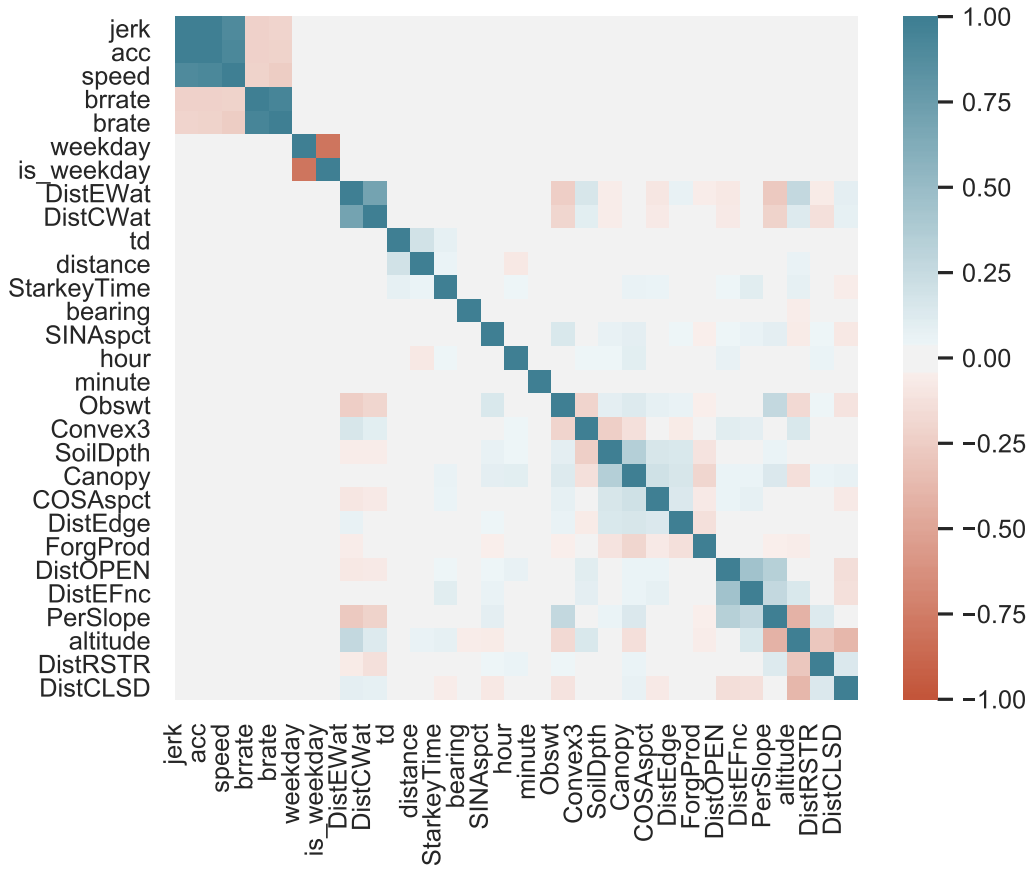

Figura 4.4: Correlação entre os atributos da base de dados de trajetórias de animais, organizados utilizando agrupamento hierárquico.

É possível observar na Figura 4.4, primeiramente, a presença de um grupo composto por aceleração, velocidade, jerk, bearing rate e rate of bearing rate. Esse primeiro grupo parece estar relacionado às características dos movimentos realizados durante as trajetórias. Em seguida, aparece um grupo adicional composto pelos atributos: dia da semana e fim de

semana (ou não). Esse grupo de atributos parece estar relacionado ao tempo. Há também um terceiro grupo que aparenta ser ligado aos atributos adicionais incluídos.

Todavia, tendo em vista os significados dos atributos, é viável reorganizar os atributos. Em outras palavras, unir os atributos que são semanticamente parecidos, mesmo que não estejam relacionados entre si. Como existem diferentes domínios associados aos dados de trajetórias, é possível que, em alguns domínios, os atributos não sejam linearmente relacionados, mas tenham informações sobre a trajetória que representam uma mesma área. A organização dos atributos em grupos, de acordo com o que valor do atributo representa em relação a trajetória, é apresentada na 4.5.

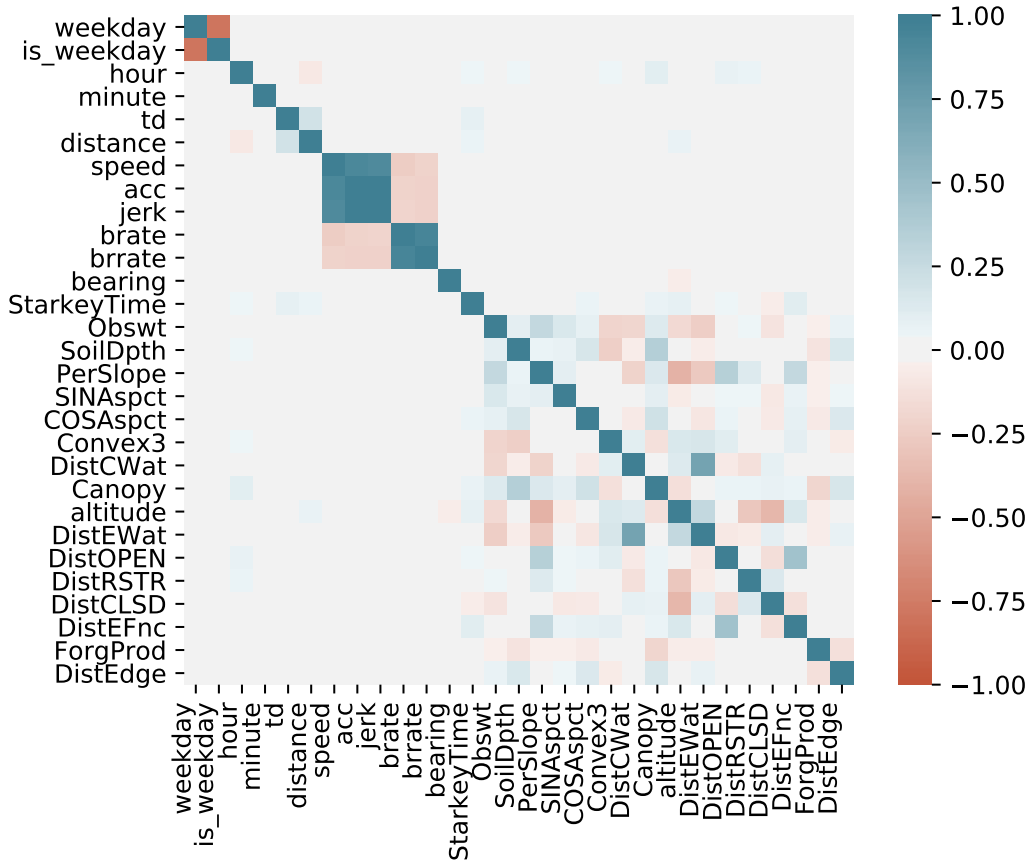

Figura 4.5: Correlação entre os atributos da base de dados de trajetórias de animais organizados, levando em conta a proximidade e a semântica dos atributos.

A primeira modificação que pode ser vista é a aproximação da hora e do minuto com o dia da semana e ao atributo que implica em ser dia de semana ou fim de semana. Todos os elementos estão ligados com o momento que o dado foi coletado. Assim, mesmo que

não estejam linearmente relacionados, faz sentido entende-los como de um mesmo grupo. A segunda mudança na organização dos atributos é em relação a posição da distância espacial e da distância temporal ao ponto anterior, além do *bearing* para próximo dos atributos que descrevem o movimento realizado na trajetória.

Com as mudanças feitas, ficou mais clara a relação entre os dados adicionados devido à natureza do experimento. No primeiro caso, existia bastante espaço cinza no meio dos dados, o que indica a não existência de correlação linear entre eles. Esses dados podem ser julgados informações contextuais, pois não estão diretamente associados à trajetória, mas são informações adicionais que os enriquecem.

O mesmo processo foi aplicado aos dados de mobilidade urbana. O número de atributos presentes é menor pois antes das adições dos atributos com a biblioteca Trajlib, a base de dados consistia em informações de localização, tempo, altitude e meio de transporte utilizado. Os elementos foram organizados, assim como no caso anterior, considerando o agrupamento deles e as informações relacionadas à semântica desses dados. A matriz de correlação resultante é apresentada na Figura 4.6.

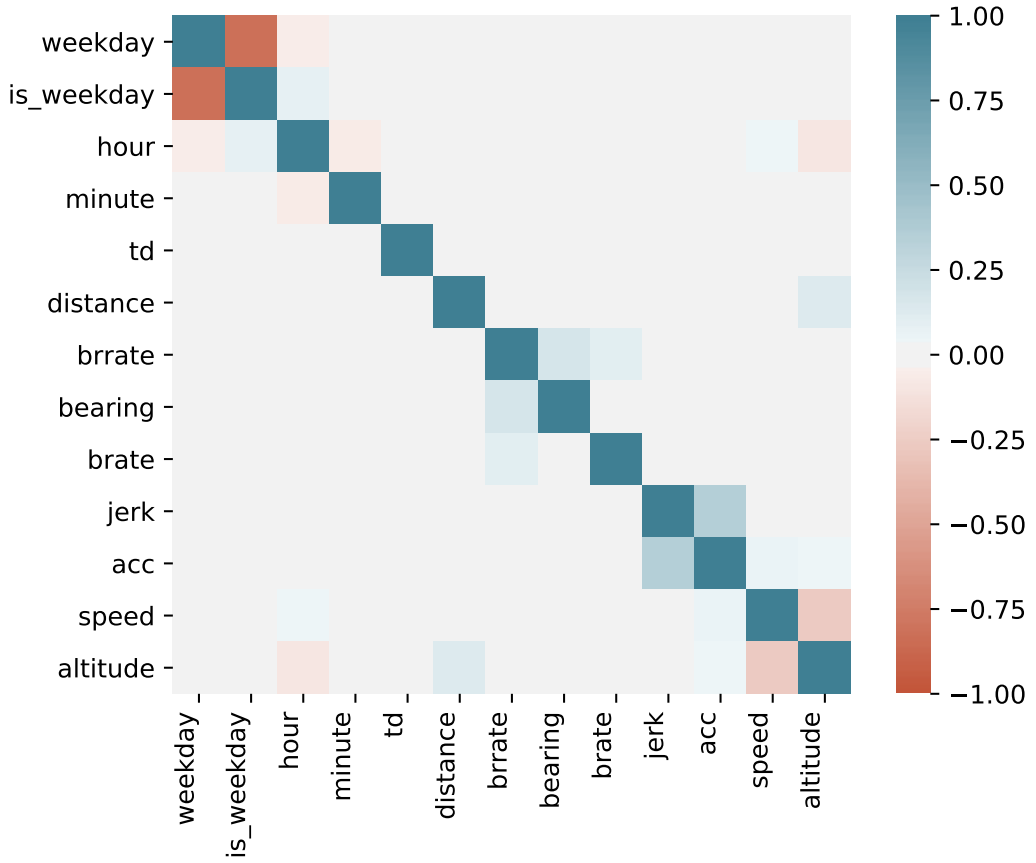

Figura 4.6: Correlação entre os atributos da base de dados de trajetórias representativas da mobilidade urbana organizados considerando a proximidade e a semântica dos atributos.

Como esperado, a correlação dos dados não é igual à do caso anterior, pois cada domínio tem suas especificidades. Todavia, é possível destacar pontos em comum diante dos dados. Apesar de menor, na relação entre os atributos desta base de dados, é possível ver a correlação entre alguns atributos que dão a mesma ideia dos grupos anteriores. Primeiramente, as variáveis de tempo (dia da semana, semana ou fim de semana, hora e minuto), apesar de não ser forte, estão correlacionadas. Em seguida, os valores de distância temporal e distância geográfica ao ponto anterior unidos as informações de bearing, bearing rate, rate of bearing rate, aceleração, jerk e velocidade representam o movimento. Por último, o valor de altitude pode ser considerado um valor contextual.

Por fim, os dados contendo trajetórias de barcos também foram avaliados no que tange aos coeficientes de correlação de Pearson dos seus atributos. Assim como a base de dados associada à mobilidade urbana, esta base de dados inicialmente era composta da localização

geográfica dos pontos, o tempo relacionado ao instante que foram coletados e o valor de *speed over groud (SOG)*. Para a melhor visualização dos dados, tal qual nos casos anteriores, os dados foram ordenados de acordo com o grau de correlação e rearranjados a partir do contexto vinculado aos atributos. A matriz de correlação resultante é apresentada na Figura 4.7.

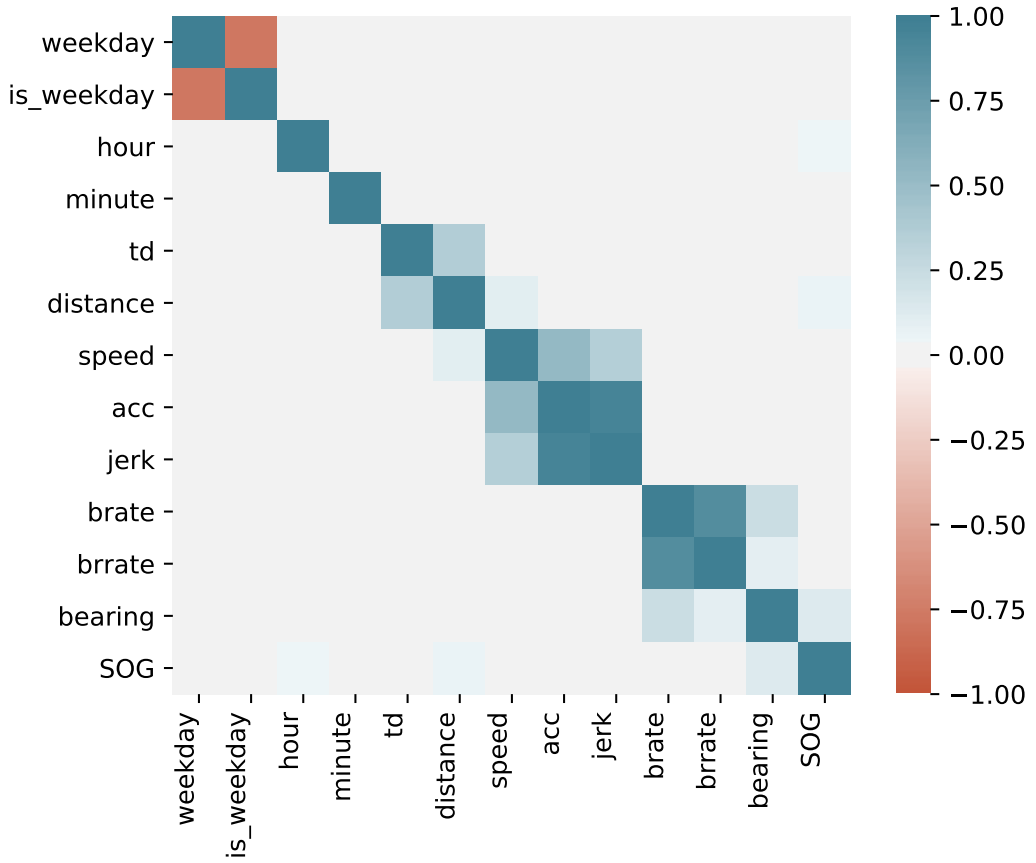

Figura 4.7: Correlação entre os atributos da base de dados de trajetórias de barcos organizados a partir da proximidade e da semântica dos atributos.

Assim como nos casos anteriores, é possível destacar o grupo de atributos que descrevem o movimento representado nas trajetórias. Os dados relacionados ao tempo não são todos linearmente dependentes, todavia, as semânticas dos seus valores são próximas. Finalmente, o dado de speed over ground pode ser considerado um atributo contextual. Esses três grupos estão presentes nas bases de dados e podem servir como ponto de partida para entendimento dos domínios associados às trajetórias.

Os atributos agrupados têm traços em comum que podem explicar de modo geral quais

as áreas estão possivelmente relacionadas aos atributos de trajetórias. É necessário pontuar que existem restrições no uso da matriz de correlação de Pearson, como não cobrir relacionamento, ou o fato de que valores muito correlacionados linearmente não agregam em um processo de aprendizagem de máquina. Entretanto, para a aplicação necessária nesse contexto, as restrições não impedem a utilização da distância, uma vez que a teoria ligada aos atributos também é considerada na organização deles em grupos. Na subseção seguinte são discutidas as dimensões e características derivadas deste experimento.

Como resultado da avaliação empírica, tem-se os grupos formados por atributos que possuem traços semelhantes e, assim, é possível se chegar às características e dimensões a serem usadas na taxonomia. Essa fase do processo consiste em, a partir dos resultados, avaliar a parte conceitual relacionada à cada uma das categorias. Uma vez que os conceitos são avaliados, as características são agrupadas de acordo com as suas semânticas.

a) Dimensão Área

A primeira característica derivada dos dados está relacionada à dinâmica do movimento. A forma como um objeto se movimenta pode ser trabalhada em diferentes áreas como entrada para modelos de aprendizagem de máquina. Muitos fatores estão ligados ao padrão de movimentação de um objeto. Por exemplo, um veículo em condições normais pode andar muito mais rápido do que uma pessoa caminhando, porém, ele tem seu caminho restringido por uma estrada. Ainda nesse mesmo exemplo, outra característica que pode ser vista é que uma pessoa pode caminhar livremente enquanto um automóvel precisa respeitar as leis associadas ao trânsito. A união de todos esses fatores relevantes pode explicar diferentes trajetórias com relação ao movimento de um objeto.

Desta forma, atributos associados à dinâmica do movimento podem acrescentar informações importantes ao processo de aprendizagem de máquina. Esses atributos aqui descritos podem estar vinculados à forma do movimento, como por exemplo, a curvatura presente em uma trajetória, como também com as características físicas do movimento tal qual a velocidade com em que o objeto móvel se movimenta. Como esses atributos estão diretamente ligados ao movimento, podem, em sua maioria, ser calculados a partir dos dados dos pontos espaço-temporais e do tempo associado a eles.

A segunda característica é sobre os atributos temporais. Eles estão associados ao mo-

mento em que o movimento ocorre. Em certos casos, a hora na qual uma trajetória acontece pode estar diretamente relacionada com o propósito da mesma. Uma pessoa, por exemplo, em um dia normal de trabalho, sai de casa no horário padrão do seu trajeto de deslocamento, ou uma pessoa sai de casa no meio da noite em um dia de semana, pois precisa de assistência médica. Por essa razão, os intervalos de tempo são importantes fontes de informação.

Além disso, em alguns casos, o horário em que uma trajetória se dá pode afetar as características do movimento. Por exemplo, um carro anda mais devagar à noite devido à menor visibilidade na estrada. Uma pessoa, por outro lado, tende a caminhar o mais rápido possível no escuro para que possa chegar a um local iluminado. Ou ainda, uma pessoa prefere pegar uma rota alternativa, pois sabe que neste horário a rota normal está parada devido ao trânsito. Assim, atributos temporais são importantes para o entendimento geral do movimento em uma trajetória.

Por fim, a última característica está associada aos atributos conceituais. Este grupo de atributos adiciona informações contextuais aos dados da trajetória, enriquecendo os dados para a obtenção de melhores resultados nos processos de aprendizagem de máquina. Eles, normalmente, agregam informações semânticas à trajetória que podem ser obtidas por meio de sensores adicionais ou combinando bases de dados que contêm informações relacionadas ao espaço e ao tempo em que a trajetória aconteceu. Um exemplo de informação contextual que pode ser utilizada em diferentes casos, é a condição climática nos pontos da trajetória.

Uma vez que as características foram avaliadas, elas recebem agrupamento de acordo com a semântica associada. As características derivadas representam conceitualmente os possíveis domínios vinculados aos atributos. Dessa maneira, como resultado da avaliação empírica, tem-se a dimensão da taxonomia nomeada área. A dimensão tem como características: dinâmica do movimento, temporal e contextual. Essas características encaixam-se na definição formal da taxonomia, pois são mutuamente excludentes e coletivamente exaustivas. Assim, a primeira versão da taxonomia (denominada de taxonomia parcial 1) é apresentada na Figura 4.8.

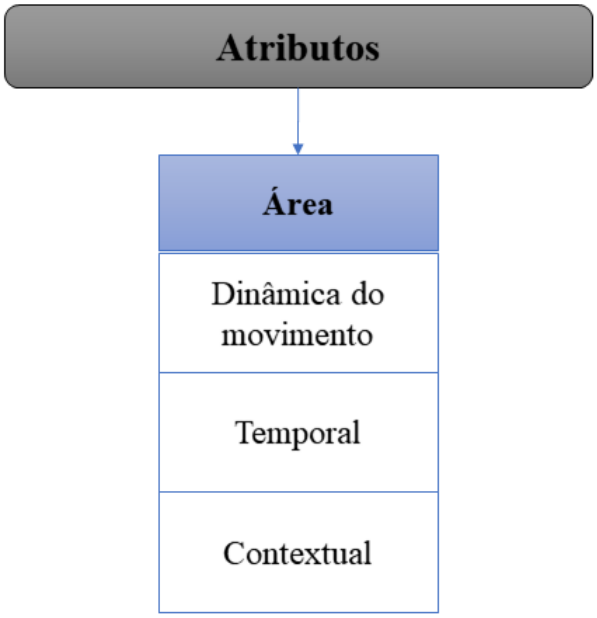

Figura 4.8: Taxonomia parcial 1

Definidas as características e a dimensão, é possível representá-los através de um diagrama de classes. Desta forma, é possível representar o relacionamento entre atributos e a dimensão área de forma estruturada. Na Figura 4.9 tem-se a representação da dimensão área como um diagrama de classes.

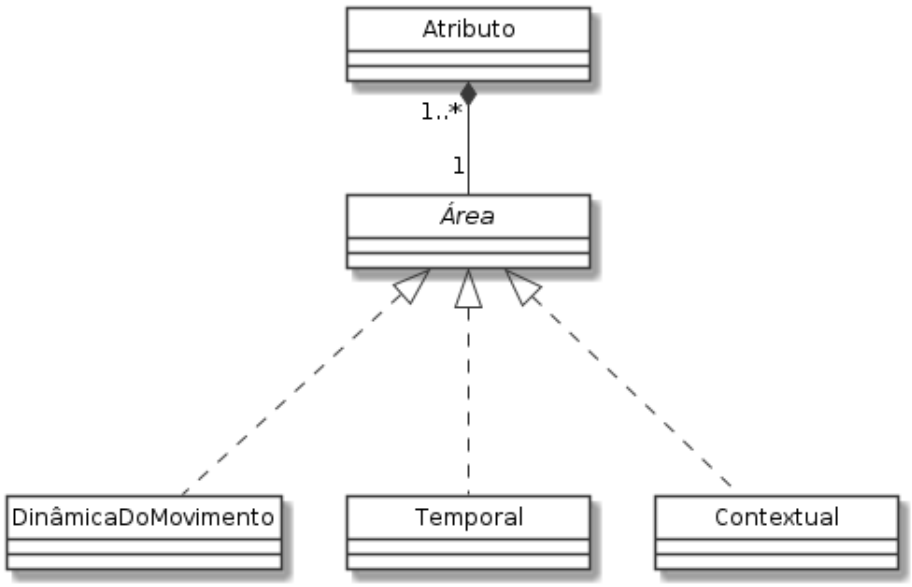

Figura 4.9: Diagrama da dimensão área utilizando um diagrama de classes

Como exemplo de aplicação, é possível descrever a ida de uma pessoa ao trabalho como uma trajetória percorrida em 48 minutos (tempo), com velocidade média de 30 km/h (dinâmica do movimento), com trânsito intenso (contexto); ou em 24 minutos (tempo), com velocidade média de 60 km/h (dinâmica do movimento), sem trânsito (contexto).

A taxonomia da Figura 4.8, gerada a partir da avaliação empírica, obedece às condições objetivas para satisfazer o critério de parada, porém ainda não pode ser considerada robusta, uma vez que explica apenas uma dimensão do problema. Por isso, o passo 2 do processo de geração da taxonomia deve ser repetido. A partir daqui, o processo foi repetido levando em consideração a abordagem conceitual para empírica, visto que a abordagem empírica para conceitual já foi realizada.

4.5.3 Passo 2: Conceitual para empírica

A parte conceitual para empírica, da segunda fase do processo, tem por objetivo levar também em conta os conceitos associados aos atributos de modo que a taxonomia seja a mais completa possível. Esta fase consiste em transformar os conceitos em características e então agrupá-las em dimensões. Uma vez que as dimensões e características são propostas, elas são avaliadas com relação a um conjunto de dados. Nesta seção, são, portanto, apresentadas as dimensões propostas nesta etapa.

Neste ponto da pesquisa, conceitos importantes foram levantados: o primeiro deles é sobre aquilo que a informação do atributo carrega. Então, deve-se entender qual é o escopo que está sendo considerado no projeto de aprendizagem de máquina. Outro ponto importante a ser considerado é se é possível o cálculo do valor do atributo a partir dos pontos espaço-temporais da trajetória. Por fim, é necessário compreender as restrições relacionadas, por exemplo, se um atributo pode ser usado sem problemas para todos os domínios de trajetória. Dentre as possíveis restrições associadas a um atributo, é relevante considerar se os atributos são dependentes de domínio, assim como se são dependentes de parâmetros. A partir dos conceitos indispensáveis destacados, as seguintes dimensões foram propostas:

a) Dimensão Aspecto

De forma geral, o atributo carrega uma informação sobre algum aspecto de um objeto estudado. No caso de trajetórias, os atributos podem ser relacionados a mais de um aspecto. Por exemplo, um atributo pode estar ligado ao tempo de um ponto espaço temporal, assim como ao espaço geográfico deste mesmo ponto. Portanto, dados os múltiplos aspectos, é importante a categorização dos atributos com relação a eles.

A primeira característica desta dimensão é relativa ao espaço geográfico. Isso é, a informação contida no atributo representa uma característica associada a um ponto geográfico pelo qual a trajetória passou. A segunda característica está relacionada com o tempo em que a trajetória ocorreu. A terceira característica é a combinação das duas primeiras onde o atributo contém uma informação espaço-temporal. Por fim, a última característica é relativa à informação contida no atributo que está vinculada ao objeto móvel.

Considerando as duas dimensões já definidas, tem-se a taxonomia parcial 2, apresentada na Figura 4.10.

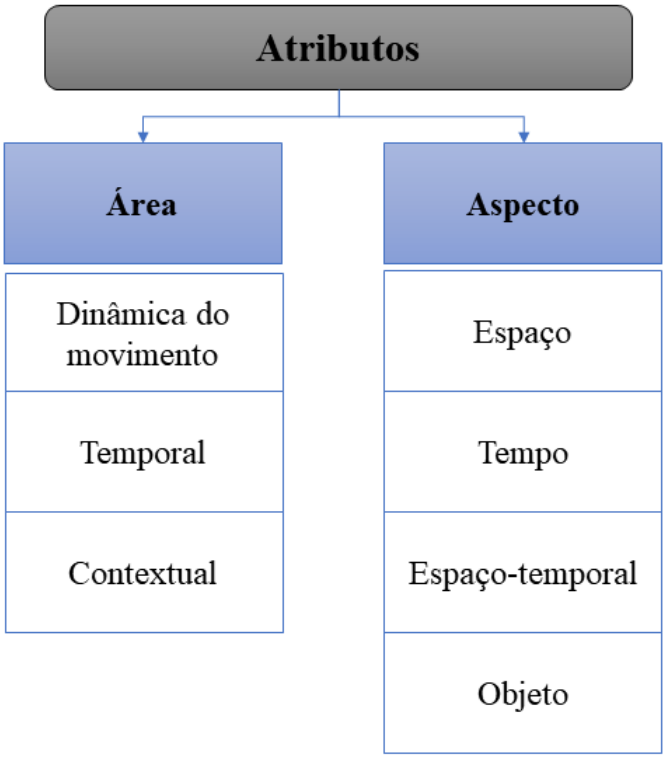

Figura 4.10: Taxonomia parcial 2

As características presentes nesta dimensão representam pontos que podem ser observados por pesquisadores para que atributos relacionados às trajetórias possam ser derivados. Assim como no caso da dimensão área, é possível associar a dimensão aspecto com os atributos através da utilização de diagrama de classes. O diagrama de classes representando esta dimensão pode ser observado na Figura 4.11

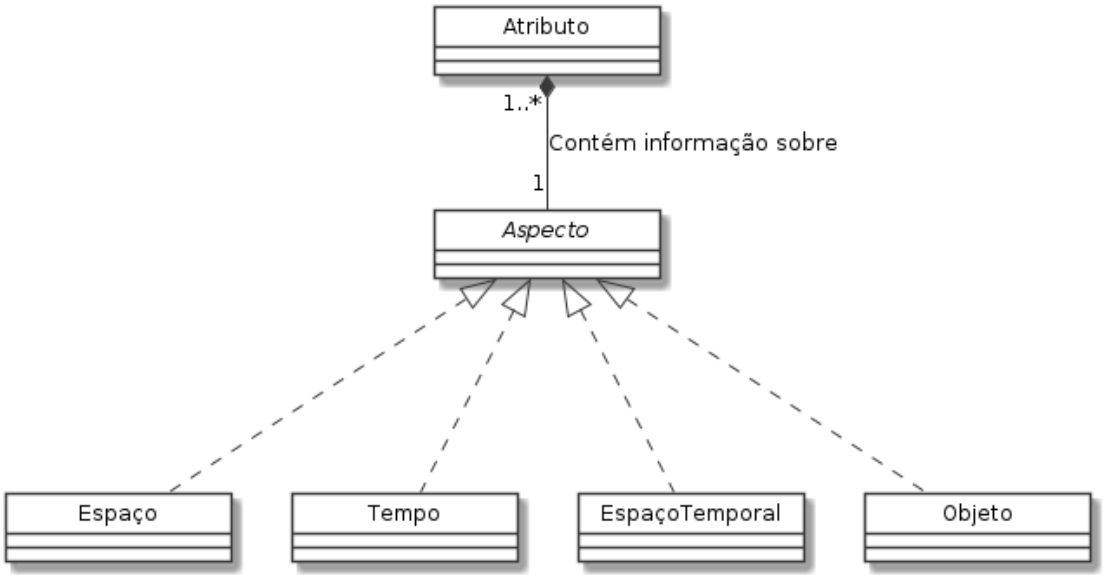

Figura 4.11: Diagrama de classes da dimensão aspecto

Os atributos velocidade, auto-interseção, rotações do motor e intervalo de tempo podem ser tomados como exemplo para simbolizar as diferenças destacadas por esta dimensão. O atributo de velocidade contém informações espaço-temporais do movimento, o atributo da auto-interseção traz informações acerca das localizações visitadas na trajetória, o atributo de rotações do motor está relacionado ao objeto em movimento e, por fim, o atributo intervalo de tempo está ligado ao tempo em que a trajetória ocorreu.

b) Dimensão Granularidade

A próxima dimensão proposta é relativa ao escopo do projeto. Um atributo pode ter informações de diferentes níveis de granularidade, de acordo com o foco do projeto. Assim, é importante entender a qual nível ele está associado para que possa ser utilizado de acordo com a necessidade. Os atributos podem conter informações sobre pontos da trajetória, trajetórias inteiras ou segmentos inteiros e múltiplas trajetórias, ou múltiplos segmentos.

Destaque-se que não é possível recorrer a informações de atributos associados às múltiplas trajetórias quando se lida com pontos. Todavia, é viável a utilização de técnicas estatísticas para utilização de dados associados a pontos para serem trabalhados no contexto de múltiplas trajetórias.

Adicionando a nova dimensão à taxonomia, tem-se a taxonomia parcial 3, apresentada na Figura 4.12.

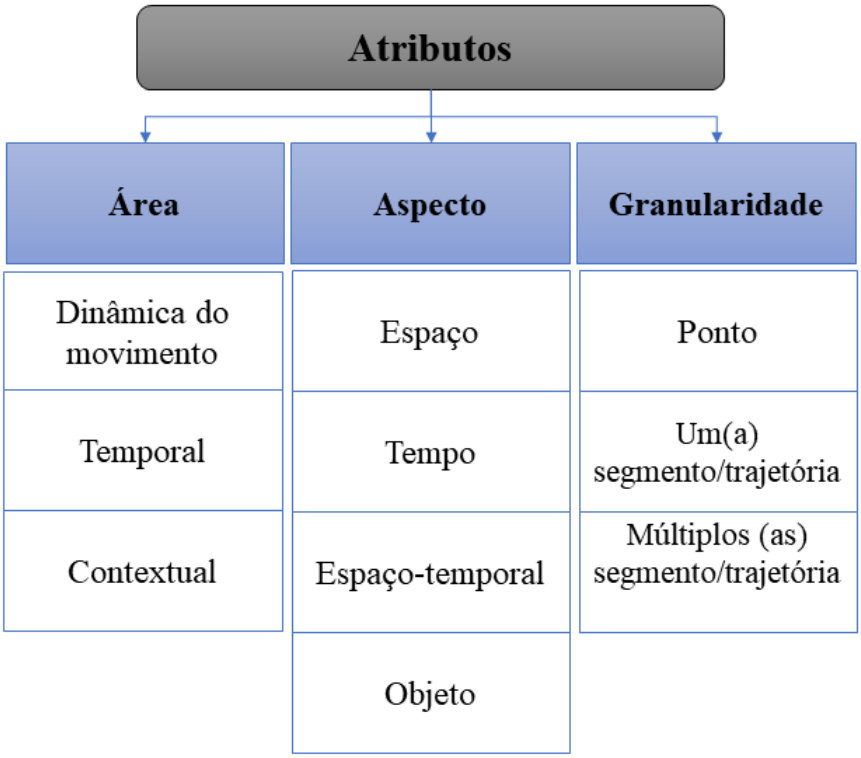

Figura 4.12: Taxonomia parcial 3

As características da dimensão granularidade podem ser organizadas em um diagrama de classes para facilitar o entendimento da sua aplicabilidade. O diagrama de classes representando a relação dos atributos acerca de diferentes granularidades pode ser observado na Figura 4.13.

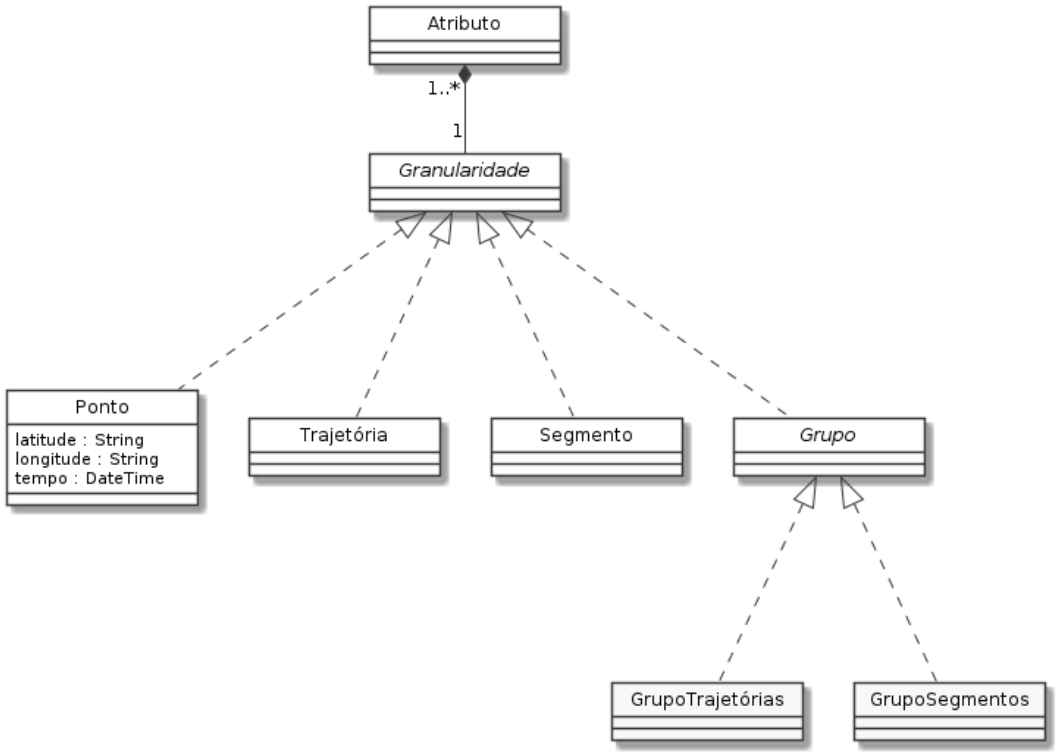

Figura 4.13: Diagrama de classes da dimensão granularidade

A granularidade do atributo está diretamente associada com o propósito do estudo. Para representar cada uma das características, é viável trabalhar exemplos de distintas aplicações. Um modelo de atributo associado a um ponto é a largura da estrada naquele ponto. Já considerando a trajetória completa, tem-se a duração do movimento. No caso de múltiplas trajetórias, é possível destacar o atributo quilômetros percorridos por mês.

c) Dimensão Fonte dos Dados

Devido à importância da clareza nos modelos no contexto de trajetórias, muitos dos atributos que foram propostos são derivações dos dados básicos que sofreram transformações manuais, que são normalmente relacionadas à teoria associada ao movimento. No entanto, também existem atributos, por exemplo, o número de rotações de um motor de um veículo usado em uma trajetória, que não podem ser obtidos por meio apenas os dados originais, mas também são importantes para utilização em modelos de aprendizagem de máquina. Assim, esta dimensão separa os atributos em duas características: calculável a partir dos dados iniciais ou não calculável.

Adicionando a nova dimensão à taxonomia, tem-se a taxonomia parcial 4, apresentada na Figura 4.14.

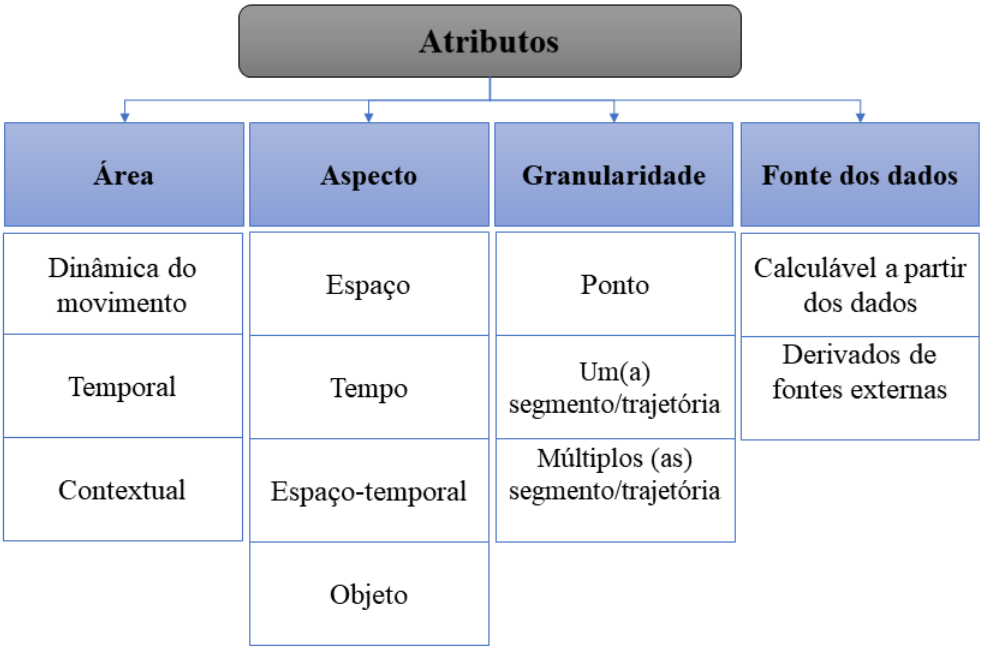

Figura 4.14: Taxonomia parcial 4

O relacionamento entre a dimensão fonte e os atributos é simples. Todo atributo está atrelado a duas opções de fonte: calculável ou derivado de uma fonte externa. A Figura 4.15 tem a representação deste relacionamento.

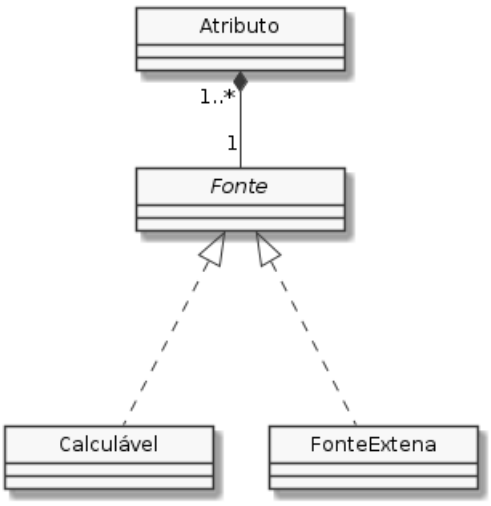

Figura 4.15: Diagrama de classes da dimensão fonte dos dados

A possibilidade de combinar os dados de trajetórias com outras fontes agrega bastante informações no processo de aprendizagem de máquina. Pode-se exemplificar a diferenciação entre estes dois tipos de atributos através dos atributos de tempo de utilização do veículo e velocidade. A velocidade pode ser calculada a partir das informações de latitude, longitude e tempo, e o tempo de uso do carro deve ser derivado de alguma fonte externa.

d) Dimensão Especificidade

Existem alguns atributos que podem ser aplicados nos mais variados domínios, pois contêm informações gerais da trajetória. Ainda, existem também atributos que estão diretamente relacionados com o domínio em que estão inseridos. Como exemplo, tem-se o atributo Velocidade Sobre a Terra (SOG) que está diretamente vinculado ao domínio em que é inserido, e, em contrapartida, tem-se altitude, que pode ser usada em diferentes domínios. Levando em consideração esses aspectos, essa dimensão consiste na separação entre as características: genérico e específico de domínio.

Adicionando a nova dimensão à taxonomia, tem-se a taxonomia parcial 5, apresentada na Figura 4.16.

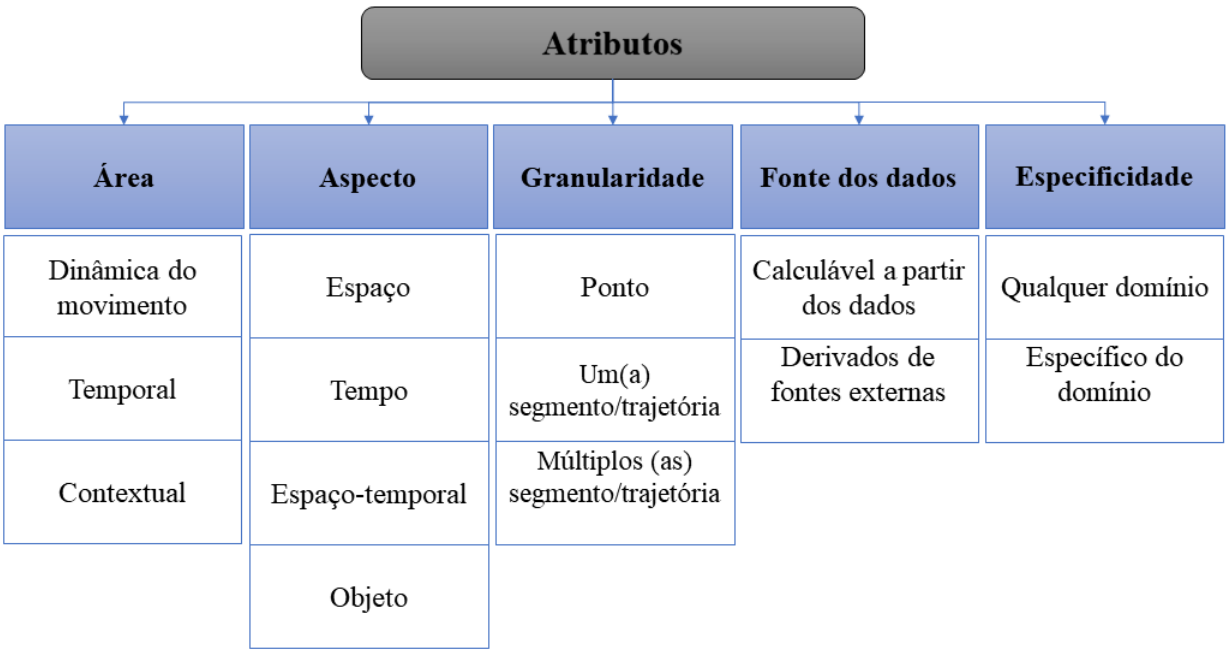

Figura 4.16: Taxonomia parcial 5

A relação entre a dimensão especificidade e suas características com os atributos pode ser representada com a utilização de um diagrama de classes. O diagrama de classes relativo à dimensão especificidade pode ser visto na Figura 4.17.

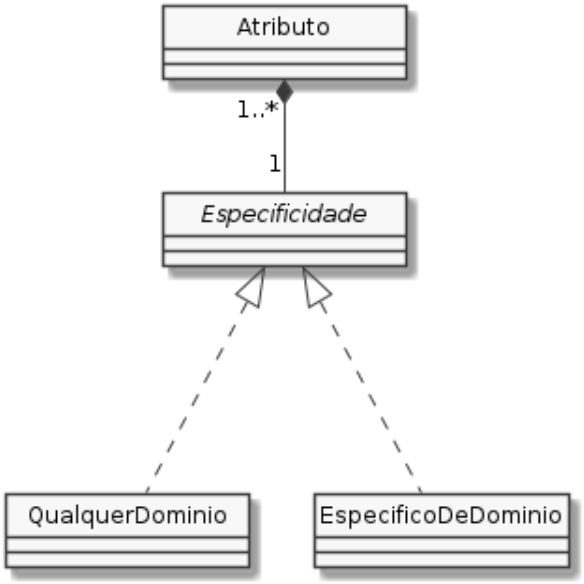

Figura 4.17: Diagrama de classes da dimensão especificidade

A comparação entres os atributos nesta dimensão, quando levamos em conta informações relevantes para o movimento de animais, em contrapartida, não auxiliam no processo de aprendizagem de máquina quando trabalhadas com relação às pessoas. É o caso do atributo distância até a fonte de água mais próxima para uma pessoa que está em um automóvel. Entretanto, a utilização do atributo velocidade aplica-se nos dois contextos.

e) Dimensão Parametrização

Por fim, o último ponto a ser destacado é se o atributo depende de uma parametrização. Esta divisão é considerada importante, pois a parametrização demanda experiência do pesquisador na área de trajetórias. Isso porque, a depender dos parâmetros, os atributos podem variar e levar a resultados diferentes. Portanto, a última dimensão separa os atributos em dependentes ou independente de parâmetros.

Adicionando a nova dimensão a taxonomia, tem-se a taxonomia final, apresentada na Figura 4.18. Com a adição da dimensão de parametrização, a taxonomia atingiu todas as

condições de parada. Logo, a mesma é entendida como satisfatória e definida enquanto a taxonomia final deste trabalho.

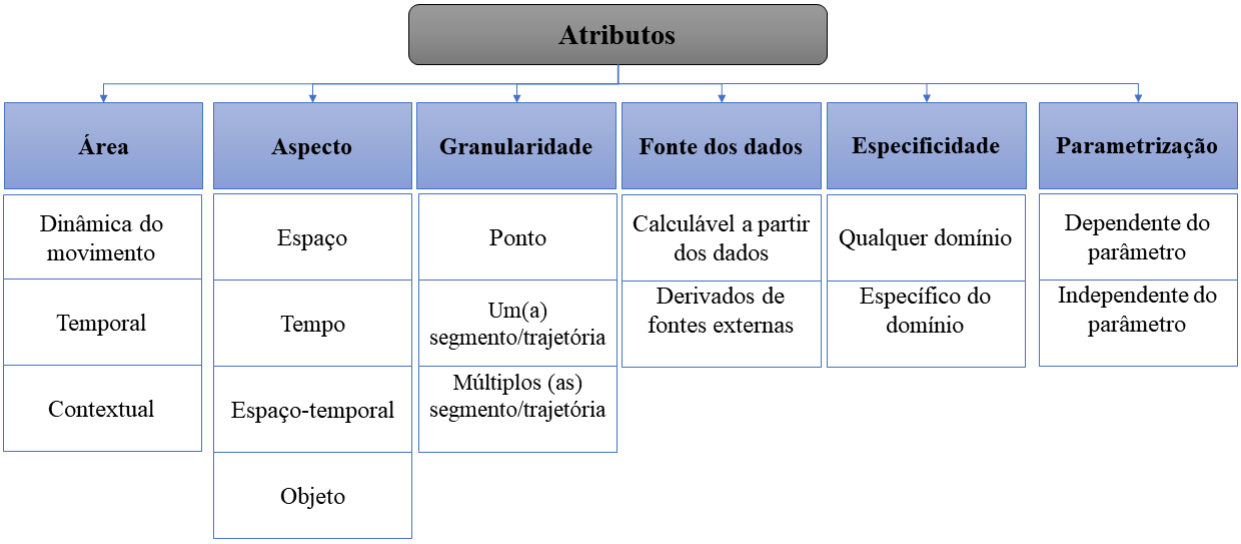

Figura 4.18: Taxonomia Trajtax

Assim como nas dimensões de especificidades e fontes de dados, a dimensão parametrização tem duas características possíveis. O diagrama de classes que relaciona o atributo e a parametrização pode ser observado na Figura 4.19

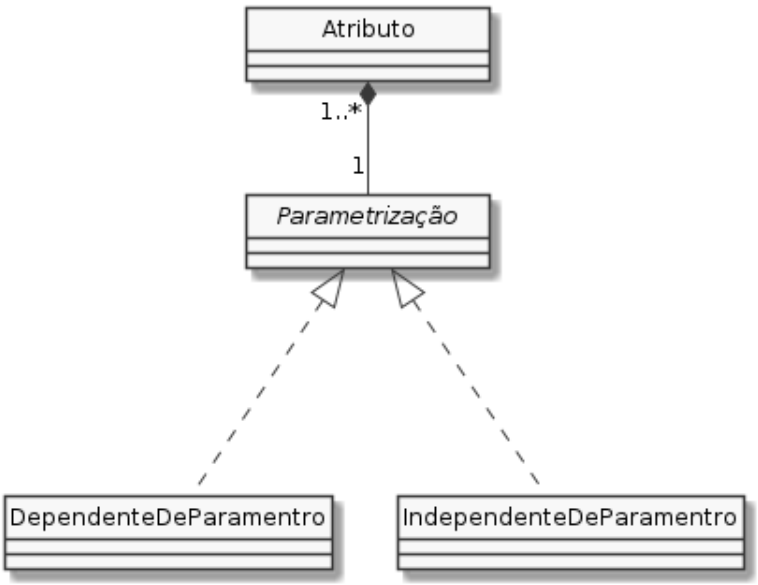

Figura 4.19: Diagrama de classes da dimensão parametrização

Para exemplificar este caso, é possível considerar um objeto que tem como atributo estar parado ou em movimento, e esse estado pode variar de acordo com o parâmetro usado como limiar. Em contrapartida, informações como aceleração não dependem de nenhum parâmetro e têm sempre o valor definido independente do contexto.

4.6 Avaliação da taxonomia resultante

Primeiramente, um passo importante é avaliar se a taxonomia está de acordo com a sua definição. Neste sentido, a taxonomia proposta está de acordo com a definição de escolhida neste trabalho, pois contém seis dimensões e cada uma delas tem pelo menos duas características. Outro ponto a ser avaliado é se as características são mutuamente excludentes e coletivamente exaustivas. Uma forma de avaliação desses critérios é a classificação de atributos nas dimensões propostas. Assim, as características foram avaliadas classificando atributos durante o processo de geração da taxonomia, para garantir que esses critérios fossem satisfeitos. A Tabela 4.5, Tabela 4.6 e Tabela 4.7 contêm a classificação dos atributos levantados nesta pesquisa de acordo com as dimensões da taxonomia Trajtax. É possível verificar que cada uma das características tenha pelo menos um objeto classificado. Para facilitar a visualização, os atributos foram agrupados por granularidade, Múltiplas(os) trajetórias/segmentos, Ponto e Trajetória/segmento completa(o), para tornar mas viável a busca pelos mesmos, de acordo com o escopo do projeto.

Tabela 4.5: Atributos com granularidade "Múltiplas(os) trajetórias/segmentos"

	Área	Aspecto	Fonte dos dados	Especificidade	Parametrização
Área	Dinâmica do movimento	Espaço	Calculável	Qualquer domínio	Independente de parâmetro
Área diária	Dinâmica do movimento	Espaço-temporal	Calculável	Qualquer domínio	Independente de parâmetro
Dias consecutivos	Temporal	Tempo	Calculável	Qualquer domínio	Independente de parâmetro
Distância diária	Dinâmica do movimento	Espaço-temporal	Calculável	Qualquer domínio	Independente de parâmetro
Distância entre os centroides diários	Dinâmica do movimento	Espaço-temporal	Calculável	Qualquer domínio	Independente de parâmetro
Distância máxima	Dinâmica do movimento	Espaço	Calculável	Qualquer domínio	Independente de parâmetro
Distância total	Dinâmica do movimento	Espaço	Calculável	Qualquer domínio	Independente de parâmetro
Duração da parada	Dinâmica do movimento	Espaço-temporal	Calculável	Qualquer domínio	Dependente de parâmetro
Localização típica do objeto	Contextual	Espaço-temporal	Calculável	Qualquer domínio	Independente de parâmetro
Número de movimentos	Dinâmica do movimento	Espaço-temporal	Calculável	Qualquer domínio	Dependente de parâmetro
Número de paradas	Dinâmica do movimento	Espaço-temporal	Calculável	Qualquer domínio	Dependente de parâmetro
Periodo de geração de trajetórias	Contextual	Tempo	Calculável	Qualquer domínio	Independente de parâmetro
Quilômetros percorridos por mês	Contextual	Espaço	Calculável	Qualquer domínio	Dependente de parâmetro
Sobreposição	Dinâmica do movimento	Espaço	Calculável	Qualquer domínio	Independente de parâmetro
Tamanho médio do segmento	Dinâmica do movimento	Espaço-temporal	Calculável	Qualquer domínio	Dependente de parâmetro
Taxa irregular de roteamento	Contextual	Espaço	Calculável	Qualquer domínio	Independente de parâmetro
Tempo de uso do veículo	Contextual	Objeto	Fontes externas	Específico de domínio	Independente de parâmetro

Tabela 4.6: Atributos com granularidade "Ponto"

	Área	Aspecto	Fonte dos dados	Especificidade	Parametrização
Aceleração	Dinâmica do movimento	Espaço-Temporal	Calculável	Qualquer domínio	Independente de parâmetro
Acelerômetro	Contextual	Espaço-temporal	Fontes externas	Qualquer domínio	Independente de parâmetro
Arranque	Dinâmica do movimento	Espaço-Temporal	Calculável	Qualquer domínio	Independente de parâmetro
Auto-interseção	Dinâmica do movimento	Espaço	Calculável	Qualquer domínio	Independente de parâmetro
Crescimento anormal da velocidade do motor	Contextual	Objeto	Fontes externas	Específico de domínio	Dependente de parâmetro
Curso sobre a Terra	Contextual	Espaço	Fontes externas	Específico de domínio	Independente de parâmetro
Dia da semana	Temporal	Tempo	Calculável	Qualquer domínio	Independente de parâmetro
Direção	Dinâmica do movimento	Espaço	Calculável	Qualquer domínio	Dependente de parâmetro
Distância entre pontos	Dinâmica do movimento	Espaço	Calculável	Qualquer domínio	Independente de parâmetro
Distância para atributos dependentes de contexto	Contextual	Espaço	Calculável	Qualquer domínio	Independente de parâmetro
Intervalo de tempo	Temporal	Tempo	Calculável	Qualquer domínio	Dependente de parâmetro
Largura da estrada	Contextual	Espaço	Fontes externas	Específico de domínio	Independente de parâmetro
Limite de velocidade da estrada	Contextual	Espaço	Fontes externas	Específico de domínio	Independente de parâmetro
Ponto de interesse	Contextual	Espaço	Fontes externas	Qualquer domínio	Independente de parâmetro
Porcentagem de inclinação	Contextual	Espaço	Fontes externas	Específico de domínio	Independente de parâmetro
Profundidade do solo	Contextual	Espaço	Fontes externas	Específico de domínio	Independente de parâmetro
Semana ou fim de semana/feriado	Temporal	Tempo	Calculável	Qualquer domínio	Independente de parâmetro
Taxa de giro	Dinâmica do movimento	Espaço	Calculável	Qualquer domínio	Independente de parâmetro
Taxa de mudança de direção	Dinâmica do movimento	Espaço	Calculável	Qualquer domínio	Dependente de parâmetro
Taxa de mudança na taxa de mudança de direção	Dinâmica do movimento	Espaço-Temporal	Calculável	Qualquer domínio	Dependente de parâmetro
Tipo de rodovia	Contextual	Espaço	Fontes externas	Específico de domínio	Independente de parâmetro
Velocidade	Dinâmica do movimento	Espaço-Temporal	Calculável	Qualquer domínio	Independente de parâmetro
Velocidade do motor	Contextual	Objeto	Fontes externas	Específico de domínio	Independente de parâmetro
Velocidade do vento	Contextual	Espaço-temporal	Fontes externas	Específico de domínio	Independente de parâmetro
Velocidade sobre a Terra	Contextual	Espaço-temporal	Fontes externas	Específico de domínio	Independente de parâmetro

Tabela 4.7: Atributos com granularidade "Trajetória/segmento completa(o)"

	Área	Aspecto	Fonte dos dados	Especificidade	Parametrização
Distância para uma sub-trajetória importante	Contextual	Espaço	Calculável	Qualquer domínio	Independente de parâmetro
Duração do movimento	Dinâmica do movimento	Tempo	Calculável	Qualquer domínio	Independente de parâmetro
Localização geografica dos pontos inicial e final	Dinâmica do movimento	Espaço	Calculável	Qualquer domínio	Independente de parâmetro
Ponto de permanência	Dinâmica do movimento	Espaço-Temporal	Calculável	Qualquer domínio	Dependente de parâmetro
Região de interesse	Contextual	Espaço	Fontes externas	Qualquer domínio	Independente de parâmetro
Tamanho da(o) trajetória/segmento	Dinâmica do movimento	Espaço	Calculável	Qualquer domínio	Independente de parâmetro
Taxa de mudança de velocidade	Dinâmica do movimento	Espaço-temporal	Calculável	Qualquer domínio	Dependente de parâmetro
Taxa de parada	Dinâmica do movimento	Espaço-temporal	Calculável	Qualquer domínio	Dependente de parâmetro
Tempo de parada de ônibus	Contextual	Espaço-temporal	Calculável	Específico de domínio	Dependente de parâmetro
Tempo no início do movimento	Temporal	Tempo	Calculável	Qualquer domínio	Independente de parâmetro
Valor comparado a um limite	Contextual	Objeto	Fontes externas	Qualquer domínio	Dependente de parâmetro
Velocidade angular do veículo	Contextual	Espaço-temporal	Fontes externas	Específico de domínio	Independente de parâmetro

Observa-se pelas tabelas 4.5, 4.6 e 4.7 que os atributos abrangem todas as categorias presentes na taxonomia. Outro ponto de destaque é que classificando e ordenando os atributos de acordo com suas características, é possível encontrar atributos com comportamento similar que podem ser trabalhados durante o processo de engenharia de atributos. Diferentes aplicações e combinações podem ser geradas com base na taxonomia de acordo com o propósito da pesquisa.

O segundo passo para avaliar a taxonomia é a verificação se ela atende às condições de parada. Neste sentido, o critério objetivo da condição de parada escolhido nesta pesquisa é que cada dimensão seja única e que cada característica seja também única dentro da sua dimensão. Esse critério foi optado para garantir que as dimensões não fossem baseadas em informações repetidas. A taxonomia gerada está de acordo com o critério. Dado que, todas as dimensões geradas são únicas, assim como as características são únicas para cada dimensão.

Uma vez que os critérios objetivos estão de acordo com o que foi proposto como critério de parada, é possível avaliar os critérios subjetivos. Os dois primeiros critérios subjetivos são que a taxonomia deve ser concisa, mas também deve ser robusta. Estes critérios são conflitantes e devem ser levados em consideração em paralelo para que haja um bom balanço entre eles. A taxonomia aqui proposta é sucinta de modo que as características presentes são facilmente visualizáveis, ainda assim, explica diferentes aspectos dos atributos. Em outras palavras, a taxonomia trabalhada não é simples demais, o que levaria a pouca explicação sobre os atributos, assim como não é complexa em demasia, o que levaria a contemplação de detalhes desnecessários.

O terceiro critério é que a taxonomia seja abrangente. Para tanto, é necessário avaliar a possibilidade de classificar todos os objetos, ou uma amostra aleatória, dentro das dimensões. Como visto nas Tabelas 4.5, 4.6 e 4.7, é possível classificar os objetos de acordo com as características propostas. Outro possível critério para que a taxonomia seja considerada abrangente é que ela cubra todas as dimensões de interesse relacionadas aos objetos. Sobre isso, as dimensões de interesse dos objetos foram levadas em conta na fase conceitual para empírica, onde foram levantados os conceitos importantes relacionados às trajetórias e aos atributos, assim a condição foi atendida.

O quarto critério é que a taxonomia seja extensível. Isso está ligado à dinâmica da taxonomia, se há a possibilidade de se adicionar novas dimensões e novas características. No

caso da taxonomia apresentada neste trabalho, o processo é simples, deve-se apenas escolher a abordagem desejada e continuar o processo de desenvolvimento utilizado neste trabalho.

E por fim, o quinto critério é relativo à explicabilidade da taxonomia. Como visto, os pontos levantados para a criação das características foram considerados importantes. Essa condição está diretamente associada às duas primeiras, pois é relativa ao nível de detalhes proveniente da taxonomia. O objetivo é que a taxonomia represente a natureza dos atributos, e não seus detalhes. A taxonomia resultante aqui discutida tem dimensões e características associadas ao contexto geral, servindo como base sobre a qual pode haver discussões sobre os distintos atributos presentes na literatura.

4.7 Considerações finais

Neste capítulo, foram discutidos os principais pontos associados aos atributos relacionados a trajetórias. O resultado do levantamento dos atributos da literatura foi descrito em detalhes e os passos seguidos no processo de desenvolvimento da taxonomia foram apresentados. A taxonomia resultante do processo foi avaliada quanto a sua utilidade em que satisfez os critérios usados como base para a avaliação. Por fim, a taxonomia Trajtax foi discutida levando em consideração suas dimensões e o que cada uma delas agrega ao conhecimento do atributo.

O propósito da taxonomia Trajtax é que ela possa ser útil para outros pesquisadores, destacando pontos importantes dos atributos vinculados à trajetórias. Os critérios de parada subjetivos aqui avaliados demonstram os potenciais associados ao uso da Trajtax em processos de aprendizagem de máquina. A partir disso, pode-se concluir que a metodologia utilizada foi aplicada com sucesso, gerando uma taxonomia robusta para ser usada no processo de engenharia de atributos com dados de trajetórias.

No próximo capítulo são apresentadas as conclusões deste trabalho e propostas de trabalhos futuros.

Capítulo 5

Conclusão

Neste trabalho, foi criada uma taxonomia relacionada aos atributos contidos em bases de dados de trajetórias, com o objetivo de explicar as especificidades associadas aos atributos deste domínio. Para o desenvolvimento desta taxonomia, foi realizada uma revisão da literatura focada nos atributos aos quais recorremos. Em especial, foram considerados os atributos de fácil interpretação, pois a criação de modelos interpretáveis ainda é um desafio com dados de trajetória. A taxonomia desenvolvida em nosso estudo foi nomeada TrajTax e tem por intuito que ela possa ser aplicada em trabalhos de aprendizagem de máquina no âmbito de trajetórias, visando melhorar seus resultados.

A revisão da literatura foi executada modelando os atributos que foram empregados em diferentes projetos com dados de trajetória. Projetos de distintos domínios foram levados em consideração e o produto gerado a partir da revisão da literatura foi um resumo detalhado dos atributos encontrados em literatura recente e relevante na área. Os atributos encontrados foram nomeados e definidos para facilitar a compreensão de como o atributo é calculado ou como ele é aferido pelos sensores. Assim, este documento pode ser utilizado, também, como uma ferramenta facilitadora para estudantes e pesquisadores que estão iniciando na área de trajetórias, e precisem procurar por uma base de referência sobre quais atributos podem ser usados para guiar suas pesquisas.

Para o desenvolvimento da taxonomia TrajTax, foram aplocadas três bases de dados e os atributos coletados da literatura. As bases de dados escolhidas para utilização nesta pesquisa advêem de diferentes domínios. Como resultado do processo, a taxonomia TrajTax contém seis dimensões denominadas: informação representada, área, escopo, fonte dos dados, espe-

cificidade e parametrização. A dimensão relacionada à informação contida no atributo tem como características: espaço, tempo, espaço-temporal e objeto. Já a dimensão relacionada à área, apresenta: dinâmica do movimento, temporal e conceitual. Com relação ao escopo, tem-se as características: ponto, um(a) trajetória/segmento, múltiplas(os) trajetórias/segmentos.

Também foi discutido que a dimensão relativa à fonte dos dados avalia se o dado é calculável a partir do que é básico ou dos derivados de fontes externas. A especificidade, por sua vez, está associada à utilidade ou não a um atributo em diferentes domínios. Por fim, a parametrização está relacionada à dependência ou não de parâmetro dos atributos.

O processo de desenvolvimento da taxonomia TrajTax foi realizado seguindo os dois caminhos propostos na literatura: avaliação empírica associando um contexto aos resultados obtidos na avaliação e avaliação contextual testada empiricamente. Na primeira, três bases de dados foram usadas com três distintos domínios. As bases de dados continham trajetórias relacionadas à mobilidade urbana com diferentes meios de transporte, trajetórias de navios que tinham suas rotas classificadas como rota de pesca ou não, e trajetórias de animais que podiam ser gado, veado ou alces. Essas bases de dados são anotadas, possibilitando a aplicação de atividades de aprendizagem de máquina supervisionada.

As bases de dados utilizadas foram incrementadas com atributos gerados pela biblioteca *Trajlib*. A biblioteca foi expandida como parte desta pesquisa para a adição de atributos temporais assim como a implementação de um algoritmo de segmentação. Os dados adicionados às bases de dados são genéricos e podem ser calculados a partir dos dados básicos de trajetórias.

Na avaliação conceitual dos atributos, foram levadas em consideração informações relacionadas ao contexto de trajetória, assim como características intrínsecas dos atributos que merecem destaque. Cada uma das características foi detalhada para que possa servir de base para próximas pesquisas e para a expansão da taxonomia resultante.

A taxonomia gerada satisfaz os critérios de avaliação de taxonomias proposto na literatura e tem potencial para ser aplicada tanto para discussão sobre o assunto quanto sendo guia para novas pesquisas. Desta forma, a TrajTax pode ser considerada uma taxonomia completa e extensível, dado que os procedimentos para sua criação foram padronizados e realizados com documentação detalhada. Os objetivos do trabalho foram, portanto, atingidos,

vislumbrando atualizações em trabalhos futuros considerando a utilização da taxonomia em trabalhos na área de aprendizagem de máquina.

5.1 Contribuições

Este trabalho tem como principais contribuições:

- O levantamento dos atributos utilizados em trabalhos na área de trajetória em diferentes domínios e a definição deles.
- A proposição de uma taxonomia para os atributos usados em estudos de aprendizagem de máquina no contexto de trajetórias.

5.2 Trabalhos futuros

Para continuação desta pesquisa, os seguintes tópicos são sugeridos:

- Avaliar a utilização da taxonomia por outros pesquisadores e propor melhorias de acordo com a utilização.
- Expandir a taxonomia para consideração de atributos que tem uma complexidade alta, onde podemos citar, por exemplo, os valores associados aos neurônios da última camada escondida de uma rede neural convolucional treinada com os dados das trajetórias, que não são levados em consideração neste trabalho pois atrapalham na interpretabilidade de modelos de aprendizagem de máquina.
- Aplicar a fase empírica em bases de dados de outros domínios, avaliando a possibilidade de existência de outros grupos.

Referências Bibliográficas

[1] Sultan Alamri, David Taniar, and Maytham Safar. A taxonomy for moving object queries in spatial databases. *Future Generation Computer Systems*, 37:232–242, 2014.

[2] Kevin Allain, Cagatay Turkay, and Jason Dykes. Towards a what-why-how taxonomy of trajectories in visualization research. 2019.

[3] Ethem Alpaydin. *Introduction to machine learning*. 2014.

[4] Mateus Barragana, Luis Otavio Alvares, and Vania Bogorny. Unusual behavior detection and object ranking from movement trajectories in target regions. *International Journal of Geographical Information Science*, 31(2):364–386, 2017.

[5] Girish Chandrashekar and Ferat Sahin. A survey on feature selection methods. *Computers & Electrical Engineering*, 40(1):16–28, 2014.

[6] Konstantinos Chatzikokolakis, Dimitrios Zissis, Giannis Spiliopoulos, and Konstantinos Tserpes. A comparison of supervised learning schemes for the detection of search and rescue (sar) vessel patterns. *GeoInformatica*, pages 1–22, 2019.

[7] Tianyi Chen, Xiupeng Shi, and Yiik Diew Wong. Key feature selection and risk prediction for lane-changing behaviors based on vehicles' trajectory data. *Accident Analysis & Prevention*, 129:156–169, 2019.

[8] Zaiben Chen, Heng Tao Shen, Xiaofang Zhou, Yu Zheng, and Xing Xie. Searching trajectories by locations: an efficiency study. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*, pages 255–266. ACM, 2010.

[9] Maximilian Christ, Nils Braun, Julius Neuffer, and Andreas W Kempa-Liehr. Time series feature extraction on basis of scalable hypothesis tests (tsfresh–a python package). *Neurocomputing*, 307:72–77, 2018.

[10] Sina Dabiri and Kevin Heaslip. Inferring transportation modes from gps trajectories using a convolutional neural network. *Transportation research part C: emerging technologies*, 86:360–371, 2018.

[11] Sina Dabiri, Chang-Tien Lu, Kevin Heaslip, and Chandan K Reddy. Semi-supervised deep learning approach for transportation mode identification using gps trajectory data. *IEEE Transactions on Knowledge and Data Engineering*, 2019.

[12] Jian Dai, Bin Yang, Chenjuan Guo, and Zhiming Ding. Personalized route recommendation using big trajectory data. In *2015 IEEE 31st international conference on data engineering*, pages 543–554. IEEE, 2015.

[13] Erico N de Souza, Kristina Boerder, Stan Matwin, and Boris Worm. Improving fishing pattern detection from satellite ais using data mining and machine learning. *PloS one*, 11(7):e0158248, 2016.

[14] Somayeh Dodge, Robert Weibel, and Ehsan Forootan. Revealing the physics of movement: Comparing the similarity of movement characteristics of different types of moving objects. *Computers, Environment and Urban Systems*, 33(6):419–434, 2009.

[15] Pedro M Domingos. A few useful things to know about machine learning. *Commun. acm*, 55(10):78–87, 2012.

[16] Yuki Endo, Hiroyuki Toda, Kyosuke Nishida, and Akihisa Kawanobe. Deep feature extraction from trajectories for transportation mode estimation. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 54–66. Springer, 2016.

[17] Mohammad Etemad, Amílcar Soares Júnior, and Stan Matwin. Predicting transportation modes of gps trajectories using feature engineering and noise removal. In *Advances in Artificial Intelligence: 31st Canadian Conference on Artificial Intelligence, Canadian AI 2018, Toronto, ON, Canada, May 8–11, 2018, Proceedings 31*, pages 259–264. Springer, 2018.

[18] Z. Feng and Y. Zhu. A survey on trajectory data mining: Techniques and applications. *IEEE Access*, 4:2056–2067, 2016.

[19] Carlos Andres Ferrero, Luis Otavio Alvares, Willian Zalewski, and Vania Bogorny. Movelets: exploring relevant subtrajectories for robust trajectory classification. In *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*, pages 849–856. ACM, 2018.

[20] Hansjörg Fromm, Thiemo Wambsganss, and Matthias Söllner. Towards a taxonomy of text mining features. 2019.

[21] Andre Salvaro Furtado, Laercio Lima Pilla, and Vania Bogorny. A branch and bound strategy for fast trajectory similarity measuring. *Data & Knowledge Engineering*, 115:16–31, 2018.

[22] Andre Salvaro Furtado, Luis Otavio Campos Alvares, Nikos Pelekis, Yannis Theodoridis, and Vania Bogorny. Unveiling movement uncertainty for robust trajectory similarity analysis. *International Journal of Geographical Information Science*, 32(1):140–168, 2018.

[23] Robert L Glass and Iris Vessey. Contemporary application-domain taxonomies. *IEEE Software*, 12(4):63–76, 1995.

[24] Bing He, Xiaolin Chen, Dian Zhang, Siyuan Liu, Dawei Han, and Lionel M. Ni. Pbe: Driver behavior assessment beyond trajectory profiling. In Ulf Brefeld, Edward Curry, Elizabeth Daly, Brian MacNamee, Alice Marascu, Fabio Pinelli, Michele Berlingerio, and Neil Hurley, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 507–523, Cham, 2019. Springer International Publishing.

[25] Zihan Hong, Ying Chen, and Hani S Mahmassani. Recognizing network trip patterns using a spatio-temporal vehicle trajectory clustering algorithm. *IEEE Transactions on Intelligent Transportation Systems*, 19(8):2548–2557, 2017.

[26] Zihan Hong, Ying Chen, and Hani S Mahmassani. Recognizing network trip patterns using a spatio-temporal vehicle trajectory clustering algorithm. *IEEE Transactions on Intelligent Transportation Systems*, 19(8):2548–2557, 2018.

[27] Amin Hosseinpoor Milaghardan, Rahim Ali Abbaspour, and Christophe Claramunt. A geometric framework for detection of critical points in a trajectory using convex hulls. *ISPRS International Journal of Geo-Information*, 7(1):14, 2018.

[28] Amin Hosseinpoor Milaghardan, Rahim Ali Abbaspour, and Christophe Claramunt. A spatio-temporal entropy-based framework for the detection of trajectories similarity. *Entropy*, 20(7):490, 2018.

[29] Amílcar Soares Júnior, Chiara Renso, and Stan Matwin. Analytic: An active learning system for trajectory classification. *IEEE computer graphics and applications*, 37(5):28–39, 2017.

[30] Amilcar Soares Junior, Valeria Cesario Times, Chiara Renso, Stan Matwin, and Lucidio AF Cabral. A semi-supervised approach for the semantic segmentation of trajectories. In *2018 19th IEEE International Conference on Mobile Data Management (MDM)*, pages 145–154. IEEE, 2018.

[31] Xiangjie Kong, Ximeng Song, Feng Xia, Haochen Guo, Jinzhong Wang, and Amr Tolba. Lotad: Long-term traffic anomaly detection based on crowdsourced bus trajectory data. *World Wide Web*, 21(3):825–847, 2018.

[32] Jae-Gil Lee, Jiawei Han, and Kyu-Young Whang. Trajectory clustering: a partition-and-group framework. In *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*, pages 593–604. ACM, 2007.

[33] Jae-Gil Lee, Jiawei Han, Xiaolei Li, and Hector Gonzalez. Traclass: trajectory classification using hierarchical region-based and trajectory-based clustering. *Proceedings of the VLDB Endowment*, 1(1):1081–1094, 2008.

[34] Huanhuan Li, Jingxian Liu, Kefeng Wu, Zaili Yang, Ryan Wen Liu, and Naixue Xiong. Spatio-temporal vessel trajectory clustering based on data mapping and density. *IEEE Access*, 6:58939–58954, 2018.

[35] H. Liu and I. Lee. End-to-end trajectory transportation mode classification using bi-lstm recurrent neural network. In *2017 12th International Conference on Intelligent Systems and Knowledge Engineering (ISKE)*, pages 1–5, Nov 2017.

[36] Jean Damascène Mazimpaka and Sabine Timpf. Trajectory data mining: A review of methods and applications. *Journal of Spatial Information Science*, 2016(13):61–99, 2016.

[37] Fabio Mazzarella, Virginia Fernandez Arguedas, and Michele Vespe. Knowledge-based vessel position prediction using historical ais data. In *2015 Sensor Data Fusion: Trends, Solutions, Applications (SDF)*, pages 1–6. IEEE, 2015.

[38] Jeffrey G Miller and Aleda V Roth. A taxonomy of manufacturing strategies. *Management Science*, 40(3):285–304, 1994.

[39] David S. Moore, William I. Notz, and Michael A. Fligner. *The Basic practice of statistics*. W.H. Freeman, 2018.

[40] Mirco Nanni and Dino Pedreschi. Time-focused clustering of trajectories of moving objects. *Journal of Intelligent Information Systems*, 27(3):267–289, 2006.

[41] Francisco Dantas Nobre Neto, Cláudio de Souza Baptista, and Claudio EC Campelo. A user-personalized model for real time destination and route prediction. In *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, pages 401–407. IEEE, 2016.

[42] Duong Nguyen, Rodolphe Vadaine, Guillaume Hajduch, René Garello, and Ronan Fablet. A multi-task deep learning architecture for maritime surveillance using ais data streams. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 331–340. IEEE, 2018.

[43] Robert C Nickerson, Upkar Varshney, and Jan Muntermann. A method for taxonomy development and its application in information systems. *European Journal of Information Systems*, 22(3):336–359, 2013.

[44] Mai Oudah and Andreas Henschel. Taxonomy-aware feature engineering for microbiome classification. *BMC bioinformatics*, 19(1):1–13, 2018.

[45] Sinan Ozdemir and Divya Susarla. *Feature Engineering Made Easy: Identify unique features from your dataset in order to build powerful machine learning systems*. Packt Publishing Ltd, 2018.

[46] Johannes Paefgen, Florian Michahelles, and Thorsten Staake. Gps trajectory feature extraction for driver risk profiling. In *Proceedings of the 2011 International Workshop on Trajectory Data Mining and Analysis*, TDMA '11, pages 53–56, New York, NY, USA, 2011. ACM.

[47] Johannes Paefgen, Florian Michahelles, and Thorsten Staake. Gps trajectory feature extraction for driver risk profiling. In *Proceedings of the 2011 International Workshop on Trajectory Data Mining and Analysis*, TDMA '11, pages 53–56, New York, NY, USA, 2011. ACM.

[48] Giuliana Pallotta, Steven Horn, Paolo Braca, and Karna Bryan. Context-enhanced vessel prediction based on ornstein-uhlenbeck processes using historical ais traffic patterns: Real-world experimental results. In *17th international conference on information fusion (FUSION)*, pages 1–7. IEEE, 2014.

[49] Andrey Tietbohl Palma, Vania Bogorny, Bart Kuijpers, and Luis Otavio Alvares. A clustering-based approach for discovering interesting places in trajectories. In *Proceedings of the 2008 ACM symposium on Applied computing*, pages 863–868, 2008.

[50] Luca Pappalardo and Filippo Simini. Data-driven generation of spatio-temporal routines in human mobility. *Data Mining and Knowledge Discovery*, 32(3):787–829, 2018.

[51] Dhaval Patel, Chang Sheng, Wynne Hsu, and Mong Li Lee. Incorporating duration information for trajectory classification. In *2012 IEEE 28th International Conference on Data Engineering*, pages 1132–1143. IEEE, 2012.

[52] Sasank Reddy, Min Mun, Jeff Burke, Deborah Estrin, Mark Hansen, and Mani Srivastava. Using mobile phones to determine transportation modes. *ACM Transactions on Sensor Networks (TOSN)*, 6(2):13, 2010.

[53] Chiara Renso, Stefano Spaccapietra, and Esteban Zimányi, editors. *Mobility Data: Modeling, Management, and Understanding*. Cambridge University Press, 2013.

[54] Mary M Rowland, Priscilla K Coe, Rosemary J Stussy, et al. The starkey habitat database for ungulate research: construction, documentation, and use. *Gen. Tech. Rep.*

PNW-GTR-430. Portland, OR: US Department of Agriculture, Forest Service, Pacific Northwest Research Station. 48 p, 430, 1998.

[55] BA Sabarish, R Karthi, and Gireesh Kumar. String-based feature representation for trajectory clustering. *International Journal of Embedded and Real-Time Communication Systems (IJERTCS)*, 10(2):1–18, 2019.

[56] Rajiv Sabherwal and William R King. An empirical taxonomy of the decision-making processes concerning strategic applications of information systems. *Journal of Management Information Systems*, 11(4):177–214, 1995.

[57] Luca Scherrer, Martin Tomko, Peter Ranacher, and Robert Weibel. Travelers or locals? identifying meaningful sub-populations from human movement data in the absence of ground truth. *EPJ Data Science*, 7(1):19, 2018.

[58] Lokesh K Sharma, Om Prakash Vyas, Simon Schieder, and Ajaya K Akasapu. Nearest neighbour classification for trajectory data. In *International Conference on Advances in Information and Communication Technologies*, pages 180–185. Springer, 2010.

[59] George Gaylord Simpson. Principles of animal taxonomy. 1961.

[60] Robert R Sokal and P. H. A Sneath. *Principles of numerical taxonomy*. W.H. Freeman, San Francisco, 1963.

[61] Robert R Sokal. Numerical taxonomy. *Scientific American*, 215(6):106–117, 1966.

[62] Han Su, Kai Zheng, Kai Zeng, Jiamin Huang, and Xiaofang Zhou. Stmaker: a system to make sense of trajectory data. *Proceedings of the VLDB Endowment*, 7(13):1701–1704, 2014.

[63] Thomas Douglas Victor Swinscow, Michael J Campbell, et al. *Statistics at square one*. Bmj London, 2002.

[64] Fred E. Szabo. C. In Fred E. Szabo, editor, *The Linear Algebra Survival Guide*, pages 47 – 77. Academic Press, Boston, 2015.

[65] George E Uhlenbeck and Leonard S Ornstein. On the theory of the brownian motion. *Physical review*, 36(5):823, 1930.

[66] Jake VanderPlas. *Python data science handbook: essential tools for working with data*. "O'Reilly Media, Inc.", 2016.

[67] Iraklis Varlamis, Konstantinos Tserpes, Mohammad Etemad, Amílcar Soares Júnior, and Stan Matwin. A network abstraction of multi-vessel trajectory data for detecting anomalies. 2019.

[68] Iraklis Varlamis, Konstantinos Tserpes, Mohammad Etemad, Amílcar Soares Júnior, and Stan Matwin. A network abstraction of multi-vessel trajectory data for detecting anomalies. In *EDBT/ICDT Workshops*, 2019.

[69] Yulong Wang, Kun Qin, Yixiang Chen, and Pengxiang Zhao. Detecting anomalous trajectories and behavior patterns using hierarchical clustering from taxi gps data. *ISPRS International Journal of Geo-Information*, 7(1):25, 2018.

[70] Zhibin Xiao, Yang Wang, Kun Fu, and Fan Wu. Identifying different transportation modes from trajectory data using tree-based ensemble classifiers. *ISPRS International Journal of Geo-Information*, 6(2):57, 2017.

[71] Jin Xu, Hai-Bo Shu, and Yi-Ming Shao. Modeling of driver behavior on trajectory-speed decision making in minor traffic roadways with complex features. *IEEE Transactions on Intelligent Transportation Systems*, (99):1–13, 2018.

[72] Di Yao, Chao Zhang, Zhihua Zhu, Qin Hu, Zheng Wang, Jianhui Huang, and Jingping Bi. Learning deep representation for trajectory clustering. *Expert Systems*, 35(2):e12252, 2018.

[73] Alice Zheng and Amanda Casari. *Feature engineering for machine learning: principles and techniques for data scientists*. "O'Reilly Media, Inc.", 2018.

[74] Yu Zheng, Like Liu, Longhao Wang, and Xing Xie. Learning transportation mode from raw gps data for geographic applications on the web. In *Proceedings of the 17th international conference on World Wide Web*, pages 247–256. ACM, 2008.

[75] Yu Zheng, Quannan Li, Yukun Chen, Xing Xie, and Wei-Ying Ma. Understanding mobility based on gps data. In *Proceedings of the 10th international conference on Ubiquitous computing*, pages 312–321. ACM, 2008.

[76] Yu Zheng, Lizhu Zhang, Xing Xie, and Wei-Ying Ma. Mining interesting locations and travel sequences from gps trajectories. In *Proceedings of the 18th international conference on World wide web*, pages 791–800. ACM, 2009.

[77] Yu Zheng, Xing Xie, Wei-Ying Ma, et al. Geolife: A collaborative social networking service among user, location and trajectory. 2010.

[78] Yu Zheng, Yukun Chen, Quannan Li, Xing Xie, and Wei-Ying Ma. Understanding transportation modes based on gps data for web applications. *ACM Transactions on the Web (TWEB)*, 4(1):1, 2010.

[79] Yu Zheng. Trajectory data mining: an overview. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 6(3):29, 2015.

[80] Jia Zhu, Changqin Huang, Min Yang, and Gabriel Pui Cheong Fung. Context-based prediction for road traffic state using trajectory pattern mining and recurrent convolutional neural networks. *Information Sciences*, 473:190–201, 2019.

[81] Xiaojin Zhu and Andrew B Goldberg. Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning*, 3(1):1–130, 2009.