UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO TECNOLÓGICO
DEPARTAMENTO DE INFORMÁTICA E ESTATÍSTICA
CIÊNCIAS DA COMPUTAÇÃO

Ricardo Giuliani

CLUSTERIZAÇÃO DE TRAJETÓRIAS MULTIASPECTO USANDO ÁRVORES DE DECISÃO

Florianópolis

2022

Ricardo Giuliani

CLUSTERIZAÇÃO DE TRAJETÓRIAS MULTIASPECTO USANDO ÁRVORES DE DECISÃO

Trabalho de Conclusão do Curso de Graduação em Ciências da Computação do Centro de Tecnologia da Universidade Federal de Santa Catarina como requisito para a obtenção do título de Bacharel em Ciências da Computação.
Orientador: Prof. Jônata Tyska Carvalho, Dr.
Coorientadora: Profa. Vania Bogorny, Dra.

Florianópolis
2022

Ficha de identificação da obra elaborada pelo autor,
através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

Giuliani, Ricardo
Clusterização de trajetórias multiaspecto usando árvores de decisão / Ricardo Giuliani ; orientador, Jônata Tyska Carvalho, coorientador, Vania Bogorny, 2022.
79 p.

Trabalho de Conclusão de Curso (graduação) - Universidade Federal de Santa Catarina, Centro Tecnológico, Graduação em Ciências da Computação, Florianópolis, 2022.

Inclui referências.

1. Ciências da Computação. 2. Trajetórias multiaspecto. 3. Trajetórias. 4. Árvores de decisão. 5. Clusterização. I. Tyska Carvalho, Jônata. II. Bogorny, Vania. III. Universidade Federal de Santa Catarina. Graduação em Ciências da Computação. IV. Título.

Ricardo Giuliani

CLUSTERIZAÇÃO DE TRAJETÓRIAS MULTIASPECTO USANDO ÁRVORES DE DECISÃO

Trabalho de Conclusão de Curso apresentado como requisito para a obtenção do Título de Bacharel em Ciências da Computação do Centro de Tecnologia da Universidade Federal de Santa Catarina e aprovado em sua forma final pela banca examinadora.

Florianópolis, 24 de março de 2022.

_____

Prof. Dr. Renato Cislaghi
Universidade Federal de Santa Catarina
Coordenador

Banca Examinadora:

_____

Prof. Dr. Jônata Tyska Carvalho
Universidade Federal de Santa Catarina
Orientador

_____

Profa. Dra. Vania Bogorny
Universidade Federal de Santa Catarina
Coorientadora

_____

Prof. Dr. Ronaldo dos Santos Mello
Universidade Federal de Santa Catarina

Dedico este trabalho à minha família e a todos que contribuíram para a conclusão desta jornada.

AGRADECIMENTOS

Agradeço, em primeiro lugar, aos meus pais Marli e Giuseppe, por todo o amor e apoio incondicional em todas as minhas decisões.

Aos meus irmãos Roberto e Rafaella, por toda parceria e ajuda em todos os momentos.

Aos professores Dr. Jônata Tyska Carvalho e Dra. Vania Bogorny pela paciência e dedicação dispensados à elaboração deste trabalho, bem como pelos ensinamentos e conhecimentos transmitidos.

Aos meus colegas e amigos de curso que estiveram comigo durante esta jornada, em especial: Maria Eduarda, Arthur (Evaristo), Alan, Bryan, Natália, Isac, Gustavo, Northon, Gabriel, Arthur (Capa), Daniela e Eduardo.

E por fim, agradeço aos professores que contribuíram com a minha formação profissional.

“Não somos apenas o que pensamos ser. Somos mais: somos também o que lembramos e aquilo de que nos esquecemos; somos as palavras que trocamos, os enganos que cometemos, os impulsos que cedemos, ‘sem querer’”.
(*Sigmund Freud*)

RESUMO

O crescente aumento e popularização de dispositivos móveis com tecnologia GPS promoveu um grande volume de dados ao permitir a captura da movimentação dos usuários ou objetos em sequências de pontos no espaço e no tempo definindo o conceito de trajetórias do objeto móvel. Estes dados espaço-temporais estão presentes em diversos problemas cotidianos, como questões sobre aquecimento global, mobilidade urbana, desastres naturais, migrações humanas e animais, e despertam um grande interesse da comunidade científica. O advento da Internet das Coisas, objetos com sensores embarcados que coletam inúmeras informações como temperatura, poluição atmosférica e batimentos cardíacos, diversas características puderam ser agregadas às trajetórias, enriquecendo-as semanticamente e gerando uma nova abordagem denominada de trajetórias multiaspecto. À medida que se adicionam mais aspectos às trajetórias, mais completa é a representação real do movimento dos objetos, e mais útil e interessante são as informações que se podem inferir sobre objetos e lugares. Os dados dessas trajetórias multiaspecto podem conter informações importantes permitindo identificar, por exemplo, comportamentos e padrões recorrentes dos objetos, como também realizar previsões sobre tendências futuras. Uma técnica interessante para analisar as trajetórias é o agrupamento, que objetiva encontrar similaridades entre trajetórias ou objetos móveis gerando entendimento sobre essas semelhanças dado o cenário de trajetórias de diferentes objetos, ou de um mesmo objeto em momentos diferentes. O agrupamento de dados apresenta diversas aplicações, como exemplo, segmentação de mercado através do perfilamento de clientes, detecção de comportamentos anômalos em conjunto de dados de trajetórias, identificação de pontos importantes baseados em trajetórias de turistas em uma determinada cidade, identificação de atividades criminais ou fraudulentas, dentre outras. O interesse por análises de trajetórias é grande, no entanto, a maioria dos trabalhos dão pouca atenção à semântica. A alta dimensionalidade e heterogeneidade de trajetórias multiaspecto representa um grande desafio quanto à forma de tratar os dados das trajetórias, integrar toda informação em uma única representação de trajetória e extrair informações valiosas. Diante disso, este trabalho tem como objetivo utilizar árvores de decisão e aprendizado não-supervisionado para identificar agrupamentos naturais, que possuam um significado, em conjuntos de dados de trajetórias multiaspecto. Os grupos encontrados são avaliados com métricas de validação internas que permitem mensurar a coesão, a separabilidade e a qualidade dos agrupamentos obtidos e métricas de validação externas que utilizam referências externas para comparação e identificação da melhor forma de agrupar as trajetórias e assim determinar a utilidade e a validade dos resultados obtidos.

Palavras-chave: Trajetórias. Trajetórias multiaspecto. Agrupamento. Árvores de decisão.

ABSTRACT

The increase and popularization of mobile devices with GPS technology promoted a large volume of data by allowing the capture of movement of users or objects in sequences of points in space and time, defining the concept of trajectories of moving objects. These spatiotemporal data are present in several daily problems, such as questions about global warming, urban mobility, natural disasters, human and animal migrations, and arouse great interest from the scientific community. The advent of the Internet of Things, objects with built-in sensors that collect numerous information such as temperature, air pollution and heartbeat, several characteristics could be added to the trajectories, enriching them semantically and generating a new approach called trajectories with multiple aspect. As more aspects are added to the trajectories, the more complete is the real representation of the movement of the objects, and the more useful and interesting is the information that can be inferred about objects and places. The data of these multi-aspect trajectories can contain important information allowing to identify, for example, recurring behaviors and patterns of objects, as well as to make predictions about future trends. An interesting technique for analyzing trajectories is clustering, which aims to find similarities between trajectories or moving objects, generating understanding of these similarities given the scenario of trajectories of different objects, or of the same object at different times. The clusterization of data presents several applications, such as, market segmentation through the profiling of customers, detection of anomalous behaviors in a set of trajectory data, identification of important points based on the trajectories of tourists in a given city, identification of criminal or fraudulent activities etc. The interest in trajectory analysis is great, however, most works pay little attention to semantics. The high dimensionality and heterogeneity of multi-aspect trajectories represent a major challenge as to how to handle the trajectory data, integrate all information in a single trajectory representation and extract valuable information. Therefore, the research aims to use decision trees and unsupervised learning to identify natural clusters, which have a meaning, in data sets of multi-aspect trajectories. The groups found are evaluated with internal validation metrics that allow measuring the cohesion, separability and quality of the obtained groupings and external validation metrics that use external references to compare and identify the best way to group trajectories and thus determine the usefulness and the validity of the results obtained.

Keywords: Trajectories. Multiple aspect trajectories. Clustering. Decision trees.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

CS	Coeficiente Silhouette
CRISP-DM	Cross Industry Standard Process for Data Mining
DB	Divisão Binária
EDR	Edit Distance on Real Sequence
GPS	Global Positioning System
ICH	Índice Calinski-Harabasz
IDB	Índice Davies-Bouldin
IOT	Internet of Things
LCSS	Longest Common Subsequence
MRV	Máxima de Redução de Variância
MSM	Multidimensional Similarity Measure
MUITAS	Multiple-Aspect Trajectory Similarity Measure
MV	Mínima Variância
POI	Point of Interest
RMV	Redução Máxima de Variância
SMSM	Stops and Moves Similarity Measure
TMA	Trajetória Multiaspecto
TraFoS	Trajectory Forest Similarity
UMS	Uncertain Movement Similarity

SUMÁRIO

1 INTRODUÇÃO

O avanço tecnológico e a popularização de dispositivos móveis com tecnologia GPS (sistema de posicionamento global, do inglês *global positioning system*) nas últimas décadas permitiu o desenvolvimento de bancos de dados com registros de mobilidade dos usuários. Conforme o dispositivo é configurado, a localização do indivíduo em um determinado instante de tempo pode ser gravada no que se denomina de ponto. Uma sequência de pontos que descrevem o movimento de uma pessoa, um animal, um meio de transporte ou de objetos em geral define o conceito de trajetória (BOGORNY; BRAZ, 2012).

O advento das redes sociais como o Facebook, o Instagram e o Twitter e de aplicativos para dispositivos móveis como o *Foursquare* e o *Yelp* promoveu um grande volume de dados ao permitir que os usuários pudessem registrar suas atividades compartilhando suas localizações e, assim, ao conjunto desses registros estabelecer trajetórias.

Os estudos sobre trajetórias evoluíram muito nos últimos anos. Spaccapietra *et al.* (2008) exploraram como uma modelagem conceitual pode prover aplicações com suporte direto a trajetórias. A visão de trajetórias como apenas pontos no espaço e no tempo passaram a ser enriquecidas com anotações semânticas, permitindo aos usuários anexar dados semânticos a partes específicas da trajetória.

Mello *et al.* (2019) introduziu uma visão de múltiplos aspectos em trajetórias, isto é, cada ponto em uma trajetória pode conter diversos atributos. Estes atributos podem ser obtidos por sensores ou dispositivos *IoT* (Internet das Coisas, do inglês *Internet of Things*), agregando informações sobre o local, como por exemplo, temperatura, poluição atmosférica, poluição sonora, dentre outros. Também pode-se obter dados inerentes ao próprio objeto: em uma pessoa, pode-se coletar dados relativos à frequência cardíaca, pressão sanguínea, estado emocional, dentre outros. Desta forma, as trajetórias são enriquecidas com inúmeras informações semânticas.

O conjunto de dados de trajetórias multiaspecto pode conter informações que podem gerar uma ampla gama de aplicações. Como exemplo, a prefeitura de uma grande cidade pode melhorar a mobilidade urbana baseada no conjunto de trajetórias considerando aspectos como o horário, o dia da semana, o mês, a condição meteorológica e a velocidade máxima permitida nas vias. Um outro exemplo, um

comerciante pode aumentar seus ganhos identificando o perfil dos consumidores baseados em suas trajetórias. Para isto, técnicas de mineração de dados são utilizadas para extrair informações valiosas que permitem descobrir padrões e encontrar anomalias e relacionamentos entre as trajetórias e podem ser usadas para fazer previsões sobre tendências futuras.

Entre as técnicas de mineração de dados, encontra-se o agrupamento ou *clustering* (em inglês). O objetivo principal de algoritmos de clusterização de trajetórias é agrupar trajetórias semelhantes e prover um melhor entendimento sobre as similaridades existentes entre trajetórias de diferentes objetos ou de um mesmo objeto em diferentes momentos (VARLAMIS *et al*., 2021).

Outra técnica de mineração utilizada são as árvores de decisão, que utilizam uma representação em formato de árvore (nodos e arestas) para classificar (tomar decisões) com base nos atributos das instâncias. A árvore de decisão é criada por um processo conhecido como divisão no valor de atributo, que cria ramificações baseadas em alguma métrica determinada. O processo de divisão prossegue até todos os ramos apresentarem apenas um único rótulo de classificação (BRAMER, 2016).

Liu, Xia e Yu (2005) introduziram uma nova técnica de mineração de dados descrevendo uma abordagem de clusterização por meio da construção de árvores de decisão. A ideia principal é utilizar a árvore de decisão para particionar o conjunto de dados em grupos densos e esparsos, no qual estes podem produzir pontos fora da curva, chamados de *outliers*, e pontos anômalos. No entanto, esta técnica é baseada em aprendizado supervisionado, no qual assume-se que o conjunto de dados possui uma classe e o problema de particionamento do conjunto de dados se torna um problema de classificação. Já Castin e Frénay (2018) propuseram uma abordagem divisiva e aglomerativa para clusterização usando árvores de decisão, considerando o aprendizado não supervisionado, utilizando conjunto de dados não rotulados. Desta forma, os dados podem ser guiados pelos ramos da árvore para descobrir a qual cluster pertencem. Porém, ambas as técnicas não foram utilizadas para avaliar dados em um conjunto de trajetórias.

Como Varlamis *et al.* (2021) apontam, o interesse da comunidade científica por análise de trajetórias, similaridade e agrupamento é alto devido a abundância de equipamentos e dispositivos de rastreamento de localização, permitindo o rastreamento de objetos em movimento. No entanto, a maioria dos trabalhos

relacionados consideram principalmente as propriedades espaço-temporais e medidas derivadas destas dimensões, apresentando pouca atenção à semântica.

Diante deste cenário, este trabalho propõe o desenvolvimento de um algoritmo de modelagem de árvore de decisão e sua utilização para identificar agrupamentos naturais que possuam um significado, em um conjunto de dados de trajetórias multiaspecto. A abordagem, até então inovadora, busca agrupar trajetórias por subconjuntos característicos comuns através de aprendizado não supervisionado.

Os resultados dos agrupamentos obtidos são avaliados permitindo mensurar sua utilidade e sua validade. A validação é feita com métricas internas e métricas externas. Em relação à primeira, utilizam-se o coeficiente de *Silhouette* e os índices de *Calinski-Harabasz* e *Davies-Bouldin* para medir a qualidade em termos de coesão, isto é, o quão próximas estão as trajetórias dentro de um mesmo cluster; separação, medindo o quão bem separado um cluster está dos demais clusters; e conectividade, ou seja, em que medida as trajetórias são colocadas no mesmo cluster em relação aos clusters vizinhos mais próximos no conjunto de trajetórias. Já em relação à segunda, busca-se comparar os clusters identificados com alguma referência externa, ligada a um problema cujo resultado seja conhecido e/ou ao um conjunto de dados específico sobre o qual se aplica o método.

1.1 OBJETIVOS

Este trabalho tem como objetivo geral realizar rotinas de análise de dados a partir de conjuntos de dados de trajetórias com múltiplos aspectos utilizando modelos de árvores de decisão a ser desenvolvido para identificação de agrupamentos naturais.

1.1.1 Objetivos Específicos

1. Desenvolver a modelagem de árvores de decisão para um conjunto de dados de trajetórias multiaspecto, otimizando o algoritmo, de forma a obter um método eficiente para identificação de agrupamentos.
2. Desenvolver uma interface em um painel de controle (*dashboard*) para prover ferramentas de visualização da árvore de decisão modelada, das trajetórias e dos grupos de trajetórias, dos conjuntos de dados em cada nodo da árvore de decisão, bem como das matrizes de frequências dos aspectos presentes nas trajetórias, e das matrizes de correlação utilizando mapas de calor (*heatmap*) para identificação de agrupamentos nos dados.
3. Definir métodos para avaliar os agrupamentos resultantes e as métricas mais adequadas para aferição da qualidade e utilidade dos agrupamentos obtidos.

1.2 ESCOPO DO TRABALHO

O trabalho tem como escopo a revisão do estado da arte de trajetórias com múltiplos aspectos, o desenvolvimento do algoritmo para geração de agrupamentos de trajetórias usando árvores de decisão, a implementação das ferramentas de visualização e das métricas utilizadas para avaliação dos resultados para conjuntos de trajetórias com múltiplos aspectos e o desenvolvimento da monografia. O ambiente usado para as atividades é o Google Colab, que oferece serviço gratuito para as necessidades deste trabalho e provê processamento e armazenamento na nuvem e possui suporte ao IPython, uma ferramenta interativa para linguagem de programação Python.

1.3 MÉTODO DE PESQUISA

O desenvolvimento deste trabalho seguiu os métodos de pesquisa divididos nas seguintes etapas:

1. Realizar uma revisão bibliográfica sobre trajetórias de múltiplos aspectos e tarefas de mineração de dados, com enfoque em árvores de decisão e análise de agrupamentos.
2. Criar e implementar algoritmo de modelagem de árvores de decisão
3. Criar painel de visualização do conjunto de dados das trajetórias, das matrizes de frequência e de correlação.
4. Pesquisar, escolher e preparar conjuntos de dados de trajetórias adequados ao problema em questão.
5. Utilizar outros conjuntos de dados de trajetórias multiaspecto para validar a implementação do algoritmo desenvolvido na etapa 2.
6. Executar o algoritmo, analisar e interpretar os resultados gerados.
7. Identificar agrupamentos naturais presentes nos conjuntos de trajetórias.
8. Avaliar os agrupamentos usando métricas para medir a utilidade e validade dos resultados obtidos.
9. Levantar e testar hipóteses sobre os resultados obtidos, identificando possíveis aplicações.

2 CONCEITOS BÁSICOS E TRABALHOS RELACIONADOS

Neste capítulo são apresentados os conceitos básicos e publicações científicas que fundamentam o desenvolvimento deste trabalho. A seção 2.1 apresenta o conceito de trajetórias brutas e trajetórias multiaspecto de objetos móveis. A seção 2.2 apresenta o conceito de mineração de dados e a seção 2.3 apresenta os conceitos de agrupamento ou clusterização de trajetórias e a definição de árvores de decisão. Na seção 2.4 são apresentadas as métricas de similaridade e a seção 2.5 apresenta as métricas de validação.

2.1 TRAJETÓRIAS BRUTAS E TRAJETÓRIAS MULTIASPECTO

Uma trajetória é definida como uma sequência de pontos localizados no espaço e no tempo e pode ser gerada por qualquer objeto, pessoa ou animal que carrega um dispositivo com tecnologia GPS. Os dados gerados por estes dispositivos são denominados de trajetórias brutas (BOGORNY; BRAZ, 2012).

A representação dos pontos de uma trajetória bruta em um conjunto de dados tipicamente se dá na forma (*tid, x, y, t*) em que *tid* representa o identificador do objeto, *x* e *y* são as coordenadas geográficas do ponto, representando a latitude e a longitude, respectivamente, e *t* refere-se ao instante de tempo que o dado foi registrado (ALVARES *et al.*, 2007). Um exemplo de trajetória bruta é ilustrado na Figura 1.

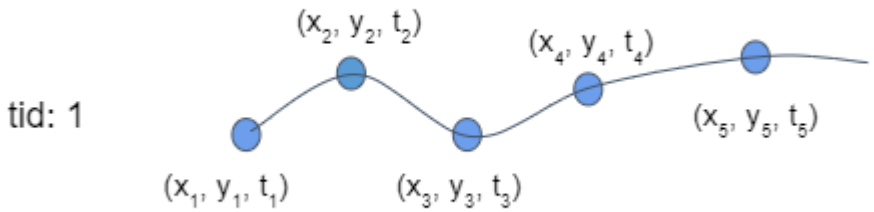


Figura 1. Exemplo de trajetória bruta.

O conjunto de dados coletados por estes dispositivos podem conter informações e conhecimentos novos interessantes e revelar padrões e comportamentos que podem ser úteis em processos de tomadas de decisão nas mais

diversas áreas (BOGORNY; BRAZ, 2012). Laube e Imfeld (2002) foram um dos pioneiros na análise destes dados, ao desenvolver métodos para análise espaço-temporal de movimento relativo dentro de grupos de objetos móveis.

No entanto, estes dados coletados podem conter uma série de informações semânticas associadas que não são capturadas por estes dispositivos. Essas informações também são importantes para extrair conhecimento das análises.

Alvares *et al.* (2007) introduziram uma abordagem para o enriquecimento semântico das trajetórias, baseado no modelo conceitual de trajetórias semânticas proposta por Spaccapietra *et al*. (2008). Nesta abordagem, é apresentado um modelo para representar trajetórias semânticas utilizando *stops* e *moves* para integrar informações geográficas aos pontos de uma trajetória. *Stops* são definidos como pontos de uma parte da trajetória que estão próximos no espaço e no tempo, representando locais de interesse ou POIs (do inglês, *points of interest*) delimitados por um horário de início e fim, uma localização espacial e uma duração mínima. *Moves* são os pontos da trajetória delimitados por duas extremidades que representam dois *stops* consecutivos ou o início da trajetória e o primeiro *stop* ou o último *stop* e o final da trajetória. A figura 2 ilustra um exemplo de trajetória semântica.

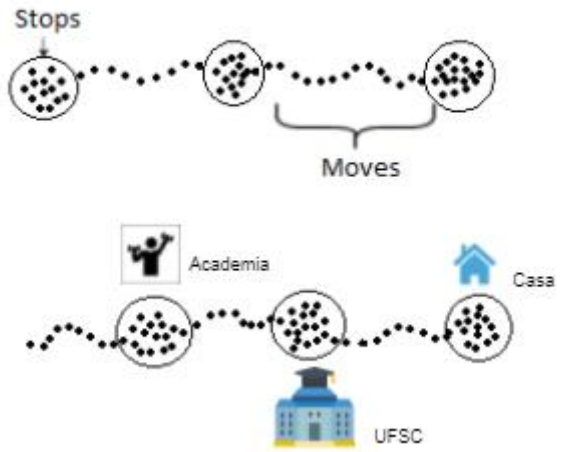


Figura 2. Exemplo de trajetória semântica.

Com o avanço tecnológico e a popularização dos dispositivos com sistemas embarcados, como sensores ou dispositivos *IoT,* os pontos de uma trajetória passaram a ser enriquecidos com inúmeras informações semânticas. Mello *et al.* (2019) propuseram uma nova visão sobre trajetórias, apresentando um novo

paradigma acerca dos objetos móveis. As trajetórias se tornaram objetos complexos com numerosas dimensões de dados contextuais ao movimento e heterogêneas na forma, na qual os autores definiram como aspectos.

Ainda segundo Mello *et al.* (2019), um aspecto é um fato do mundo real que é relevante para a análise de dados de trajetória e é caracterizado por uma categoria de aspecto. Como exemplificado no trabalho referenciado, o aspecto *trem* pertence à categoria de aspecto *meio de transporte* e o aspecto *chuvoso* pertence à categoria de aspecto *condição climática*. Desta forma, um aspecto sempre estará relacionado a pelo menos uma categoria. A associação entre um aspecto e uma categoria fornece o contexto deste aspecto e é denominado de significado semântico.

Um aspecto que não se altera durante o curso de uma trajetória é denominado de aspecto de longo prazo (ALP) e é associado à trajetória multiaspecto como um todo. Quando um aspecto varia frequentemente, o aspecto é associado ao ponto da trajetória e é denominado de aspecto volátil (AV).

Por fim, Mello *et al.* (2019) definem uma trajetória multiaspecto conforme descrito em Definição 1 (tradução nossa).

Definição 1. Uma trajetória multiaspecto tma = (P, C_ALP, om, desc) é uma sequência de pontos P = $<p_1, p_2, ... p_n>$ de um objeto móvel *om*, um conjunto de aspectos de longo prazo C_ALP (que pode ser vazio), sendo C_ALP = $\{ss_1, ss_2, ... ss_p\}$ o conjunto de significados semânticos e, uma descrição *desc*, com $p_i = (x_i, y_i, t_i,$ C_AV), $p_i \in P$, sendo *x* e *y* a posição espacial do *om* no instante de tempo *t* e *C_AV* o conjunto de aspectos voláteis relacionados ao p_i, em que C_AV = $\{ss_1, ss_2, ... ss_q\}$ é o conjunto de significados semânticos (que pode ser vazio).

A figura 3 mostra um exemplo de trajetória multiaspecto. Observa-se na imagem a trajetória de uma pessoa (considerada objeto móvel, *om*) ao longo de um dia. A cada ponto p_i da trajetória estão associadas a localização geográfica (x, y) e o tempo *t* o qual o ponto é registrado. Nota-se, em alguns pontos, aspectos voláteis relacionados. Por exemplo, na residência, determinado dispositivo, como um *smartwatch,* mede a frequência cardíaca e a qualidade do sono ao aferir os estágios do sono. Cada POI identificado (residência, escritório, cinema e restaurante) também é considerado um aspecto volátil, bem como a forma de locomoção do *om* (a pé ou

táxi). A trajetória também contém aspectos de longo prazo, como a condição climática e o dia da semana.

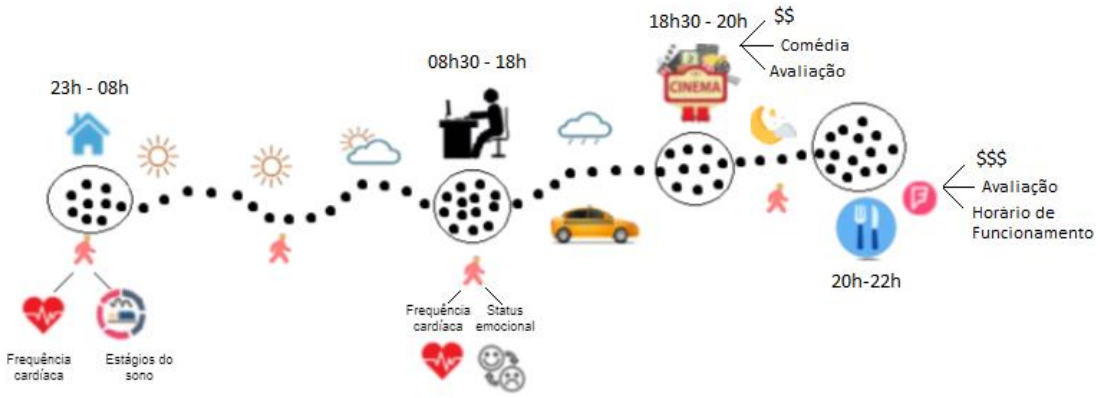


Figura 3. Exemplo de trajetória multiaspecto (adaptado de Mello *et al.*).

2.2 DATA MINING

Nas últimas décadas, notadamente a partir de 1990, a geração e a coleta de dados têm crescido muito rapidamente. Segundo Chan, Han e Yu (1996), milhões de bancos de dados têm sido usados na gestão de negócios, administração governamental, gestão de dados científicos e de engenharia e em diversas outras aplicações.

O rápido crescimento no volume de dados e de banco de dados levou à necessidade de novas técnicas e ferramentas que transformassem os dados processados em informações úteis e conhecimento. Diante disso, *data mining* (mineração de dados, em português) se tornou uma área de pesquisa com crescente importância (FAYYAD, PIATETSKY-SHAPIRO, SMYTH, 1996).

A mineração de dados, também conhecida como descoberta de conhecimento em bancos de dados, significa um processo de extração não trivial de informações implícitas, anteriormente desconhecidas e potencialmente úteis de dados em bancos de dados (FRAWLEY, PIATETSKY-SHAPIRO, MATHEUS, 1992).

O processo para minerar dados pode envolver diversos passos: seleção, limpeza e transformação de dados, busca de padrões e informações relevantes nos dados, apresentação, interpretação e avaliação dos resultados (GHEWARE, KEJKAR, TONDARE, 2014). Um processo típico é mostrado na Figura 4.

As tarefas de mineração podem ser do tipo preditivas ou descritivas. Na primeira, os dados são rotulados e o objetivo é predizer o valor de novas instâncias. Estas tarefas são conhecidas como aprendizado supervisionado, e exemplos são a classificação e a regressão. Já a segunda, os dados não são rotulados e o objetivo é extrair o máximo de informação dos dados disponíveis. As tarefas são conhecidas como aprendizado não supervisionado e, como alguns exemplos, citam-se a clusterização (agrupamento), regras de associação e sumarização.

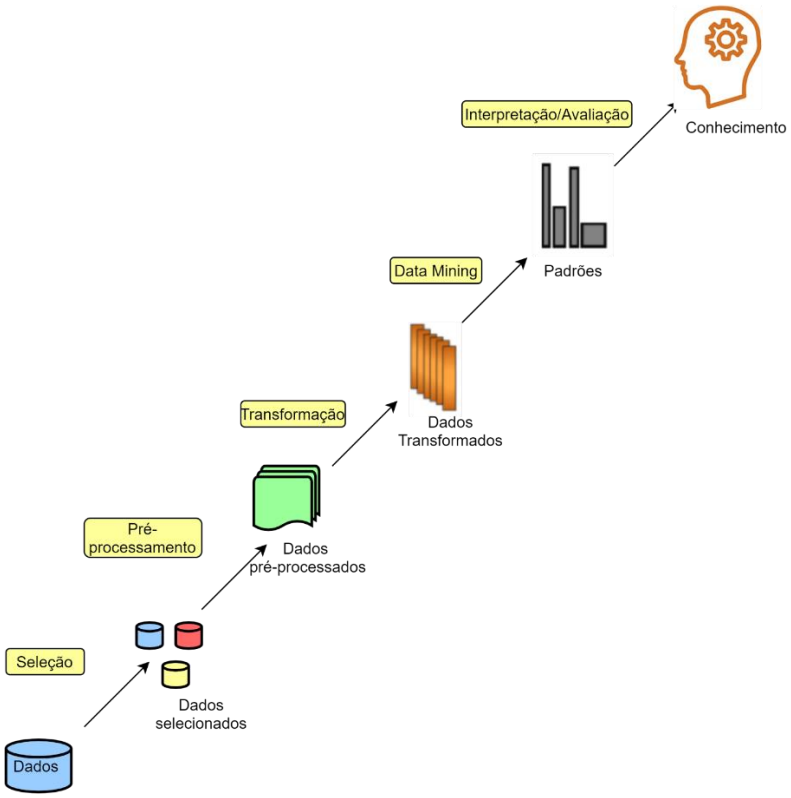


Figura 4. Processo típico de mineração de dados.

2.3 CLUSTERIZAÇÃO

A clusterização é a organização de uma coleção de padrões em clusters (grupos) baseada na similaridade. Os objetos dentro de um cluster válido são mais semelhantes entre si do que a um objeto pertencente a um cluster diferente (JAIN, MURTY, FLYNN, 1999).

Os algoritmos de clusterização podem ser particionais (Figura 5a) ou hierárquicos (Figura 5b). Algoritmos hierárquicos encontram clusters sucessivos

usando clusters previamente estabelecidos, enquanto algoritmos particionais determinam todos os clusters de uma vez. Os algoritmos hierárquicos podem ser aglomerativos (de baixo para cima, *bottom-up*) ou divisivos (de cima para baixo, *top-down*). Os algoritmos aglomerativos começam com cada elemento como um cluster separado e os mesclam em clusters sucessivamente maiores. Algoritmos divisivos começam com todo o conjunto e continuam dividindo-o em clusters sucessivamente menores (MADHULATHA, 2012) .

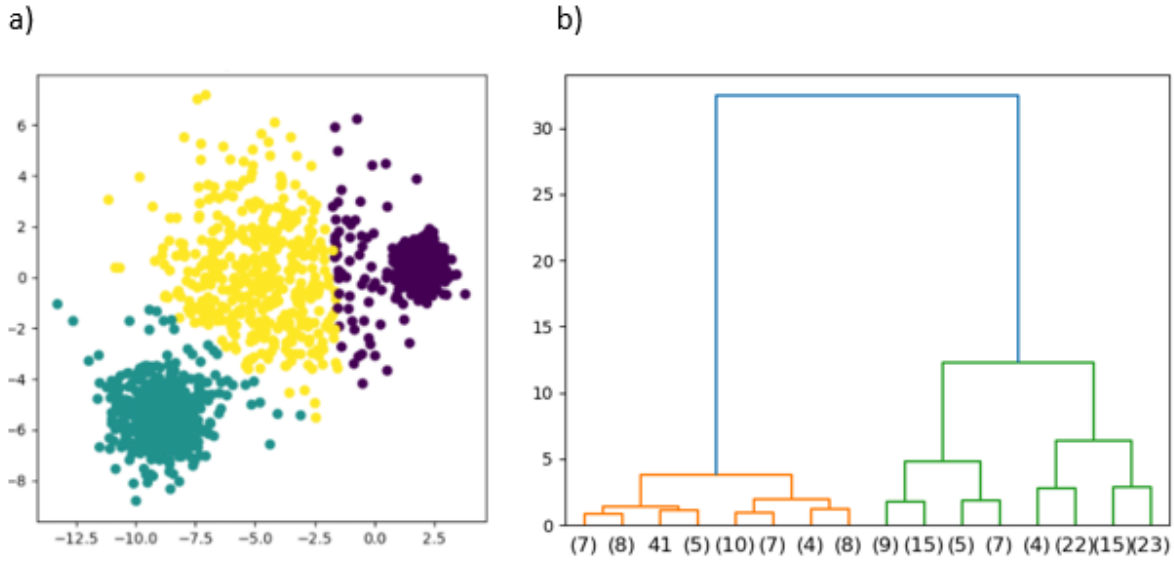


Figura 5. Exemplo de clusterização. Em (a) clusterização particional e em (b) clusterização hierárquica[1].

2.3.1 Clusterização via árvores de decisão

Uma árvore de decisão é uma estrutura de árvore semelhante a um fluxograma em que cada nó interno representa um teste em um atributo; cada ramificação da árvore representa um resultado do teste, e o rótulo da classe é representado por cada nó folha (SHARMA, KUMAR, 2016).

O aprendizado usa uma árvore de decisão como um modelo preditivo que mapeia observações sobre um item para conclusões sobre o valor-alvo do item. Portanto, um algoritmo de árvore de decisão é tipicamente uma tarefa de classificação, ou seja, é um aprendizado supervisionado.

[1] Fonte: Documentação oficial Scikit-learn. Disponível em: <https://scikit-learn.org/stable/modules/clustering.html>. Acesso em: 03 setembro 2021.

No entanto, Liu, Xia e Yu (2005) propuseram uma nova técnica de clusterização através da construção de árvores de decisão. A principal ideia é utilizar a árvore de decisão para particionar o conjunto de dados em clusters e regiões vazias (esparsas) em diferentes níveis de detalhes. A técnica é capaz de encontrar agrupamentos naturais em grandes espaços de alta dimensão de forma eficiente.

O particionamento de dados deve seguir um critério de divisão. Conforme Castin e Frénay (2018) mencionam, a maioria dos trabalhos na literatura utilizam como critério de divisão a redução da variância dos dados.

A árvore inicia com o nó raiz representando o conjunto original dos dados e a cada divisão novos nós são gerados. Cada nó da árvore representa um cluster. Um exemplo de árvore de decisão utilizada para clusterização é ilustrada na Figura 6.

Figura 6. Clusterização via árvore de decisão (adaptado de VARLAMIS *et al.)*.

2.3.2 Clusterização de trajetórias

Yuan *et al.* (2016) apontam a clusterização de trajetórias como a ferramenta mais popular em *data mining*, cujo objetivo é descobrir a similaridade entre objetos móveis, agrupando trajetórias semelhantes no mesmo cluster e encontrando os comportamentos de movimento mais comuns.

Segundo os autores, a medição de similaridade é uma das partes mais importantes em um algoritmo de agrupamento. A similaridade ou a distância entre duas trajetórias diferentes devem ser comparadas antes de serem agrupadas.

Os dados de uma trajetória são diferentes dos dados estáticos utilizados em algoritmos tradicionais de clusterização. Dessa forma, diversos trabalhos são listados na literatura, propondo algoritmos para esta tarefa.

Masciari (2009) propôs uma abordagem com base na estratégia de regionalização adequada, o qual o espaço de busca é particionado em regiões com granularidade adequada de acordo com a posição de rastreamento de um objeto em movimento. O autor propõe um algoritmo cuja métrica é baseada na distância de edição.

Jeung *et al.* (2008) desenvolveram três algoritmos para encontrar grupos de padrões de objetos móveis que viajam juntos por um determinado tempo com técnicas de clusterização baseada em densidade.

Hung *et al.* (2015) propõem um *framework* de clusterização e agregação de trajetórias (CACT – *Clustering and Aggregating Clues of Trajectories*) com objetivo de descobrir rotas de trajetórias que representam os comportamentos de movimento frequentes de um usuário. Os autores observam que trajetórias podem conter *durações silenciosas*, quando não há dados disponíveis para descrever o movimento de um usuário, implicando desafios à mineração de dados. O comportamento de movimento deixa algumas pistas (dicas) em trajetórias observadas. Essas pistas podem ser extraídas de pontos de dados co-localizados no espaço e no tempo.

Nanni e Pedreschi (2006) propõem o T-OPTICS, uma adaptação do OPTICS (baseado em densidade) para agrupar dados de trajetória em uma noção simples de distância entre as trajetórias. Com base no resultado do cluster de trajetória, a focalização temporal é esboçada para explorar a semântica intrínseca da dimensão temporal que pode melhorar a qualidade da clusterização de trajetória.

Yasodha e Ponmuthuramalingam (2012) discorrem sobre medidas de similaridade de trajetórias com dados temporais, e vários algoritmos de agrupamento de dados temporais são classificados e resumidos em diferentes representações. Os autores propõem uma medida útil para ajudar a entender a construção de algoritmos de clusterização com base em análises de agrupamento.

Lee *et al.* (2007) desenvolvem o *framework* de partição e agrupamento (*Partition-and-group*) para clusterização de trajetórias. Os autores observam que subtrajetórias comuns podem acabar ignoradas quando trajetórias em sua totalidade são analisadas. O particionamento das trajetórias segue o princípio da descrição de

comprimento mínimo, o qual a melhor hipótese para um determinado conjunto de dados é a que leva a máxima compressão dos mesmos. O agrupamento das subtrajetórias é feito com algoritmo de clusterização baseado em densidade.

Pelekis *et al.* (2011) estudaram a presença de incerteza em bancos de dados de trajetórias, como erros em dispositivos GPS. Os autores introduzem uma abordagem em três passos para lidar com a incerteza: o primeiro passo é a representação vetorial de ponto de trajetórias que engloba a incerteza subjacente e introduzem uma métrica de distância eficaz para lidar com a incerteza; o segundo passo é a utilização de um novo algoritmo, CenTra, desenvolvido para resolver o problema de encontrar a trajetória central no conjunto de dados e, o terceiro passo, propor uma variante do algoritmo de clusterização Fuzzy C-Means que engloba o CenTra em seu processo de atualização.

Chen *et al.* (2012) propõem um algoritmo de agrupamento baseado em esboço para trajetórias incertas. Em seu algoritmo, um conjunto de segmentos candidatos é construído para representar o modelo de trajetórias incertas com base no particionamento espacial da curva de Hilbert.

Os trabalhos referenciados consideram principalmente as propriedades espaço-temporais e medidas derivadas dessas dimensões, conforme mencionado por Varlamis *et al.* (2021), atribuindo pouca atenção à dimensão semântica. A literatura ainda carece de trabalhos que se aprofundam na tarefa de clusterização de trajetórias semânticas, sobretudo na mais recente abordagem de trajetórias com múltiplos aspectos.

Varlamis *et al.* (2021) propõem uma nova medida de similaridade denominada TraFoS (Trajectory Forest Similarity) que define um novo método para comparar trajetórias multiaspecto. O objetivo supera as limitações das medidas existentes e proporciona uma integração com os algoritmos de agrupamento disponíveis. O TraFoS inclui uma representação multivetorial das trajetórias que melhora a comparação de similaridade. O TraFos permite comparar as trajetórias em cada aspecto e, em seguida, combinar semelhanças em uma única medida.

2.4 MÉTRICAS DE SIMILARIDADE

Similaridade é um valor que reflete a força de relação entre dois objetos, representando o quão semelhante são dois padrões de dados. A clusterização é baseada em alguma métrica de similaridade para agrupar objetos de dados semelhantes. Os clusters são formados de tal forma que quaisquer dois objetos de dados dentro de um cluster têm um mínimo valor de distância e quaisquer dois objetos de dados em diferentes clusters têm um valor de distância máxima (IRANI, PISE, PHATAK, 2016).

A distância euclidiana, a distância de Manhattan, a distância de Minkowski e a similaridade por cosseno são alguns exemplos de métricas de similaridade utilizados em algoritmos de clusterização tradicionais.

Em relação às trajetórias, as métricas de similiaridade listadas na literatura incluem: *Uncertain Movement Similarity* (UMS), *Stops and Moves Similarity Measure* (SMSM), *Longest Common Subsequence* (LCSS), *Edit Distance on Real Sequence* (EDR), *Multidimensional Similarity Measure* (MSM), *Multiple-aspect Trajectory Similarity* (MUITAS) e *Trajectory Forest Similarity* (TraFos).

A métrica UMS proposta por Furtado *et al.* (2017) se baseia na dimensão espacial de trajetórias brutas. Esta métrica não requer a definição de um limite de distância, sendo a distância entre cada dois pontos da trajetória computada com elipses. Dessa forma, as trajetórias são representadas como sequências de elipses e a similaridade é obtida pela proporção da interseção das elipses.

Lehmann, Alvares e Bogorny (2019) propuseram a métrica SMSM, métrica de similaridade que lida com *stops* e *moves* e as dimensões espacial, temporal e semântica. A métrica avalia a correspondência entre duas trajetórias através de uma distância máxima estipulada e utiliza uma pontuação (*score*) tanto para os *stops* quanto para os *moves*, atribuindo pesos diferentes referentes ao grau de importância, para encontrar a paridade entre duas trajetórias.

O LCSS proposto por Vlachos *et al.* (2002) é uma métrica no qual dois pontos se correspondem quando suas distâncias são menores que um determinado limite estipulado. Quanto maior for a subsequência comum de correspondência entre duas trajetórias, mais similares elas são. O LCSS considera todos os aspectos igualmente

importantes e relevantes e requer uma correspondência estrita para todos os aspectos de cada *stop* ou *move* para considerá-los como semelhantes.

O EDR proposto por Chen, Özsu e Oria (2005) é uma métrica baseada na distância de edição, métrica utilizada para comparação de *strings*. O algoritmo mede o número mínimo de inserções, deleções e substituições de pontos (*stops* e *moves*) necessários para transformar uma trajetória em outra. Quanto menor for o valor obtido, mais similares são as trajetórias.

Furtado *et al.* (2016), em outro trabalho, propuseram o MSM, em que dadas duas trajetórias, para cada ponto da primeira, o MSM procura a melhor correspondência com a segunda. As pontuações ponderadas das correspondências são adicionadas para compor a paridade entre as duas trajetórias. O MSM examina cada aspecto de uma trajetória separadamente e oferece suporte a pesos diferentes para cada aspecto, atribuindo mais ou menos importância aos aspectos, de acordo com as necessidades da aplicação. No entanto, o MSM não considera as possíveis relações que possam existir entre os atributos ou os aspectos, o que o torna menos robusto no caso de trajetórias multi-aspecto.

As limitações do MSM são contornadas pela métrica MUITAS, proposta por Petry *et al.* (2019), ao oferecer suporte a aspectos compostos que agregaram vários atributos e fornecem uma comparação de trajetória mais abrangente. Os múltiplos aspectos consideram a relação semântica entre os atributos de uma trajetória. Além disso, ele suporta o uso de limiares em cada valor de aspecto e o uso de pesos diferentes entre os aspectos, aumentando assim a flexibilidade da métrica de similaridade de trajetória.

O MUITAS apresenta algumas limitações, conforme apontam Varlamis *et al*. (2021). A popularidade dos atributos compartilhados por vários pontos ou a frequência de ocorrência dos próprios pontos não são explicitamente considerados.

Varlamis *et al*. (2021) propõem uma nova medida de similaridade, o TraFoS, que contornam algumas limitações apresentadas em trabalhos relacionados. Os autores apontam que as medidas de similaridade mencionadas ou pressupõem que os aspectos são totalmente alheios entre si, ou combinam indiscriminadamente todos os atributos, mesmo aqueles que não se relacionam com outros. Elas também são limitadas à ordem de *stops* e/ou *moves*, e procuram por subtrajetórias mais curtas ou

mais longas. Além disso, eles não consideram a frequência de ocorrência de cada valor de aspecto na trajetória.

O TraFoS leva em consideração a frequência dos valores dos aspectos da trajetória, gera uma representação de todas as trajetórias em um mesmo espaço vetorial e considera todo o conjunto de trajetórias antes de compará-las aos pares.

2.5 MÉTRICAS DE VALIDAÇÃO

Uma das questões mais importantes na análise de clusterização é a avaliação dos resultados do clustering para encontrar o particionamento que melhor se adapta aos dados (HALKIDI, BATISTAKI, VAZIRGIANNI, 2001).

O procedimento de avaliação dos resultados de um algoritmo de agrupamento é conhecido como validação de agrupamento. Segundo Theodoridis e Koutroubas (2009) existem três abordagens utilizadas para avaliar a validação de clusters. A primeira é baseada em critérios externos, em que se avaliam os resultados de um algoritmo de agrupamento com base em uma estrutura pré-especificada, que é imposta a um conjunto de dados e reflete nossa intuição sobre a estrutura de agrupamento do conjunto de dados. A segunda abordagem é baseada em critérios internos, em que avaliam-se os resultados de um algoritmo de agrupamento em termos de quantidades que envolvem os vetores do próprio conjunto de dados (por exemplo, matriz de proximidade). A terceira abordagem é baseada em critérios relativos. Aqui, a ideia básica é a avaliação de uma estrutura de agrupamento comparando-a com outros esquemas de agrupamento, resultantes do mesmo algoritmo, mas com valores de parâmetros diferentes.

Existem dois critérios propostos para avaliação de clusterização e seleção de um esquema de clusterização ideal, propostos por Berry e Linoff (1996):

- **Compactação ou coesão**: os membros de cada cluster devem estar o mais próximo possível uns dos outros;
- **Separabilidade**: mede o quão bem separado um cluster está dos demais clusters.

2.5.1 Validação Interna

2.5.1.1 Coeficiente Silhouette

O coeficiente ou índice Silhouette (CS) é um valor que mede o quão similar um objeto é em relação ao seu próprio cluster (coesão) em comparação com os demais clusters (separação). O coeficiente varia entre -1 a +1, em que um alto valor indica que o objeto está bem relacionado ao seu cluster. Este coeficiente pode ser calculado utilizando qualquer métrica de similaridade.

O CS é definido para cada objeto *i* do conjunto de dados e o coeficiente s_i é calculado da seguinte forma:

1. Para cada objeto *i* calcula-se a média de distância a_i entre *i* e todos os demais pontos do cluster o qual *i* pertence.
2. Para todos os demais clusters *C*, o qual *i* não pertence, calcula-se a média de distância *d(i,C)* de *i* para todos os objetos de *C*. O objetivo é buscar a menor distância *d(i,C)* que é definida como $b_i = min_C\, d(i,C)$. O valor de b_i pode ser visto como a distância entre *i* e o cluster vizinho mais próximo.
3. Finalmente, o coeficiente Silhouette de *i* é definido pela fórmula:

$$S_i = \frac{(b_i - a_i)}{max(a_i, b_i)} \tag{1}$$

2.5.1.2 Índice Calinski-Harabasz

O índice de Calinski-Harabasz (ICH) é a razão da soma das distâncias intercluster e das distâncias intracluster para todos os clusters. Quanto maior a pontuação, melhor a clusterização. Uma pontuação alta significa um cluster denso e bem separado.

Dado um conjunto de dados *E* de tamanho n_E o qual foi clusterizado em *k* clusters, o ICH *S* é definido em (2), em que $tr(B_k)$ é a matriz de distância intercluster (3) e $tr(W_k)$ é a matriz de distância intracluster (4). Em (3) e (4), C_q representa o conjunto de objetos do cluster *q*, c_q é o centro do cluster *q*, C_E é o centro do *E* e n_q é o número de objetos no cluster *q*.

$$S_i = \frac{tr(B_k)}{tr(W_k)} \cdot \frac{n_E - k}{k - 1} \tag{2}$$

$$Wk = \sum_{q=1}^{k} \sum_{x \epsilon Cq} (x - Cq)(x - Cq)^T \tag{3}$$

$$B_k = \sum_{q=1}^{k} n_q (c_q - c_E)(c_q - c_E)^T \tag{4}$$

2.5.1.3 Índice Davies-Bouldin

O índice de Davies-Bouldin (IDB) representa a média de similaridade entre clusters, em que a similaridade é uma medida que compara a distância entre os clusters com o tamanho dos próprios clusters. Quanto menor o índice IDB, melhor é a separação entre os clusters.

O índice é definido como a média de similaridade entre cada cluster C_i para i = 1, 2, ... k com o seu cluster mais próximo C_j. A similaridade é definida como R_{ij} (5) no qual:

- s_i é a média de distância entre cada ponto do cluster i com o centroide do seu cluster;
- d_{ij} é a distância entre os centroides do clusters i e j.

$$R_{ij} = \frac{s_i + s_j}{d_{ij}} \tag{5}$$

Por fim, o índice DB é calculado como:

$$DB = \frac{1}{K}\sum_{i=1}^{k} max_{i \neq j} R_{ij} \quad (6)$$

2.5.2 Validação Externa

2.5.2.1 Completeness Score, Homogeneity Score e V-measure score

As três métricas de validação externa denominadas, em inglês, *Completeness Score*, *Homogeneity Score* e *V-measure* (HIRSCHBERG, ROSENBERG, 2007) são métricas que fazem referência à rotulagem dos clusters dada uma verdade fundamental, sendo mais usual o emprego do termo em inglês: *ground truth*. O *ground truth* é a informação que se sabe ser real ou verdadeira, fornecida por observação direta e medição (evidência empírica) em oposição à informação fornecida por inferência.

As métricas avaliam se a inferência inicial do rótulo de um determinado cluster corresponde à informação real (*ground truth*), indicando ao final se houve ou não uma clusterização satisfatória.

Em relação à primeira, o resultado de agrupamento satisfaz a completude se todos os pontos de dados que são membros de uma determinada classe são elementos do mesmo agrupamento. A pontuação dessa métrica varia entre 0 (zero) e 1, sendo que esta denota uma rotulagem perfeitamente completa.

A segunda métrica, o resultado de agrupamento satisfaz a homogeneidade se todos os seus agrupamentos contiverem apenas pontos de dados que são membros de uma única classe. A pontuação é semelhante à anterior, sendo que o valor 1 representa uma rotulagem perfeitamente homogênea.

Finalmente, a terceira métrica, *V-measure*, consolida as duas métricas anteriores por meio da média harmônica entre elas. A pontuação também varia entre 0 (zero) e 1, sendo que 1 representa uma rotulagem perfeitamente completa.

2.5.2.2 Adjusted Mutual Info Score

Segundo os autores Vinh, Epps, Bailey (2010), dados dois agrupamentos U e V, a Informação Mútua (MI) mede a informação que os clusters compartilham entre si,

considerando suas entropias (conjunta e condicional). Em outras palavras, o IM mede a dependência mútua entre os clusters, quantificando a quantidade de informações obtidas por um dos clusters por meio da observação do outro cluster. Ao se conhecer um dos agrupamentos, reduz-se a incerteza sobre o outro.

A Informação Mútua Ajustada (AMI) é um ajuste da pontuação da Informação Mútua (MI) para levar em conta o acaso. A comparação é feita considerando uma clusterização aleatória dos dados

Em relação ao particionamento dos dados, pode-se então perguntar: se um conjunto fosse particionado aleatoriamente, qual seria a distribuição de probabilidades? Qual seria o valor esperado da informação mútua? A informação mútua ajustada ou AMI subtrai o valor esperado do MI, de modo que o AMI é 0 (zero) quando duas distribuições diferentes são aleatórias e um quando duas distribuições são idênticas.

Isso explica o fato de que o MI geralmente é maior para dois clusters com um número maior de clusters, independentemente de haver realmente mais informações compartilhadas.

2.5.2.3 Adjusted Rand Score

O *Rand Index* é uma medida de similaridade entre dois agrupamentos considerando todos os pares de amostras, contando os pares que são atribuídos em um mesmo agrupamento ou em diferentes agrupamentos que foram previstos e os que foram gerados. A rotulagem prevista pode ocorrer ao acaso, isto é, de forma aleatória e, para estes casos, utiliza-se o coeficiente ajustado denominado de *Adjusted Rand Score* (ARI).

O ARI é, então, usado para garantir um valor próximo de zero para rotulagem aleatória independentemente do número de clusters e amostras e exatamente 1,0 quando os clusters são idênticos.

2.5.2.4 Fowlkes Mallows Score

Fowlkes e Mallows (1983) é uma métrica de similaridade definida como a média geométrica entre precisão e sensibilidade (*recall*). A medida de precisão-sensibilidade

é uma medida útil para avaliar o sucesso da previsão quando se tem classes ou rotulagens muito desequilibradas. Na recuperação de informações, a precisão é uma medida da relevância do resultado, enquanto a recuperação é uma medida de quantos resultados realmente relevantes são retornados.

O índice varia entre 0 e 1. Um resultado alto indica uma grande similaridade entre os clusters.

3 ÁRVORE DE DECISÃO PARA CLUSTERIZAÇÃO DE TRAJETÓRIAS MULTIASPECTO

Este capítulo apresenta o método proposto para a modelagem da árvore de decisão com o objetivo de realizar a clusterização de trajetórias multiaspecto com a abordagem particional dos dados. O algoritmo conta com o desenvolvimento de uma interface para visualização dos resultados, sendo utilizada na rotina de análises dos agrupamentos encontrados.

A seção 3.1 apresenta as etapas para o desenvolvimento da árvore de decisão e a seção 3.2 apresenta a interface de visualização (*dashboard*) dos resultados.

3.1 MODELAGEM DA ÁRVORE DE DECISÃO

As etapas iniciais anteriores à modelagem da árvore de decisão são fundamentais para o êxito do algoritmo.

A primeira etapa refere-se ao entendimento dos dados, analisando os atributos e os aspectos relacionados às trajetórias coletadas com o intuito de identificar possíveis problemas nos dados e possíveis *insights* interessantes que podem ser utilizados nas análises dos resultados. Essa etapa visa selecionar quais os aspectos que serão utilizados na modelagem da árvore, tendo em vista que alguns aspectos podem não ser relevantes ou gerar resultados pouco informativos.

A segunda etapa refere-se à preparação dos dados que envolve a limpeza de dados, buscando tratar questões como dados faltantes e dados ruidosos (dados sem sentido que não podem ser interpretados por máquinas), e a transformação de dados que visa formatá-los de maneira mais apropriada e adequada para a tarefa de mineração.

A terceira etapa refere-se à modelagem da árvore de decisão. Inicialmente, a árvore começa pelo nó denominado *raiz* que contém o conjunto de todas as trajetórias multiaspecto. Sobre este conjunto de dados, uma matriz de frequência é elaborada a partir dos aspectos selecionados na primeira etapa. Essa matriz representa a quantidade de vezes que cada aspecto aparece em uma determinada trajetória. A Figura 5 exemplifica uma matriz de frequência para duas trajetórias (a, b) dos aspectos

selecionados (dias da semana, categoria do POI e condição climática) mostrando a quantidade de vezes que cada aspecto foi registrado para cada trajetória.

traj_id	day_Monday	day_Tuesday	day_Saturday	type_College	type_Food	type_Nightlife	weather_Clouds	weather_Clear
a	6	7	5	6	15	1	6	12
b	9	4	5	0	13	8	4	13

Figura 7. Exemplo de matriz de frequência.

A etapa a seguir refere-se à divisão do conjunto de dados em dois grupos, isto é, a divisão do nó raiz em dois nós filhos. Para isto, deve-se escolher a forma como as trajetórias serão particionadas. As métricas propostas para a divisão são a média ou a mediana dos valores de cada aspecto da matriz de frequência.

Considerando o exemplo da Figura 7 que apresenta oito aspectos, o algoritmo irá gerar oito subárvores diferentes. A Figura 8 ilustra um exemplo de uma subárvore gerada pelo aspecto condição climática nublado (*weather_Clouds*), considerando a média como métrica para divisão. O nó raiz representa o conjunto total de trajetórias antes da divisão, o nó filho da esquerda representa as trajetórias cuja frequência do aspecto nublado é inferior à média e o nó filho da direita representa as trajetórias em que a frequência de nublado é superior à média.

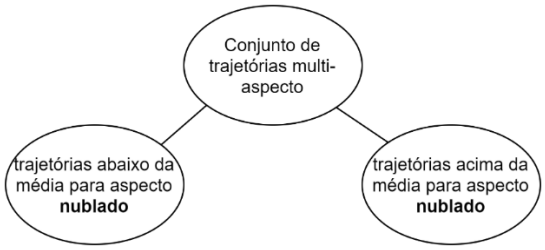


Figura 8. Subárvore gerada pelo particionamento das trajetórias pelo aspecto Nublado.

A última etapa do algoritmo é determinar qual aspecto será escolhido para divisão dos dados, isto é, das subárvores geradas na etapa anterior, apenas uma é selecionada para dividir os dados.

Os métodos propostos para seleção do aspecto são:

- divisão binária (**DB**);
- mínima variância (**MV**);
- redução máxima da variância (**RMV**) e;

- redução máxima da variância através da relação entre todos aspectos (**MRV**).

O método de **divisão binária** avalia a diferença da quantidade de trajetórias entre os grupos resultantes (nós filhos) do particionamento de cada subárvore. O aspecto escolhido para a divisão é aquele que gera a menor diferença da quantidade entre os grupos.

Considere dois aspectos no conjunto de dados: *lazer* e *alimentação.* Suponha que para a subárvore gerada por *lazer* o nó esquerdo tenha sete trajetórias e o nó direito tenha oito trajetórias e, para a subárvore gerada por *alimentação*, o nó esquerdo tenha cinco trajetórias e o nó direito tenha 10 trajetórias. Nesse exemplo, a diferença da quantidade de trajetórias para o aspecto *lazer* é de uma trajetória e para o aspecto *alimentação* de cinco trajetórias. Portanto, o método de divisão binária irá selecionar o aspecto *lazer* para efetuar a divisão dos dados, uma vez que apresentou a menor diferença entre os grupos.

O método da **mínima variância** calcula a média da variância entre os grupos resultantes da divisão para cada subárvore. A variância é uma medida de dispersão que mostra o quão distante cada valor do conjunto de dados está do valor central (média ou mediana). Nesse método, o aspecto que gerar a menor média de variância será escolhido.

Tomando o exemplo anterior, suponha que o nó esquerdo da subárvore de *lazer* tenha apresentado uma variância de 3,5 e o nó direito uma variância de 7. Para a subárvore de *alimentação*, os valores da variância foram 2,3 e 9, para os nós esquerda e direita, respectivamente. O aspecto *lazer* tem média de 5,25 e o aspecto *alimentação* tem média 5,65. Neste caso, *lazer* apresentou a menor média de variância e é escolhido para dividir os dados.

O método de **redução máxima da variância** também calcula a média da variância entre os grupos resultantes da divisão para cada subárvore, mas nesse método o valor resultante é subtraído da variância inicial (nó raiz). O aspecto escolhido para dividir os dados é aquele que gera a maior redução.

Seguindo o exemplo supracitado, suponha que no nó raiz, *lazer* apresentou variância inicial 6 e *alimentação* apresentou variância 6,5. A redução de variância para *lazer* é então calculada subtraindo-se a variância inicial da média de variância dos nós

filhos (5,65) que resulta em 0,75. De forma análoga, *alimentação* (5,65) resulta em 0,85. Logo, considerando a máxima redução de variância, o aspecto *alimentação* será o escolhido.

O método **redução máxima da variância através da relação entre todos os aspectos** avalia para cada subárvore gerada a média de redução de todos os demais aspectos. O aspecto que gera a maior redução é escolhido.

Considere um conjunto de dados com três aspectos para a categoria de POI: *lazer*, *alimentação* e *universidade*. Para cada subárvore gerada, este método irá utilizar o método anteriormente descrito (redução máxima da variância) para todos os demais aspectos. Por exemplo, para a subárvore de *lazer*, a redução máxima de variância também será calculada para os aspectos *alimentação* e *universidade*. Em seguida, o método irá efetuar a média dos valores obtidos. Após calcular as médias para todas as subárvores, aquela que apresentar o maior valor indicará o aspecto selecionado para a divisão das trajetórias.

Após selecionado o aspecto por algum dos métodos propostos e o conjunto de dados ser particionado em dois grupos ou nós filhos, o algoritmo retorna à etapa três calculando novas matrizes de frequência para cada nó filho e seguindo os passos para novas divisões. A métrica selecionada (média ou mediana) e o método de seleção do aspecto não se alteram.

O algoritmo deve ser configurado para estabelecer um critério de parada, onde o conjunto de dados não será mais particionado. Os critérios de parada podem ser definidos pela profundidade da árvore e/ou pela quantidade mínima de trajetórias que um nó deve ter, isto é, após uma divisão se o nó resultante apresentar valor menor ou igual de trajetórias ao determinado pelo critério de parada, este nó não será mais particionado.

A Figura 9 resume o processo envolvido na tarefa de mineração de dados e a modelagem da árvore de decisão. O processo orientou-se pelo modelo *Cross Industry Standard Process for Data Mining* (Processo Padrão Interindústrias para Mineração de Dados, em português), abreviado como CRISP-DM.

As principais etapas do CRISP-DM descritas nesta seção referem-se ao entendimento dos dados, à preparação dos dados e à modelagem. A avaliação e o entendimento do negócio fazem parte da seção seguinte que descreve a interface de visualização, e que fornece suporte para a análise e compreensão dos resultados.

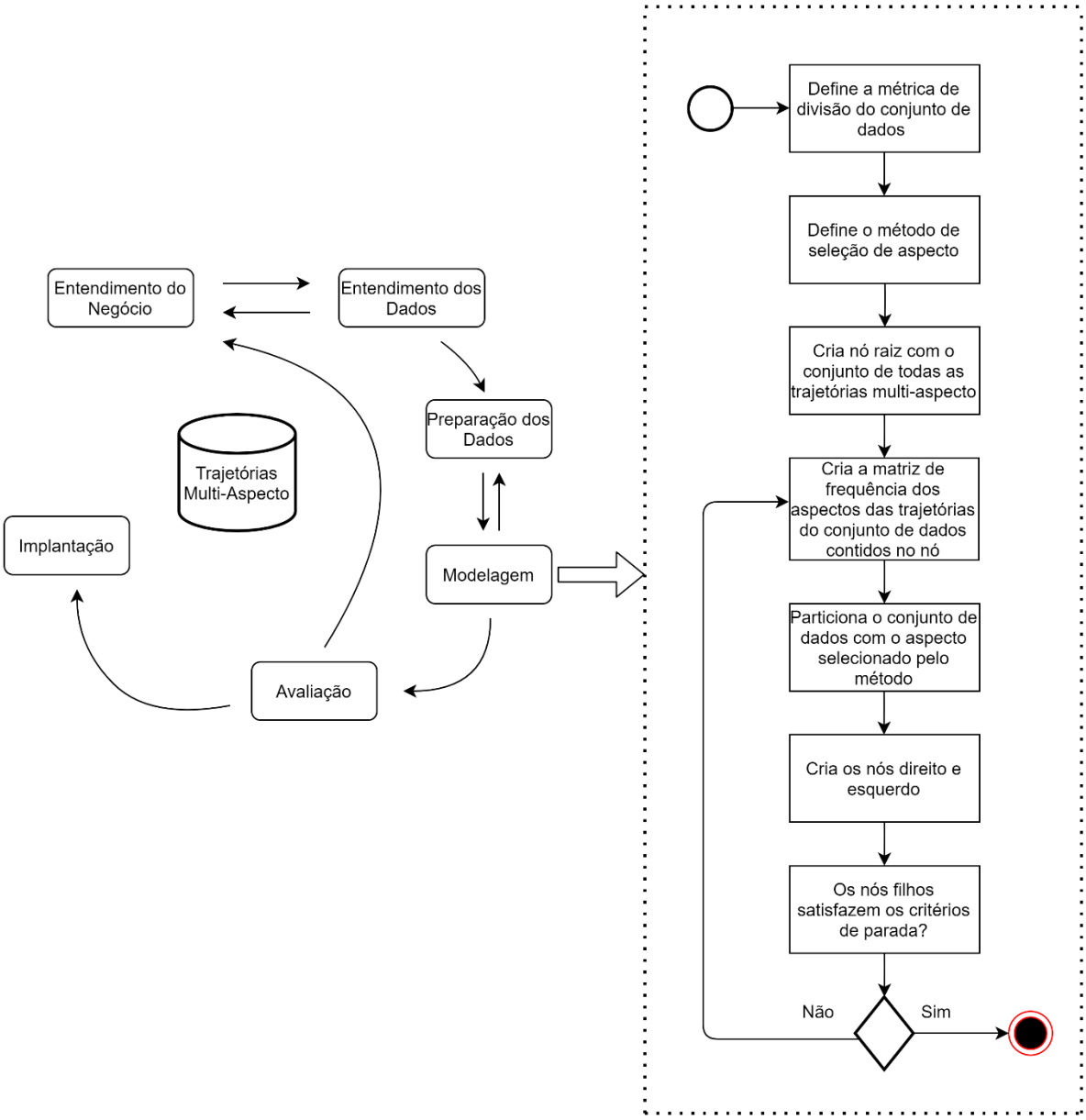


Figura 9. Modelo CRISP-DM.

3.2 INTERFACE DE VISUALIZAÇÃO

A interface de visualização foi desenvolvida com o objetivo de auxiliar a avaliação dos resultados obtidos da modelagem da árvore de decisão, tendo em vista que existem diversas formas de se construí-la. Dessa forma, a interface facilita a comparação e a identificação do método de particionamento dos dados de trajetórias que gera os melhores agrupamentos.

A Figura 10 ilustra o painel e as opções de visualização descritas a seguir. O menu superior oferece as opções de configuração dos dados a serem mostrados, como o nó da árvore (*Tree Node*), a quantidade de registros do conjunto de dados a ser exibido (*# rows*) e as trajetórias específicas de um usuário (*Select User*). O menu inferior oferece as ferramentas de visualização.

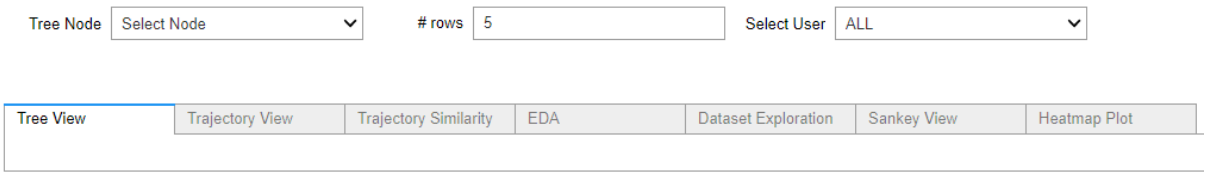


Figura 10. Interface da ferramenta de visualização da árvore de decisão modelada e das trajetórias multi-aspecto.

- *Tree Node*: permite selecionar um nó específico da árvore para detalhar as trajetórias pertencentes àquele cluster;
- *# rows*: permite determinar a quantidade de registros da tabela do conjunto de trajetórias que serão mostrados;
- *Select User*: permite selecionar as trajetórias de um determinado usuário dentro do cluster;
- *Tree View* (Figura 11): mostra a árvore de decisão criada pelo algoritmo descrito anteriormente;
- *Trajectory View* (Figura 12): detalha as trajetórias pertencentes ao cluster selecionado em *Tree Node*. Nesta opção, são identificados os usuários presentes no grupo, o total de trajetórias agrupadas, a entropia e a tabela de todos os pontos das trajetórias multiaspecto. A entropia calculada é uma forma de medir o grau médio de incerteza, permitindo a quantificação da informação. Em outras palavras, a entropia mede a relação entre as trajetórias e os usuários do cluster, informando se o cluster gerado agrupou mais trajetórias de apenas um determinado usuário, ou se o agrupamento foi mais distribuído entre todos os usuários. O cálculo da entropia é mostrado na Fórmula 7, que representa a soma das probabilidades de ocorrência de uma trajetória de cada usuário do cluster multiplicadas pelo logaritmo da probabilidade. Em cada cluster, o resultado da entropia é normalizado

considerando a entropia máxima de acordo com o número de usuários agrupados, a fim de ajustar os valores dos diferentes clusters para uma escala comum, no intervalo entre 0 e 1. Valores mais próximos a 0 são desejáveis, uma vez que representa um cluster que agrupou mais trajetórias de apenas um usuário, enquanto que valores próximos a 1 indica as trajetórias estão mais distribuídas entre dois ou mais usuários.

$$H(x) = -\sum_{i=1}^{n} P(x_i) log P(x_i) \tag{7}$$

- *Trajectory Similarity* (Figura 13): mostra a média de similaridade entre as distâncias das trajetórias do cluster usando as métricas de similaridade MUITAS e MSM, bem como as matrizes de distância de ambas as métricas.
- *EDA (Figura 14)*: esta opção refere-se à análise exploratória dos dados, mostrando alguns gráficos informativos e relevantes das trajetórias do cluster;
- *Dataset exploration* (Figura 15): mostra a matriz de frequência das trajetórias do cluster.
- *Sankey View (*Figura *16)*: mostra o diagrama aluvial ou de *Sankey*. Este diagrama é uma outra forma de visualizar a árvore de decisão de maneira dinâmica, que permite observar de forma mais intuitiva como as trajetórias estão sendo agrupadas a cada divisão. A estrutura se dá na forma de uma árvore na horizontal, estando a raiz mais à esquerda e os nós folhas mais à direita. A espessura da linha que liga um cluster ao outro indica uma noção da quantidade de trajetórias após uma divisão.
- *Heatmap plot* (Figura 17): representa o mapa de calor que mostra por meio da diferença da magnitude de cor a visualização dos dados da relação entre as trajetórias dos usuários e os aspectos das trajetórias.

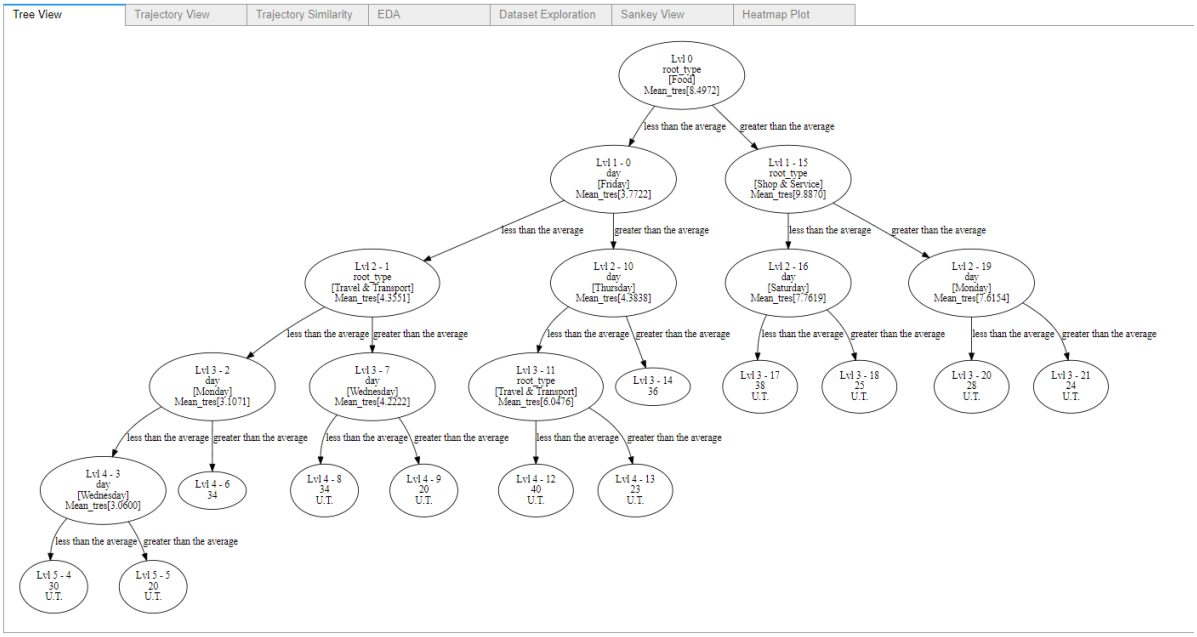
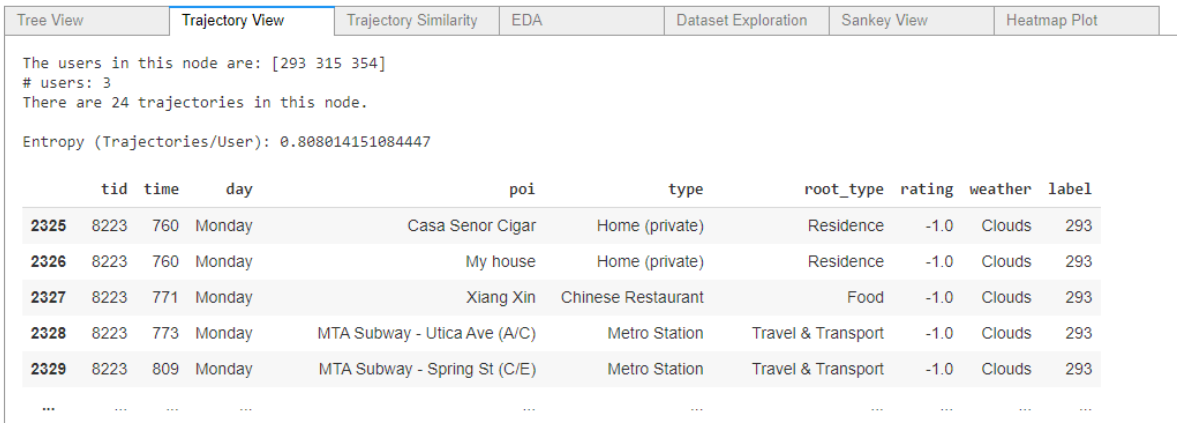
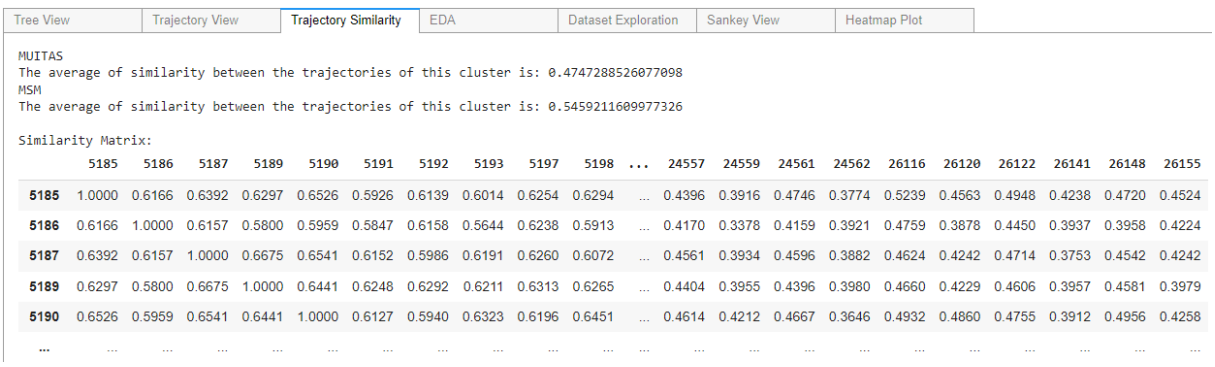


Figura 11. Visualização da árvore de decisão modelada.



	tid	time	day	poi	type	root_type	rating	weather	label
2325	8223	760	Monday	Casa Senor Cigar	Home (private)	Residence	-1.0	Clouds	293
2326	8223	760	Monday	My house	Home (private)	Residence	-1.0	Clouds	293
2327	8223	771	Monday	Xiang Xin	Chinese Restaurant	Food	-1.0	Clouds	293
2328	8223	773	Monday	MTA Subway - Utica Ave (A/C)	Metro Station	Travel & Transport	-1.0	Clouds	293
2329	8223	809	Monday	MTA Subway - Spring St (C/E)	Metro Station	Travel & Transport	-1.0	Clouds	293
...	...	...	...	...	...	...	...	...	...

Figura 12. Visualização do *dataset* do cluster de trajetórias multi-aspecto e informações sobre usuários, trajetórias e entropia.



	5185	5186	5187	5189	5190	5191	5192	5193	5197	5198	...	24557	24559	24561	24562	26116	26120	26122	26141	26148	26155
5185	1.0000	0.6166	0.6392	0.6297	0.6526	0.5926	0.6139	0.6014	0.6254	0.6294	...	0.4396	0.3916	0.4746	0.3774	0.5239	0.4563	0.4948	0.4238	0.4720	0.4524
5186	0.6166	1.0000	0.6157	0.5800	0.5959	0.5847	0.6158	0.5644	0.6238	0.5913	...	0.4170	0.3378	0.4159	0.3921	0.4759	0.3878	0.4450	0.3937	0.3958	0.4224
5187	0.6392	0.6157	1.0000	0.6675	0.6541	0.6152	0.5986	0.6191	0.6260	0.6072	...	0.4561	0.3934	0.4596	0.3882	0.4624	0.4242	0.4714	0.3753	0.4542	0.4242
5189	0.6297	0.5800	0.6675	1.0000	0.6441	0.6248	0.6292	0.6211	0.6313	0.6265	...	0.4404	0.3955	0.4396	0.3980	0.4660	0.4229	0.4606	0.3957	0.4581	0.3979
5190	0.6526	0.5959	0.6541	0.6441	1.0000	0.6127	0.5940	0.6323	0.6196	0.6451	...	0.4614	0.4212	0.4667	0.3646	0.4932	0.4860	0.4755	0.3912	0.4956	0.4258
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

Figura 13. Matriz de similaridade das métricas MUITAS e MSM e médias de similaridade entre as trajetórias de um cluster.

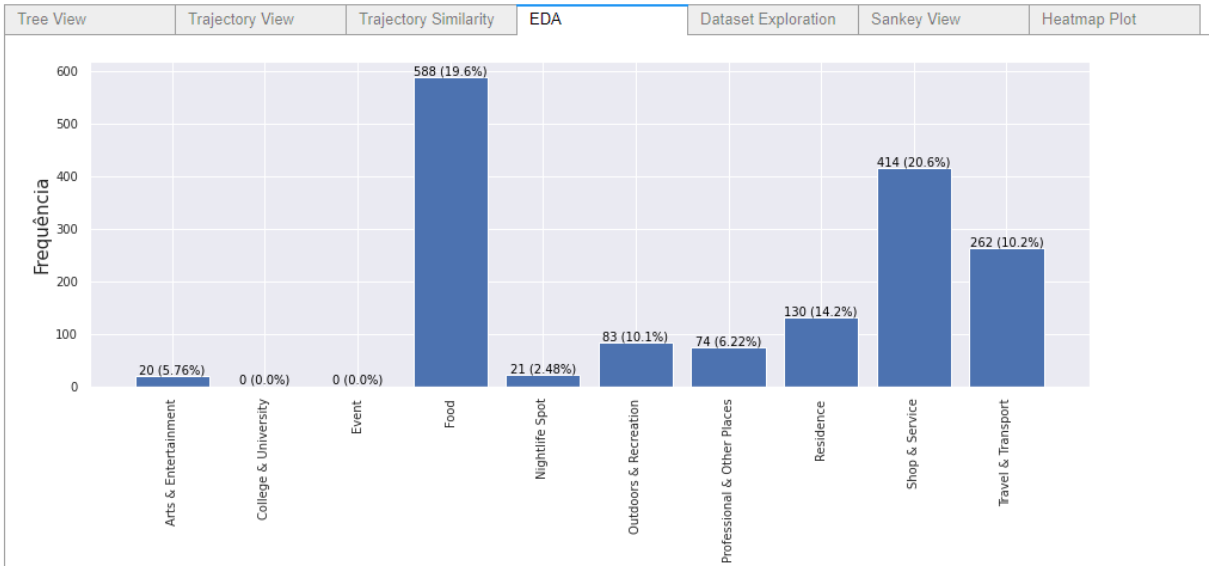
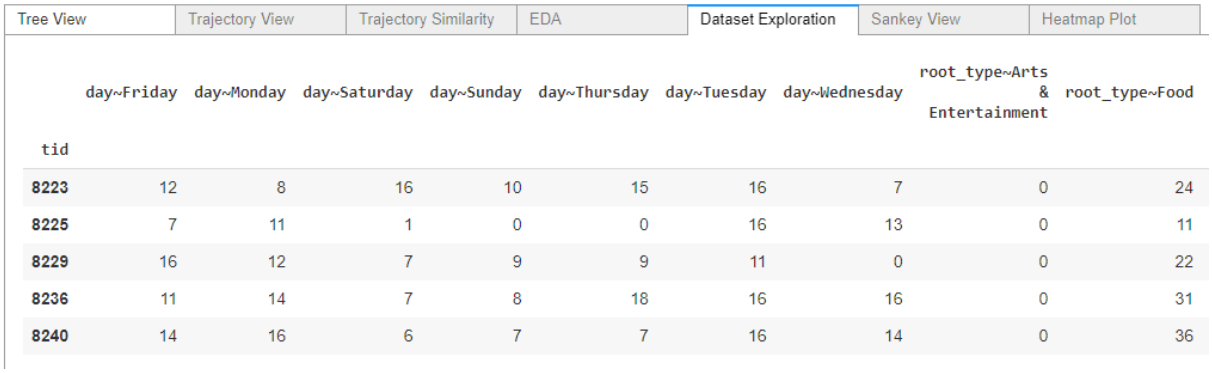


Figura 14. Ferramenta para análise exploratória de dados.



tid	day~Friday	day~Monday	day~Saturday	day~Sunday	day~Thursday	day~Tuesday	day~Wednesday	root_type~Arts & Entertainment	root_type~Food
8223	12	8	16	10	15	16	7	0	24
8225	7	11	1	0	0	16	13	0	11
8229	16	12	7	9	9	11	0	0	22
8236	11	14	7	8	18	16	16	0	31
8240	14	16	6	7	7	16	14	0	36

Figura 15. Matriz de frequências dos aspectos das trajetórias de um determinado cluster.

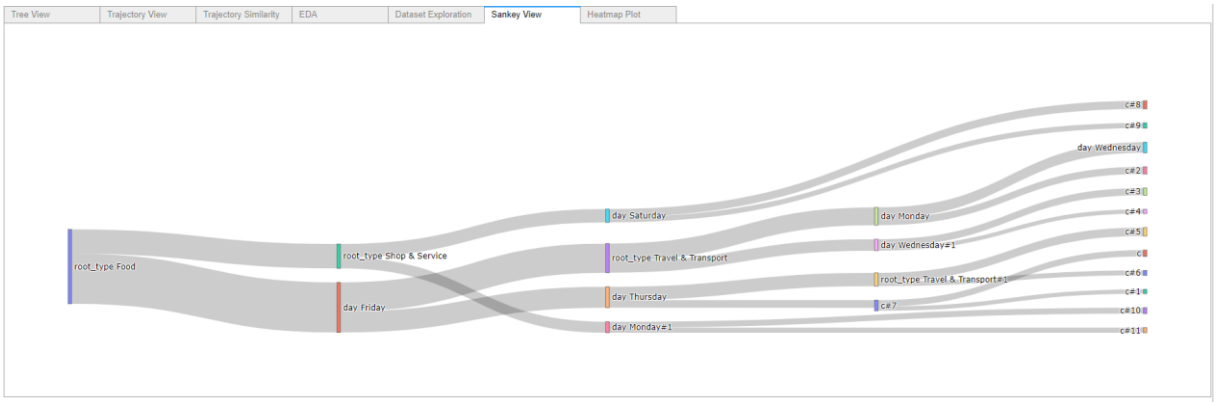


Figura 16. Diagrama de Sankey (aluvial).

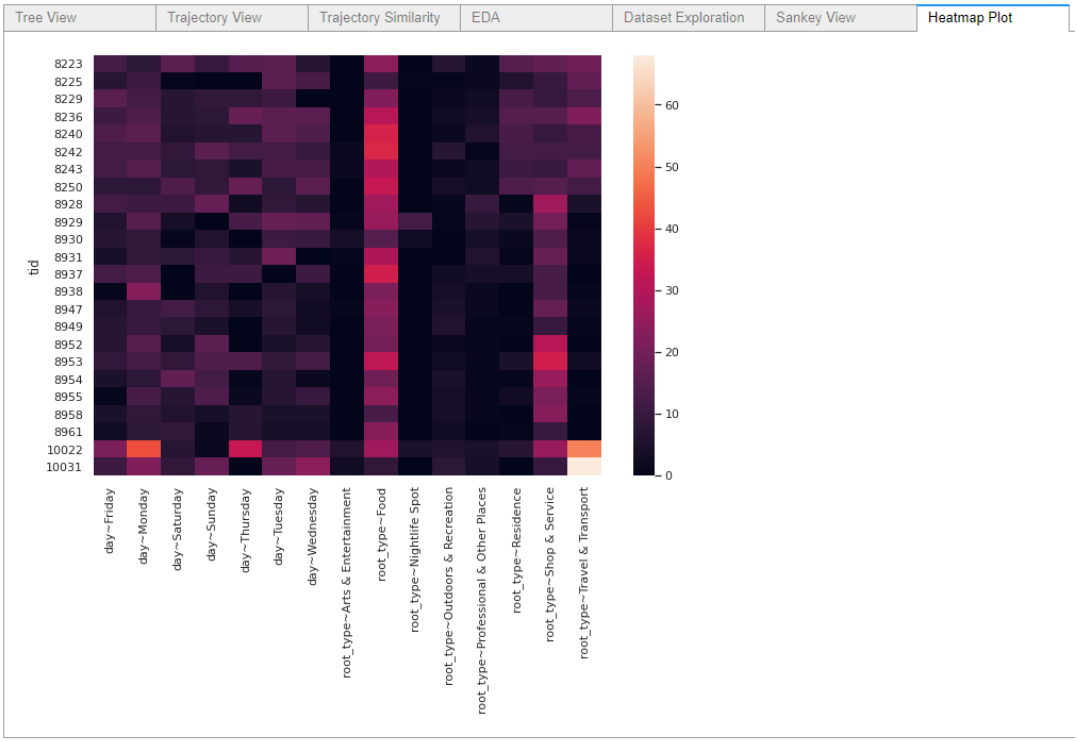


Figura 17. Relação entre as trajetórias dos usuários e os aspectos das trajetórias através de mapas de calor.

4 EXPERIMENTOS E RESULTADOS

Este capítulo apresenta os experimentos realizados para avaliar a modelagem da árvore de decisão para tarefa de clusterização de trajetórias multiaspecto proposto neste trabalho. A seção 4.1 apresenta o *dataset* (conjunto de dados) de trajetórias utilizado no experimento e a seção 4.2 apresenta os resultados obtidos a partir dos experimentos realizados.

4.1 CONJUNTO DE DADOS DE TRAJETÓRIAS MULTIASPECTO

Os dados das trajetórias multiaspecto utilizados nos experimentos foram coletados do *Foursquare*, um aplicativo de dispositivos móveis que provê recomendações personalizadas de lugares próximos à localização dos usuários baseadas nos históricos de busca e de *check-ins*. Um *check-in* é definido como o processo pelo qual as pessoas anunciam sua chegada a um local, como um escritório, hotel, aeroporto, hospital, evento, escola, dentre outros, e refere-se a um ponto de trajetória.

O *dataset* contém 3.079 trajetórias multiaspecto de 193 usuários do aplicativo, contabilizando o total de 66.962 *check-ins*. De maneira geral, o *dataset* apresenta uma média aproximada de 22 *check-ins* por trajetória e 16 trajetórias por usuário. A trajetória de um usuário é definida como o conjunto de pontos ou *check-ins* no período de uma semana.

O *dataset* apresenta 14 atributos que são detalhados a seguir:

- *checkin_id*: refere-se ao identificador do *check-in;*
- *venue_id*: refere-se ao identificador do local onde o usuário efetuou o *check-in*. O *dataset* possui no total 13.848 locais diferentes;
- *tid*: refere-se ao identificador da trajetória;
- *lat_lon*: refere-se às coordenadas geográficas – latitude e longitude – onde o *check-in* foi efetuado;
- *data_time*: representa a data (dia, mês e ano) e a hora exata (hora, segundos e milissegundos) do ponto registrado;

- *time*: representa a quantidade de minutos, em relação ao início do dia, que o ponto foi registrado. Por exemplo, se um ponto foi registrado às 17:32:25, a quantidade de minutos a partir do início do dia (00:00:00) é de 1052 minutos.
- *day*: representa o dia da semana;
- *poi*: representa o ponto de interesse, isto é, o local visitado pelo usuário. O total de POIs registrados no *dataset* é de 10.808. Alguns exemplos de POIs são: Starbucks, George Washington Bridge Bus Station, Yakitori Sun-Chan, NYCT Transit Survey Unit;
- *type*: refere-se ao tipo específico do POI. Considerando os exemplos anteriores citados, os tipos de POI são cafeteria, estação de ônibus, restaurante japonês e escritório, respectivamente. O total de tipos de POIs no *dataset* é de 491;
- *root_type*: refere-se ao tipo genérico do POI. Seguindo os exemplos anteriores, os tipos genéricos são, respectivamente, alimentação, viagem & transporte, alimentação e trabalho & profissões. Existem 10 tipos genéricos de POIs;
- *price:* refere-se à avaliação do POI em relação à média de gasto que um usuário pode ter ao visitá-lo. O valor -1 significa que não há nenhum custo envolvido. Valores entre 1 e 4 representam a escala do gasto, sendo 1 um POI barato e 4 um POI caro;
- *rating:* refere-se à média de avaliação pela comunidade de usuários sobre determinado POI. Os valores variam entre 0 e 10, e o -1 é utilizado para os POIS não avaliados;
- *weather:* refere-se à condição climática no momento em que o ponto é registrado. Existem seis diferentes valores no *dataset*: ensolarado, nublado, chuvoso, nevando, neblina e desconhecido (para os casos em que por algum motivo não foi possível detectar a condição do tempo);
- *label:* refere-se ao rótulo ou identificador dado ao usuário do aplicativo.

4.2 EXPERIMENTOS REALIZADOS

O algoritmo proposto possui a característica da subjetividade, uma vez que cabe ao analista escolher os hiperparâmetros do critério de parada (número mínimo de trajetórias em um nodo-folha e altura máxima da árvore), bem como qual critério de seleção do aspecto será utilizado (divisão binária, mínima variância, redução máxima da variância, redução máxima da variância através da relação entre todos aspectos). Além disso, o analista também pode determinar quais aspectos serão utilizados na clusterização, a depender do domínio do problema que deseja resolver.

A preparação dos dados, que antecede a etapa de modelagem da árvore, seguiu os mesmos critérios utilizados nos trabalhos de Varlamis *et al.,* Petry *et al.* e Yang *et al..* Os registros de informações no conjunto de dados foram pré-processados para remover ruídos, duplicatas e registros incompletos. Para fornecer um *ground-truth* de base supervisionada para avaliação dos agrupamentos, as trajetórias foram separadas em trajetórias semanais para cada usuário e rotuladas com o ID do usuário. Este *ground-truth* assume alta autossemelhança entre as trajetórias semanais de cada usuário e a dissimilaridade com as trajetórias de outros usuários. Trajetórias com menos de 10 *check-ins* e usuários com menos de 10 trajetórias foram removidos do conjunto de dados (VARLAMIS *et al.,* 2021).

Os critérios de parada levaram em conta apenas a quantidade mínima de trajetórias desejadas nos nodos-folhas. Neste caso, o valor escolhido foi de 25. Em relação à altura da árvore, optou-se por não estabelecer um limite a fim de captar todos os possíveis clusters.

Os aspectos das trajetórias utilizados na clusterização foram o *root_type*, *weather* e *day*. A escolha desses aspectos baseou-se no fato destes serem fortemente correlacionados e de característica permanente, isto é, associados a todos os pontos de qualquer trajetória.

A métrica estatística utilizada como critério de divisão das trajetórias em subgrupos menos frequentes e mais frequentes em relação a determinado aspecto escolhido foi a média.

A Tabela 1 elucida a quantidade de grupos totais gerados para cada critério de escolha do aspecto, contando também com a quantidade de nodos-folha e a altura máxima obtida.

Tabela 1. Relação do número de grupos gerados e altura da árvore modelada para cada critério de seleção de aspecto.

Método	Grupos totais	Grupos folha	Altura da Árvore
DB	385	193	8
MV	387	194	14
RMV	393	197	14
MRV	389	195	10

Como se depreende da Tabela 1, os resultados foram parecidos numericamente, apresentando uma média de 388 clusters gerais, sendo uma média de 195 clusters folhas e alturas variando entre 8 e 14. No entanto, as estruturas das árvores geradas são diferentes. Essas diferenças entre os métodos são interessantes, pois podem capturar informações e comportamentos distintos no conjunto de trajetórias. Os dendrogramas ilustrados nas Figuras 18 e 19 exemplificam essas divergências. A Figura 18, a árvore gerada foi modelada pelo critério de escolha de aspecto **MV**, enquanto que a Figura 19, o critério foi o **RMV**. Ao comparar os dois diagramas, nota-se que o primeiro capturou um comportamento atípico das trajetórias, evidenciado no ramo extremo à direita. Na ocasião, o algoritmo selecionou o aspecto '*Event'* do atributo *root_type*, no nodo raiz e agrupou todas as trajetórias dos usuários que efetuaram *check-in* para este tipo de aspecto. Este grupo contou com 20 usuários, do total de 193, somando um total de 23 trajetórias, do total de 3.079 trajetórias. Dentre todos os aspectos presentes no conjunto de dados, '*Event'* representou apenas 26 *check-ins*, no universo de 66962.

As Figuras 20 e 21 mostram uma análise exploratória dos dados do grupo acima mencionado. Os gráficos de barras mostram a frequência **relativa** da ocorrência dos aspectos da categoria *root*_type e *day*, respectivamente, isto é, a frequência de ocorrência dos aspectos deste cluster específico em relação ao conjunto total das trajetórias. Como se pode observar na Figura 20, a frequência relativa para '*Event'* equiparou-se à sua frequência absoluta, ou seja, todas as trajetórias que efetuaram *check-in* relacionado a '*Event*' ficaram em um mesmo grupo.

Acerca dos dias da semana, nota-se, na Figura 21, que sábado foi o dia mais frequente para o grupo em questão. Esse é um comportamento esperado, uma vez que, normalmente, eventos costumam ocorrer aos finais de semana.

Em relação aos demais aspectos, nota-se que houve poucos pontos referentes à '*College & University*', o que pode indicar que a maioria dos usuários agrupados neste cluster não são parte da comunidade universitária. Por outro lado, percebe-se que os usuários deste grupo estão mais relacionados às atividades que envolvem arte e entretenimento e à vida noturna, com os dias de sexta-feira e sábado mais frequentes em relação aos demais.

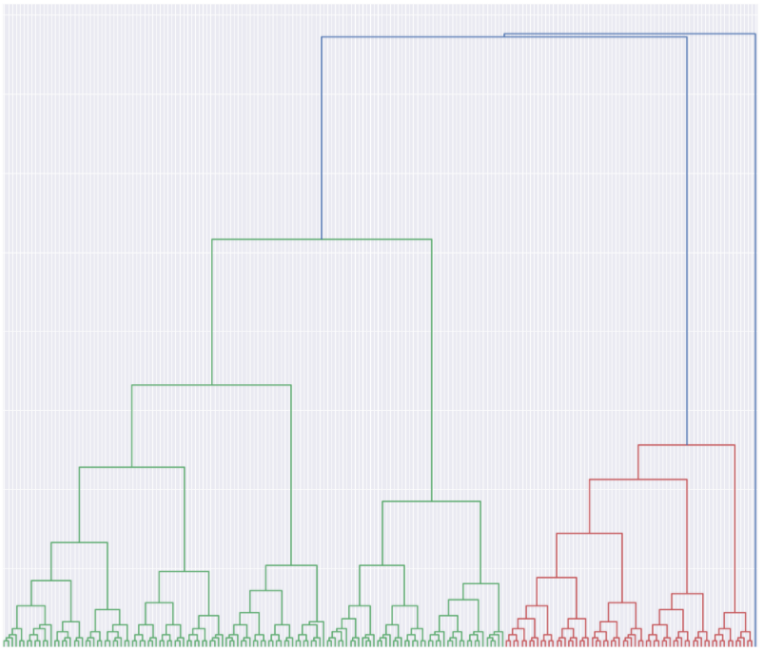


Figura 18. Árvore de clusters gerada pelo critério de Mínima Variância para escolha do aspecto.

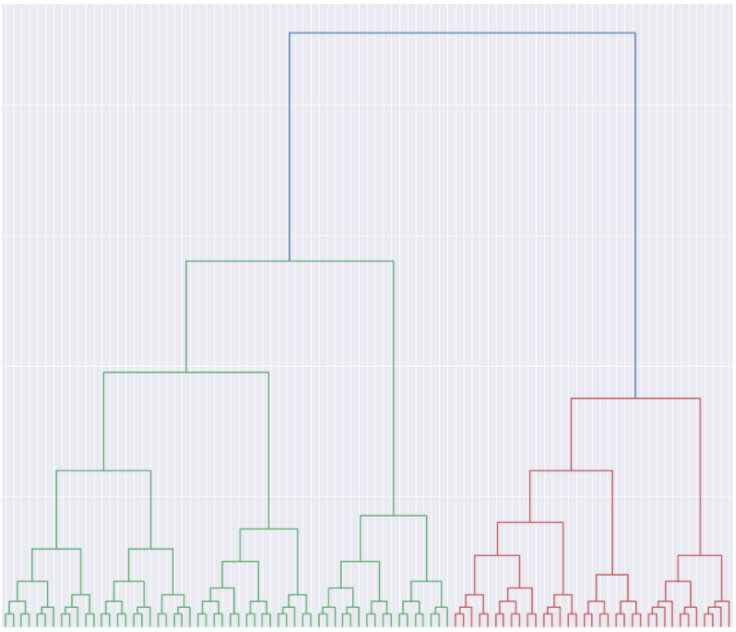


Figura 19. Árvore de clusters gerada pelo critério de Redução Máxima da Variância para escolha do aspecto.

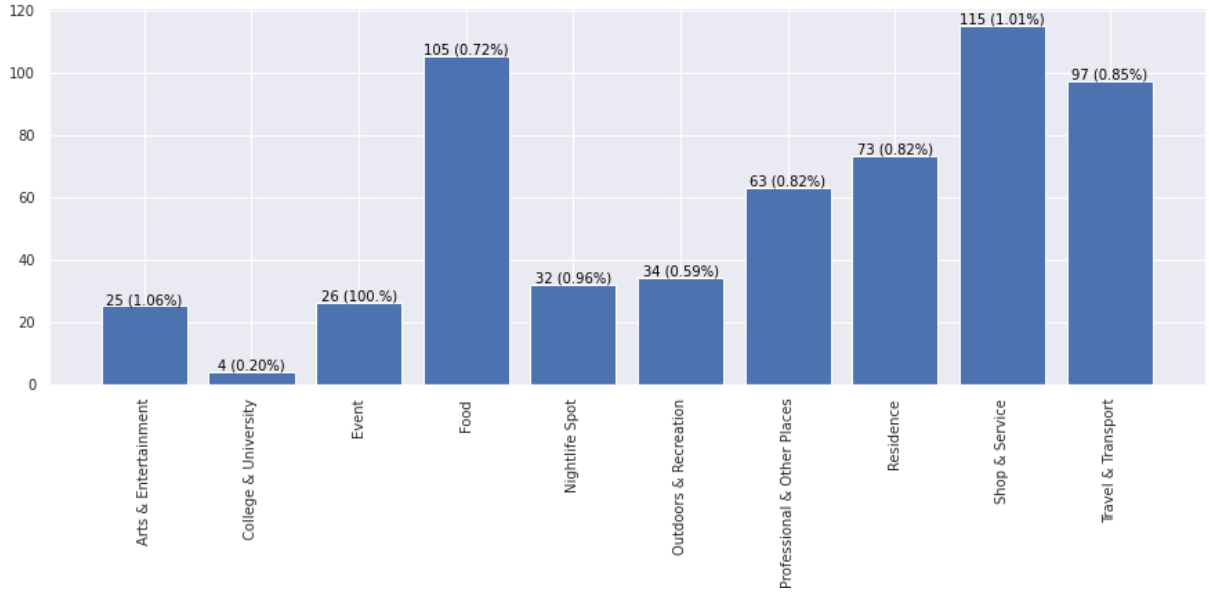


Figura 20. Frequências relativas dos aspectos para categoria *root_type.*

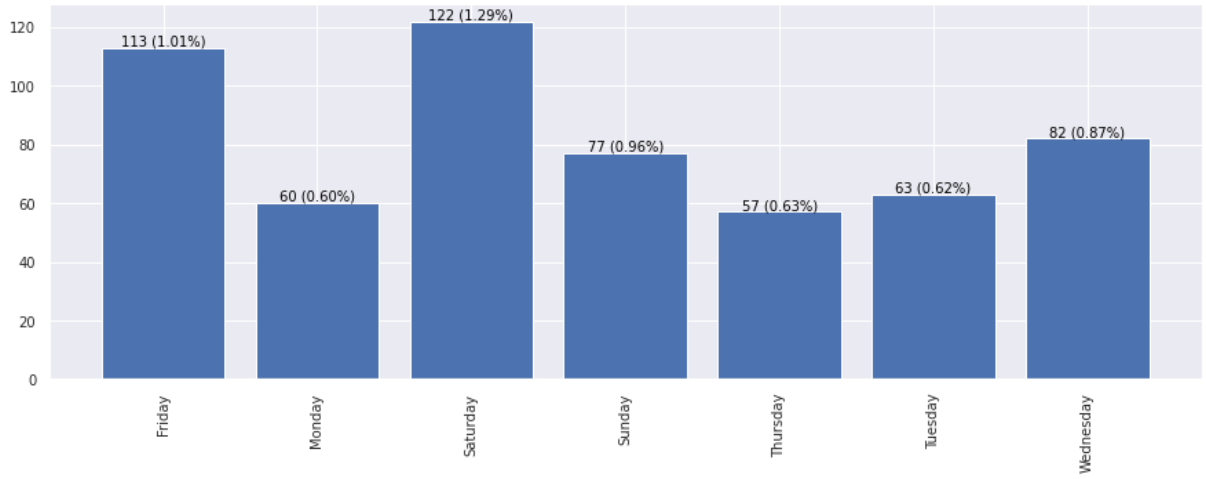


Figura 21. Frequências relativas dos aspectos para categoria *day.*

A etapa posterior à análise dos resultados foi a validação, interna e externa, dos clusters gerados. O intuito da validação foi aferir a qualidade do resultado em termos de coesão e separabilidade das trajetórias agrupadas. O objetivo é agrupar trajetórias com características e comportamentos semelhantes em um mesmo grupo, observando a coesão do grupo, e agrupar as trajetórias menos semelhantes em grupos diferentes, observando-se a separabilidade.

A validação interna foi realizada com as seguintes métricas: *Coeficiente de Silhouette* (CS), *Índice Calisnki-Harabasz* (ICH) e *Índice Davies-Bouldin* (IDB). Tipicamente, essas métricas utilizam em suas fórmulas matemáticas alguma métrica de distância, como a euclidiana ou a Manhattan. No entanto, devido às características dos dados das trajetórias multiaspecto, essas métricas não podem ser utilizadas. A validação foi realizada por meio das métricas de similaridade MSM e MUITAS.

Para cada métrica de similaridade, foi construída uma matriz, relacionando o *score* de similaridade entre todas as trajetórias. O *score* obtido entre duas trajetórias substitui a distância entre elas.

Os resultados obtidos da validação interna para os quatro critérios de escolha de aspecto propostos constam nas Tabelas 2 e 3, para as métricas MUITAS e MSM, respectivamente.

Os valores em destaque em ambas as tabelas evidenciam os melhores resultados obtidos. O critério da redução máxima da variância (RMV), para o *dataset* utilizado, trouxe a melhor clusterização das trajetórias, obtendo o melhor resultado para todas as métricas de validação interna utilizadas.

Tabela 2. Validação interna usando a métrica de similaridade MUITAS.

Método	CS	ICH	IDB
DB	-0,31	14,41	0,37
MV	-0,28	5,20	0,45
RMV	**-0,36**	**23,66**	**0,29**
MRV	-0,34	4,41	0,46

Tabela 3. Validação interna usando a métrica de similaridade MSM.

Método	CS	ICH	IDB
DB	-0,29	12,79	0,48
MV	-0,31	1,64	0,51
RMV	**-0,32**	**18,44**	**0,41**
MRV	-0,31	3,32	0,57

A validação externa consistiu em comparar os resultados da análise dos clusters com um resultado conhecido externamente. Para isto, assumiu-se que as trajetórias de um mesmo usuário provavelmente pertenceriam ao mesmo cluster e, portanto, a avaliação externa da clusterização utilizou o usuário como um rótulo de

referência, ou seja, o *ground-truth*. A decisão da escolha desse *ground-truth* baseou-se no trabalho de González, Hidalgo e Barabási (2008) que demonstraram que as trajetórias humanas possuem um alto grau de regularidade temporal e espacial, em que cada indivíduo é caracterizado por uma distância de viagem característica independente do tempo apresentando uma probabilidade significativa de retornar a alguns locais altamente frequentados. Essa ideia também foi utilizada em outros trabalhos relacionados, como o de Varlamis *et al.*

As métricas para essa avaliação foram: *homogeneity score* (Homogem.), *completeness score* (Complet.), *v-measure score*, *adjusted mutual info score* (Mut. Info), *adjusted rand score* (Rand), e *Fowlkes Mallows score* (F.M).

Os resultados constam na Tabela 4. As métricas MUITAS e MSM apresentaram resultados muito similares e, portanto, foram dispostas na mesma tabela. Os melhores resultados aparecem em destaque, seguidos pelos resultados que aparecem sublinhados.

Dentre os métodos propostos, a mínima variância (MV) e redução máxima da variância (RMV) demonstraram a melhor clusterização.

Tabela 4. Resultado da validação externa.

Método	Homogen.	Complet.	V-measure	Mut. Info	Rand	F.M
DB	0,557	0,556	0,557	0,168	0,065	0,070
MV	**0,630**	0,568	**0,597**	0,122	0,057	0,063
RMV	0,587	**0,581**	0,584	**0,210**	**0,085**	**0,090**
MRV	0,590	0,561	0,575	0,144	0,066	0,071

Considerando as duas formas de validação – interna e externa –, dentre os métodos propostos, a redução máxima da variância (RMV) foi o melhor método de seleção de aspecto.

4.3 AVALIAÇÃO DOS RESULTADOS

Conforme mencionado anteriormente, ainda existe uma carência de trabalhos relacionados à clusterização de trajetórias multiaspecto na literatura. Para avaliar os resultados, buscou-se realizar comparações com o método proposto por Varlamis *et*

al. e com a abordagem tradicional de clusterização aglomerativa hierárquica. Esta última, a métrica de ligação (*linkage affinity*) utilizada foi a distância euclidiana e os critérios de ligação foram: (i) ***ward***, que minimiza a variação dos clusters que estão sendo agrupados; (ii) ***average***, que utiliza a média das distâncias de cada observação dados dois conjuntos; (iii) ***complete***, que usa as distâncias máximas entre todas as observações dados dois conjuntos, e; (iv) ***single***, que usa o mínimo das distâncias entre todas as observações dados dois conjuntos.

O trabalho de Varlamis *et al.* buscou comparar a métrica desenvolvida, o TraFoS, com as métricas MUITAS e MSM, empregando a abordagem de clusterização aglomerativa usando o *single* e o *average linkage*. Os resultados obtidos pelos autores são mostrados na Tabela 5.

Tabela 5. Validação externa (esquerda) e interna (direita) da clusterização conduzida por Varlamis *et al*.

Métrica de Similaridade	**Homogen.**	**Complet.**	**V-measure**	**Mut. Info**	**Rand**	**F.M**	**CS**	**ICH**	**IDB**
Clusterização Aglomerativa usando *Single Linkage*									
MUITAS	0,06	0,58	0,11	0,11	0,00	0,07	-0,19	0,73	1,18
MSM	0,06	0,58	0,11	0,11	0,00	0,07	-0,13	0,70	1,22
TraFoS	**0,08**	**0,72**	**0,14**	**0,27**	**0,00**	**0,08**	**-0,94**	**0,99**	**1,01**
Clusterização Aglomerativa usando *Average Linkage*									
MUITAS	0,08	**0,56**	0,13	0,13	0,00	**0,07**	-0,13	1,21	2,21
MSM	0,07	**0,58**	0,12	0,12	0,00	**0,07**	-0,07	1,63	**1,83**
TraFoS	**0,24**	0,30	**0,23**	**0,23**	0,00	0,02	**-0,95**	**0,96**	3,67

Em relação a Tabela 5, nota-se que no trabalho dos autores citado, a validação externa para as métricas MUITAS e MSM também foram muito similares. Os melhores resultados para a validação interna foram destacados.

De maneira geral, a clusterização aglomerativa usando *single linkage* produziu os melhores resultados.

Quando comparamos a métrica da validação externa **V-Measure**, que reflete a média harmônica entre as pontuações *Completeness* e *Homogeneity*, o melhor valor obtido foi de 0,23 (*Average Linkage*). Considerando o método proposto neste trabalho, para a mesma métrica, o melhor valor obtido foi de 0,597 (MV), o que demonstra que a clusterização das trajetórias multiaspecto por meio de árvores de decisão foi aproximadamente 2,6 vezes mais preciso. Quando comparado com *Single Linkage* (0,14) a clusterização por meio de árvores de decisão é aproximadamente 4,3 vezes mais preciso.

Já em relação à validação interna, o melhor resultado para a métrica **Silhouette** foi de -0,95, enquanto que no método proposto o resultado obtido foi -0,36. Para essa métrica o TraFos apresentou melhor resultado que a clusterização por meio de árvores de decisão, no entanto, ao se analisar as métricas ICH e IDB, o TraFos apresentou, respectivamente, 0,99 e 1,01, enquanto que o método por árvores de decisão apresentou, respectivamente, 23,66 e 0,29, mostrando resultados melhores. Para o ICH, quanto maior o valor, melhor é a clusterização e, para o IDB, quanto menor o valor, melhor é a clusterização.

Importante destacar que os valores negativos obtidos em ambos os trabalhos se devem ao uso de matrizes de similaridade como cálculo das distâncias entre as trajetórias, e não as distâncias entre elas.

A Tabela 6 mostra os resultados com a clusterização aglomerativa hierárquica descrita anteriormente, o qual utilizou-se a matriz de frequência dos aspectos como *dataset*.

Tabela 6. Resultados da validação externa (esquerda) e interna (direita) da clusterização com abordagem aglomerativa hierárquica.

Critério Ligação	Homogen.	Complet.	V-measure	Mut. Info	Rand	F.M	CS	ICH	IDB
Single	0,060	0,570	0,115	-0,002	0,000	0,070	**0,20**	12,34	**0,34**
Ward	**0,507**	0,556	**0,530**	**0,206**	**0,061**	**0,073**	0,05	**54,07**	1,57
Average	0,171	**0,580**	0,264	0,059	0,001	0,063	0,12	26,33	0,85
Complete	0,323	0,525	0,400	0,155	0,008	0,049	0,05	41,26	1,18

O critério de ligação *Ward* apresentou os melhores valores, de maneira geral, seguindo pelo critério *Complete*.

Considerando as mesmas métricas utilizadas para comparar os resultados do método proposto com a clusterização aglomerativa de Varlamis *et al.* – V-measure e Silhouette – os melhores valores foram 0,530 e 0,18, respectivamente. A clusterização por meio de árvores de decisão foi aproximadamente 12% mais precisa, em relação a validação externa, e gerou clusters 80% mais coesos e homogêneos.

4.4 ANÁLISE DOS RESULTADOS

A interpretação dos resultados obtidos na clusterização das trajetórias também possui um caráter subjetivo, uma vez que as análises devem ir ao encontro dos problemas os quais o analista deseja resolver.

Um exemplo de análise dos resultados pode ser visto na Figura 22. A Figura ilustra um trecho da clusterização, que foi simplificado para melhor visualização.

Na análise qualitativa dos resultados obtidos no *dataset Foursquare* foi possível observar que em dias chuvosos e em relação a atividades de lazer, os usuários tendem a frequentar lugares referentes a artes e entretenimento, tais como museus e cinemas ao invés de *nightclubs*. Da Figura 22, nota-se que o cluster (representado pela barra em preto) mais à esquerda teve suas trajetórias divididas pelo aspecto *Rain* (chuva). Do cluster em questão, saem duas colunas, com diferentes espessuras que representam o quantitativo de trajetórias que foram separados em face à métrica estatística utilizada – média –, a barra superior representa as trajetórias mais frequentes para o aspecto *Rain*, e a barra inferior as trajetórias menos frequentes.

A seguir, notou-se que o aspecto mais relevante para as trajetórias com frequências maiores que a média para *Rain* foi a categoria de POI *Arts & Entertainment* e o aspecto mais relevante para as trajetórias menos frequentes foi *Nightlife Spot,* conforme mostram os clusters localizados no centro do diagrama.

A análise exploratória dos dados em relação aos aspectos selecionados mostrou que *Arts & Entertainment*, além de ser mais frequente para o aspecto chuvoso, é mais frequente aos domingos, independentemente da condição climática, enquanto que em relação a *nightclubs*, o dia mais relevante é a quinta-feira. Ainda em relação a artes e entretenimento, notou-se que os usuários preferem atividades recreativas ao ar livre a atividades em locais fechados.

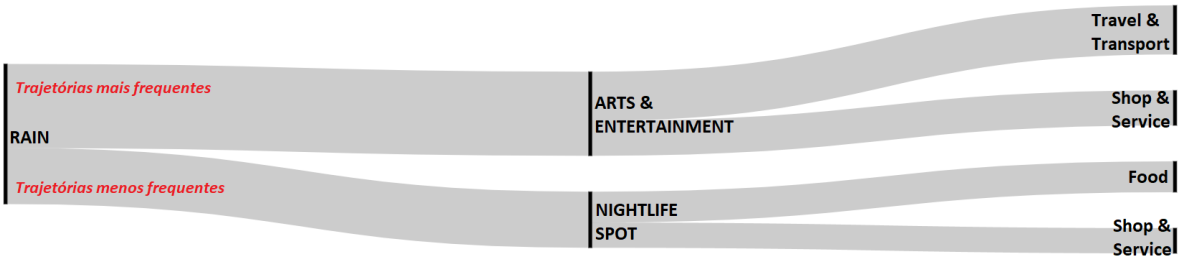


Figura 22. Análise qualitativa da clusterização

Outro comportamento interessante encontrado das análises realizadas foi que quarta-feira foi o dia mais frequente em que os usuários do aplicativo saíram para algum restaurante e para fazerem compras em *malls & services*. Nos finais de semana, os usuários tendem a fazer apenas um dos dois, isto é, ou sair para algum restaurante ou para fazer compras. Também se notou que os aspectos *travel & transport* e *food* foram os aspectos mais fortemente correlacionados, independentemente do dia da semana.

Conforme mencionado em tópico anterior, a clusterização resultou em diversos grupos com trajetórias similares, o que permite fazer inúmeras análises dos resultados. Esses foram apenas alguns dos exemplos em que se pôde encontrar algum comportamento presente no conjunto das trajetórias multiaspecto.

5 CONCLUSÃO

Com o enorme volume de dados e o crescente poder computacional disponível atualmente, torna-se mais evidente a necessidade de novos métodos de análise de dados complexos. Trajetórias de múltiplos aspectos trazem muitas oportunidades em relação à mineração de dados, mas a natureza da sequência, a alta dimensionalidade, heterogeneidade e volume de dados também apresentam novos desafios.

Embora vários métodos tenham sido propostos para a classificação de trajetórias multiaspecto (TMA), apenas alguns trabalhos se concentram na clusterização de TMA. Neste trabalho, foi proposto um novo método para agrupamento de trajetórias de múltiplos aspectos usando uma abordagem baseada em árvore de decisão. A principal contribuição deste algoritmo é uma nova abordagem para clusterização de trajetórias considerando todas as suas dimensões, uma vez que os estudos sobre a dimensão semântica das trajetórias são bastante recentes e limitados. Os resultados experimentais mostraram um desempenho superior quando comparado a um método de agrupamento de última geração para TMA, denotando que o novo método é capaz de gerar agrupamentos mais coesos, compactos e conectados.

As diferentes opções de clusterização e visualizações fornecem uma ferramenta flexível para análise exploratória de dados e aplicações que se adaptam facilmente à tarefa em questão. Como o agrupamento é uma tarefa de mineração de dados que é inerente e altamente dependente da aplicação e exploratória, a característica flexível do método proposto é uma característica importante.

Como trabalho futuro, planeja-se fornecer uma investigação detalhada sobre o impacto dos diferentes critérios de divisão e a avaliação. Esses critérios são uma direção promissora, uma vez que cada um dos critérios parece ser mais adequado para diferentes aplicações e análises. Por exemplo, com base na análise qualitativa vemos que enquanto o critério de avaliação DB é bem adequado para gerar clusters que são mais ou menos do mesmo tamanho, o critério VM parece encontrar facilmente trajetórias discrepantes, ou seja, trajetórias que apresentam aspectos que raramente aparecem em outras trajetórias.

REFERÊNCIAS

ALVARES, L.O *et. al*. **A model for enriching trajectories with semantic geographical information.** Proceedings of the 15th acm international symposium on advances in geographic information systems (acm gis 2007). 162-169 p, USA, 2007.

BERRY, M.J.A.; LINOFF, G. **Data Mining Techniques For Marketing, Sales and Customer Support**. John Wiley & Sons, Inc., USA, 1996.

BOGORNY, V; BRAZ, F. J.. **Introdução a trajetórias de Objetos Móveis:** conceitos, armazenamento e análises de dados. 1. ed. Joinville: Univille, v. 500. 116p, 2012.

BRAMER, M. **Principles of Data Mining.** 3. ed. Springer. 2016.

CASTIN, L., FRÉNAY, B.. **Clustering with decision trees:** divisive and agglomerative approach. ESANN, 2018.

CHEN, J.Y, *et. al..* **Sketch-based uncertain trajectories clustering.** Proceedings of the 9th international conference on fuzzy systems and knowledge discovery, pp 747–751, 2012.

CHEN, L.; ÖZSU, M.T.; ORIA, V.. **Robust and fast similarity Search for moving object trajectories.** Proceedings of the 2005 ACM SIGMOD international conference on Management of data. ACM, 491–502, 2005.

CHEN, M., HAN, J., YU, P.S.. **Data Mining:** na overview from a database perspective. IEEE Transactions on knowledge and data engineering, Vol. 8, N. 6, 1996.

FAYYAD, U., PIATETSKY-SHAPIRO, G., SMYTH, P.. **From data mining to knowledge discovery in databases.** AI Magazine, Volume 17, Number 3, 1996.

FOWLKES, E.B., MALLOWS, C.L.. **A Method for Comparing Two Hierarchical Clusterings.** Journal of the American Statistical Association, Volume 78, Number 383, pp: 553-569, 1983.

FRAWLEY, W.J., PIATETSKY-SHAPIRO, G., MATHEUS, C.J.. **Knowledge Discovery in Databases:** An overview. AI Magazine, Volume 13, Number 3, 1992.

FURTADO, A.S. *et. al..* **Unveiling movement uncertainty for robust trajectory similarity analysis.** International Journal of Geographical Information Science, 2017.

GHEWARE, S.D., KEJKAR, A.S., TONDARE, S.M.. **Data mining:** task, tools, techniques and applications. International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 10, 2014.

GONZÁLEZ, M.C, HIDALGO, C.A, BARABÁSI, A.. **Understanding individual human mobility patterns.** Nature, Vol. 453, Issue 7196, pp: 779-782, 2008.

HALKIDI, M.; BATISTAKIS, Y.; VAZIRGIANNIS, M.. **On clustering validation techniques.** Journal of Intelligent Information Systems, 17:2/3, 107–145, 2001

HUNG, C.C.; PENG, W.C; LEE, W.C. **Clustering and aggregating clues of trajectories for mining trajectory patterns and routes.** VLDB J 24(2):169–192, 2015.

IRANI, J.; PISE, N.; PHATAK, M.. **Clustering techniques and the similarity measures used in clustering:** a survey. International Journal of Computer Applications (0975 – 8887) Vol. 134 – No.7, 2016.

JAIN, A.K; MURTY, M.N; FLYNN, P.J.. **Data Clustering:** A Review. ACM Computing Surveys, Vol. 31, No. 3, 1999.

JEUNG, HY et. al. **Discovery of convoys in trajectory databases.** Proceedings of the 34th international conference on very large databases, pp 1068–1080, 2008.

LAUBE, P; IMFELD, S. Analyzing relative motion within groups of trackable moving point objects. GIScience, 132-144, 2002.

LEHMANN, A.L.; ALVARES, L.O.; BOGORNY, V.. **SMSM:** a similarity measure for trajectory stops and moves. International Journal of Geographical Information Science, 2019.

LIU, B., XIA, Y., YU, P.S. **Clustering via decision tree construction.** Foundations and Advances in Data Mining. Studies in Fuzziness and Soft Computing, vol 180. Springer, Berlin, Heidelberg, 2005.

MADHULATHA, T.S.. **Na overview on clustering methods.** IOSR Journal of Engineering, Vol. 2(4) pp: 719-725, 2012.

MASCIARI, E.. **A framework for trajectory clustering.** Lecture notes in computer science, vol 5659, pp 102–111, 2009.

MELLO, R. S. *et al..* **MASTER: a multiple aspects view on trajectories.** Transactions in GIS, 23:805–822, 2019.

NANNI, M; PEDRESCHI, D. **Time-focused clustering of trajectories of moving objects.** J Intell Inf Syst 27(3):267–289, 2006.

PELEKIS, N. *et.al.* **Clustering trajectories of moving objects in a uncertain world.** Ninth IEEE International Conference on Data Mining, 2009.

PETRY, L.M. *et. al..* **Towards semantic-aware multiple-aspect trajectory similarity measuring.** Transactions in GIS 23, 5 (2019), 960–975, 2019.

ROSENBERG, A.; HIRSCHBERG, J. **V-Measure: A conditional entropy-based external cluster evaluation measure.** Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 410–420, Prague, June 2007

SPACCAPIETRA, S. *et al..* **A conceptual view on trajectories.** Data & Knowledge Engineering. Vol. 65, No. 1, 2008.

SHARMA, H.; KUMAR, S.. **A survey on decision tree algorithms of classification in data mining.** International Journal of Science and Research (IJSR), Volume 5 Issue 4, 2016.

THEODORIDIS, S.; KOUTROUBAS, K.. **Pattern Recognition.** Capítulo 16. 4ª edição. Elsevier. 2009.

VARLAMIS, I. *et al..* **A novel similarity measure for multiple aspect trajectory clustering.** SAC '21: Proceedings of the 36th Annual ACM Symposium on Applied Computing, pp. 551-558, 2021.

VINH, N.X.; EPPS, J., BAILEY, J.. **Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance.** Journal of Machine Learning Research 11, 2010.

VLACHOS, M.; KOLLIOS, G.; GUNOPULOS, D.. **Discovering similar multidimensional trajectories.** Proceedings of the 18th International Conference on Data Engineering, 2002.

YANG, D, ZHANG, D., ZHENG, V. W., YU, Z.. **Modeling User Activity Preference by Leveraging User Spatial Temporal Characteristics in LBSNs.** *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 45, no. 1, pp. 129-142, 2015.

YASODHA, M; PONMUTHURAMALINGAM, D.R.P. **A survey on temporal data clustering.** Int J Adv Res Comput Commun Eng 1(9):772–786, 2012.

APÊNDICES

APÊNDICE A – CÓDIGO DESENVOLVIDO

O desenvolvimento do algoritmo de clusterização baseado em árvore de decisão foi realizado no ambiente Google Colaboratory – Colab –, que hospeda o serviço Jupyter Notebook, uma plataforma interativa da linguagem de programação Python.

O código encontra-se no repositório Github sob o nome MAT-Tree e pode ser acessado através do seguinte link: https://github.com/ricardogiuliani/MAT-Tree.

APÊNDICE B – ARTIGO

MAT-Trees: A tree-based method for Multiple Aspect Trajectory Clustering

Ricardo Giuliani[1], **Vania Bogorny**[1], **Jônata Tyska Carvalho**[1]

[1]Department of Informatics and Statistics – Federal University of Santa Catarina (UFSC)
Mailbox 476 – 88.040-900 – Florianópolis – SC – Brazil

ricardo.giuliani@grad.ufsc.br, vania.bogorny@ufsc.br, jonata.tyska@ufsc.br

Abstract. *Multiple aspect trajectory is a relevant concept that enables mining interesting patterns and behaviors of moving objects for different applications. This new way of looking at trajectories includes a semantic dimension, which presents the notion of aspects that are relevant facts of the real world that add more meaning to spatio-temporal data. Given the inherent complexity of this new type of data, the development of new data mining methods are needed. In this paper, we propose a hierarchical clustering algorithm for multiple aspect trajectories using a decision tree structure that chooses the best aspect to branch and group the most similar trajectories according to different criteria. This form of grouping revealed a formation of more cohesive and homogeneous groups in almost 89% compared to the base method and 5 times more precise according to an external metric, denoting better data clusters.*

1. Introduction

Mining hidden and useful patterns from data is an activity that has been strongly performed and researched for at least three decades [Fayyad et al. 1996]. In the era of big data, the capability of discovering these patterns in complex data types, has enabled the construction of many interesting applications. The development of techniques for analyzing trajectory data is growing rapidly [Zheng et al. 2013, Tortelli et al. 2022, Sadahiro et al. 1996, Wu et al. 2021], leveraged by the popularization of GPS-equipped devices which is allowing the collection of spatial data from moving objects in unprecedented volumes. Moreover, the advancement of the Internet of Things has allowed the extraction of numerous information beyond the data of spatiotemporal dimensions, which are called aspects and which contribute to the enrichment of the semantic dimension of mobility data. This type of trajectory is called Multiple Aspect Trajectory (MAT) [Mello et al. 2019].

Regarding data mining of mobility data, several techniques have been proposed for raw trajectory Classification and Clustering. Some of the classification tasks involve predicting the transportation modes of the moving objects, animal categories, the type of a vessel, and so on [Silva et al. 2019] Some interesting applications of this type can be found in the works presented by [Dabiri and Heaslip 2018] that take advantage of Convolutional Neural Networks (CNNs) architectures to predict transportation modes based on only raw GPS trajectories, and by [Etemad et al. 2018] that predicts transportation modes of GPS trajectories using feature engineering and noise removal by extracting global and local trajectory features from all pairs of trajectory points. Regarding moving object

clustering, the algorithms explore the spatial dimension, as in [Hung et al. 2015] that proposed a framework for clustering and aggregating clues of trajectories for finding routes; or explore the temporal dimension, as in [Nanni and Pedreschi 2006] that proposes an adaptation of a density-based clustering algorithm to trajectory data, exploiting the temporal dimension to improve the quality of trajectory clustering. These works present techniques that provide very interesting results when analyzing spatiotemporal data, but they cannot be used when the semantic dimension is taken into account. Techniques for analyzing multiple aspect trajectories require the capability of dealing with heterogeneous data and an even bigger data volumes due to the additional semantic data.

In the last few years, many works are focusing on the challenges of mining patterns from multi-aspect trajectory data. Among the works related to classification tasks we find HiPerMovelets [Tortelli et al. 2022], a method for extracting movelets [Ferrero et al. 2018] which consists of finding the best, i.e. most discriminative, dimension combination and subtrajectory size for trajectory classification, and MARC which is an approach for classifying multiple-aspect trajectories based on attribute embedding and Recurrent Neural Networks (RNNs) that addresses all trajectory properties: space, time, semantics, and sequence. In relation to clustering, [Varlamis et al. 2021] proposes TraFos, a similarity measure for comparing MATs, including a multi-vector representation of MATs that enables performing cluster analysis on them. In [Santos et al. 2021] SS-OCoClus is proposed, a co-clustering method for mining semantic trajectories that consists of solving the problem of discovering groupings of objects without having to evaluate all of their attributes. While the results achieved so far by semantic trajectories classification methods are very solid, the few works focusing on MAT clustering showed that there are plenty of space for new unsupervised techniques capable of grouping and describing multiple semantic aspects together.

The few methods proposed in the literature for MAT clustering [Varlamis et al. 2021, Santos et al. 2021], i.e., works that consider all three dimensions of space, time and semantics, deal with only one aspect at a time. As they do not deal with diverse and heterogeneous characteristics, most works use only a single representation for trajectory similarity and cluster analysis.

Regarding MAT clustering, these methods are important in order to build applications based on the capability of finding groups of moving objects that are similar, i.e. move together, visit similar places, behave similarly, perform similar activities and so on. Overall, these clusters can provide insights into the relationships between the objects and the trajectories, allowing pattern extraction and detection of similar or anomalous behavior of moving objects [Varlamis et al. 2021].

In this paper we propose a divisive hierarchical clustering algorithm that considers the three dimensions of trajectories using a decision tree-based approach. Based on the frequency that each of the aspects appear in the the trajectories of a given dataset, the algorithm seeks to iteratively select the ideal aspect to divide the set of trajectories in two, using a threshold that could be the mean or median of this aspect. This leads to the formation of hierarchical clusters of trajectories that are naturally more similar to each other, i. e., clusters that present a higher frequency for certain aspects and lower frequency for others. It is noteworhty that since the proposed method uses a tree approach, there is a hierarchy between the formed clusters. Therefore, in the bottom of the tree, at the

leaf nodes, we have more specific and detailed clusters, while closer to the top of the tree, clusters have more general and generic information, being the root node the original dataset.

With the proposed method we seek to provide analysts the capacity to answer questions as: ”in which aspects the moving objects are more similar?”

The proposed method was validated using a Foursquare dataset, a mobile app that allows users to check in to the places they visit. The dataset contains user trajectories and, in addition to geographic location and time, includes a series of related characteristics such as the type of place visited, rating, price tier and weather condition. A series of Foursquare were performed comparing the proposed method with the state-of-the-art method for MAT clustering. For validating the proposed method, we have used internal and external validation metrics. The results obtained with the experiments showed that the new method is quantitatively and qualitatively better than the previous state-of-the-art method [Varlamis et al. 2021]. The clusters generated by our method were approximately 89% more cohesive and more separable than the base method. Also, regarding the external metric used, the proposed method was five times more precise according to external validation metrics. From the qualitative point of view, the method provide different options of clustering and visualizations, being a valuable tool for data analysts performing exploratory analysis on multi-aspect trajectory data.

The rest of this paper is organized as follows. In Section 2, we present the basics of defining multiple aspect trajectories and related works. In Section 3 we describe the proposed approach to clustering MAT by decision trees. In Section 4 we discuss the experimental results and evaluation of the proposed method and in Section 5 we bring the conclusions of our work.

2. Basic Concepts and Related Works

In this section we provide a brief description of the definition of MAT and a brief review of MAT clustering approaches.

2.1. Multiple Aspect Trajectory

Multiple aspect trajectories are defined by their three-dimensional nature, i.e., the sequences of points composed by space and time, in addition to the semantic dimension. The concept of semantic dimension is the representation of any context information or relevant meanings that are of fundamental importance for understanding the data obtained in a trajectory [Spaccapietra et al. 2008]. The first approach that brought semantic data enrichment to trajetory data was the one of stops and moves [Alvares et al. 2007a]. Moves are made of sample locations between stops, that could also be at the beginning or at the end of a trajectory, while stops are groups of sample points that are close in space and time and that reflect interesting spatial places known as Points of Interest (POI). Every stop has a beginning and ending time, as well as a spatial position and a minimum duration.

More recently, the semantic dimension started to represent the vast set of characteristics that each point of a trajectory can present, which is called aspect, thus bringing the idea of multiple aspect trajectories [Mello et al. 2019]. Aspect can be described as a real world fact relevant to the analysis of moving object data. Thus, a point on a trajectory can contain several aspects, as exemplified in Figure 1, which shows the trajectory and

POIs, represented in circles, of a given person, such as weather information, heart rate, emotional status while working at the office, the ticket price, the genre of the film and its rating in a cinema session and a restaurant with aspects representing its reviews, opening hours, price range and the restaurant type. As can be seen, the trajectories can present numerous aspects that enrich the semantic dimension.

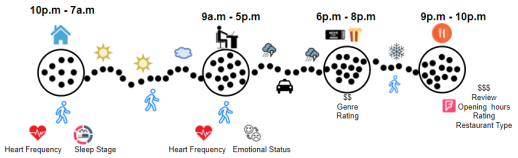


Figure 1. Example of a Multiple Aspect Trajectory.

2.2. Multiple Aspect Trajectory Clustering

Trajectory clustering is a frequently used unsupervised data analysis approach for grouping trajectories based on a similarity function. The main idea is to group trajectories that have similar characteristics and behaviors in the same clusters while keeping dissimilar trajectories in different clusters.

The vast majority of trajectory clustering algorithms in the literature take into account only the spatiotemporal dimensions, which employ classical clustering approaches based on distance or density on a similarity metric [Yuan et al. 2017], [Yao et al. 2017], such as Euclidean distance, Manhattan distance and so on.

Regarding the semantic dimension, some works can be cited as the SMoT algorithm [Alvares et al. 2007a], which proposes the extraction of stops and moves from trajectories. An algorithm for the extraction of patterns present in moving objects is proposed in [Alvares et al. 2007b] and in [Ying et al. 2011] a prediction model is proposed based on a novel cluster-based prediction strategy which evaluates the next location of a mobile user based on the frequent behaviors of similar users in the same cluster, determined by analyzing users' common behavior in semantic trajectories.

Despite it is possible to find in the literature some works investigating how to cluster spatiotemporal trajectories, there is a lack of methods enabling cluster analysis of trajectories using the semantic dimension and its multiple aspects.

Since similarity metrics are a core concept for clustering techniques, the works related to multiple aspect trajectory clustering are also related to proposals of similarity metrics. In this context, the main similarity metrics for MAT are MSM [Furtado et al. 2016], MUITAS [Petry et al. 2016] and TraFos [Varlamis et al. 2021]. The Multidimensional similarity measure (MSM) analyses each feature individually and supports various weights for each aspect, which assign greater or lesser priority to each aspect depending on the application requirements. On the other hand, the MUltIple-aspect TrAjectory Similarity (MUITAS) can aggregate multiple attributes considering the relationship between those attributes in a trajectory. Lastly, the Trajectory Forest Similarity (TraFoS) proposes a new vector representation model for semantic trajectories that captures the frequency of semantic information across the trajectory points and allows fast similarity comparisons.

3. Proposed Approach

The main idea of the proposed algorithm is the development of a hierarchical tree in which an aspect among the set of all aspects present in the trajectories dataset is selected to split the total dataset into two subsets of trajectories, called nodes. This process is repeated for each new node until the resulting clusters (nodes) present a small amount of trajectories, called leaf nodes. Figure 2 illustrates an example of a tree generated by the clustering algorithm using Sankey diagram representation. The vertical bars indicate the clusters generated by the division through the chosen aspect, noting that the leftmost bar represents the complete dataset with all trajectories and the rightmost bars represent the leaf nodes.

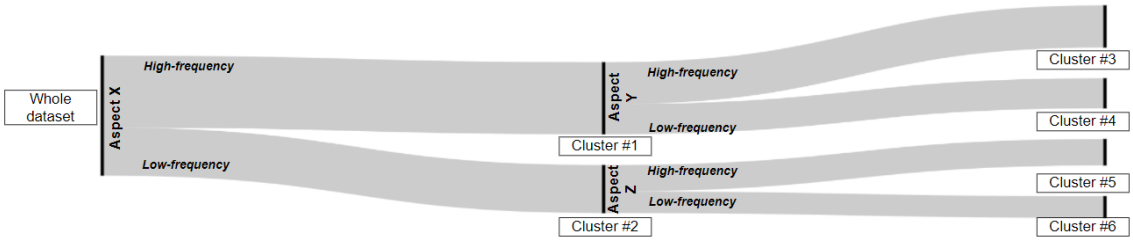


Figure 2. Example of a multi-aspect trajectory clustering decision tree.

The whole process is described on Algorithm 1. The main parts and characteristics of the algorithm are: (i) the construction of the frequency matrix; (ii) the criterion chosen for splitting the dataset at each node/iteration; (iii) the criterion chosen for evaluating the quality of the split tested; (iv) the stop criterion.

The algorithm receives as input a dataset D of the multi-aspect trajectories, the statistical metric that will be used for dividing the trajectories *stat_met*, the criterion for choosing the aspect *asp_criterion*, described later, the maximum height desired for the tree *max_tl* and the minimum number of trajectories targeted for a leaf node *mt*.

The first node of the tree, called the root, is responsible for calling the MAT_TREE function (line 39). This function takes as arguments the set of trajectories, the set of multiple aspects and the height of the tree.

We extract the trajectories and aspects of dataset D into data structures, such as lists or arrays (lines 37 and 38) which the root node will use as parameters to initialize the MAT_TREE function.

MAT_TREE is a recursive function that starts its execution by creating the tree's node structure (line 2) that will hold the trajectories received as a parameter of the function (line 3). Next, child nodes are created, both initialized empty, called left_node (line 4) and right_node (line 5), which may contain the trajectories resulting from the division process.

The construction of the tree is based on the creation of frequency matrices of occurrences of the multi-aspects of each of the trajectories, meaning that for each new node generated in the tree, a frequency matrix of each aspect is created for each trajectory belonging to this node (line 8). Figure 3 depicts an example of a frequency matrix, showing the occurrence of four aspects of three MATs.

Initially, it is necessary to define which criterion will be used to choose the aspect

Algorithm 1: Proposed tree-based clustering algorithm

Input: Trajectory dataset D, Aspect selection criterion asp_criterion, Statistical metric stat_met, Max Tree Level max_tl, Min trajectories per node mt

Output: Cluster Tree

Function MAT_TREE $(trajectories, aspects, tree\_level)$ **:**
  $node \leftarrow New\ TreeNode$
  $node.data \leftarrow trajectories$
  $node.left\_node \leftarrow NULL$
  $node.right\_node \leftarrow NULL$
  $freq\_matrix \leftarrow CALC\_ASPECT\_FREQUENCY(node.data)$
  aspect_threshold $\leftarrow$ [ ]
  foreach *aspect* $a_i \in freq\_matrix$ **do**
    $asp \leftarrow CALC\_ASPECT\_STATISTICS(a_i, stat\_met)$
    $aspect\_threshold.add(asp)$
  end
  $left\_group \leftarrow$ [ ][ ]
  $right\_group \leftarrow$ [ ][ ]
  $aspect\_chosen \leftarrow \emptyset$
  foreach *aspect* $a_i \in freq\_matrix$ **do**
    foreach *trajectory* $t_i \in trajectories$ **do**
      if *freq_matrix*$[t_i] <$ *aspect_threshold*$[a_i]$ **then**
        left_group$[a_i][t_i]$
      else
        right_group$[a_i][t_i]$
      end
    end
  end
  $aspect\_chosen \leftarrow ASP\_SELECTION(asp\_criterion, aspect, left\_group, right\_group)$
  $tree\_level \leftarrow tree\_level + 1$
  if *size_of(left_group)* $<$ *mt* $\vee$ *tree_level* $>$ *max_tl* **then**
    $root.left\_node \leftarrow NULL$
  else
    $root.left\_node \leftarrow MAT\_TREE(left\_group, aspects - aspect\_chosen, tree\_level)$
  end
  if *size_of(right_group)* $<$ *mt* $\vee$ *tree_level* $>$ *max_tl* **then**
    $root.right\_node \leftarrow NULL$
  else
    $root.right\_node \leftarrow MAT\_TREE(right\_group, aspects - aspect\_chosen, tree\_level)$
  end
  return node
$trajectories \leftarrow D.trajectories$
$aspects \leftarrow D.aspects$
$root \leftarrow MAT\_TREE(trajectories, aspects, 0)$
return root

that will be used for the partitioning of trajectories and how this partitioning will occur.

The way in which the partitioning takes place is determined using the statisti-

cal metric *stat_met* passed as an argument to the function, such as the average; in this scenario, it is calculated the average occurrence of each aspect in the set of trajectories (lines 9-13). Thus, after some aspect has been chosen for the division, trajectories that present a frequency higher than the *stat_met* of this aspect will be grouped in a cluster, i.e. *node.right_node*, while the trajectories that present a frequency lower than the *stat_met* will be grouped in another cluster, *node.left_node* (lines 26-35). Here, other approaches can be employed, such as using the median instead of the average.

traj_id	Aspect A	Aspect B	Aspect C	Aspect D
tid_01	3	0	12	10
tid_02	14	12	14	11
tid_03	12	3	17	12

Figure 3. Frequency matrix of four aspects from three MAT.

We propose four evaluation criteria for the selection of the aspect, namely: (i) binary division (BD), (ii) minimum variance (MV), (iii) maximum variance reduction (MVR) and (iv) maximum variance reduction considering the largest reduction average among all the aspects present in the trajectories (LRA). The algorithm will test the partitioning of trajectories for all aspects and according to the chosen evaluation criterion it will decide which aspect will be chosen (line 24). Below we provide details about each of the proposed criterion.

BD will analyze the absolute difference in the amount of trajectories between the subsets formed after partitioning. The chosen aspect will be the one that generates the smallest difference value.

MV evaluates the mean variance of the aspect in the two subsets formed, selecting the one with the smallest mean variance.

MVR also considers the mean variance of the aspect in the two subsets formed, however, here we seek to evaluate the variance reduction considering its initial value, that is, the variance of the parent node. The aspect chosen is the one that generated the greatest value.

LRA is calculated in a similar way to the previous one, however, instead of considering the reduction of the aspects individually, this approach considers the average reduction of all other aspects, selecting the one that generated the greatest mean of variance reduction.

The criteria involving the variance, it is important to highlight that the variance represents the dispersion of the distribution of trajectories according to their aspects, that is, how far from the mean the trajectories are. With a smaller variance, we are interested in a more homogeneous partitioning of the trajectories regarding the frequency of its aspects. The hierarchical tree algorithm generates increasingly specific and detailed clusters as the trajectories are partitioned. It should be noted that, by design, neighboring nodes will have more similar trajectories, while distant nodes will have less similar trajectories.

The algorithm stops and returns the generated tree when the stipulated minimum number of trajectories or maximum tree level (or both parameters) are reached. These hyperparameters may vary and depend on the application domain to be analyzed.

4. Experimental Evaluation

The experiments carried out to evaluate the clustering algorithm with decision trees used a dataset with data from the Foursquare social network, a platform based on the location and check-ins of users, which was also used in previous works such as [Varlamis et al. 2021], [Petry et al. 2016] and [Yang et al. 2015]. The data represent the set of trajectories of 193 users, who accounted for a total of 66962 points in 3079 trajectories. The aspects present in the dataset are i) the geographic coordinates (latitude and longitude, being a numerical attribute), ii) the time (numeric) at which the user checked in, iii) the day of the week (nominal), iv) the point of interest (POI), which can be understood as the name of the establishment or store, such as Starbucks and McDonald's (nominal), iv) the type of POI (nominal), such as coffeehouse and fast food, v) check-in category (nominal), called the root type (for example, Food, Outdoors Recreation), vi) the rating assigned by the user in the application (ordinal) and vii) the weather condition (nominal) at the time of check-in.

The assembly of the cluster decision tree was created for each proposed partitioning aspect choice method, using the average occurrence of the aspects as a form of division. The parameters used to create the clusters included as a stop criterion the minimum number of 25 trajectories in the final cluster (leaf node). This stop criterion can be viewed as a hyperparameter of the proposed method, therefore it is highly application-dependent and arbitrary. An eventual optimal value for this parameter needs to be defined empirically for the application at hand.

The experiments were carried out using the most strongly correlated aspects, similar to the experiments conducted in [Varlamis et al. 2021]. The aspects used were the weather, the day of the week and the POI category. The algorithm is very flexible and allows for any other combination of aspects, and this combination also depends on the application at hand and on what kind of patterns the analyst wants to mine.

The four trees resulting from each method showed numerical similarities, in terms of number of nodes, leaf nodes and height. In relation to the number of nodes, the trees had an average of 388 nodes and in relation to the leaf nodes the average was 195. The heights of the trees varied between 8 and 14. Table 1 shows the detailed results of clustering.

However, the structures of the generated trees were different. These differences between the methods are interesting, as they can capture different information and behaviors in the set of trajectories.

The dendrograms illustrated in Figures 4 and 5 exemplify these divergences. In Figure 18, the generated tree was modeled by the MV aspect choice criterion, while in Figure 19, the criterion was the MVR. When comparing the two diagrams, it is noted that the first captured an atypical behavior of the trajectories, evidenced in the extreme right branch. On that occasion, the algorithm selected the *Event* aspect of the root_type attribute, in the root node, and grouped all the trajectories of the users who checked-in for this type of aspect. This group had 20 users, out of a total of 193, adding up to a total of 23 trajectories, out of a total of 3079 trajectories. Among all the aspects present in the dataset, *Event* represented only 26 check-ins, in the universe of 66962.

Figures 6 and 7 show an exploratory data analysis from the aforementioned group. The bar graphs show the relative frequency of occurrence of aspects of the *root_type* and

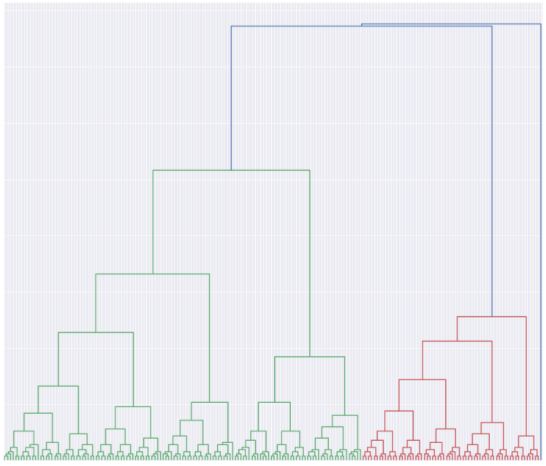


Figure 4. Cluster tree generated by the Minimum Variance criterion.

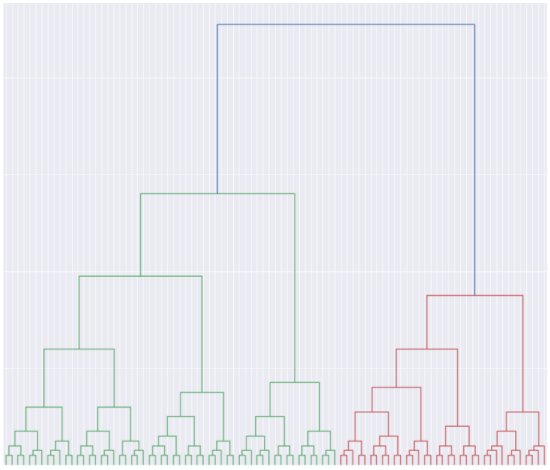


Figure 5. Cluster tree generated by the Maximum Reduction of Variance.

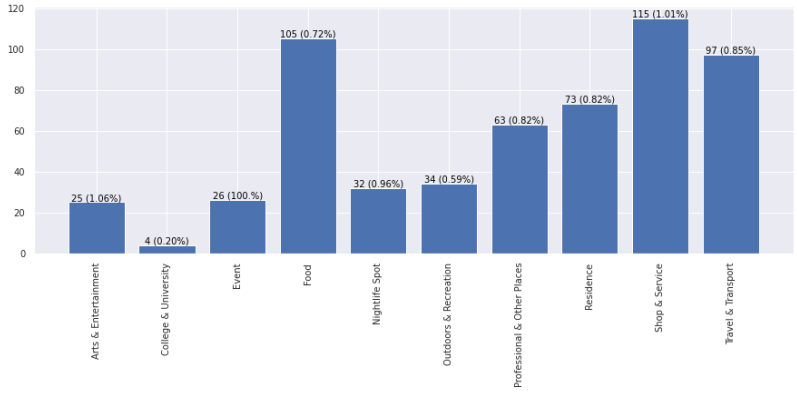


Figure 6. Relative frequencies for *root type* category aspects.

day category, respectively, that is, the frequency of occurrence of aspects of this specific cluster in relation to the total set of trajectories. As can be seen in Figure 5, the relative frequency for *Event* was equivalent to its absolute frequency, that is, all trajectories that

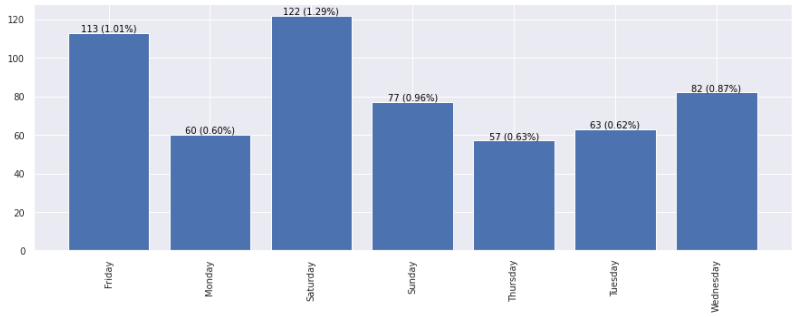


Figure 7. Relative frequencies for *day* category aspects.

checked in related to *Event* were in the same group.

Regarding the days of the week, Figure 6 shows that Saturday was the most frequent day for the group in question. This is an expected behavior, since events usually take place on weekends.

Regarding the other aspects, it is noted that there were few points referring to 'College & University', which may indicate that most users grouped in this cluster are not part of the university community. On the other hand, it can be seen that users in this group are more related to activities involving art and entertainment and nightlife, with Friday and Saturday being more frequent than the others.

Table 1. Relation of the number of generated groups and height of the modeled tree for each aspect selection criterion.

Aspect Selection	Total	Leaf-nodes	Tree Height
BD	385	193	8
MV	387	194	14
LRA	389	195	10
MVR	393	195	10

Table 2. Internal and External Clustering Validation using MUITAS similarity metric.

Aspect Selection	Sil.	CH	DB	Homogen.	Complet.	V-measure	Mut. Info	Rand	F.M
BD	-0.305	14.411	0.369	0.557	0.556	0.557	0.168	0.065	0.070
MV	-0.284	5.199	0.449	**0.630**	0.568	**0.597**	0.122	0.057	0.063
LRA	-0.336	4.408	0.464	0.590	0.561	0.575	0.144	0.066	0.071
MVR	**-0.360**	**23.656**	**0.292**	0.587	**0.581**	0.584	**0.210**	**0.085**	**0.090**

To quantitatively evaluate the clustering result, we employ internal validation metrics to verify the goodness of a clustering structure without referring to external data by using the internal knowledge of the clustering process. As described in [Rendón et al. 2011], the best results are well-separated and compact clusters, meaning that the trajectories in the same cluster are as much similar as possible and the trajectories in different clusters are highly distinct. The evaluation used MSM [Furtado et al. 2016]

Table 3. Internal and External Clustering Validation using MSM similarity metric.

Aspect Selection	Sil.	CH	DB	Homogen.	Complet.	V-measure	Mut. Info	Rand	F.M
BD	-0.289	12.785	0.476	0.557	0.556	0.557	0.168	0.065	0.070
MV	-0.309	1.640	0.509	**0.630**	0.568	**0.597**	0.122	0.057	0.063
LRA	-0.305	3.321	0.566	0.590	0.561	0.575	0.144	0.066	0.071
MVR	**-0.317**	**18.440**	**0.413**	0.587	**0.581**	0.584	**0.210**	**0.085**	**0.090**

and MUITAS [Petry et al. 2016] as similarity metrics to build the distance matrix between the dataset trajectories.

We also employ external validation that consists in comparing the cluster analysis results to an externally known outcome. As mentioned in [Varlamis et al. 2021], we assume that trajectories of the same user are likely to belong to the same cluster, and therefore, the external evaluation of the clustering uses the user as a reference label, that is, the ground truth.

For the internal metrics, we used the silhouette score, the Calinski Harabaz score (C.H), and the Davies-Bouldin Index (D.B), and for the external metrics, we used the homogeneity score, the completeness score, the v-measure score, the adjusted mutual info score, the adjusted rand score, and the Fowlkes Mallows score (F.M).

As demonstrated below, the proposed method achieved very promising results. In comparison to the state-of-the-art method for MAT clustering presented in [Varlamis et al. 2021], the values of the internal and external validation metrics were better, indicating that there was a good clustering. Among the proposed methods for selecting the attribute for dividing the trajectories, the best approach was the maximum reduction of variance, although all of them presented better results than the classic clustering.

Considering the similarity metrics MUITAS and MSM, and considering Agglomerative Clustering using Single Linkage in [Varlamis et al. 2021], when we compare the values of the silhouette coefficient, for example, which presented as a result -0.19 and -0.13 respectively, our method generated as results the values -0.36 and -0.32 respectively. It is important to note that the negative value obtained in this coefficient is due to the use of similarity matrices as a computation of the distances between the trajectories, and not the distances between them. This means that the trajectories are better clustered together, i.e., clusters are more cohese than those found on the previous method. Regarding external validation, our results were also superior to the classical approach, denoting that the resulting clusters were more homogeneous and better complete labeling, as shown in Table 2 and Table 3. In terms of internal validation such as the Silhouette coefficient, and considering the MUITAS similarity metric, we obtained a result approximately 89% better and in relation to the external validation, considering the V-measure metric, for instance, we obtained a 5 times greater precision.

5. Conclusion

With the huge volume of data and the increasing computational power currently available, the necessity of new methods for analyzing complex data becomes more evident. Multiple aspect trajectories bring lots of opportunities regarding data mining, but the sequence nature, the high dimensionality, heterogeneity and data volume, also pose new challenges.

Although a number of methods have been proposed for MAT classification, only

a few works focus on MAT clustering. In this paper, we proposed MAT-Trees, a novel method for multiple aspect trajectory clustering using a decision tree-based approach. The main contribution of this algorithm is a new approach to cluster trajectories considering all their dimensions, since studies on the semantic dimension of trajectories are quite recent and limited. The experimental results showed a higher performance when compared to a state-of-the-art clustering method for MAT, denoting that the new method is capable of generating more cohesive, compact and connected clusters.

The different options of clustering and visualizations provides a flexible tool for exploratory data analysis and applications that easily adapt to the task at hand. Since clustering is a data mining task that is inherently highly application-dependent and exploratory [Tan et al. , chapter 1], the flexible characteristic of the proposed method is an important characteristic.

As future work, we plan to provide a detailed investigation regarding the impact of the different split and evaluation criteria is a promising direction, since each of the criterion seems to be better suited for different applications and analysis. For instance, based on our qualitative analysis we see that while the BD evaluation criterion is well suited for generating clusters that are more or less the same size, the MV criterion seems to easily find outliers trajectories, i.e., trajectories presenting aspects that rarely appear on other trajectories.

References

Alvares, L. O., Bogorny, V., Kuijpers, B., Macedo, J. A. F., Moelans, B., and Vaisman, A. (2007a). A model for enriching trajectories with semantic geographical information. In *Proceedings of the 15th Annual ACM International Symposium on Advances in Geographic Information Systems*. Seattle, Washington, USA,.

Alvares, L. O., Bogorny, V., Macedo, J. A. F., Moelans, B., and Spaccapietra, S. (2007b). Dynamic modeling of trajectory patterns using data mining and reverse engineering. In *26th International Conference on Conceptual Modeling*, volume 83, pages 149–154. Auckland, New Zealand.

Dabiri, S. and Heaslip, K. (2018). Inferring transportation modes from gps trajectories using a convolutional neural network. In *Transportation Research Part C: Emerging Technologies*, volume 86, pages 360–371.

Etemad, M., Júnior, A. S., and Matwin, S. (2018). Predicting transportation modes of gps trajectories using feature engineering and noise removal. In *Lecture Notes in Computer Science*, pages 259–264.

Fayyad, U. M., Piatetsky-Shapiro, G., and Smyth, P. (1996). Knowledge discovery and data mining: towards a unifying framework. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96)*, pages 82—-88. AAAI Press.

Ferrero, C., Alvares, L. O., Zalewski, W., and Bogorny, V. (2018). Movelets: Exploring relevant subtrajectories for robust trajectory classification. In *33rd ACM/SIGAPP Symposium On Applied Computing*, pages 267–289. Pau, France.

Furtado, A. S., Kopanaki, D., Alvares, L. O., and Bogorny, V. (2016). Multidimensional similarity measuring for semantic trajectories. In *Transactions in GIS*, volume 20, pages 280–298.

Hung, C., Peng, W., and Lee, W. (2015). Clustering and aggregating clues of trajectories for mining trajectory patterns and routes. In *The VLDB Journal*, volume 24, pages 162–192.

Mello, R. S., Bogorny, V., Alvares, L. O., Santana, L. H. Z., Ferrero, C. A., Frozza, A. A., Schreiner, G. A., and Renso, C. (2019). Master: A multiple aspects view on trajectories. In *Transactions in GIS*, volume 23, pages 805–822.

Nanni, M. and Pedreschi, D. (2006). Time-focused clustering of trajectories of moving objects. In *J. Intell. Inf. Syst.*, volume 27, pages 267–289.

Petry, L. M., Ferrero, C., Alvares, L. O., Renso, C., and Bogorny, V. (2016). Towards semantic-aware multiple-aspect trajectory similarity measuring. In *Transactions in GIS*, volume 23.

Rendón, E., Abundez, I., Arizmendi, A., and Quiroz, E. M. (2011). Internal versus external cluster validation indexes.

Sadahiro, Y., Lay, R., and Kobayashi, T. (1996). Knowledge discovery and data mining: towards a unifying framework. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96)*, pages 82–88. AAAI Press.

Santos, Y., Carvalho, J. T., and Bogorny, V. (2021). Mining frequency-based sequential trajectory co-clusters.

Silva, C. L., Petry, L. M., and Bogorny, V. (2019). A survey and comparison of trajectory classification methods. In *8th Brazilian Conference on Intelligent Systems (BRACIS)*, volume 23, pages 788–793. Salvador, BA, Brazil.

Spaccapietra, S., Parent, C., Damiani, M. L., Macedo, J. A., Porto, F., and Vangenot, C. (2008). A conceptual view on trajectories. In *Data Knowl. Eng.*, volume 65, pages 126–146.

Tan, P., Steinbach, M., Karpatne, A., and Kumar, V. Introduction to data mining. 2nd ed., Pearson.

Tortelli, T. P., Carvalho, J. T., and Bogorny, V. (2022). Hipermovelets: high-performance movelet extraction for trajectory classification. In *International Journal of Geographical Information Science*, pages 1–25.

Varlamis, I., Sardianos, C., Bogorny, V., Alvares, L. O., Carvalho, J. T., Renso, C., Perego, R., and Violos, J. (2021). A novel similarity measure for multiple aspect trajectory clustering. In *Proceedings of the 36th Annual ACM Symposium on Applied Computing*, pages 551–558. New York, NY, USA.

Wu, S. X., Wu, Z., Zhu, W., Yang, X., and Li, Y. (2021). Mining trajectory patterns with point-of-interest and behavior-of-interest. In *IEEE International Conference on Communications Workshops (ICC Workshops)*, pages 1–6. Montreal, Canada.

Yang, D., Zhang, D., Zheng, V. W., and Yu, Z. (2015). Modeling user activity preference by leveraging user spatial temporal characteristics in lbsns. In *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, volume 45, pages 129–142.

Yao, D., Zhang, C., Zhu, Z., Huang, J., and Bi, J. (2017). Trajectory clustering via deep representation learning. In *International Joint Conference on Neural Networks (IJCNN)*, pages 3880–3887.

Ying, J. J., Lee, W., Weng, T., and Tseng, V. (2011). Semantic trajectory mining for location prediction. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, volume 83, pages 34–43. Chicago, Illinois, USA.

Yuan, G., Sun, P., Zhao, J., Li, D., and Wang, C. (2017). A review of moving object trajectory clustering algorithms. In *Artificial Intelligence Review*, volume 47.

Zheng, K., Zheng, Y., Yuan, N. J., and Shan, S. (2013). On discovery of gathering patterns from trajectories. In *IEEE 29th International Conference on Data Engineering (ICDE)*, pages 242–253. Brisbane, QLD, Australia.