UNIVERSIDADE FEDERAL DE SÃO CARLOS

CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

ANÁLISE DE DADOS SEQUENCIAIS HETEROGÊNEOS BASEADA EM ÁRVORE DE DECISÃO E MODELOS DE MARKOV: APLICAÇÃO NA LOGÍSTICA DE TRANSPORTE

STEVE ATAKY TSHAM MPINDA

ORIENTADOR: PROFA. DRA. MARILDE TEREZINHA PRADO SANTOS

São Carlos – SP

Outubro/2015

UNIVERSIDADE FEDERAL DE SÃO CARLOS

CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

ANÁLISE DE DADOS SEQUENCIAIS HETEROGÊNEOS BASEADA EM ÁRVORE DE DECISÃO E MODELOS DE MARKOV: APLICAÇÃO NA LOGÍSTICA DE TRANSPORTE

STEVE ATAKY TSHAM MPINDA

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de São Carlos, como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação, área de concentração: Banco de Dados

Orientador: Profa. Dra. Marilde Terezinha Prado Santos

São Carlos – SP

Outubro/2015

Ficha catalográfica elaborada pelo DePT da Biblioteca Comunitária da UFSCar

A862ad

Ataky, Steve Tsham Mpinda.
Análise de dados sequenciais heterogêneos baseada em árvore de decisão e modelos de Markov : aplicação na logística de transporte / Steve Ataky Tsham Mpinda. -- São Carlos : UFSCar, 2015.
182 f.

Dissertação (Mestrado) -- Universidade Federal de São Carlos, 2015.

1. Data mining (Mineração de dados). 2. Análise de dados. 3. Classificação automática. 4. Árvore de decisão. 5. Markov, Processos de. 6. Logística - transporte. I. Título.

CDD: 005.8 (20ª)

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Ciências Exatas e de Tecnologia
Programa de Pós-Graduação em Ciência da Computação

Folha de Aprovação

Assinaturas dos membros da comissão examinadora que avaliou e aprovou a Defesa de Dissertação de Mestrado do candidato Steve Ataky Tsham Mpinda, realizada em 16/10/2015:

Profa. Dra. Marilde Terezinha Prado Santos
UFSCar

Profa. Dra. Marcela Xavier Ribeiro
UFSCar

Prof. Dr. Alexandre César Muniz de Oliveira
UFMA

Dedico este trabalho à minha família ”les Ataky”e ao meu dileto amigo irmão Helvécio V. Perreira

AGRADECIMENTOS

Durante quase um ano e sete meses de mestrado, encontrei pessoas que contribuíram de alguma forma a este trabalho e às quais endereço minhas gratulações mais sinceras. Algumas consigo lembrar os nomes, contudo quantos anônimos ou esquecidos têm, de ricochete, contribuído. Que aquelas cujos nomes não foram mencionados recebam, da mesma forma, a expressão de todos os meus pensamentos, respeito e gratidão.

Avant tout agradeço a Deus Altíssimo, por sua misericórdia, soberania, fidelidade e bondade, pois até aqui ajudou-me.

A honra vai à minha orientadora professora doutora Marilde Terezinha Prado Santos, a qual me colocou em condições ideais para a realização deste trabalho e tem sido tão generosa em me transmitir a parte de seu conhecimento, sem a qual seria impossível a realização desta dissertação. Estou-lhe grato deveras.

Meus sinceros agradecimentos aos pesquisadores e especialistas Plínio Vilela e Sérgio Brocchetto pelo acompanhamento e colaboração para a concretização deste trabalho.

Minha gratidão, meu profundo respeito a todos os membros da banca, de antemão, por seu trabalho e atenção consagrada a este trabalho.

Agradeço aos membros do meu laboratório LaBDES e turma que pude frequentar durante o período do meu trabalho e que souberam tornar meu ambiente agradável, com sua presença e convívio saudável. Cito Arthur Andrade, Claudineia Gonçalvez, João Paulo, Alex Guido, Bruno Silva, Ana Bratz, Amir J., Alessandro P., Rafael S., Pablo Botton, Rogers C., Rafael S., Suzane Carol, Bento S.e Kathiane. Mestre Luíz Pacini pelos almoços e jantares assim como o prazer de compartilharmos a ciência. Foi um grande prazer poder contar com sua companhia.

Minha gratidão aos professores Estevam R. e Maria do Carmo, pela urbanidade , disponibilidade e seus conhecimentos em diversas áreas, os quais puderam compartilhar com a minha pessoa para superar certas dificuldades.

Aos mais nobres confrades e insubstituíveis Patrick A. Claeys, Mike Tshibe, Cédrick B

(*father of Computer Vision*), Hans Cutrim, Leopold Mulumba, Ghislain Tshibangu, Joel Bafumba, Arnold Banza, Samir Souza, Fabrice Mutumbo (*le grand Baobá*), Leonardo Melo (capela), Mário Henrique, Higo F., Arthur Pinheiro (rei), Diego Saqui, Serge Lewula, Philippe T, Junior Bibefu, Débora Castro, Olavo Castro, Tio Domingo (meu pai) e dileta Raíssa Everton pelas motivações e encorajamentos.

Deveras agradecido ao meu caro Roussian Gaioso pelo provimento logístico para o meu deslocamento seguro às dependências da UFSCar.

Dedico do fundo meu coração este trabalho, a meus nobres pais Guillaume Kisseng Ataky e Annie Kafuti, a meus irmãos Amy, Geoffrey, Heritier, Jered, Patrícia, Ezechel e Carolina, a meu diletíssimo amigo irmão Helvécio Perreira. É com seu total apoio, paciência e amor que estou aqui hoje. Estou-lhes assaz grato pelos sacrifícios que fizeram a meu favor durante esses tempos dos meus estudos e de ausência.

A todos que me vilipendiaram, pois graças a eles tive mais motivação e ânimo para fazer melhor as coisas.

A todos que me disseram NÃO, pois graças a eles aprendi a fazer as coisas sozinho, a correr atrás e a valorizar mais a vida.

Aos professores do DC - UFSCar e DEINF - UFMA que participaram direta e indiretamente da minha formação.

A CAPES e ao PPGCC pelo apoio financeiro.

A PIB-São Luis pelo apoio espiritual.

A todos que ajudaram, torceram, ou de alguma forma, contribuíram para meu sucesso na realização deste trabalho.

Por fim, muito obrigado a minha futura esposa que, mesmo ainda não nos conhecendo como esposo e esposa, está em algum lugar do universo torcendo pelo meu sucesso.

Je te raconterai les difficultés que j'ai dû affronter et te dirai que ça a valu la peine.
I'll tell you the difficulties I had to face and tell you it was worth it.
Steve Ataky T. Mpinda

RESUMO

Nos últimos anos aflorou o desenvolvimento de técnicas de mineração de dados em muitos domínios de aplicação com finalidade de analisar grandes volumes de dados, os quais podendo ser simples e/ou complexos. A logística de transporte, o setor ferroviário em particular, é uma área com tal característica em que os dados disponíveis são muitos e de variadas naturezas (variáveis clássicas como velocidade máxima ou tipo de trem, variáveis simbólicas como o conjunto de vias percorridas pelo trem, etc). Como parte desta dissertação, aborda-se o problema de classificação e previsão de dados heterogêneos, propõe-se estudar através de duas abordagens principais. Primeiramente, foi utilizada uma abordagem de classificação automática com base na técnica por árvore de classificação, a qual também permite que novos dados sejam eficientemente integradas nas partições inicial. A segunda contribuição deste trabalho diz respeito à análise de dados sequenciais. Propôs-se a combinar o método de classificação anterior com modelos de Markov para obter uma partição de sequências temporais em grupos homogêneos e significativos com base nas probabilidades. O modelo resultante oferece uma boa interpretação das classes construídas e permite estimar a evolução das sequências de um determinado veículo. Ambas as abordagens foram então aplicadas nos dados do sistema de informação ferroviário, no espírito de dar apoio à gestão estratégica de planejamentos e previsões aderentes. Este trabalho consiste em fornecer inicialmente uma tipologia mais fina de planejamento para resolver os problemas associados com a classificação existente em grupos de circulações homogêneos. Em segundo lugar, buscou-se definir uma tipologia de trajetórias de trens (sucessão de circulações de um mesmo trem) para assim fornecer ou prever características estatísticas da próxima circulação mais provável de um trem realizando o mesmo percurso. A metodologia geral proporciona um ambiente de apoio à decisão para o monitoramento e controle da organização de planejamento. Deste fato, uma fórmula com duas variantes foi proposta para calcular o grau de aderência entre a trajetória efetivamente realizada ou em curso de realização com o planejado.

Palavras-chave: Classificação automática, análise de dados sequências, dados heterogêneos, Mineração de dados, planejamento de trens, aderência, replanejamento, previsão de planejamento, árvore de classificação, modelos de Markov

ABSTRACT

Latterly, the development of data mining techniques has emerged in many applications' fields with aim at analyzing large volumes of data which may be simple and / or complex. The logistics of transport, the railway setor in particular, is a sector with such a characteristic in that the data available in are of varied natures (classic variables such as top speed or type of train, symbolic variables such as the set of routes traveled by train, degree of tack, etc.). As part of this dissertation, one addresses the problem of classification and prediction of heterogeneous data; it is proposed to study through two main approaches. First, an automatic classification approach was implemented based on classification tree technique, which also allows new data to be efficiently integrated into partitions initialized beforehand. The second contribution of this work concerns the analysis of sequence data. It has been proposed to combine the above classification method with Markov models for obtaining a time series (temporal sequences) partition in homogeneous and significant groups based on probabilities. The resulting model offers good interpretation of classes built and allows us to estimate the evolution of the sequences of a particular vehicle. Both approaches were then applied onto real data from the a Brazilian railway information system company in the spirit of supporting the strategic management of planning and coherent prediction. This work is to initially provide a thinner type of planning to solve the problems associated with the existing classification in homogeneous circulations groups. Second, it sought to define a typology of train paths (sucession traffic of the same train) in order to provide or predict the next movement of statistical characteristics of a train carrying the same route. The general methodology provides a supportive environment for decision-making to monitor and control the planning organization. Thereby, a formula with two variants was proposed to calculate the adhesion degree between the track effectively carried out or being carried out with the planned one.

Keywords: Automatic classification, sequence data analysis, heterogeneous data, Data mining, train planning, adherence, replanning, planning, forecasting, classification tree, Markov models

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

Capítulo 1

INTRODUÇÃO

Este capítulo apresenta o contexto em que esta pesquisa está envolvida, os pontos motivacionais para investimento de tempo e recurso no projeto em foco, assim como o estabelecimento dos objetivos a serem atingidos ao fim da dissertação de mestrado. Por fim, também é exposta a estrutura e organização do presente documento.

1.1 Contexto

O setor ferroviário está passando por uma fase de aceleração de sua evolução (expansão de trens, a concorrência, desenvolvimento de tráfico transnacional, a preferência ambiental, etc.). Como resultado, o tráfego se intensificou e os métodos de produção estão sendo racionalizados. Desta forma, alguns setores da produção ferroviária tornam-se mais complexos e exigem novas ferramentas de apoio à decisão. É particularmente o caso da construção de mapa de planejamento (agendamento) e gestão operacional de tráfegos.

A produção ferroviária

O produto final da produção ferroviária é o resultado de uma sucessão de operações industriais, cuja maioria permanece fora do domínio público. Essa sucessão começa muito cedo (grandes projetos de infraestrutura são concebidos 10 anos antes de serem colocados em serviço) e termina com a gestão operacional, que consiste em garantir o bom funcionamento das operações e gerenciar facilmente os inevitáveis incidentes que surgem.

Por conseguinte, encontra-se tanto na literatura (por exemplo, (CAPRARA et al., 2007; CORDEAU et al., 1998) como na própria organização de equipes responsáveis por otimizar as

principais fases da produção ferroviária, uma partição em várias categorias de problemas de otimização cujos principais seguem:

- *A identificação da demanda e optimização* de ofertas que podem responder. Isto representa as fases mais a montante da produção para uma empresa ferroviária;
- *A otimização do investimento em infraestrutura* permite propor novas ofertas com base na infraestrutura existente, e estudar os trabalhos necessários para atender às novas exigências (por exemplo, pedido de aumento do tráfego nas linhas regionais);
- *O planejamento* corresponde à fase de construção do conjunto de horários de circulações (também chamado de ”agendamento”);
- *O roteamento ou encaminhamento (platforming)* consiste em atribuir vias para diferentes circulações;
- *Gestão do material* abrange o estabelecimento de horários de cenários, otimização de manutenção (retornos aos depósitos) do material circulante;
- *A gestão operacional do tráfego* deve responder em tempo real aos perigos que tornam a realização do horário teórico impossível na ausência de ação da regulação do tráfego ferroviário. Também chamado de replanejamento.

Estes problemas não são independentes; os resultados de um influenciam nos outros. Outrossim, todos esses problemas se alimentam entre si. No entanto, não é possível hoje tratar todos os eles globalmente.

Todavia, optou-se por não separar completamente o roteamento da gestão operacional do tráfego, na medida em que se trata de um problema importante na prática.

O quadro desta dissertação restringe-se ao estudo da gestão operacional do tráfego, tratando os problemas que tornam impossível a realização do horário planejado. O intuito é tornar o planejamento mais aderente ao que acontece na prática, assim como prever uma provável realização e evolução da circulação dos trens.

A gestão operacional do tráfego

A densidade de tráfego tende a crescer a ponto de causar uma saturação da infraestrutura em muitas áreas. Esta densificação, antes de tudo, tornou-se possível, em parte, através das ferramentas de otimização de construção de horários. No entanto, a mesma densificação torna

os incidentes mais frequentes e, especialmente, suas consequências mais importantes ou, pelo menos, muito mais difíceis de gerenciar de forma eficaz.

O replanejamento e a gestão de tráfego em tempo real necessitam, portanto, cada vez mais das ferramentas de apoio à decisão apropriada.

Esquematicamente, o problema poderia se resumir como segue: devido à ocorrência de um ou mais incidentes (acidentes, problemas meteorológicos, etc.), horários teóricos já não são mais realizáveis, e uma decisão precisa ser tomada, uma vez que na mesma malha (via) há diversas circulações planejadas, e o incidente pode de alguma forma influenciar a realização do planejamento das outras circulações e, eventualmente, originar um custo maior de negócio. O primeiro passo para enfrentar essa perturbação é, basicamente, identificar quais foram as causas ou os eventos geradores do incidente (tratamento estatística de categorias de incidente, envolvendo o lugar, data, etc.), para com base neles analisar as consequências diretas ou indiretas e, em seguida, fazer uma estimativa de replanejamento otimizado de circulações de modo a manter a aderência entre o planejamento inicial e o novo planejamento, considerando as restrições de negócio a serem absolutamente obedecidas.

De acordo com Vilela et al. (2014), os especialistas do domínio, a representação de uma solução para a tarefa de planejamento requer a modelagem dos elementos que representam a sequência de eventos que o trem tem de enfrentar para realizar o objetivo esperado. Um trem contém informações específicas, tais como: seu comprimento, tipo de carga, velocidade, etc. A Figura 1.1 ilustra as classes para modelar a solução de planejamento de trens.

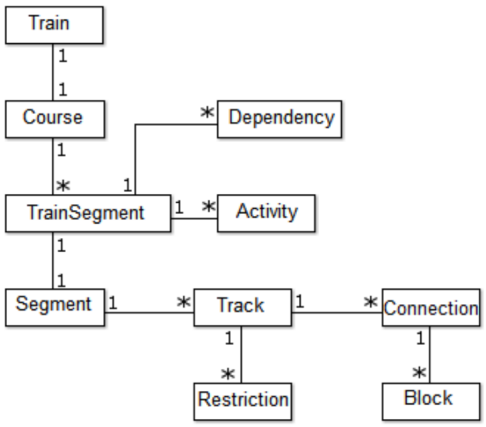

Figura 1.1: Classes para modelar o problema de planejamento (visão microscópica)(VILELA et al., 2014)

O itinerário do trem especifica quais são os segmentos que o trem deve usar para ir desde a sua origem até o seu destino. Este itinerário contém um conjunto de segmentos que são específicos para esse trem. Um segmento de trem especifica o conjunto de possíveis vias através

das quais o trem pode passar, as atividades que o trem deverá realizar nesse segmento, as dependências entre o trem que está sendo planejado e outros trens, e o conjunto de vias através do qual o trem não pode passar (Figura 1.2), visto como as relações entre os elementos que definem um trem e o seu itinerário.

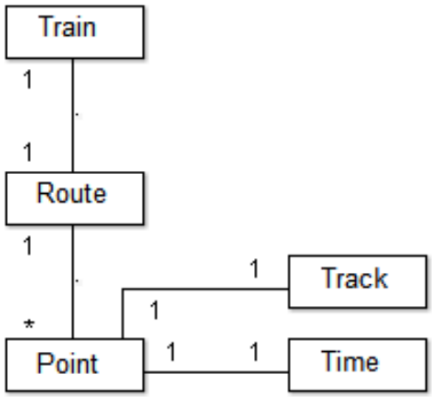

Figura 1.2: Classes para modelar o problema de planejamento (visão macroscópica)(VILELA et al., 2014)

Esta solução inclui o horário e a via em que o trem deve passar, que doravante chamar-se-á de posição, na rota toda a partir da origem até o destino. Tendo em vista diferentes posições em que o trem pode se encontrar, depreende-se que a rota pela qual esta circulação é realizada é um conjunto de pontos onde cada ponto é identificado separadamente. As publicações referentes afirmam que cada ponto é dada por uma referência em um instante de tempo e em uma via associada. Deste modo, é possível representar os movimentos de parada, onde há vários pontos no mesmo local, mas em diferentes instantes de tempo. A Figura 1.2 mostra como estes elementos estão relacionados com o modelo.

1.2 Motivação e Objetivo

A empresa ABC [1] desenvolve soluções de inteligência para otimização, planejamento e suporte à tomada de decisão nas operações e nos negócios. Igualmente, ela desenvolve pesquisas relacionadas com a melhoria da qualidade do planejamento de trens. A solução atual adotada pela empresa implementa um algoritmo baseado em simulação de eventos discretos e heurísticas para calcular o planejamento dos trens. Esse algoritmo tem uma série de vantagens em relação aos algoritmos baseados em, por exemplo, otimização matemática. A principal delas é a capacidade de representação das entidades de domínio e das restrições consideradas no cálculo da

[1] Por razão de confidencialidade, fomos solicitados para não divulgar o nome da empresa

circulação, isso permite que a solução seja mais aderente às necessidades dos clientes. Além disso, também é capaz de produzir resultados válidos de circulação de trens (sem bloqueios na malha) em um tempo de processamento que permite o seu uso em ambiente operacional, ou seja, computando um resultado válido em alguns segundos e não minutos ou horas como no caso de outros algoritmos encontrados na literatura (CORDEAU et al., 1998; BUSSIECK et al., 1997).

Essas características permitem à empresa uma vantagem competitiva no mercado. Por outro lado, os princípios implementados no algoritmo trazem algumas desvantagens pontuais que são difíceis de serem tratadas sem subverter o funcionamento do algoritmo e sua estabilidade. De certa forma, a empresa fornece uma solução válida considerada boa, mas não necessariamente ótima. Algumas decisões tomadas em escopo local podem não representar as melhores opções quando considera-se o escopo de planejamento sob uma perspectiva mais global.

A empresa também enfrenta um problema no que tange à aderência entre circulações planejadas e realizadas, isto é, nem sempre o planejamento realizado corresponde ao planejado. Devido à ocorrência de um evento imprevisto que, geralmente, não são automaticamente identificados (por falta de um mecanismo apropriado para tal fim), uma série de consequências podem afetar a realização de uma circulação conforme inicialmente planejada. Quando isso acontece, faz-se necessário tomar decisões com a maior celeridade possível, uma vez que existe uma dependência (concorrência) entre diferentes circulações tendo em vista que estas compartilham do mesmo recurso, a saber, a malha ferroviária. Para tal, é preciso encontrar uma forma de se remediar.

De fato, os possíveis incidentes podem ser analisados, ao estudar o comportamento da evolução de circulações, e, em seguida, categorizados conforme seus impactos no planejamento. Ao conseguir distinguir as consequências de cada categoria de incidente, poder-se-á prever uma nova evolução com base nos dados recolhidos cada vez que ocorrer um incidente que se enquadre em algumas dessas categorias. Os fatores (incidentes) que geralmente influenciam na não aderência de trens, especialmente trens de carga nas áreas ou regiões não urbanas, são, por exemplo, as condições climáticas (vento, chuva, precipitações, temperatura, etc.) e não climáticas (condição da malha, folhagem na malha, água estacionada na malha, lama, etc.)

No planejamento de trem, tem-se as informações relacionas ao seu horário de saída, itinerário (conjunto de pontos ”vias”, velocidade ”km/h”previsto, sentido), tipo de trem (velocidade ”min/max”, comprimento, prioridade). Por outro lado, o realizado final fornece o conjunto de horários relacionados com o conjunto de vias (posições) pelas quais passou e, eventualmente, o tipo de trem.

A empresa alega que, ao chegar ao destino, as informações disponíveis para um determinado trem, atualmente, são o planejamento e o realizado final, conforme ilustra a Figura 1.3.

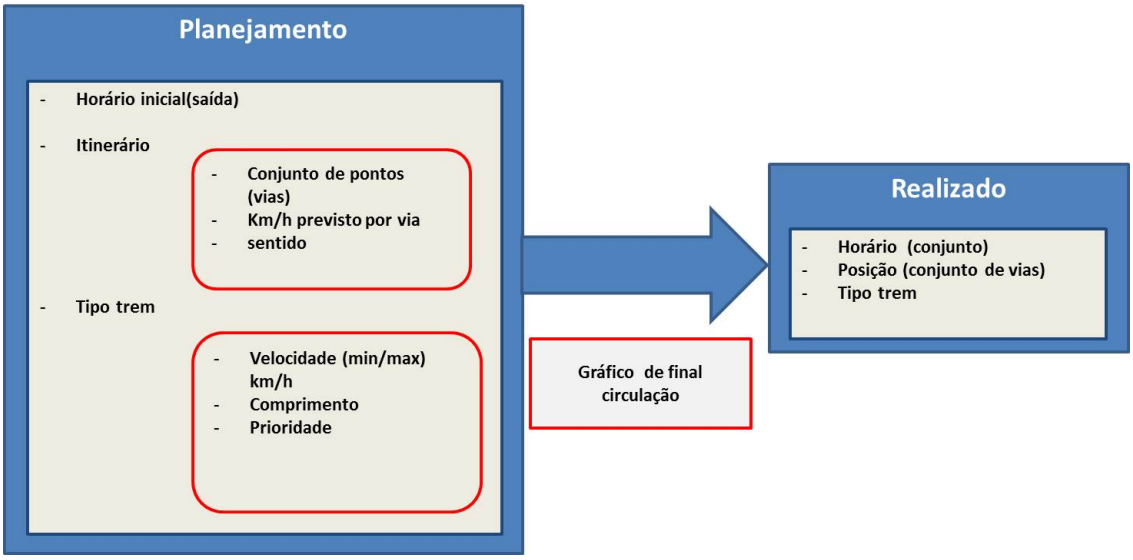

Figura 1.3: Cenário atual entre o planejamento inicial e realizado final

Hipóteses

É imprescindível explicitar os motivos motivadores (qual o problema a ser resolvido e por quê?), assim como a forma em que se espera desenvolver ou propor uma solução. Para tal, começa-se por apresentar uma justificativa sucinta.

Justificativa: justifica-se abordar este tema por ser um problema real e que ainda não foram encontradas soluções viáveis na literatura no sentido direto do contexto estudado.

Quais são as hipóteses? Quais são as questões de pesquisa?

H1 - a partir de dados de circulações realizadas é possível identificar incidentes ocorridos durante sua realização;

H2 - a partir dos incidentes identificados é possível correlacioná-los a impactos no planejamento (categorizá-los);

H3 - a partir da ocorrência de um incidente, ou dispondo de informação relacionadas, é possível determinar outros incidentes derivados (um incidente pode provocar outros incidentes);

H4 - a partir de informações de previsão de ocorrência de incidentes e do impacto dos incidentes no planejamento de circulações, é possível melhorar a aderência entre circulação planejada e realizada.

Tendo em vista a identificação de um problema a ser resolvido, o conjunto de hipóteses e

questões de pesquisa relacionados, o objetivo principal deste trabalho de mestrado é propor uma abordagem que permita melhorar a aderência entre o planejamento de circulação de veículos frente ao efetivamente realizado, assim como realizar uma previsão sobre o comportamento futuro dos veículos.

Os objetivos específicos que permitirão atingir ao objetivo da previsão consiste em responder às seguintes perguntas:

O que deve permitir a previsão de planejamento?

Moon et al. (2000), distingue três preocupações em torno de previsões, as quais foram adaptadas no caso do planejamento:

- previsão de planejamento, definida como uma projeção para o futuro do planejamento, dado um número de condições ambientais;
- o planejamento operacional, definida como um conjunto de decisões e ações de gestão tomadas para atender ou aproximar as previsões do planejamento;
- As metas de circulação a serem alcançadas.

Na verdade, existem relações de dependência entre a previsão de circulação, o planejamento operacional e as metas de circulação. É imprescindível observar que a previsão de circulação é a condição *sine qua non* para todo planejamento e que, daí, ela também é uma ferramenta essencial para o controle da empresa.

Quais são as fontes de dados de previsão?

Para se fazer previsões, é preciso dispor de um histórico de grandezas a prever. Isto pode parecer óbvio, mas ainda há alguns anos, a empresa aqui estudada não armazenava sistematicamente todas as informações. De qualquer forma, a operação mais importante no processo de previsão é a recuperação do histórico e seu saneamento, diz Robert Husset, CEO da Aldata Solution [1]

Quais os dados disponíveis na empresa que podem ser utilizados na previsão? Quais são os dados externos que explicam a grandeza de se esperar? Qual é a qualidade dos dados? Qual é a confiança? Qual é o nível de detalhe? Com que frequência são coletados? Qual é o tipo desses dados? A resposta a essas perguntas irá determinar os modelos e técnicas de se usar.

[1] Success Story : Le système de prévision des ventes chez Match, Aldata Solution 2008

As circulações são muitas vezes influenciadas por eventos externos (temporada, meteorológicos, manutenção, etc.). Ao conseguir controlar essas variáveis e prevê-las, sê-lo-á vantajoso usar as técnicas mais adequadas, tais como as usadas neste trabalho, para enfrentar seus impactos.

Quais são os parâmetros de previsão de circulações?

Considerando-se que a previsão deva ser a montante do planejamento, os conceitos básicos da previsão baseiam-se, principalmente, na programação.

Quais são as técnicas a escolher para prever os planejamentos das circulações?

Mentzer e Gomes (1994), alegam que não existe uma técnica superior a outra. A razão vem da grande variedade de séries temporais e horizontes de previsão. Uma máquina, no entanto, poderia armazenar o comportamento de circulações em determinada posição de forma contínua e que poderia permitir dispor-se de informações (infinitamente) mais numerosas. Estas séries temporais contêm as características específicas e, portanto, parece difícil reter uma única técnica como sendo a melhor.

Por outro lado, as técnicas mais complexas nem sempre são os mais bem sucedidos. Muitas técnicas utilizam a mesma abordagem e, por conseguinte, apresentam a mesma eficácia. Por exemplo, o método de amortecimento exponencial dupla, o método de Holt Winters e ARIMA podem ser idênticos, sob certas condições (PETER; RICHARD, 1996). Isso leva a concluir que muitas técnicas são semelhantes, embora tenham nomes diferentes. Finalmente,(MENTZER; GOMES, 1994), recomenda a escolher uma ou duas que melhor se adequam ao cenário.

1.3 Visão geral do Trabalho

A visão geral deste trabalho, conforme a Figura 1.4, consiste na coleta de dados heterogêneos, os quais passam pelo processo de homogeneização e, uma vez homogeneizados, aplica-se, por um lado, a técnica de árvore de classificação para agrupá-los nas respectivas classes e, em seguida, calcula-se a aderência quando for necessário. Por outro lado, aplica-se os modelos de Markov, também alimentados pela árvore de classificação, para proceder às possíveis previsões.

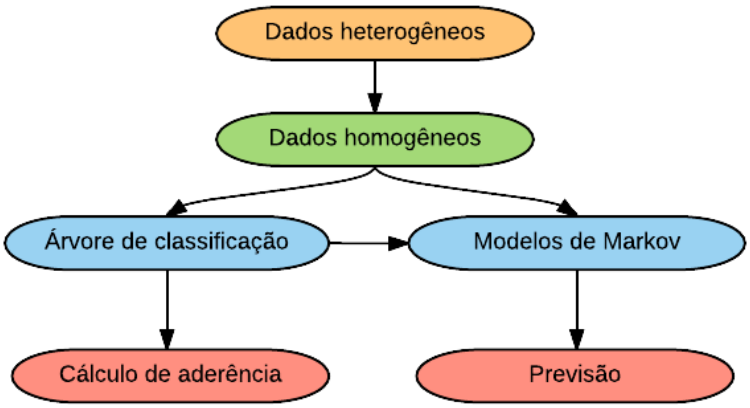

Figura 1.4: Cenário atual entre o planejamento inicial e realizado final

1.4 Estrutura do Trabalho

Este trabalho é organizado da seguinte forma:

- **Capítulo 2**: apresenta a teoria sobre a Classificação de Dados Heterogêneos;
- **Capítulo 3**: aborda o Estado da arte da Análise de dados Sequenciais;
- **Capítulo 4**: resume os Modelos de Markov aplicados no contexto deste trabalho;
- **Capítulo 5**: apresenta a Arquitetura para gerenciamento de dados na logística de transporte proposta neste trabalho;
- **Capítulo 6**: aborda os Conceitos e a aplicação de análise de dados sequencial para este trabalho;
- **Capítulo 7**: apresenta os Experimentos e validação realizados da abordagem proposta;
- **Capítulo 8**: apresenta as conclusões e contribuições deste trabalho.

Capítulo 2

AGRUPAMENTO E CLASSIFICAÇÃO DE DADOS HETEROGÊNEOS

Neste capítulo, faz-se uma revisão sistemática respeitante à agrupamento automático, onde apresenta-se um resumo de trabalhos relacionados à técnica sobrejacente. Precisamente, apresentam-se os diferentes tipos de dados que podem ser objeto de abordagem de agrupamento automático, o problema da definição de um índice de similaridade com finalidade de poder classificar os indivíduos de descrição heterogênea. Além disso, são apresentados os métodos clássicos de agrupamento automático (hierárquica, particionamento e árvore de decisão) encontrados na literatura. O objetivo é apresentar e posicionar o método baseado em árvore de decisão aplicado nesta dissertação.

2.1 Estado da arte: o agrupamento automático

Neste capítulo, faz-se uma revisão sistemática respeitante à agrupamento automático, onde apresenta-se um resumo de trabalhos relacionados à técnica sobrejacente. Precisamente, apresentam-se os diferentes tipos de dados que podem ser objeto de abordagem de agrupamento automático, o problema da definição de um índice de similaridade com finalidade de poder classificar os indivíduos de descrição heterogênea. Além disso, são apresentados os métodos clássicos de agrupamento automático (hierárquica, particionamento e árvore de decisão) encontrados na literatura. O objetivo é apresentar e posicionar o método baseado em árvore de decisão aplicado nesta dissertação.

2.1.1 Introdução

O aumento de números de bases de dados, a diversidade de formatos de dados e a precisão destes últimos em todas as áreas de atividade são campos genuínas em que as empresas têm a oportunidade de minerar para extrair conhecimentos. Estes conhecimentos são disponíveis, mas nem sempre são fáceis de extrair e representar sem inteligência. A inteligência, esta substância, permite-nos, observando as formas, estabelecendo regras, encontrar novas ideias que valem a pena tentar e fazer previsões.

A mineração de dados, do inglês "*data mining*", aparece em meados da década de 1990 nos Estados Unidos como uma nova disciplina na interface da Estatística e tecnologia da informação: bancos de dados, inteligência artificial, aprendizado de máquina ("machine learning"), e pode ser definida como a seguir:

Definição 1

Conjunto de técnicas e métodos de áreas de estatística, matemática e ciência da computação permitindo a extração, a partir de um grande volume de dados brutos, conhecimentos originais outrora desconhecidos. Este é "mineração"para descobrir "informações ocultas"que os dados contêm e que se descobrem buscando associações, tendências, relacionamentos ou padrões. Técnica de mineração de dados é encontrar estrutura e relacionamento escondido em grande número da população (HUANG; BENESTY, 2004).

Definição 2

Data Mining, ou Mineração de Dados é o conjunto de métodos e técnicas para a exploração e análise de grandes bases de dados computacionais, de forma automática ou semiautomática, para detectar nestes dados regras de associações, de tendências desconhecidas ou ocultas, de estruturas especiais restaurando a maior parte da informação útil, enquanto reduz a quantidade de dados para apoiar a tomada de decisão (TUFFÉRY, 2012).

Em suma, a Mineração de Dados é a arte de extrair informações (ou conhecimentos) a partir de dados.

Consoante (TUFFÉRY, 2012), a mineração de Dados é caracterizada por dois tipos de técnica:

- *As técnicas descritivas (busca de "patterns"*: essas técnicas também chamadas exploratórias visam destacar informações presentes, mas escondidas pelo volume de dados.

Elas reduzem, resumem e sintetizam os dados; não há uma variável ”alvo”a prever.

Os principais métodos aplicacados neste contexo são: análise fatorial, agrupamento automática (”clustering”) e busca de associação.

- *As técnicas preditivas (modelagem)*: igualmente conhecidas como explicativas, destinam-se a extrapolar novas informações a partir de informações presentes (caso de pontuação “*scoring*”). Elas explicam os dados; neste caso há uma variável ”alvo”a prever.

  Ademais, seus principais métodos podem se classificar em:

  – *agrupamento/descriminação (variável ”alvo”qualitativa)*: análise descriminante / regressão logística, árvore de decisão e redes neurais;

  – *predição (variável ”alvo”quantitativa)*: regressão linear (simples e múltipla), ANOVA, MANOVA, ANCOVA, MANCOVA (GLM), árvore de decisão e redes neurais.

A mineração de dados (TUFFÉRY, 2012; JAMBU, 2000), ou análise inteligente de dados refere-se ao conjunto de métodos para a exploração e análise de dados computacionais, de forma automática ou semi-automática, a fim de detectar nestes dados regras, tendências desconhecidas ou ocultas e estruturas especiais.

Como parte deste trabalho, aborda-se o problema de agrupamento automática. A agrupamento automática ou análise de agrupamento, é vista como a tarefa que classifica uma população heterogênea em uma série de grupos mais homogêneos, chamados *clusters*.

Neste processo, não há nenhuma variável alvo privilegiada, e não tem-se nenhuma outra informação prévia senão a descrição de dados em uma lista de variáveis comuns. Destarte, poder-se-á dizer que a clusterização é uma tarefa de aprendizagem ”não supervisionada”onde os registros são agrupados de acordo com a sua similaridade, de modo a satisfazer as duas seguintes propriedades:

- Os elementos pertencentes à mesma classe são tão similares quanto possível; em outras palavras, caracterizam uma homogeneidade intraclasse (coesão), que reflete essa característica.

- Os elementos pertencentes às diferentes classes são tão diferentes quanto possíveis; em outras palavras, caracterizam uma heterogeneidade interclasses.

Após a fase de agrupamento, deve-se determinar o significado, se houver, a ser atribuído aos *clusters* resultantes. À vista disto, na área médica, por exemplo, a clusterização é utilizada

para determinar grupos de pacientes suscetíveis a estarem sujeitos a determinada terapias; cada classe agrupando pacientes reagindo de forma idêntica.

Ademais, tal como acontece com todos os métodos de mineração, os dados não estruturados não são diretamente analisados por meio de técnicas de agrupamento automático, todavia, apresentam-se mais frequentemente sob-forma de uma matriz retangular onde nas linhas têm-se indivíduos (objetos, entidades, instâncias, etc.) e nas colunas as variáveis (atributos, características, etc.)

Em alguns casos, o usuário tem a saída de uma matriz de similaridade (similaridade, dissimilaridade ou distâncias) entre os objetos a serem classificados, ou, caso contrário, a construir a partir de seus dados. Estas medidas de similaridade entre os objetos dependem da natureza das variáveis medidas.

2.1.2 Diferentes tipos de varáveis

O agrupamento ocorre nos dados resultantes de uma série de escolhas que irão influenciar os resultados da análise. Tipicamente, os dados são descritos em uma matriz de indivíduos-variáveis por um valor único. Em aplicações reais, onde a principal preocupação é levar em conta a variabilidade e riqueza de informações nos dados, é comum lidar com dados complexos e heterogêneos (ou mistos). O que resulta em que cada posição na matriz de descrições pode conter não apenas um único valor, mas também um conjunto de valores, um intervalo de valores ou uma distribuição de um conjunto de valores. Dir-se-á, portanto, que o agrupamento baseia-se em uma *"matriz de descrições simbólicas"*.

Descrição clássica de uma variável

Chama-se variável qualquer característica de uma entidade (pessoa, organização, objeto, evento, etc.), que pode ser expressa como um valor numérico (medida) ou codificado (atributo).

Os possíveis valores de uma variável, para o conjunto de indivíduos estudados, são chamados modalidade da variável. Em outras palavras, as modalidades correspondentes aos possíveis valores da variável estatística. Em estatística, uma variável estatística define uma partição sobre uma população, cada indivíduo pertence a uma e uma única modalidade.

As informações sobre o problema a ser resolvido se apresenta, na maioria das vezes, sob a forma de tabelas ou matrizes; onde as linhas representam exemplos ou casos a serem estudados ou tratados. Além disso, variáveis, também chamados atributos, que descrevem um caso podem

ser de vários tipos.

A seguir uma descrição dos diferentes tipos de variáveis e suas características:

- Disjuntivas: podem admitir dois estados (exemplo: verdadeiro ou falso);
- Categóricas não ordenadas: as diferentes categorias contêm nenhuma noção de ordem (exemplo, a cor de cabelos);
- Categóricas ordenadas: as diferentes categorias podem ser classificadas (exemplo: faixa de atribuição de conceitos);
- Contínuas: podem tomar valores numéricos em que os cálculos, tais como a média, podem ser realizados.

Observação 1: Os tipos de variáveis influenciam fortemente nas técnicas utilizadas no processo de Mineração de Dados.

Descrição simbólica de uma varável

Como parte da análise de dados simbólicos introduzida por (DIDAY; KODRATOF, 1991), a definição de uma variável foi estendida afim de poder descrever um indivíduo por variáveis Y_h com várias modalidades de domínio de observação O_h (CHAVENT, 1997; EL-GOLLI, 2004). O domínio de chegada de uma varável Y_h a descrição simbólica será então modificada em relação à variável clássica ou convencional O_h. Neste contexto, distinguem-se, geralmente, três tipos de variáveis a descrição, a saber: multi-valoradas, modais e variáveis taxonômicas ou estruturadas.

Variáveis a descrições multi-valoradas

Considera-se uma variável Y_h que pode ser descrita por vários valores do domínio de observação O_h.

- Se o domínio de observação O_h for quantitativo (contínuo ou discreto), a descrição multi-valorada de Y_h é um intervalo de valores, e o domínio de chegada Δ_h de Y_h é o conjunto de intervalos fechados e limitados sobre O_h. Por exemplo, a variável Y_h = *tempo ideal* (em minutos) que um trem pode gastar circulando em um determinado segmento da malha com determinadas condições climáticas = [117,162].
- Se o domínio de observação O_h for qualitativo nominal, a descrição multi-valorada de Y_h é um conjunto de valores, e o domínio de chegada Δ_h de Y_h é conjunto de

subconjuntos de O_h. Por exemplo, a variável Y_h = grau de aderência de circulações realizadas pelo trem AA1 com relação ao planejamento pode pegar os valores Y_h (AA1) = {convergente, divergente}.

Chavent (1997) afirma que no nível semântico, as descrições multi-valoradas permitem traduzir os conceitos de imprecisão e variabilidade na descrição dos indivíduos.
Suponha-se que Y_i seja um indivíduo que tenha como descrição relativa à variável *tipo de cônica* o conjunto de valores $Y_h(X_i)$ = {*círculo, elipse*}. Isto pode corresponder a uma imprecisão devido a uma dúvida: a cônica em questão é um círculo ou elipse. Se, de outra parte, a variável *velocidade prevista do trem*, Y_h (*velocidade*)=[12.42, 14.50], isto pode corresponder a um ruído: a velocidade do trem pode variar entre 12,42km/h e 14,50km/h. No primeiro caso, o conceito da verdadeira cônica não faz muito sentido, pois depende do ponto de vista de cada indivíduo. Já no segundo caso, pode-se supor que a verdadeira velocidade do trem pertence ao intervalo [12.42, 14.50].

Sob outra perspectiva, um intervalo ou conjunto de valores pode permitir introduzir o conceito de variabilidade na descrição. Por exemplo, o conjunto *Ensolarado, nublado, chuvoso* pode exprimir a lista de todas os estados do tempo. O intervalo [122,158] pode exprimir a variação do tempo, em minutos, que um trem pode gastar num percorrendo um determinado segmento. Trata-se aqui de variabilidade devido ao caráter temporal da variável.

Variáveis a descrição modais

Diz-se de uma variável Y_h que pode descrever-se por uma função definida sobre o domínio de observação O_h em [0,1].

Esta função pode ser uma distribuição de probabilidade sobre O_h ou uma função de composição de conjunto *fuzzy* de O_h. Exemplificando, poder-se-á indicar que a velocidade de um trem é uniformemente distribuída sobre o intervalo [12.42, 14.50], ou ainda normalmente distribuída em torno do valor 13.46. Neste caso, a velocidade do trem é descrita pela função da densidade da lei normal de média 13.46 e de desvio padrão σ.

Ao contrário do caso multi-valorado, em que os valores que uma variável assume traduz a imprecisão sem contudo dar um grau de certeza sobre esses valores, as variáveis modais são usadas para converter o conceito de imprecisão para o conceito de incerteza. Por exemplo, para a variável Y_h = *tipo de cônica* onde o domínio de observação O_h é definido por um conjunto de valores precisos, um grau de incerteza pode ser fornecido para converter o conceito de imprecisão na descrição de dados. Poder-se-á, desta forma, por exemplo, dizer que o tipo do objeto é ”circular”com 2/3 de certeza e ”elipsoidal”com 1/3

de certeza.

Variáveis taxonômicas ou estruturadas

Os domínios de observação das variáveis de agrupamento às vezes podem ser munidos de conhecimentos adicionais chamados *conhecimentos de domínio*. Estes conhecimentos adicionais são definidos no caso de descrições mono-valorado, entretanto podem ser considerados no tratamento sobre descrições multi-valorado (por exemplo, no cálculo da medida de similaridade entre indivíduos em um processo de agrupamento automático).

Às vezes, acontece que um especialista possa fornecer uma estrutura de valores do domínio de observação como uma árvore ordenada, um grafo direcionado, etc. De acordo com (ICHINO; YAGUCHI, 1994; MICHALSKI; STEPP, 1983), uma variável cujo domínio de observação é representado por uma estrutura hierárquica é chamado variável taxonômico ou estruturada.

2.1.3 Medidas de similaridade

Em estatística e matemática, uma medida associada a um conceito estabelece uma correspondência entre objetos e números, o que permite comparar os objetos e determinar o valor de verdade de um ou mais relações $=$, $\neq$, $<$ ou $>$. Se, como muitas vezes acontece, um conceito contém várias dimensões, e ainda se queira tratá-lo como um todo, é preciso superar o problema da multidimensionalidade, o que pode-se fazer através da construção de um índice.

O processo que consiste em classificar os objetos pode ser grandemente facilitado se puder-se formalizar o conceito de similaridade e nele associar uma medida. Na verdade, existem procedimentos de agrupamento automatizado baseados em medidas de similaridade.

Observe-se que, primeiramente, o conceito da similaridade aplica-se a um par de objetos. A similaridade é, portanto, uma propriedade de nenhuma dos dois objetos: é uma propriedade do par (ambos os objetos). Em seguida, o conceito de similaridade é um conceito geral que abrange uma infinidade de conceitos específicos: pois, quando examina-se a similaridade entre dois objetos, é sempre relativa a um determinado atributo. Um conceito específico de similaridade é definido pelo atributo a que se refere para comparar os objetos que quer-se medir a similaridade.

Isto posto, um sistema, qualquer que seja, que tem como finalidade analisar ou organizar de forma automática um conjunto de dados ou conhecimento, deve usar, de uma forma ou outra, um operador capaz de avaliar com precisão as similaridades ou dissimilaridades entre esses dados.

O conceito de similaridade originou bastante pesquisas em diversos domínio ou diversas

áreas. Para qualificar este operador, vários conceitos, tais como similaridade, dissimilaridade ou distância podem ser usados.

A seguir, apresentam-se mais detalhadamente conceitos acima (similaridade, dissimilaridade e distância), as relações entre eles, assim como suas diferenças.

Definições

Chama-se similaridade ou dissimilaridade qualquer aplicação a valores numéricos que permite medir a relação entre indivíduos do mesmo conjunto. Para uma similaridade, a relação entre dois indivíduos será tão forte quanto seu valor for grande. Para uma dissimilaridade, ao contrário, a relação será tão forte quanto menor for seu valor de dissimilaridade (CELEUX et al., 1989).

1. **Índice de dissimilaridade**

   O índice de dissimilaridade, além de ser usado para medir a dissimilaridade entre duas distribuições observadas, pode ser igualmente usado para medir a desigualdade ou concentração. Além disso as medidas de desigualdade ou de concentração são tipicamente medidas de dissimilaridade entre uma distribuição observada e uma distribuição de referência. Para medir a desigualdade ou a concentração, trata-se de comparar uma distribuição observada com a distribuição de referência que representa a igualdade perfeita ou concentração nula.

   Assim, um operador de similaridade $d = X * X \rightarrow \mathbb{R}^+$ definido sobre o conjunto de indivíduos $X = \{X_1, \ldots, X_n\}$ é dito índice de dissimilaridade, se verificar as seguintes propriedades:

   1. $\forall X_i, X_j; d(X_i, X_j) = d(X_j, X_i)$ (propriedade de simetria)
   2. $\forall X_i \in X; d(X_i, X_j) \geq d(X_i, X_i) = 0$ (propriedade de positividade)

2. **Distância**

   Entre as medidas de similaridade, algumas são medidas de distâncias, generalizadas por mais de duas ou três dimensões, na medida em que têm as propriedades que uma medição de distância deve possuir. Por outro lado, pode-se considerar a distância como um caso especial da dissimilaridade: a distância é uma dissimilaridade entre dois objetos com relação à sua situação no espaço ou simplesmente entre dois pontos no espaço.

   Dois objetos similares têm, portanto, uma distância nula entre si, a distância máxima separa dois objetos diferentes.

Esta transformação da similaridade em distância permite dar uma representação gráfica. Trata-se de uma primeira abordagem para posicionar objetos em um espaço; quanto mais os pontos estão próximos mais os indivíduos não são similares. Esse predicado é a base das técnicas de classificações. Estas usam o mesmo princípio de distância para construir o agrupamento de objetos em grupos.

Assim, um operador de similaridade $d = X * X \rightarrow \mathbb{R}^+$ definido sobre o conjunto de indivíduos $X = \{X_1, \ldots, X_n\}$ é dito distância se, além de verificar as propriedades 1 e 2, verificar as seguintes propriedades de *identidade* e *desigualdade triangular*:

3. $\forall X_i, X_j \in X; d(X_i, X_j) = 0 \Rightarrow X_i = X_j$ (propriedade de idêntidade)
4. $\forall X_i, X_j, X_k \in X; d(X_i, X_j) \leq d(X_i, X_k) + d(X_k, X_j)$ (desigualdade triangular)

3. **Índice de similaridade**

O problema de medir a similaridade muitas vezes surge nas estatísticas, onde o coeficiente de correlação mais simples é uma medida da similaridade entre dois conjuntos de dados.

Do mesmo modo, para avaliar a precisão de um modelo com relação aos dados utilizados para estimar os seus parâmetros, mede-se a similaridade entre os valores observados e os valores previstos pelo modelo.

Destarte, um operador de similaridade $d = X * X \rightarrow [0, 1]$ definido sobre o conjunto de indivíduos $X = \{X_1, \ldots, X_n\}$ é dito índice de similaridade se, além de verificar a propriedade de simetria (1), verificar as seguintes propriedades:

5. $\forall X_i \in X; s(X_i, X_j) \geq 0$ (propriedade de positividade)
6. $\forall X_i, X_j \in X; X_i \neq X_j; d(X_i, X_i) = s(X_j, X_j) > s(X_i, X_j)$ (propriedade de maximização) Deve-se notar aqui que a passagem do índice de similaridade s no conceito dual do índice de dissimilaridade (que denotar-se-á d) é trivial. Tendo S_{max} a similaridade de um indivíduo consigo mesmo ($S_{max} = 1$ caso de uma similaridade normalizada), basta supor:

$$\forall X_i, X_j \in X; d(X_i, X_j) = S_{max} - s(X_i, X_j) \tag{2.1}$$

Medida de similaridade entre indivíduo a descrição clássica

O processo de agrupamento objetiva estruturar os dados contidos em $X = \{X_1, \ldots, X_n\}$ em função de suas similaridades, sob forma de um conjunto de classes tanto homogêneos como contrastados.

O conjunto de indivíduos X é geralmente descrito sobre um conjunto de m variáveis $Y = \{Y_1, \ldots, Y_n\}$ é definidos cada um por:

$$Y_h : X \rightarrow \Delta_h$$

$$X_i \in X \rightarrow Y_h(X_i)$$

Onde Δ_k é o domínio de chegada da variável Y_h.

Por conseguinte, os dados de agrupamento são descritos em uma matriz de Indivíduos-Variáveis em que cada célula da matriz contém a descrição de um indivíduo sobre uma das m variáveis. Esta matriz Indivíduos-Variáveis é geralmente uma matriz homogênea que pode ser do tipo quantitativo (onde todas as variáveis são quantitativas) ou qualitativo (onde todas as variáveis são qualitativas).

1. **Matriz de dados numéricos (contínuos ou discretos)**

   Existe uma variedade de medidas de similaridade, dissimilaridade e distância. Para a distância de dados de tipo quantitativo contínuo ou discreto, a distância mais usada é a distância de *Minkowski* de ordem ∞ definida em R^m pela raiz enésima da soma das diferenças absolutas entre os valores relativos aos elementos à enésima potência como segue:

   $$\forall X_i, X_j \in X; d(X_i, X_j) = \left(\sum_{h=1}^{m} \left| Y_h(X_i) - Y_h(X_j) \right|^{\infty} \right)^{\frac{1}{\infty}} \tag{2.2}$$

   onde $\infty \geq 1$, com se:

   - $\infty = 1$, d é a distância de *Manhattan* definida pela soma das diferenças absolutas entre os valores do elemento. Também conhecida como a distância de *City Block*.

     $$d(X_i, X_j) = \sum_{h=1}^{m} \left| Y_h(X_i) - Y_h(X_j) \right| \tag{2.3}$$

   - $\infty = 2$, d é a distância *Euclidiana* clássica definida pela raiz quadrada da soma dos quadrados das diferenças entre os valores relativos aos elementos. Este é o valor padrão para os dados de intervalo.

     $$d(X_i, X_j) = \sqrt{\sum_{h=1}^{m} \left(Y_h(X_i) - Y_h(X_j) \right)^2} \tag{2.4}$$

– $\infty = +\infty$, d é a distância de *Chebyshev* definida pela diferença máxima absoluta entre os valores relativos aos elementos, definida como segue:

$$d(X_i, X_j) = \max_{1 \leq h \leq m} \left| Y_h(X_i) - Y_h(X_j) \right| \tag{2.5}$$

Ainda que geralmente se use a distância euclidiana, a distância de Manhattan também por vezes é usada, especialmente para reduzir o efeito de largas diferenças devido aos pontos discrepantes (*outliers*), pois suas coordenadas não são elevadas ao quadrado. Deve-se notar que na maioria dos casos, a distância de Manhattan fornece resultados semelhantes aos da distância Euclidiana.

2. **Matriz de dados binários**

A matriz de dados binários consiste em classificar n indivíduos descritos por m variáveis binários codificados por 0 ou 1. A similaridade entre dois indivíduos X_i e X_j se calcula a partir de informações da matriz de confusão $2x2$ da Figura 2.1. Tal matriz permite contar o número de concordâncias $(a+d)$ e o número de discordâncias $(b+c)$ entre os indivíduos.

	1	0
1	a	b
0	c	d

Tabela 2.1: Matriz de confusão

Deve-se notar que o papel das modalidades de uma variável binária é muito importante no cálculo de uma medida da similaridade entre os indivíduos. Na prática, uma variável binária pode ser *simétrica* (as modalidades 0 e 1 desta variável têm a mesma importância) ou *assimétrica* (os dois termos não têm a mesma importância).

As medidas de similaridade (também conhecidas com *coeficiente de associação*) mais conhecidas entre dois indivíduos X_i e X_j são:

– Índice de Jaccard (JACCARD, 1908): este é um índice em que as ausências comuns não são consideradas. Também conhecido como a razão de similaridade, ele é definido para os valores assimétricos.

$$\forall X_i, X_j \in X; d(X_i, X_j) = \frac{a}{a+b+c} \tag{2.6}$$

– Índice de Sokal e Sneath (SOKAL; SNEATH, 1963): este é um índice com o qual as não-discordâncias são dobradas, e as ausências conjuntas não são consideradas.

$$\forall X_i, X_j \in X; d(X_i, X_j) = \frac{a}{a + 2 * (b + c)} \tag{2.7}$$

– Índice de Russel e Rao: esta é uma versão binária do produto interno (ponto). A mesma importância é dada para as concordâncias e discordâncias. Este é o valor padrão para os dados de similaridade binários. Além disso, à ausência conjunta não é considerada como uma similaridade:

$$\forall X_i, X_j \in X; d(X_i, X_j) = \frac{a}{a + b + c + d} \tag{2.8}$$

– Índice Simple Match de Sokal e Michener: esta é a razão entre as concordâncias e o número total de valores. A mesma importância é dada para as concordâncias e discordâncias. Além disso (SOKAL; MICHENER, 1958), está definido para as variáveis simétricas e a ausência conjunta é considerada como uma similaridade.

$$\forall X_i, X_j \in X; d(X_i, X_j) = \frac{a + d}{a + b + c + d} \tag{2.9}$$

Pode-se também usar a distância euclidiana para definir similaridade entre indivíduos de uma matriz de dados totalmente binários.

3. **Medida de similaridade entre variáveis aleatórias**

Se, em vez de analisar indivíduos, quer-se analisar variáveis, neste caso faz-se necessário definir um operador capaz de avaliar a proximidade entre essas variáveis de análise.

A incerteza sobre uma variável aleatória Y_i, ou um par de variáveis aleatórias (Y_i, Y_j) pode ser medida pela entropia denotada $H(Y_i)$ ou $H(Y_i, Y_j)$ respectivamente.

A quantidade denotada $I(Y_i : Y_j)$, chamada *informação mútua*, mede a informação transmitida entre (Y_i e Y_j).

$$I(Y_i : Y_j) = H(Y_i) + H(Y_j) - H(Y_i, Y_j) \tag{2.10}$$

Como pode-se mostrar que $I(Y_i : Y_j) \leq H(Y_i, Y_j)$, e que a independência entre Y_i e Y_j faz com que $I(Y_i : Y_j) = 0$ e que uma dependência determinista bijetiva entre Y_i e Y_j conduz a $I(Y_i : Y_j) = H(Y_i, Y_j)$, Dussauchoy (1982) propôs uma similaridade normalizada e uma dissimilaridade entre as variáveis aleatórias que são escritas da seguinte forma:

$$s(Y_i, Y_j) = \frac{I(Y_i : Y_j)}{H(Y_i, Y_j)} \tag{2.11}$$

$$d(Y_i, Y_j) = \frac{H(Y_i, Y_j) - I(Y_i : Y_j)}{H(Y_i, Y_j)} \tag{2.12}$$

Além disso, o autor, igualmente, generalizou essa noção de dissimilaridade para medir a dissimilaridade entre dois vetores aleatórios, e aplicou essas ideias à decomposição de sistemas complexos modelados usando vetores aleatórios.

Medida de similaridade entre indivíduos a descrições simbólicas

A seção anterior ilustra a diversidade das medidas existentes e a importância da escolha da distância ou similaridade no processo de agrupamento automático afim de não influenciar o seu desenvolvimento. A escolha da dissimilaridade / similaridade é facilitada quando na presença de um único tipo de dados. No entanto, em aplicações reais, é comum ter que lidar com diferentes tipos de dados chamados heterogêneos ou mistos. O que resulta em que o agrupamento refere-se a uma ”tabela de dados simbólicos ou complexos”que contém tanto as variáveis mono-valoradas (quantitativas ou qualitativas) e multi-valoradas.

Neste caso, as medidas usuais de proximidade entre dois indivíduos não são diretamente aplicáveis, é necessário desenvolver novas abordagens. Existem duas principais estratégias para resolver este problema:

- **Primeira abordagem (homogeneização da matriz de descrições)**: trata-se de transformar as variáveis para homogeneizá-las (de modo que tenha o mesmo tipo no final), em seguida, utilizar uma função de comparação global que leve em consideração todas as variáveis para calcular um índice de proximidade entre os indivíduos.

  No caso clássico, onde os indivíduos são descritos por variáveis mono-variadas (quantitativas e qualitativas), várias operações de conversão que permitem passar de um tipo para outro são definidas na literatura, mas estas transformações induzem uma perda de informações e uma distorção nos resultados. Por exemplo, uma variável nominal pode ser transformada em tantas variáveis binárias quantas modalidades ela tem. Uma variável quantitativa pode ser transformada em uma variável categórica ordinal efetuando uma divisão do domínio de observação (R ou N), utilizando os limites definidos pelo usuário, atribuindo a cada indivíduo o número da de decisão a que pertence.

Uma vez realizada tais transformações q, a matriz de dados torna-se homogênea e encontrar-se-á no quadro clássico da seção anterior.

No caso *simbólico* onde os indivíduos são descritos tanto pelas variáveis mono-valoradas (quantitativas e qualitativas) e multi-valoradas não se trata de transformar as descrições multi-valoradas em uma modalidade única a fim de não perder a informação contida nestas descrições, mas em vez disso passar de uma descrição mono-valorada para uma descrição multi-valorada. Chavent (1997) exemplifica um valor v visto como o intervalo $[v,v]$, no caso de uma variável quantitativa ou o conjunto $\{v\}$, no caso de uma variável qualitativa (CHAVENT, 1997). Essas diferentes transformações permitem que se obtenha uma matriz homogênea, onde todas as descrições são multi-valoradas. Resta definir um índice global de proximidade que leva em conta todas estas descrições para medir a similaridade entre indivíduos.

- **Segunda abordagem: agregação de comparações sobre as variáveis**: trata-se de uma abordagem geralmente utilizada para comparar dois indivíduos a descrições simbólicas. Esta abordagem não requer qualquer transformação prévia da matriz de dados. O seu princípio é como segue:

  – Definir para cada variável Y_h do conjunto das m variáveis características $Y = \{Y_1, Y_2, ..., Y_m\}$ uma função de comparação g_h

  – Utilizar a função de agregação proposta por (ICHINO; YAGUCHI, 1994) com base na *métrica de Minkowski* para combinar as diferentes comparações obtidas em cada variável para o mesmo grau de similaridade.

  Esta é a abordagem que se está interessado neste trabalho.

Função de comparação entre descrições mono-valoradas

Uma variável a descrição mono-valorada pode ser quantitativa ou qualitativa. As funções de comparação elementares mais utilizadas para os dados mono-valorados são os seguintes:

– Para uma variável quantitativa Y_h:

$$\forall X_i, X_j \in X; g_h\left(Y_h\left(X_i\right), Y_h\left(X_j\right)\right) = \left[Y_h\left(X_i\right) - Y_h\left(Y_j\right)\right] \tag{2.13}$$

A função de comparação pode ser normalizada dividindo-a por um coeficiente m_h de normalização calculando o desvio máximo da variável Y_k. Este é definido como:

$$m_h = \underset{x_i \in X}{max} Y_h\left(X_i\right) - \underset{x_i \in X}{min} Y_h\left(X_i\right) \tag{2.14}$$

– Para uma variável qualitativa Y_h:

$$\forall X_i, X_j \in X; g_h\left(Y_h\left(X_i\right), Y_h\left(X_j\right)\right) = \begin{cases} 0, & Y_h(X_i) = Y_h(X_j) \\ 1, & Y_h(X_i) \neq Y_h(X_j) \end{cases} \tag{2.15}$$

Função de comparação entre descrições multi-valoradas

Como mencionado anteriormente, nesta deste trabalho tem-se interesse em caso de variáveis multi-valoradas cujo domínio de observação O é nominal (ou seja, domínio de chegada Δ é um conjunto de modalidades). Dada a descrição destas variáveis, pode-se então definir algumas funções de comparações entre conjuntos.

Para uma lista de funções de comparação a se usar no caso de variáveis a descrições multi-valoradas (intervalo de valores) e modais, poder-se-á encontrar mais detalhes em (MALERBA et al., 2001; BOCK; DIDAY, 2000).

a. ***A distância de jaccard***

O índice e a distância de Jaccard são duas métricas usadas em estatística para comparar a similaridade e a diversidade entre as amostras. Eles são nomeados a partir do nome do botânico suíço Paul Jaccard.

O índice de Jaccard (ou coeficiente de Jaccard) é a relação entre o cardinal (tamanho) da interseção de conjuntos considerados e o cardinal da união dos conjuntos. Ele permite avaliar a similaridade entre os conjuntos.

Uma lista de extensões de índices de similaridade entre dados binários foi proposto em (CARVALHO, 1994) para calcular a similaridade entre conjuntos de valores. O mais usado e mais simples para calcular é o *índice de Jaccard*.

Sejam duas descrições $Y_h(X_i)$ e $Y_h(X_j)$ correspondentes à variável Y_h para os dois indivíduos (X_i) e (X_j) respectivamente, o índice de similaridade de Jaccard é definido, no caso simbólico, como:

$$\forall X_i, X_j \in X; g_h\left(Y_h\left(X_i\right), Y_h\left(X_j\right)\right) = \frac{\left|\left(Y_h\left(X_i\right) \cap Y_h\left(X_j\right)\right)\right|}{\left|\left(Y_h\left(X_i\right) \cup Y_h\left(X_j\right)\right)\right|} \tag{2.16}$$

Onde $||$ é o cardinal.

A ideia subjacente é que duas descrições se assemelham tanto quando sua intersecção é importante e sua união reduzida. Esta similaridade torna possível obter o valor máximo de 1, quando os dois conjuntos são idênticos, e o valor mínimo (0) se eles forem completamente separados.

b. ***A dissimilaridade de Gowda e Diday***

Gowda e Diday (1991) definem uma dissimilaridade entre dois conjuntos de modalidades levando-se em conta as similaridades entre os seus valores comuns e também as similaridades de *cardinal*, com base nas duas funções que seguem:

$$\forall X_i, X_j \in X; D_s\left(Y_h\left(X_i\right), Y_h\left(X_j\right)\right) = \frac{\left|l_i - l_j\right|}{l_s} \tag{2.17}$$

$$\forall X_i, X_j \in X; D_c\left(Y_h\left(X_i\right), Y_h\left(X_j\right)\right) = \frac{l_i + l_j - 2 * \left|\left(Y_h\left(X_i\right) \cap Y_h\left(X_j\right)\right)\right|}{l_s} \tag{2.18}$$

Onde

- l_i é o número de valores em $Y_h(X_i) = |Y_h(X_i)|$.
- l_j é o número de valores em $Y_h(X_j) = |Y_h(X_j)|$.
- l_s é o número de valores em $Y_h(X_i) \cup Y_h(X_j) = |Y_h(X_i) + Y_h(X_j)| - |Y_h(X_i) \cap Y_h(X_j)|$

Enquanto a função D_s compara o *cardinal* dos dois conjuntos de valores, a função D_c compara seus conteúdo.

Com base nas duas funções, a dissimilaridade de Gowda e Diday entre as duas descrições $Y_h(X_i)$ e $Y_h(X_j)$ é dada pela formula a seguir:

$$\forall X_i, X_j \in X; g_h\left(Y_h\left(X_i\right), Y_h\left(X_j\right)\right) = D_s\left(Y_h\left(X_i\right), Y_h\left(X_j\right)\right) + D_c\left(Y_h\left(X_i\right), Y_h\left(X_j\right)\right) \tag{2.19}$$

ou

$$\forall X_i, X_j \in X; g_h\left(Y_h\left(X_i\right), Y_h\left(X_j\right)\right) = \left(D_s + D_c\right)\left(Y_h\left(X_i\right), Y_h\left(X_j\right)\right)$$

Os autores também definiram uma função de comparação para medir a similaridade entre dois intervalos de valores (GOWDA; DIDAY, 1991).

2.1.4 As técnicas clássicas de agrupamento automático

Entre os métodos de estatística exploratória multidimensionais, cujo objetivo é extrair informação útil a partir de uma massa de dados de ”informações úteis”, distinguem-se métodos de análise fatorial (AF) dos métodos de agrupamento. O objetivo dos métodos de AF é entre outros a visualização de dados, a redução do número de variáveis. O objetivo do agrupamento automático é formar grupos de indivíduos ou variáveis para estruturar um conjunto de dados.

Outrossim, o agrupamento é um método matemático para a análise de dados: para facilitar o estudo de uma população de trabalho importante (animais, plantas doentes, genes, etc ...), grupam-nos em classes de modo que os indivíduos da mesma classe sejam tão similares quanto possíveis e as classes sejam mais distintas possíveis. Para isso existem várias maneiras de se fazer (o que pode levar a resultados diferentes).

O agrupamento automático por sua vez vem sendo usado há muito tempo em contextos variados por pesquisadores de diferentes disciplinas, como processo de análise exploratória dos dados. Foi a causa de inúmeros trabalhos teóricos e aplicados, e ainda é o foco de revistas especializadas e comunidades ativas em todo o mundo.

Os métodos de agrupamento automático de um conjunto de indivíduos podem ser divididos em duas principais famílias: as abordagens hierárquicas e abordagens *partitioning* (JAIN et al., 1999; BERKHIN, 2002). As *abordagens hierárquicas*, que produzem uma sequência de partições aninhados de heterogeneidades, conduzem a resultados sob forma de *árvore hierárquica indexada* também conhecida como o dendrograma, que visualiza este sistema de classes organizadas por inclusão. Ao contrário de abordagens hierárquicas, abordagens por *partitioning* (ou particionamento) buscam a melhor partição em k classes disjuntas de dados, o número de classes (*clusters* ou *grupos*) com k sendo fixado a priori. As abordagens por *particionamento* utilizam um processo iterativo com base no número k que consiste em atribuir cada indivíduo à classe mais próxima no sentido de uma distância, ou índice de similaridade, otimizando alguma função objetivo.

Abordagens hierárquicas de agrupamento

Ao se tratar do agrupamento automático de dados, o conceito de agrupamento hierárquico abrange diferentes métodos de *clustering*, ou seja, de agrupamento por algoritmo de agrupamento. A construção de um agrupamento hierárquico pode ser feita de duas maneiras: em primeiro lugar, a partir de uma matriz simétrica de similaridade entre os indivíduos, um *algoritmo de aglomeração* inicialmente forma classes pequenas que contenham apenas indivíduos muito parecidos, então, a partir destes, construiu classes menos homogêneas, até que se obtenha a classe inteira. Esta construção é chamada de *agrupamento Hierárquico Ascendente* (CHA). A segunda forma de construir um agrupamento hierárquico inversa o processo anterior. Baseia-se em um algoritmo de *divisão* com um critério de divisão de um subconjunto de variáveis, que procede por dicotomias sucessivas do conjunto inteiro de todos os indivíduos até um nível que satisfaça algumas de parada e cujos elementos constituem uma partição do conjunto de indivíduos a serem classificados. Esta construção é chamado de *agrupamento Hierárquico*

Descendente (CHD). Outrossim, Diday (1986) propôs uma outra forma de construção como uma generalização de modelos hierárquicos, e é chamado a *agrupamento piramidal*. Assim como as hierárquicas, as representações piramidais são conjuntos de partes também chamadas *classes* ou *níveis* do conjunto de todos os indivíduos a classificar. No entanto, a representação piramidal é uma estrutura de dados mais complexa. de fato, contrariamente ao caso hierárquico convencional, duas classes da pirâmide podem ter uma interseção não vazia, e por isso alguns indivíduos a classificar, podem pertencer a duas classes que não estão aninhadas uma dentro da outra. A hierarquia obtida neste caso é dito *hierarquia de recobrimento* (ou *pirâmide*). Como parte deste trabalho, tem-se interesse apenas nos casos em que os indivíduos pertencem a uma classe (partição). Portanto, a seguir detalhar-se-á as abordagens hierárquicas convencionais que leve a uma hierarquia de partições. Sem embargo, pode-se encontrar no trabalho (BERTRAND; DIDAY, 1990) devidos detalhes sobre a abordagem piramidal.

1. **agrupamento Hierárquico Ascendente**

   A CHA permite construir uma hierarquia de objetos, sob a forma de ”árvore”em ordem ascendente. Começa-se por considerar cada indivíduo como uma classe e tenta-se mesclar duas ou mais classes apropriadas (de acordo com uma similaridade) para formar uma nova classe. O processo é iterado até que todos os indivíduos se encontrem numa mesma classe. Este agrupamento gera uma árvore que pode ser cortada em diferentes níveis para se obter um número de classes mais ou menos maior.

   Diferentes medidas de distância entre as classes podem ser usadas: a distância euclidiana, a distância inferior (que favorece a criação de classes de baixa inércia) ou maior distância (que favorece a criação de classes de maiores inércia) etc ...

   No caso do agrupamento hierárquico, a partir dos elementos, forma-se classes pequenas que contêm apenas os indivíduos muito semelhantes, em seguida, a partir destes, constrói-se classes menos homogêneas até obter a classe inteira.

   Para mais detalhes recomenda-se ler (SAPORTA, 2006; BENZECRI, 1973; JAMBU, ; CELEUX et al., 1989)

   O esquema de um algoritmo de agrupamento Hierárquico Ascendente (CHA) (ELGHAZEL, 2007) é como segue:

   1. As classes iniciais são os próprios indivíduos
   2. Calcula-se as distâncias entre as classes
   3. As duas classes mais próximas são mescladas e substituídas por uma só.

4. O processo repete o passo 2 até ter apenas uma classe, que contém todas as observações.

Um algoritmo de *aglomeração* funciona assim, buscando para cada etapa as classes mais próximas para mesclá-las, e o passo mais importante no algoritmo é a escolha da distância entre duas classes. os algoritmos mais convencionais definem a distância entre duas classes a partir da medida de dissimilaridade entre objetos que constituem cada grupo.

Este procedimento baseia-se em duas escolhas:

– A determinação de um critério de similaridade entre os indivíduos. O método dá ao usuário a escolha de dissimilaridade.

– Determinar uma dissimilaridade entre classes: um processo chamado critério de agregação. O método dá ao usuário a escolha deste critério.

Critérios de agregação

Muitos critérios de agregação foram propostos e os mais conhecidos são:

- **O critério de salto mínima**
  A distância entre 2 classes C_1 e C_2 é definida pela menor distância que separa um indivíduo de C_1 e um indivíduo de C_2.

$$D(C_1,C_2) = min\left(\{d(x,y)\}, x \in C_1, y \in C_2\right) \tag{2.20}$$

- **O critério de salto máxima**
  A distância entre 2 classes C_1 e C_2 é definida pela maior distância que separa um indivíduo de C_1 e um indivíduo de C_2.

$$D(C_1,C_2) = max\left(\{d(x,y)\}, x \in C_1, y \in C_2\right) \tag{2.21}$$

- **O critério da média**
  Este critério consiste em calcular a distância média entre todos os elementos de C_1 e todos os elementos de C_2.

$$D(C_1,C_2) = \frac{1}{nc_1 nc_2} \sum_{x \in C_1} \sum_{y \in C_2} d(x,y) \tag{2.22}$$

  Com:
  - nc_1 : o cardinal de C_1
  - nc_2 : o cardinal de C_2

- **O critério de Ward**

O critério de Ward consiste em escolher em cada fase o agrupamento de classes tal que o aumento da inércia intra-classe seja mínima.

$$D(C_1, C_2) = \frac{nc_1 nc_2}{nc_1 + nc_2} d^2(gc_1, gc_2) \tag{2.23}$$

Com:

- gc_1 : o centro de gravidade de C_1

- gc_2 : o centro de gravidade de C_2

- **O critério de centros de gravidades**

A distância entre 2 classes C_1 e C_2 é definida pela maior distância entre seus centros de gravidade.

$$D(C_1, C_2) = d(gc_1, gc_2) \tag{2.24}$$

A dificuldade em escolher o critério de agregação é o fato de que esses critérios podem levar a resultados diferentes. De acordo com a literatura, critério mais utilizado é o de Ward.

Obervação: Ao se tratar da estratégia de salto mínimo, quanto mais elementos uma classe contém, mais atraente é para os elementos isolados. Pelo contrário, no caso de uma estratégia de salto máximo, quanto mais elementos uma classe contém, menos atraente é para elementos isolados.

Abaixo algumas possíveis distâncias.

– O algoritmo *Single-linkage* em que a distância entre dois clusters é representada pela distância mínima entre todos os pares de dados entre os dois clusters (par composto de um elemento de cada cluster), fala-se então de *salto mínimo*. A vantagem desta abordagem é que ela sabe detectar as classes alongadas, mas o seu ponto fraco é que é sensível ao efeito de cadeia[1] (TUFFÉRY, 2012) e, portanto, menos adaptada para detectar as classes esféricas.

– O algoritmo *Complete-linkage* em que a distância entre dois clusters é representada pela distância máxima entre todos os pares de dados dos dois clusters, fala-se então de *salto máximo* ou critério de diâmetro. Por definição, esta abordagem é muito sensível a *outliers* e, portanto, é pouco utilizada (TUFFÉRY, 2012).

– O algoritmo *Average-linkage* propõe o cálculo da distância entre dois clusters, tendo o valor médio das distâncias entre todos os pares de objetos de dois clusters. Fala-se

[1] Chama-se efeito de cadeia quando dois pontos distantes entre si, mas ligados por uma série de pontos muito próximos uns dos outros estão incluídos na mesma classe.

também de *salto médio*. Esta abordagem tende a produzir as classes com a mesma variância.

– O algoritmo *Centroid-linkage* (ou *salto baricêntrico*) define, por sua vez, a distância entre dois clusters como a distância entre os seus *centros de gravidade*. Tal método é o mais robusto ao se tratar de *outliers*. No entanto, é limitado aos dados numéricos quantitativos para os quais é possível o cálculo do centro de gravidade.

Estes métodos convencionais são interessantes no sentido em que a maior parte deles é baseada em uma ligação métrica que os torna aplicáveis a qualquer tipo de dados, desde que haja a possibilidade de construir uma matriz de similaridade entre os indivíduos a classificar.

Nas últimas décadas, novos métodos de agrupamento hierárquico foram desenvolvidos para evitar a maioria dos problemas relacionados aos algoritmos sobrejacentes, fornecendo partições das classes de forma e tamanho arbitrários. Entre estes são CURE (*Clustering Using Representatives*) (GUHA et al., 1998), CHAMELEON (KARYPIS et al., 1999), ROCK (FISHER, 1987b).

A principal característica do algoritmo CURE é de propor um método original de indexação de clusters. Em vez de indexar um cluster pelo conjunto de objetos que o compõem (indexação gráfica) ou por um centroide (indexação geométrica), CURE determina um número constante de objetos representativos deste cluster.

CHAMELEON procede em dois passos: o primeiro passo consiste na partição do conjunto de objetos em pequenos clusters homogêneos, em seguida uma fase de aglomerativa permite alcançar uma hierarquia. Para isso, CHAMELEON utiliza um formalismo de representação baseado em grafos.

Finalmente os métodos como Rock (*RObust Clustering using linKs*) e COBWEB (FISHER, 1987a) constroem uma árvore a partir de dados descritos por variáveis qualitativas.

De fato, as abordagens de agrupamento hierárquico acima foram propostas para resolver os problemas associados com as abordagens convencionais, proporcionando assim partições de classes de forma e tamanho arbitrários. Todavia, a fiabilidade e a eficácia destas abordagens dependem fortemente da escolha da *amostra representativa* (para CURE e ROCK) e *parâmetros de controles* (CHAMELEON), e esta escolha nem sempre é óbvio em relação aos dados a classificar. Um apoio à decisão às vezes é necessário, o qual pode ser baseado, por exemplo, em uma abordagem mais abrangente com base na avaliação das partições obtidas em função da escolha.

As vantagens e desvantagens de CHA

- *Vantagem:* fácil de se implementar

-*Desvantagem*: método muito custoso, complexidade em $O(n^2)$.

2. **agrupamento Hierárquico Descendente**

O agrupamento hierárquico descendente (CHD) procede da maneira oposta do agrupamento hierárquico ascendente: ela consiste em dividir o conjunto em grupos homogêneos que, por sua vez, são subdivididos para chegar às n unidades elementares.

Igualmente, ao contrário de agrupamento hierárquico, em cada passo do algoritmo existem dois processos para:

1. Buscar uma classes para dividir.
2. Escolher um modo de atribuição de objetos às subclasses.

dentre os algoritmos mais antigos, o algoritmo *Williams* e *Lambert* (WILLIAMS; LAMBERT, 1959) divide a maior classe em duas classes. *Hubert* propôs dividir a classe do maior diâmetro. Nenhum dos dois justificou sua escolha de divisão.

Este método de agrupamento constrói a sua hierarquia no sentido inverso, começando com uma classe maior contendo todos os objetos. A cada passo, ele divide um classe em duas classes menores até que todas as classes contenham apenas um indivíduo. Isto significa que, para n indivíduos, a hierarquia está construída em $N-1$ passos.

No primeiro passo, os dados são divididos em duas classes, por meio de dissimilaridades. Em cada um dos seguintes passos, a classe com o maior diâmetro é dividido da mesma forma. Após $n-1$ divisões, todos os indivíduos são bem separados. A dissimilaridade média entre o indivíduo "x"que pertencia à classe C, que contém n indivíduos e todos os outros indivíduos da classe C é definida como:

$$d_x = \frac{1}{n-1} \sum_{x \in C, y \neq x} d(x,y) \tag{2.25}$$

As vantagens e desvantagens de CHD

Vantagem

Em comparação com a maioria dos algoritmos de agrupamento automático, o algoritmo de agrupamento hierárquico descendente não requer a utilização de um limiar arbitrário para a formação das classes que podem, eventualmente, levar à busca de uma partição, numa direção não realista.

Desvantagem

Os resultados são geralmente densos, os níveis de nós de hierarquia não são mais definidos senão pela ordem em que aparecem.

3. **Abordagem simbólica de agrupamento hierárquico**

Em 2003, uma abordagem simbólica de agrupamento hierárquico foi proposta por Mali e Mitra (MALI; MITRA, 2003). Ela segue o mesmo princípio de funcionamento que as abordagens convencionais, mas difere pelo critério de agregação que ela usa. De facto, define a distância entre duas classes C_i e C_j como segue:

$$d_{agr}(C_i,C_j) = \frac{\sum_{u=1}^{|C_i|}\sum_{q=1}^{|C_j|}}{|C_i|\,|C_j|}\left(\frac{|C_i|.|C_j|}{|C_i|+|C_j|}\right)^{\frac{1}{2}} \tag{2.26}$$

Onde d representa a medida de dissimilaridade de *Gowda e Diday* definida sobre o conjunto de indivíduos X e $|C_i|$ o cardinal (número de indivíduos) da classe $|C_i|$.

Notar-se-á que o termo "ponderação"frequentemente utilizada por esta distância assume o valor de $\sqrt{50}$ (para $|C_i| = |C_j| = 100$), de 101/100 (para $|C_i| = 1$) e $\sqrt{0.5}$ (para $|C_i| = |C_j| = 1$). A distância de agregação terá, portanto, grandes valores para clusters de de maior tamanho e pequenos valores para clusters menores. Por conseguinte, a abordagem de agrupamento hierárquico tende a favorecer a fusão das classes *singletons*, ou pequenas e grandes classes, em detrimento da função das classes medianas.

4. **Conclusão**

Estruturas hierárquicas por sucessivas divisões têm um aspecto interessante: Elas começam pelo topo da árvore, ou seja, a parte em que tem-se essencialmente a interpretação. Nada obstante, simplificações drásticas de que necessitam, para manter o tempo de cálculo razoável, muitas vezes tornam os resultados decepcionantes. Entretanto, as dicotomias baseadas em variáveis bem escolhidas têm a vantagem de ser rápidas de proporcionar interpretações fáceis. Assim, elas permitem de tratar facilmente grandes conjuntos de dados com poucas variáveis.

As vantagens e desvantagens de métodos hierárquicos

Vantagens

- Flexibilidade em relação ao nível de granularidade,
- Facilidade de lidar com qualquer tipo de similaridade ou distância,

- Aplicabilidade a qualquer tipo de atributo.

Desvantagens

- A dificuldade de escolher a a reta que define os critérios,
- A maioria dos algoritmos hierárquicos não revisam as classes (intermediário) uma vez que elas são construídas.

Abordagens por particionamento

Ao contrário dos métodos de classificações hierárquicas que constroem as classes gradualmente, os algoritmos de particionamento constroem diretamente uma partição do conjunto de indivíduos em k classes. Em geral, por definição de uma partição, isto significa que cada classe deve conter pelo menos um indivíduo, e que cada indivíduo deve pertencer a uma única classe (no entanto, os algoritmos *"fuzzy"* não impõem essa condição). Para se fazer, dado o número k de classes requiridas, estes algoritmos geram uma partição inicial, e, em seguida, procuram melhorá-la realocando os indivíduos de uma classe para outra. Sem dúvida, não é facilmente possível listar todas as possíveis partições.

Esses algoritmos, portanto, procuram os máximos locais por meio da otimização de uma função objetiva ($\boldsymbol{F}$ abaixo) que determina que os objetos devem ser "similares"dentro da mesma classe, e "dissimilares"de uma classe para outra.

$$F = \sum_{i=1}^{k} \sum_{x \in C_1} d(x, m_i) \tag{2.27}$$

com:

- m_i: os centroides atuais (correntes)
- x: objeto a classificar

Os algoritmos de particionamento são divididos em três sub-famílias: os métodos *k-means*, os métodos *k-medoids* e os métodos de *dynamic clustering*, de acordo com a definição de representantes da classe. As duas primeiras famílias destes métodos tendem a construir as classes convexas sem levar em consideração os pontos discrepantes (aberrantes). Nessas famílias de algoritmos, os mais utilizados são: *centros móveis (k-means)*, PAM, CLARA e CLARANS. Além disso, estas duas famílias de abordagens se enunciam como variantes (casos especiais) do método *dynamic cluestering* que fornece, por sua vez, diversas categorias de modos de representação (chamados núcleos), de acordo o objetivo de análise desejada.

1. **Métodos *k-means***

O método de centros móveis (FORGY, 1965) é o mais clássico e muito utilizado. Procede da seguinte forma: constrói k classes a partir de um conjunto de n indivíduos, maximizando-se:

$$\sum_{i=1}^{k} \sum_{x \in C_r} d(x_i, g_r)^2$$

com:

- C_r: a classe número r
- x_i: um indivíduo em uma classe
- g_r: o centro de classe C_r.

Esses métodos tem a lógica algorítmica a seguir:

Dados: k o número máximo de classes desejado.

Início

(1) Escolher k indivíduos aleatoriamente (como centro de classes iniciais)

(2) Atribuir cada indivíduo ao centro mais próximo

(3) recalcular o centro de cada uma destas classes

(4) Repetir o passo (2) e (3) até que os centros se estabilizam

(5) Editar a partição obtida.

FIM

Na prática, os métodos de centro móvel procuram minimizar a *inércia intra-classe* definida pela soma dos desvios dos centroides aos pontos das suas classes e, por conseguinte, maximizar também a *inércia inter-classe* da partição dada pela soma das diferenças entre o centroide das classes e o centroide da população total (pelo teorema de *Huygens*: inércia total = intra-classe + inércia inter-classe). Minimizando a *inércia intra-classe*, o método dos centros móveis tendem a buscar as classes esféricas de igual volume e de baixa inércia(CELEUX et al., 1989).

Com abordagem *k-means*, os centros são recalculados após cada atribuição de um indivíduo a uma classe, ao invés de esperar para a atribuição de todos os indivíduos antes de atualizar os centros. Este abordagem geralmente leva a melhores resultados do que os métodos de centros móveis e convergência, e também é mais rápido (MACQUEEN et al., 1967; HARTIGAN; WONG, 1979).

Esses algoritmos tem como vantagens, antes de tudo, sua simplicidade, mas também sua complexidade computacional que é razoável. No entanto, estes métodos sofrem de certos inconvenientes: por um lado, o cálculo da média que eles utilizam é muito sensível aos

dados discrepantes *outliers* e restringe sua aplicação a dados numéricos. Por outro lado, o resultado final obtido é altamente dependente da escolha dos centros iniciais.

2. **Métodos *k-medoids***

   Em métodos *k-medoids* uma classe é representada por um dos seus indivíduos *medoid*. É um método iterativo que combina a reafetação dos indivíduos nas classes com intervenção de *medoids* e de outros indivíduos. É um método simples porque abrange qualquer tipo de variáveis. Quando os *medoids* são escolhidos, as classes são definidas como subconjuntos dos indivíduos ao redor dos *medoids* mais próximos para uma medida de distância escolhida.

   Em comparação como os métodos anteriormente apresentados, a principal diferença reside na escolha de um representante de uma classe. Os métodos *k-medoids* têm a vantagem de ser aplicável a qualquer tipo de dados, e são geralmente mais robusto a *outliers* do que os métodos *k-means*, em vista que recorrem às medianas *(medoids)*, em vez dos *(centroids)* para estimar a distância aos centros.

3. **Métodos de agrupamentos dinâmicos (*dynamic clustering*)**

   Desenvolvido em grande parte por (DIDAY, 1971), método *dynamic clustering* (clusters dinâmicos) difere, principalmente, das abordagens anteriores pelo modo de representação de classe também chamado de núcleo. Este pode ser o seu centro de gravidade (neste caso, refere-se à abordagem dos centros móveis), um conjunto de indivíduos (caso da abordagem *k-medoids* com um único indivíduo), a uma distância (a abordagem de distâncias adaptativas (DIDAY; GOVAERT, 1974)), uma distribuição de probabilidade (a decomposição das misturas (SCHROEDER, 1976)), etc.

   O algoritmo de clusters dinâmicos busca otimizar um critério objetivo medindo o ajuste (adequação) entre uma partição e um modo de representação daquela classe partição. Na prática, o algoritmo converge quando este critério a ser otimizado cessa de diminuir significativamente, ou quando um determinado número de iterações é alcançado. Celeux et al. (1989) afirmam que *"o problema de otimização, neste caso surge em termo de busca simultânea do agrupamento e representação das classes desse mesmo agrupamento dentre um conjunto de classificações e representações possíveis, que minimizam o critério estabelecido"*. Para minimizar este critério, o algoritmo de clusters dinâmicos utiliza principalmente um passo de representação seguido por um passo de atribuição de forma iterativa até a convergência que dá uma solução localmente ótima para o problema dado.

   Assim como as abordagens anteriores de agrupamento automático por particionamento,

o método de clusters dinâmicos fornece uma solução dependente da configuração inicial, geralmente feita por acaso.

4. Conclusão

   Observa-se que na maioria dos casos, as k classes encontradas por uma abordagem de agrupamento automático por particionamento são de melhor qualidade do que aquelas geradas por uma abordagem hierárquica (NG; HAN, 2002). No entanto, os algoritmos de agrupamento por particionamento sofrem pelo fato de que usam ao todo um único ponto como representante de uma classe (o problema do *"representante único"*). De fato, como o objetivo destes algoritmos é encontrar classes que reduzem ao mínimo uma função objetiva igual à soma dos quadrados das distâncias para os núcleos de classes, eles falham para certos conjuntos de dados em que alguns indivíduos são mais próximos do núcleo de outra classe do que do núcleo de sua própria classe. Por conseguinte, eles não podem capturar facilmente as classes com formas arbitrárias (Figura 2.1) ou com tamanhos muito diferentes (Figura 2.2).

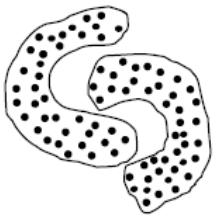

Figura 2.1: Conjuntos de dados para os quais as abordagens *k-means* e *k-medoids* falham: Classes de forma arbitrária (KARYPIS et al., 1999).

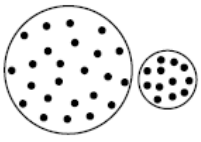

Figura 2.2: Conjuntos de dados para os quais as abordagens *k-means* e *k-medoids* falham: Classes de tamanhos diferentes (KARYPIS et al., 1999).

2.1.5 Avaliação das abordagens de agrupamento

O objetivo das abordagens de agrupamento automático é produzir classes com uma coesão máxima (*similaridade intra-classe*) ao realizar uma separação máxima (*dissimilaridade inter-classes*) entre as classes da partição obtida.

Ao observar ou analisar as técnicas de agrupamento hierárquico, percebe-se que o corte do dendrograma por uma reta horizontal fornece uma partição de todos os indivíduos a classificar.

O número de classes de partição é definido pelo nível deste corte, o que nem sempre é fácil de determinar. Para as técnicas de agrupamento por particionamento, o número de classes a descobrir deve ser fixado a priori, o que nem sempre é o caso dos conjuntos de dados. É por isso que, geralmente, fixa-se vários valores para o número k de classes, e os particionamentos correspondentes, são comparados em seguida.

Observa-se, de fato, que um certo número de abordagens requer a avaliação da qualidade das partições de dados. No entanto, a análise e comparação das partições não é um processo imediato. Assim como existem muitos algoritmos de agrupamento automático, a literatura dispõem de critérios relacionados com a qualidade de *clustering*. Forcar-se-á, nesta seção, apenas nos chamados critérios *internos*, ou seja, baseados nos dados e similaridades entre eles (similaridade / dissimilaridade). Os critérios *externos*, por sua vez, são baseados em informações externas, tais como o *label* da classe ou a opinião do especialista. Ao se usar os critérios internos, o problema de agrupamento automático é então considerado como um problema de optimização, cujo desempenho pode ser determinado.

Muitos procedimentos, também chamados de *índices de validade de clustering*, têm sido propostos na literatura, a fim de determinar a melhor partição de um conjunto de dados numéricos (BEZDEK; PAL, 1998). Estes procedimentos são bem adequados para o caso simbólico pelo fato de que necessitam apenas da definição de uma medida de dissimilaridade para funcionar (MALI; MITRA, 2003).

No que concerne à apresentação dos diferentes índices, faz-se necessário descrever algumas notações preliminares. Seja P uma partição com k classes $\{C_1,\ldots,C_k\}$ do conjunto de indivíduos $X = \{X_1,\ldots,X_n\}$, a *dispersão intra-classe* $s_a(C_i)$ e a separação *inter-classe* $d_a(C_i,C_j)$ são dadas pelas seguintes formulas:

$$\forall C_i \in P; s_a\left(C_i\right) = \frac{1}{|C_i|\left(|C_i|-1\right)} \sum_{u=1}^{|C_i|} \sum_{q=1}^{|C_i|} d\left(X_u, X_q\right) \tag{2.28}$$

$$\forall C_i, C_j \in P; d_a\left(C_i, C_j\right) = \frac{1}{|C_i|\,|C_i|} \sum_{u=1}^{|C_i|} \sum_{q=1}^{|C_j|} d\left(X_u, X_q\right) \tag{2.29}$$

Onde d representa a medida de dissimilaridade definida sobre o conjunto de indivíduos X e C_i o cardinal (número de indivíduos) da classe C_i.

Índice de Davies-Bouldin

O índice de *Davies-Bouldin* (BEZDEK; PAL, 1998) baseia-se na maximização da razão entre as *dispersões intra-classes* e a *separação inter-classe*. Este é calculado como segue:

$$DB(P) = \frac{1}{k}\sum_{i=1}^{k} \underset{1\leq j\leq k}{max} \left\{ \frac{s_a(C_i) + s_a(C_j)}{d(C_i, C_j)} \right\}; j \neq i \tag{2.30}$$

Observa-se, portanto, que quanto mais compactas e distintas forem as classes umas das outras, menor será a razão. Consequentemente, a partição da melhor qualidade será a que minimizar o índice de *Davies-Bouldin*.

Índice de Dunn

O índice de Dunn (DUNN, 1973) baseia-se na identificação de clusters compactos e bem separados. Ele é definido pela razão entre a menor *dissimilaridade inter-classe* d_min (isto é, entre dois indivíduos de duas classes diferentes) e a maior *dissimilaridade intra=classe* s_max (isto é, entre dois indivíduos da mesma classe).

$$Dunn(P) = \frac{d_{min}}{s_{max}} \tag{2.31}$$

Este índice tem como objetivo principal maximizar a *dissimilaridade interclasse* e minimizar a *dissimilaridade intra-classe*. Portanto, o objetivo consiste em maximizar o índice de *Dunn*.

Índice de Silhouette

Definido por (ROUSSEEUW, 1987), o índice de silhouette para todo indivíduo X_i do conjunto X é representado pela seguinte formula:

$$\forall X_i \in X; s(X_i) = \frac{b(X_i) - a(X_i)}{max(a(X_i), b(X_i))} \tag{2.32}$$

Onde:

- $a(X_i)$ é a similaridade média entre o indivíduo X_i e todos os outros indivíduos da classe à

qual pertence $C(X_i)$.

$$\forall X_i \in X; a(X_i) = \frac{1}{|C(X_i)| - 1} \sum_{\substack{X_i \in C(X_i) \\ X_i \neq X_j}} d\left(X_i, X_j\right) \tag{2.33}$$

- $b(X_i)$ é o mínimo das dissimilaridades médias entre o indivíduo X_i e todos os outros indivíduos da classe da partição P diferente de $C(X_i)$.

$$\forall X_i \in X; b(X_i) = \max_{\substack{C \in P \\ C \neq C(X_i)}} d\left(X_i, X_j\right) \tag{2.34}$$

onde

$$d(X_i, C) = \frac{1}{|C|} \sum_{X_j \in C} d\left(X_i, X_j\right)$$

notar-se-á que o índice de silhouette é limitado: $-1 \leq s(X_i) \leq 1$. Ademais, quando $s(X_i)$ é próximo de 1, X_1 é dito bem classificado em $C(X_i)$. Quanto $s(X_i)$ é próximo de 0, então X_i se situa entre duas classes. Por fim, se $s(X_i)$ é próximo de -1, X_i é dito mal classificado em $C(X_i)$ e deve ser anexado a um outro cluster mais próximo.

De fato, cada classe também é representada por uma *silhouette* que mostra quais objetos são corretamente classificados dentro desta classe e quais tem apenas uma posição intermédia. Para uma determinada classe C_i, o seu índice de silhouette é definido pela média dos índices de silhouette dos indivíduos que nela pertencem:

$$\forall C_i \in P; s(C_i) = \frac{\sum_{X_j \in C_i} s\left(X_j\right)}{|C_i|} \tag{2.35}$$

O índice de silhouette global da partição P é dado pela média global da largura de silhouettes nas diferentes classes C_i qui compõem a partição:

$$s(P) = \frac{\sum_{C_i \in P} s(C_i)}{k} \tag{2.36}$$

A melhor partição mantida é então a que permite obter uma silhouette global máxima.

2.1.6 Conclusão

Vários métodos são propostos para o problema geral de agrupamento. Eles têm por diferença as medidas de proximidade que eles usam, os tipos de dados que processam e os objetivos finais do agrupamento. Cada um desses métodos tem seus pontos fortes e fracos. Os métodos hierárquico ascendentes são utilizados no caso de dados de tamanho pequeno uma vez que a complexidade é muito alta. Se, no entanto, problemas de tempo de execução surgir, então são os métodos *k-means* que são utilizados.

Além disso, a escolha de um método adequado depende muito da aplicação, os tipos de dados e os recursos disponíveis. Uma análise cuidadosa dos dados contribui para melhor escolher o algoritmo. Não existe um algoritmo que pode satisfazer todas aplicações.

Esta divisão focou-se no estado da arte dos métodos de agrupamento automático, dependendo do tipo de dados a serem agrupados (classificados), assim como a abordagem escolhida. Apresentou-se o painel de abordagens hierárquicas e abordagens de particionamento. Para cada uma, tentou-se de mostrar os pontos fortes e fracos.

Introduziu-se o conjunto de conceitos necessários para o resto desta dissertação, os conceitos de similaridade e dissimilaridade e os diferentes índices para avaliar a qualidade de uma partição de dados, o que é necessário ao tratar de uma abordagem iterativa.

2.2 Classificação automática por árvore de classificação e árvore de decisão

Esta seção apresenta as abordagens para a classificação automática de dados por árvore de agrupamento e árvore de decisão, as quais são quasi-similares, com as peculiaridades apresentadas.

2.2.1 classificação automática por árvore de classificação

Baseada no princípio de segmentação em um problema multivariado para constituir grupos homogêneos, a árvore de classificação relaciona o processo de constituição dos grupos homogêneos com a construção da regra de atribuição. Definido como:

$(Y_1,\ldots,Y_2) = f(X_1,\ldots,X_j;\alpha)$, onde Y_i serve para caracterizar a homogeneidade e $f(X_i)$ a construir os grupos.

A árvore de classificação consiste em (RAKOTOMALALA, 2007):(a) medir a homogeneidade dos grupos: generalizando o conceito de variância com o conceito de inércia, (b) e tem como critério a generalização multivariada da decomposição da variância, o teorema de Huygens. Isto leva a:

- Escolher a variável que maximiza o ganho de inércia, isto é, a inércia intraclasse;
- Produzir os subgrupos homogêneos;
- Obter os centros de gravidade distantes uns dos outros.

Igualmente às árvores de decisão, para decompor uma variável X, as técnicas de árvore de classificação consiste em tentar todas as segmentações possíveis (entre dois pontos sucessivos) e escolher o limite que maximiza o ganho de inércia. Por convenção, a árvore é binária para evitar a fragmentação dos dados. Para proceder à classificação das modalidades das variáveis categoriais, agrupa-se em prioridade as modalidades correspondentes aos grupos de indivíduos mais próximos (minimizar a perda de inércia consecutiva a um agrupamento - critério de WARD).

Assim como os métodos de agrupamento Hierárquica Ascendente (CHA), a árvore de classificação é uma técnica de classificação automática que visa criar grupos homogêneos em conformidade com um certo número de variáveis ativas (CHAVENT et al., 1999). Além disso, ela permite produzir diretamente uma regra de atribuição ”industrializável”, que proporciona uma interpretação direta dos grupos usando as mesmas regras. Ela possui velocidade / capacidade de lidar com grandes bases de dados (semelhante a árvores de decisão), possibilidade de guiar à busca das classes de investigação (construção interativa - apoiando-se sobre o conhecimento de domínio para produzir grupos relevantes). Outro fator é a possibilidade de separar variáveis ativas (explicadas), para avaliar a homogeneidade dos grupos anteriores, por exemplo, comportamento de compras, e variáveis de segmentação (explicativas), para desenvolver e explicar os grupos, por exemplo, características de identificação de indivíduos.

2.2.2 Árvore de decisão

Para algumas áreas de aplicação, é essencial produzir procedimentos de classificação compreensíveis pelo usuário. Este é, particularmente, o caso para o auxílio ao diagnóstico médico onde o médico deve ser capaz de interpretar as razões do diagnóstico. As árvores de decisão respondem a essa restrição porque elas representam graficamente um conjunto de regras e são facilmente interpretáveis. Para grandes árvores, o processo global pode ser difícil de compreender, no entanto, a classificação de um determinado elemento é sempre compreensível. Os

algoritmos de aprendizagem para árvores de decisão são eficazes, disponíveis na maioria dos ambientes de mineração de dados.

Uma árvore de decisão é uma árvore no sentido computacional. Lembre-se que os nós da árvore são rotulados por posições que são palavras de $\{1,\ldots,p\}*$, onde P é o número máximo de nós.

Os nós internos são chamados de nós de decisão. Tal nó é marcado por um teste que pode ser aplicado a qualquer descrição de um indivíduo na população. Em geral, cada teste examina o valor de um único atributo do espaço de descrições. Possíveis respostas do teste correspondem aos rótulos das arestas deste nó. No caso dos nós de decisão binária, arestas, rótulos de arestas são omissos e, por convenção, aresta esquerda corresponde a uma resposta positiva ao teste. As folhas são marcadas por uma classe chamada *classe por padrão*. A figura 2.3 apresenta um exemplo ilustrativo da árvore de decisão.

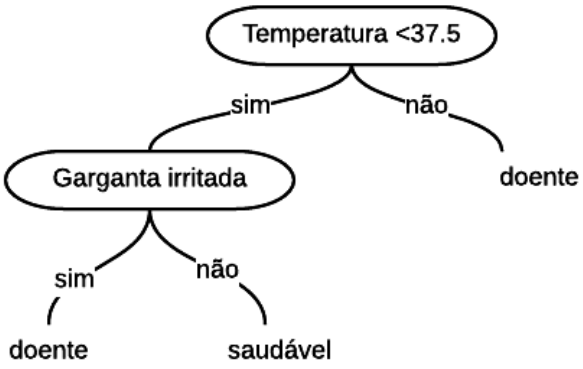

Figura 2.3: Exemplo da árvore de decisão

Existem diversos algoritmos na literatura utilizados para a construção de árvores de decisão, tais como ID3, C4.5 (SALZBERG, 1994) e CHAID. De forma resumida pode-se dizer que os algoritmos para classificação são recursivos e constroem a árvore utilizando uma abordagem *top-down*. Os algoritmos classificadores possuem como meta a construção de árvores que possuam o menor tamanho e a maior acurácia possíveis. Uma questão chave para a construção de uma árvore de decisão consiste na estratégia para a escolha dos atributos que estarão mais próximos da raiz da árvore (ou seja, os atributos que são inicialmente avaliados para determinar a classe à qual uma observação pertence).

Quando não se consegue realizar a discriminação nos ramos referindo-se à classe à qual uma observação pertence, pode-se dizer que, a escolha deste teste não faz nada ganhar, ele

será rejeitado, caso contrário pode ser considerado como interessante. Este raciocínio informal deve ser automatizado. Portanto, deve-se introduzir as quantidades que comparam as diferentes opções possíveis.

Existem várias funções que satisfazem essas propriedades, podem-se citar duas: a função entropia e a função de Gini (ganho de informação).

Entropia

Ao se falar da entropia, trata-se de uma medida comumente usada em teoria de informação para definir o ganho de informação que caracteriza a impureza de uma coleção arbitrária de exemplos. Dada uma coleção S que contém exemplos positivos e negativos de algum conceito objetivo, a entropia de S relativa a esta classificação lógica é:

$$Entropia = -p\oplus \log_2 p\oplus - p\ominus \log_2 p\ominus - \tag{2.37}$$

Onde p$\oplus$ é a proporção de exemplos positivos em S e p$\ominus$ é a proporção de exemplos negativos em S.

O melhor atributo é aquele com o ganho de informação maior, como será definido adiante. Além disso, em todos os cálculos que envolvem entropia a expressão 0 log 0 é definida como sendo 0.

Por outro lado, se o atributo designado pode assumir n valores diferentes, então a entropia de S relativa a esta classificação é definida como:

$$Entropia \equiv \sum -p_i \log_2 p_i \tag{2.38}$$

Onde p_i é a proporção de S necessária para classificar i.

Função de Gini (ganho de Informação)

Após a definição de entropia como uma medida da impureza na coleção de exemplos de treinamento, pode-se definir agora a medida da efetividade de um atributo para classificar os dados de treinamento. Usou-se uma medida, chamada ganho de informação, que é simplesmente a redução esperada na entropia causada pelo particionamento dos exemplos por este atributo. Mais precisamente, o ganho de informação, **Ganho(*S*, *A*)** de um atributo A, relativo a uma coleção de exemplos S, é definido como:

$$Ganho(S,A) \equiv Entropia(S) - \sum_{v \in valores(A)} \frac{|S_v|}{|S|} Entropia(S_v) \tag{2.39}$$

onde Valores(A) é o conjunto de todos possíveis valores para o atributo A, e S_v é o subconjunto de S para qual o atributo A tem valor v (isto é, $S_v = s \in S \mid A(s) = v$).

Observe-se que o primeiro termo na equação é a entropia da coleção original S, e o segundo termo é o valor esperado da entropia S dividido pelo atributo A. A entropia esperada descrita por este segundo termo simplesmente é a soma das entropias de cada subconjunto S_v, com peso igual à fração de exemplos $\frac{|Sv|}{|S|}$ que pertence a S_v. Ganho(S,A) é então a redução esperada na entropia causada pelo conhecimento do valor do atributo A. Isto é, Ganho(S, A) é a informação dada sobre o valor da função-objetivo, dado o valor de algum atributo A. O valor de Ganho(S, A) é o número de bits economizados quando codifica-se o valor-objetivo de um membro arbitrário de S, sabendo-se o valor do atributo A.

2.2.3 Conclusão

As árvores de decisão fornecem métodos eficazes que funcionam bem na prática. Elas têm a vantagem de ser compreendida por qualquer utilizador (se o tamanho da árvore produzida é razoável) e tem uma tradução imediata em termos de regras de decisão. Para o sistema baseado em regras induzidas, as regras são mutuamente exclusivas e a ordem em que os atributos são examinados é congelada. Os métodos são não-ótimos: as arvores produzidas não são os melhores. Na verdade, as escolhas na construção de árvores, baseadas em variadas heurísticas, nunca são questionadas (sem volta (ou *backtraking*)). Finalmente, é possível modificar os valores de vários parâmetros para escolher entre muitas variações e fazer a escolha certa nem sempre é fácil. Os tamanhos de amostra afetam os critérios de poda a escolher (do conjunto de treinamento, um conjunto de testes, validação cruzada, ...).

Capítulo 3

ANÁLISE DE DADOS SEQUENCIAIS: ESTADO DA ARTE

Este capítulo apresenta as diferentes tarefas da análise de dados sequencial, o problema da classificação automática e agrupamento automático de sequências temporais evocando as abordagens mais comentadas na literatura (as abordagens por aproximação e os modelos de mistura). Relata-se para cada uma destas abordagens, princípios, vantagens e desvantagens.

3.1 Introdução

Se a mineração de dados, tal como definida por (KODRATOFF et al., 2001), pode ser comparada a *"um processo interativo e iterativo de análise de um grande conjunto de dados brutos definido para extrair conhecimento exploráveis através do analista que desempenha um papel central"*, a mineração de dados sequenciais, por sua vez, fornece mais capacidade para sugerir as causas e os efeitos.

Pode-se identificar dois tipos de objetos sobre os quais algoritmos de mineração de dados sequenciais operam: (1) as séries sequenciais representando dados provenientes de fontes contínuas, e (2) as sequências temporais representando listas ordenadas de eventos. É por este último tipo de dados que motivou a continuação deste trabalho.

As principais fontes estatísticas de séries temporais, descrevendo a evolução dos indivíduos em diferentes escalas de tempo, geralmente são levantamentos retrospectivos. Nestes inquéritos são coletados para uma amostra de indivíduos, trajetórias definidas pelas mudanças de variáveis de estado que representam o seu comportamento.

Durante a última década, diferentes técnicas de mineração de dados sequenciais foram propostas e mostram ser úteis em várias áreas de aplicação (marketing, sociologia, medicina). Como a mineração de dados sequenciais explora mecanismos e técnicas próprias para diferentes áreas, a saber a *estatística*, o *aprendizado de máquina* e os *bancos de dados*, os trabalhos de várias fontes (ANTUNES; OLIVEIRA, 2001; LAXMAN; SASTRY, 2006). No entanto, os principais objetivos (tarefas e operações) da mineração dos dados da sequenciais são as mesmas e podem ser agrupadas da seguinte forma: (1) a *previsão*, (2) *classificação*, (3) a *descoberta de padrões* e (4) a *classificação automática*. Propõe-se de explicar cada uma dessas tarefas nas seções a seguir.

Previsão

Também chamada de predição por abuso de linguagem no contexto de séries temporais (ou tradução literal do termo inglês), a previsão consiste em avaliar (ou prever) o estado de uma sequência temporal em um instante t tendo conhecimento dos seus valores nos instantes $t_i < t$. As abordagens de previsão em sequências temporais geralmente operam em duas etapas: inicialmente, procuram desenvolver um modelo de dados que de alguma forma resume as relações entre os diferentes estados das sequências tratadas, e permite, numa segunda fase realizar previsões ótimas. A previsão de séries temporais é um problema que foi tratado, por muito tempo, por métodos estatísticos convencionais. Os modelos como auto-regressivos (ARMA, ARIMA) (BOX; JENKINS, 1994), foram usados com sucesso em diversas áreas de aplicação (econômico, industrial, etc.), com o objetivo de prever os valores futuros em uma sequência temporal, por combinação linear de valores anteriores (BOX; JENKINS, 1994; CHATFIELD, 2004; HASTIE et al., 2008). Enquanto os modelos ARMA (*AutoRegressive Moving Avarage*) não permitem tratar as séries temporais estacionárias [1], os modelos ARIMA (*Autoregressive Integrated Moving Average*) foram propostos para tratar as sequências não-estacionárias depois ter determinado o número de vezes que é necessário para diferenciar a série antes de torná-lo estacionário. Os limites dos modelos auto-regressivos residem essencialmente na natureza dos dados: valores qualitativos nominais, por exemplo, tornam essas técnicas inaplicáveis. Neste contexto, os modelos de Markov (*cadeias de Markov ou modelos ocultos de Markov*) podem ser usados para fazer previsões sobre o estado futuro do processo modelado. Recentemente, as técnicas estatísticas não paramétricas, tais como redes neurais também têm sido propostos (KOSKELA et al., 1996).

[1] Uma série ou sequência temporal é dita estacionária se as suas realizações são derivadas a partir do mesmo processo aleatório cujos parâmetros (média, variância, autocorrelação ...) permanecem os mesmos ao longo do tempo.

Classificação supervisionada

Considerada uma das técnicas de mineração de dados mais antigas e mais utilizadas, quer em medicina para determinar a patologia de um paciente, em *marketing* para identificar o perfil de um cliente, na sociologia para conhecer a categoria demográfica de um indivíduo, a classificação de dados sequenciais assume que cada sequência temporal de uma dada população pertence a uma classe (ou categoria) pré-definida e procura determinar automaticamente a categoria apropriada para qualquer nova sequência introduzida no sistema. Nos últimos anos, várias técnicas de classificação de dados sequenciais têm sido desenvolvidas. As mais comuns são abordagens baseadas em *protótipos* e abordagens baseadas em *modelos*. Com a primeira família, cada uma das classes pré-definidas é caracterizada por uma sequência representativa "tipo"da classe *(protótipo)* e a classificação de uma nova sequência é feita observando entre os protótipos das classes a que está mais próxima a esta sequência, de acordo com a métrica escolhida.

Nas próximas seções, detalha-se algumas medidas de similaridade entre as sequências temporais.

Descoberta de padrão

Outra técnica descritiva, menos comum, entretanto, do que a classificação automática, mas que interessa mais e mais setores estratégicos tais como *marketing*, finanças ou medicina (identificação de sintomas antes da doença) é a extração automática de padrões sequencias. A operação de extração de tais padrões, introduzida por Agrawal em (AGRAWAL; SRIKANT, 1995) pode ser considerada uma extensão da recuperação de regras de associação em bancos de dados transacionais. A busca de padrões sequenciais consiste em extrair sequências, ou seja, conjuntos de símbolos comumente associados ao longo de um período de tempo específico, e a probabilidade de reincidência. Este mecanismo pode permitir identificar algum comportamento típico de indivíduos ao longo do tempo.

Classificação automática

A classificação automática procura identificar uma tipologia de indivíduos em grupos homogêneos e bem separados. Aplicada às sequências de dados temporais, os indivíduos manipulados são desta vez sequências temporais que procura-se corresponder estatisticamente seus diferentes registros. Em geral, a classificação automática é útil como um preliminar para outras operações de mineração de dados. Na verdade, uma abordagem comum de mineração de dados é encontrar classes de indivíduos que têm o mesmo tipo de comportamento, a fim de, em seguida, ser capaz de identificar a classe à qual pertence um novo indivíduo, explorando seu histórico (classificação) e avaliar o comportamento futuro

(previsão). É por estas razões que insiste-se, neste trabalho, nos algoritmos exploratórios de dados sequenciais para a classificação automática.

3.2 Abordagens de classificação de dados sequenciais

Várias técnicas de classificação automática de dados sequenciais foram desenvolvidas nos últimos anos. Elas têm sido aplicadas em diferentes áreas tais como a análise de sequências biológicas (CADEZ et al., 2000b), o estudo da mobilidade dos objetos em vídeos (BUZAN et al., 2004), ou outras áreas onde o indivíduo tem um papel primordial, modelar o comportamento dos usuários na Web (CADEZ et al., 2000a), etc. Dentro desta gama de métodos de classificação, as mais comumente usadas são abordagens por proximidade e abordagens por modelos de mistura.

3.2.1 As abordagens de classificação com base na noção de proximidade

Muitos métodos em análise de dados são baseados no conceito de similaridade e distância entre os objetos a serem analisados. As abordagens de classificação automáticas de dados sequenciais naturalmente precisam saber a proximidade entre as sequências para agrupá-las. Esta seção apresenta diferentes técnicas para avaliar a proximidade entre sequências temporais, que causa igualmente o problema de sequências de comprimento variados. A distâncias mais utilizadas são a distância *euclidiana*, alinhamento dinâmico temporal (**DTW: Dynamic Time Warping**) e a maior subsequência comum (**LCS: Longest Common Subsequence**).

Uma vez que o índice de proximidade é definido para todas as sequências da população, as abordagens convencionais para classificação automática apresentadas anteriormente podem ser facilmente aplicadas. É, particularmente, o caso das abordagens de classificação hierárquicas e das abordagens baseadas em árvore de classificação. Estas abordagens requerem o uso de uma medida de similaridade entre as sequências (ou as trajetórias) com base na qual tentam-se construir uma partição destas sequências em classes homogêneas e bem separadas. Estes aspectos são detalhados nas seções a seguir.

3.2.2 Algumas distâncias adotadas às sequências temporais

A distância euclidiana

A distância euclidiana é uma das distâncias mais utilizadas. Ela apresenta a vantagem de ser intuitiva e simples de se implementar, no entanto, encontra-se rapidamente limitada frente aos dados ruidosos, periódicos ou de comprimentos diferentes. A distância euclidiana $d(S_i, S_j)$ entre duas sequências temporais $S_i = e_{i,1}, \ldots, e_{i,T_i}$ e $S_j = e_{j,1}, \ldots, e_{j,T_j}$ ($e_{i,t}$ é a t-esima observação da sequência S_j) de comprimentos diferentes ($T_i \neq T_j$) é definida como segue:

$$d(S_i, S_j) = \sqrt{\sum_{t=1}^{min(T_i,T_j)} (e_{i,t} - e_{j,t})^2} \tag{3.1}$$

Dynamic Time Warping (DTW)

Para superar os problemas associados à distância Euclidiana, Sakoe (1979) introduziu a distância *DTW: Dynamic Time Warping* (alinhamento dinâmico temporal) no domínio de reconhecimento de fala. Esta foi utilizada para medir a similaridade entre qualquer palavra falada por um locutor humano e uma várias palavras de referência, permitindo, em particular, superar o ritmo de pronúncia.

A DTW (KRUSKAL, 1983) é reconhecida posteriormente como uma medida muito fiável para avaliar a distância entre duas sequências não necessariamente de comprimento idêntico, tendo em conta o efeito de translação (dilatação) presentes nos dados, isto é, a presença ou ausência de estados intermediários entre estados estudados nas duas sequências. Semanticamente, para comparar duas sequências temporais com a distância DTW, o processo consiste em deformar as duas sequências através da inserção de $'-'$ (concretamente isto equivale a um estiramento de uma e/ou da outra sequência) até que se obtenha o "melhor"*matching* (correspondência) entre sequências modificadas. Este processo é chamado de *alinhamento temporal*.

O algoritmo de cálculo de DTW realiza esse alinhamento, buscando entre todos os alinhamentos possíveis, qual minimiza uma função de custo γ integrando a diferença entre os dados alinhados e um custo de deformação temporal. A distância escolhida é a correspondente ao custo mínimo de alinhamento.

Longest Common Subsequence: LCS

Proposta, inicialmente, para a comparação de cadeia de caracteres, a medida da mais longa subsequência comum (*LCS, Longest Common Subsequence*) de Paterson (PATERSON; DANČÍK, 1994) foi considerada em seguida como um caso particular da *Dynamic Time Warping* específica aos dados qualitativos (categoriais). Utilizando o mesmo princípio que a DWT, o algoritmo de busca da mais longa subsequência comum reduz a distância de cúmulo para cada comparação entre os símbolos das sequências a 1 ou 0, consoante a presença ou a ausência do mesmo símbolo.

Definição

Sejam S_1 e S_2 duas sequências de dados sequenciais (ditas cadeias de caracteres). Uma subsequência comum a S_1 e S_2 é uma cadeia de caracteres c cujos elementos aparecem tanto em S_1 como em S_2 respeitando a ordem pré-estabelecida nestas duas sequências. Nota-se $LCS(S_1, S_2)$, o comprimento de uma subsequência comum relativa a S_1 e S_2.

O problema da avaliação da *distância* entre duas cadeias de caracteres é um generalização do problema da avaliação do comprimento da mais longa subsequência comum a estas duas cadeias de caracteres. Esta distância chamada *distância de edição* é um meio típico de abordagens de reconhecimento de escrita manuscrita, mas também foi utilizado para medir a quantidade de evolução entre duas sequências biológicas na classificação automática de diferentes tipos de trajetórias (BUZAN et al., 2004).

Mais precisamente, a distância de edição entre duas sequências de dados categoriais ou simbólicos S_i e S_j se escreve:

$$d_E(S_i, S_j) = |S_i| + \left|S_j\right| - 2 * LCS(S_i, S_j) \tag{3.2}$$

Algoritmo

A medida da mais longa subsequência comum relativa a duas sequência de dados simbólicos pode ser igualmente calculada por um algoritmo de programação dinâmica, de complexidade $O(T_i, T_j)$, da forma que segue (a matriz de cúmulo é chamada, neste caso, **L**):

Algorithm 1 Algoritmo de busca da mais longa subsequência comum

Entrada: Duas sequências de dados categoriais $S_i = e_{i,1}, \ldots, e_{i,T_i}$ e $S_j = e_{j,1}, \ldots, e_{j,T_j}$

Saída: o comprimento máximo de uma subsequência comum relativa a S_i e S_j: $LCS\,(S_i, S_j)$

1: $L[0][0] \leftarrow 0$;
2: $L[0 \ldots T_i][0] \leftarrow 0$;
3: $L[0][\ldots T_j][0] \leftarrow 0$;
4: **para** u $\leftarrow 1$ **até** T_i **faça**
5: **para** v $\leftarrow 1$ **até** T_j **faça**
6: **se** $(e_{i,u} = e_{j,v})$ **então**
7: **devolve** $L[u][v] \leftarrow L[u-1][v-1] + 1$;
8: **senão**
9: **se** $L[u-1][v] > L[u][v-1]$ **então**
10: **devolve** $L[u][v] \leftarrow L[u-1][v]$;
11: **senão**
12: **devolve** $L[u][v] \leftarrow L[u][v-1]$;
13: **fim se**
14: **fim se**
15: **fim para**
16: **fim para**
17: **devolve** $(L[T_i][T_j])$

A título de exemplo, seja as duas sequências S_1 = CATCAGTA e S_2 = ACTCCATGCA. A tabela 3.1 ilustra a matriz de cúmulo das distâncias L[][] de tamanho $9x11$. Tem-se $LCS(S_1, S_2) = 6$ e $d_E(S_i, S_j) = |S_i| + \left|S_j\right| - 2 * LCS(S_i, S_j) = 7 + 10 - 2 * 6 = 6$.

As subsequências máximas comuns relativas a S_1, S_2 são, por exemplo, *CTCAGA* e *ATCATA*.

		0	1	2	3	4	5	6	7	8	9	10
	S_2 \ S_1		A	C	T	C	C	A	T	G	C	A
0		0	0	0	0	0	0	0	0	0	0	0
1	C	0	0	1	1	1	1	1	1	1	1	1
2	A	0	1	1	1	1	1	2	2	2	2	2
3	T	0	1	1	2	2	2	2	3	3	3	3
4	C	0	1	2	2	3	3	3	3	3	4	4
5	A	0	1	2	2	3	3	4	4	4	4	5
6	G	0	1	2	2	3	3	4	4	5	5	5
7	T	0	1	2	3	3	3	4	5	5	5	5
8	A	0	1	2	3	3	3	4	5	5	5	**6**

Figura 3.1: Matriz de cúmulo de distância L para calcular LCS (ELGHAZEL, 2007)

3.3 Conclusão

As abordagens de classificação automática baseadas em índice de proximidade são particularmente adaptadas para busca de diferentes perfis de indivíduos que constituem a população. Para isso, elas buscam descobrir uma partição de dados em classes homogêneas e bem separadas, de modo que as sequências mais próximas (no sentido da métrica utilizada) encontrarem-se na mesma classe (coesão intra-classe), enquanto sequências dissimilares são distribuídas em classes diferentes (separação inter-classes). No entanto, as classes obtidas por classificação de abordagens baseadas em um índice proximidade nem sempre são fáceis de interpretar. Na verdade, a maioria desses métodos são muitas vezes capazes de fornecer uma descrição das classes através de sequências ditas ”tipos”(ou seja, as sequências centrais da classe, por exemplo), mas não conseguem desenvolver modelos que resumem as informações contidas nas sequências da classe e as relações entre elas. Mas para muitas aplicações de apoio à decisão, faz-se necessário ser-se capaz de descrever as classes da população sob uma forma compacta permitindo uma possível abstração de dados. Por conseguinte, destaca-se a dificuldade de considerar as novas sequencias introduzidas no sistema para deduzir suas classes e prever a progressão do comportamento dos seus indivíduos correspondentes.

Capítulo 4

MODELOS DE MARKOV

Este capítulo apresenta uma síntese sobre a teoria das cadeias de Markov, assim como a identificação de técnicas de modelização em diferentes domínios de aplicação.

4.1 Introdução

A modelagem estocástica permite usar modelos probabilísticos para resolver problemas com informações incertas ou incompletas. Assim, os Modelos de Markov despertam um interesse em ambos os aspectos teóricos e aplicados.

A teoria de cadeias de Markov nasceu em 1913, e cuja primeira aplicação foi desenvolvida por Markov para analisar a linguagem. Este trabalho foi regularmente utilizado, mas as primeiras aplicações utilizáveis foram realizadas nos anos 60, como os modelos probabilísticos de urnas por Neuwirtch, cálculo direto da máxima verossimilhança ou a observação da progressão da série de estados em uma cadeia de Markov. Isso permitiu a comunidade científica explorar todo o potencial desses modelos. Foi nos anos 70 que os investigadores fizeram algoritmos poderosos para resolver os problemas de reconhecimento, de análise e de aprendizagem.

Desde 1975, os Modelos Ocultos de Markov (*Hidden Markov Models* em inglês ou HMMs) são usados em muitas aplicações, principalmente no domínio de reconhecimento de voz. Essas aplicações não se contentam com o apoio a apenas da teoria dos HMMs, mas desenvolvem várias extensões teóricas, a fim de melhorar os modelos. Isto é o que os tornou bem sucedidos.

Neste capítulo é apresentada uma síntese sobre a teoria das cadeias de Markov, Modelos Ocultos de Markov e identificação de técnicas de modelagem em diferentes domínios de aplicação.

4.2 Teoria de Cadeias de Markov

Um processo estocástico é um fenômeno onde ocorre o acaso (RABINER, 1989). Define-se $X(t)$ uma variável aleatória evoluindo em função do tempo.

Seja uma sequência 1, 6, 2, 5 donde $X_0 = 1, X_1 = 6, X_2 = 2, X_3 = 5$.

Este processo é chamado markoviano se sua evolução não depende de seu passado, mas apenas de seu estado atual (chama-se a propriedade de Markov).

Um processo de Markov pode ser modelado por um modelo teórico denominado ”Modelo de Markov”. Existem dois tipos de modelos: observáveis e ocultos.

4.2.1 Cadeia Observável

A evolução do processo de Markov pode ser apresentada por um grafo de transição de estados (Figura 4.1) que mostra a estrutura do processo de acordo com as seguintes regras:

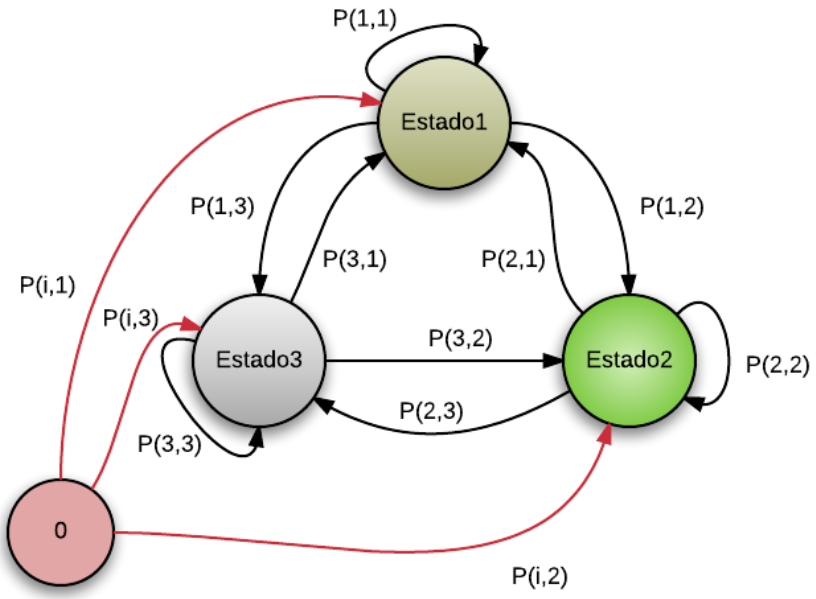

Figura 4.1: Grafo de um Modelo de Markov Observável

Observa-se, por exemplo, que $P(32)$ ou P_{32} é a probabilidade de passar para o estado 3 ao estado 2. Da mesma forma, P_{i2} é a probabilidade de iniciar no estado 2.

- Os estados são representados por nós (Estado n). Fala-se de alfabeto de estados : S :

$\{s_1, \ldots, s_N\}$, os estados da cadeia de Markov.

- As transições (possibilidades de passar de um estado para o outro estado) são representadas pelas arestas (direcionadas), elas são ponderadas por suas probabilidades.

  As probabilidades são agrupadas em uma matriz de transição:

$$A = \left\{a_{i,j} = P(s_j \mid s_i)\right\} \sum_{j=1}^{N} a_{i,j} = 1 \tag{4.1}$$

- As probabilidades de partida: são as probabilidades de iniciar em um estado ou em outro (ponto 0). Elas são agrupadas em um vetor de inicialização: $\prod = \{\pi = P(s_i)\}$ nota-se: $\sum_{i=1}^{N} \pi_i = 1$.

Um modelo λ é dito observável pois os estados são diretamente observáveis, e é caracterizado por uma matriz de transição A e um vetor de inicialização π, nota-se:

$$\lambda = \{\Pi, A\}$$

4.2.2 Cadeia Oculta

Em um Modelo Oculto de Markov os estados $S : \{s_1, \ldots, s_N\}$ são não observáveis (DENG; ZHENG, 2006). No entanto, eles emitem sinais observáveis $O : \{o_1, \ldots, o_k\}$ que são ponderados por suas probabilidades. O modelo λ pode ser representado graficamente conforme a Figura 4.2.

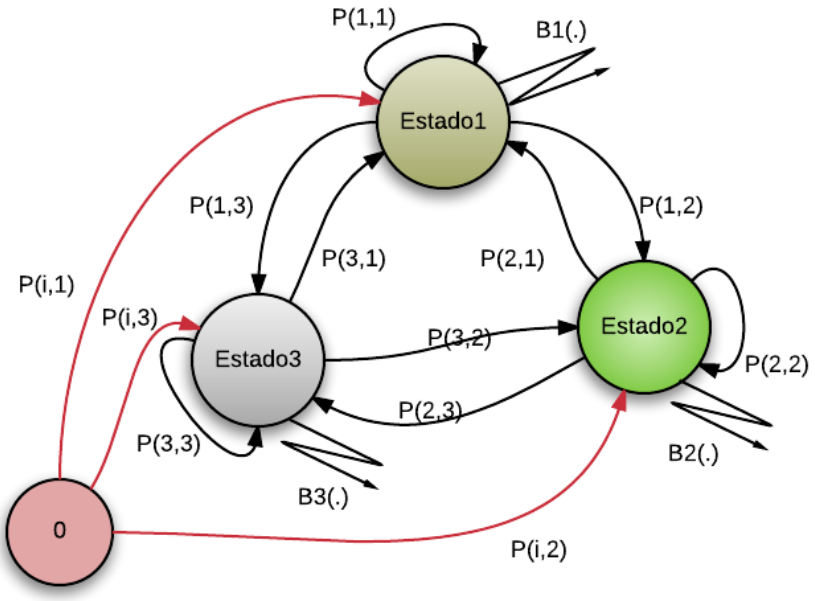

Figura 4.2: Grafo de um Modelo Oculto de Markov

- Os estados são $S : \{s_1, \ldots, s_N\}$
- A matriz de transição $A = \{a_{i,j} = P(s_j \mid s_i)\}$; $\sum_{j=1}^{N} a_{i,j} = 1$;
- O vetor de inicialização $\prod = \{\pi = P(s_i)\}$; $\sum_{i=1}^{N} \pi_i = 1$;
- As probabilidades que o estado s_i emite o sinal de observação o_k (direcionado). Estas observações são agrupadas em uma matriz de emissão $B = \{b_i(o_k) = P(o_k \mid s_i)\}$; $\sum_{j=1}^{T} b_i(o_j) = 1$;

Doravante, usar-se-á as representações matemáticas abaixo com seus respectivos significados:

- N: o número de estados $N = \sum S$;
- T: o número de observações possíveis $T = \sum O$;
- q_t: o estado do sistema no instante t;
- M: comprimento da sequência observada.

Um Modelo Oculto de markov λ é caracterizado por uma matriz de transição A, uma matriz de observação B e um vetor de inicialização π, donde nota-se:

$$\lambda = \{\Pi, A, B\}$$

Nota-se que um Modelo de Markov Observável pode ser modelado sob forma de um Modelo Oculto onde os estados correspondem aos eventos observados, isto é, cada estado s_i tem uma probabilidade "1"de emitir a observação b_i.

Exemplificando (RABINER, 1989), seja um Modelo de Markov observável representando o clima/tempo do dia.

As observações são os estados definidos por Nublado, chuva, Sol. O Modelo de Markov permite definir:

- A probabilidade de realizar uma sequência. Exemplo: Nublado, Nublado, Chuva = NNC.
- A previsão de um estado futuro conhecendo ou não o estado atual.
- A probabilidade de se ter "d"dias com as mesmas condições.

Esclarecendo, este processo tem como modelagem a seguir:

- $X(t)$ = clima/tempo do dia;
- O alfabeto $S = O = \{s_{1=}O_1, s_{2=}O_2, \ldots, s_{n=}O_m\}$
- A matriz de transição $A = \left\{a_{i,j} = P(s_j \mid s_i)\right\}$; $\begin{pmatrix} 0.4 & 0.3 & 0.3 \\ 0.2 & 0.6 & 0.2 \\ 0.1 & 0.1 & 0.8 \end{pmatrix}$

  $A = \left\{a_{i,j} = P\left[q_t = S_j \mid q_{t-1} = S_j\right] = P(o_j \mid o_i)\right\}$, $\forall i \in [1,N]$ e $j \in [1,N]$.
- As probailidades iniciais $\Pi = \{\pi_i = P(s_i)\} = \begin{pmatrix} 0.02 \\ 0.4 \\ 0.58 \end{pmatrix}$

A Figura 4.3 ilustra o modelo gráfico correspondente aos dados acima.

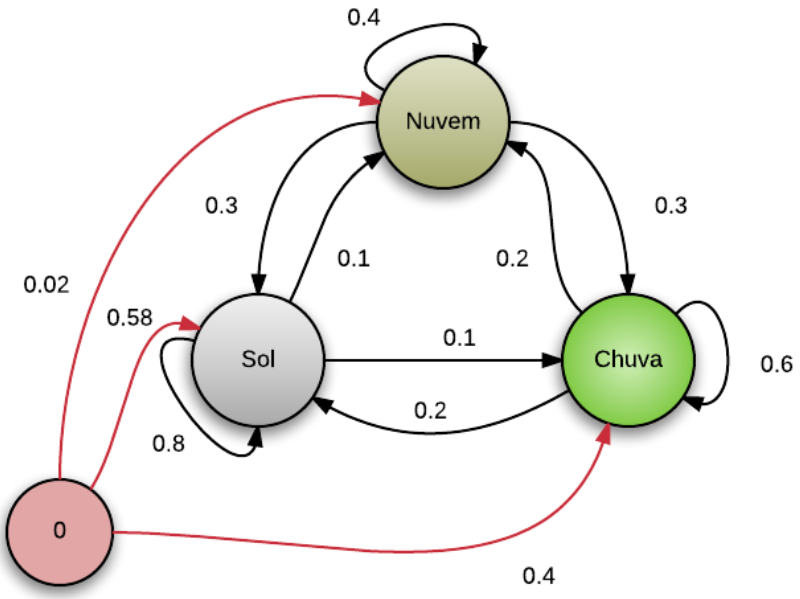

Figura 4.3: Grafo do Modelo de Markov Observável. Adaptado do (RABINER, 1989)

No entanto, um modelo de Markov, onde cada estado corresponde a um evento observável é demasiado restritivo para ser generalizado para um grande número de problemas em que os estados não são diretamente observáveis. Na seção 4.2.3, estuda-se o princípio de Markov para incluir o caso em que a observação é uma função probabilística do estado.

4.2.3 Modelos Ocultos de Markov

Um Modelo Oculto de Markov ou HMM (Hidden Markov Model) é um processo duplamente estocástico cujo um dos componentes é uma cadeia de Markov não observável (ZHANG et al., 2013). Este processo pode ser observado através de um outro conjunto de processos que produz uma série de observações. Mais precisamente, é um modelo que descreve os estados de um processo de Markov com probabilidades de transição e probabilidades de observação por estados.

Definição

Ao criar um HMM, existem três problemas para resolver: o reconhecimento, a análise e a aprendizagem (DEQUIER, 2005). A fim de melhor entendê-los, suponha-se o caso a seguir:

Considera-se que se queira conhecer ou determinar a Estação em que se encontra. Sejam

as quatro estações (Primavera, Verão, Outono, Inverno), como sendo os estados do modelo. O estado "estação"não é diretamente observável, mas emite observações, o tempo do dia. Este estado é definido por Nublado, Chuva e Sol.

Admita-se uma cadeia de observações como sendo o tempo da semana. Neste caso, considerando apenas três dias para limitar o número de cálculo (Sol, Sol, Nublado) ou (SSN).

O Modelo de Markov permitirá definir qual cadeia de Markov tem mais probabilidade de ter gerado a sequência observada, por exemplo, Verão, Verão, Primavera, ou (VVP).

Seja a modelagem do processo a seguir:

- $X(t)$ = clima/tempo do dia;
- O alfabeto $S = \{s_1, s_2, \ldots, s_m\}$ = { Primavera, Verão, Outono, Inverno }
- A matriz de transição $A = \left\{a_{i,j} = P(s_j \mid s_i)\right\}$; $\begin{pmatrix} P & V & O & I & \\ 0.3 & 0.5 & 0 & 0.2 & P \\ 0.2 & 0.5 & 0.2 & 0.1 & V \\ 0 & 0.1 & 0.8 & 0.1 & O \\ 0.3 & 0.1 & 0.2 & 0.4 & I \end{pmatrix}$

  $A = \left\{a_{i,j} = P\left[q_t = S_j \mid q_{t-1} = S_j\right] = P(o_j \mid o_i)\right\}, \forall i \in [1,N]$ e $j \in [1,N]$.
- As probailidades iniciais $\Pi = \{\pi_i = P(s_i)\}$ = $\begin{pmatrix} P \\ V \\ O \\ I \end{pmatrix} = \begin{pmatrix} 0.25 \\ 0.25 \\ 0.25 \\ 0.25 \end{pmatrix}$
- O alfabeto $O = (o_1, o_2, \ldots, o_k)$ = *(Nublado, Chuva, Sol)* = *(N, C, S)* símbolos emitidos pelos s_i
- As probabilidades de emissão $B = \{b_i(O_k) = P(O_k \mid s_i)\}$ = $\begin{pmatrix} N & C & S & \\ 0.1 & 0.45 & 0.45 & P \\ 0.01 & 0.13 & 0.86 & V \\ 0.05 & 0.55 & 0.4 & O \\ 0.2 & 0.5 & 0.3 & I \end{pmatrix}$

A representação gráfica deste modelo pode ser vista na Figura 4.4.

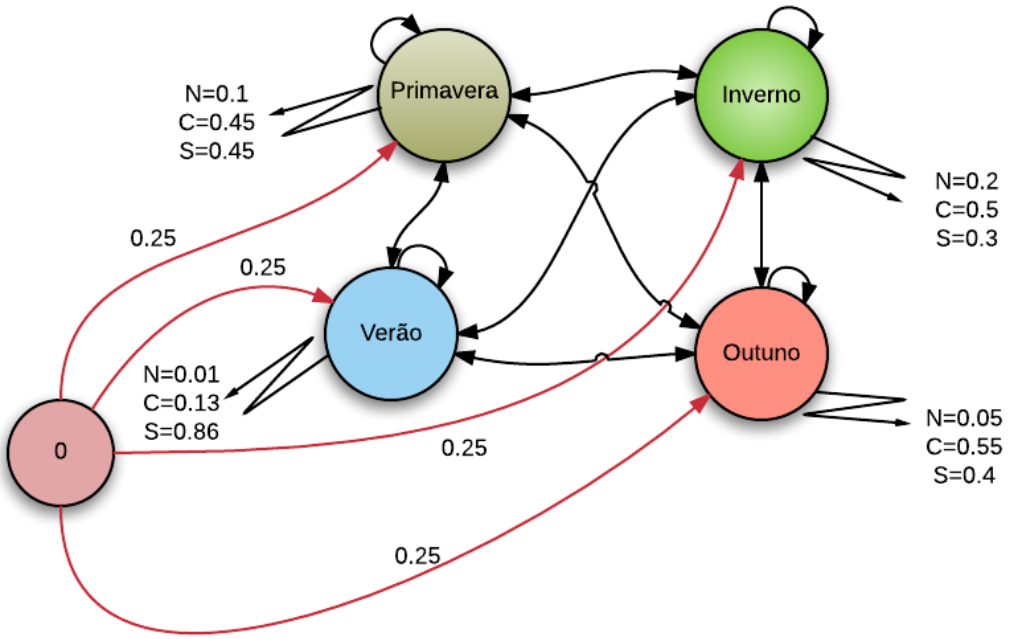

Figura 4.4: Modelagem HMM para determinar a provável estão em um instante *t*.

Os três problemas fundamentais dos HMM

Os HMM têm três problemas fundamentais (KRIOUILE, 1990), os quais é necessário resolver a fim de poder usar os HMM na modelagem de um processo real.

- **Problema 1: Reconhecimento**: dado um HMM $\lambda = \{\Pi, A, B\}$ e uma sequência observada $O = \{o_1, o_2, \ldots, o_n\}$ qual é a verossimilhança $P(O \mid \lambda)$ que o modelo λ gere O?
- **Problema 2: Análise**: dado um HHM λ e uma sequência observada O, qual é a sequência dos estados que tem a probabilidade máxima de ter gerado O?
- **Problema 3: Aprendizagem**: a partir de uma cadeia de observações $O = \{o_1, o_2, \ldots, o_n\}$ como ajustar os parâmetros do HMM $\lambda = \{\Pi, A, B\}$ para maximizar a verossimilhança do conjunto de aprendizagem $P(O \mid \lambda)$?

Problema 1: Reconhecimento

Dada uma sequência $O = \{o_1, o_2, \ldots, o_n\}$ e um modelo λ, como pode-se calcular eficazmente a probabilidade (verossimilhança) que a sequência de observação O seja produzida por λ, isto é, $P(O \mid \lambda)$. Em outras palavras, como avaliar o modelo a fim de escolher dentre muitos aquele que melhor gera a sequência de observação. Muitas técnicas permitem resolver esse

problema: método de avaliação direta, procedimento "Forward-Background"e Algoritmo de Viterbi.

- **Avaliação direta**: a probabilidade $P(O \mid \lambda)$ de uma sequência de observações $O = \{o_1, o_2, \ldots, o_n\}$, sabendo o modelo $\lambda = \{\Pi, A, B\}$, é o somatório sobre todos os caminhos da probabilidade que o caminho $Q = \{q_1, q_2, \ldots, q_n\}$ em curso tenha gerado a observação, seja : $P(O \mid \lambda) = \sum_Q P(O \mid Q, \lambda) P(Q \mid \lambda)$

  A probabilidade de optar pelo caminho Q é definida como: $P(Q \mid \lambda) = P(s_1, s_2, \ldots, s_n \mid \lambda) = \pi_1 * a_{1,2} * a_{2,3} * \cdots * a_{n-1,n}$.

  E a probabilidade para que esta sequência Q emita os sinais de observaçoes O: $P(O \mid Q, \lambda) = P(s_1, s_2, \ldots, s_n, \lambda) = b_1(o_1) * b_2(o_2) * \ldots b_n(o_n)$.

  Obtém-se:

$$P(Q \mid \lambda) = \sum_Q \pi_1 * a_{1,2} * b_1(o_1) * a_{2,3} * b_2(o_2) * \cdots * a_{n-1,n} * b_n(o_n) \tag{4.2}$$

  Isto gera $(2T-1) * N^T$ multiplicações e $N^2 - 1$ adições. seja em torno de $2T * N^T$ operações. Por exemplo, com $N = 5$ (estados), $T = 100$ observações, deve-se fazer operações de ordem de $(2 * 100 * 5^100) \simeq 10^{72}$, o que, com 1Ghz e supondo que um cálculo é igual a uma operação elementar do CPU, ter-se-á a resposta em $\simeq 3 * 10^{55}$ anos.

  Constata-se que inúmeras multiplicações são repetidas (porção de subsequência comum). A ideia é de calcular $P(O \mid \lambda)$ de forma incremental: Algoritmo Forward-Background.

- **Algoritmo Forward-Background**: nesta abordagem, considera-se que a observação pode ser feita em duas etapas (TAN-JAN; CHEN, 2006):

  1. A emissão da sequência de observações $\{o_1, o_2, \ldots, o_t\}$ e a realização do estado q_t no tempo t: **Forward**;
  2. A emissão da sequência de observações $\{o_{t+1}, o_{t+2}, \ldots, o_T\}$ partindo do estado q_t no tempo t: **Backward**.

  $P(O \mid \lambda)$ pode ser definido a cada instante $t \in [1, T]$ por :

$$P(O \mid \lambda) = \sum_{i=1}^{N} \alpha_i(i) * \beta_i(i) \tag{4.3}$$

  Onde $\alpha_i(i)$ é a probabilidade de emitir a sequência $\{o_1, o_2, \ldots, o_t\}$ e chegar a q_t no instante t e $\beta_i(i)$ a probabilidade de emitir a sequência $\{o_{t+1}, o_{t+2}, \ldots, o_T\}$ partindo do estado q_t no instante t, conhecendo λ.

O cálculo de $\alpha_i(i)$ faz-se com t crescente enquanto o de $\beta_i(i)$ faz-se com t decrescente, daí a expressão Forward-Background.

a. **A variável Forward**

Seja a probabilidade $\alpha_i(i) = P(O, q_t = s_i \mid \lambda)$ de gerar $O = \{o_1, 0_2, \ldots, o_n\}$ e de se encontrar no estado q_t no instante t.

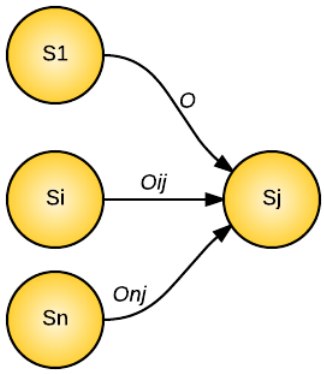

Figura 4.5: Forward: $\alpha_t(i) \rightarrow \alpha_{t+1}(j)$

O algoritmo abaixo permite calcular esta probabilidade:

1. **Inicialização**: $\alpha_1(i) = \pi_i * b_i(o_1)$
2. **Iteração**: $\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij}\right] * b_j(o_{t+1}), \forall t \in [1, T-1], j \in [1, N]$
3. **Término**: $P(O \mid \lambda) = \sum_{i=1}^{N} \alpha_T(i)$

b. **A variável Backward**

Seja a probabilidade $\beta_i(i) = P(O \mid q_1 = s_i, \lambda)$ de gerar $O = \{o_1, 0_2, \ldots, o_T\}$ sabendo-se que está no estado q_t no instante t.

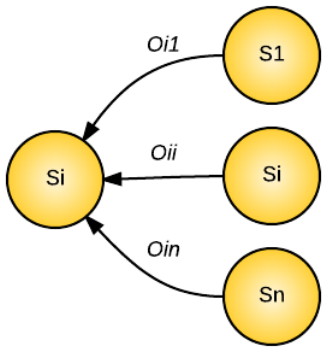

Figura 4.6: Backward: $\beta(i) \leftarrow \beta_{t+1}(j)$

O algoritmo abaixo permite calcular esta probabilidade:

1. **Inicialização**: $\beta_T(i) = 1, 1 \leq i \leq N$
2. **Iteração**: $\beta_t(i) = \left[\sum_{j=1}^{N} a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)\right], \forall t \in [T-1, 1]$ e $\forall i \in [1, N]$
   Para estar no estado q_t no instante t, tendo em conta a sequência de observações de $O = \{o_{t+1}, o_{t+2}, \ldots, o_T\}$, é preciso considerar todos os estados possíveis s_j (todas as transições a_{ij}) e a observação O_{t+1} no estado j (Os $b_j(o_{t+1})$), em seguida a sequência de observações parciais restante a partir do estado $j(\beta_{t+1}(j))$.

Problema 2: Análise

Dada uma sequência de observações $\{o_1, o_2, \ldots, o_T\}$ e um modelo $\lambda = \{\Pi, A, B\}$, como pode-se escolher uma sequência de estados $Q = \{q_1, q_2, \ldots, q_T\}$ que seja ótima de acordo com um critério apropriado. A dificuldade reside na sequência ótima dos Estados. Existem vários métodos: o critério local, o critério global e o algoritmo de Viterbi (para mais detalhes sobre o algoritmo pode-se consultar (TAN-JAN; CHEN, 2006; BOBBIN, 2007; SIU; CHAN, 2006)).

O critério local é uma técnica que consiste em escolher o estado q_t que é o mais provável e isto independentemente dos outros estados para cada t. Este método não é viável, pois não garante que as transições entre cada estado de Q_t sejam válidos.

No critério global busca-se a única trajetória ótima da sequência de estados, portanto, maximizar $P(Q \mid O, \lambda)$ o que vem a ser o mesmo que maximizar $P(O, Q \mid \lambda)$, pois $P(O, Q \mid \lambda) = P(Q \mid O, \lambda) * P(O \mid \lambda)$

Problema 3: Aprendizagem

Como pode-se ajustar os parâmetros do modelo $\lambda = \{\Pi, A, B\}$ para maximizar $P(O \mid \lambda)$? O fato de que o comprimento da sequência de observações (dados de aprendizagem) seja finito, não existe uma solução analítica direta (de otimização global) para construir o modelo. Não obstante, pode-se escolher $\lambda = \{\Pi, A, B\}$ tal que $P(O_t \mid \lambda)$ é um máximo local utilizando um procedimento iterativo tal como de Baum-Welch (detalhado em: (BAGGENSTOSS, 2000; HSIAO et al., 2009; OUDELHA; AINON, 2010; CHESHOMI et al., 2010))

A ideia da aplicação é de usar procedimentos de re-estimação que refina o modelo, pouco a pouco, de acordo com as seguintes etapas:

1. Escolher um conjunto inicial de parâmetros λ_0;
2. Calcular λ_1 a partir de λ_0;
3. Repetir o processo até um critério de fim.

No algoritmo do Baum-Welch, a probabilidade de passar por s_i em t e s_j em $t+1$, ou seja, a probabilidade de passar de um estado 1 ao estado 2, gerando O com λ:

$$\xi_1(i,j) = P(q_t = s_1, q_{t+1} = s_j \mid O, \lambda)$$

As varáveis backward e forward permitem escrever:

$$\xi_1(i,j) = \frac{P(q_t = s_1, q_{t+1} = s_j \mid O, \lambda)}{P(O \mid \lambda)}$$

daí:

$$\xi_1(i,j) = \frac{\alpha_t(i) * a_{i,j} * b_j(o_{t+1}) * \beta_{t+1}(j)}{P(O \mid \lambda)}$$

A escolha de um modelo inicial influi sobre os resultados: todos os valores nulos de A e de B no início, permanecem zero ao fim da aprendizagem.

Para se ter uma estimativa adequada do modelo, as re-estimativas são feitas em um conjunto de várias sequências de observações chamadas corpus de aprendizagem. Assim, o comprimento do corpus de aprendizagem influi também sobre os resultados.

4.3 Topologias dos Modelos de Markov

A escolha da topologia influi sobre a qualidade do reconhecimento:

4.3.1 Modelo ergódico

Conforme ilustra a Figura 4.7 um modelo ergódico é um modelo onde todo estado é alcançável a partir de qualquer outro estado em número finito de transições. Este tipo de modelo é mais geral e interessante quando o modelo representa um processo que se quer seguir ou acompanhar as evoluções dos estados.

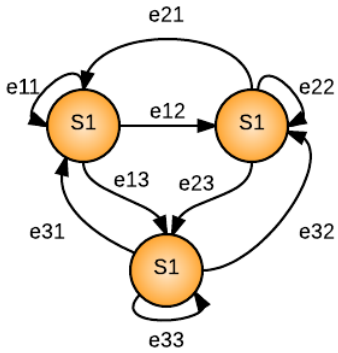

Figura 4.7: Exemplo da topologia ergótica com três estados

4.3.2 Modelo Esquerda-Direta

Conforme ilustra a Figura 4.8, se t aumenta, então os índices dos estados aumentam com igual intensidade. É usado para acompanhar as observações cuja evolução se faz em um determinada ordem tal como o reconhecimento de voz:

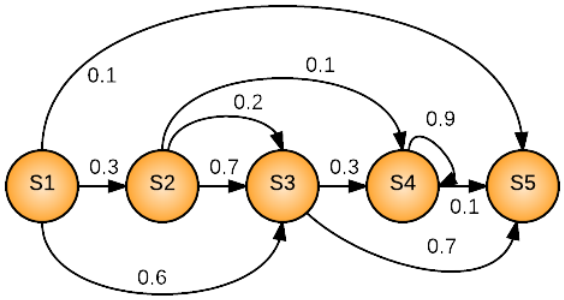

Figura 4.8: Exemplo da topologia Esquerda-Direita

Formalmente: $a_{ij} = 0$ se $j < i$;

$$\pi_i = \begin{cases} 0, & i \neq 1 \\ 1, & i = 1 \end{cases}$$

Da mesma forma, tem-se geralmente as restrições suplementares como: $a_{ij} = 0$ se $j > i + \Delta$ ($\Delta = 2$ no modelo de Bakis). Estes modelos permitem modelar sinais que evoluem com o tempo (é o caso da voz).

4.4 Extensão dos HMM

Os HMM podem ser extendidos como:

1. **Densidade de Probabilidade**

   De acordo com o tipo de densidade de probabilidade de observações, discreta ou contínua, é possível construir dois tipos de modelos de HMM: seja um HMM discreto e um HMM contínuo.

   - HMM discreto (*Discret Hidden Markov Models (DHMM)*): as observações em geral são contínuas, pois elas provêm de fenômenos físicos contínuos. No caso de um HMM discreto, as observações contínuas são quantificadas mediante um dicionário (alfabeto de observações contáveis).
   - HMM contínuo (*Continuous Hidden Markov Models (CHMM)*): embora seja possível quantificar as observações contínuas, pode-se ter séria degradação de informação associada a esta quantificação. Far-se-á, portanto, necessário escolher uma função de densidade de probabilidade de observações contínuas, condicionadas pelos estados do processo.

2. **Tempo de permanência em um estado**

   Uma das desvantagens básicas dos HMM é a falta de informação sobre a variabilidade no tempo de permanência em um estado favorecendo as curtas durações.

   No entanto, é da maior importância em certos processos físicos, por exemplo, na variabilidade da duração dos sons em voz. Dois métodos principais foram desenvolvidos:

   - O método de Ferguson (MITCHELL; JAMIESON, 1993) baseado em um HMM de duração variável Discreta (*Variable Duration Hidden Markov Model VDHMM*)
   - e o método de Levinson (DELSARTE; GENIN, 1986) baseado em um HMM de duração variável contínua (*Continuous Duration Hidden Model CDHMM*).

- **Ordem de uma cadeia**

  Uma limitação básica de Modelos Ocultos de Markov está em sua definição. Eles modelam um processo Markoviano, isto quer dizer que a sua evolução não depende do seu passado, mas apenas do seu estado atual. O processo é de primeira ordem, que não é o caso para muitas aplicações. Em um processo, se o estado futuro depende de k estados precedentes, a cadeia é de ordem k.

4.5 Conclusão

O estudo dos Modelos de Markov e HMM permitiu, primitivamente, fundamentar as bases teóricas associadas a esses modelos. Em segundo lugar, permitiu tomar conhecimento do seu uso em diferentes domínios, tal como a previsão meteorológica.

Em conclusão, duas dificuldades surgem. A primeira dificuldade reside no fato de que os HMM permitem uma integração simples de fontes de conhecimento, mas não fornecem grande parte da visão sobre o processo de reconhecimento que as abordagens de Inteligência Artificial. Portanto, é muito difícil analisar os erros de um sistema HMM, de modo a melhorar o seu desempenho. No entanto, uma incorporação segura do conhecimento pode melhorar os sistemas de uma forma significativa.

A segunda é que a modelagem com base no HMM, em um processo real é eficaz se os parâmetros do modelo são estimados corretamente. Estas estimativas são muitas vezes imprecisas por duas razões: a primeira é que o processo não obedece às restrições do HMM (os estados não são coerentes), a segunda é devido à dificuldade de se obter estimativas fiáveis de todos os parâmetros.

Várias soluções foram desenvolvidas para trazer melhorias em nível de aprendizagem. Estas soluções são variantes e alternativas do algoritmo de Baum-Welch, mas não há nenhuma resposta teórica para a seleção de uma solução ou outro. Geralmente, somente a experiência permite determinar a eficácia e eficiência de uma solução em relação às outras.

Nesta dissertação, usou-se os Modelos de Markov e HMM para emitir as probabilidades de realização ou ocorrência de uma sequência, a previsão de um estado futuro, a previsão de um estado futuro a partir de um estado conhecido e a descoberta (derivação) de um conhecimento (parâmetro faltante) a partir de um parâmetro conhecido ou mais provável de originar ou provocá-lo.

Capítulo 5

ARQUITETURA PARA GERENCIAMENTO DE DADOS NA LOGÍSTICA DE TRANSPORTE

Neste capítulo, propõe-se uma arquitetura genérica que permite, além dos dados do planejamento do tráfego inicial e do realizado final, a obtenção de dados intermediários, ou seja, em um instante t, processo esse chamado também de monitoramento, a fim de poder verificar se tudo ocorre conforme planejado. Além disso, consubstanciou-se a possibilidade de prever os possíveis planejamentos sequenciais do tráfego com base nas informações anteriores, de monitoramento como as condições climáticas.

5.1 Arquitetura geral

Nesta seção, apresentam-se a arquitetura geral proposta para gerenciar os dados e as fontes de dados relativos aos **planejamentos**, as possíveis operações sobre eles, assim como o detalhamento dos principais componentes da presente arquitetura. Esta solução consiste em trazer uma forma de se dinamizar um planejamento, tornando-o mais aderente ao que acontece na prática.

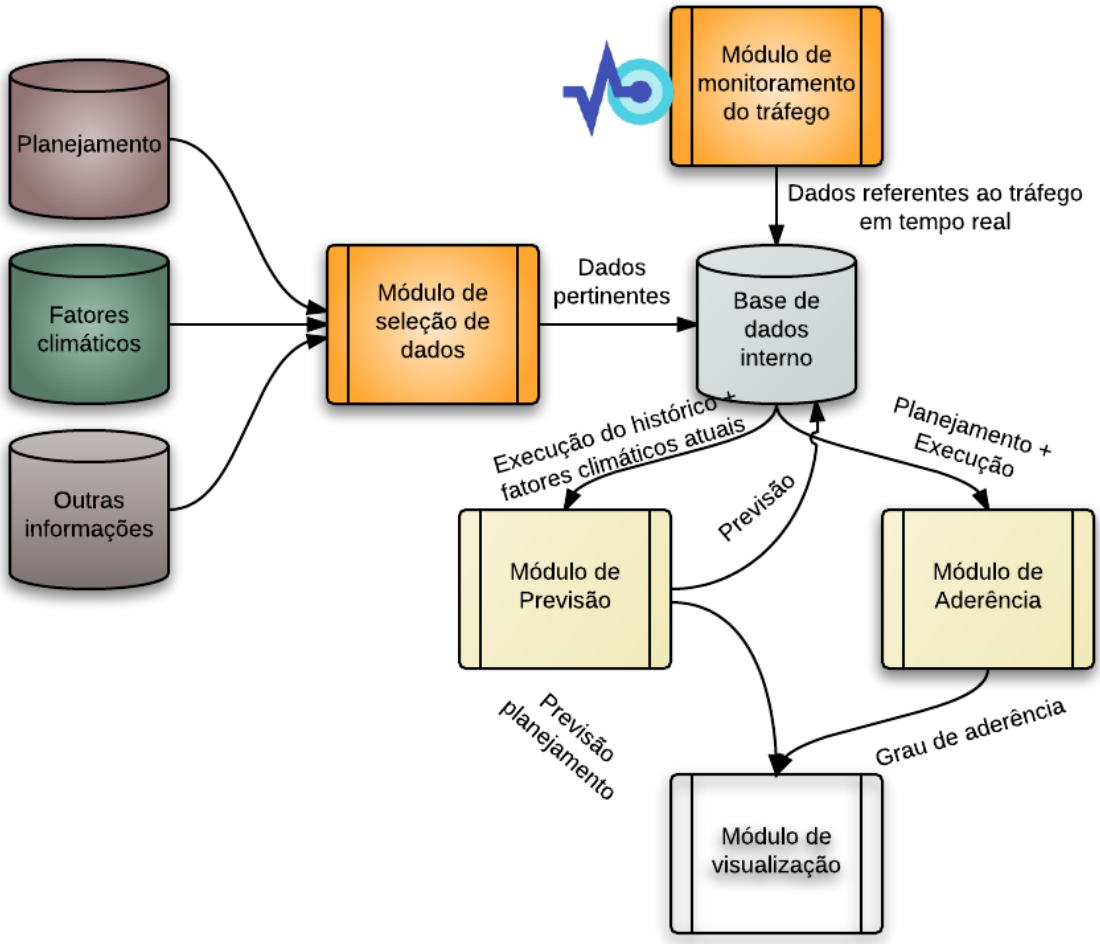

Figura 5.1: Arquitetura geral proposta

A presente arquitetura contém cinco módulos, sendo eles o **Módulo de Seleção de Dados**, **Módulo de Monitoramento de Tráfego**, **Módulo de Aderência**, **Módulo de Previsão** e **Módulo de Visualização**, os quais são apresentados detalhadamente a seguir.

5.1.1 Módulo de Seleção de Dados

O planejamento de rotas para trens envolve vários cenários, de ricochete, variadas informações ou dados. A base de dados **Planejamento** contém variados dados referentes aos cenários de planejamentos (tanto atuais como anteriores) armazenados nas bases empresariais, assim como os cenários realizados correspondentes aos planejamentos anteriores. A base de dados **Fatores climáticos**, por sua vez, fornece as informações climáticas tanto anteriores como atuais (em tempo real) relacionadas ao tempo e espaço em que os veículos se encontram. Enquanto base de dados **Outras informações** fornece demais informações, tal como a condição da malha, etc.

Desta forma, o principal papel do presente módulo é o de selecionar, a partir das fontes de dados mencionadas, os dados pertinentes para o processo de previsão e cálculo de aderência, tais como, para o planejamento de veículos, seu horário de saída, itinerário (conjunto de pontos ”vias”, velocidade ”km/h”previsto, sentido), tipo de veículo (velocidade ”min/max”, comprimento, prioridade). Por outro lado, o realizado final que fornece o conjunto de horários relacionados com o conjunto de vias (posições) pelas quais os trens passaram. Além disso, aos dados climáticos coletados durante as circulações anteriores.

5.1.2 Módulo de Monitoramento de Tráfego

Com o objetivo de tornar o planejamento mais aderente ao que acontece na prática, este módulo consiste na coleta de informações (preferencialmente em tempo real) relativas ao movimento (circulação) de trens em um segmento, assim como seu estado atual, o que implica na obtenção dos horários de chegada e partida de cada veículo em cada segmento, assim como sua velocidade média. Neste fito, o monitoramento de veículos em movimento (circulações) no segmento é um elemento-chave. Pela definição, um segmento liga dois pontos distintos, seja um de origem e outro de destino, possibilitando assim a circulação. Em teoria, esta conexão gera leis de causa e efeito.

Uma maneira de monitorar o fluxo é pegar informações de veículo, um por um. Sabendo-se que existem muitos veículos, não é simples, mesmo com as ferramentas computacionais vigentes, analisá-los individualmente. Além disso, monitorar os diferentes fluxos constituídos de mais de duas posições nas vias também é custoso. Por esta razão, este trabalho se concentra em prever o planejamento ou um possível replanejamento aderente, com base nas informações recolhidas durante o monitoramento, estabelecendo como novo ponto de partida ou origem a posição dos trens até completar seu ciclo previsto.

5.1.3 Repositório de Dados Internos

Ainda com o intuito de tornar o planejamento mais aderente ao que acontece na prática, o Repositório de Dados Internos consiste em uma coleção de dados e/ou informações provenientes dos módulos de *Seleção* (dados pertinentes ao planejamento corrente e os planejamentos passados), *Monitoramento* (dados referentes ao cenário que está ocorrendo na prática).

É inevitável dispor de dados ou informações para que alguma providência possa ser tomada. Por esse motivo, percebe-se a necessidade de um sistema de armazenamento de informações do cenário da circulação de trem, de origem ao destino, nos pontos (vias ou segmentos) delimitados

e precisos. A Figura 5.2 ilustra o diagrama de classes do repositório de dados proposto.

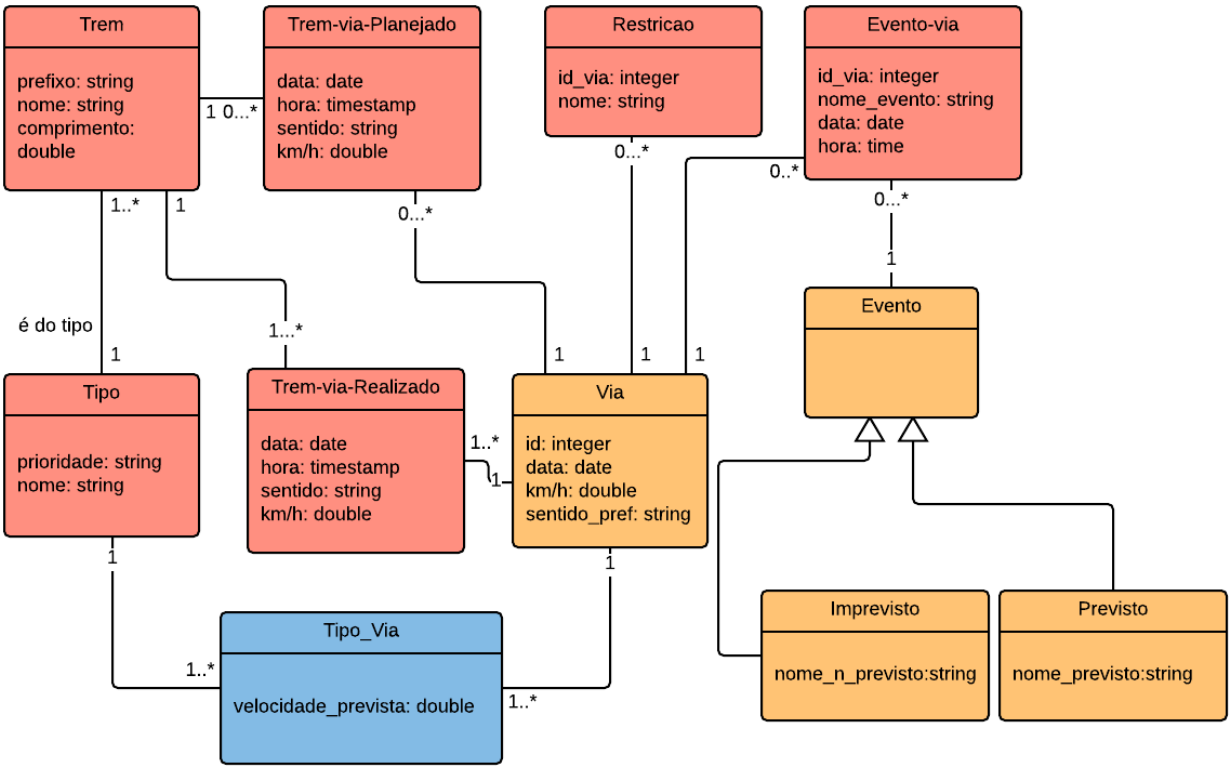

Figura 5.2: Diagrama de Classe da base de dados proposta

5.1.4 Módulo de aderência

Dispondo-se da planilha de informações, tais como itinerários planejados dos trens, dados referentes à evolução da trajetória realizada até um determinado instante *t*, ambos procedentes, indiretamente, do Repositório de Dados Internos, pode-se realizar os procedimentos relativos ao *cálculo de aderência*, isto é, verificar se o cenário planejado corresponde ao que está acontecendo ou não.

Chama-se de não aderência a não correspondência entre dois eventos comparados com base nas mesmas métricas. Os eventos comparados neste trabalho são o planejamento de rotas para um trem e o efetivamente realizado pelo trem, as métricas utilizadas neste contexto são a velocidade média, o tempo gasto por segmento e a posição atual.

A não aderência de um trem pode interferir na aderência de outros trens. Por isso, é preciso sempre verificar as causas para não aderência, sendo as possíveis causas a manutenção, condição da malha e questões meteorológicas.

Partindo do fluxograma (Figura 5.3), foi elaborada a arquitetura do módulo de aderência (Figura 5.4) como primeiro passo da proposta deste trabalho para permitir analisar a aderência no determinado instante *t*, e não apenas no final da circulação como foi apresentado na Figura 1.3.

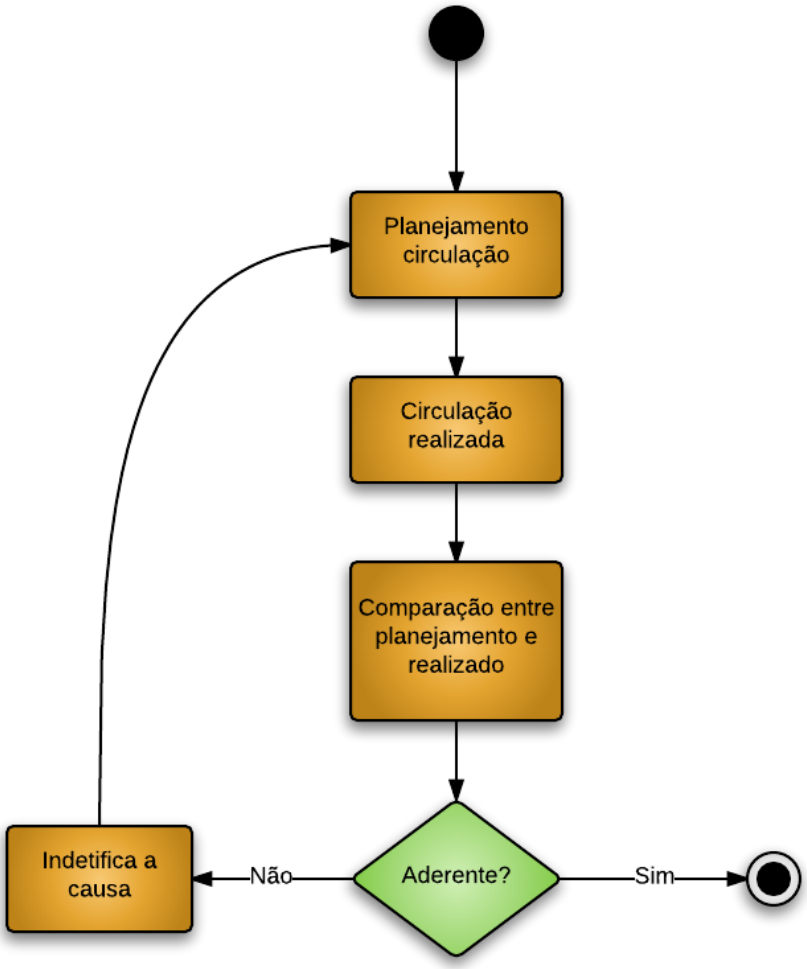

Figura 5.3: Fluxograma de idealização da proposta

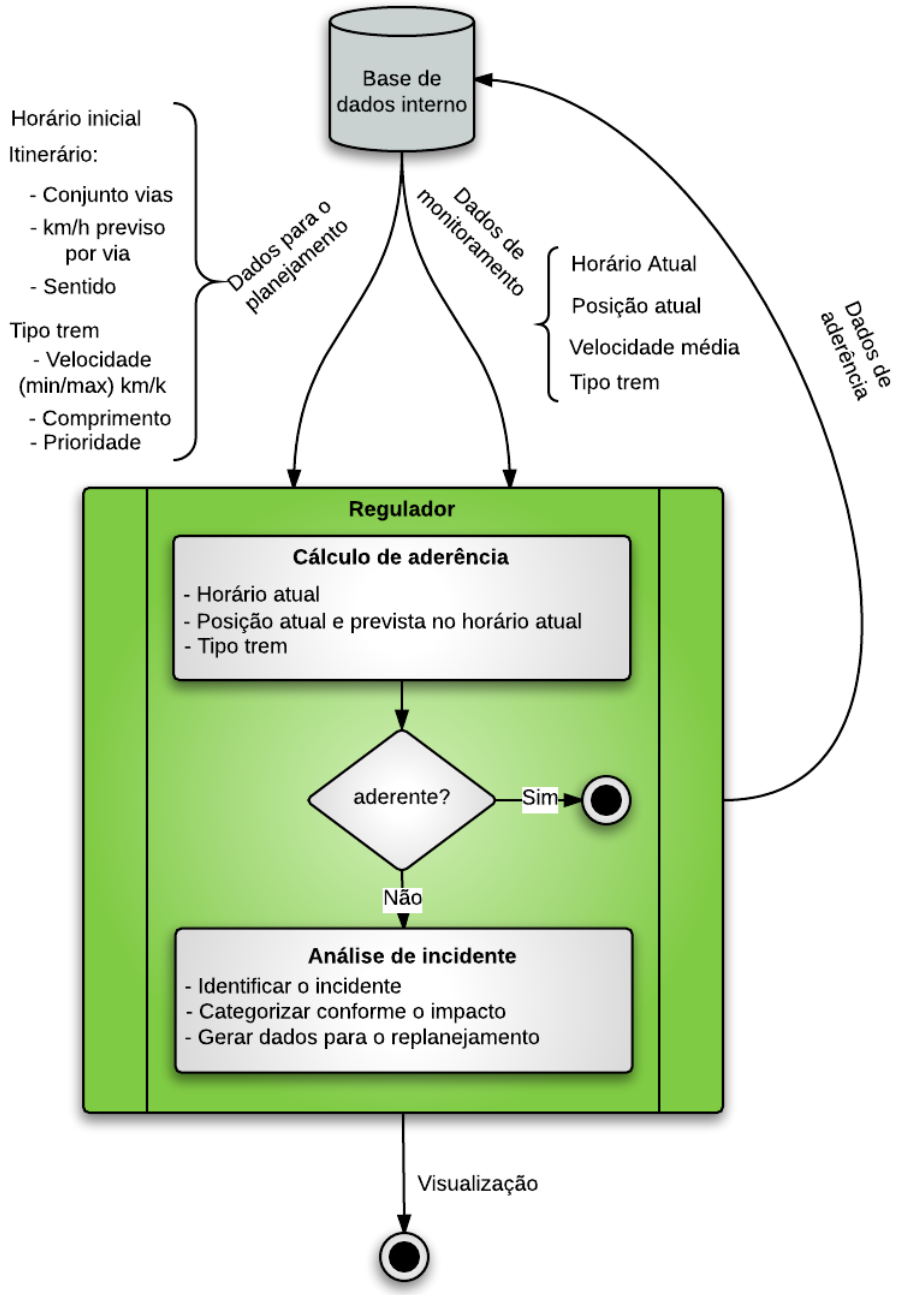

Figura 5.4: Módulo de aderência proposta

Onde:

H1: identificar o incidente;

H2: categorizar o impacto conforme o impacto.

Para se ter um entendimento claro sobre esta arquitetura, faz-se necessário explicar com detalhes a função de cada um dos componentes contidos na Figura 5.4, os quais formam o módulo de aderência.

1. *Base de dados*: É indispensável dispor de dados ou informações para tomar certas providências ou realizar certas operações. Por esse motivo, percebe-se a necessidade de um sistema de armazenamento de informações do cenário da circulação de veículos, de origem ao destino, nos pontos (vias ou segmentos) delimitados e precisos. Como pode-se notar (Figura 5.4), tanto as informações oriundas do monitoramento do tráfego realizado quanto do *Planejamento* são enviadas para a base de dados, e o *Regulador* serve-se dessas informações para efetuar seu processamento.

   Além disso, cada vez que não houver aderência, acontecerá uma geração de dados para o replanejamento que, por sua vez, serão armazenados na base e recuperados pelo *Planejamento* a fim de gerar, se necessário, um replanejamento aderente.

2. *Dados para o Planejamento*: é o planejamento inicial de circulações de trens. Este contém as informações concernentes ao seu horário de saída, itinerário, nome do veículo e seu tipo.

3. *dados de monitoramento*: representa a circulação efetivamente realizada em um determinado tempo; não apenas no final do percurso, mas a cada instante que se solicita o monitoramento do trem, onde obtém-se informações relativas ao horário atual, posição atual (via em que o trem se encontra), estado atual (velocidade ”km/h”), o nome do trem e seu tipo.

4. *Regulador*: Este submódulo, uma das principais contribuições desta dissertação, tem como objetivo principal dinamizar o replanejamento.

   Inicialmente, realiza-se a checagem de aderência, isto é, confere se existe uma correspondência biunívoca entre o planejamento e o realizado. Este processo consiste em comparar cada dado inicial com seu correspondente realizado. Por exemplo, ao saber a posição e o horário atual do trem, verificar se converge ao que foi planejado. O termo *converge* aplicou-se, pois tendo em vista a não perfeição de tudo que não segue as leis da natureza, como por exemplo o ciclo realizado pelo trem, deve-se considerar a aproximação ao invés da igualdade. Para tal, torna-se imprescindível definir um certo grau de liberdade. Um grau de liberdade é um teste estatístico para avaliar a equação de uma série de dados a uma família de leis de probabilidade ou testar a independência entre duas variáveis aleatórias. Se, levando-se em conta o grau de liberdade, também chamado *intervalo de confiança*, o teste apresentar uma aderência, então o ”realizado”corresponde ao ”planejado”, em vista disso, não há necessidade de um replanejamento. Caso contrário, será então preciso identificar a causa, o que também é previsto no Regulador.

Obviamente, a não aderência deixa bem claro que ocorreu um evento, podendo ser previsto ou imprevisto, cujo impacto influenciou negativamente na realização do planejamento inicial. Neste intento, é indispensável identificar o incidente (o que?), categorizar conforme o impacto (consequências?) e, por fim, gerar dados para o replanejamento.

O cenário de planejamento atual, do contexto ferroviário estudado neste trabalho, envolve:

- Uma entidade veículo
- A trajetória a ser percorrida, que por sua vez consiste em:
  - ponto de origem;
  - ponto de destino;
  - conjunto de vias pelas quais o veículo irá circular de origem até o destino;
  - tempo médio gasto por cada via (dependendo do tipo de trem, e as restrições referentes). Consideram-se os fatores climáticos e a condição da malha nas melhores condições possíveis.

Para exemplificar, considere-se a configuração da malha (Figura 5.5).

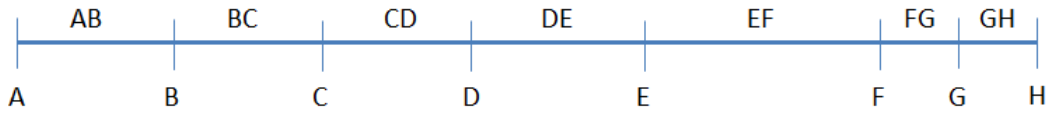

Figura 5.5: Exemplo de configuração da malha

Os pontos (A, B, C, D, E, F, G, H) representam o início e/ou fim de um segmento (via). E, os pares (AB, BC, CD, DE, EF, FG, GH) representam os conjuntos de vias, onde o primeiro símbolo representa o início da via e o segundo seu fim. Por exemplo, a anotação AB representa a via que tem como ponto inicial o ponto A e ponto final B. A cada via pode-se associar sua distância em quilômetro, conforme a Figura 5.6.

Figura 5.6: Exemplo de configuração da malha associada com as distâncias de vias

O planejamento de um veículo, neste caso um trem, do tipo x_i na malha sobrejacente obedecerá, além de suas próprias restrições, as restrições referentes à malha, como por exemplo, a

velocidade prevista desse tipo x_i de trem numa determinada via $v_i \in V$, onde V é o conjunto de vias.

Seja **T1-A** o trem **T1** do tipo **A**, considere-se o planejamento a seguir sobre a malha da figura 5.6:

- ponto de saída **A**;
- ponto de destino **F**;

ou seja, $T1 - A(A \rightarrow F)$. Usar-se-á T1-Ai, $i \in \{1, 2, \ldots, n\}$ para indicar a i-éssima circulação realizada pelo trem T1-A.

Constata-se que o conjunto (AB, BC, CD, DE, EF) é o conjunto de vias pelas quais o trem **T1-A** irá circular, com base no seu itinerário planejado. Admitindo-se que a velocidade máxima do trem é estimada a 8 km/h; a tabela 5.1 contém as restrições de velocidade atribuídas ao trem T1-A, indicando as velocidades previstas para esse tipo de veículo nas respectivas vias.

Vias	Velocidade máxima
AB	8kmh
BC	6kmh
CD	8kmh
DE	7kmh
EF	6kmh
FG	8kmh
GH	6kmh

Tabela 5.1: Exemplo de restrições de velocidades previstas para um tipo de trem sobre determinadas vias.

Tendo em vista as configurações acima, suponha-se o seguinte planejamento com respectivos tempos de duração de circulação e a velocidade máxima:

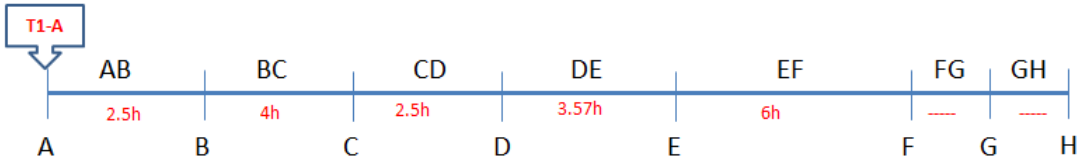

Figura 5.7: Exemplo do planejamento associado com o tempo gasto (previsto) em cada segmento

Tabulando esses dados, temos a Tabela 5.2.

Vias	Tempo gasto
AB	2h50 (150min)
BC	4h00 (240min)
CD	2h50 (150min
DE	3h57 (214.2min)
EF	6h00 (360min)

Tabela 5.2: Representação tabular do exemplo do planejamento associado com o tempo gasto (previsto) em cada segmento referente à figura 5.7

Portanto, o tempo total da viagem é aproximadamente 18h57, ou seja, 1114.2 minutos.

Entretanto, nem sempre o trem circula com a velocidade máxima. Neste caso, ainda considerando todos os fatores (ambientais e não ambientais) favoráveis, pode-se fazer um planejamento com certo grau de tolerância de variação de velocidade, ou seja, intervalo de confiança. Assim sendo, supondo que o grau de liberdade seja de 20% para trás (atraso), e 10% para frente (antecipação) se $vTmax > vPvia$, e 0% caso contrário. Onde vTmax e vPvia, representam a velocidade máxima do trem e sua velocidade prevista sobre a via em curso, respectivamente.

Outrossim, estabelecem-se certas condições indicando que o trem está circulando normalmente, em atraso ou adiantado com relação ao tempo estimado para percorrer uma via v_i. Relações essas que chamar-se-ão, doravante, de *aderência ou grau de aderência*, conforme ilustra a tabela 5.3 a seguir:

Via	Tempo estimado (min)	Normal	Atrasado	Adiantado
AB	150	$120 < t \leq 165$	$t > 165$	$t < 120$
BC	240	$192 < t \leq 264$	$t > 264$	$t < 192$
CD	150	$120 < t \leq 165$	$t > 165$	$t < 120$
DE	214.2	$171.36 < t \leq 235.62$	$t > 235.62$	$t < 171.36$
EF	360	$288 < t \leq 396$	$t > 396$	$t < 288$

Tabela 5.3: Tabela de grau de liberdade

Admitindo que o planejamento para o trem $T1 - A$ seja de acordo com a figura 5.8, e que o trem esteja em curso de circulação conforme a Figura 5.9.

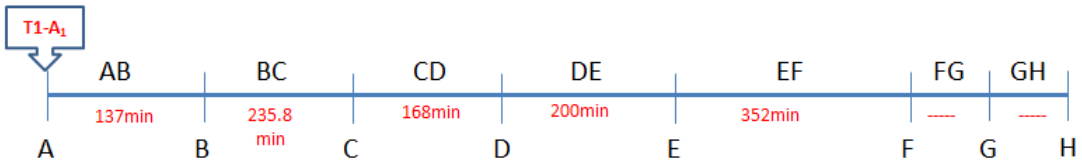

Figura 5.8: Planejamento da circulação do trem T1-A

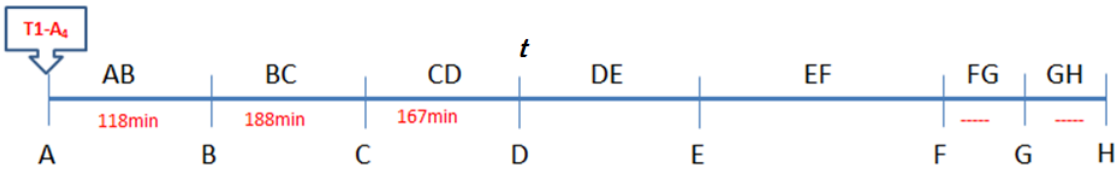

Figura 5.9: Circulação realizada do trem T1-A no instante *t*

Neste instante *t* em que o trem se encontra é possível realizar a checagem de aderência, isto é, conferir se existe uma correspondência entre a circulação planejada e o que já foi realizado, levando-se em consideração o grau de liberdade da Tabela 5.3. Falar-se-á de aderência se os valores, por exemplo, o tempo gasto nos segmentos do realizado e planejado, pertencerem ao mesmo intervalo definido. Estas operação são realizadas no *módulo de aderência* no *Regulador*.

No cenário acima, pode-se perceber que, no segmento **AB** o tempo previsto foi de **137** minutos, mas na prática, o trem gastou **118** minutos. Após checagem, observa-se que, neste caso, não houve aderência, uma vez que o tempo realizado pertence ao intervalo de grau aderência *Adiantado*, portanto, diz-se que o *realizado* não corresponde ao *planejado*; em vista disso, há necessidade de descobrir as causas dessa não aderência e, se necessário (por exemplo, se a prioridade do trem for alta, deve-se melhorar as condições para atingir a meta colocada), fornecer dados para um possível replanejamento. Caso contrário, não será preciso efetuar qualquer operação. Vale ressaltar que tanto o atraso quanto o adiantamento podem provocar distúrbios na malha, pois existem outros trens disputando as mesmas vias.

Na próxima seção apresenta-se outro módulo, **Módulo de Predição**, responsável por registrar as ocorrências de eventos inesperados, assim como os fatores ambientais(climáticos e não-climáticos) ao longo do percurso dos trens, e, com base neles, prever futuros comportamentos do tráfego.

5.1.5 Módulo de Predição

O sistema de previsão, conforme indica a Figura 5.10, baseia-se em:

- Os dados de entrada de um sistema de informação e os resultados da previsão anterior;

- Processo de correção e validação de previsões fornecidas;
- Um modelo matemático construído a partir de hipóteses estatísticas e informações que descrevem o fenômeno a prever;
- Um desenvolvimento computacional do sistema de previsão, permitindo uma atualização regular e automatizada de previsões; e
- As práticas de gestão para integrar o negócio de previsão na gestão operacional.

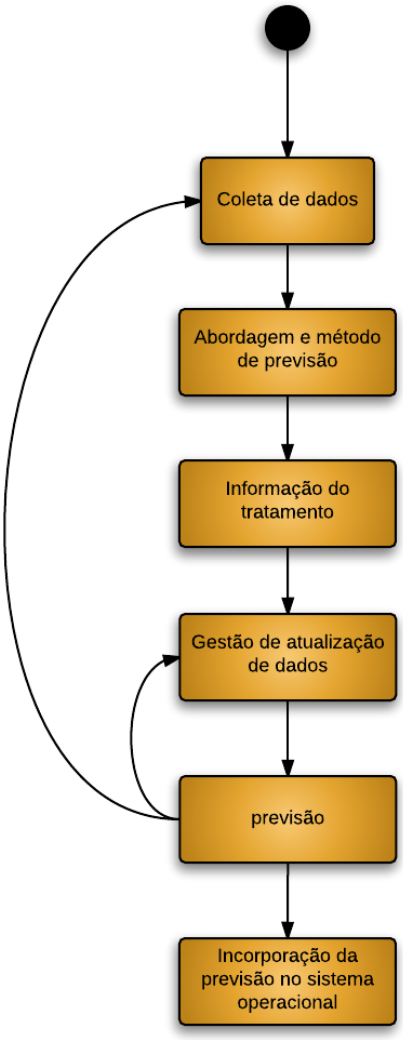

Figura 5.10: Fluxograma do sistema de previsão.

Deste modo, antes de realizar uma série de processamentos, de análise de dados e cálculos de previsão, é preciso responder às perguntas: O que deve permitir a previsão de planejamento? Quais são as fontes de dados de previsão? Quais são os parâmetros de previsão de circulações?

Quais são as técnicas a escolher para prever o planejamento (replanejamento) das circulações? Qual grau de erro de previsão pode ser percebido como aceitável? Depois de responder às perguntas, os dados internos são coletados, analisados e processados, a fim de fornecer um histórico confiável. Na sequência-se, busca-se modelos matemáticos de previsão compatíveis com os dados disponíveis. Tais modelos são informatizados para serem executados em intervalos regulares de acordo com os dados atualizados. A informação da previsão pode ser corrigida se necessário. No final, a previsão é retornada ao módulo de visualização.

A implantação deste sistema de previsão requer que se tenha o conhecimento e as competências que abranjam:

- a capacidade de identificar as necessidades reais em termos de previsão e restrições a elas associadas;
- a aplicação de diferentes métodos de previsão;
- os procedimentos permitindo selecionar os métodos de previsão adequados às situações específicas;
- suporte organizacional para a aplicação e utilização de métodos de previsão.

A implementação do Módulo de Previsão neste trabalho, exigiu que as seguintes perguntas fossem devidamente respondidas para guiar as decisões de projeto do módulo:

- **O que deve permitir a previsão de planejamento?**
  Encontrou-se três objetos de reflexão em torno de previsões no que concerne ao planejamento:
  - previsão de planejamento, definida como uma projeção para o futuro do planejamento, dado um conjunto de condições ambientais;
  - o planejamento operacional, definido como um conjunto de decisões e ações de gestão tomadas para atender ou adequar as previsões do planejamento;
  - as metas de circulação a serem alcançadas.

  Na verdade, existem relações de dependência entre a previsão de circulação, o planejamento operacional e as metas de circulação. É imprescindível observar que a previsão de circulação é a condição *sine qua non* para todo o planejamento e que, daí, ela também é uma ferramenta essencial para o controle operacional. Portanto, a previsão deve permitir ter o controle daquilo que vai se realizar com maior grau de probabilidade.

- **Quais são as fontes de dados de previsão?**
  Para se fazer previsões com base em métodos estatísticos e computacionais, é preciso dispor de um histórico do que se deseja prever. Isto pode parecer óbvio, mas nem sempre as operadoras de tráfego armazenam sistematicamente todas as informações. De qualquer forma, a operação mais importante no processo de previsão é a recuperação do histórico de ocorrências passadas.

  Outra questão é saber quais os dados disponíveis na empresa que podem ser utilizados na previsão? Quais são os dados externos que explicam como tais dados externos podem auxiliar na previsão? Qual é a qualidade dos dados? Qual é a confiança? Qual é o nível de detalhe? Com que frequência foram obtidos? Qual é o tipo dos dados? A resposta a essas perguntas irá determinar os modelos e técnicas de se usar.

  De fato, as circulações são muitas vezes influenciadas por eventos externos (sazonalidade de uso dos trens e vias, meteorologia, manutenção, ...). Ao conseguir controlar essas variáveis, prevê-las e até mesmo provocar ou influenciá-las, sê-lo-á vantajoso usar as técnica mais adequadas, tais como as usadas neste trabalho, para minimizar seus impactos.

  Neste trabalho, as fontes de dados de previsão são a base de dados da empresa ABC (com dados reais), os sensores que tratam de coletar as informações sobre a condição da malha, presença de obstáculos (por exemplo, folhas mortas e água na malha), assim como as bases de dados e os serviços web meteorológicos.

- **Quais são os parâmetros de previsão de circulações?**
  Os parâmetros a serem considerados, os quais foram identificados como os principais fatores que influenciam no planejamento são:

  - **Parâmetros climáticos**
    - Temperatura (em Celsius)
    - Chuva
    - Vento (em km/h)
    - Precipitação (em porcentagem)
    - Umidade
    - Tempo
  - **Parâmetros não climáticos**
    - Condição da malha
    - Folhas mortas na malha

- Água na malha
- Período do dia

De fato, as condições climáticas adversas podem ter consequências graves para o tráfego ferroviário. Faz-se necessário estar em constante contato com os serviços meteorológicos a fim de antecipar esses riscos meteorológicos. Além disso, deve-se dispor de sensores ou agentes de infra-estrutura para monitorar constantemente as instalações sensíveis (sinais mecânicos, comutadores, etc.).

Todos os meios de transporte são suscetíveis a diversos a algum grau, às condições meteorológicas e ao clima afirma o Ministério de Recursos Naturais Canadense (RNC: *Ressources naturelles Canada*) [1]. De acordo com a Sociedade Nacional de Ferrovia Francesa [2] (SNCF: *Société Nationale de Chemin de fer*), existem fatores que podem levemente ou gravemente influenciar ou perturbar a circulação dos trens e, em seguida, levá-los a não-aderência do planejamento inicial. Entre eles são a os parâmetros climáticos e não climáticos a seguir.

Parâmetros climáticos

As condições meteorológicas enfraquecem as infra-estruturas ferroviárias, assim como os equipamentos. Com a aproximação do inverno, frio, neve, gelo, geada, vento ou chuva pesada, todas estas condições meteorológicas são de sérias consequências sobre a circulação de trens. Perturbações que podem variar de pequenos atrasos a uma interrupção total do tráfego para reparar as vias e catenária. Os parâmetros aqui apresentados são as principais causas de tais perturbações.

– *Temperatura*

  A deformação das malhas é um problema grave de segurança e redução do desempenho. Quanto mais alta a temperatura mais deformação há na malha. O trem pode sofrer restrições de velocidade devido às temperaturas elevadas, afirmam os especialista da **SNCF**.

  Vários limites de temperatura podem ser estabelecidos, permitindo que se avalie o estado da malha remotamente, em tempo real, e reagir para cada variação de acordo com os procedimentos predefinidos. Por exemplo, quando a temperatura da malha

[1] http://www.rncan.gc.ca/environnement/ressources/publications/impacts-adaptation/rapports/evaluations/2004/ch8/10218#ar acessado em 15-08-2015

[2] http://www.sncf.com/fr/presse/article

na faixa de 49-53 ° C, os trens podem ser submetidos a restrições de velocidade devido às altas temperaturas.

No inverno, também pode-se ter avisos quando a temperatura ferroviária é abaixo de 0 ° C (ou outras variações definidas pelo usuário) para ajudar a agendar a manutenção invernal.

– *Precipitação*

Deve-se coletar as informação a respeito da precipitação, sua intensidade e suas quantidades acumuladas. Estes dados permitirão que os especialistas sejam avisado com antecedência dos possíveis impactos no caso de chuvas pesadas, por exemplo.

– *Chuva*

Se as chuvas são violentas, elas podem causar deslizamentos de terra que por sua vez provoca poluição e obstáculos na malha. Estes fluxos de lama, portanto, interrompem o tráfego de trens.

A duração das reparações é difícil de se avaliar se a chuva continua caindo, os reparos podem ser prorrogados e o tráfego pode ser interrompido em qualquer dos eixos para se obter uma segurança máxima das vias.

As chuvas podem ser fortes, moderadas e fracas. Dependendo da intensidade, podem levar à perda de velocidade do trem, no pior caso, à interrupção.

– *Vento*

Um vento forte acompanhado de chuva pesada provoca a presença de detritos ou resíduos nas plataformas e vias que podem interromper o tráfego enquanto se retiram os escombros.

O vento pode ser fraco, moderado ou violento (podendo virar o trem). Neste caso, a circulação pode até ser interrompida.

– *Neve*

No inverno, o excesso de neve pode causar projeções de gelo em trens. Este fenômeno provoca regularmente danos materiais e impacto sobre o tráfego. Por exemplo, a neve pode reduzir a velocidade do trem até 50 %.

No inverno, fortes nevascas podem cobrir as vias. Quando um trem passa, a neve levantada devido à alta velocidade se acumula abaixo do trem. Com o frio, a neve endurece e se transforma em um bloco denso de gelo. Quando dois trens de alta velocidade se cruzam, o apelo do ar causado pelo cruzamento retira os blocos de neve. Levantados, eles agem como projéteis que danificam janelas e equipamentos ferroviários.

Este fenômeno é observado principalmente em linhas de alta velocidade, com impactos medidos a 600 km/h. De acordo com a velocidade de projeção do gelo, os danos podem variar desde a fissura de uma janela à deterioração de elementos essenciais que levam à rescisão ou parada do trem, afirma a **SNCF**. Dentre esses elementos: os sensores abaixo do trem permitem a transmissão de informações para a cabina do condutor.

Parâmetros não climáticos

– *Folha mortas ou árvore caída*

Durante os tempos severos as árvores e folhas podem cair sobre as vias, criando obstáculos na malha. Isso cria perturbações na tração do trem. Uma árvore sobre as vias é muito difícil de se retirar porque não se deve danificar as instalações ao redor de vias. A folhas mortas na malha pode afetar a tração do trem até 30 % abaixo do normal.

- **Quais são as técnicas a escolher para prever o planejamento (replanejamento) das circulações?**

  Para se escolher uma técnica, é preciso ter conhecimento do tipo de dados com que se lida, a quantidade de informações, assim como sua complexidade. Assim, visto que neste contexto trata-se de um grande volume de dados heterogêneos, sendo eles qualitativos e quantitativos (discretos ou contínuos), escolheu-se como técnica de classificação de dados a *árvore de classificação*, levando-se em consideração as suas características.

  Além disso, tendo em vista que a previsão é uma *aproximação* e não uma exatidão, faz-se necessário fornecer uma previsão com um certo índice de confiança, isto é, o grau de probabilidade para que aquilo que se prevê ocorra. Outrossim, se faltar um determinado dado necessário para a previsão, o qual pode ser derivado a partir do outro dado disponível, faz-se esta derivação considerando-se o índice de confiança. Logo, torna-se imprescindível o uso das técnicas probabilísticas, dentre as quais escolheu-se os modelos de Markov, tendo em vista a sua eficiência.

5.2 Conclusão

Este capítulo apresentou uma arquitetura capaz de realizar as operações requeridas para um contexto de planejamento e aderente. Esta arquitetura, composta por cinco módulos, monitora a evolução do tráfego de um veículo, verifica se a evolução corresponde ao planejado e, com

base em outras informações tais como dados anteriores e previsão do tempo, prevê uma futura evolução, indicando o índice de confiança ou grau de probabilidade de ocorrência de tal evolução.

O capítulo a seguir conceitualiza a aplicação de análise de dado sequencial empregada nos dois módulos principais, a saber, cálculo de aderência e previsão.

Capítulo 6

CONCEITO E APLICAÇÃO DE ANÁLISE DE DADOS SEQUENCIAL

Neste capítulo é apresentada a fundamentação conceitual empregada nos módulos de Aderência e Previsão para análise e previsão de dados sequenciais heterogêneos, com base no acoplamento entre uma abordagem de classificação por árvore de decisão, os Modelos de Markov e cálculo de distância de similaridade.

6.1 Introdução

Neste capítulo, é apresentada a fundamentação conceitual empregada nos módulos de **Aderência** e **Previsão** para análise e previsão de dados sequenciais heterogêneos. Propõe-se uma abordagem que combina a classificação por árvore de decisão e os Modelos de Markov, fornecendo para cada uma das classes de sequências obtidas uma dupla descrição: (a) proporciona um conjunto de *tipos de sequências (com relação ao tipo de aderência)*[1] que refletem as propriedades de suas classes e garantam também uma separação *disjunta* das mesmas *vis-à-vis* de outras classes de partição, (b) corresponder um *modelo probabilístico* de geração de dados (cadeia de Markov) que resume as relações entre os diferentes estados das sequências da classe. Este modelo proporciona, assim, uma melhor *interpretabilidade* das classes construídas e pode ser aplicado para classificar novas sequências (*classificação*) e estimar a sua evolução futura (*previsão*).

[1] No contexto deste trabalho, tem-se três tipos de classes diferenciadas pelo grau de aderência. Os graus de aderência encontrados são: Normal (o trem realizou a circulação no tempo determinado), Atrasado (o trem realizou a circulação com atraso com relação ao tempo previsto) e Adiantado (o trem realizou a circulação com antecedência com relação ao tempo previsto)

6.2 Visão da abordagem para análise de dados sequenciais

Nesta seção, é apresentada uma abordagem híbrida que combina as abordagens probabilísticas com as abordagens de classificação com base na noção de proximidade, buscando atender as expectativas da análise de dados sequenciais. Neste contexto, uma visão para análise de dados sequenciais foi elaborada, a qual explora as vantagens de ambas as abordagens: árvore de classificação e os modelos *Markovianos* para interpretar facilmente as classes, identificar a classe de pertencimento de novas sequencias e prever a evolução de determinadas sequências. Vale ressaltar que, neste caso, sabe-se com antecedência o número de classes que caracterizam as trajetórias estudadas. A abordagem aplica-se a sequências de qualquer comprimento cujos estados são descritos por variáveis heterogêneas e usa a Distância de Edição *DEFlex*[2], apresentada na seção 6.2.2, para refletir melhor a presença de estados comuns entre as sequências de tamanho diferente.

Seja $S = \{S_1, S_2, \ldots, S_n\}$ um conjunto de n sequências que podem ser de comprimentos diferentes, onde cada sequência $S_i = \left\{e_{i,1}, e_{i,2} \ldots, e_{i,T_i}\right\}$ é dada por um conjunto de T_i estados ($e_{i,j}$) observados sucessivamente por um trem i, e j de 1 até T_i, informa o tempo de observação do evento $e_{i,j}$. Cada estado $e_{i,j}$ é descrito sobre um conjunto de p variáveis heterogêneas $Y = \left\{Y_1, Y_2, \ldots, Y_p\right\}$. O processo de classificação considerado neste capítulo visa a estruturar as sequências contidas em $S = \{S_1, S_2, \ldots, S_n\}$ em função de suas similaridades, sob forma de um conjunto de classes homogêneas e significativas.

Neste caso em que tem-se sequências complexas (Figura 6.1), onde está previsto, por exemplo, no segmento AB o tempo de 137min conforme as seguintes condições: temperatura média, sem chuva, vento moderado, tempo ensolarado, boa condição, sem folhas mortas na malha, sem água na malha, e a circulação ocorrendo no período da manhã. Nesta situação, faz-se necessário estruturá-las e homogeneizá-las a fim de explorá-las para a tomada de decisão, que é o objetivo deste trabalho. Assim, considera-se, a seguir, que cada estado $e_{i,j}$ não é mais descrito por um conjunto de p variáveis heterogêneas, mas por um grupo $g_{i,j}$ obtido por classificação das p variáveis descritivas. A sequência construída torna-se, portanto, $S_i = \left\{g_{i,1}, g_{i,2} \ldots, g_{i,T_i}\right\}$.

[2]DEFlex: Distância de edição flexível

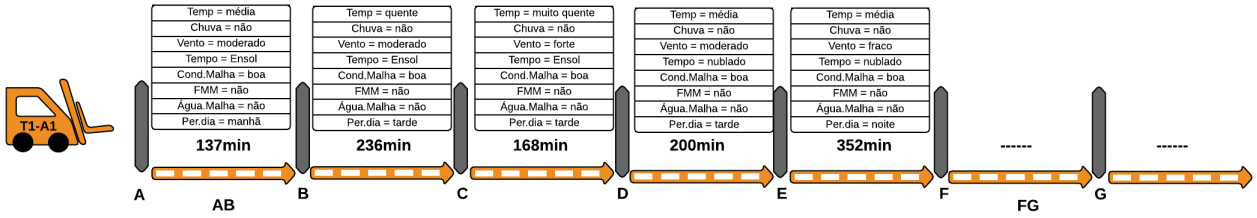

Figura 6.1: Exemplo de sequências com dados heterogêneos complexos.

6.2.1 Geração do Grupo Homogêneo de Aderência (GHA)

Conforme apresentado na seção anterior (Seção 6.2), onde considera-se cada estado $e_{i,j}$ não mais como um conjunto de p variáveis heterogêneas, mas por um grupo $g_{i,j}$ obtido por classificação das p variáveis descritivas. Assim, a nova representação da sequência construída como sendo, $S_i = \{g_{i,1}, g_{i,2} \ldots, g_{i,T_i}\}$, neste trabalho define-se como homogenização de *grupo homogêneo de aderência* (GHA).

O caso estudado nesta dissertação, esta visão de análise foi aplicada nos dados sequenciais relativos às bases de dados de trajetórias de trens. Trata-se de um conjunto de trajetórias de circulações de trens dos anos 2014/2015, fornecidas pela Empresa ABC, envolvendo algumas regiões brasileiras. A Figura 5.2 ilustra o diagrama de classe desenvolvido para coletar as informações relativas ao cenário de planejamento.

Este conjunto de dados sequenciais contém o conjunto de circulações efetuadas para vários trens. Para manter o termo de confidencialidade, associamos cada trem a um identificador regional único gerado a partir de suas informações (tipo, prefixo, velocidade máxima). Esse identificador é utilizado para a trajetória de trem em um determinado período.

Com base nas planilhas de dados construídos a partir das bases de dados do sistema ferroviário, foi possível obter para cada uma das circulações de trens um conjunto de características *estáticas* e *dinâmicas*. As características estáticas são informações sistematicamente cadastradas, concernentes ao trem (tipo, velocidade máxima, prefixo) e as informações dinâmicas são armazenadas durante a trajetória percorrida (dados climáticos, condição da malha, vento, tempo gasto (*duração em minuto*) por segmento (via), etc.), às quais se acrescenta o grupo homogêneo de aderência (GHA) da circulação num determinado segmento.

O GHA constituí a classe de aderência obtida a partir da instância formada pelo conjunto dos parâmetros coletados durante a circulação. Estas informações são coletadas em cada segmento, levando, portanto, às conclusões parciais a tempo real. No final da circulação, pode-

se, se necessário, verificar o GHA global. Entretanto, as informações coletadas durante uma circulação não consistem apenas na duração ou no tempo gasto por um trem em um determinado segmento.

A Figura 6.2 ilustra como são representadas a malha e diferentes circulações realizadas (com apenas o tempo gasto), referente ao planejamento inicial apresentado na Figura 5.8.

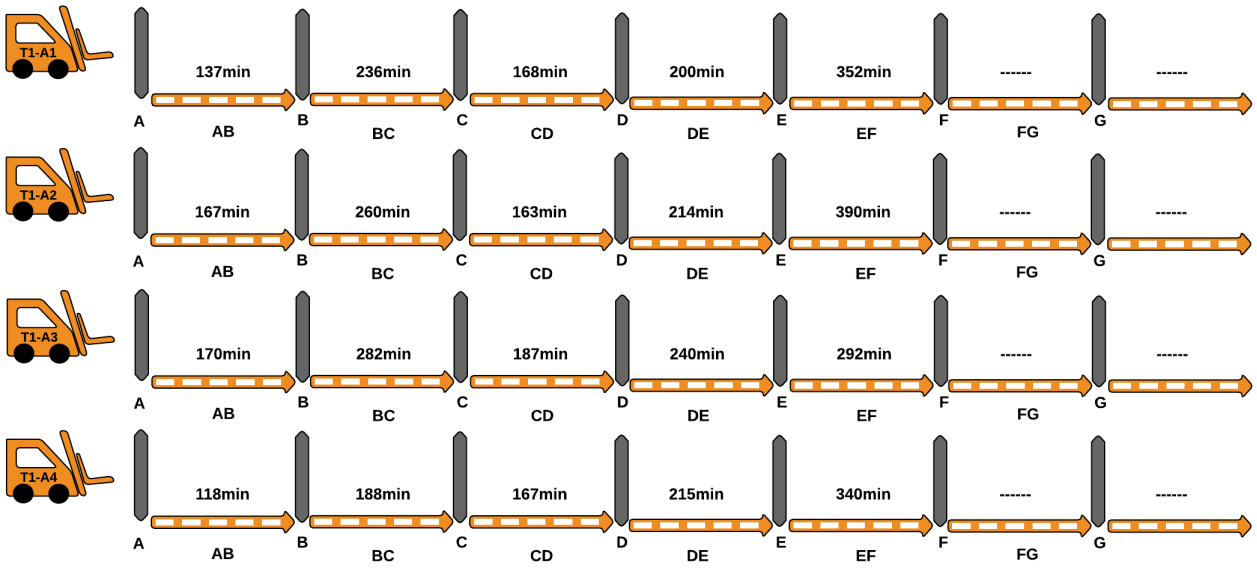

Figura 6.2: Representação de trajetórias percorridas por um trem com o tempo gasto em cada segmento.

Pode observar-se uma diferença importante nos tempos gastos nas mesmas vias para o mesmo trem em diferentes circulações. Além disso, pretende-se descobrir porque não houve aderência, com base na Tabela 5.3 em certos casos, entre o efetivamente realizado e o planejado. A pergunta é: por quê?!

No Anexo A.6 foram apresentados alguns fenômenos climáticos e não climáticos suscetíveis a perturbar as circulações, portanto, influenciar sua aderência com relação ao planejamento. Destarte, com o intuito de descobrir os fatores que causaram a não aderência e, em seguida, categorizá-los, verificou-se na base de dados os dados referentes aos fenômenos ambientais (climáticos e não climáticos) conforme a Tabela 6.1 que apresenta os fenômenos ambientais e seus respectivos domínios de valores, onde uma combinação desses valores formam uma instância que pode levar a uma das classes de aderência apresentada na Tabela 6.2.

Temp.	Chuva	Vento	Tempo	CM	FMM	Àgua	P.dia
muito quente	forte	violento	ensolarado	boa	sim	sim	manhã
quente	moderada	forte	nublado	ruim	não	não	tarde
média	fraca	moderado	chuvoso	—	—	—	noite
frio	não	fraca	neve	—	—	—	—

Tabela 6.1: Fenômenos ambientais e seus respectivos domínios de valores

Aderência
Normal
a**T**rasada
a**D**iantada

Tabela 6.2: Classes de aderência

Onde **Temp**, **CM** e **FMM** significam temperatura, condição da malha e folhas mortas na malha respectivamente. O atributo **Aderência** indica o grau de aderência de uma circulação com relação ao planejamento. Este pode ser **Normal** e representado pela letra **N**, a**T**rasado e representado pela letra **T**, ou a**D**iantado e representado pela letra **D**.

Tendo em vista o exposto, e para melhor explorar essas trajetórias, levando-se em consideração a natureza *heterogênea* e complexa das informações (clássica e simbólica) caracterizando seus estados-circulações, faz-se necessário reduzir essas informações, assim como estruturá-las. Um particionamento desses estados-circulações pela abordagem de classificação justifica-se como uma solução apropriada para esta fase preparatória das trajetórias realizadas. Essa abordagem permite construir uma partição fina do conjunto de circulações em classes homogêneas (classes de aderências) e disjuntas. As circulações serão, portanto, associadas a uma classe construída com base nas variáveis *ferroviárias* previamente mencionadas.

Diversas técnicas podem ser utilizadas para a construção de classificadores, tais como métodos Bayesianos, redes neurais, árvores de decisão, etc. Neste trabalho, optou-se por usar a árvore de decisão que possui certas vantagens, tais como a produção de procedimentos de classificações compreensíveis; resultado mais facilmente interpretável e, portanto, explorável; saída de resultados sob forma de regras lógicas de classificação; mais precisamente, o fato desta técnica suportar as variáveis tanto qualitativas como quantitativas (discretas ou contínuas) que usou-se neste trabalho. O objetivo é gerar uma sequência hierárquica de testes, tão curta quanto

possível, que divide sucessivamente todos os dados de treinamento em subconjuntos disjuntos, tais como os casos de subgrupos pertencentes à mesma classe são detectados rapidamente.

Sendo assim, obteve-se a árvore de decisão apresentada na Figura 6.3 com a acurácia média de 91% após 100 treinamentos. Essa árvore indica as condições de não aderência de uma trajetória realizada com base nos atributos preditivos. A estrutura possui as seguintes características:

- cada nó interno é um teste em um atributo preditivo;
- uma ramificação partindo de um nó interno representa um resultado para o teste (por exemplo, Temperatura = “quente”);
- uma folha da árvore representa um rótulo de classe (por exemplo, Aderência = “Normal” ou Aderência = “Atrasado”);

Observa-se que esta árvore de decisão pode ser utilizada com duas finalidades: **previsão** (exemplo: descobrir se circulação será aderente ou não em função das informações climáticas e não climáticas) e **descrição** (fornecer informações interessantes a respeito das relações entre os atributos preditivos e o atributo classe numa base de dados).

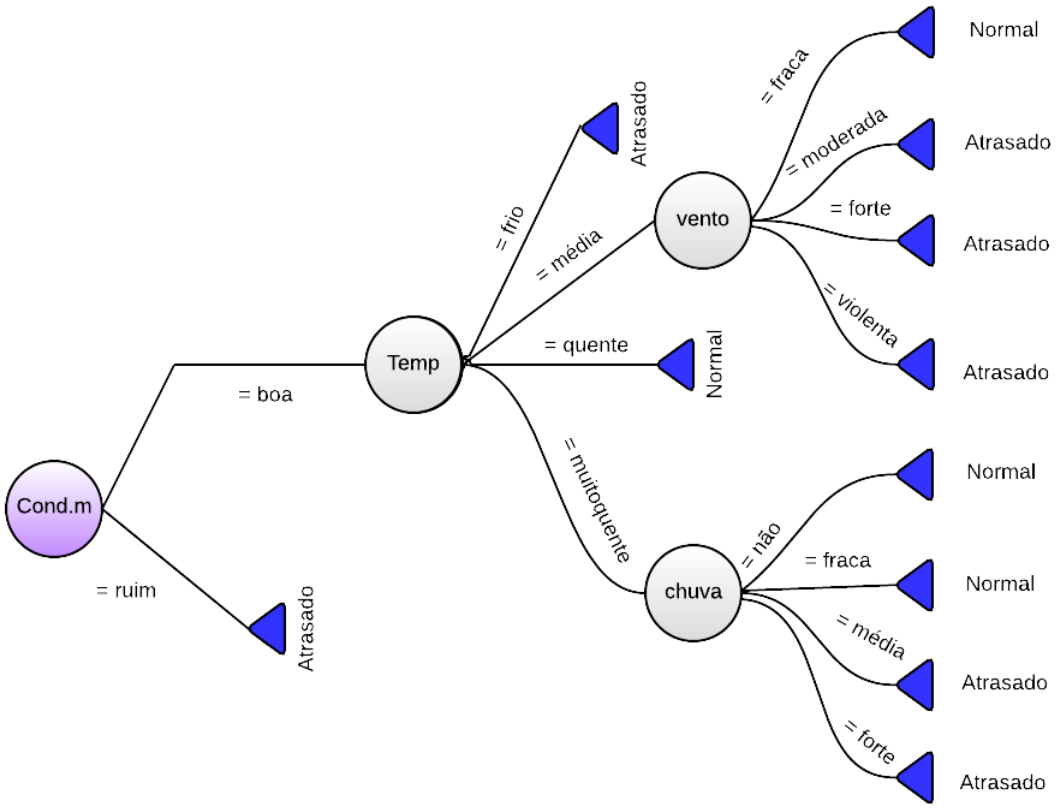

Figura 6.3: Árvore de decisão de aderência com poda

Uma árvore de decisão é formada por um conjunto de regras de classificação. Cada caminho da raiz até uma folha representa uma destas regras. A árvore de decisão deve ser definida de forma que, para cada observação da base de dados, haja um e apenas um caminho da raiz até a folha. Por exemplo, as quatro regras de classificação a seguir, compõem a árvore de decisão da Figura 6.3.

- (Condição da Malha = "ruim")→ (Aderência = "Atrasado")
- (Condição da Malha = "boa") & (Temperatura = "média") & (Chuva = "forte") → (Aderência = "Atrasado")
- (Condição da Malha = "boa") & (Temperatura = "média") & (Chuva = "fraca") → (Aderência = "Normal")
- (Condição da Malha = "boa") & (Temperatura = "muito quente") & (Vento = "fraca") → (Aderência = "Normal")

Uma regra de classificação é uma expressão da forma X $\rightarrow$ Y, onde X é denominado antecedente e Y é denominado consequente. O antecedente deve ser formado por um ou mais atributos preditivos, enquanto o atributo classe aparece no lado do consequente. Uma regra do tipo X $\rightarrow$ Y indica que a classe Y pode ser determinada pelos atributos preditivos indicados no antecedente. Medidas como a probabilidade condicional podem ser utilizadas para avaliar a qualidade de uma regra de classificação. Observa-se que na Figura 6.3, o atributo "Condicao-Malha" encontra-se na raiz da árvore, pois foi considerado pelo algoritmo classificador como o atributo mais importante para determinar se uma circulação é aderente ou não. Geralmente são utilizadas medidas baseadas na entropia para tratar este problema.

Dentre os diversos algoritmos utilizados para construção de árvores de decisão, decidiu-se pelo emprego do C4.5. Para verificar a coerência da árvore de decisão gerada, foi necessário calcular a entropia (com base na Equação 2.38) de cada um dos atributos considerados neste trabalho, conforme segue:

Atributo	Entropia (S, Atributo(valor))
Chuva	Entropia ($S_{forte} = 0$)
	Entropia ($S_{media} = 0$)
	Entropia ($S_{fraca} = 0.873$)
	Entropia ($S_{nao} = 0.894$)
Temperatura	Entropia ($S_{muitoquente} = 0.917$))
	Entropia ($S_{quente} = 0.0$)
	Entropia ($S_{media} = 0.870$)
	Entropia ($S_{frio} = 0.345$)
Vento	Entropia ($S_{violento} = 0.0$)
	Entropia ($S_{forte} = 0.37$)
	Entropia ($S_{moderado} = 1.00$)
	Entropia ($S_{fraca} = 0.79$)
Tempo	Entropia ($S_{ensolarado} = 0.665$)
	Entropia ($S_{nublado} = 0.927$)
	Entropia ($S_{chuvoso} = 0.818$)
Condição da malha	Entropia ($S_{boa} = 0.933$)
	Entropia ($S_{ruim} = 0.00$)
Folha morta na malha	Entropia ($S_{sim} = 0.00$)
	Entropia ($S_{nao} = 0.980$)
Água na malha	Entropia ($S_{sim} = 0.524$)
	Entropia ($S_{nao} = 0.904$)
Período do dia	Entropia ($S_{manha} = 0.997$)
	Entropia ($S_{tarde} = 0.997$)
	Entropia ($S_{noite} = 0.917$)

Tabela 6.3: Cálculo de entropia com relação a cada atributo considerados neste trabalho.

Calculando o Ganho de informação nos dados, tem-se:

Ganho (S, Atributo)	Valor
Ganho(S, Temperatura)	$0.999 - 0.66 = 0.339$
Ganho(S, Chuva)	$0.999 - 0.661 = 0.338$
Ganho(S, Condição da malha)	$0.999 - 0.783 = 0.261$
Ganho(S, Vento)	$0.999 - 0.766 = 0.233$
Ganho(S, Água na malha)	$0.999 - 0.807 = 0.192$
Ganho(S, Tempo)	$0.999 - 0.814 = 0.185$
Ganho(S, Folha morta na malha)	$0.999 - 0.868 = 0.131$
Ganho(S, período do dia)	$0.999 - 0.981 = 0.018$

Tabela 6.4: Análise dos ganhos de informação conseguido classificando-se os dados da base de dados reais do cenário ferroviária de todos os atributos

A entropia e o ganho de informação permitem a seleção inteligente de nós para formar a hierarquia de uma árvore de decisão. Além disso, o ganho de informação mede a efetividade de um atributo em classificar um conjunto de treinamento.

Após cálculo de entropia e ganho de informação, observou-se que a árvore de decisão obtida inicialmente poderia ser melhorada conforme ilustrada na Figura 6.4.

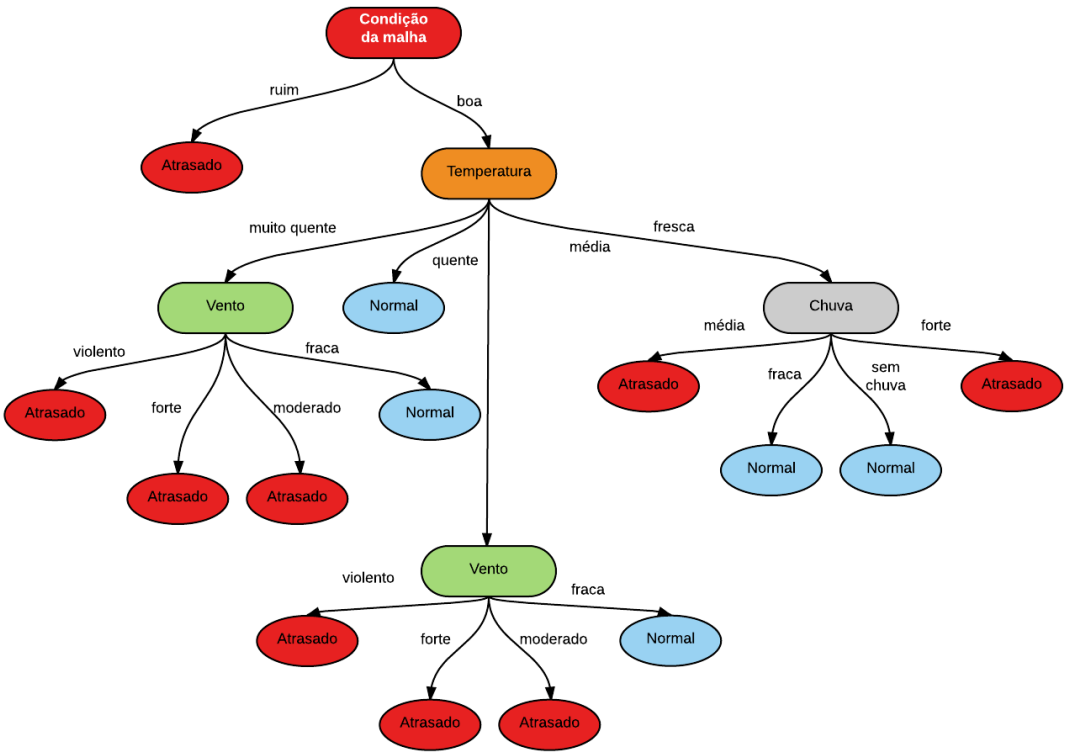

Figura 6.4: Árvore de decisão melhorada

Para facilitar as tarefas de interpretação, decisão e validação dos resultados, usam-se as classes de aderência de circulações (GHA: Grupo Homogênea de Aderência) já existentes nas informações fornecidas e simuladas do sistema ferroviário, sendo que tais informações foram obtidas a partir da árvore de decisão sobre as variáveis relacionadas às circulações de trens. As figuras 6.5 e 6.6 ilustram a representação de trajetória realizada por um trem em um dia e representação de trajetórias realizadas por um trem em um diferentes dias, porém no mesmo percurso, respectivamente, onde o cenário agora considera o GHA.

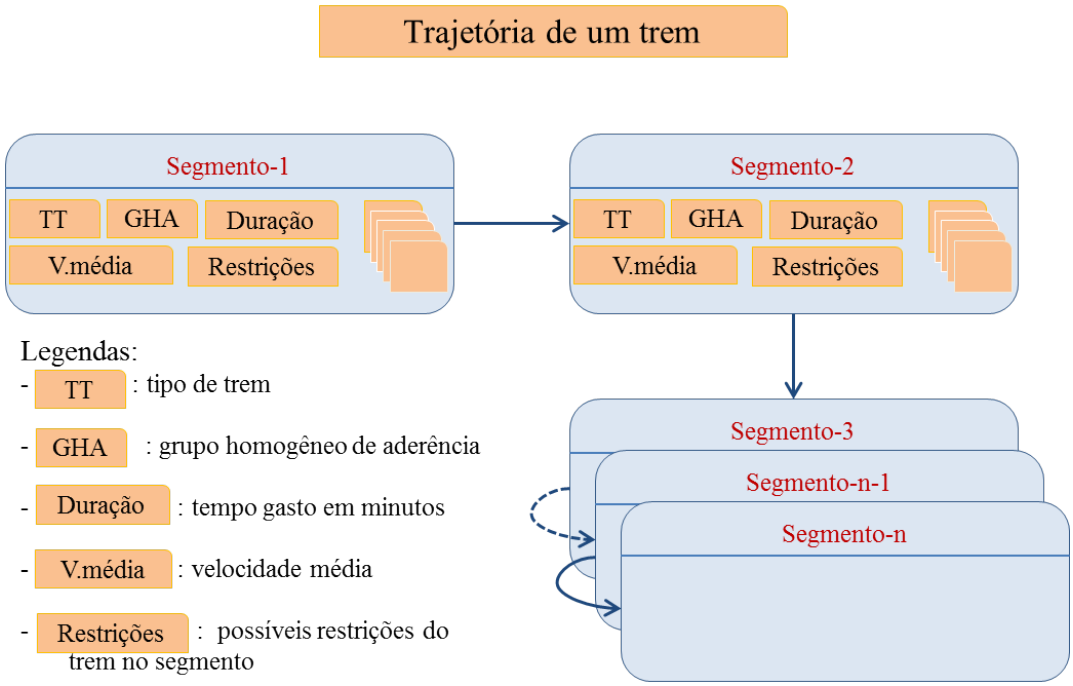

Figura 6.5: Representação de trajetória realizada por um trem em um dia.

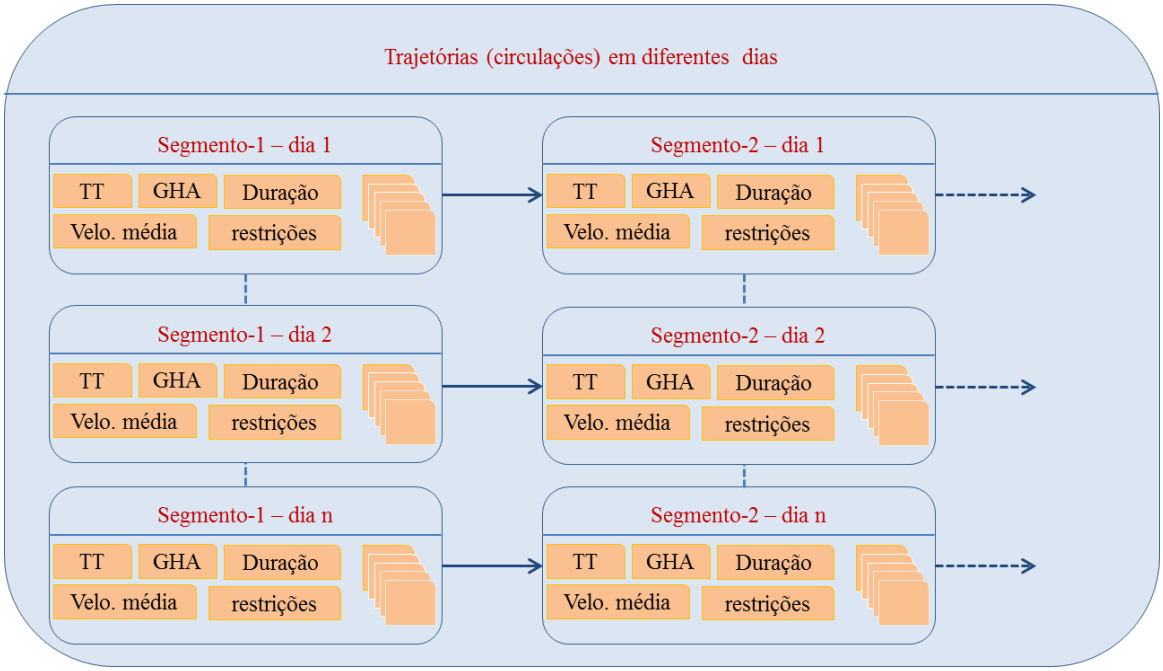

Figura 6.6: Representação de trajetórias realizadas por um trem em um diferentes dias, porém no mesmo percurso.

Sob outra perspectiva, para melhor explicar as dimensões relativas aos estados-circulações, faz-se necessário cruzar a variável *grupo de aderência* **(GHA)** com a variável *velocidade média* **(VM)** na descrição dos estados das sequências realizadas. Na prática, o GHA representa mais o aspecto operacional, enquanto a VM fornece a melhor ideia sobre o estado do trem. Na tabela 6.5, apresenta-se o exemplo de algumas trajetórias de circulações.

Trajetória (S_i)	Estado 1	Estado 2	Estado 3	...	Estado n
S_1	GHA-VM-1	GHA-VM-2	GHA-VM-3	...	GHA-VM-N
S_2	GHA-VM-1	GHA-VM-2	GHA-VM-3	...	GHA-VM-N
S_3	GHA-VM-1	GHA-VM-2	GHA-VM-3	...	GHA-VM-N
...	...	...	...	...	...
S_n	GHA-VM-1	GHA-VM-2	GHA-VM-3	...	GHA-VM-N

Tabela 6.5: Trajetória de circulações com informações cruzadas

Onde $Si = \{S_1, S_2, \ldots, S_n\}$ representa a diferentes de circulações realizadas pelo mesmo trem, anteriormente representado como $T1 - A1, T1 - A2, ..., T1 - A_n$, e os estados $e_{i,j}$ refletem a observação o comportamento do trem em determinado instante, dados pelo GHA em determinado segmento observado.

Com o objetivo de preservar a particularidade de cada uma destas duas informações (GHA e

VM) presente nas trajetórias, propôs-se considerá-las separadamente no cálculo de similaridade entre trajetórias, consoante a tabela a 6.6.

Trajetória (S_i)	$S_{i,j}$	Estado 1	Estado 2	Estado 3	...	Estado n
S_1	Série GHA:($S_{1,1}$)	GHA1	GHA2	GHA2	...	GHA3
	Série VM:($S_{1,2}$)	vm_1	vm_2	vm_3	...	vm_n
S_2	Série GHA:($S_{2,1}$)	GHA2	GHA1	GHA1	...	GHA1
	Série VM:($S_{2,2}$)	vm_1	vm_2	vm_3	...	vm_4
...	...	...	...	...	...	...
S_n	Série GHA:($S_{n,1}$)	GHA3	GHA3	GHA2	...	GHA1
	Série VM:($S_{n,2}$)	vm_1	vm_2	vm_3	...	vm_4

Tabela 6.6: Descrição dos estados das trajetórias de circulações de trens.

Onde :

GHA: Grupo Homogêneo de aderência pode assumir os valores (N: *normal*; T: *atrasado*; D: *adiantado*), do tipo de dados simbólicos.

VM: velocidade média, do tipo de dados clássicos.

A partir da Tabela 6.6, a medida de similaridade entre duas sequências como S_i e S_j proposta neste caso é dada pela agregação de distâncias de edição obtidas em cada conjunto de atributos (GHA e VM). Para se calcular a VM pode-se proceder por distancia euclidiana, porém, GHA exige uma nova formulação para cálculo da distância, devido à sua natureza simbólica.

Para fazer isso, a distância Euclidiana é usada do seguinte modo:

$$d(S_i, S_j) = \sqrt{\sum_{h=1}^{2} d_E(S_{i,h}, S_{j,h})^2} \tag{6.1}$$

Onde $d_E(S_{i,h}, S_{j,h})$ ($h \in \{1,2\}$) é a distância de edição entre duas séries de atributos $S_{i,h}$ e $S_{j,h}$ ($h = 1$ para GHA e $h = 2$ para VM) correspondentes respectivamente às trajetórias S_i e S_j.

6.2.2 Cálculo de dissimilaridades entre sequências de GHA

Tendo em vista a natureza qualitativa, obtida, dos estados $e_{i,j}$ em grupos $g_{i,j}$ das sequências S_i a analisar, escolheu-se usar a *distância de edição* apresentada na Seção 3.2.2, reproduzida em 6.2, para avaliar a similaridade entre as sequências temporais. Na prática, o problema da

avaliação da distância de edição entre duas sequências temporais (consideradas como duas cadeias de caracteres) é uma generalização do problema de avaliação do comprimento da maior subsequência comum (LCS) entre estas duas sequências. Todavia, esta distância apresenta certas limitações para ser aplicada no contexto estudado neste trabalho.

$$d_E(S_i, S_j) = |S_i| + \left|S_j\right| - 2 * LCS(S_i, S_j) \tag{6.2}$$

Limitações da distância de edição para o caso estudado

A distância de edição é dada pela Equação 6.2. Esta formulação permite a comparação entre sequencias de tamanhos distintos, entretanto, proporciona resultados não conclusivos nesta situação, permitindo variação na interpretabilidade dos resultados para o caso estudado nesta dissertação, o que torna sua aplicação direta não apropriada à problemática de classificação. Como ilustração desta fragilidade, suponha-se as três sequências: $U = (x, y, x, y)$; $V = (l, m, l, m)$ e $W = (x, y, o, p, q, r, s, t)$. Ao utilizar a distância de edição conforme a Equação 6.2, entre U e V $(4 + 4 - 2 * 0 = 8)$ é encontrada a distância de 8 unidades e a mesma distância é encontrada entre U e W $(4 + 8 - 2 * 2 = 8)$. No entanto, o que se esperaria é que a distância entre U e W (com dois símbolos em comum : x e y) deveria ser menor que a entre U e V (que apresentam nenhum símbolo em comum).

A limitação observada para a aplicação entre sequencias de tamanhos variáveis, foi enfrentada com a definição de uma nova distância de edição nomeada Distância de Edição Flexível (DeFlex), conforme apresentado na Equação 6.3. A *distância de edição* ***DeFlex*** entre duas sequências S_i e S_j, *independe do comprimento*, é assim definida por 1 menos o produto da distância de edição original (onde a soma dos comprimentos das duas sequências é representada por uma variável X) e o inverso da variável X (ver equação 6.3). Os valores desta medida são, portanto,limitado entre $[0, 1]$. Uma vez que neste trabalho não se trata apenas de calcular a distância ou similaridade para sequências de comprimentos diferentes, mas também, e especialmente, a similaridade entre duas sequências de *comprimento igual* (o mesmo percurso realizado por um trem em tempos diferentes) onde considera-se o comprimento mínimo entre as duas sequências, a Equação (6.3) foi proposta para tratar ambos os casos, para isso a definição da variável X varia de acordo com as restrições, a saber, caso as sequências tenham tamanhos iguais ou tamanhos diferentes.

$$d_{Flex}(S_i, S_j) = 1 - \left[\left\{X - 2 * LCS(S_i, S_j)\right\}\right] * X^{-1} \tag{6.3}$$

$$d_{Flex}(S_i,S_j) = 1 - \frac{\{X - 2 * LCS(S_i,S_j)\}}{X}$$

Onde:

$$\begin{cases} X = |S_i| + |S_j|, & considerando\,|S_i| \neq |S_j| \\ X = 2 * min(|S_i,S_j|), & considerando\,|S_i| = |S_j| \end{cases}$$

Onde $min(|S_i,S_j|)$ é o menor comprimento entre as sequências S_i e S_j, sabendo que $|S_i,S_j| \leq |S_i| + |S_j|$. Isto é útil para calcular a distância entre duas sequências levando-se em consideração as trajetórias equidistantes.

Como exemplo ilustrativo: sejam as sequências $U = (x,y,x,y)$ e $W = (x,y,o,p,q,r,s,t)$.

1. a distância de edição, considerando a fórmula original (equação 3.2), entre as duas sequências resulta $(4+8-2*2=8)$.
2. pela equação 6.3, onde $X = |S_i| + |S_j|$) (comprimentos diferentes), tem-se: $1 - \frac{((4+8-2*2)}{12} = 1 - 0.66 = 0.33)$.
3. pela equação 6.3, onde $X = 2 * min(|S_i|, |S_j|)$ (comprimentos iguais), tem-se: $1 - \frac{((4+4-2*2)}{8} = 1 - 0.5 = 0.5)$, isto restringindo ao mesmo comprimento da sequencia U, a partir do inicio. Isto é, $U = (x,y,x,y)$ e $W = (x,y,o,p)$.

Os valores encontrados nos itens 2 e 3 respondem efetivamente ao que se esperava.

Assim sendo, aplicou-se, doravante, a distância DeFlex para calcular a aderência entre dois Grupos Heterogêneos de Aderência (GHA), uma vez que estes são simbólicos, ou seja, forma uma sequência de caracteres.

6.2.3 Classificação de sequências

As sequências a serem analisadas são doravante associadas a uma matriz de dissimilaridade simétrica $D = \{d(S_i,S_j) \mid S_i,S_j \in S\}$ de tamanho *nxn*. Trata-se nesta parte da definição de um método híbrido de classificação automática do conjunto de sequências $S = \{S_1,S_2,\ldots,S_n\}$. O intuito é de não apenas definir e construir uma tipologia de sequências em classes homogêneas e bem separadas, mas também resumir a informação que elas contêm nos modelos para interpretar e aplicá-las mais tarde para fins de classificação e previsão. O método é composto pelos processos:

Etapa a: ***Classificação automática por árvore de classificação***

Esta abordagem visa estruturar as sequências contidas em $S = \{S_1, S_2, \ldots, S_n\}$ em função de suas similaridades, sob forma de um conjunto de classes *homogêneas* e *disjuntas*, cada uma caracterizada por um conjunto de *sequências dominantes*. Estas são um reflexo das propriedades de sua classe, mas também garantem uma clara separação das últimas com relação a outras classes de partição.

Etapa b: ***Modelo de mistura e árvore de decisão***

Visto as performances significativas consideráveis das cadeias de Markov no que concerne à elaboração de modelos probabilísticos de geração de dados que resumem as relações entre os estados das sequências tratadas, e tendo em vista as dificuldades de classificação automática por modelo de mistura no que tange às probabilidades iniciais e o número de classes, propõe-se utilizar os resultados de classificação obtido na etapa (a), como alternativa ao problema de inicialização dos parâmetros dos modelos a estimar. Além disso, usar-se as probabilidades com base na modelagem de um processo estocástico markoviano observável que representa o comportamento do tempo do dia. A inicialização se faz então a partir da classificação obtida por árvore de classificação para distribuir os indivíduos nas classes. Assim sendo, as probabilidades iniciais $P(c_i = c \mid S_i, \Phi)$ são iguais a 1 para a classe de pertencimento dada pela árvore de classificação, e nulas para todas as outras classes da partição.

Etapa c: ***Previsão da progressão de sequência***

Uma vez que o modelo de mistura tenha aprendido sobre o conjunto das sequências temporais $S = \{S_1, S_2, \ldots, S_n\}$, pode servir-se disto para fazer a previsão em *tempo real* da progressão de uma sequência S_a (nova ou existente) após ter observado seu histórico $\left\{g_{a,1}, g_{a,2}, \ldots, g_{a,T_a}\right\}$ ($g_{a,1}$ é o grupo associado ao estado $e_{a,j}$). Os passos do modelo de mistura (markoviano observável) são:

1. Atribuir esta sequência para classe c_a mais suscetível de reproduzi-la (a classe à qual S_a tem a maior probabilidade de pertencer $P(S_a \mid c_a = c, \Phi_a)$): propriedade em linha de classificação.

$$c_a = argmax_{1 \leq c \leq k} \{P(S_a \mid \Phi_c) = P(S_a \mid c_a = c, \Phi_c)\} \tag{6.4}$$

2. Utilizar A_{c_a}, a matriz *nxn* (*n* é o número de estados possíveis) de transição, associada à cadeia de Markov Φ_{c_a} da classe c_a, afim de prever a progressão da sequência

temporal S_a.

$$g_{a,T_{a+1}} = argmax_{1 \leq z \leq m} \{a_{c_a}(e_{a,T_a}, Z)\} \tag{6.5}$$

6.3 Conclusão

O presente capítulo apresentou a fundamentação conceitual empregada nos módulos de Aderência e Previsão para análise e previsão de dados sequenciais heterogêneos, com base no acoplamento entre uma abordagem de classificação por árvore de classificação, os Modelos de Markov e cálculo de distância de similaridade. A árvore de classificação permite que se classifique dados em conjuntos de classes homogêneas e disjuntas, por exemplo, uma trajetória percorrida em um instante t pode pertencer em apenas uma classe de aderência, isto é, ou à classe ”Normal”, ”Atrasado”ou ”Adiantado”. Enquanto o cálculo de similaridade entre duas circulações ou sequências é feito com base na fórmula de distância DeFlex proposta, que resulte em quão aderente uma circulação realizada é com relação ao planejado. Além do mais, apresentou-se os modelos de mistura de Markov para a elaboração de modelos probabilísticos para se fazer a previsão do comportamento futuro de veículos em tempo real após ter observado seu histórico.

O capítulo a seguir apresenta um estudo de caso onde aplica-se detalhadamente os conceitos aqui expostos.

Capítulo 7

EXPERIMENTOS E VALIDAÇÃO

Neste capítulo, apresentam-se a aplicação completa da abordagem proposta nesta dissertação, os procedimentos de coleta de dados, e o protótipo de uma ferramenta de apoio a decisão baseada nesta abordagem. Tal abordagem foi aplicada ao planejamento de rotas para trens, a análise de trajetórias, assim como as previsões de planejamentos e comportamentos futuros do trem. Além de avaliar a relevância da abordagem proposta, esta aplicação ferroviária permite desenvolver novas perspectivas para o apoio a decisão para o planejamento de rotas. Portanto, na última etapa, apresenta-se o simulador, construído como um protótipo para realizar uma prova de conceito, para o apoio a decisão de planejamentos de trens desenvolvido para caracterizar, a nível técnico, as diferentes contribuições teóricas deste trabalho. Conclui-se o capítulo expondo as perspectivas que já foram consideradas para a metodologia.

7.1 Introdução

Neste capítulo, aplica-se a metodologia de análise de dados sequenciais proposta. Para isso, partiu-se de um conjunto de dados ferroviários relativos à Empresa ABC, contendo um conjunto de trajetórias (circulações) de trens em diversas regiões brasileiras. Uma trajetória $S_i = \left\{e_{i,1}, e_{i,2} \ldots, e_{i,T_i}\right\}$ é definida como um conjunto des T_i circulações efetuadas sucessivamente por um trem i, uma circulação $e_{i,j}$ sendo caracterizada pelo conjunto de informações gerais sobre o trem (prefixo, tipo) e as informações respeitantes a seus percursos (a classe de aderência de cada segmento (via) percorrido, diferentes informações climáticas, condição da malha, o tempo gasto em cada segmento, etc.)

Por outro lado, tem-se implementado uma plataforma de simulação chamada ”Análise

de planejamento de rotas de trens", dedicada à análise de planejamento de rotas e previsão de circulações com base no acoplamento entre a abordagem de classificação por *árvore de decisão* e *cadeias de Markov*. A abordagem por árvore de classificação fornece classes de circulações homogêneas, cada uma caracterizada por um conjunto de trajetórias padrões (*tipo de aderência*), ao passo que as Cadeias de Markov permitem interpretar e descobrir (caso faltem algumas informações) as classes por intermédio dos modelos probabilísticos, que formam, assim, uma estrutura automática de previsão de comportamento de trajetórias de trens. De Fato, para um trem com uma série de circulações, trata-se, inicialmente, de identificar as classes de trajetórias que se aproxima mais, levando-se em consideração as suas características e outras informações relevantes que influenciam uma circulação (condições climáticas, por exemplo). Em seguida, se necessário, pode-se prever como será o comportamento mais provável da próxima circulação, e estimar suas características principais (Tipo de aderência, tempo gasto, etc.). A cada propriedade é atribuída as probabilidades obtidas a partir do modelo de Markov estabelecido para a classe de trajetórias.

7.2 Coleta de dados

Para avaliar o desempenho da proposta deste trabalho, foram feitos um conjuntos de experimentos sobre as amostras de trajetórias (circulações) de trens coletadas na base de dados da empresa ABC.

Com o intuito de descobrir os padrões sequencias, o comportamento de trens em diferentes circulações do mesmo percurso, categorizar os acontecimentos (incidentes) com base em seus impactos, antes de mais nada, foi necessário definir uma coleção de dados em forma matricial conforme indicado na Figura 7.1, onde para cada trem se indicam suas características. Cada coluna representa uma variável particular (no caso, os parâmetros climáticos e não climáticos). Cada linha corresponde a um determinado membro do conjunto de dados em questão. O conjunto de dados pode incluir dados para um ou mais trens, correspondente ao número de linhas. A planilha de dados representa todos os cenários da amostra de trajetórias de trens obtidas na base de dados.

As duas amostras de 321 registros (AmV) e 1235 registros (AmGHA) respectivos foram construídas por um procedimento de amostragem estratificada realizada na população total (de todo o banco de dados). Cada uma das amostras foi desenvolvidas de forma diferente: Para a amostra AmV selecionou-se as trajetórias com base na **V.média** *(Velocidade média)* dos trens por segmentos (vias) e, a amostra AmGHA foi construída com base no **GHA** *(Grupo hete-*

rogêneo de Aderência). Além disso, as trajetórias em questão dizem respeito a distintos tipos de trens e suas diferentes circulações no mesmo percurso, mas em dias diferentes.

A construção da planilha de dados apresentada na Figura 7.1 baseia-se, inicialmente, na coleta de informações disponíveis na base de dados fornecida. Nada obstante, a base armazenava apenas os dados relativos ao trem (tipo, velocidade máxima, restrições), à circulação (duração em cada segmento, velocidade média por segmento), e à condição da malha. Mas, não armazenava-se os dados relacionados às informações climáticas. Neste fito, foi necessário buscar essas informações e acrescentá-las manualmente. Este processo empírico foi realizado buscando as informações meteorológicas correspondentes a cada dia, hora e local em que o trem circulou, nos sites mais confiáveis como do INPE (Instituto Nacional de Pesquisas Espaciais) [1] e da WMO (Organização mundial de Meteorologia) [2]. Tendo essas informações, tornou-se possível elaborar uma planilha mais concisa a fim de proceder a operações de treinamento para se descobrir os padrões e o comportamento de trens em valores diferentes das variáveis consideradas.

Na prática, encontrou-se certos fatores que desempenham um papel fundamental na determinação do comportamento do trem durante a circulação, levando a grau de aderência (Normal, Atrasado, Adiantado) ao chegar ao final da trajetória. Estes fatores são: temperatura ambiente, chuva, vento, clima, condição da malha, presença de folhas mortas na malha, água na malha, e, mas não necessariamente, o período do dia, sendo que cada um dos fatores pode ter valores variados. Por exemplo, a temperatura pode ter valores (simbólicos) como: muito quente, quente, médio, com intervalos numéricos definidos para cada valor. A combinação desses atributos, ou melhor, de seus valores, por linha, forma o que chama-se, doravante, de instância. Uma instância, por sua vez, determina um possível grau de aderência ao planejamento de uma determinada trajetória. Para se ter essa certeza, procedeu-se ao processo de mineração de dados, o treinamento da planilha construída, a fim de descobrir conhecimentos por trás dessa gama volumosa de dados.

Nesta finalidade, foi indispensável o uso de um software para mineração de dados, usou-se o **Weka**[3], que possui algoritmos eficientes para a extração de classificadores em bases de dados, os quais identificam a classe de pertencimento de uma determinada observação.

A mineração de modelos de classificação na planilha de dados construída consistiu em um processo composto por duas fases: aprendizado e teste. Na fase de aprendizado, um algoritmo

[1] www.inpe.br

[2] www.wmo.int

[3] O sistema Weka é um software livre (de código aberto) para mineração de dados, desenvolvido em Java, dentro das especificações da GNU (General Public License. Mais informações: www.cs.waikato.ac.nz/ml/weka/)

classificador, neste caso a árvore de decisão (algoritmo C4.5), foi aplicada sobre o conjunto de dados de treinamento. Como resultado, obteve-se a construção do modelo classificador propriamente dito. O conjunto de treinamento correspondeu a um subconjunto de observações selecionadas de maneira aleatória a partir da base de dados da empresa. Cada observação do conjunto de treinamento é caracterizada por dois tipos de atributos: o atributo classe, que indica a classe a qual a observação pertence, neste caso a **aderência** podendo ser (Normal, Atrasado ou Adiantado); e os atributos preditivos (temperatura ambiente, chuva, vento, tempo, condição da malha, presença de folhas mortas na malha, presença água na malha, e o período do dia), cujos valores serão analisados para que seja descoberto o modo como eles se relacionam com o atributo classe.

A aplicação deste processo pode ser observada na Figura 7.1, onde considera-se o conjunto de dados de treinamento da planilha construída neste trabalho.

#T-via	Duração-min	Temperatura	Chuva	Vento	Tempo	Condição Malha	FMM	Água atmosf.	Periodo dia	Grau de aderência
T1-AB1	137	médio	não	moderado	Ensolerado	bom	não	não	manhã	Normal
T1-BC1	235.5	quente	não	moderado	Ensolerado	bom	não	não	tarde	Normal
T1-CD1	168	muito quente	não	Forte	Ensolerado	bom	não	não	tarde	Atrasado
T1-DE1	200	médio	não	moderado	nublado	bom	não	não	tarde	Normal
T1-EF1	352	médio	não	Fraca	nublado	bom	não	não	noite	Normal
T1-FG1						#				
T1-AB2	167	médio	média	moderado	chuvoso	bom	não	não	manhã	Atrasado
T1-BC2	260	médio	fraca	Fraca	chuvoso	bom	não	não	tarde	Normal
T1-CD2	163	médio	não	Fraca	nublado	bom	não	não	tarde	Normal
T1-DE2	214	médio	não	Fraca	nublado	bom	não	não	tarde	Normal
T1-EF2	390	médio	não	Fraca	nublado	bom	não	não	noite	Normal
T1-FG2						#				

Figura 7.1: Estrutura da planilha construída

Após o classificador ter sido construído, iniciou-se a etapa de teste, que visou avaliar a sua acurácia através do emprego de um conjunto de dados de teste. O conjunto de teste contém observações que também foram selecionadas aleatoriamente a partir da planilha de dados. No entanto, estas observações devem ser diferentes das que foram selecionadas para compor o conjunto de treinamento. A acurácia do classificador representa a porcentagem de observações do conjunto de teste que são corretamente classificadas por ele. Caso a acurácia seja alta, o modelo de classificação é considerado eficiente e pode ser utilizado para classificar novos casos. Nos treinamentos realizados, obteve-se a acurácia de 91% após 100 treinamentos, portanto, o modelo de classificação gerado é considerado eficiente.

7.3 Operações sobre os dados

Tendo em vista a disponibilidade da planilha de informações, tais como itinerários planejados dos trens, dados referentes à evolução da trajetória realizada até um determinado instante *t*, ambos procedentes, indiretamente, do Repositório de Dados Internos, pode-se realizar os procedimentos relativos ao *cálculo de aderência*, isto é, verificar se o cenário planejado corresponde ao que está acontecendo ou não. Além do mais, pode-se fazer a previsão com base nas informações contidas na planilha, assim como os padrões descobertos, e as informações ambientais atuais. As subseções 7.3.1 e 7.3.2 apresentam o cálculo de aderência e previsão de planejamentos respectivamente.

7.3.1 Cálculo de aderência

O cálculo de aderência faz uso das fórmulas propostas na Subseção 6.2.2 desta dissertação, retornando o grau de similaridade(aderência) entre duas circulações. As decisões sobre a aderência ou não aderência dependem do intervalo de confiança determinado pelo setor operacional.

Conforme apresentado na Seção 6.2.1, as informações coletadas durante uma circulação de trens não consistem apenas na duração ou no tempo gasto em um determinado segmento. Existem outros fatores ou fenômenos climáticos e não climáticos capazes de perturbar as circulações, portanto, influenciar sua aderência com relação ao planejamento. Destarte, com o intuito de descobrir os fatores que causaram a não aderência e, em seguida, categorizá-los, verificou-se na base de dados os dados referentes aos fenômenos ambientais, onde cada atributo tem valores.

Agora, considerando o planejamento a seguir (Figura 7.2) com base nos dados coletados:

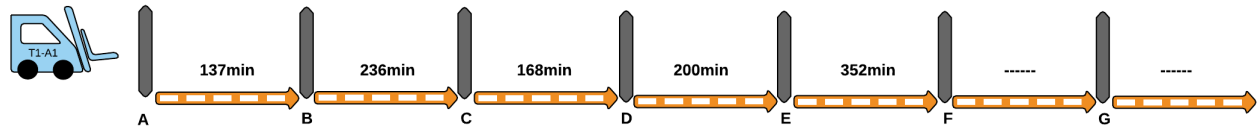

Figura 7.2: Planejamento da trajetória a ser percorrida pelo trem T1

Uma amostra de três circulações realizadas, T1-A2, T1-A3, T1-A4, foi coletada, e elaborou-se os cenários ilustrada na Figura 7.3.

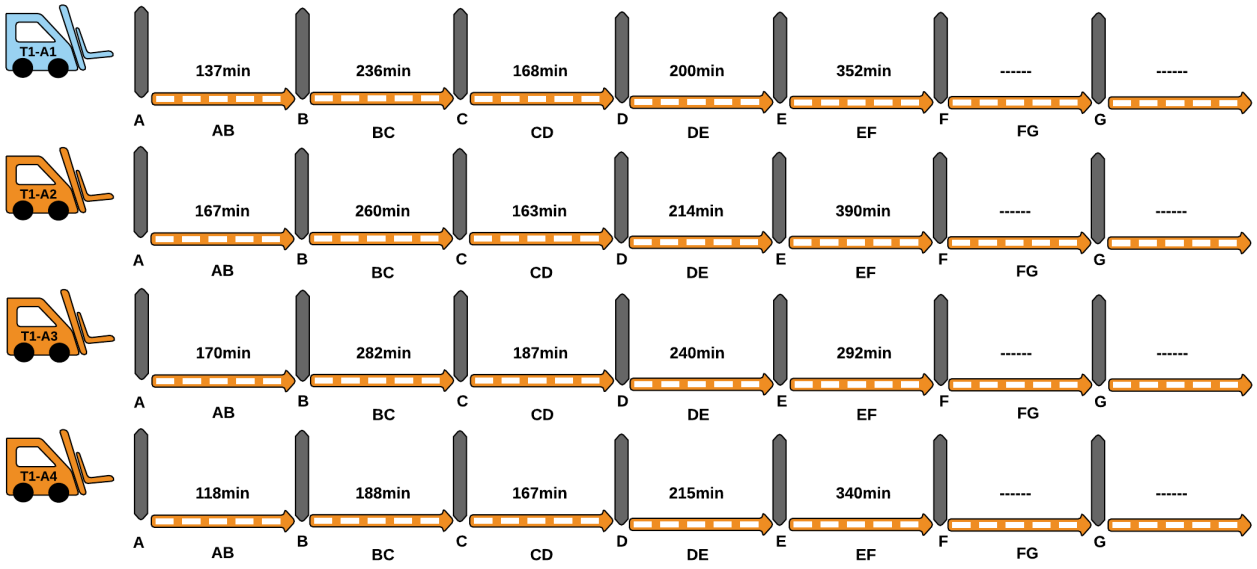

Figura 7.3: Planejamento da trajetória a ser percorrida pelo trem T1 e diferentes circulações realizadas

Pode-se constatar que a Figura 7.3 contém apenas informações relacionadas ao tempo gasto por segmento. Entretanto, existem outras informações (climáticas e não climáticas) que se associam a cada trajetória, as quais servem para identificar o conhecimento por trás de seus acontecimentos, tais como: seus impactos em uma circulação. A Figura 7.4 mostra como isto é realizado.

Estabeleceu-se intervalos de confiança, conforme apresentado na Tabela 7.1, para determinar se o trem está circulando dentro das velocidades aceitáveis ao percorrer uma via uma via v_i. Tais intervalos indicam se o trem está circulando normalmente, em atraso ou adiantado com relação ao tempo estimado no planejamento inicial, e considerando os fatores ambientais registrados.

Via	Tempo estimado (min)	Normal	Atrasado	Adiantado
AB	150	$120 < t \leq 165$	$t > 165$	$t < 120$
BC	240	$192 < t \leq 264$	$t > 264$	$t < 192$
CD	150	$120 < t \leq 165$	$t > 165$	$t < 120$
DE	214.2	$171.36 < t \leq 235.62$	$t > 235.62$	$t < 171.36$
EF	360	$288 < t \leq 396$	$t > 396$	$t < 288$

Tabela 7.1: Tabela de grau de liberdade

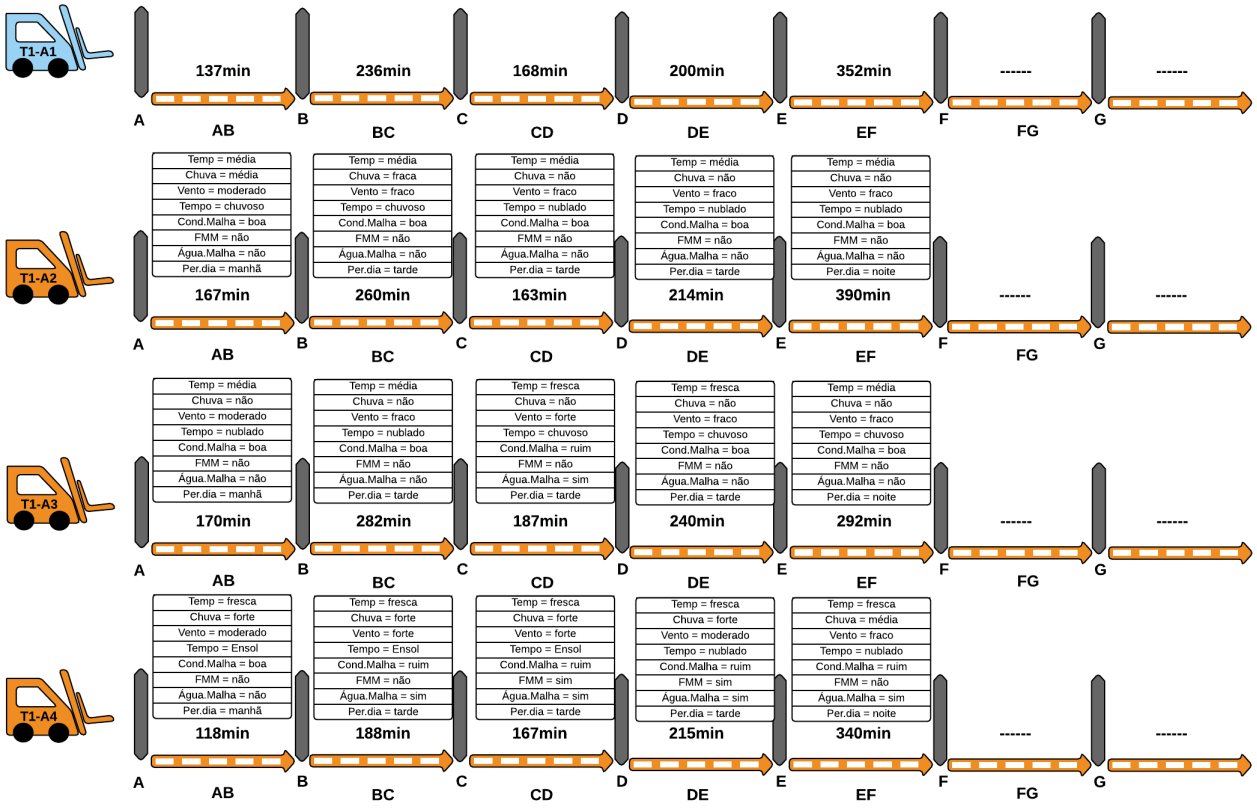

Figura 7.4: Planejamento da trajetória a ser percorrida pelo trem T1 e diferentes circulações realizadas com informações relativas aos fatores climáticos e não climáticos relevantes.

Conforme apresentado na seção 6.2, para facilitar as tarefas de interpretação, decisão e validação dos resultados, usou-se as classes de aderência de circulações (GHA: Grupo Homogênea de Aderência) correspondentes a cada trajetória percorrida. Têm-se três GHA: Normal **N**, Atrasado **T** e Adiantado **D**.

Sob outra perspectiva, para melhor explicar as dimensões relativas aos estados-circulações, cruzou-se a variável *grupo de aderência* **(GHA)** com a variável *velocidade média* **(VM)** na descrição dos estados das sequências realizadas. Na prática, o GHA representa mais o aspecto de evolução da circulação, enquanto a VM fornece a melhor ideia sobre o estado do trem. Na tabela 6.5, apresenta-se o exemplo de algumas trajetórias de circulações.

Isto posto, com base na Tabela 7.1 e considerando as informações ambientais coletadas, a Figura 7.4 veio a ser representada consoante a Figura 7.5.

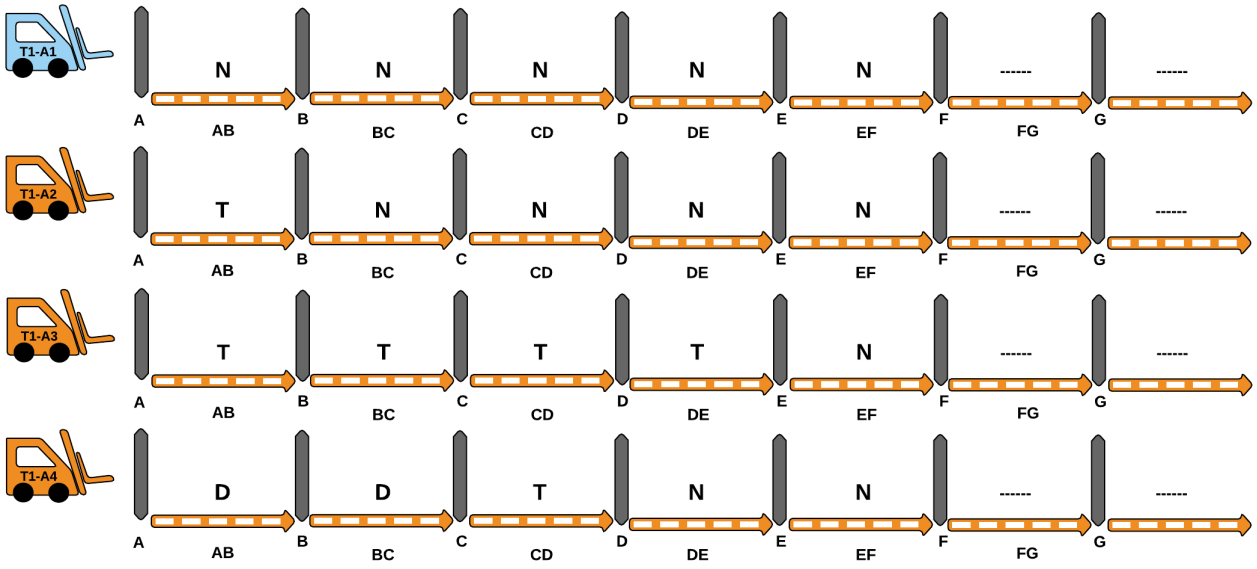

Figura 7.5: Planejamento e trajetórias percorridas pelo trem T1 e diferentes circulações realizadas com informações relativas ao grau de aderência

Dessarte, pode-se, a partir da Figura 7.5, formar cadeias ou sequências de caracteres agrupando os graus de aderências respeitantes a cada circulação. Neste caso, teremos quatro cadeias de caracteres conforme segue:

1. *Planejamento* **T1-A1**: (NNNNN), determina o que se espera.
2. *Primeira circulação* **T1-A2**: (TNNNN)
3. *Segunda circulação* **T1-A3**: (TTTTN)
4. *Terceira circulação* **T1-A4**: (DDTNN)

Para descobrir a aderência de cada circulação realizada em relação ao planejado, faz-se necessário verificar quão similares são. Para tal, tendo estas cadeias de caracteres, usou-se a distância proposta conforme:

(1)
$$d_{Flex}(S_i, S_j) = 1 - \frac{|S_i| + |S_j| - 2 * LCS(S_i, S_j)}{|S_i| + |S_j|}$$

(2)
$$d_{Flex}(S_i, S_j) = 1 - \frac{2 * min(|S_i|, |S_j|) - 2 * LCS(S_i, S_j)}{2 * min(|S_i|, |S_j|)}$$

Como as trajetórias são de igual comprimento, aplicou-se a fórmula (1), e obteve-se:

- O grau de aderência entre a primeira circulação realizada **T1-A2** em relação ao planejamento **T1-A1** é dada por:

$$d_{Flex}(T1-A2,T1-A1) = 1 - \frac{|T1-A2|+|T1-A1|-2*LCS(T1-A2,T1-A1)}{|T1-A2|+|T1-A1|}$$

$$d_{Flex}(T1-A2,T1-A1) = 1 - \frac{|5|+|5|-2*4}{|5|+|5|} = 0,8$$

  Isto é, T1-A2 é 0,8 ou 80% aderente a T1-A1.

- O grau de aderência entre a segunda circulação realizada **T1-A3** em relação ao planejamento **T1-A1** é dada por:

$$d_{Flex}(T1-A3,T1-A1) = 1 - \frac{|T1-A3|+|T1-A1|-2*LCS(T1-A3,T1-A1)}{|T1-A3|+|T1-A1|}$$

$$d_{Flex}(T1-A3,T1-A1) = 1 - \frac{|5|+|5|-2*1}{|5|+|5|} = 0,2$$

  Isto é, T1-A3 é 0,2 ou 20% aderente a T1-A1.

- O grau de aderência entre a terceira circulação realizada **T1-A4** em relação ao planejamento **T1-A1** é dada por:

$$d_{Flex}(T1-A4,T1-A1) = 1 - \frac{|T1-A4|+|T1-A1|-2*LCS(T1-A4,T1-A1)}{|T1-A4|+|T1-A1|}$$

$$d_{Flex}(T1-A4,T1-A1) = 1 - \frac{|5|+|5|-2*2}{|5|+|5|} = 0,4$$

  Isto é, T1-A3 é 0,4 ou 40% aderente a T1-A1.

Por outro lado, se fosse utilizada a fórmula de edição genérica, para cada cálculo de efetuado, ter-se-ia os seguintes valores:

- A distância entre **T1-A1** e **T1-A2** é dada por:

$$d_E(T1-A2,T1-A1) = |T1-A2|+|T1-A1|-2*LCS(T1-A2,T1-A1)$$

$$d_E(T1-A2,T1-A1) = |5|+|5|-2*LCS(4) = 2$$

- A distância entre **T1-A3** e **T1-A1** é dada por:

$$d_E(T1-A3, T1-A1) = |T1-A3| + |T1-A1| - 2*LCS(T1-A3, T1-A1)$$

$$d_E(T1-A3, T1-A1) = |5| + |5| - 2*LCS(1) = 8$$

- A distância entre **T1-A3** e **T1-A1** é dada por:

$$d_{E_m}(T1-A4, T1-A1) = |T1-A4| + |T1-A1| - 2*LCS(T1-A4, T1-A1)$$

$$d_E(T1-A4, T1-A1) = |5| + |5| - 2*LCS(2) = 6$$

- Nisto, suponha-se que se tenha mais uma trajetória, seja T1-A5 = (DDTTDD), cuja intersecção com T1-A0 (NNNNNN), planejamento inicial, é nula. Aplicando a distância de edição genérica tem-se:

$$d_E(T1-A5, T1-A0) = |T1-A5| + |T1-A1| - 2*LCS(T1-A5, T1-A0)$$

$$d_E(T1-A5, T1-A0) = |6| + |6| - 2*0 = 12$$

Percebe-se que tanto os valores anteriormente obtidos como o valor 12 neste último caso em que houve dissimilaridade total, não deixaram claro o seu significado, portanto, inconclusivos no contexto deste trabalho.

Mas, ainda neste último caso, ao aplicar a distância proposta, obtém-se:

$$d_{Flex}(T1-A5, T1-A0) = 1 - \frac{|T1-A5| + |T1-A0| - 2*LCS(T1-A5, T1-A0)}{|T1-A5| + |T1-A0|}$$

$$d_{Deflex}(T1-A5, T1-A0) = 1 - \frac{|6| + |6| - 2*0}{|6| + |6|} = 1 - 1 = 0$$

O que explica claramente que as duas sequências são disjunta, ou completamente dissimilares, uma vez que o intervalo de grau de aderência varia de: $[0,1]$, onde 0 refere-se à não-aderência total e 1 aderência total.

Além do cenário anterior, existe outro caso em que se queira verificar a aderência entre uma

circulação em andamento e seu planejamento inicial. A Figura 7.6 apresenta graficamente o cenário do planejamento inicial e a Figura 7.7 o monitoramento do trem em curso de circulação.

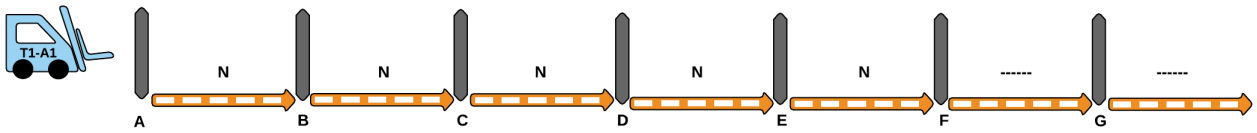

Figura 7.6: Planejamento inicial de trajetória a ser percorrida pelo trem T1

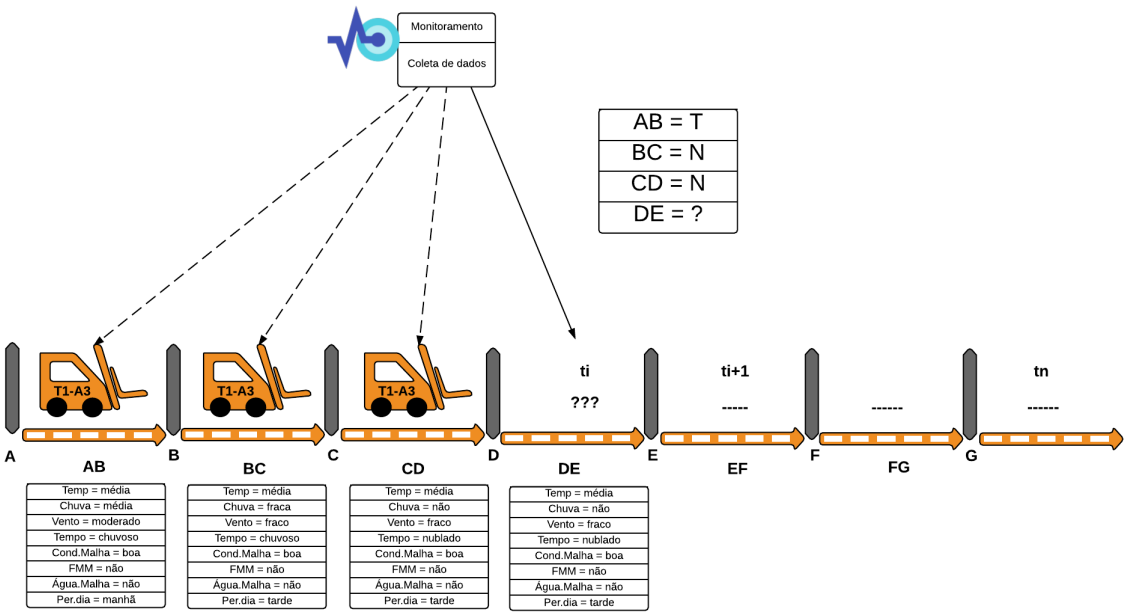

Figura 7.7: Informações referentes ao monitoramento do trem T1 na terceira circulação

Ao se realizar o cálculo de aderência com o trem em andamento, obtendo-se os dados via monitoramento em um momento qualquer da circulação, tem-se duas sequências de comprimentos diferentes. No exemplo da Figura 7.7, a sequência planejada é **NNNNN** e a sequência monitorada até então é **TNN**.

Neste ponto, ao calcular a aderência do planejado e realizado, com a variante (1), tem-se:

$$d_{Flex}(T1-A3, T1-A1) = 1 - \frac{|T1-A3| + |T1-A1| - 2 * LCS(T1-A3, T1-A1)}{|T1-A3| + |T1-A1|}$$

$$d_{Flex}(T1-A3, T1-A1) = 1 - \frac{|3| + |5| - 2 * 2}{|3| + |5|} = 1 - 0.75 = 0.25$$

Ou 25%. Todavia, não se pode decidir nada com base nesse resultado, uma vez que não se sabe como será o comportamento do resto da circulação. Portanto, o certo será verificar ou calcular a aderência de realizado com o planejado correspondente, isto é, de igual comprimento, o que leva a usar a segunda variante, a qual condicionada o cálculo de aderência e dissimilaridade entre duas sequências com comprimentos iguais :

$$d_{Flex}(S_i, S_j) = 1 - \frac{2 * min(S_i, S_j) - 2 * LCS(S_i, S_j)}{2 * min(S_i, S_j)}$$

Ao aplicá-la para o caso anterior, tem-se:

$$d_{Flex}(T1-A3, T1-A1) = 1 - \frac{2 * min(T1-A3, T1-A1) - 2 * LCS(T1-A3, T1-A1)}{2 * min(T1-A3, T1-A1)}$$

$$d_{Flex}(T1-A3, T1-A1) = 1 - \frac{(2*3) - 2*2}{2*3} = 1 - 0.33 = 0,66$$

Ou seja, o realizado é 66% aderente ao planejado correspondente. Portanto, pode-se tomar certas decisões, caso necessário, pois tem-se percepção da real situação.

7.3.2 Previsão de planejamentos

Como ilustrado na Figura 7.7, onde a sequência planejada é **NNNNN** e a sequência monitorada no instante *t* é **TNN**, indicando que o trem está em curso de uma nova trajetória cujo GHA ainda não é conhecido, mas pode ser previsto. A questão é como proceder à realização da previsão dos novos estados ou grau de aderência que o trem vai realizar? Como, dispondo-se dos dados que influenciam na realização de uma circulação conforme planejado, pode-se prever os possíveis comportamentos dos veículos nas próximas viagens?

Na circulação de um trem, em princípio, as leis da física deveriam prever o resultado ou grau de aderência de uma dada trajetória. Entretanto, seria necessário saber a *velocidade média* com a maior precisão e a condição tanto do trem como da malha no início, a velocidade do vento e suas variações com a severa precisão, etc. Isto quer dizer dispor-se das informações completas. Levando-se em consideração que todos esses fatores mudam com o tempo, não se pode, portanto, ter um resultado como total precisão ou exatidão. Para isso, esse procedimento, realizado no módulo de planejamento, baseia-se nas técnicas de probabilidades e nos Modelos de Markov para aproximar o resultado, conforme apresentadas nas seções 7.3.2 e 7.3.3.

Previsão probabilísticas com base nos dados históricos

Esta seção tem como objetivo aplicar as probabilidades com base nas ocorrências anteriores de circulações de trens, probabilidades estas podendo permitir que se faça uma previsão de comportamento de veículo quando ocorrer eventos registrados no passado. Para fazer isso, foi indispensável:

1. Compreender o que é uma probabilidade e chegar a quantificá-la;
2. Estudar as relações entre probabilidade e ignorância das causas (ou falta de informação);
3. Distinguir causalidade e correlações.

Compreender o que é uma probabilidade e como quantificá-la

Quando um trem percorre uma trajetória conforme planejado, pode-se dizer que tem-se "uma chance por três"de que sua aderência seja *Normal*, até mesmo "duas chances por seis para realizar duas trajetórias aderentes consecutivamente"; Qual então o sentido disto? A média estatística aqui é o número de graus de aderência *Normal* obtido dividido pelo número total de trajetórias percorridas. A probabilidade corresponde à "*tendência*"desta média estatística quando o número de trajetórias aumenta indefinidamente. Dizer que a probabilidade de se obter uma aderência Normal é de $\frac{1}{3}$ significa que a média estatística oscila em torno de $\frac{1}{3}$; e que é mais próximo deste valor quanto maior for o número de trajetórias: esta é a lei dos grandes números.

É assim que são definidas as probabilidades: constata-se que quando o número de trajetórias aumenta, as médias estatísticas se estabilizam em torno de um valor limite. Este valor limite é a probabilidade p.

Se p está próximo de 1, quase todos os eventos são bem sucedidos. Dir-se-á que o evento é quase certo acontecer. Se p está próximo de 0, quase nenhum é bem sucedido, dir-se-á que é quase impossível acontecer.

Vale ressaltar que, no contexto ferroviário estudado neste trabalho, identificou-se certos parâmetros, julgados mais influenciadores pelo fato de suas variações terem impactos na realização de uma circulação. Para tal, com base nas bases de dados disponíveis da empresa ABC, estabeleceu-se tabelas de correlação de cada um dos parâmetros considerados e o número de ocorrência relativo aos graus de aderência.

- **Correlação com relação à Temperatura**

Grau de aderência	Muito quente	Quente	Média	Fresco
Normal	33%	100%	70%	6.4%
Atrasado	66%	0.0%	19.6%	93%
Adiantado	0.0%	0.0%	10%	0.0%

Tabela 7.2: Tabela de Correlação com relação à Temperatura

Observando a Tabela 7.2, percebe-se, por exemplo, que todas as vezes que o parâmetro **temperatura** assumiu o valor *"muito quente"*, 33% de ocorrências foram consideradas de aderência *Normal*, 66% *Atrasado* e 0.0% *adiantado*, do total das amostras de testes usadas. Por outro lado, quando assumiu o valor *"fresco"*, 6.4% de ocorrências foram consideradas de aderência *Normal*, 93% *Atrasado* e 0.0% *adiantado*, de total das amostras de testes usadas.

- **Correlação com relação à Chuva**

Grau de aderência	Forte	Média	Fraca	Não
Normal	0.0%	0.0%	70%	68.8%
Atrasado	100%	100%	29%	21%
Adiantado	0.0%	0.0%	0.0%	10%

Tabela 7.3: Tabela de Correlação com relação à Chuva

Por outro lado, na Tabela 7.3, percebe-se, por exemplo, que todas as vezes que o parâmetro **chuva** assumiu o valor *"forte"*, 0.0% de ocorrências foram consideradas de aderência *Normal*, 100% *Atrasado* e 0.0% *adiantado*, do total das amostras de testes usadas. Em alternativa, quando *não choveu*, 68.8% de ocorrências foram consideradas de aderência *Normal*, 21% *Atrasado* e 10% *adiantado*, de total das amostras de testes usadas.

- **Correlação com relação ao Vento**

Grau de aderência	Violento	Forte	Moderado	Fraco
Normal	0.0%	7%	50%	76%
Atrasado	100%	93%	50%	10%
Adiantado	0.0%	0.0%	0.0%	14%

Tabela 7.4: Tabela de Correlação com relação ao Vento

O parâmetro *vento* na tabela 7.4, por sua vez, mostra que todas as vezes que assumiu o valor *"Violento"*, 0.0% de ocorrências foram consideradas de aderência *Normal*, 100% *Atrasado* e 0.0% *adiantado*, do total das amostras de testes usadas. Em contrapartida, quando *moderado*, 68.8%, apresentou uma equi-probabilidade entre o grau de aderência *Normal*, 50%, e *Atrasado* 50% e, por fim, 0.0% *adiantado*, de total das amostras de testes usadas.

- **Correlação com relação ao Clima**

Grau de aderência	Ensolarado	Nublado	Chuvoso
Normal	82.6%	65.7%	25.5%
Atrasado	17%	17%	74.4%
Adiantado	0.0%	17%	0.0%

Tabela 7.5: Tabela de Correlação com relação ao Clima

Quanto ao parâmetro *clima* da tabela 7.5, notou-se um equi-probabilidade entre o *Atrasado* e *Adiantado* com 17% de probabilidade quando o *tempo* assumir o valor *"quente"*. Entretanto, apresentou uma maior probabilidade de aderência *Normal* com 65.7%.

- **Correlação com relação à Condição da malha**

Grau de aderência	Boa	Ruim
Normal	65%	0.0%
Atrasado	27.7%	100%
Adiantado	7%	0.0%

Tabela 7.6: Tabela de Correlação com relação à Condição da malha

No que tange à condição da malha (tabela 7.6), no total de vezes que o parâmetro assumiu o estado *"bom"* obteve-se 65% de aderência *Normal*, 27.7% *Atrasado* e 7% *Adiantado*. Entretanto, atrasou com 100% de ocorrências cada vez que a condição da malha foi *"ruim"*.

A mesma interpretação segue para os casos da correlação com relação às Folhas mortas na malha (tabela 7.7), Água na malha (tabela 7.8) e Período do dia (tabela 7.9).

- **Correlação com relação às Folhas mortas na malha**

Grau de aderência	Sim	Não
Normal	0.0%	58%
Atrasado	100%	35%
Adiantado	0.0%	6%

Tabela 7.7: Tabela de Correlação com relação às Folhas mortas na malha

- **Correlação com relação à Água na malha**

Grau de aderência	Sim	Não
Normal	3%	68%
Atrasado	96%	24%
Adiantado	0.0%	7%

Tabela 7.8: Tabela de Correlação com relação à Água na malha

- **Correlação com relação ao Período do dia**

Grau de aderência	Manhã	Tarde	Noite
Normal	47.6%	46.7%	66.6%
Atrasado	38%	47.6%	33.3%
Adiantado	14%	4.7%	0.0%

Tabela 7.9: Tabela de Correlação com relação ao Período do dia

Distinguir causalidade e correlação

Sobre a confusão correlação-causalidade, observou-se que a temperatura fresca atrasa a circulação. Será isso devido à baixa temperatura? Isto não é evidente, pois também observou-se que houve atrasos significativos nos períodos de chuva; e sabe-se as relações chuva-temperatura. Neste fato, por si só, esta análise não permite tirar conclusões.

Outro fator é que ao constatar que a intensidade da chuva influencia na aderência, pode-se concluir - rapidamente - que quanto menor a intensidade de chuva, maior chance de ter aderência normal. Pode ser que seja verdade, mas a ”demonstração”é falsa: com a chuva fraca teve-se maior porcentagem de aderência normal que sem chuva. Por conseguinte, conclui-se que dois eventos podem ser correlacionados (ligados) sem ter relações de causa e efeito.

Demonstrar uma teoria com apenas estatísticas pode ser enganador. Muitas vezes, a teoria preexiste e os números são então utilizados para reforçá-la ”cientificamente”. Neste fito, três advertências são necessárias:

1. Sempre verificar se os dados são significativos. Em número, como visto, é óbvio; mas também em termo de qualidade.

2. A teoria deve ter o poder explicativo, ainda que seja apenas para saber em qual direção ler as correlações. Por exemplo, é agora bem estabelecido que historicamente, mudanças de temperatura estão intimamente relacionadas com alterações na concentração de dióxido de carbono na atmosfera. Mas não se pode fazer a economia de compreender pela teoria se é o aquecimento que cria o excesso de dióxido de carbono, ou o inverso;

3. Finalmente, embora a média seja significativa, sua operação pode ser complicada. Sabendo a média é importante, mas igualmente importante é saber se os resultados são muitas vezes longe ou perto desta média.

Estudo de correlação entre os parâmetros

Precisa-se descobrir a dependência ou independência entre os parâmetros. Para fazer isto, aplicou-se um critério de decisão chamado **”d-separação”**[4], o qual é aplicado a partir de um grafo causal, para se verificar se um conjunto de variáveis X é independente do outro conjunto Y, dado um terceiro conjunto Z. A ideia é associar a ”dependência”com ”ligação”(isto é, a existência de um caminho de ligação) e a ”independência”com ”separação”. Esta técnica baseia-se nos conceitos de *Redes Bayesianas*, que resulta em um modelo de grafos probabilísticos.

Começa-se por considerar a separação entre duas variáveis x e y; a extensão de conjuntos de variáveis é simples (dois conjuntos são separados se e somente se cada elemento de um conjunto é separado de cada elemento no outro).

Para se verificar a independência, optou-se por representar a árvore de decisão melhorada da Figura 7.8 em forma de grafos de dependência, ou rede Bayesiana, para melhor aplicar as condições sobrejacentes. A Figura 7.9 ilustra esta transformação.

[4]do ingês: d-separation

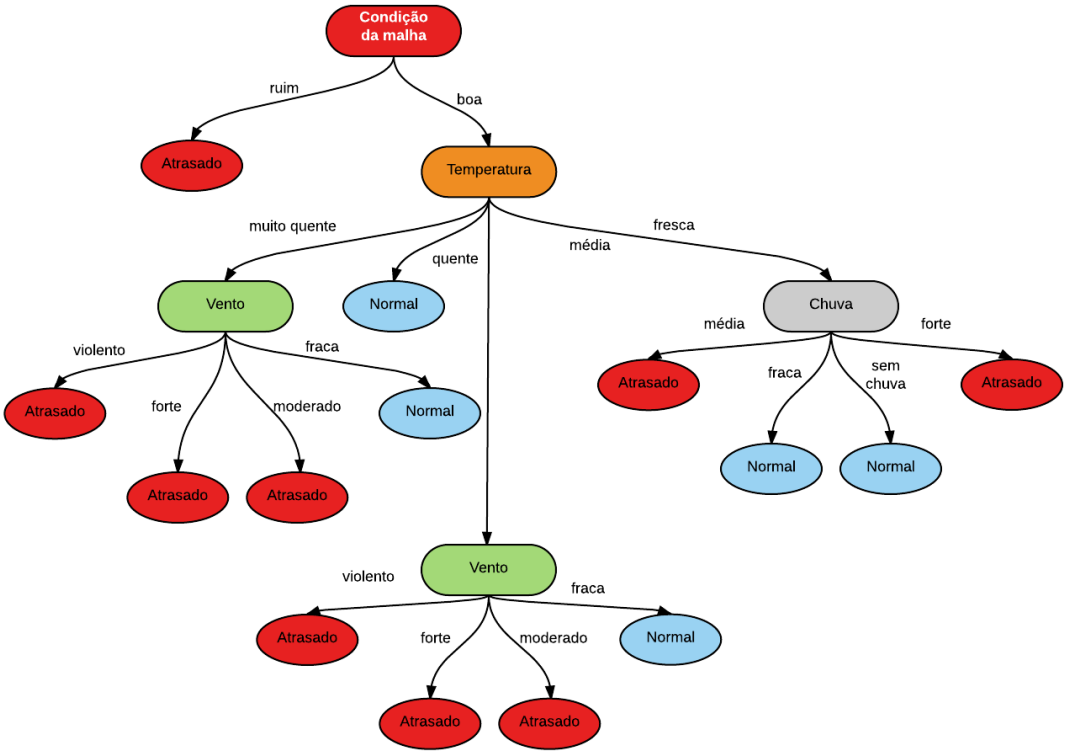

Figura 7.8: Árvore de decisão melhorada

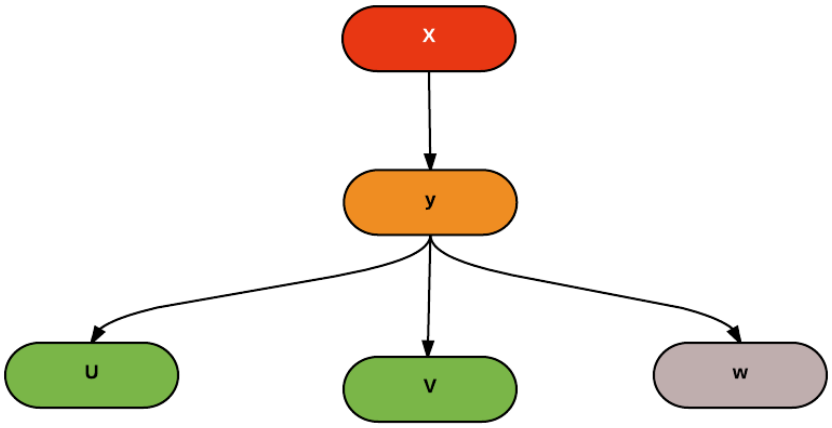

Figura 7.9: Representação da transformação da árvore de decisão obtida em grafo de dependência

Onde X: condição da malha, Y: temperatura, U: vento (caso temperatura muito quente), V: vento (caso temperatura média) e W: chuva.

Diz-se que dois vértices distintos A e B de V se estão d-separados por um conjunto de vértices $N \subseteq V$ quando, para todos os caminhos (não direcionados) entre A e B, pelo menos uma das três afirmações seguintes for satisfeita:

- Existe um vértice $V \in N$ no caminho entre A e B de modo que a conexão entre A e B através de V é serial;
- Existe um vértice $V \in N$ no caminho entre A e B de modo que a conexão através de V é divergente;
- Existe um vértice V de modo que V e todos os seus descendentes não estão em N e a conexão de A e B através de V é convergente;

A condição de *Markov* garante que se dois vértices A e B estão d-separados por um conjunto de vértices $N \subseteq V$ no grafo G, então A e B são condicionalmente independentes dado N.

Verificando a independência condicional, encontrou-se que:

- U é d-separado (independente) de X dado Y, pois tem-se uma conexão serial. Neste caso a probabilidade $P(U \mid Y,V,W) = \frac{P(U,V,W,Y)}{P(Y,V,W)} = \frac{P(U|Y)*P(Y)*P(V,W|V)}{P(Y,U,W)} = P(U \mid Y)$;
- V é d-separado (independente) de X dado Y, pois tem-se uma conexão serial. Neste caso a probabilidade $P(V \mid Y,U,W) = \frac{P(V,U,W,Y)}{P(Y,U,W)} = \frac{P(V|Y)*P(Y)*P(U,W|U)}{P(Y,V,W)} = P(V \mid Y)$;
- W é d-separado (independente) de X dado Y, pois tem-se uma conexão serial. Neste caso a probabilidade $P(W \mid Y,V,U) = \frac{P(W,V,U,Y)}{P(Y,V,U)} = \frac{P(W|Y)*P(Y)*P(V,U|V)}{P(Y,W,U)} = P(W \mid Y)$.
- U, V e W estão d-separados por N = V, onde tem-se uma conexão divergente.

Tendo em vista as independências condicionais entre as variáveis, pode-se proceder aos cálculos da distribuição conjunta de probabilidades. Para isso, associa-se a árvore de decisão com as devidas probabilidades calculadas nos dados históricos, onde tem-se a nova árvore probabilística da Figura 7.10.

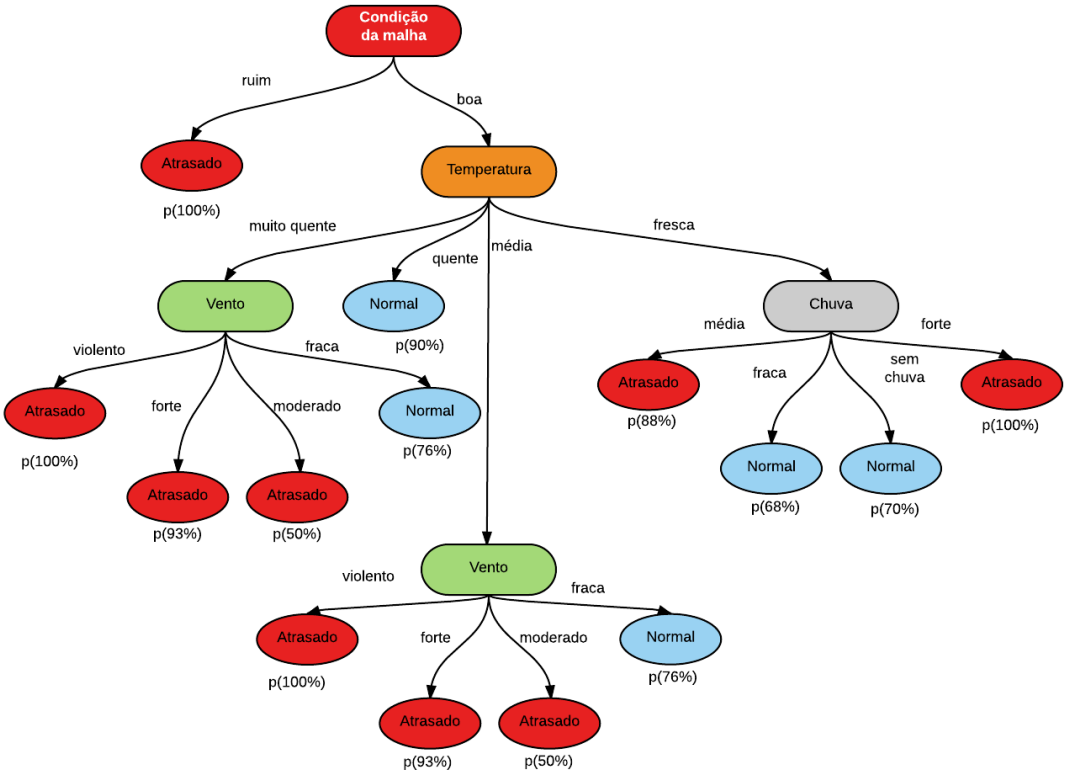

Figura 7.10: Árvore de decisão associada com as probabilidades

Onde, por exemplo:

- a probabilidade de um veículo atrasar quando a intensidade da chuva é média sabendo que a temperatura é fresca resulta: $\frac{(0.88)\cap P(temperatura-fresca)}{P(temperatura-fresca)}$
- a probabilidade de um veículo circular normalmente quando a intensidade da chuva é média sabendo que a temperatura é fresca resulta: $\frac{(1-0.88)\cap P(temperatura-fresca)}{P(temperatura-fresca)}$
- a probabilidade de um veículo circular normalmente quando a intensidade do vento é fraca sabendo que a temperatura é muito quente resulta: $\frac{(0.76)\cap P(temperatura-muitoquente)}{P(temperatura-muitoquente)}$

Observação: as probabilidades obtidas quando a temperatura está "Muito quente"e "Média"são iguais. Isto significa que o comportamento de um veículo depende muito mais da intensidade do "vento"do que da temperatura.

Demonstração:

- A sequência da árvore *Temperatura (muito quente)– Vento (violento)* remete a:

$\sum_{i=1}^{3} x_i * P(x_i \mid Vento = "violento") = 0.33 * 0.99 + 0.66 * 0.99 + 0.0 * 0.99 = 0.98$

A sequência da árvore *Temperatura (média)– Vento (violento)* remete a:

$\sum_{i=1}^{3} x_i * P(x_i \mid Vento = \text{"}violento\text{"}) = 0.70 * 0.99 + 0.196 * 0.99 + 0.1 * 0.99 = 0.98$

- A sequência da árvore *Temperatura (muito quente)– Vento (Forte)* remete a:

  $\sum_{i=1}^{3} x_i * P(x_i \mid Vento = \text{"}violento\text{"}) = 0.33 * 0.93 + 0.66 * 0.93 + 0.0 * 0.93 = 0.92$

  A sequência da árvore *Temperatura (média)– Vento (Forte)* remete a:

  $\sum_{i=1}^{3} x_i * P(x_i \mid Vento = \text{"}violento\text{"}) = 0.70 * 0.93 + 0.196 * 0.93 + 0.1 * 0.93 = 0.92$

Isto vale para outros casos onde a intensidade do vento é moderada e fraca, haja vista o exposto, pode-se concluir que a intensidade do vento tem maior pujança sobre a temperatura.

Com base nisto, faz-se uma previsão não apenas do provável comportamento ou aderência de um veículo, como também com o grau de probabilidade.

7.3.3 Previsão probabilística com base nos Modelos de Markov

Conforme introduzido, os processos estocásticos descrevem a evolução de um sistema que sofre variações com o tempo. E, conforme o princípio do Markov, o futuro depende apenas do presente e não do passado. Isto posto, a previsão probabilística com base nos modelos de Markov são utilizadas para fornecer uma previsão tendo em vista as condições ambientais do momento e, se disponíveis, as condições futuros, por exemplo, as previsões meteorológicas, neste caso as *precipitações* e informações sobre a condição da malha capturadas pelos sensores. Além do mais, graças as propriedades de probabilidades de transições, o modelo permite deduzir as informações faltantes.

Estes modelos foram utilizados para fornecer as previsões relativas tanto ao provável grau de aderência como às próximas ocorrências das sequências de grau de aderência. O que contempla à hipótese H3.

Previsão com base em precipitação

Considera-se:

$$X = \begin{cases} 0, & \text{se não tem precipitação no dia } t \\ 1, & \text{se tem precipitação no dia } t \end{cases}$$

Este processo estocástico é chamado cadeia de Markov se é somente se: $P(X_{t+1} = j \mid X_0 = k_0, X_{t-1} = k_{t-1}, X_t = i) = P(X_{t+1} = \mid X_t = i)$. Esta propriedade significa que a probabilidade

de uma precipitação futura, dispondo-se dos dados de precipitações passadas e a precipitação presente, não depende do passado, mas unicamente da precipitação atual.

A probabilidade de transição entre os estados i e j, $p_{ij} = P(X_{t+1} = j \mid X_t = i)$. Esta probabilidade é estacionária se $P(X_{t+1} = j \mid X_t = i) = P(X_1 = j \mid X_0 = i), t = 1, 2, ...$

A partir das probabilidades de transição, forma-se:

- A matriz de transição, tendo $M + 1$ (os estados das precipitações presentes) e $M + 1$ colunas (os estados das precipitações futuras), cada entrada da matriz corresponde a p_{ij}.
- Se $p_{ij} > 0$, há, no grafo de transição com $M + 1$ vértices, uma aresta entre o estado i e j.

Sabendo que a probabilidade de não ter precipitações hoje é, por exemplo, 0.6 e de ter precipitações hoje é 0.4, estas probabilidade não mudam, ainda que se considere o que ocorreu no dia anterior.

Observação: Estas probabilidades são fornecidas pelos especialistas ou pelos serviços meteorológicos, tais como o INPE e WMO.

A propriedade markoviana é satisfeita contanto que se tenha: $P(X_{t+1} = 0 \mid X_0 = k_0, X_1 = k_1, X_{t-1} = k_{t-1}, X_0) = P(X_{t+1} = 0 \mid X_t = 0)$ e $P(X_{t+1} = 0 \mid X_0 = k_0, X_1 = k_1, X_{t-1} = k_{t-1}, X_0) = P(X_{t+1} = 0 \mid X_t = 1)$.

Daí, tem-se uma cadeia de Markov cujas probabilidades de transição são:

$$p_{00} = P(X_{t+1} = 0 \mid X_t = 0) = 0.6$$

$$p_{10} = P(X_{t+1} = 0 \mid X_t = 1) = 0.4$$

Onde:

- P_{00} : a probabilidade de se não ter precipitação no dia *t+1* sabendo que não se têm precipitação no dia t;
- P_{10} : a probabilidade de se não ter precipitação no dia *t+1* sabendo que têm precipitação no dia t.

Com base nas propriedades de probabilidades de transição, deduz-se as informações faltantes:

$$p_{01} = P(X_{t+1} = 1 \mid X_t = 0) = 1 - 0.6 = 0.4$$

$$p_{11} = P(X_{t+1} = 1 \mid X_t = 1) = 1 - 0.4 = 0.6$$

Com:

- P_{01} : a probabilidade de se ter precipitação no dia *t+1* sabendo que não se têm precipitação no dia *t*;
- P_{11} : a probabilidade de se ter precipitação no dia *t+1* sabendo que têm precipitação no dia *t*.

Desta forma, tem-se a matriz de transição: $P = \begin{bmatrix} p_{00} & a_{01} \\ a_{10} & a_{11} \end{bmatrix} = \begin{bmatrix} 0.6 & 0.4 \\ 0.4 & 0.6 \end{bmatrix}$

E, graficamente tem-se:

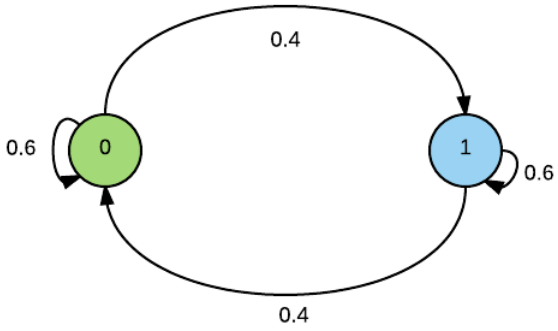

Figura 7.11: Grafo de transição de modelo estocástico da variação da precipitação

Uma vez descobertas as estimativas de se ter precipitação ou não entre um instante presente t e futuro $t+1$, o segundo passo consiste em coletar informações relativas à intensidade de precipitações, interpretá-las (de acordo com os detalhes de especialistas encontrados no anexo A), deduzir as informações por trás e, por conseguinte, prever um planejamento e o comportamento de veículos levando-se em consideração as possíveis ocorrências.

As precipitações são fornecidas em percentual, por exemplo, 89% que representa a probabilidade de ter chuva (ou neve no inverno). Desta forma, dada a precipitação, pode-se deduzir tanto a variação da temperatura (muito quente, quente, média ou fresca) como da chuva (forte, moderada, fraca ou sem chuva), e, em seguida, proceder ao processo de previsão ainda que não se disponha de informações sobre os parâmetros mencionados. A Tabela 7.10 fornece as precipitações e suas interpretações.

%	Caráter das precipitações	Ação a tomar	Duração	Significado
0-20 %	Muito isoladas	Nenhuma	–	risco muito baixo
30 %	Isoladas	Esconder-se	Breve	risco baixo
40-60 %	Dispersas	Guarda-chuva	Breve	risco moderado
60-70 %	Frequentes	Plano de substituição	durável	risco elevado
80 % e +	Chuva ou neve	Cancelar	Contínua	risco muito elevado

Tabela 7.10: Probabilidades de precipitações

Outro fator é que recebendo informações oriundas dos sensores, se, por exemplo, houver folhagem na malha ou água, isto implica na condição da malha, onde considera-se ”não boa”. Portanto, neste caso também pode-se proceder ao processo de previsão dispondo-se de pelo menos uma dessas informações.

Para tanto, a árvore de decisão associada com as probabilidades da Figura 7.10, transforma-se na árvore baseada em precipitação e condição da malha observadas da Figura 7.12.

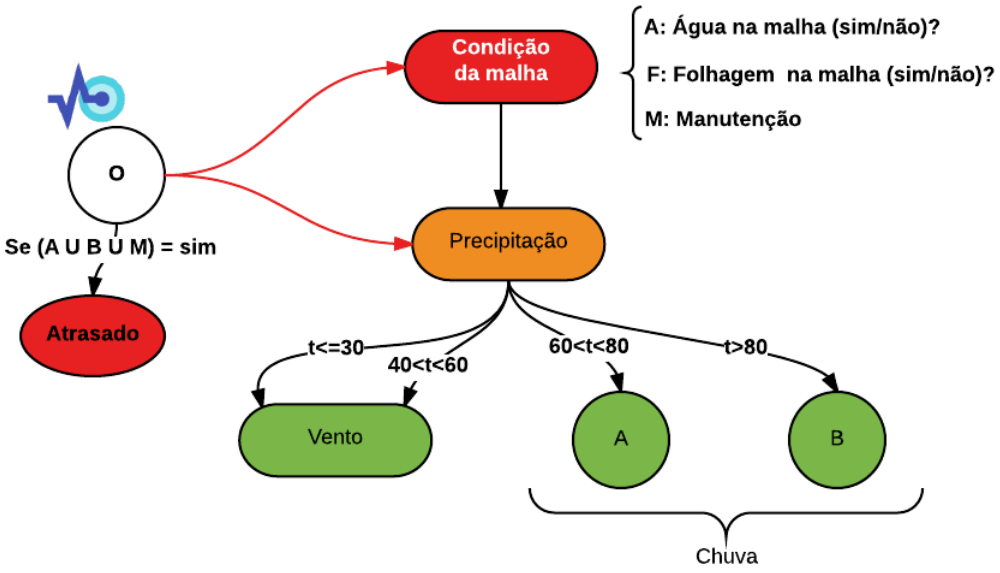

Figura 7.12: Árvore de decisão probabilística baseada em precipitação e condição da malha observadas

Onde o estado *O* é o estado de observação da condição da malha e intensidade das precipitações. A condição da malha a partir de agora pode ser determinada observando a água e folhagem na

malha, assim como a manutenção, sendo os três fatores que mais influenciam esta condição. Quanto à precipitação, é observado seu valor percentual, o qual leva a deduzir a temperatura e chuva. Manteve-se o parâmetro vento, pois quando tem-se risco mínimo ou menor de precipitação, faz-se necessário associar este parâmetro para se ter uma previsão mais realística dado o impacto deste último conforme a árvore apresentada na Figura 7.10.

O estado A, conforme a Tabela 7.12, indica um risco elevado de precipitação, e o estado B indica um risco muito elevado de precipitação. Quanto se tem A, as intensidades mais suscetíveis da chuva são (moderada e fraca), já no estado B (forte e moderada), daí a necessidade de recalcular as probabilidades tendo em vista a fusão de dois estados resultando em apenas um. Princípio de minimização de autômatos.

Para isso, tem-se:

- Se a precipitação for > 80, considera-se a possibilidade de que a chuva assuma dois estados (forte e moderado). Neste caso, calula-se a probabilidade $P(C_f \cup C_m)$, onde C_f e C_m simbolizam chuva forte e chuva moderada, respectivamente. Isto leva a $P(C_f \cup C_m) = P(C_f) + P(C_m) - P(C_f \cap C_m)$. Sabendo que a $P(C_f \cap C_m) = P(C_f) * (C_m) \simeq 0.87$, logo $P(C_f \cup C_m) = 0.88 + 0.99 - 0.87 \simeq 0.99$ ou 99%. Uma vez que ambos estados levam ao grau de aderência atrasado, pode-se concluir, portanto, que se a precipitação for > 80 é quase certeza que o veículo vai atrasar.

- Se a precipitação for entre $[60, 80]$, considera-se a possibilidade de que a chuva assuma dois estados (moderado e fraco). Neste caso, calula-se a probabilidade $P(C_m \cup C_{fr})$, onde C_m e C_{fr} simbolizam chuva moderada e chuva fraca, respectivamente. Isto leva a $P(C_m \cup C_{fr}) = P(C_m) + P(C_{fr}) - P(C_m \cap C_{fr})$. Tendo em vista que quando a chuva assume o estado moderado tem-se a aderência atrasado e quando assume o estado fraco tem-se a aderência normal, é imprescindível, antes de mais nada, convergir as duas tendência. Desta forma, tem-se $1 - 0.68 = 0.32$ que representa a conversão do grau de aderência normal do estado "fraco"ao grau "atrasado". Sabendo, agora, que a $P(C_m \cap C_{fr}) = P(C_m) * (C_{fr}) \simeq 0.28$, logo $P(C_m \cup C_{fr}) = 0.88 + 0.32 - 0.28 \simeq 0.92$ ou 92%. Uma vez que ambos estados agora levam ao grau de aderência atrasado, pode-se concluir, portanto, que se a precipitação for entre $[60, 80]$ tem probabilidade de 92% que o veículo atrase.

- Se a precipitação for entre $[40, 60]$, considera-se a possibilidade de que a chuva assuma dois estados (fraco e sem chuva). Neste caso, calula-se a probabilidade $P(C_{fr} \cup C_{sc})$, onde C_{fr} e C_{sc} simbolizam chuva fraca e sem chuva, respectivamente. Isto leva a $P(C_{fr} \cup C_{sc}) = P(C_{fr}) + P(C_{sc}) - P(C_{fr} \cap C_{sc})$. Sabendo que a $P(C_{fr} \cap C_{sc}) = P(C_{fr}) * (C_{sc}) \simeq 0.48$, logo

$P(C_{fr} \cup C_{sc}) = 0.68 + 0.70 - 0.48 \simeq 0.9$ ou 90%. Uma vez que ambos estados levam ao grau de aderência normal, pode-se concluir, portanto, que se a precipitação for entre $[40, 60]$ tem probabilidade de 90% que o veículo circule normalmente.

- E se a precipitação for entre $[0, 30]$, neste caso verifica-se as informações sobre o vento e procede-se nos cálculos de probabilidades da mesma forma como feito acima.

Modelagem de um processo observável

Esta parte ilustra um caso completo da modelagem e cálculos necessários para definir uma cadeia de Markov.

Seja um Modelo de Markov observável representando a circulação do trem. As observações são os estados definidos por Normal, Atrasado, Adiantado. O Modelo de Markov permite definir, por exemplo:

- A probabilidade de realização de uma sequência. Exemplo: Normal, Normal, Atrasado = **NNT**, no período do clima chuvoso com a intensidade de chuva fraca;

- A previsão de um estado futuro tendo conhecimento ou não do estado atual;

- A probabilidade de ser ter para k circulações as mesmas aderências.

A modelagem deste processo corresponde a:

- $X(t)$ = circulação do dia;

- O alfabeto $S = O = \{s_1 = o_1, s_2 = o_2, \ldots, s_n = o_m\}$ = { Normal, Atrasado, Adiantado} possíveis realizações de $X(t)$;

- A matriz de transição A ???

  Conforme mencionado na etapa B da Seção 6.2.3, a inicialização das probabilidade se faz a partir da classificação obtida por árvore de classificação o qual distribuiu os indivíduos, no caso as circulações, nas classes. Estas probabilidades são calculadas com base nos dados históricos com relação aos parâmetros temperatura, chuva e vento. As Tabelas 7.11, 7.12 e 7.13 ilustram as respectivas probabilidades.

  O primeiro símbolo indica o estado inicial, enquanto o segundo indica o estado de transição. Além disso, os símbolos N, T e d representam o grau de aderência Normal, Atrasado e Adiantado respectivamente.

Aderência	Forte	Moderada	Fraca	Sem chuva
NN	0.0%	0.0%	66.6%	78%
NT	0.0%	0.0%	33.3%	16%
ND	0.0%	0.0%	0.0%	0.7%
—	—	—	–	—
TT	99%	99%	20%	38%
TN	0.0%	0.0%	60%	53%
TD	0.0%	0.0%	0.0%	0.0%
—	—	—	–	—
DD	0.0%	0.0%	0.0%	50%
DN	0.0%	0.0%	0.0%	16%
DT	0.0%	0.0%	0.0%	33%

Tabela 7.11: Probabilidades de transições com base no parâmetro Chuva

Aderência	Violento	Forte	Moderada	Fraca
NN	0.0%	0.0%	66.6%	78%
NT	0.0%	99%	33.3%	1.0%
NN	0.0%	0.0%	0.0%	1.0%
—	—	—	–	—
TT	99%	85%	66.6%	0.0%
TN	0.0%	0.7%	28.5%	99%
TD	0.0%	0.0%	0.0%	0.0%
—	—	—	–	—
DD	0.0%	0.0%	0.0%	50%
DN	0.0%	0.0%	0.0%	33%
DT	0.0%	0.0%	0.0%	16.6%

Tabela 7.12: Probabilidades de transições com base no parâmetro Vento

Para a temperatura, decidiu-se simplificar ao reduzir a tabela com apenas dois valores, a saber, fresca e muito quente.

Aderência	Muito quente	Fresca
NN	0.0%	0.0%
NT	50%	99%
ND	0.0%	0.0%
—	—	—
TT	50%	83%
TN	50%	0.6%
TD	0.0%	0.0%
—	—	—
DD	0.0%	0.0%
DN	0.0%	0.0%
DT	0.0%	0.0%

Tabela 7.13: Probabilidades de transições com base no parâmetro Temperatura

Escolheu-se o caso onde o clima é chuvoso e a intensidade da chuva é fraca, o que resulta em $A = \{a_{ij} = P(s_j \mid s_i)\} = \begin{bmatrix} 0.78 & 0.53 & 0.16 \\ 0.16 & 0.38 & 0.33 \\ 0.07 & 0.0 & 0.5 \end{bmatrix}$

$A = \{a_{ij} = P[q_i = S_j \mid q_{t-1} = S_i] = P(o_j \mid o_i)\}, \forall i \in [1,N]$ e $j \in [1,N]$.

- As probabilidades inciais calculadas $\Pi = \{\pi = P(s_i)\} = \begin{bmatrix} 0.69 \\ 0.21 \\ 0.1 \end{bmatrix}$

- Por fim, o cenário pode representado graficamente conforme a Figura 7.13.

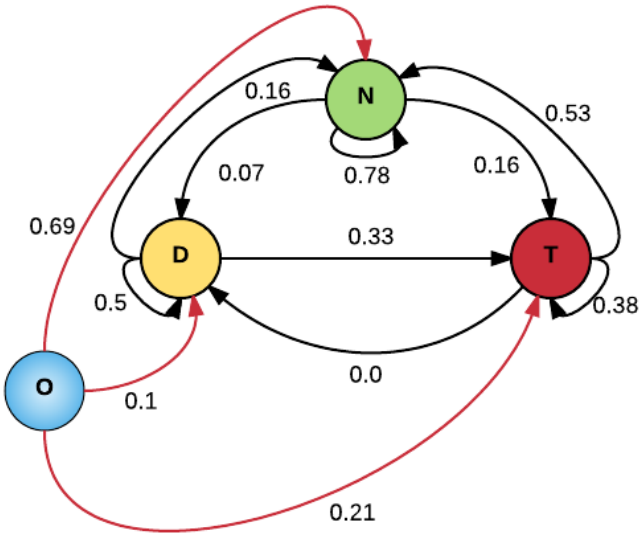

Figura 7.13: Modelo gráfico do Modelo de Markov observável das probabilidades de sucessão das sequências com relação ao período chuvoso com chuva fraca.

Previsões com base nas cadeias de Markov

Após ter modelado o processo, pode-se fazer previsões sobre o estado futuro deste processo. Desta forma, é possível realizar as operações tais como:

1. **Probabilidade de realização de uma sequência**

   Usou-se a fórmula $P(O) = P(o_n, o_n - 1, \ldots, o_1) = P(o_1) * \prod_{i=2}^{n} P(o_t \mid o_{t-1})$.

   Qual é a probabilidade que se tenha as sequências a seguir?

   - A probabilidade da sequência P(Normal, Normal, Normal, Normal) = P(N, N, N, N) ou P(NNNN) = $0.69 * 0.78 * 0.78 * 0.78 = 0.32744$
   - A probabilidade da sequência P(Normal, Normal, Normal, Normal, Normal) = P(N, N, N, N, N) ou P(NNNNN) = $0.69 * 0.78 * 0.78 * 0.78 * 0.78 = 0.255$
   - A probabilidade da sequência P(Normal, Normal, Normal, Normal, Normal, Normal) = P(N, N, N, N, N, N) ou P(NNNNNN) = $0.69 * 0.78 * 0.78 * 0.78 * 0.78 * 0.78 = 0.1212$
   - A probabilidade da sequência P(Atrasado, Atrasado, Normal, Normal, Atrasado, Normal, Adiantado) = P(T, T, N, N , T, N, D) ou P(TTNNTND) = $0.21 * 0.38 * 0.53 * 0.78 * 0.16 * 0.53 * 0.07 = 0.0036948$

- A probabilidade da sequência P(Adiantado, Normal, Normal, Atrasado) = P(D, N, N , T) ou P(DNNT) = $0.1 * 0.16 * 0.78 * 0.16 = 0.0199$
- A probabilidade da sequência P(Atrasado, Atrasado, Atrasado) = P(T, T, T) ou P(TTT) = $0 * 21 * 0.38 * 0.38 = 0.03$
- A probabilidade da sequência P(Normal, Normal) = P(N, N) ou P(NN) = $0 * 69 * 0.78 = 0.53$, que significa que a probabilidade de ter a sequência (NT) e (ND) é $1 - 0.53 = 0.47$. Portanto, a sequência (NN) tem maior probabilidade de se realizar.

2. **Previsão de um estado futuro**

   O sistema de eventos consiste em :

   $P(X(t) = o_i) = \sum_{j=1}^{m} \left[P(X(t) = o_i \mid X(t-1) = o_j) * P(X(t-1) = o_j\right]$

   Considerando a hipótese de estacionaridade, isto é, a mudança de um estado independe do tempo, este sistema torna-se: $P(X(t)) = o_i \mid X(t-1) = o_j) = P(o_i \mid o_j) = a_{ij}$.

   Isto posto, pode-se calcular, por exemplo, a probabilidade de ser ter aderência "Normal"no instante $t+3$:

   $P(X(3) = N) = 0.78 * (0.78 * 0.69 + 0.16 * 0.21 + 0.07 * 0.1) + 0.16 * (0.53 * 0.69 + 0.38 * 0.21 + 0 * 0.1) + 0.07 * (0.16 * 0.69 + 0.33 * 0.21 + 0.5 * 0.1) = 0.5380$

3. **Previsão de um estado futuro a partir de um estado conhecido**

   Usou-se a fórmula:

   $P(X(t_3) = o_k \mid X(t_1) = o_k) = \sum_{j=1}^{m} P(X(t_3) = o_k \mid X(t_2) = o_j) P(X(t_2) = o_j \mid X(t_1) = o_1)$

   Pode-se calcular a probabilidade de se ter a aderência "Normal"no instante $t+4$, sabendo que o grau de aderência atual é "Atrasado"no instante t:

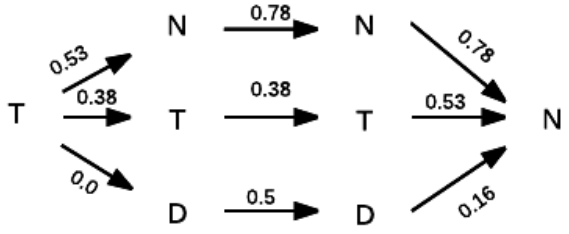

Figura 7.14: Previsão do estado "Normal"no instante *t+4* a partir do estado "Atrasado"

Analiticamente tem-se: $P(X(t_4) = "N" \mid X_1 = "T") = 0.78 * 0.78 * 0.53 + 0.53 * 0.38 * 0.38 + 0.16 * 0.5 * 0 = 0.40$

E ao calcular a probabilidade de se ter a aderência ”Normal”no instante $t+3$, sabendo que o grau de aderência atual é ”Atrasado”no instante t:

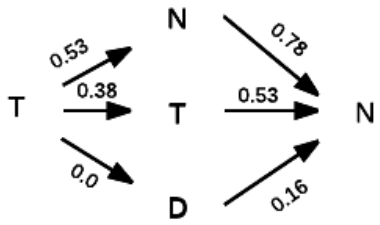

Figura 7.15: Previsão do estado ”Normal”no instante *t+3* a partir do estado ”Atrasado”

Analiticamente tem-se: $P(X(t_3) = \text{”}N\text{”} \mid X_1 = \text{”}T\text{”}) = 0.78*0.53+0.53*0.38+0.16*0 = 0.61$

4. **A Esperança matemática de se ter em *k* circulações as mesma sequências.**

   Por fim, pode-se calcular a Esperança matemática de números de sequências seguidas de graus de aderência de uma circulação considerando a chuva ”fraca”e o trem em *bom estado*. A Esperança é calcula por $E[X_n] = \frac{1}{1-a_{ij}}$.

   1. Para o estado ”Normal”tem-se: $E[Normal] = \frac{1}{1-0.69} = 3.22 \simeq 3$. Isto é, o número médio de se ter estados ”Normal”um após do outro é 3.
   2. Para o estado ”Atrasado”tem-se: $E[Atrasado] = \frac{1}{1-0.21} = 1.26 \simeq 1$. Isto é, o número médio de se ter estados ”Atrasado”um após do outro é 1.
   3. Para o estado ”Adiantado”tem-se: $E[Adiantado] = \frac{1}{1-0.1} = 1.11 \simeq 1$. Isto é, o número médio de se ter estados ”Adiantado”um após do outro é 1.

7.4 Avaliação de desempenho

Para validar a abordagem de análise proposta, foi realizado um conjunto de experimentos analíticos sobre as amostras de dados relativos às trajetórias de trens extraídos da base de dados ferroviários e dados simulados. Ambas as amostras foram construídas por um procedimento de amostragem estratificada efetuado sobre a população global, isto é, a base completa.

A qualidade das classes de trajetórias obtidas pela abordagem deste trabalho foi comparada com a qualidade das classes identificadas nas bases de dados fornecidas e simuladas, ou seja,

validação com os próprios dados dos cenários.

Dado que o objetivo da abordagem para a análise de dados sequenciais é de fornecer resultados interpretáveis e exploráveis pelo usuário, usou-se algumas medidas objetivas para uma melhor avaliação dos resultados. Os seguintes dois índices de qualidade foram então considerados:

1. **Índice de desempenho de previsão (DP)**

   Este índice foi utilizado para examinar a taxa de *boa previsão* da evolução das sequências $S_I = (g_{i,1}, g_{i,2}, \ldots, g_{i,T_i})$, onde $(g_{i,j} = (GHA_{i,j}, VM_{i,j}))$ em um conjunto de dados sequenciais Y. A principal ideia desse processo de avaliação consiste em selecionar as sequências S_i separadamente e em:

   - Eliminar o último estado g_{i,T_i} da sequência S_i
   - Classificar a nova sequência truncada (S_{tr}) nas k classes existentes, usando a Equação 7.1. Onde nota-se c_i a classe escolhida.

   $$c_i = argmax_{1 \leq c \leq k} \{P(S_{tr} \mid c_i = c, \phi_c)\} \tag{7.1}$$

   - Prever o estado q mais provável de ocorrer ou aparecer na extremidade desta sequência truncada. Isto é realizado utilizando a matriz de transição A_{c_i} associada à classe c_i escolhida. Este estado q sera, em seguida, comparado com o estado real g_{i,T_i} que foi suprimido da sequência como segue:

   $$DP_X = \frac{\sum_{i \in |x|} \omega_i}{|X|} \tag{7.2}$$

   Onde

   $$\omega_i = \begin{cases} 1, & \text{se} g_{i,T_i} = argmax_{1 \leq q \leq m} \left\{a_{c_i}(g_{i,T_{i-1}}, q)\right\} \\ 0, & \text{caso contrário} \end{cases}$$

   Os desempenhos de previsão foram examinados para duas amostras de circulações de trens, com os dados reais e simulados. Os resultados são obtidos com o auxílio de um processo de validação cruzado. Na prática, cada conjunto de dados sequenciais (trajetórias de trens) foi dividido em quatro partes distintas. Duas destas partes (50% das trajetórias (reais e simulados)) foram usadas como um conjunto de treinamento e o resto (50% (reais e simulados)) para a fase de teste. Este processo foi repetido seis vezes e, em seguida, definiu-se a média dos resultados. A amostra de aprendizagem serve para gerar uma tipologia de trajetórias e modelar as classes obtidas a partir da abordagem aplicada. A amostra de teste é em seguida utilizada para avaliar os resultados de previsão.

Para uma boa estimativa dos resultados obtidos, o desempenho de previsão também foi avaliado para a base de aprendizagem.

2. **Índice de homogeneidade Intraclasse (HI)**

   Proposto por (ESTACIO-MORENO et al., 2004), este índice tem um papel fundamental para a validação de um problema de classificação automática de dados sequenciais. Neste trabalho, a homogeneidade intraclasse é considerada como um índice probabilístico que reflete a estabilidade, a coesão e a facilidade de interpretação das classes obtidas por um processo de classificação automatizada. Onde, quanto maior o valor de homogeneidade intraclasse, maior a possibilidade das classes de partição serem compactas e de fácil interpretação.

   O índice HI é dado como: para uma partição p em k classes $\{C_1,\ldots,C_k\}$ do conjunto de sequências $S=\{S_1,\ldots,S_n\}$, a homogeneidade intraclasse HI(P) é definida pela média das homogeneidades intraclasse das k classes da partição P como segue:

$$HI(P)=\frac{\sum_{c=1}^{k} HI_c}{k} \tag{7.3}$$

   Ou

$$HI_c=\sum_{S_i\in c} \tag{7.4}$$

   Onde:

   - $\delta_i = 0$ se $P(c_i=c \mid S_i,\phi)<0.5$
   - $\delta_i = 1$ se $P(c_i=c \mid S_i,\phi)\geq 0.5$
   - c_i é a classe da sequência S_i
   - $\phi=\{\phi_1,\ldots,\phi_k\}$ representa os parâmetros de todas as classes.

As tabelas 7.14 e 7.15 mostram os resultados obtidos sobre as amostras de trajetórias de dados reais e simulados em termo de desempenho de ***previsão*** e ***homogeneidade intraclasse***. Este resultado indica com clareza que a abordagem proposta mostrou bom desempenho.

Outrossim, pôde-se confirmar a eficácia dos modelos decisionais da classificação e de previsão propostos. A presente abordagem permite fornecer uma tipologia de trajetórias com as classes homogêneas e de fácil interpretação para um processo de previsão.

HI	DP (d. aprend-real)	DP (d. aprend-simul)	DP (d. teste reais)	DP (d. de teste simul)
87,0%	72,6%	75,2%	69,8 %	73,1%

Tabela 7.14: Desempenho relativo à base de dados com 217

HI	DP (d. aprend-real)	DP (d. aprend-simul)	DP (d. teste reais)	DP (d. de teste simul)
92,0%	74,2%	79,7%	73,8 %	75,3%

Tabela 7.15: Desempenho relativo à base de dados com 623

Os trabalhos correlatos (MOHAMMAD et al., 2012; ANNABELL et al., 2011; KECMAN et al., 2015b, 2015a) também foram realizados neste contexto, entretanto, estes tratam apenas os trens de passageiros, de alta velocidade e nas regiões onde as infra-estruturas são consideras melhores e plataformas impecáveis. Além disto, embora tenham utilizados os métodos estocásticos que também foram utilizados neste trabalho, não consideram fatores como água na malha, folhagem na malha, diferentes intensidades da chuva e vento, os quais foram estudados neste trabalho. Além do mais, este trabalho tratou especialmente o contexto de trens de cargas, os quais têm mais restrições. A tabela da Figura 7.16 ilustra a comparação entre os trabalhos sobrejacentes e o presente trabalho. O único trabalho que forneceu o desempenho resultante da validação é o de (KECMAN et al., 2015a), portanto, no presente trabalho foram utilizadas as mesmas métricas a fim de comparar o desempenho dos dois trabalhos.

Trabalho	Tipo tráfego	Fatores	Métodos	Validação
(KECMAN et al, 2015a)	**Passageiros**, alta velocidade, Beijing - Shanghai	**Evolução do atraso**, tempo real	Processo estocástico	Histórico-Dados reais - 71%
(KECMAN et al, 2015b)	**Passageiros**, alta velocidade, Sweden	Predição **Evolução do atraso**, **tempo real** - **independência de trens**	Rede Bayesiana	Hist. Dados reais – resultado parcial
(MOHAMMAD et al, 2012)	**Passageiros,** rede ampliada, Germany	Distribuição de atraso nas conexões de trens	Abordagem Probabilística	Dados reais
(ANNABELL et al, 2011)	**Passageiros,** rede ampliada, Germany, políticas de Deutsche Bahn AG	**Propagação de atraso, evento de partida e chegada**	Modelo Estocástico de previsão	Dados artificiais discretos
Este trabalho	**Passageiros**, **cargas**	**Evolução do atraso**, **Tempo/clima**, **ambiente**, **tempo real**	Processo Estocástico, Abordagem Probabilística, Árvore de classificação	Histórico- Dados reais e simulados 72.6% RD 75.8% SD

Figura 7.16: Comparação do presente trabalho com os trabalhos correlatos.

Onde RD e SD significam dados reais e dados simulados respectivamente.

A avaliação dos experimentos realizados, assim como a comparação com os trabalhos correlatos, mostrou a utilidade da abordagem proposta nesta dissertação como uma ferramenta de apoio à tomada de decisão no contexto ferroviário do tipo estudado, permitindo responder aos problemas de classificação e previsão de trajetórias ou comportamentos dos veículos. Além do mencionado, o setor operacional pode, com base nesta ferramenta, prever os comportamentos futuros do tráfego, o que permitirá uma boa organização a nível operacional do tráfego, gestão do material e roteamento ou planejamento de rotas.

7.5 Protótipo da ferramenta de Simulação

Neste seção apresenta-se o protótipo de uma ferramenta de apoio a decisão baseado na abordagem apresentada neste capítulo, onde foi aplicado no planejamento de rotas para trens, a análise de trajetórias assim como nas previsões de planejamentos e comportamentos futuros do trem. Este protótipo serve, além da prova de conceito, para caracterizar, a nível técnico, as diferentes contribuições teóricas deste trabalho.

A Figura 7.17 permite visualizar as diferentes trajetórias de trens existentes na base de dados com relação ao prefixo de trem selecionado. As Figuras 7.18 e 7.19 realizam o cálculo de grau de aderência entre o efetivamente realizado com o planejado, fornecendo o resultado correspondente. Já os ambientes das Figuras 7.20 e 7.21 fornecem a previsão do GHA com base nas informações ambientais fornecidas. As Figuras 7.22 e 7.23 realizam a previsão da probabilidade de realização de uma sequência.

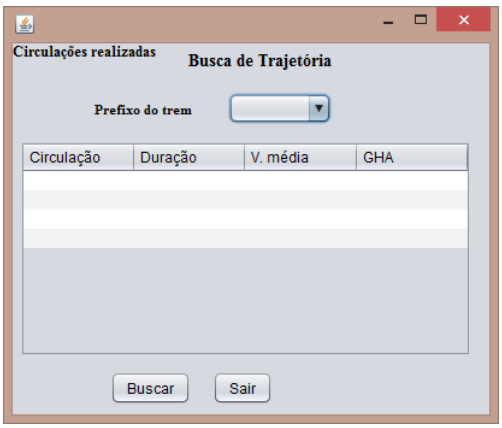

Figura 7.17: Busca de Trajetória

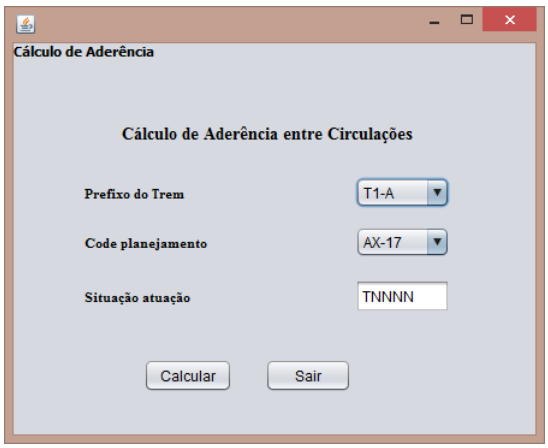

Figura 7.18: Cálculo de Grau de Aderência

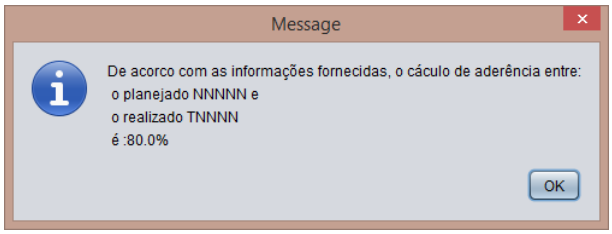

Figura 7.19: Resultado do Cálculo de Grau de Aderência

Figura 7.20: Previsão do GHA com base nas informações ambientais

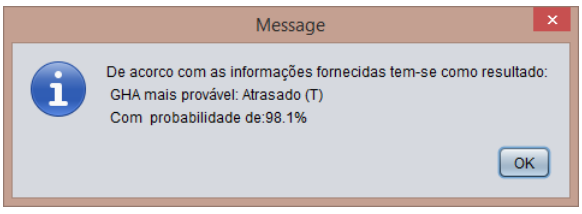

Figura 7.21: Resultado da previsão do GHA com base nas informações ambientais

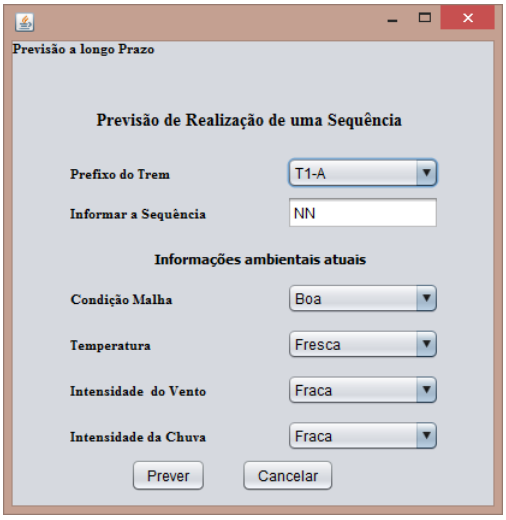

Figura 7.22: A previsão da probabilidade de realização de uma sequência

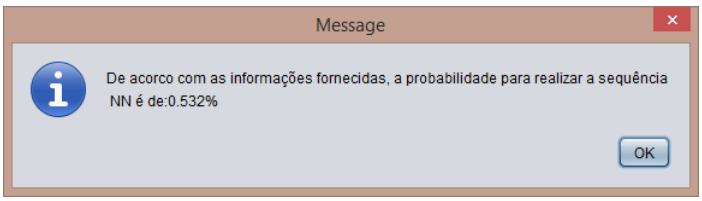

Figura 7.23: Resultado da previsão da probabilidade de realização de uma sequência

7.6 Conclusão

Apresentou-se, neste capítulo, uma abordagem de análise de dados sequências heterogêneos que permite a classificação automática e a previsão de sequências. Tal abordagem baseou-se em uma metodologia mista que envolve a abordagem por árvore de classificação e as cadeias de Markov, a qual fornece uma tipologia das sequências em classes de pertencimento homogêneas. As operações realizadas sobre os dados coletados retornam tanto o grau de aderência (classe de pertencimento), como um modelo probabilístico de geração de dados (cadeia de Markov) informando as relações entre estados de sequências da classe, assim como as probabilidades de ocorrências. Este modelo favorece melhor interpretabilidade das classes, e, para tanto, pode ser aplicado para agrupar novas sequências (agrupamento) e estimar suas evoluções futuras (previsão).

Com finalidade de validar a abordagem proposta neste trabalho, aplicou-la-se em conjuntos de dados de trajetórias extraídas na base de dados de circulações de trens com os dados reais e simulados. Esta aplicação permitiu apreciar as vantagens e a pertinência dos resultados obtidos com relação aos objetivos e às hipóteses levantadas. A aplicação desta abordagem nos dados

históricos reais e simulados permitiu avaliar seu desempenho, pela sua capacidade de ilustrar quão aderente esta sendo uma circulação com relação ao planejado, assim como pelo fato de atribuir as probabilidade de ocorrências de um evento qualquer no determinado instante.

Por último, apresentou-se um protótipo como ferramenta de apoio à decisão afim de consubstanciar as propostas teóricas no acoplamento entre a abordagem de classificação por árvore de classificação e as cadeias de Markov. Trata-se de um protótipo de apoio à decisão dedicado à classificação e previsão de trajetórias de trens. O intuito desta ferramenta é antecipar a atividades respeitantes às circulações, monitorar um determinado veículo para descobrir quão aderente é sua evolução com relação ao planejamento, e, quando necessário, realizar as previsões do comportamento nas próximas viagens.

Capítulo 8

CONCLUSÃO

Este capítulo apresenta as conclusões gerais das contribuições deste trabalho e as perspectivas futuras.

8.1 As contribuições

Como parte desta dissertação, procurou-se fornecer soluções para as questões da logística de transporte, especificamente do transporte ferroviário, questões essas consistindo em classificação de dados complexos e análise de sequências temporais, onde aplicou-se as soluções propostas nos dados de circulações de trens da empresa ABC.

Ao se tratar da classificação dos dados ferroviários, o objetivo era proporcionar uma tipologia apurada das circulações de trens, sendo estas circulações descritas por um conjunto de dados complexos, como alternativa à classificação em Grupo Homogêneos de Aderência (GHAs) hodierno. Para tal, foi necessário definir um método de classificação automática para processar dados e descrições complexos e heterogêneas para permitir interpretar facilmente as classes obtidas com base em um conjunto representativo de indivíduos (trens).

Para isso, baseou-se na técnica de árvore de classificação que possui a vantagem de fornecer uma partição podada de dados onde a separação interclasse é realizada conjuntamente com a coesão intraclasse. Além do mais, a árvore de classificação possui também outra característica importante, a qual diz respeito à aplicabilidade tanto nos dados clássicos como simbólicos.

Um experimento foi realizado sobre os conjuntos de dados tanto reais como simulados e os resultados foram comparados com as ocorrências observadas no final de cada trajetória. Esta abordagem pode ser utilizada para diferentes tipos de regiões e contextos logísticos de

transporte, no entanto, deve-se adequar os parâmetros.

Foi proposta, em seguida, uma fórmula com duas variantes que foi chamada DeFlex (Distância de edição flexível), derivada da Distância de Edição ou Distância de Levenshtein, que concerne ao cálculo de similaridade entre duas sequências, neste caso de grau de aderência entre tanto uma circulação completa realizada com o planejado como uma circulação em curso (monitorada) com o planejado. Uma vez que tem-se duas situações, onde a primeira consistindo em duas sequências de comprimentos iguais e a segunda consistindo em duas sequências de comprimentos distintos, propôs-se a primeira variante da fórmula para calcular o grau de aderência do primeiro caso, e a segunda variante para calcular o grau de aderência do segundo caso, onde impõe-se a condição de equidistância. Esta fórmula foi aplicada convertendo os Grupos Homogêneos de Aderências (GHA) em caracteres, a saber, N para o GHA ”Normal”, T para o GHA ”Atrasado”e D para o GHA ”Adiantado”, as quais combinações formam uma sequências de caracteres. Tal fórmula mostrou-se eficiente, pois levou aos resultados analiticamente eficientes e esperados. Além do mais, foi proposta uma abordagem de análise de dados sequências que permite a classificação automática assim como a previsão das sequências. Com base em uma metodologia mista combinando a abordagem por árvore de classificação e as cadeias de Markov, que fornece uma tipologia das sequências das classes homogêneas. Estas classes são representadas por tanto um grau de aderência como por um modelo probabilístico (cadeia de Markov) explicando as probabilidades de pertencimento de uma sequência a uma classe. Tendo em vista a melhor interpretabilidade e eficiência dos modelos Markovianos, pode-se, neste nível estimar as evoluções futuros (previsão) das sequências. Outrossim, com base nos modelos Markovianos aplicados, é possível calcular a probabilidade de realização de uma sequência, realizar a previsão de um estado futuro e realizar a previsão de um estado futuro a partir de um estado conhecido, isto independe do instante, entretanto, considerando as mesmas condições ambientais.

Para validar a abordagem proposta, aplicou-se as propostas aqui apresentadas sobre os cenários de trajetórias de trens e comparou-se os resultados com as recomendações de especialistas, assim como com a efetividade dos próprios dados de ocorrências na base de dados. Além disto, a abordagem proposta inclui ou fornece informações sobre eventos que possuem uma maior probabilidade de ocorrer. Com base nisto, os especialistas podem tomar decisões, até mesmo fazer o planejamento ou replanejamento com o conhecimento prévio das prováveis ocorrências. É imprescindível lembrar que as hipóteses inicialmente definidas foram contempladas.

Sob a perspectiva técnica, desenvolveu-se uma ferramenta em java como protótipo para

permitir avaliar consubstancialmente a importância da abordagem teórica proposta. Protótipo esse consiste em apoiar à tomada de decisão quanto à aderência, à classificação e à previsão de circulações de trens.

Ademais, as propostas iniciais deste trabalho têm gerado uma publicação em uma conferência com Qualis B1, um trabalho em submissão em uma conferência com Qualis B1 e um trabalho em curso de submissão em um periódico com Qualis A2. Além do mais, um trabalho não diretamente relacionada com a abordagem proposta desta dissertação foi publicada em periódico.

8.2 Trabalhos futuros

Como parte do trabalho sobre a análise de dados sequencias, visa-se procurar a melhoria da qualidade das previsões da proposta deste trabalho incluindo as técnicas *Fuzzy* para aprimorar as inferências estatísticas no que concerne à diferenciação dos valores que os parâmetros utilizados neste trabalho podem assumir. Por exemplo, avaliar a temperatura (muito quente, quente, médio, etc...), o vento (forte, moderado, etc...) tratando certas incertezas.

Outro aspecto seria a introdução do Raciocínio Baseado em Caso (RBC) para o aprendizado com base em experiência passada. Este seria usada como a solução de novos problemas por meio da utilização de casos anteriores já conhecidos. Neste caso, uma situação ou evento apresentado é resolvido com a reutilização da solução de um problema anterior parecido com o atual; tal solução podendo ser aplicada em sua completude ou apenas parcialmente no novo problema, com tal possibilidade de ser *modificada consoante os requisitos da nova situação*, o que contorna os desafios de uma modelagem baseada no Modelo Oculto de Markov (HMM) de um processo real, onde este mostra-se eficaz se os parâmetros do modelo são corretamente estimados. Estas estimativas são, na maioria da vezes, imprecisas e isto por dois motivos: inicialmente o processo não obedece às restrições do HMM (os estados não são coerentes) e a dificuldade de obter as estimativas confiáveis de todos os parâmetros, o que na ocorrência de novos eventos não analisados nem encontrados na bases de dados históricos daria problema, mas que poderia ser resolvido com o RBC visto sua vantagem de adaptação e modificação de parâmetros consoante os requisitos da nova situação.

Outra questão levantada é a escolha da melhor circulação realizada dentre todas as circulações realizadas e consideradas melhores sob certa perspectiva. Para isso, pensou-se em incrementar o método de **Marie Jean Antoine Nicolas de Caritat, marquês de Condorcet** chamado ”Paradoxo de Condorcet (le paradoxe et gagnant de Condorcet)”no século XVIII, onde um sistema é

considerado melhor (vencedor) se pertencer ao grupo dos melhores (vencedores) e for o melhor (vencedor) dos melhores (vencedores).

Perfazendo, este trabalho foi realizado em cima dos dados de um tempo curto do sistema ferroviário. Outra etapa poder-se-á consistir em aumentar o tamanho da base, isto é, coletar dados de mais anos afim de explorar amplamente os possíveis cenários e permitir, portanto, a melhor precisão nos resultados.

REFERÊNCIAS

AGRAWAL, R.; SRIKANT, R. Mining sequential patterns. In: *conference on data engineering (ICDE'95)*. [S.l.: s.n.], 1995. p. 3–14.

ANNABELL, B.; GEBHARDT, A.; MÜLLER-HANNEMANN, M.; OSTROWSKI, M. Stochastic Delay Prediction in Large Train Networks. In: CAPRARA, A.; KONTOGIANNIS, S. (Ed.). *11th Workshop on Algorithmic Approaches for Transportation Modelling, Optimization, and Systems*. Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2011. (OpenAccess Series in Informatics (OASIcs), v. 20), p. 100–111. ISBN 978-3-939897-33-0. ISSN 2190-6807. Disponível em: ⟨http://drops.dagstuhl.de/opus/volltexte/2011/3270⟩.

ANTUNES, C.; OLIVEIRA, A. L. Temporal data mining: An overview. In: *KDD Workshop on Temporal Data Mining*. [S.l.: s.n.], 2001. p. 1–13.

BAGGENSTOSS, P. A modified baum-welch algorithm for hidden markov models with multiple observation spaces. In: *Acoustics, Speech, and Signal Processing, 2000. ICASSP '00. Proceedings. 2000 IEEE International Conference on*. [S.l.: s.n.], 2000. v. 2, p. II717–II720 vol.2. ISSN 1520-6149.

BENZECRI, J. *L'Analyse des données: La taxinomie*. [S.l.]: Paris:Dunod, 1973., 1973. v. 1. ISBN 9782040108915.

BERKHIN, P. *Survey Of Clustering Data Mining Techniques*. San Jose, CA, 2002. ⟨http://www.accrue.com/products/rp_cluster_review.pdf⟩.

BERTRAND, P.; DIDAY, E. Une généralisation des arbres hiérarchiques : les représentations pyramidales. *Revue de Statistique Appliquée*, Revue de Statistique Appliquée, v. 38-3, 1990.

BEZDEK, J. C.; PAL, N. R. Some new indexes of cluster validity. *IEEE Transactions on Systems, Man and Cybernetics*, IEEE, v. 28, n. 3, p. 301–315, 1998.

BOBBIN, J. An incremental viterbi algorithm. In: *Machine Learning and Applications, 2007. ICMLA 2007. Sixth International Conference on*. [S.l.: s.n.], 2007. p. 104–111.

BOCK, H.; DIDAY, E. *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data*. Springer Berlin Heidelberg, 2000. ISBN 9783540666196. Disponível em: ⟨https://books.google.com.br/books?id=1qqBvpkMU80C⟩.

BOX, G. E. P.; JENKINS, G. M. *Time Series Analysis: Forecasting and Control*. 3rd. ed. Upper Saddle River, NJ, USA: Prentice Hall PTR, 1994. ISBN 0130607746.

BUSSIECK, M.; WINTER, T.; ZIMMERMANN, U. Discrete optimization in public rail transport. *Mathematical Programming*, Mathematical Programming, v. 79, p. 415–444, 1997.

BUZAN, D.; SCLAROFF, S.; KOLLIOS, G. Extraction and clustering of motion trajectories in video. In: IEEE. *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*. [S.l.], 2004. v. 2, p. 521–524.

CADEZ, I.; HECKERMAN, D.; MEEK, C.; SMYTH, P.; WHITE, S. Visualization of navigation patterns on a web site using model-based clustering. In: ACM. *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*. [S.l.], 2000. p. 280–284.

CADEZ, I. V.; GAFFNEY, S.; SMYTH, P. A general probabilistic framework for clustering individuals and objects. In: ACM. *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*. [S.l.], 2000. p. 140–149.

CAPRARA, A.; KROON, L.; MONACI, M.; PEETERS, M.; TOTH, P. Passenger railway optimization. In: BARNHART, C.; LAPORTE, G. (Ed.). *Handbooks in Operations Research and Management Science*. Elsevier, 2007. v. 14, cap. 3, p. 129—187. Disponível em: ⟨http://arrival.cti.gr/index.php/Documents/0035⟩.

CARVALHO, F. de A. T. de. Proximity coefficients between boolean symbolic objects. In: DIDAY, E.; LECHEVALLIER, Y.; SCHADER, M.; BERTRAND, P.; BURTSCHY, B. (Ed.). *New Approaches in Classification and Data Analysis*. Springer Berlin Heidelberg, 1994, (Studies in Classification, Data Analysis, and Knowledge Organization). p. 387–394. ISBN 978-3-540-58425-4. Disponível em: ⟨http://dx.doi.org/10.1007/978-3-642-51175-2_44⟩.

CELEUX, G.; DIDAY, E.; LECHEVALLIER, Y.; GOVAERT, G.; H, R. *Classification automatique des données*. Paris:Dunod, 1989., 1989. ISBN 9782040187989. Disponível em: ⟨http://www.worldcat.org/title/classification-automatique-des-donnees/oclc/19990958⟩.

CHATFIELD, C. *The analysis of time series: an introduction*. 6th. ed. Florida, US: CRC Press, 2004.

CHAVENT, M. *Analyse des données symboliques: une méthode divisive de classification*. Tese (Doutorado) — Université Paris Dauphine, 1997.

CHAVENT, M.; GUINOT, C.; LECHEVALLIER, Y.; TENENHAUS, M. Méthodes divisives de classification et segmentation non supervisée : recherche de la typologie de la peau humaine saine. In: NUMDAM. *RSA*. [S.l.], 1999. p. 87–99.

CHESHOMI, S.; RAHATI-Q, S.; AKBARZADEH-T, M.-R. Hmm training by a hybrid of chaos optimization and baum-welch algorithms for discrete speech recognition. In: *Digital Content, Multimedia Technology and its Applications (IDC), 2010 6th International Conference on*. [S.l.: s.n.], 2010. p. 337–341.

CORDEAU, J.-F.; TOTH, P.; VIGO, D. A survey of optimization models for train routing and scheduling. *CoRR*, Transportation Science, v. 32(4), 1998. Disponível em: ⟨http://dx.doi.org/10.1287/trsc.32.4.380⟩.

DELSARTE, P.; GENIN, Y. The split levinson algorithm. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, v. 34, n. 3, p. 470–478, Jun 1986. ISSN 0096-3518.

DENG, C.; ZHENG, P. A new hidden markov model with application to classification. In: *Intelligent Control and Automation, 2006. WCICA 2006. The Sixth World Congress on*. [S.l.: s.n.], 2006. v. 2, p. 5882–5886.

DEQUIER, J. *Chaînes de Markov et applications*. 2005.

DIDAY, E. Une nouvelle méthode en classification automatique et reconnaissance des formes la méthode des nuées dynamiques. *Revue de statistique appliquée*, v. 19, n. 2, p. 19–33, 1971.

DIDAY, E. Une représentation visuelle des classes empiétantes : les pyramides. *RAIRO*, Revue RAIRO APII, v. 20, 1986.

DIDAY, E.; GOVAERT, G. Classification with adaptive distance. *COMPTES RENDUS HEBDOMADAIRES DES SEANCES DE L ACADEMIE DES SCIENCES SERIE A*, GAUTHIER-VILLARS 120 BLVD SAINT-GERMAIN, 75280 PARIS, FRANCE, v. 278, n. 15, p. 993–995, 1974.

DIDAY, E.; KODRATOF, Y. Des objets de l'analyse des données à ceux de l'analyse des connaissances. *CEPADUES*, CEPADUES EDITIONS, 1991.

DUNN, J. C. A fuzzy relative of the isodata process and its use in detecting compact, well-seperated clusters. *Journal of Cybernetics*, Taylor & Francis, v. 3, n. 3, p. 32–57, 1973.

DUSSAUCHOY, A. Generalized information theory and decomposability of systems. *International Journal on General System*, International Journal on General System, v. 9, 1982.

EL-GOLLI, A. *Extraction de données symboliques et cartes topologiques : Application aux données ayant une structure complexe*. Tese (Doutorado) — Université Paris Dauphine, 2004.

ELGHAZEL, H. *Classification et Prévision des Données Hétérogènes: Application aux Trajectoires et Séjours Hospitaliers*. Tese (Doutorado) — Université Claude Bernard Lyon 1, 2007.

ESTACIO-MORENO, A.; BARBARY, O.; GALLINARI, P.; PIRON, M. Classification de données biographiques : application à des trajectoires migratoires vers cali (colombie). In: *Revue de Statistique Appliquée*. [S.l.: s.n.], 2004. v. 54, p. 33–54.

FERRO, C.; STEPHENSON, D. Uncertainty and inference for verification measures. *Weather and Forecasting*, Weather and Forecasting, v. 22, p. 637–650, 2007.

FERRO, C.; STEPHENSON, D. Extremal dependence indices: improved verification measures for deterministic forecasts of rare binary events. *Monthly Weather Revue*, Monthly Weather Revue, v. 26, p. 699–713, 2011.

FISHER, D. Knowledge acquisition via incremental conceptual clustering. *Machine Learning*, Kluwer Academic Publishers-Plenum Publishers, v. 2, n. 2, p. 139–172, 1987. ISSN 0885-6125. Disponível em: ⟨http://dx.doi.org/10.1023/A\%3A1022852608280⟩.

FISHER, D. Rock : A robust clustering algorithm for categorical attributes. *Machine Learning*, Information Systems, v. 2, 1987.

FORGY, E. W. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics*, v. 21, p. 768–769, 1965.

GILLELAND, E. *CONFIDENCE INTERVALS FOR FORECAST VERIFICATION*. [S.l.]: OpenSky, 2010.

GOWDA, K. C.; DIDAY, E. Symbolic clustering using a new dissimilarity measure. *Pattern Recogn.*, Elsevier Science Inc., New York, NY, USA, v. 24, n. 6, p. 567–578, abr. 1991. ISSN 0031-3203. Disponível em: ⟨http://dx.doi.org/10.1016/0031-3203(91)90022-W⟩.

GUHA, S.; RASTOGI, R.; SHIM, R. Cure : an efficient clustering algorithm for large databases. In: ACM SIGMOD. *proceedings of ACM SIGMOD International Conference on Management of Data*. [S.l.], 1998. p. 73–84.

HAKIM, M. *La prévision du temps*. 2015. Disponível em: ⟨http://www.meteofrance.fr/prevoir-le-temps/la-prevision-du-temps/les-techniques-de-prevision⟩.

HARTIGAN, J. A.; WONG, M. A. Algorithm as 136: A k-means clustering algorithm. *Applied statistics*, JSTOR, p. 100–108, 1979.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The elements of statistical learning: data mining, inference and prediction*. 2. ed. Springer, 2008. Disponível em: ⟨http://scholar.google.com/scholar.bib?q=info:roqIsr0iT4UJ:scholar.google.com/&output=citation&hl=en&ct=citation&cd=0⟩.

HSIAO, R.; TAM, Y.-C.; SCHULTZ, T. Generalized baum-welch algorithm for discriminative training on large vocabulary continuous speech recognition system. In: *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*. [S.l.: s.n.], 2009. p. 3769–3772. ISSN 1520-6149.

HUANG, Y.; BENESTY, J. *Audio Signal Processing for Next-Generation Multimedia Communication Systems: For Next-Generation Multimedia Communication Systems*. Springer, 2004. ISBN 9781402077685. Disponível em: ⟨https://books.google.com.br/books?id=YIXRVNAaBdUC⟩.

ICHINO, M.; YAGUCHI, H. Generalized minkowsky metrics for mixed feature type data analysis. *IEEE Transactions on Systems, Man and Cybernetics*, IEEE, v. 24, 1994.

IRM. *Caractéristiques de quelques paramètres climatiques*. 2015. Disponível em: ⟨http://www.meteo.be/meteo/view/fr/360361-Parametres.html\#ppt_7369963⟩.

JACCARD, P. Nouvelle recherche sur la distribution florale. *Bulletin Society Vaud*, Science Natural, v. 44, 1908.

JAIN, A. k.; MURTY, M. N.; FLYNN, P. J. Data clustering: A review. *ACM Computing Surveys*, ACM, v. 31, 1999.

JAMBU, M. *Méthodes de base de l'analyse des données*. [S.l.: s.n.]. ISBN 978-2212052565.

JAMBU, M. *Introducion au Data Mining; Analyse intelligence des données*. Eyrolles, 2000. ISBN 978-2212052558. Disponível em: ⟨http://www.amazon.fr/INTRODUCTION-MINING-Analyse-intelligente-donn\%C3\%A9es/dp/2212052553⟩.

KARYPIS, G.; HAN, E.-H.; KUMAR, V. Chameleon : Hierarchical clustering using dynamic modeling. *Revue de Statistique Appliquée*, Computer, v. 32, 1999.

KECMAN, P.; CORMAN, F.; MENG, L. Train delay evolution as a stochastic process. In: *Proceedings of the 6th International Conference on Railway Operations Modelling and Analysis : RailTokyo2015*. [S.l.: s.n.], 2015.

KECMAN, P.; CORMAN, F.; PETERSON, A.; JOBORN, M. Stochastic prediction of train delays in real-time using bayesian networks. In: *Conference on Advanced Systems in Public Transport : CASPT2015*. [S.l.: s.n.], 2015.

KODRATOFF, Y.; NAPOLI, A.; A, Z. D. *Bulletin Association Française de l'Intelligence Artificielle*. 2001. Disponível em: ⟨www.afia.asso.fr⟩.

KOSKELA, T.; LEHTOKANGAS, M.; SAARINEN, J.; KASKI, K. Time series prediction with multilayer perceptron, fir and elman neural networks. In: INNS PRESS SAN DIEGO, USA. *Proceedings of the World Congress on Neural Networks*. [S.l.], 1996. p. 491–496.

KRIOUILE, A. *La reconnaissance automatique de la parole et les modèles de Markoviens cachés*. Tese (Doutorado) — Thèses de Doctorat, Université de Nancy I, 1990.

KRUSKAL, J. B. An overview of sequence comparison: Time warps, string edits, and macromolecules. *SIAM review*, SIAM, v. 25, n. 2, p. 201–237, 1983.

LAXMAN, S.; SASTRY, P. A survey of temporal data mining. *Sadhana*, Springer India, v. 31, n. 2, p. 173–198, 2006. ISSN 0256-2499. Disponível em: ⟨http://dx.doi.org/10.1007/BF02719780⟩.

MACQUEEN, J. et al. Some methods for classification and analysis of multivariate observations. In: OAKLAND, CA, USA. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. [S.l.], 1967. v. 1, n. 14, p. 281–297.

MALERBA, D.; ESPOSITO, F.; GIOVIALE, V.; TAMMA, V. Comparing dissimilarity measures for symbolic data analysis. In: ETK-NTTS'01. *Proceedings of the Joint Conferences on New Techniques and Technologies for Statistics and Exchange of Technology and Knowhow*. 2001. p. 473–481. Disponível em: ⟨http://cgi.csc.liv.ac.uk/valli/Papers/ntts-asso.pdf⟩.

MALI, K.; MITRA, S. Clustering and its validation in a symbolic framework. *Pattern Recognition Letters*, Elsevier, v. 24-14, 2003.

MENTZER, J.; GOMES, R. Further extensions of adaptive extended exponential smoothing and comparison with the m-competition. *Journal of the Academy of Marketing Science*, Springer-Verlag, v. 22, n. 4, p. 372–382, 1994. ISSN 0092-0703. Disponível em: ⟨http://dx.doi.org/10.1177/0092070394224006⟩.

MICHALSKI, R. S.; STEPP, R. E. Learning from observations : Conceptual clustering, in machine learning: An artificial intelligence approach. v. 24, 1983.

MITCHELL, C.; JAMIESON, L. Modeling duration in a hidden markov model with the exponential family. In: *Acoustics, Speech, and Signal Processing, 1993. ICASSP-93., 1993 IEEE International Conference on*. [S.l.: s.n.], 1993. v. 2, p. 331–334 vol.2. ISSN 1520-6149.

MOHAMMAD, H. K.; SCHNEE, M.; WEIHE, K.; ZORN, H.-P. Reliability and Delay Distributions of Train Connections. In: DELLING, D.; LIBERTI, L. (Ed.). *12th Workshop on Algorithmic Approaches for Transportation Modelling, Optimization, and Systems*. Dagstuhl,

Germany: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2012. (OpenAccess Series in Informatics (OASIcs), v. 25), p. 35–46. ISBN 978-3-939897-45-3. ISSN 2190-6807. Disponível em: ⟨http://drops.dagstuhl.de/opus/volltexte/2012/3701⟩.

MOON, M. A.; MENTZER, J. T.; JR., D. E. T. Customer demand planning at lucent technologies: A case study in continuous improvement through sales forecast auditing. *Industrial Marketing Management*, v. 29, n. 1, p. 19 – 26, 2000. ISSN 0019-8501. Disponível em: ⟨http://www.sciencedirect.com/science/article/pii/S001985019900108X⟩.

NG, R. T.; HAN, J. Clarans: A method for clustering objects for spatial data mining. *Knowledge and Data Engineering, IEEE Transactions on*, IEEE, v. 14, n. 5, p. 1003–1016, 2002.

OLIVIER, J. *mieux comprendre les précipitations*. 2015. Disponível em: ⟨http://www.meteomedia.com/nouvelles/articles/mieux-comprendre-les-probabilites-de-precipitations-dans-vos-previsions/31271/⟩.

OUDELHA, M.; AINON, R. Hmm parameters estimation using hybrid baum-welch genetic algorithm. In: *Information Technology (ITSim), 2010 International Symposium in*. [S.l.: s.n.], 2010. v. 2, p. 542–545. ISSN 2155-897.

PAILLEUX, J.; STRAUSS, B. *La Météorologie. Numéro spécial "Prévision numérique du temps", 8ème série, n 30*. la Société météorologique de France et par Météo France, 2000. Disponível em: ⟨http://www.meteo.fr⟩.

PATERSON, M.; DANČÍK, V. *Longest common subsequences*. [S.l.]: Springer, 1994.

PETER, J. B.; RICHARD, A. D. *Introduction to Time Series and Forecasting*. Springer New York, 1996. ISBN 978-1-4757-2526-1. Disponível em: ⟨http://books.google.com.br/books?id=_w5AJtbfEz4C&printsec=frontcover&hl=pt-PT&source=gbs_ge_summary_r&cad=0\#v=onepage&q&f=false⟩.

RABINER, L. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, v. 77, n. 2, p. 257–286, Feb 1989. ISSN 0018-9219.

RAKOTOMALALA, R. Interactive clustering tree : Une méthode de classification descendante adaptée aux grands ensembles de données. In: NUMDAM. *Data Mining et Apprentissage statistique*. [S.l.], 2007. p. 75–94.

ROUSSEEUW, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, Elsevier, v. 20, p. 53–65, 1987.

SAKOE, H. Two-level dp-matching–a dynamic programming-based pattern matching algorithm for connected word recognition. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, IEEE, v. 27, n. 6, p. 588–595, 1979.

SALZBERG, S. C4.5: Programs for machine learning by j. ross quinlan. morgan kaufmann publishers, inc., 1993. *Machine Learning*, Kluwer Academic Publishers, v. 16, n. 3, p. 235–240, 1994. ISSN 0885-6125. Disponível em: ⟨http://dx.doi.org/10.1007/BF00993309⟩.

SAPORTA, G. *Probabilités, analyse des données et statistique*. TECHNIP, 2006. ISBN 9782710809807. Disponível em: ⟨http://www.amazon.fr/Data-mining-statistique-d\%C3\%A9cisionnelle-\%C3\%A9dition/dp/2710810174⟩.

SCHROEDER, A. Analyse d'un mélange de distributions de probabilité de même type. *Revue de Statistique Appliquée*, [s.n.], v. 24, n. 1, p. 39–62, 1976. Disponível em: ⟨http://eudml.org/doc/106012⟩.

SIU, M.; CHAN, A. A robust viterbi algorithm against impulsive noise with application to speech recognition. *Audio, Speech, and Language Processing, IEEE Transactions on*, v. 14, n. 6, p. 2122–2133, Nov 2006. ISSN 1558-7916.

SOKAL, R.; MICHENER, C. Statistical method for evaluating systematic relationships. *University of Kansas science bulletin*, University of Kansas science bulletin, v. 38, 1958.

SOKAL, R. R.; SNEATH, P. H. A. Principles of numerical taxonomy. *Freeman*, Freeman, 1963.

TAN-JAN, H.; CHEN, B.-S. Novel extended viterbi-based multiple-model algorithms for state estimation of discrete-time systems with markov jump parameters. *Signal Processing, IEEE Transactions on*, v. 54, n. 2, p. 393–404, Feb 2006. ISSN 1053-587X.

TUFFÉRY, S. *Data Mining et Statistique Décisionnelle*. TECHNIP, 2012. ISBN 9782710810179. Disponível em: ⟨http://www.amazon.fr/Data-mining-statistique-d\%C3\%A9cisionnelle-\%C3\%A9dition/dp/2710810174⟩.

VILELA, R. P.; LUCIANO, M.; RAFAEL, L.; ANDERSON, P. An heuristic approach to train dispatch planning. In: *Joint Rail Conference*. [S.l.: s.n.], 2014.

WEIJS, S.; R, v. N.; N, v. d. G. Kullback-leibler divergence as a forecast skill score with classic reliability-resolution-uncertainty decomposition. *Monthly Weather Revue*, Monthly Weather Revue, v. 138, p. 3387–3399, 2010.

WILLIAMS, W.; LAMBERT, J. Multivariate methods in plan ecology. *Journal of Ecology*, Journal of Ecology, v. 47-1, 1959.

WWOSC, . *SEAMLESS PREDICTION OF THE EARTH SYSTEM: FROM MINUTES TO MONTHS*. World Meteorological Organization, 2014. Disponível em: ⟨http://s2sprediction.net/file/documents_publications/wmo_1156_en.pdf⟩.

ZHANG, S. na; WU, D. an; WU, L.; LU, Y. bin; PENG, J. yan; CHEN, X. yang; YE, A. dang. A markov chain model with high-order hidden process and mixture transition distribution. In: *Cloud Computing and Big Data (CloudCom-Asia), 2013 International Conference on*. [S.l.: s.n.], 2013. p. 509–514.

ÈVE, C. *La météo au quotidien*. 2015. Disponível em: ⟨http://www.meteo.org/phenomen/nuage.htm\#QU'EST-CEQU'UNNUAGE?⟩

Apendice A

CONHECIMENTO METEOROLÓGICO

Este capítulo apresenta uma ideia geral sobre a meteorologia, seus impactos em diversos domínios, especialmente no contexto do tráfico ferroviário, assim como os elementos meteorológicos úteis para se fazer as possíveis previsões do planejamento.

A.1 Introdução

Com o desenvolvimento de instrumentos cada vez mais eficientes, a previsão do tempo melhorou muito ao longo dos últimos trinta anos, especialmente nos casos que variam de dois a sete dias. Pode-se pensar que o progresso vai continuar. Mas todos os fenômenos não são igualmente previsíveis: a ”previsibilidade”do tempo e seus progressos dependem muito do fenômeno meteorológico considerado.

Observa-se que as atividades humanas são cada vez mais dependentes das condições meteorológicas, assim como a segurança das pessoas e dos bens em caso de fenômenos climáticos extremos. Há muito tempo que procurou-se controlar e prever o comportamento da atmosfera, com variados graus de sucesso. A maioria dos países tem serviços meteorológicos responsáveis, entre outras atividades, pelo desenvolvimento e manutenção de redes de observação, bem como pela previsão do tempo. Essas atividades meteorológicas são altamente federadas, especialmente a nível europeu, mas também a nível mundial.

A previsão numérica do tempo consiste em aplicar à atmosfera as leis da hidrodinâmica que impulsionam seu desenvolvimento através de um sistema de equações diferenciais parciais, sistema que se resolve por métodos numéricos. Este princípio de cálculo do estado futuro da atmosfera a partir de um estado presente foi concebido no final da Primeira Guerra Mundial por

um meteorologista britânico, Richardson, trinta anos antes da invenção do computador.

Anteriormente, os meteorologistas esforçavam-se para analisar subjetivamente os fenômenos considerados relevantes para a previsão do tempo em uma região, utilizando as melhores observações disponíveis, antes de extrapolar estes fenômenos no tempo por leis empíricas.

Em seguida, os modelos têm desempenhados um papel cada vez mais importante no desenvolvimento da previsão, na medida em que foram melhorados nos seguintes aspectos (PAILLEUX; STRAUSS, 2000):

- Modelagem de processos físicos cada vez mais avançados da atmosfera e suas fronteiras;
- Consideração de observações cada vez mais variadas para descrever os modelos de estado inicial dos modelos e algoritmos matemáticos mais sofisticados;
- Aumento da resolução espacial.

A cadeia de previsão meteorológica prossegue, em seguida, com o exame e interpretação dos resultados do modelo por especialistas. Trata-se, por um lado, de traduzir os resultados numéricos em uma forma utilizável e, por outro lado, submeter esses resultados a um exame crítico a fim de discriminar a informação confiável da informação incerta e, quando necessária, detectar sinais provenientes de eventos perigosos. Tendo em conta os múltiplos fatores que afetam a qualidade dos modelos, este exame crítico pode ser muito diferente de um dia para outro, de modo que um desafio fundamental para os meteorologistas é de detectar em meio à enorme quantidade de dados produzidos pelos modelos, quais são importantes ou relevantes. Os especialistas dispõem, portanto, de um sistema de processamento de dados e visualização, chamada **Sinergia**, desenvolvido especificamente por este fim.

A última etapa do intervalo de previsão é moldar resultados de uma forma adaptada às necessidades do usuário. Neste caso, a diversidade é extrema, uma vez que refere-se ao termo genérico de ”público em geral”, assim como o usuário profissional que requere uma apresentação específica.

A.2 Previsão probabilística

Hodiernamente, a previsão probabilística está sendo usada em todas as escalas para explicitamente representar incertezas previsionais relacionadas com as condições iniciais e incertezas do modelo. A previsão probabilística pode ser avaliada em diferentes formas com a escolha de abordagem dependente da forma como a previsão se destina a ser utilizada (WWOSC, 2014).

Especificamente, os elementos desse conjunto podem ser avaliados individualmente como previsões determinísticas ou o conjunto pode ser resumido usando um membro representante tal como a média do conjunto; eles pode ser avaliados como previsões probabilísticas (por exemplo, ao converter o conjunto de previsão para uma distribuição de probabilidade ou estimando as probabilidades para eventos específicos); ou eles podem ser avaliadas como uma distribuição. Enquanto os métodos voltados para as duas primeiras opções são relativamente bem estabelecida, métodos para a avaliação da distribuição como um todo ainda são relativamente novas e melhoradas de diagnóstico e abordagens intuitivas para avaliação da previsão probabilística ainda são necessários (WWOSC, 2014).

A verificação tradicional de previsões probabilísticas e distribuição da *previsão probabilística* são baseadas principalmente em métricas, como *Brier skill score* (para previsões de probabilidade) (WEIJS et al., 2010) e *CRPS* (para o conjunto distribuição), e diagnóstico, como a confiabilidade e característica de operação relativa *(ROC: relative operating characteristic)*, para avaliar a consistência de propagação de erros e confiabilidade, assim como a discriminação de probabilidade e conjunto de previsões.

Ferro e Stephenson (2011) considera a ”justeza”dos *scores* como o *Brier Score* e *CRPS* para avaliação do conjunto de previsões. No estudo de Ferro um *score* é definido como justo se ”a expectativa do score com relação às distribuições de ambos os membros do conjunto e da verificação de observação é otimizado quando essas distribuições coincidem”.

A.3 Verificação de incerteza nos resultados

A incerteza nos resultados surge de muitas fontes. Muitas vezes, as observações são inerentemente incertas devido à medição, bem como erros de representatividade espacial e temporal, e aplicação de verificação de previsão de amostras limitadas de previsões leva a incerteza relacionada à variabilidade da amostragem. A variabilidade da amostragem é um pouco mais simples de se dar conta do que a incerteza relacionada à observação, e métodos para estimar intervalos de confiança estatísticos foram definidos para muitas medidas de verificação (por exemplo (FERRO; STEPHENSON, 2007; GILLELAND, 2010) e estão incluídos em pelo menos alguns pacotes de verificação (por exemplo, Ferramentas de Modelo de avaliação (MET): http://www.dtcenter.org/met/users/).

Estas abordagens geralmente levam em conta os efeitos de correlações temporais; representando os impactos das correlações espaciais sobre os intervalos de confiança é um pouco mais problemático e geralmente não é tratado adequadamente. Métodos para aplicar os interva-

los de confiança para diferenças de desempenho para amostras pareadas levam a comparações estatísticas mais poderosas do modelo de desempenho de previsão.

Enquanto considerando a incerteza de observação em estudos de verificação ainda é um tema de pesquisa, alguns conhecimentos foram adquiridos nos últimos anos. No entanto, muito mais conhecimento e novos recursos são necessários. Fundamentalmente, como modelos têm melhorados, não é mais apropriado ignorar erro de observação; Na verdade, como modelos melhoram, o erro resultante em previsões se tornará cada vez mais perto do erro nas observações. Idealmente, os vieses em observações podem ser removidos (quando conhecidos), mas é mais difícil explicar os erros aleatórios, que levam á verificações mais pobres de *scores* para as previsões determinísticas. Os resultados da verificação para o conjunto de previsões são caracterizadas pela confiabilidade mais pobres.

A.4 As técnicas de previsão

Existem várias técnicas de previsão complementares. Estas técnicas baseiam-se em modelos numéricos que simulam o comportamento da atmosfera. Elas exigem a esperteza de meteorologistas para analisar os resultados do modelo e traduzi-los em termos compreensíveis pelos usuários (HAKIM, 2015).

A.4.1 A previsão determinística

De acordo com Hakim (2015), para um período de algumas horas a 3 ou 4 dias, pode-se utilizar uma técnica chamada previsão ”determinística”. Esta técnica baseia-se na utilização de modelos de Previsão Numérica do Tempo, que simulam o comportamento da atmosfera com base nas equações da física e da termodinâmica.

O primeiro passo é estabelecer previsão determinística a partir de observações de uma representação cartográfica do tempo, isto é, um estado inicial da atmosfera. O modelo calcula, em seguida, a evolução dos parâmetros meteorológicos (pressão, temperatura, vento) ao longo do tempo. A partir de um estado da atmosfera determinado, o modelo desenvolve um cenário de evolução desses parâmetros, é por isso que fala-se de previsão ”determinística”.

As simulações são então analisadas por um especialista que conhece os limites do modelo. Ele ajusta, modifica e traduz os resultados em termos de tempo ”observáveis”como a duração e intensidade da precipitação, as temperaturas mínimas e máximas, a ocorrência de tempestades, rajadas de vento ou névoas.

Contudo, esta abordagem determinística não permite avaliar as incertezas sobre o único cenário de previsão utilizado. No entanto, é essencial para os usuários de previsão meteorológica ter acesso a esta informação. Isto é o que permite a previsão global, que prevê, além do cenário mais provável, as incertezas associadas (de confiança, os cenários alternativos, ...).

A.4.2 A previsão global

Hakim (2015) ainda afirma que cada passo da previsão do tempo está sujeita a incertezas que podem afetar a qualidade da previsão final. As observações são heterogêneas no espaço e no tempo, os modelos são representações únicas necessariamente imperfeitas do comportamento da atmosfera, e a atmosfera em si tem um comportamento caótico: dois estados iniciais muito próximos podem levar a situações muito diferentes em poucos dias ou algumas horas.

Inovações na área de medição e pesquisa sobre processos atmosféricos permitem reduzir gradualmente as primeiras duas fontes de incerteza. Mas a terceira é uma realidade física que escapa, uma propriedade do ambiente.

Em vez de focar-se em apenas uma abordagem determinística que produz um único cenário de evolução exclusivo para cada cartografia do tempo, os meteorologistas, portanto, usam cada vez mais um método que leva em consideração estas incertezas: a previsão geral (ou probabilística). Este consiste em realizar simulações a partir de variadas descrições do estado inicial da atmosfera. Estas não são escolhidas ao acaso: elas são representativas das incertezas identificadas nas medidas. A previsão global oferece, portanto, vários cenários para a evolução da atmosfera. Suas convergências ou divergências informam os especialistas sobre a probabilidade de ocorrência de cada cenário: convém a eles escolher o mais provável e quantificar a incerteza em torno destas previsões. Esta quantificação da incerteza permite os meteorologistas combinarem suas previsões para além de quatro dias com um índice de confiança. A figura A.1 ilustra exemplo de previsões.

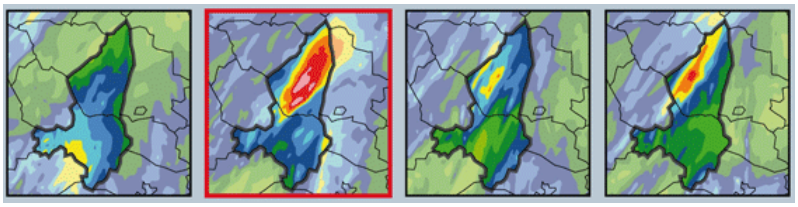

Figura A.1: Previsão global (exemplo de previsões obtidas) (HAKIM, 2015)

A.4.3 Previsão imediata

A mesma fonte ressalta que a previsão imediata diz respeito a prazo muito curto: contando alguns minutos a algumas horas. Essas previsões são usadas por exemplo, por meteorologistas para monitorar os eventos espeço-temporais perigosos, como os episódios do Mediterrâneo, assim como a alerta meteorológica vermelha ou laranja. Além do mais, afirma também ser necessário para monitorar o risco de chuva durante eventos desportivos ou culturais, por exemplo: trata-se de fornecer cronologias muito precisas de episódios chuvosos das próximas horas. A previsão imediata também é usada em outros domínio, como a aviação ou *transporte ferroviário*, domínio estudado nesta dissertação de mestrado.

O desenvolvimento destas previsões baseia-se na utilização de observações. O método de assimilação de dados, que sejam suficientes para assegurar a qualidade das previsões além de algumas horas, limita a qualidade para as primeiras horas. Para prever o tempo nas próximas horas, os meteorologistas, portanto, combinam derivadas informações com extrapolações de dados observacionais, incluindo radar e imagens de satélite.

A.5 Inclusão dos elementos meteorológicos na previsão

A.5.1 Características de alguns parâmetros climáticos

Estado do céu

É evidente que ao se falar do estado do céu, faz-se referência, preferencialmente, às nuvens. De acordo com (ÈVE, 2015), especialista em meteorologia canadense, *Uma nuvem é uma massa que consiste em gotas muito finas de água; no entanto, a água encontra-se por vezes na forma de pequenos cristais de gelo ou flocos de neve.* Abaixo, alguns exemplos de nuvens na vida quotidiana:

1. O ar que sai da boca e do nariz no tempo frio. Isto é devido à umidade (vapor de água) contida no ar dos pulmões; em contato com o ar frio do exterior o vapor de água condensa-se em gotas finas.
2. O vapor de água que escapa de uma chaleira. O ar mais frio do ambiente produz uma condensação que resulta em uma pequena nuvem.

Todas as nuvens são constituídas por água no estado gasoso, líquido ou sólido. A quantidade de água necessária para a formação de uma nuvem é relativamente baixa. Uma nuvem do

tamanho de um campo de futebol contém menos água do que uma banheira cheia. O seu peso é equivalente à de um homem adulto.

Há 200 anos, as nuvens não tinham nomes específicos. Em 1803 Luke Howard inventou o sistema de classificação ainda em uso hoje.

O termo nuvem é baseada em dois princípios simples: a altitude de sua base e sua forma. Primeiro, as nuvens são divididas em quatro grupos. Os três primeiros, de acordo com a altura média da base; o quarto grupo reflete a forte extensão vertical de algumas nuvens que podem se encontrar em mais do que um andar de cada vez.

Define-se três tipos de nuvens (ÈVE, 2015):

1. Os Cirrus, ou ciclo em Latim, têm aspecto de filamentos brancos e delicados. Eles são chamados Cirrus (Ci), Cirrostatus (Cs) e Cirrocumulus(Cc).
   Quanto à previsão, para o (Ci) tem-se tempo bom se os ventos são de setores oeste, noroeste ou norte. Por outro lado, se os ventos são persistentes do nordeste, leste ou sul, haverá previsões de precipitação em 20-30 horas.
   Já no caso do (Cs), prováveis precipitações em 15-25 horas, se os ventos forem persistentes do nordeste, sudeste, ou se os ventos são do sudeste ao sul. Qualquer outra direção do vento fará com que um céu seja nublado.
   O (Cs) por sua vez está presente logo no início das manhãs de verão, esta nuvem muitas vezes leva a trovoadas à tarde. Precipitação provável em 15-20 horas, se os ventos são do nordeste para o sul. Qualquer outra direção do vento fará com que um céu nublado.

2. Os Status, que significa estendido ou ampliado, são cinzas e formam uma camada baixa e uniforme. Estes são o Status (St), Nimbostratus (Nb) e Stratocumulus (SC).
   Quanto á previsão, com relação ao (St), ventos do nordeste ao sul podem causar chuvas fortes. Qualquer outra direção do vento provocará um céu nublado.
   Já o (Sc) apresenta ameaça imediata de mau tempo, algumas gotas de chuvas fortes. Precedendo uma frente fria, que irá causar tempestades e ventos fortes. Se os Scs são vermelho indiano ao pôr do sol e os ventos são do nordeste para o sul, haverá precipitação em 12-20 horas. Qualquer outra direção do vento fará com que o céu seja nublado. Se, ao invés disso, as nuvens são dourada, rosa, laranja, não haverá precipitação dentro de 20-24 horas.

3. Os Cumulus, ou pilha em Latim, tem uma grande extensão vertical e podem afetar vários níveis. Estes são o Cumulus (Cu) e cumulonimbus (Cb). Quanto ao (Cu), se não crescem na vertical, eles anunciam o bom tempo. Se eles crescem verticalmente a partir do

sudoeste para noroeste, haverá probabilidade de precipitação em 5-10 horas com ventos fortes, tempestades ou simplesmente linhas de instabilidade.

No que tange ao (Cb), em geral, as precipitações chegarão, provavelmente, brevemente do sudoeste, do oeste ou do norte. As nuvens distantes muitas vezes têm a forma de uma bigorna com um topo desfiados como Ci.

Enquanto o (Ns) apresenta nuvem de chuva ou neve. A precipitação será de longo prazo, se os ventos são do nordeste para o sul, e de curta duração, se o sudoeste, o oeste ou norte.

Precipitações

Qualquer precipitação requer a condensação de vapor de água. Mas quando as gotículas de água das nuvens têm crescidas o suficiente, elas tornam-se pesadas demais para serem suportadas na nuvem; Então elas começam a cair no terra. Três fatores determinam a forma final em que ela se apresenta: são as correntes de ar, temperatura e umidade.

Existem dois tipos de precipitações (ÈVE, 2015);

- precipitação estratiforme: cobrindo uma grande área, que dura muito tempo, mas de baixa intensidade, o que ocorre em áreas de vales de baixa pressão, que está associada com tipos de nuvens ”stratus”;
- precipitação convectiva: cobrindo áreas pequenas, que não dura mas é intensa, que é muito localizada e produzida pela instabilidade convectiva do ar, e, finalmente, que está associada com tipos nuvem ”cumulus”.

A precipitação pode cair sob três formas:

- Precipitação líquida: chuva e garoa;
- Precipitação congelada: chuva congelada e garoa congelada;
- Precipitação sólida: neve, grãos de neve, cristais de gelo, granizo e saraiva.

As previsões meteorológicas desempenham um papel importante em decisões de cancelar ou não uma atividade. Neste fito, (OLIVIER, 2015) apresenta as probabilidades de precipitação, a fim de medir o risco de chuvas, fator motor de perturbações de várias atividades, inclusive do *tráfico ferroviário*. A seguir são apresentados os métodos por trás das estatísticas.

O termo comumente utilizado para **probabilidade de precipitação** é **PDP**, e é expresso em **porcentagem** nas previsões meteorológicas. Esta estatística corresponde à probabilidade que

uma **quantidade mensurável** (**ao menos 0,2 mm**) de precipitações de produz na região em que se encontra.

O percentual representa as **probabilidades** de ter
textbfchuva (ou **neve** no inverno) e não as **quantidades**, afirma Olivier (2015). Assim que os modelos indicam 0,2 mm de precipitação, existe um impacto sobre a percentagem. A tabela A.1 mostra diferentes precipitações e ações a serem tomadas.

%	Caráter das precipitações	Ação a tomar	Duração
20 %	Muito isoladas	nenhuma	–
30 %	Isoladas	Esconder-se	Breve
40 %	Dispersas	Guarda-chuva	Breve

Tabela A.1: Probabilidades de precipitações 20 % - 40 %

Olivier (2015) afirma que geralmente, quando as P.D.P são 10%, 20% ou 30%, a maioria das pessoas sabe que o risco de receber precipitações é baixa. Muitas pessoas, no entanto, estão relutantes em manter ou cancelar uma atividade, quando previsão mostra 40% de probabilidade. Deve-se, portanto, lembrar que **40% de probabilidade também significa que há 60% de probabilidade de não receber precipitações (chuvas)**.

A 60%, pode-se esperar um pouco de chuva, mas ainda pode-se ter uma atividade. A partir de 70%, é melhor ter um plano B, explica Didier Robert Lacroix, meteorologista do MéteoMédia [1].

A *80%*, *90%* e *100%*, a **chuva contínua** é praticamente garantida durante o período indicado na sua previsão (manhã, tarde, final da tarde, noite). Se esta probabilidade estender-se ao longo de várias horas ou dias, fala-se então de um regime de chuva consistente.

%	Caráter das precipitações	Ação a tomar	Duração
60 %	Frequentes	Plano de substituição (plano B)	durável
80 % e +	Chuva ou neve	Cancelar	Contínua

Tabela A.2: Probabilidades de precipitações 60 % -80 % e +(OLIVIER, 2015)

Porquê não **50 %**?
Os especialistas afirmam que não utilizam a estatística **50 %** nas previsões por uma razão

[1] http://www.meteomedia.com/

precisa. ”De uma perspectiva científica, é possível, mas os meteorologistas, por exemplo, de MétéoMédia tomar uma posição que usando 40% ou 60%. Para eles, o uso de 50% em uma predição é muito arriscado. É como dizer uma chance em duas”, acrescenta Robert Didier Lacroix.

Ventos

O vento nasce como resultado de diferenças de temperaturas e pressão. A pressão sobre a terra é alta se o ar pesado e frio desce e baixa se o ar quente e leve sobe (ÈVE, 2015). O ar quente (mais leve) sobe naturalmente nas camadas superiores da atmosfera, a fim de criar uma área de baixa pressão ”L”. No entanto, o ar quente atrai para si outra massa de ar: a zona de alta pressão. É esta diferença de pressão entre as duas massas de ar que origina o vento. Só porque o ar contido na ”H”naturalmente tende a entrar na ”L”mais próxima. O ar que se move, é o vento!

A principal responsável por esse fenômeno é o sol. Ele aquece os mares e continentes, mas não no mesmo ritmo. Uma vez aquecido, este último por sua vez, aquece as massas de ar que as saliência. O ar, então, começa a mover-se, uma vez que aumenta em volume quando aquecido. Torna-se mais leve e sobe. Um exemplo para a escala terrestre deste deslocamento das massas de ar é que o ar quente sobe a partir do equador e as massas do ar mais frias dos polos substituí-lo (Estes são os ventos alísios).

Classifica-se o vento em várias escalas. Deve-se notar que as velocidades de terra e mar não são equivalentes. A velocidade do vento é medida em quilômetros por hora ou em nós (cerca de 1,85 km/h por nó). Em previsão náutica, duas unidades são muitas vezes usadas. Primeiro, numa escala de 0-12 que chama-se de *”escala Beaufort”*.

- 0 ($<$2 km/h): tudo é calmo, fumaça sobe em linha reta, nem uma folha se move.
- 4 (21 a 29 km/h): É uma brisa agradável. A poeira é levantada e galhos se agitam.
- 8 (entre 63 e 75 km/h): É uma rajada. Pequenos galhos se quebram; torna-se difícil para avançar contra o vento.
- 12 ($>$ 117 km/h): Este é um vento de furacão; árvores desenraizadas, telhados arrancados, casas derrubadas. Muito raro, o caos generalizado.

Para as previsões públicas (terrestres), usa-se os seguintes termos (ÈVE, 2015):

- Leve (0 a 9 km/h).
- Moderado (10 a 40 km/k).
- Forte (41 a 60 km/h).
- Muito forte (61 a 90 km/h)
- Muito forte / força de tempestade (mais de 91 km/h)
- Força de furacão (mais de 115 km/h)

E essas unidades associadas aos fenômenos mais ou menos violentos, tornam-se:

- Ventos de menos de 35 km/h: perturbações tropicais.
- Ventos de 36 km/h a 60 km/h: depressões tropicais.
- Ventos de 61 km/h a 114 km/h: tempestades tropicais.
- Ventos de mais de 115 km/h furações.

Um vento de 160 kmh pode permitir que uma pessoa se incline na direção do vento, estique as pernas e toque o chão com as mãos, sem cair! ...?

Temperatura

A distribuição da temperatura média do ar é determinada essencialmente por dois factores: a distância do mar e altitude.

Além destes fatores determinantes, há também factores estritamente meteorológicos e outros factores geográficos (origem e frequência local das diferentes massas de ar, a subsidência, a radiação infravermelha emitida pela superfície do substrato e da composição de solo, a influência do relevo local), que também influenciam a distribuição espacial da temperatura (IRM, 2015).

A temperatura sofre igualmente de uma variação anual que segue a variação de radiação incidente e função da posição geográfica com relação aos oceanos e continentes.

De uma forma proporcionalmente mais sensível, observa-se igualmente, uma variação diurna da temperatura do ar resultante da radiação solar e da radiação do solo. A radiação solar é máxima ao meio dia local, mas tendo em vista que a radiação do solo culmina por volta das 14h

local, a temperatura média do ar apresenta seu máximo cerca de 2 horas depois da culminação do sol.

A temperatura mínima se observa ao momento em que a energia solar torna-se inferior á energia terrestre. Este fenômeno se observa cerca de 1 hora depois do nascer do sol.

Conclusão

Esta seção apresentou as regras gerais para a integração de elementos meteorológicos na previsão. A informação necessária para se realizar uma previsão é composta dos seguintes elementos meteorológicos:

- *Temperatura*: valor que representa a temperatura prevista para a próxima hora.
- *Tempo*: imagem e resumo que indicam as condições meteorológicas mais prováveis para a próxima hora.
- *Eventuais precipitações (EdP)*: categoria (baixa, média ou alta), indicando a probabilidade de precipitação prevista durante a hora seguinte. Categorias de eventualidades de precipitações são dadas conforme abaixo e são diretamente ligadas com a possibilidade de precipitação de previsões escritas.
  - Nula : 0 %
  - Baixa : 40 % e menos
  - Média : 60 ou 70 %
  - Elevada : 80 % e mais
- Vento : Velocidade e direção do vento prevista para a próxima hora. Em tempo tempestuoso, estas serão mencionadas na previsão. A abreviatura ”VR”indica que a direção do vento vai variar durante o período em questão.

Ao dispor-se desses elementos, suas características e variabilidades, poder-se-á, então estimar, com base nas probabilidades, as próximas *prováveis* previsões.

A.6 Influência dos fatores climáticos no transporte ferroviário

As condições climáticas adversas podem ter consequências graves para o tráfego ferroviário. Faz-se necessário estar em constante contato com os serviços meteorológicos a fim de anteci-

par esses riscos meteorológicos. Além disso, deve-se dispor de sensores ou agentes de infra-estrutura para monitorar constantemente as instalações sensíveis (sinais mecânicos, comutadores, etc.).

Todos os meios de transporte são suscetíveis a diversos graus, às condições meteorológicas e ao clima afirma o Ministério de Recursos Naturais Canadense (RNC: *Ressources naturelles Canada*) [2]. De acordo com a Sociedade Nacional de Ferrovia francesa [3] (SNCF: *Société Nationale de Chemin de fer*), existem fatores que podem levemente ou gravemente influenciar ou perturbar a circulação dos trens e, em seguida, levá-los a não-aderência do planejamento inicial. Entre eles são a os parâmetros climáticos e não climáticos a seguir.

A.6.1 Parâmetros climáticos

As condições meteorológicas enfraquecem as infra-estruturas ferroviárias, assim como os equipamentos. Com a aproximação do inverno, frio, neve, gelo, geada, vento ou chuva pesada, todas estas condições meteorológicas são de sérias consequências sobre a circulação de trens. Perturbações que podem variar de pequenos atrasos a uma interrupção total do tráfego para reparar as vias e catenária. Os parâmetros aqui apresentados são as principais causas de tais perturbações.

- *Temperatura*

  A deformação das malhas é um problema grave de segurança e redução do desempenho. Quanto mais alta a temperatura mais deformação há na malha. O trem pode sofrer restrições de velocidade devido às temperaturas elevadas, afirmam os especialista da **SNCF**.

  Vários limites de temperatura podem ser estabelecidos, permitindo que se avalie o estado da malha remotamente, em tempo real, e reagir para cada variação de acordo com os procedimentos predefinidos. Por exemplo, quando a temperatura da malha na faixa de 49-53 ° C, os trens podem ser submetidos a restrições de velocidade devido às altas temperaturas.

  No inverno, também pode-se ter avisos quando a temperatura ferroviária é abaixo de 0 ° C (ou outras variações definidas pelo usuário) para ajudar a agendar a manutenção invernal.

- *Precipitação*

[2] http://www.rncan.gc.ca/environnement/ressources/publications/impacts-adaptation/rapports/evaluations/2004/ch8/10218#ar acessado em 15-08-2015

[3] http://www.sncf.com/fr/presse/article

Deve-se coletar as informação a respeito da precipitação, sua intensidade e suas quantidades acumuladas. Estes dados permitirão que os especialistas sejam avisado com antecedência dos possíveis impactos no caso de chuvas pesadas, por exemplo.

- *Chuva*

  Se as chuvas são violentas, elas podem causar deslizamentos de terra que por sua vez provoca poluição e obstáculos na malha. Estes fluxos de lama, portanto, interrompem o tráfego de trens.

  A duração das reparações é difícil de se avaliar se a chuva continua caindo, os reparos podem ser prorrogados e o tráfego pode ser interrompido em qualquer dos eixos para se obter uma segurança máxima das vias.

  As chuvas podem ser fortes, moderadas e fracas. Dependendo da intensidade, podem levar à perda de velocidade do trem, no pior caso, à interrupção.

- *Vento*

  Um vento forte acompanhado de chuva pesada provoca a presença de detritos ou resíduos nas plataformas e vias que podem interromper o tráfego enquanto se retiram os escombros.

  O vento pode ser fraco, moderado ou violento (podendo virar o trem). Neste caso, a circulação pode até ser interrompida.

- *Neve*

  No inverno, o excesso de neve pode causar projeções de gelo em trens. Este fenômeno provoca regularmente danos materiais e impacto sobre o tráfego. Por exemplo, a neve pode reduzir a velocidade do trem até 50 %.

  No inverno, fortes nevascas podem cobrir as vias. Quando um trem passa, a neve levantada devido à alta velocidade se acumula abaixo do trem. Com o frio, a neve endurece e se transforma em um bloco denso de gelo. Quando dois trens de alta velocidade se cruzam, o apelo do ar causado pelo cruzamento retira os blocos de neve. Levantados, eles agem como projéteis que danificam janelas e equipamentos ferroviários.

  Este fenômeno é observado principalmente em linhas de alta velocidade, com impactos medidos a 600 km/h. De acordo com a velocidade de projeção do gelo, os danos podem variar desde a fissura de uma janela à deterioração de elementos essenciais que levam à rescisão ou parada do trem, afirma a **SNCF**. Dentre esses elementos: os sensores abaixo do trem permitem a transmissão de informações para a cabina do condutor.

A.6.2 Parâmetros não climáticos

- *Folha mortas ou árvore caída*

  Durante os tempos severos as árvores e folhas podem cair sobre as vias, criando obstáculos na malha. Isso cria perturbações na tração do trem. Uma árvore sobre as vias é muito difícil de se retirar porque não se deve danificar as instalações ao redor de vias. A folhas mortas na malha pode afetar a tração do trem até 30 % abaixo do normal.

A.6.3 Conclusões

Deve haver instalações e estruturas de todos os tipos para garantir a circulação de pessoas e transporte de mercadorias ou mineiros - estradas, ferrovias, pistas, terminais marítimos, canais e pontes são exemplos. As condições climáticas e meteorológicas influenciam o planejamento, concepção, construção, manutenção e operação dessas instalações ao longo da sua vida útil. As redes de transporte atuais podem bem ser robustas, todavia, partes de seus componentes dificilmente resistirão a certas condições meteorológicas.

O propósito deste capítulo foi apresentar os fenômenos climáticos, seus principais parâmetros, e seus impactos na malha ferroviária ou na circulação dos trens, para permitir as medidas de prevenção e possíveis previsões sobre o comportamento do tráfego na presença dos fatores apresentados. Esta parte é importante, pois irá compor o módulo de previsão que é uma das propostas deste trabalho.