



Introdução à Mineração de Dados

Eduardo Ogasawara

eduardo.ogasawara@cefet-rj.br
<https://eic.cefet-rj.br/~eogasawara>

Por que Mineração de Dados?

O Cenário do Big Data

Vivemos em uma era de **crescimento explosivo de dados**, passando de terabytes para petabytes em poucos anos. Ferramentas automatizadas de coleta, sistemas de banco de dados e a Web transformaram a maneira como armazenamos informações.



Principais Fontes de Dados

- **Negócios:** Web, e-commerce, transações financeiras
- **Ciência:** Sensores, astronomia, bioinformática, simulações
- **Sociedade:** Notícias, fotos, vídeos, dados abertos, IoT

"Estamos nos afogando em dados, mas morrendo de sede por conhecimento!"

A necessidade é a mãe da invenção — a mineração de dados surgiu como resposta à análise automatizada de conjuntos massivos de dados.

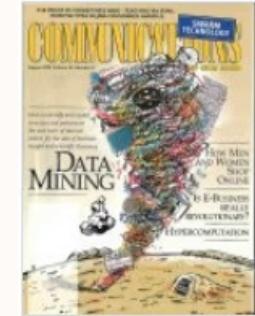
O que é Mineração de Dados?

Descoberta de Conhecimento em Dados

A mineração de dados (ou *knowledge discovery from data*) é a **extração de padrões interessantes ou conhecimento** a partir de quantidades massivas de dados.

Características dos Padrões

- **Não triviais:** Não óbvios à primeira vista
- **Implícitos:** Ocultos nos dados
- **Previvamente desconhecidos:** Não armazenados explicitamente
- **Potencialmente úteis:** Acionáveis para tomada de decisão



Nomes Alternativos

- Knowledge Discovery in Databases (KDD)
- Knowledge Extraction
- Business Intelligence
- Data Analysis

O Processo KDD

O processo de Descoberta de Conhecimento em Bases de Dados (KDD) representa uma visão clássica de sistemas de banco de dados. A mineração de dados desempenha um **papel essencial** neste processo, mas é apenas uma das etapas de uma jornada maior.

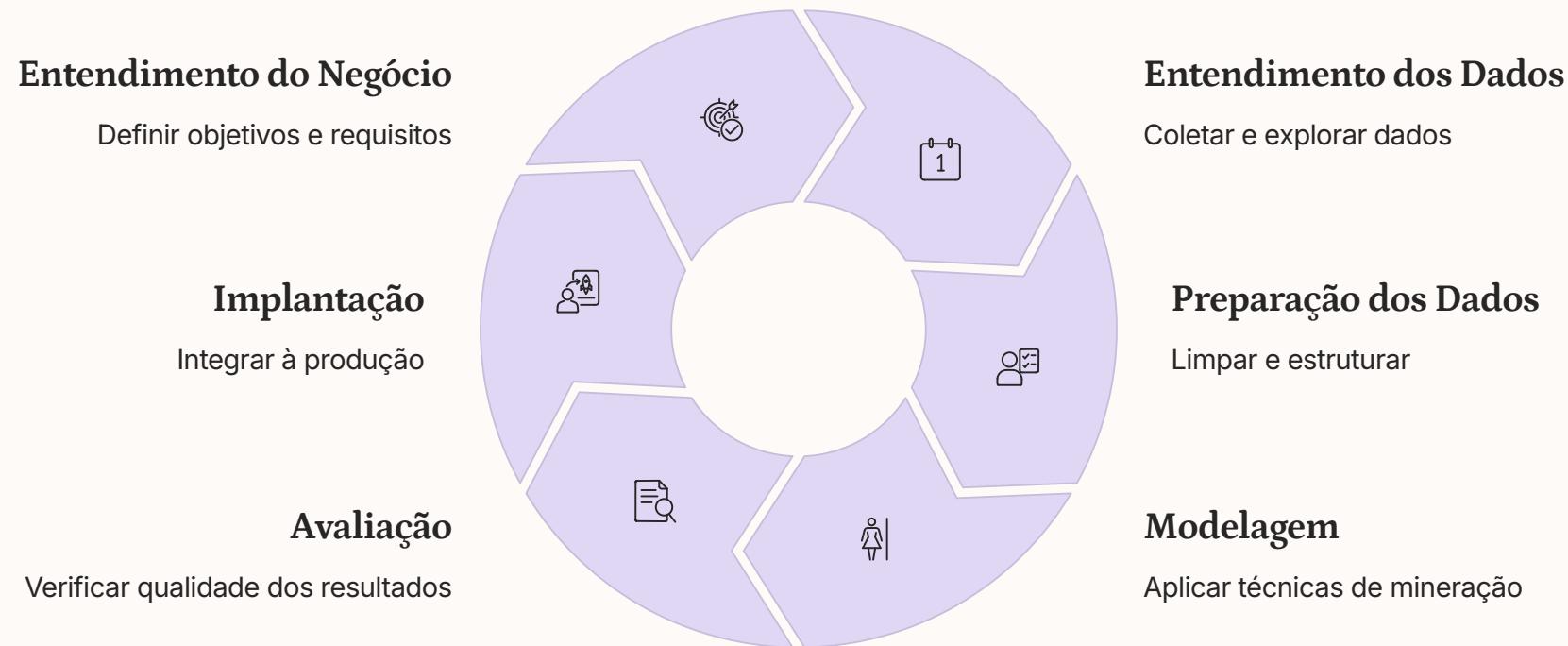


Cada etapa é crucial para garantir que o conhecimento extraído seja válido, confiável e útil para a tomada de decisões estratégicas.

Visão Geral do CRISP-DM

Cross-Industry Standard Process for Data Mining é uma metodologia amplamente utilizada e padronizada pela indústria para projetos estruturados de mineração de dados.

Diferentemente de abordagens lineares, o CRISP-DM é **cíclico e iterativo**, reconhecendo que projetos de dados raramente seguem um caminho direto do início ao fim.



KDD vs. CRISP-DM: Comparação

Embora o KDD e o CRISP-DM compartilhem objetivos semelhantes, suas abordagens e origens diferem significativamente. O **KDD** é o processo conceitual mais amplo de descoberta de conhecimento em dados, enquanto o **CRISP-DM** fornece um guia prático e cíclico para gerenciar projetos de mineração de dados em ambientes reais.

KDD	CRISP-DM
<ul style="list-style-type: none">• Origem acadêmica• Processo linear• Ênfase na extração de conhecimento• Abordagem mais teórica• Foco em algoritmos e técnicas	<ul style="list-style-type: none">• Origem industrial• Processo cíclico• Ênfase na resolução de problemas de negócios• Abordagem mais prática• Foco em entregas e resultados

Característica	KDD	CRISP-DM
Origem	Academia	Indústria
Processo	Linear	Cíclico
Ênfase	Extração de Conhecimento	Solução de Problemas de Negócio
Aplicação	Pesquisa e Desenvolvimento	Projetos Empresariais

Enquanto o KDD é mais acadêmico e linear, o CRISP-DM é iterativo e fundamentado em aplicações do mundo real.

Tudo é "Mineração de Dados"?

Não!

É importante distinguir a mineração de dados de outras atividades relacionadas ao processamento de informações. Embora essas áreas se sobreponham na prática, cada uma possui características e objetivos distintos.

Consultas (Queries)

Recuperação de dados específicos usando linguagens como SQL. As consultas retornam informações conhecidas e explícitas, não descobrem padrões ocultos.

Relatórios (Reporting)

Apresentação estruturada de dados agregados e métricas. Os relatórios resumem informações existentes, mas não identificam novos insights automaticamente.

Estatística Tradicional

Análise de dados usando métodos estatísticos clássicos. Focada em testes de hipóteses pré-definidas, enquanto mineração de dados busca padrões sem hipóteses iniciais.

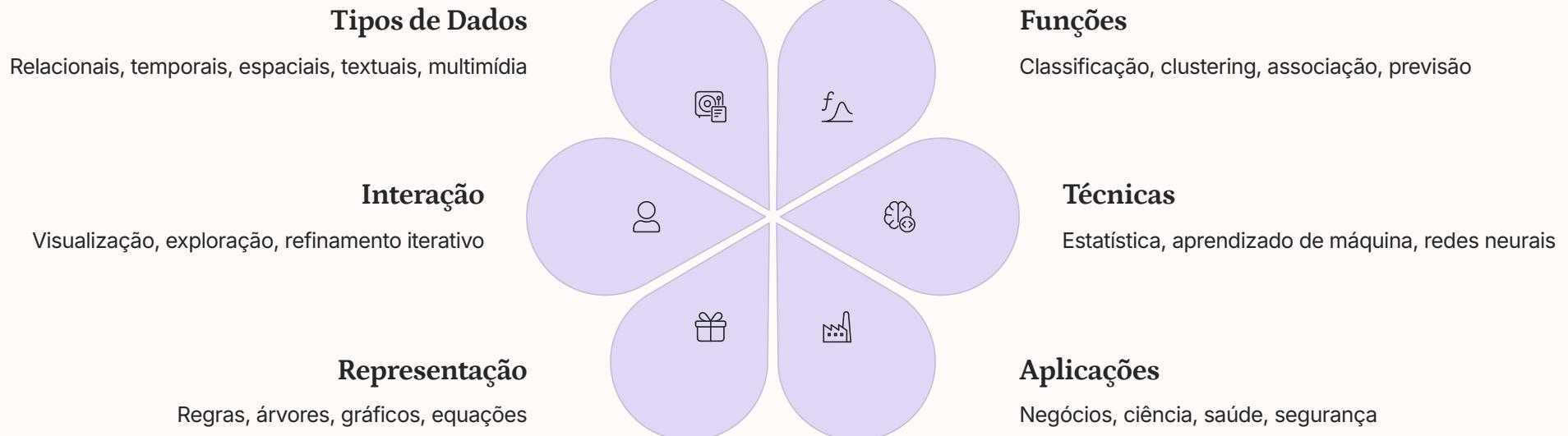
Mineração de Dados

Descoberta automática de padrões não triviais, implícitos e potencialmente úteis em grandes volumes de dados, usando algoritmos avançados de aprendizado.

- A mineração de dados se distingue por sua capacidade de descobrir conhecimento novo e açãoável sem orientação explícita sobre o que procurar.

Visão Multidimensional da Mineração de Dados

A mineração de dados pode ser compreendida através de múltiplas dimensões que definem seu escopo e aplicabilidade. Esta visão multidimensional ajuda a categorizar e entender a amplitude da área.



Dimensões Principais

- Modelos de Dados:** Que tipo de dados estamos minerando?
- Conhecimento a Descobrir:** Que tipo de padrões buscamos?
- Técnicas Utilizadas:** Quais algoritmos aplicamos?
- Aplicações:** Onde utilizamos os resultados?

Esta perspectiva multidimensional ressalta que a mineração de dados não é uma técnica única, mas um **ecossistema complexo** de métodos, aplicações e tipos de dados inter-relacionados.

Mineração em Dados Orientados a Bancos de Dados

A mineração de dados começou com estruturas tradicionais de banco de dados e continua sendo fundamental em muitas aplicações empresariais. Esses conjuntos de dados são caracterizados por sua **estrutura bem definida** e relações claras entre entidades.



Bancos de Dados Relacionais

Estruturas tabulares com esquemas definidos, chaves primárias e estrangeiras. Permitem consultas SQL e suportam transações ACID.



Bancos Transacionais

Sistemas OLTP focados em processamento rápido de transações. Capturam eventos de negócio em tempo real com alta concorrência.



Bancos Heterogêneos

Integração de múltiplos sistemas de banco de dados diferentes. Requerem técnicas especiais de integração e limpeza.

Esses tipos de dados formam a base de muitas aplicações empresariais, desde sistemas de gestão até plataformas de e-commerce, e continuam sendo alvos principais para técnicas de mineração de dados.



Data Warehouses

Repositórios centralizados otimizados para análise e relatórios. Integram dados de múltiplas fontes com esquemas dimensionais (estrela/floco de neve).



Bancos Objeto-Relacionais

Combinam características relacionais com orientação a objetos. Suportam tipos de dados complexos e herança.



Sistemas Legados

Bancos de dados antigos ainda em operação. Apresentam desafios de compatibilidade e modernização.

Mineração em Conjuntos de Dados Avançados

À medida que a tecnologia evolui, novos tipos de dados emergem, exigindo técnicas especializadas de mineração. Esses **dados avançados** apresentam características únicas que desafiam métodos tradicionais.



Dados de Streaming e Sensores

Fluxos contínuos de dados em tempo real que requerem processamento imediato



Séries Temporais

Dados ordenados no tempo com dependências sequenciais e tendências



Grafos e Redes Sociais

Estruturas complexas de nós e conexões representando relacionamentos



Dados Espaciais e Espaço-Temporais

Informações georreferenciadas com dimensões de localização e tempo



Bancos Multimídia

Imagens, vídeos, áudio requerendo análise de conteúdo não estruturado



Bases Textuais

Documentos, artigos, posts processados por text mining e NLP



World Wide Web

Dados semi-estruturados com hiperlinks, tags e metadados



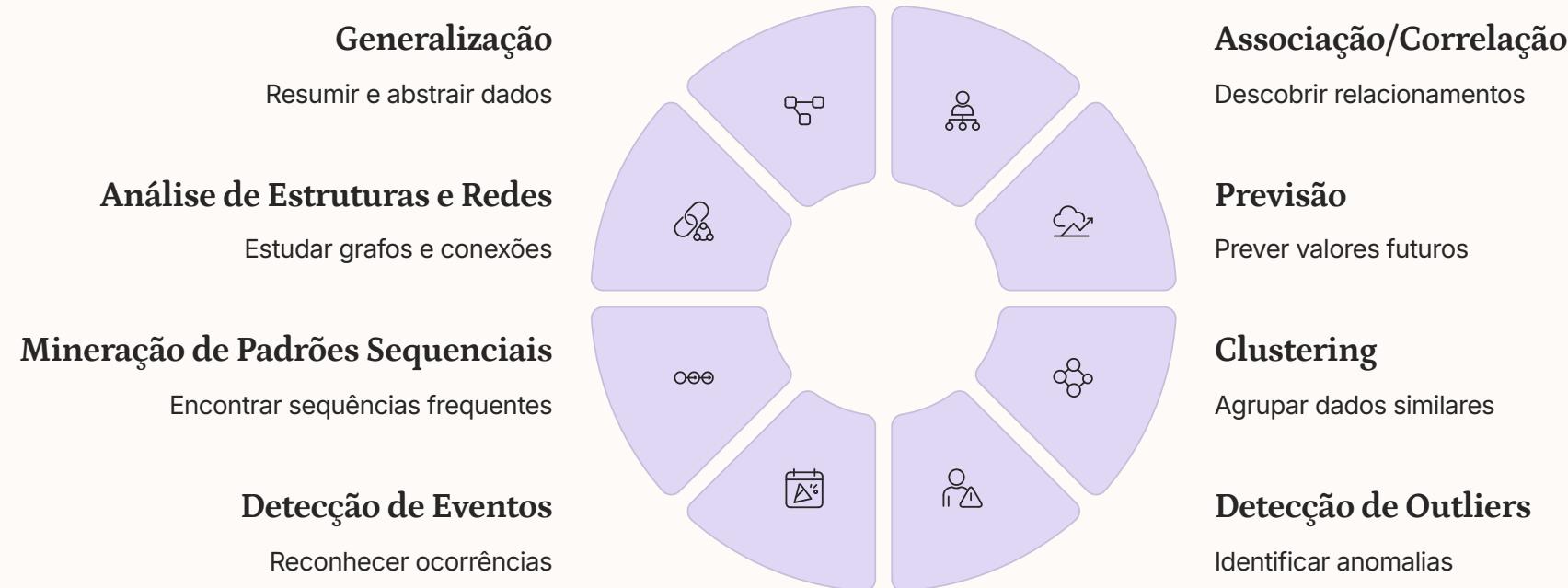
Sequências Biológicas

DNA, proteínas e dados genômicos com padrões complexos

Cada tipo de dado avançado requer algoritmos especializados, técnicas de pré-processamento únicas e considerações especiais de escalabilidade. A mineração eficaz desses dados é crucial para aplicações modernas em IoT, bioinformática, análise de redes sociais e muito mais.

Funções Centrais da Mineração de Dados

A mineração de dados engloba um conjunto fundamental de funções que representam diferentes formas de extrair conhecimento dos dados. Cada função aborda um tipo específico de problema analítico e utiliza técnicas distintas.



- Estas oito funções centrais formam a base de praticamente todas as aplicações de mineração de dados, desde análise de cestas de compras até detecção de fraudes financeiras.

Compreender quando e como aplicar cada função é essencial para o sucesso de projetos de mineração de dados. Nas próximas seções, exploraremos cada uma dessas funções em detalhes.

Função de Mineração de Dados: Generalização

Resumindo e Abstraindo Dados

A **generalização** é o processo de abstrair dados detalhados em representações de nível superior que facilitam a compreensão e análise. Esta função é fundamental para data warehousing e análise OLAP.

Componentes Principais

- **Integração de Informações:** Consolidação de múltiplas fontes de dados
- **Data Warehouse:** Repositórios otimizados para consultas analíticas
- **Limpeza e Transformação:** Preparação e padronização dos dados
- **Modelo Multidimensional:** Organização em cubos de dados com dimensões e medidas

01

Tecnologia de Cubo de Dados

Métodos escaláveis para computar e materializar agregações multidimensionais, permitindo análises rápidas em diferentes níveis de granularidade.

02

OLAP (Online Analytical Processing)

Ferramentas interativas para exploração multidimensional de dados, incluindo operações de drill-down, roll-up, slice e dice.

03

Descrição Conceitual Multidimensional

Caracterização e discriminação de dados. Por exemplo, generalizar medidas de pluviometria em categorias como "região seca" vs. "região úmida".

- A generalização permite que analistas vejam o "panorama geral" antes de mergulhar em detalhes específicos, facilitando a descoberta de tendências e padrões macro.

Função de Mineração: Associação e Correlação

Descobrindo Relacionamentos em Dados

A análise de **associação e correlação** identifica relacionamentos interessantes entre itens em grandes conjuntos de dados transacionais.

Conceitos Fundamentais

- **Padrões Frequentes (Itemsets):** Conjuntos de itens que aparecem juntos com frequência
- **Regras de Associação:** Implicações do tipo "Se X, então Y"
- **Supor te:** Frequência de ocorrência do padrão
- **Confiança:** Probabilidade condicional da regra



Exemplo Clássico

Fraldas Cerveja [0.5%, 75%]

Esta regra indica que 0.5% das transações contêm ambos os itens (supor te), e 75% das compras de fraldas incluem cerveja (confiança).



Associação vs. Correlação vs. Causalidade

Itens fortemente associados nem sempre são correlacionados. É crucial distinguir entre correlação estatística e causalidade real.

Questões Principais

1. Como minerar tais padrões e identificar regras eficientemente em grandes datasets?
2. Como usar e rankear esses padrões para tomada de decisão?
3. Quais medidas além de supor te e confiança são relevantes?

Aplicações incluem análise de cestas de compras, recomendação de produtos, detecção de fraudes e análise de clickstream em websites.

Função de Mineração: Previsão

Classificação e Predição de Rótulos

A **previsão** envolve construir modelos baseados em exemplos de treinamento para prever valores futuros ou classificar novos dados. Esses modelos são *data-driven*, aprendendo padrões diretamente dos dados.

O processo típico inclui:

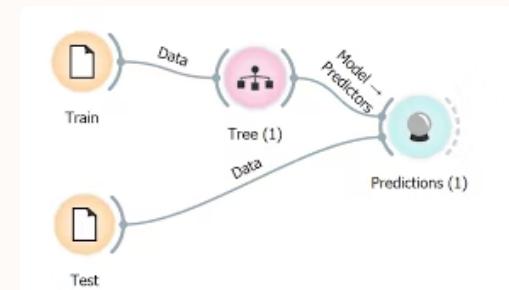
1. Coletar dados de treinamento rotulados
2. Extrair características relevantes
3. Treinar modelo usando algoritmo de aprendizado
4. Avaliar desempenho em dados de teste
5. Aplicar modelo para previsões em novos dados

Métodos Típicos

- Árvores de Decisão
- Naive Bayes
- Máquinas de Vetores de Suporte (SVM)
- Redes Neurais e Deep Learning
- Random Forest
- Regressão Linear/Logística
- CNNs, RNNs, Transformers

Aplicações Típicas

- **Científicas:** Previsão climática, análise genômica
- **Industriais:** Manutenção preditiva, controle de qualidade
- **Empresariais:** Churn prediction, detecção de fraudes
- **Governamentais:** Análise de crédito, segurança pública



Função de Mineração: Análise de Clusters

Aprendizado Não Supervisionado

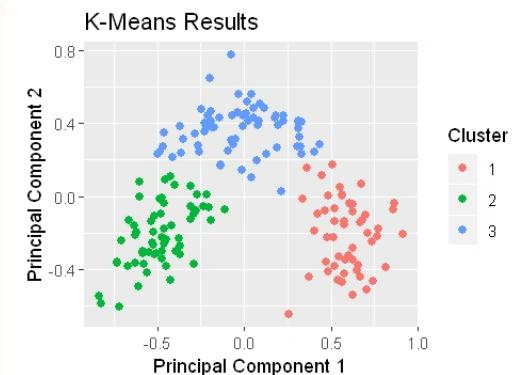
A **análise de clusters** é uma técnica de aprendizado não supervisionado onde os rótulos de classe são desconhecidos. O objetivo é agrupar dados para formar novas categorias baseadas em similaridade.

Princípio Fundamental

Maximizar a similaridade intra-classe e minimizar a similaridade inter-classes. Objetos dentro do mesmo cluster devem ser similares, enquanto objetos de clusters diferentes devem ser distintos.

Aplicações Práticas

- Segmentação de clientes
- Análise de padrões de compra
- Agrupamento de documentos
- Identificação de comunidades em redes sociais
- Análise de imagens e padrões visuais



Métodos de Particionamento

K-means, K-medoids: dividem dados em k grupos distintos

Métodos Hierárquicos

Agglomerative, Divisive: criam árvores de clusters

Métodos Baseados em Densidade

DBSCAN: identificam clusters de formas arbitrárias

Diferentemente da classificação, o clustering não requer dados de treinamento rotulados, tornando-o ideal para descoberta exploratória de padrões em dados desconhecidos.

Função de Mineração: Análise de Outliers

Um **outlier** é um objeto de dados que não está em conformidade com o comportamento geral dos dados. A questão fundamental é: trata-se de ruído ou exceção? O "lixo" de uma pessoa pode ser o "tesouro" de outra.

Abordagens de Detecção

Estatística: Baseada em distribuições probabilísticas e desvios padrão

Clustering: Pontos que não pertencem a nenhum cluster ou formam clusters muito pequenos

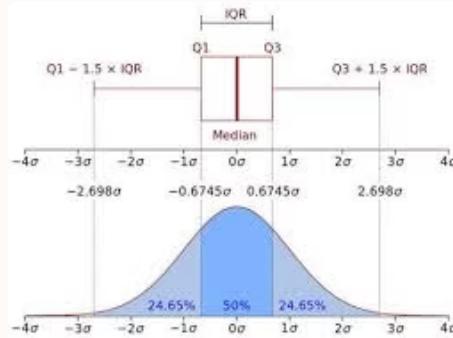
Análise de Regressão: Pontos que se desviam significativamente do modelo ajustado

Aplicações Valiosas

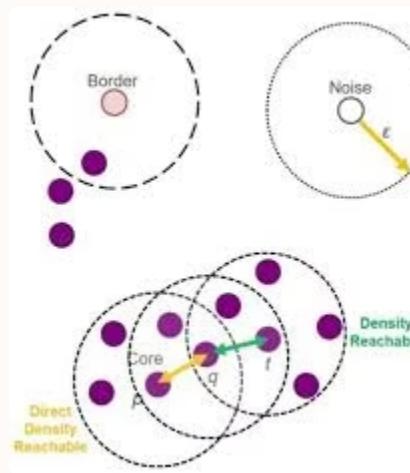
Detecção de Fraudes: Transações financeiras anômalas

Análise de Eventos Raros: Falhas em sistemas, doenças raras

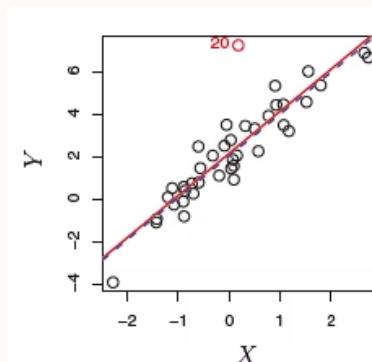
Controle de Qualidade: Produtos defeituosos em manufatura



Distribuição



Modelo



Densidade

A detecção de outliers é crucial em muitos domínios, pois anomalias frequentemente indicam eventos importantes, erros críticos ou oportunidades valiosas que merecem investigação especial.

Função de Mineração: Detecção de Eventos

A **detecção de eventos** em séries temporais identifica ocorrências significativas que desviam do comportamento esperado ou padrões recorrentes de interesse.

Anomalias

Padrões ou observações que não se conformam ao comportamento esperado. Podem ser baseados em distribuição, distância de um modelo ou volatilidade. Construídos a partir de processos distintos dos dados normais.

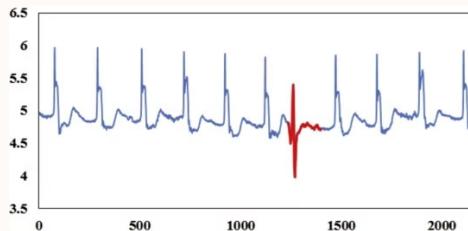
Motifs

Padrões (inicialmente desconhecidos) que ocorrem um número significativo de vezes em um dataset. Representam estruturas repetitivas que caracterizam o comportamento típico dos dados.

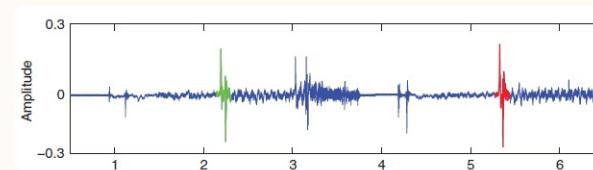
Change Points / Concept Drifts

Pontos ou intervalos de tempo que marcam mudanças significativas no comportamento do dataset. Separam diferentes estados no processo que gera os dados.

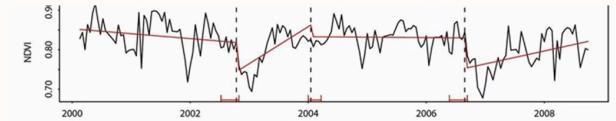
Anomalia



Motif



Change Point



A detecção de eventos é fundamental para monitoramento de sistemas, análise de mercados financeiros, previsão de falhas em equipamentos e compreensão de mudanças em processos complexos.

Função de Mineração: Padrões Sequenciais

Mineração de Padrões Sequenciais

A **mineração de padrões sequenciais** detecta e analisa subsequências frequentes de eventos, itens ou tokens que ocorrem em um espaço métrico ordenado.

Diferentemente da análise de associação, que ignora ordem, a mineração sequencial considera a **ordem temporal ou posicional** dos elementos.

Características Principais

- Elementos ordenados no tempo ou espaço
- Subsequências que ocorrem frequentemente
- Padrões podem ter gaps (não necessariamente consecutivos)
- Considera restrições temporais entre eventos

Análise de Clickstream

Padrões de navegação em websites



Análise de DNA

Subsequências genéticas frequentes

Comportamento de Compra

Sequências de produtos comprados ao longo do tempo

Padrões Sazonais

Eventos que se repetem em ciclos temporais

Exemplo: *Compra notebook → 2 semanas → Compra mouse → 1 mês → Compra impressora* representa um padrão sequencial de compras relacionadas a computadores.

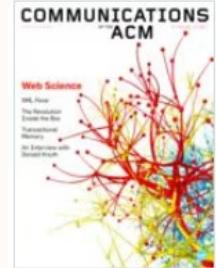
Função de Mineração: Análise de Estruturas e Redes

Mineração de Estruturas e Redes

A análise de estruturas e redes explora dados que possuem **relacionamentos complexos** representados como grafos, árvores ou outras estruturas interconectadas.

Mineração de Estruturas

Extrai padrões de dados estruturados como árvores XML, grafos moleculares e layouts de páginas web. Aplicações incluem bioinformática, química e análise de documentos.



Análise de Redes

Minera nós e arestas em redes como sociais, de citações e comunicação. Tarefas incluem detecção de comunidades, centralidade, influência e predição de links.



Detecção de Comunidades Web

Identificar grupos de páginas ou usuários com interesses similares e forte conectividade interna.



Proteínas Similares em Redes Biológicas

Encontrar proteínas com estruturas ou funções semelhantes baseadas em interações moleculares.



Descoberta de Influenciadores em Redes Sociais

Identificar usuários com maior alcance, engajamento e capacidade de disseminar informação.

Métricas de Centralidade

Degree, Betweenness, Closeness, PageRank

Algoritmos de Comunidade

Louvain, Label Propagation, Modularity

Análise de Difusão

Propagação de informação, epidemias, cascatas

Natureza Multidisciplinar da Mineração de Dados

A mineração de dados transcende fronteiras disciplinares tradicionais, atuando como ponto de convergência entre diversas áreas do conhecimento. Esta característica multidisciplinar não é acidental — ela reflete a complexidade inerente ao desafio de extrair conhecimento significativo de grandes volumes de dados.

Estatística, ciência da computação, matemática, aprendizado de máquina, visualização de dados e conhecimento de domínio específico se entrelaçam para formar o tecido fundamental desta disciplina. Cada área contribui com perspectivas únicas e ferramentas essenciais para o processo de descoberta de conhecimento.



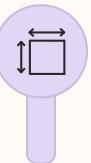
Por Que Tantas Disciplinas?

A convergência de múltiplas disciplinas na mineração de dados não é meramente uma escolha metodológica — é uma necessidade ditada pela natureza dos desafios enfrentados. Os dados modernos apresentam características que demandam expertise variada e abordagens sofisticadas.



Volume Tremendo de Dados

Os algoritmos devem ser escaláveis para lidar com big data, processando terabytes ou petabytes de informação de forma eficiente.



Alta Dimensionalidade

Dados com centenas ou milhares de atributos exigem técnicas especializadas para evitar a "maldição da dimensionalidade".



Complexidade Elevada

Fluxos de dados, dados de sensores, informações espaço-temporais, texto e multimídia apresentam estruturas únicas e desafiadoras.



Aplicações Sofisticadas

Desde diagnóstico médico até detecção de fraudes, as aplicações modernas demandam soluções cada vez mais refinadas.

Disponibilidade e Acesso aos Dados

Antes de embarcar em qualquer projeto de mineração de dados, questões fundamentais sobre acesso e uso dos dados devem ser cuidadosamente consideradas. Estas considerações não são apenas técnicas — elas envolvem aspectos legais, éticos e práticos que podem determinar a viabilidade de um projeto.

Questões Cruciais

- **Você tem acesso aos dados?** Disponibilidade física e permissões necessárias
- **Você pode usar os dados?** Aspectos legais e contratuais
- **Você pode publicar seus resultados?** Propriedade intelectual e confidencialidade
- **É big data ou small data?** Volume suficiente para justificar mineração de dados?

Small data pode ser extremamente valioso. A distinção entre big e small data não é apenas sobre volume, mas sobre a relevância e qualidade da informação disponível.

Referência: R. Kitchin e T.P. Lauriault, 2015, "Small data in the era of big data", GeoJournal, v. 80, n. 4, p. 463–475.

Principais Desafios na Mineração de Dados

O campo da mineração de dados enfrenta desafios significativos que vão muito além da simples aplicação de algoritmos. Estes desafios refletem tanto limitações técnicas quanto preocupações sociais crescentes sobre o uso responsável de dados.



Metodologia e Interação

- Mineração em espaços multidimensionais
- Descoberta de conhecimento em múltiplos níveis de abstração
- Incorporação de conhecimento de domínio
- Mineração interativa e exploratória



Desempenho e Escalabilidade

- Eficiência e escalabilidade de algoritmos
- Mineração paralela, distribuída e incremental
- Otimização para hardware moderno



Diversidade de Tipos de Dados

- Dados estruturados, semi-estruturados e não estruturados
- Dados complexos: grafos, sequências, dados espaciais
- Multimídia e fluxos de dados em tempo real



Mineração de Dados e Sociedade

- Propriedade e privacidade dos dados
- Segurança e sensibilidade das informações
- Questões éticas e impacto social
- Mineração responsável e justa
- Interpretabilidade dos modelos

Data Mining, Data Science e Analytics

A mineração de dados existe dentro de um ecossistema mais amplo de disciplinas relacionadas. Compreender as distinções e conexões entre data mining, data science e analytics é fundamental para posicionar adequadamente o trabalho de pesquisa e prática profissional.

Data Science: Visão Disciplinar

Data Science é um campo interdisciplinar emergente que sintetiza e se baseia em Estatística, Ciência da Computação, Comunicação, Gestão e Sociologia para estudar dados e seus ambientes (incluindo aspectos de domínio e contextuais) de forma a transformar dados em conhecimento e decisões.



Mineração de Dados como Núcleo

- **Data Mining:** Algoritmos e modelos para descoberta de padrões
- **Data Science:** Dados + contexto + tomada de decisão
- Áreas complementares, não concorrentes



Analytics Descriptiva

Usa estatísticas para descrever e sumarizar os dados analisados, revelando o que aconteceu.

Analytics Preditiva

Faz previsões sobre eventos futuros desconhecidos através de análise avançada, revelando por que algo aconteceu.

Analytics Prescritiva

Otimiza recomendações e sugere ações para tomada de decisões inteligentes.

□ **Data Analytics:** Teorias, tecnologias, ferramentas e processos que permitem a compreensão e descoberta a partir de dados — abrangendo todo o processo de descoberta de conhecimento: seleção, pré-processamento, transformação, mineração de dados e interpretação.

Business Intelligence: Aplicação de Data Analytics para suportar decisões de negócio.

Onde Publicar sua Pesquisa?

A escolha adequada de veículos de publicação é crucial para disseminar pesquisas e construir uma carreira acadêmica sólida. O ecossistema de publicações em mineração de dados é rico e diversificado, abrangendo conferências de prestígio e periódicos de alto impacto.

Data Mining, KDD e Data Science

Conferências: SIGKDD, IEEE-ICDM, SIAM-DM, PKDD, IEEE-DSAA

Periódicos: Data Mining and Knowledge Discovery, Statistical Analysis and Data Mining, ACM Transactions on Knowledge Discovery from Data

Sistemas de Banco de Dados

Conferências: SIGMOD, PODS, VLDB, IEEE-ICDE, EDBT, ICDT, SSDBM

Periódicos: IEEE-TKDE, VLDB Journal, Information Systems

IA e Aprendizado de Máquina

Conferências: AAAI, ML, IJCNN, IJCAI, NeurIPS

Periódicos: Machine Learning, Artificial Intelligence, Knowledge and Information Systems

Estatística

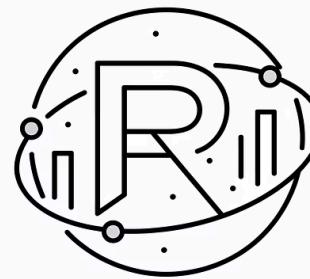
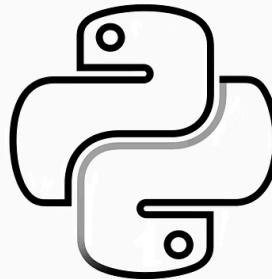
Periódicos: Journal of Applied Statistics, Annals of Data Science

Aplicações Específicas

Periódicos: Diversos periódicos especializados por domínio de aplicação

Linguagens em Alta para Data Mining

A escolha da linguagem de programação impacta significativamente a produtividade e as possibilidades de implementação em projetos de mineração de dados. Três ecossistemas dominam o cenário atual, cada um com suas forças particulares.



Python

Machine Learning Course

- Scikit-learn: biblioteca completa para ML clássico
- PyTorch: framework para deep learning
- TensorFlow: plataforma end-to-end para ML

Ecossistema rico, comunidade ativa, ideal para prototipagem rápida e produção.

R

Data Mining Course

- Miríade de pacotes especializados
- Forte orientação estatística
- Excelente para análise exploratória

Linguagem de escolha para estatísticos e cientistas de dados focados em análise.

Spark

Parallel and Distributed Computing

- MLlib: biblioteca de ML escalável
- Processamento distribuído
- Ideal para big data

Solução empresarial para processar volumes massivos de dados em clusters.

Ferramentas de Data Mining

Ferramentas de mineração de dados com interface gráfica democratizam o acesso a técnicas sofisticadas, permitindo que profissionais de diversas áreas apliquem algoritmos avançados sem programação extensiva. O ecossistema open source oferece opções robustas e profissionais.

RapidMiner

Plataforma visual para ciência de dados com interface intuitiva de arrastar e soltar. Suporta todo o ciclo de vida de ML.

rapidminer.com

Weka



Coleção clássica de algoritmos de ML para tarefas de mineração de dados. Interface Java robusta e extensível.

cs.waikato.ac.nz/ml/weka

Apache Mahout

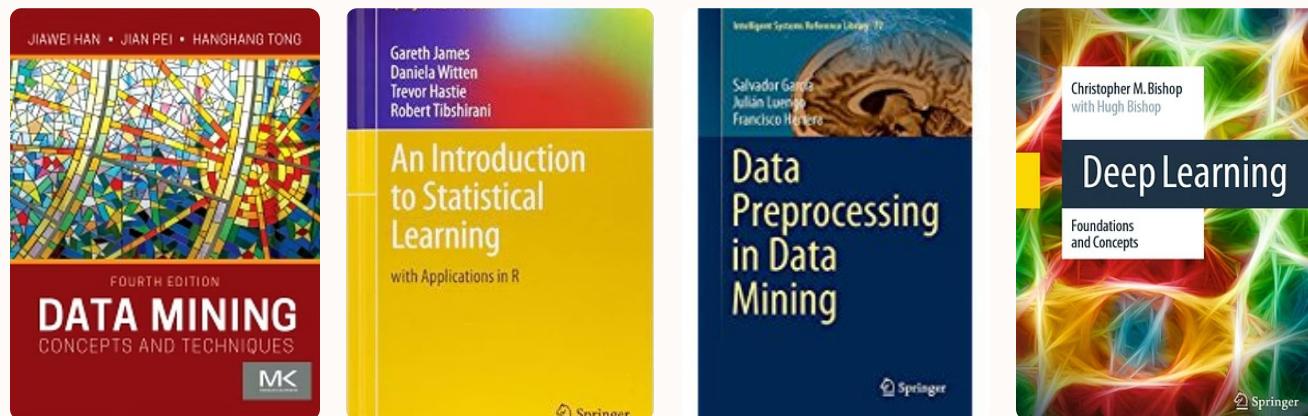
Biblioteca escalável de ML construída sobre Hadoop e Spark. Focada em recomendações, classificação e clustering.

mahout.apache.org



Referências Principais

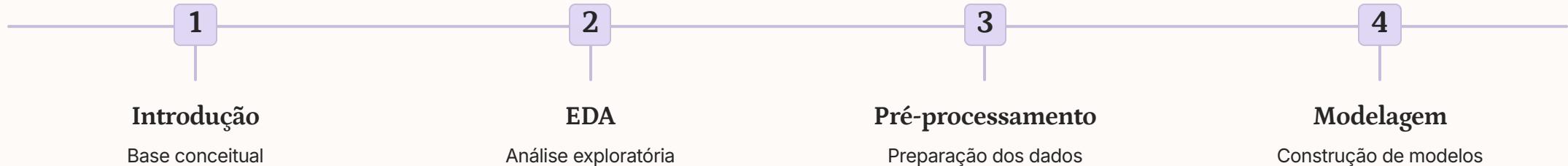
Esta seleção de referências representa os pilares fundamentais para o estudo aprofundado de mineração de dados, cobrindo desde conceitos básicos até técnicas avançadas e aplicações contemporâneas.



1. **J. Han, J. Pei, and H. Tong**, *Data Mining: Concepts and Techniques*, 4th edition. Cambridge, MA: Morgan Kaufmann, 2022.
2. **G. M. James, D. Witten, T. Hastie, and R. Tibshirani**, *An Introduction to Statistical Learning: With Applications in R*. Springer Nature, 2021.
3. **S. Garcia, J. Luengo, and F. Herrera**, *Data Preprocessing in Data Mining*. Springer, 2014.
4. **C. M. Bishop and H. Bishop**, *Deep Learning: Foundations and Concepts*. Springer Nature, 2023.

Roadmap

Material completo disponível em: <https://eic.cefet-rj.br/~eogasawara/jornada>



[Introdução à Mineração]

Base conceitual

- Processo KDD e CRISP-DM
- Funções centrais da mineração
- Motivação e desafios reais

[Análise Exploratória (EDA)]

- Estatísticas descritivas
- Histogramas e densidades
- Boxplots e outliers
- Correlação e matrizes de dispersão

Implementações completas em código

Pré-processamento

- Tratamento de missing
- Outliers (boxplot, 3σ)
- PCA e seleção de atributos
- Normalização
- Feature engineering

Roteiro técnico com DALToolbox

Modelagem por Regressão

- Regressão simples e múltipla
- Polinomial e overfitting
- Diagnóstico e VIF
- Avaliação com RMSE/MAE

Experimentos reproduzíveis



Conectando Conceitos e Pesquisa

Análise de Séries Temporais - Data Analytics

Eduardo Ogasawara

eduardo.ogasawara@cefet-rj.br

<https://eic.cefet-rj.br/~eogasawara>

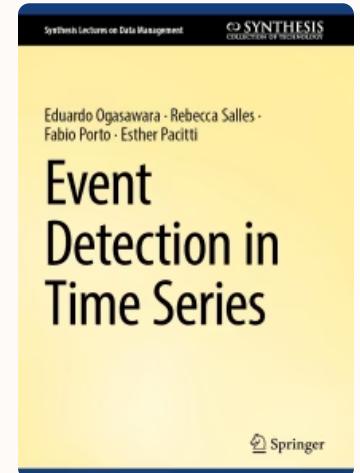
BIOGRAFIA

Professor no CEFET/RJ

Líder do DAL, atuando em análise de séries temporais e detecção de eventos.

Sua atividade docente abrange os cursos: BCC, PPCIC, PPPRO

Autor do livro "Event Detection in Time Series", publicado pela Springer Nature, consolidou sua expertise na área. É membro **IEEE (Sênior)**, SBC e ACM.



Temas de Pesquisa em Séries Temporais

A pesquisa se concentra em técnicas avançadas de análise de séries temporais, com ênfase em três pilares fundamentais que formam a base de sistemas inteligentes de processamento temporal de dados.

Pré-processamento

Preparação resiliente a não-estacionariedade

Predição

Modelos adaptativos e robustos

Detecção de Eventos

Identificação inteligente de padrões

A detecção de eventos abrange múltiplas subáreas especializadas:



Detecção de anomalias

Identificação de padrões incomuns e outliers



Pontos de mudança e desvios

Análise de transições e alterações de comportamento



Descoberta de motifs

Identificação de padrões recorrentes



Detecção online

Processamento em tempo real



Avaliação de detecções

Métricas e validação de resultados

- Referência:** E. Ogasawara, R. Salles, F. Porto, and E. Pacitti, *Event Detection in Time Series*, 1st ed. Springer Nature Switzerland, 2025. doi: 10.1007/978-3-031-75941-3.

Pré-processamento de Dados em Séries Temporais

O pré-processamento é uma etapa crítica que determina a qualidade das análises subsequentes. A pesquisa desenvolve técnicas especializadas para lidar com os desafios únicos de séries temporais não-estacionárias.

Abordagens Principais

- **Normalização adaptativa:** Técnicas que se ajustam às mudanças nas características dos dados ao longo do tempo
- **Transformações resilientes:** Métodos que mantêm a integridade dos padrões temporais
- **Filtragem inteligente:** Remoção de ruídos preservando informações relevantes
- **Aumento de dados:** Geração sintética para enriquecer conjuntos limitados

Extração de Componentes

Decomposição utilizando transformadas de Fourier (FFT), wavelets e decomposição empírica modal (EMD) para revelar estruturas ocultas nos dados.

- Publicação:** E. Ogasawara, L. C. Martinez, D. De Oliveira, G. Zimbrão, G. L. Pappa, and M. Mattoso, "Adaptive Normalization: A novel data normalization approach for non-stationary time series," *Proceedings of the International Joint Conference on Neural Networks*, 2010. doi: 10.1109/IJCNN.2010.5596746.

Predição de Séries Temporais

A predição de séries temporais envolve a aplicação de diversos paradigmas de modelagem, desde técnicas estatísticas clássicas até modelos de aprendizado profundo e foundation models especializados. A pesquisa explora a **adaptabilidade e robustez** desses modelos diante de mudanças nos padrões temporais.



Modelos Lineares

Regressão e ARIMA para padrões estatísticos

Machine Learning

Algoritmos de aprendizado supervisionado

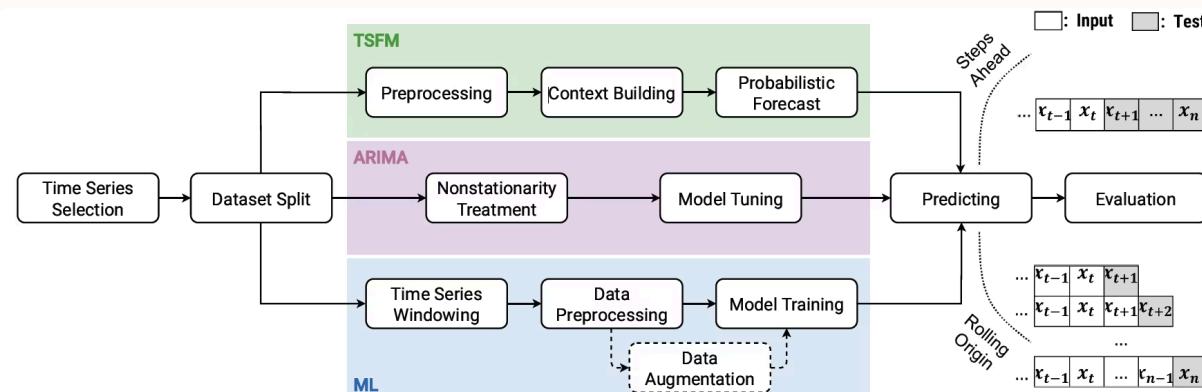
Deep Learning

Redes neurais profundas (LSTM, GRU)

Foundation Models

TSFM de última geração

Predições usando ARIMA, ML e TSFM



A otimização e seleção de hiperparâmetros são aspectos fundamentais para garantir o desempenho ideal dos modelos preditivos.

- ❑ **Publicação:** R. Parracho, F. Alexandrino, L. de Souza Figueiredo, B. Dutra de Macedo, A. Lamblet Vaz, D. Louback, V. C. Desouzart, R. Salles, D. Carvalho, F. Porto, and E. Ogasawara, "Leveraging Large Language Models for Time Series Prediction on Low-Frequency Data," *Simpósio Brasileiro de Banco de Dados (SBBD)*, SBC, 2025.

Detecção de Anomalias em Séries Temporais

A detecção de anomalias identifica padrões incomuns que se desviam significativamente do comportamento esperado.

Formalização e Taxonomia

Desenvolvimento de frameworks conceituais para classificar e caracterizar diferentes tipos de anomalias: pontuais, contextuais e coletivas.

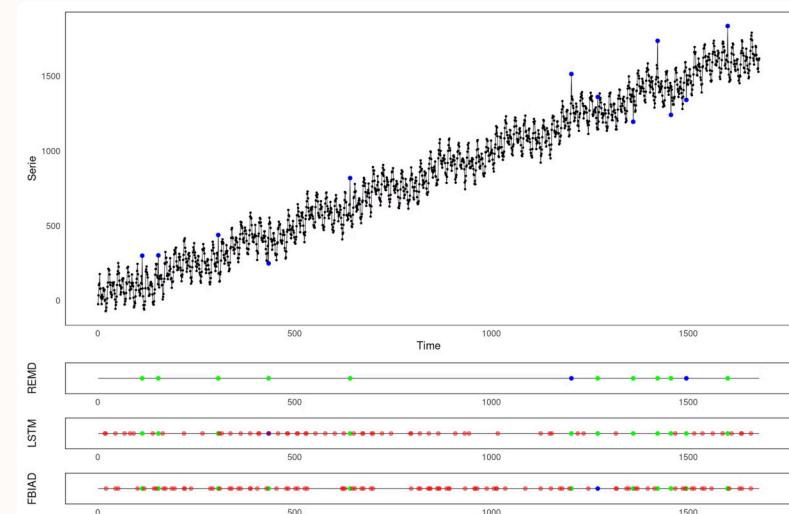
Métodos Avançados

- **Híbridos:** Combinação de múltiplas técnicas para detecção robusta
- **Preditivos:** Baseados em desvios de previsões
- **Reconstrução:** Utilizando autoencoders e modelos generativos

Desafios Principais

1. Não-estacionariedade dos dados
2. Alta dimensionalidade
3. Escassez de rótulos confiáveis

Detector usando REMD, FIBAD, LSTM



- ❑ **Publicação:** J. Souza, E. Paixão, F. Fraga, L. Baroni, R. F. S. Alves, K. Belloze, J. Dos Santos, E. Bezerra, F. Porto, and E. Ogasawara, "REMD: A Novel Hybrid Anomaly Detection Method Based on EMD and ARIMA," *Proceedings of the International Joint Conference on Neural Networks*, 2024. doi: 10.1109/IJCNN60899.2024.10651192.

Detecção de Pontos de Mudança e Desvio de Conceito

Pontos de mudança marcam transições fundamentais no comportamento de séries temporais, enquanto desvios de conceito representam alterações graduais ou abruptas nas relações subjacentes dos dados. A detecção precisa desses eventos é crucial para manter a acurácia de sistemas preditivos.

Métodos Estatísticos

Testes de hipótese e análise de distribuições

Autoencoders

Detecção via espaço latente



Abordagens Inovadoras

Modelos ensemble proporcionam detecção mais robusta ao combinar múltiplos detectores, enquanto a análise espectral em camadas latentes de redes profundas revela mudanças sutis não aparentes no espaço original.

- ❑ **Publicação:** L. G. Tavares et al., "Fuzzy-Based Ensemble Method for Robust Concept Drift Detection in Multivariate Time Series," *Proceedings of the International Joint Conference on Neural Networks*, 2025.

Descoberta de Motifs em Séries Temporais

Motifs são padrões recorrentes em séries temporais que representam comportamentos ou eventos significativos. Sua descoberta e indexação permite a interpretação de eventos complexos e a identificação de estruturas temporais importantes.

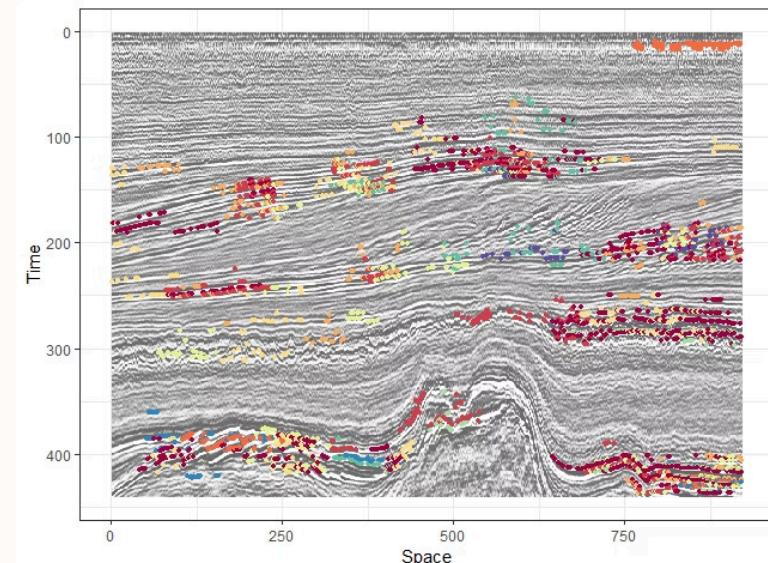
Contribuições Principais

- **Indexação eficiente:** Estruturas de dados para busca rápida de motifs
- **Rotulação semântica:** Atribuição de significado aos padrões descobertos
- **Motifs espaço-temporais:** Extensão para dados com dimensão espacial

Aplicações

A descoberta de motifs tem aplicações em diversas áreas, incluindo monitoramento sísmico, análise de sinais biomédicos, detecção de fraudes e manutenção preditiva.

Detecção em dados sísmicos



- Publicação:** H. Borges, M. Dutra, A. Bazaz, R. Coutinho, F. Perosi, F. Porto, F. Masseglia, E. Pacitti, and E. Ogasawara, "Spatial-time motifs discovery," *Intelligent Data Analysis*, vol. 24, no. 5, pp. 1121–1140, 2020. doi: 10.3233/IDA-194759.

Detecção Online de Eventos

A detecção online processa fluxos contínuos de dados com restrições rigorosas de latência, essencial para aplicações que exigem respostas imediatas. Esta área apresenta desafios únicos relacionados à eficiência computacional e adaptação contínua.

Processamento Contínuo

Análise em tempo real sem interrupções

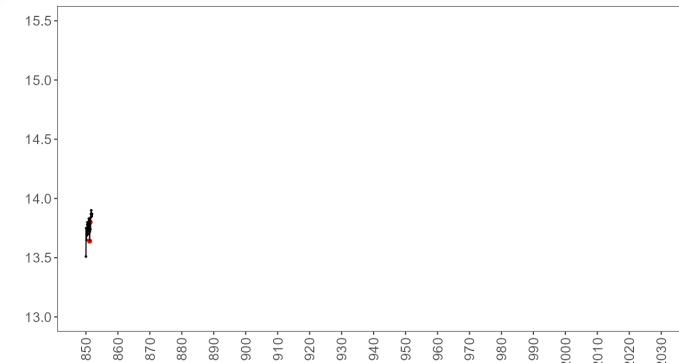
Janelas Deslizantes

Atualização incremental de modelos

Adaptação Dinâmica

Ajuste automático a mudanças

Exemplo: Detecção Online de Anomalia



O sistema ilustrado demonstra a capacidade de identificar anomalias em tempo real, permitindo respostas imediatas a eventos críticos.

- ❑ **Publicação:** J. Lima, L. G. Tavares, E. Pacitti, J. E. Ferreira, I. Santos, I. G. Siqueira, D. Carvalho, F. Porto, R. Coutinho, and E. Ogasawara, "Online Event Detection in Streaming Time Series: Novel Metrics and Practical Insights," *Proceedings of the International Joint Conference on Neural Networks*, 2024. doi: 10.1109/IJCNN60899.2024.10650809.

Avaliação de Detecções em Séries Temporais

A avaliação de detectores de eventos em séries temporais requer métricas especializadas que considerem a natureza temporal dos dados. Diferentemente de problemas de classificação tradicionais, pequenos desalinhamentos temporais não devem ser penalizados severamente.

Desafios na Avaliação

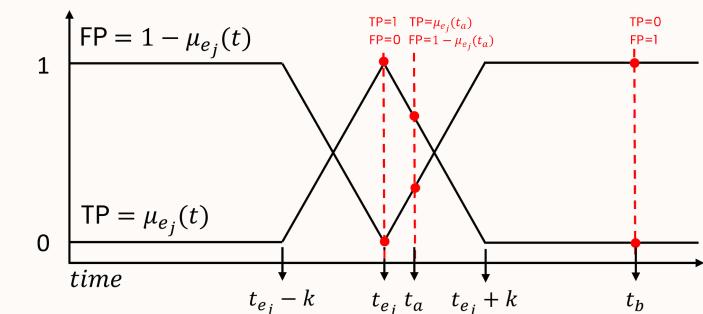
- Tolerância temporal:** Detecções próximas devem ser consideradas corretas
- Curadoria contínua:** Validação de detecções em fluxos de dados
- Imprecisão inerente:** detecção de eventos raramente são exatas

Contribuições

Desenvolvimento de métricas baseadas em lógica fuzzy que incorporam tolerância temporal, permitindo avaliação mais justa e realista de sistemas de detecção.

A construção de benchmarks com rótulos consistentes e interpretações padronizadas é fundamental para o avanço da área, permitindo comparações justas entre diferentes métodos.

Tolerância temporal fuzzy



- **Publicação:** R. Salles, J. Lima, M. Reis, R. Coutinho, E. Pacitti, F. Masseglia, R. Akbarinia, C. Chen, J. Garibaldi, F. Porto, and E. Ogasawara, "SoftED: Metrics for soft evaluation of time series event detection," *Computers and Industrial Engineering*, vol. 198, 2024. doi: 10.1016/j.cie.2024.110728.

Desenvolvimento Tecnológico

Esses softwares implementam diretamente as técnicas apresentadas nos slides anteriores. A produção tecnológica materializa a pesquisa científica em ferramentas práticas que beneficiam pesquisadores, profissionais e a sociedade. Os pacotes de software desenvolvidos são **disponibilizados como código aberto**, promovendo transparência, colaboração e amplo impacto social.



daltoolbox

Framework para ciência de dados em R, com ferramentas integradas para pré-processamento, modelagem, visualização e avaliação. Serve de base para o harbinger e tsredit.



harbinger

Pacote especializado em detecção de eventos em séries temporais. Implementa algoritmos de estado da arte e as métricas SoftED, facilitando a pesquisa reproduzível e aplicações industriais em monitoramento e alertas.



tsredit

Conjunto de ferramentas para previsão de séries temporais, que integram métodos estatísticos clássicos e técnicas de aprendizado de máquina. Permite a comparação de modelos e a seleção automatizada de hiperparâmetros.

Referências e Materiais Complementares



Eduardo Ogasawara

@eduardo.ogasawara • 987 inscritos • 121 vídeos

Sou professor do Departamento de Ciéncia da Computação do Centro Federal de Educac... [...mais](#)
eic.cefet-rj.br/~eogasawara

Personalizar o canal Gerenciar vídeos Acessar Comunidade

Início Vídeos Playlists Posts



Vídeos de Destaque ► Reproduzir tudo

Esta playlist reúne vídeos selecionados que refletem a expertise do professor Eduardo Ogasawara, docente do Departamento de Ciéncia da Computação do CEFET-RJ, com ampla experiência em...

Pesquisa em Análise de Sérias Temporais para Data... Fundamentos da ciéncia e do método científico na... O que é Python na ciéncia de dados: algoritmos, variáveis... Fundamentos de Sérias Temporais e Processos... Introdução à Linguagem R para Análise de Dados e... Problemas, algoritmos e organização de passos na...

Eduardo Ogasawara 90 visualizações • há 2 semanas Eduardo Ogasawara 51 visualizações • há 4 semanas Eduardo Ogasawara 45 visualizações • há 3 semanas Eduardo Ogasawara 40 visualizações • há 11 dias Eduardo Ogasawara 22 visualizações • há 3 semanas Eduardo Ogasawara 16 visualizações • há 9 dias



Jornada de Ciéncia de Dados - Mineração de Dados

Acesse o material completo do curso de Análise de Dados, incluindo conceitos fundamentais, técnicas estatísticas e aplicações práticas. O curso aborda desde os fundamentos até métodos avançados de análise exploratória e inferencial.

<https://eic.cefet-rj.br/~eogasawara/jornada>

Mineração de Dados

Explore slides detalhados e videoaulas do curso de Mineração de Dados. O material cobre algoritmos de descoberta de padrões, técnicas de clustering, classificação e outras metodologias essenciais para extração de conhecimento de grandes volumes de dados.

<https://eic.cefet-rj.br/~eogasawara/data-mining>

- Slides de aula organizados por tópico e módulo
- Videoaulas explicativas com demonstrações práticas
- Exemplos de código e implementações

Todos os recursos estão organizados de forma didática para facilitar o estudo autônomo e servir como material de consulta durante as disciplinas. Recomenda-se acompanhar os conteúdos na sequência proposta para melhor aproveitamento.



Até logo!



Eduardo Ogasawara

Professor Titular

CEFET/RJ

eduardo.ogasawara@cefet-rj.br

Interessado em colaborar?

<https://eic.cefet-rj.br/~eogasawara>

Entre em contato para discutir oportunidades de pesquisa, orientação ou parcerias

Visite nosso laboratório

<https://eic.cefet-rj.br/dal>

Conheça os projetos em andamento e as ferramentas desenvolvidas pelo Data Analytics Lab