



Avaliação como Problema Científico

A comparação científica de detectores de eventos exige comparabilidade experimental rigorosa para validar a superioridade de um sobre o outro.

Eduardo Ogasawara

eduardo.ogasawara@cefet-rj.br

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

Introdução

A avaliação como problema científico exige rigor experimental. Para garantir a comparabilidade e validade da superioridade de um detector de eventos sobre outro, é fundamental que a avaliação seja conduzida sob condições estritamente controladas. A seguir, os pilares dessa comparabilidade:

Mesma Base de Dados

Os detectores devem ser testados usando exatamente os mesmos dados. Isso elimina vieses causados por diferenças na natureza, tamanho ou qualidade dos dados.

Mesmas Métricas

As métricas de avaliação (e.g., precisão, recall, F1-score) devem ser idênticas para todos os detectores, garantindo uma medida de desempenho consistente e comparável.

Mesmo Protocolo

O processo de avaliação (e.g., divisão treino/teste, técnicas de validação cruzada, métodos de otimização de hiperparâmetros) deve ser uniforme.

Considere dois detectores D_1 e D_2 . A comparação justa entre eles só é possível quando submetidos às mesmas condições experimentais.

Benchmark em Detecção de Eventos

Um benchmark é uma estrutura formal de validação que torna a comparação repetível e científica. Qualquer pesquisador pode testar diferentes detectores no mesmo benchmark e obter resultados comparáveis.

Essa formalização é essencial para transformar avaliação subjetiva em processo científico robusto.

Componentes do Benchmark

$$\mathcal{B} = (\mathcal{X}, \mathcal{E}, \mathcal{M})$$

- \mathcal{B} = benchmark completo
- \mathcal{X} = conjunto de séries temporais
- \mathcal{E} = eventos anotados (ground truth)
- \mathcal{M} = conjunto de métricas de avaliação



Incerteza do Ground Truth

Em muitos domínios, o que chamamos de ground truth é uma aproximação que depende de especialistas, instrumentos e critérios. A avaliação precisa reconhecer que a referência pode ter incerteza e não ser uma "verdade absoluta".

Eventos Incompletos

Observações podem não capturar todos os eventos que realmente ocorreram no sistema

Anotações Subjetivas

Diferentes especialistas podem identificar ou classificar eventos de formas distintas

Divergência Real vs. Observado

O conjunto de eventos reais \mathcal{E}_{real} pode diferir significativamente do observado $\mathcal{E}_{observado}$

$$\mathcal{E}_{real} \neq \mathcal{E}_{observado}$$

Benchmarks Sintéticos

Dados sintéticos são gerados por modelos controláveis onde os eventos são conhecidos por construção. Você sabe exatamente onde estão os eventos porque os colocou no processo gerador.

$$x_t = f_\theta(t) + \varepsilon_t$$

Vantagens e Limitações

Controle total: Diagnósticos limpos e eventos perfeitamente conhecidos

Realismo limitado: Pode não refletir complexidades e imperfeições de dados reais

- **Símbolos:** x_t = observação no instante t; $f_\theta(t)$ = componente determinística; ε_t = ruído; θ = parâmetros do gerador

Benchmarks Reais



Finanças

Dados de mercados financeiros com eventos reais de volatilidade e mudanças de regime



Sensores Industriais

Monitoramento de processos com falhas e anomalias autênticas



Redes

Tráfego de rede com intrusões e comportamentos anômalos reais



Biomédico

Sinais fisiológicos com eventos clínicos documentados

Benchmarks reais testam detectores em condições que importam na prática. O custo é que o ground truth pode ser incompleto ou incerto — a avaliação mede desempenho "sob referência imperfeita". Assume-se que \mathcal{E}_{real} é parcialmente observável.

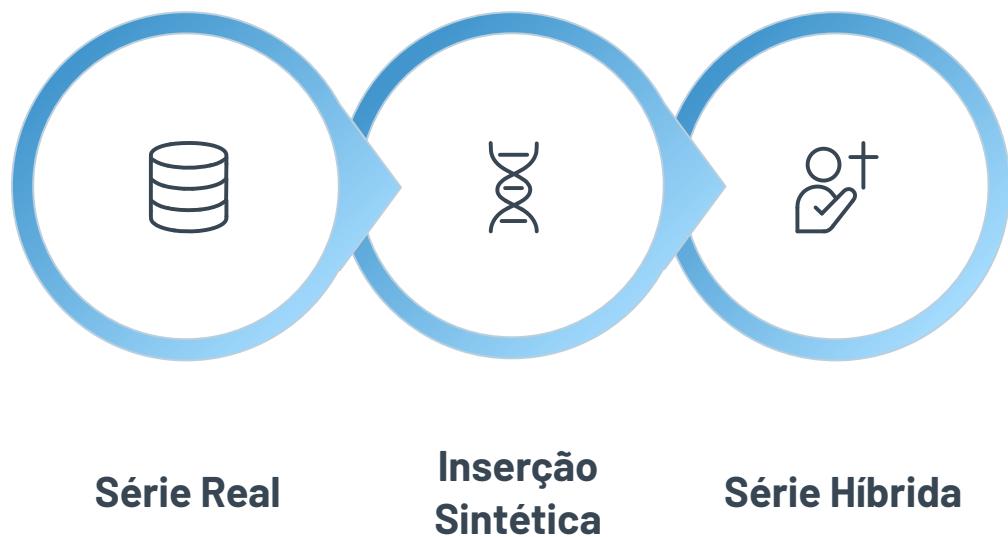
Benchmarks Híbridos

A abordagem híbrida tenta capturar o melhor dos dois mundos: manter a textura de dados reais e, ao mesmo tempo, ter eventos controlados para avaliação. Isso ajuda a calibrar sensibilidade e robustez sem perder realismo.

Eventos artificiais são inseridos em séries reais, permitindo que o ground truth do componente inserido seja perfeitamente conhecido enquanto o contexto permanece realista.

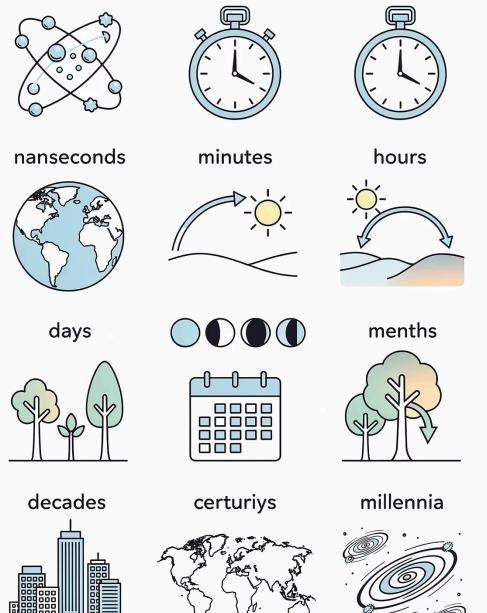
$$x_t = x_t^{real} + e_t^{sintético}$$

- x_t = série híbrida resultante
- x_t^{real} = componente real observado
- $e_t^{sintético}$ = componente sintético inserido



Escalas Temporais

Eventos podem existir em diferentes janelas temporais. Mudar a escala muda fundamentalmente o que conta como evento. Um detector pode parecer ótimo numa escala curta e ruim numa escala longa, ou vice-versa.



1

Escala Curta (w_1)

Eventos de alta frequência, picos locais, anomalias pontuais

2

Escala Média (w_2)

Padrões intermediários, tendências de curto prazo

3

Escala Longa (w_k)

Mudanças de regime, tendências de longo prazo

Benchmarks e protocolos precisam explicitar a escala temporal, senão a comparação fica inconsistente.

Formalmente: $w_1 < w_2 < \dots < w_k$ e $\mathcal{E}(w_1) \neq \mathcal{E}(w_2)$

Hierarquia de Eventos

Eventos não são necessariamente independentes. Uma anomalia pode ocorrer dentro de um período de mudança mais amplo, que por sua vez faz parte de um regime maior. Essa estrutura hierárquica complica a avaliação porque "acertar" pode depender do nível que você quer medir.

	<h3>Nível 1: Local</h3> <p>Eventos pontuais e anomalias específicas</p>
	<h3>Nível 2: Padrão</h3> <p>Sequências e comportamentos recorrentes</p>
	<h3>Nível 3: Regime</h3> <p>Mudanças estruturais amplas no sistema</p>

Relação de Contenção

$$e_i^{(\text{nível 1})} \subset e_j^{(\text{nível 2})}$$

Onde $e_i^{(\text{nível 1})}$ representa eventos em nível mais "fino" e $e_j^{(\text{nível 2})}$ eventos em nível mais "amplo".

A avaliação pode precisar respeitar e medir desempenho em múltiplos níveis simultaneamente.

⚠ ATENÇÃO

Significado dos Eventos e Limitação das Métricas

Dois detectores podem ter métricas parecidas e ainda assim estarem capturando "tipos" de fenômenos completamente diferentes. Um pode identificar picos locais enquanto outro detecta mudanças de regime. Isso mostra que métrica não é sinônimo de significado.

O Problema da Equivalência Métrica

$$\mathcal{E}_1 \neq \mathcal{E}_2, \quad F_1(\mathcal{E}_1) \approx F_1(\mathcal{E}_2)$$

Detectores podem "explicar" coisas distintas mesmo com desempenho numérico similar. A avaliação não captura automaticamente a semântica dos eventos.

Métricas Parecidas

Valores numéricos similares de F_1 , precisão ou recall

Eventos Diferentes

Fenômenos distintos sendo capturados por cada detector

Interpretação Necessária

Análise qualitativa complementa avaliação quantitativa

Função Geral de Avaliação

A qualidade da avaliação não é um único número "universal". Você mede múltiplas dimensões que se combinam para formar uma visão completa do desempenho. A avaliação é multidimensional por natureza e depende do objetivo científico.

$$Q = f(\text{tempo, estrutura, escala, semântica})$$



Tempo

Atrasos de detecção e tolerância temporal



Estrutura

Intervalos, padrões e relações entre eventos



Escala

Janela temporal de observação e análise



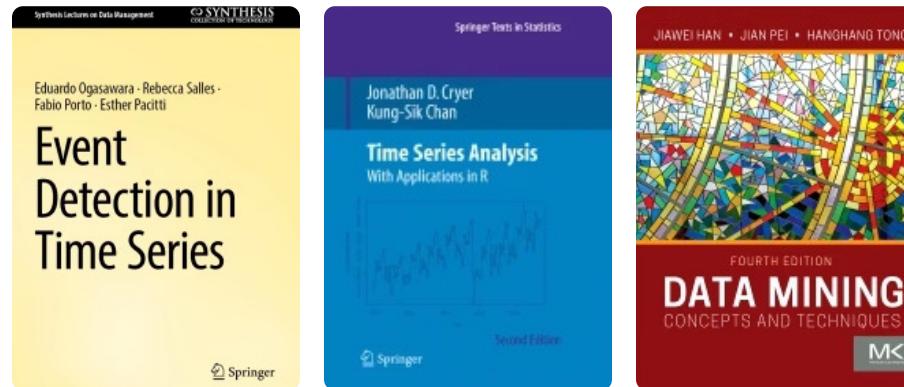
Semântica

Significado do evento no domínio de aplicação

Onde Q representa a qualidade da avaliação e $f(\cdot)$ é a regra que combina as dimensões de avaliação de forma apropriada ao contexto.

Referências Bibliográficas

Uma coleção cuidadosamente selecionada de obras fundamentais que abordam análise de séries temporais e mineração de dados.



Event Detection in Time Series

Ogasawara, E.; Salles, R.; Porto, F.; Pacitti,

E. (2025). Publicação recente da Springer Nature Switzerland que explora técnicas avançadas de detecção de eventos em séries temporais.

Time Series Analysis: With Applications in R

Cryer, J. D.; Chan, K.-S. (2008). Obra clássica da Springer que combina fundamentação teórica sólida com implementações práticas.

Data Mining: Concepts and Techniques

Han, J.; Pei, J.; Tong, H. (2022). Quarta edição publicada pela Morgan Kaufmann que consolida conceitos fundamentais e técnicas avançadas de mineração de dados