

O Problema da Avaliação de Eventos

Medir a correspondência entre eventos reais e detectados ao longo do tempo é crucial para a validação de algoritmos.

Eduardo Ogasawara

eduardo.ogasawara@cefet-rj.br

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

Introdução

Avaliar detecção de eventos significa medir a correspondência entre dois conjuntos temporais: eventos reais e eventos detectados. O objetivo é comparar o fenômeno no tempo versus a estimativa algorítmica.

Considere dois conjuntos de instantes de evento:

$$\mathcal{E} = \{\tau_1, \dots, \tau_n\}, \quad \hat{\mathcal{E}} = \{\hat{\tau}_1, \dots, \hat{\tau}_m\}$$

Onde:

- \mathcal{E} representa eventos reais
- $\hat{\mathcal{E}}$ representa eventos detectados
- τ_i é o instante do i-ésimo evento real
- $\hat{\tau}_j$ é o instante do j-ésimo evento detectado
- n é o número de eventos reais
- m é o número de eventos detectados

Paradigma Clássico de Avaliação

O machine learning clássico utiliza classificação binária com contagens discretas: TP (verdadeiros positivos), FP (falsos positivos), TN (verdadeiros negativos) e FN (falsos negativos).

Este paradigma pressupõe rótulos exatos por amostra, funcionando bem em uma grade discreta onde cada amostra é classificada como "evento" ou "não evento". O problema surge quando aplicamos isso ao tempo como fenômeno contínuo.

O Tempo Como Problema

Eventos têm localização temporal, podem ter duração e incerteza. O erro não é apenas "acertou ou errou", mas "quão longe no tempo" a detecção ocorreu.

Localização Temporal

Eventos ocorrem em instantes específicos no tempo

Duração e Incerteza

Eventos podem ter extensão temporal e imprecisão

Erro Contínuo

Deslocamento temporal não implica erro absoluto

Se um detector marca um instante diferente do real, isso pode ser apenas um deslocamento temporal:

$$\tau_i \neq \hat{\tau}_j \text{ não implica, por si só, erro absoluto}$$

Precision, Recall e F₁

Métricas fundamentais da classificação que dependem de TP, FP e FN. São úteis, mas precisam de definição temporal clara do que constitui um "acerto".

Fórmulas das Métricas

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Definições

- **Precision:** fração de detecções que são corretas
- **Recall:** fração de eventos reais que foram detectados
- **F₁:** média harmônica entre Precision e Recall

- Em séries temporais, a pergunta central é: o que conta como TP? Sem uma regra temporal de correspondência, TP/FP/FN ficam indefinidos ou inconsistentes.



Limitações no Contexto Temporal

Eventos raramente coincidem exatamente no tempo. TP, FP e FN ficam ambíguos sem tolerância temporal, e a avaliação clássica ignora completamente o "quanto errou no tempo".

Coincidência Exata é Rara

Em geral, vale $\tau_i \neq \hat{\tau}_j$

Ambiguidade nas Contagens

Sem tolerância, TP/FP/FN perdem significado prático

Perda de Informação

Avaliação binária ignora magnitude do erro temporal

Se você exigir coincidência exata, quase tudo vira erro, mesmo quando o detector está "praticamente certo". A avaliação precisa incorporar proximidade temporal para não punir detecções úteis.

SOLUÇÃO

Correspondência Entre Eventos

Define-se uma regra de matching (pareamento) onde um acerto significa estar dentro de uma janela de tolerância. Isso aproxima a avaliação do que é relevante no tempo.

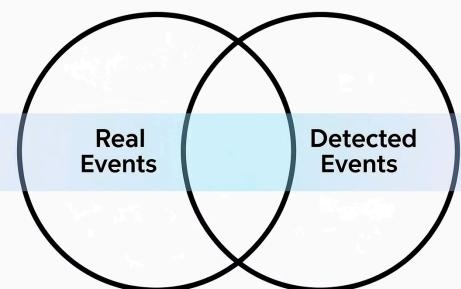
Um pareamento simples é definido como:

$$M(\tau_i, \hat{\tau}_j) = \begin{cases} 1, & \text{se } |\tau_i - \hat{\tau}_j| \leq \delta \\ 0, & \text{caso contrário} \end{cases}$$



Onde $M(\tau_i, \hat{\tau}_j)$ é o indicador de correspondência (1 ou 0) e δ é a tolerância temporal máxima aceitável. Esta é a ponte entre métrica de classificação e erro temporal.

TP, FP e FN com Tolerância



Nova Definição

TP passa a depender do matching temporal, FP são detecções sem pareamento aceitável, e FN são eventos reais não pareados.

Fórmulas de Contagem

$$TP = \sum_{i,j} M(\tau_i, \hat{\tau}_j)$$

$$FP = |\hat{\mathcal{E}}| - TP$$

$$FN = |\mathcal{E}| - TP$$

Onde $|\hat{\mathcal{E}}|$ é o número de eventos detectados e $|\mathcal{E}|$ é o número de eventos reais.

A grande mudança: TP deixa de ser "acertou exatamente" e vira "pareou dentro da tolerância". FP e FN vêm naturalmente como sobras.

Tolerância Temporal δ

O parâmetro δ controla o que é considerado acerto. Um δ grande facilita pareamentos, enquanto um δ pequeno exige maior precisão temporal.



δ Grande

Aceita atrasos maiores, detector parece melhor em TP/Recall

δ Pequeno

Exige precisão temporal rigorosa, mede qualidade mais dura

A condição de aceitação é:

$$|\tau_i - \hat{\tau}_j| \leq \delta$$

- ❑ A tolerância é a "régua" da avaliação. Por isso δ é parte do próprio problema de avaliação, não apenas um parâmetro técnico.

Erro Temporal de Detecção

Além de contar acertos, mede-se o deslocamento no tempo. O erro é contínuo (pode ser positivo ou negativo), e o MAE no tempo resume a magnitude média do erro.

Erro por Evento

$$\Delta\tau_i = \hat{\tau}_i - \tau_i$$

Onde $\Delta\tau_i$ é o erro temporal do i -ésimo evento, τ_i é o instante real e $\hat{\tau}_i$ é o instante detectado.

Erro Médio Absoluto

$$MAE_\tau = \frac{1}{|\mathcal{E}|} \sum_i |\Delta\tau_i|$$

O MAE_τ resume a magnitude média do erro e conversa diretamente com latência e precisão temporal.

Esta segunda camada de avaliação não apenas diz "pareou ou não pareou", mas mede o quanto a detecção se deslocou no tempo.

Métricas como Função de δ

Precision, Recall e F_1 dependem do critério de pareamento. Variar δ muda TP, FP e FN, portanto a qualidade do detector não é um número único.

Precision(δ)

Varia conforme tolerância temporal

Recall(δ)

Depende da janela de aceitação

$F_1(\delta)$

Função da exigência temporal

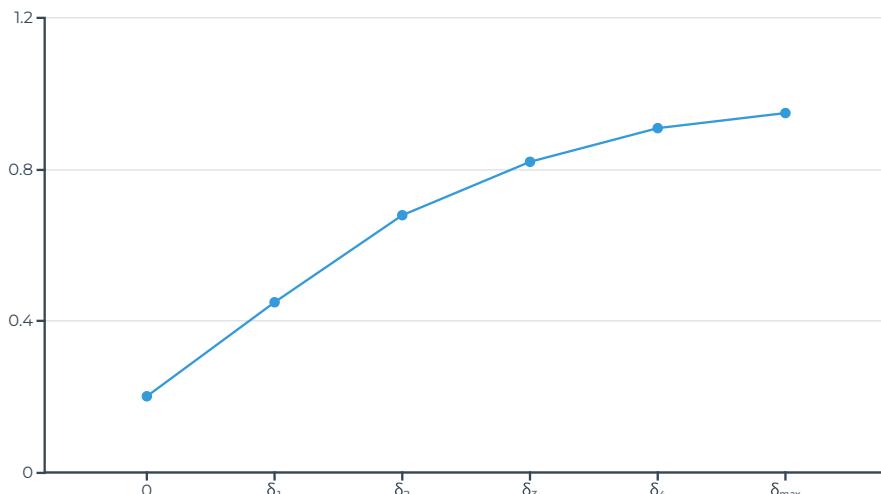
As métricas tornam-se funções:

$$\text{Precision}(\delta), \quad \text{Recall}(\delta), \quad F_1(\delta)$$

Quando você muda δ , você muda a regra do jogo. Um detector pode ser ótimo para tolerâncias grandes mas ruim para tolerâncias pequenas.

Curva de Desempenho Temporal

Em vez de avaliar um único ponto, avalia-se uma curva completa. O parâmetro δ varre níveis de exigência temporal, e a curva descreve a robustez do detector ao erro temporal.



Interpretação da Curva

Uma curva típica é expressa como:

$$F_1(\delta), \quad \delta \in [0, \delta_{\max}]$$

Esta curva mostra como o detector se comporta quando você vai "apertando" ou "afrouxando" a tolerância temporal.

É uma forma muito mais informativa do que reportar um único F_1 , porque revela quão sensível o método é ao tempo.

Avaliação como Problema Contínuo

A avaliação depende de \mathcal{E} (eventos reais), $\hat{\mathcal{E}}$ (eventos detectados) e δ (tolerância). O resultado é função de escolhas temporais, e o tempo entra como variável de controle da qualidade.

Uma forma abstrata de representar a avaliação:

$$Q = f(\mathcal{E}, \hat{\mathcal{E}}, \delta)$$

Onde Q é a medida de qualidade da detecção e $f(\cdot)$ é a regra de avaliação.

Dependência de Contexto

Não existe qualidade sem especificar como você aceita erro no tempo

Problema Parametrizado

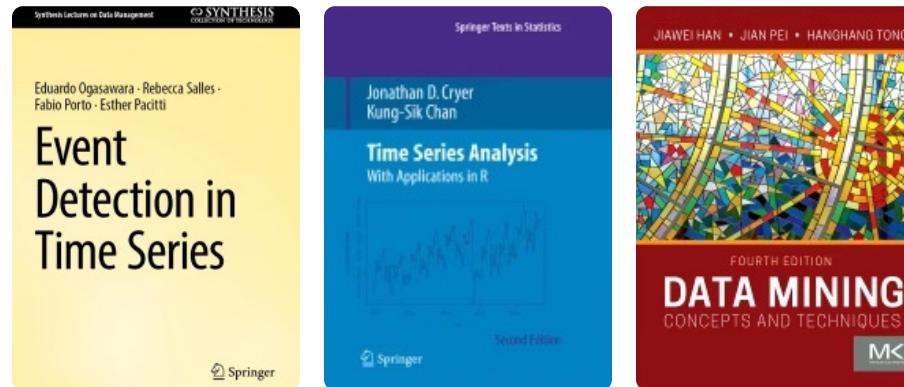
A avaliação é contínua e depende do que você considera aceitável

Controle de Qualidade

O tempo é variável fundamental na definição de desempenho

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