

GerModel: Gestão de Modelos em Dados em Fluxo com Conceitos Recorrentes

Rodrigo Machado · Lucas Tavares · Laís Baroni ·
Fabio Porto · Rafaelli Coutinho · Eduardo Ogasawara

Motivação

Dados meteorológicos chegam continuamente e mudam de comportamento ao longo do tempo — o que funcionava bem no verão pode não funcionar no inverno. Manter modelos preditivos atualizados nesse cenário é um desafio constante.

Retreinamento Custoso

Toda vez que o comportamento dos dados muda, a solução mais simples é treinar um modelo do zero. Mas isso consome tempo e processamento — especialmente quando os dados chegam em alta frequência.

Padrões Sazonais

Padrões climáticos se repetem: o verão de 2024 se parece com o de 2023. Isso sugere que modelos antigos podem ser reaproveitados quando o padrão volta a aparecer.

Recorrência como Oportunidade

Se padrões se repetem, o problema não é só detectar mudanças — é decidir de forma inteligente quando retreinar e quando reaproveitar um modelo já treinado.

Problema

Quando o comportamento dos dados muda, o modelo preditivo pode perder qualidade. A pergunta central é: **vale a pena treinar um modelo novo do zero, ou existe um modelo antigo que já funciona bem para esse padrão?**

Decisão Pós-Mudança

Após detectar que os dados mudaram, é preciso escolher: retreinar completamente (caro) ou buscar um modelo histórico que já viu esse padrão antes (mais rápido).

Equilíbrio Desempenho × Custo

Retreinar sempre garante atualização, mas é caro. Reutilizar é mais rápido, mas só funciona se o padrão atual for realmente parecido com o passado.

Critério de Seleção

Falta uma regra prática para comparar o padrão atual com os históricos e decidir qual modelo antigo é o mais adequado para reutilizar.

Contribuição

O **GerModel** propõe uma abordagem para gerenciar um repositório de modelos treinados, decidindo automaticamente entre reutilizar um modelo antigo ou treinar um novo — avaliada com dados reais de precipitação.

Gestão em Fluxo

Em vez de descartar modelos antigos, o GerModel os armazena e os mantém disponíveis para reuso futuro, como uma biblioteca de modelos históricos.

Similaridade como Critério

Para decidir qual modelo reutilizar, o GerModel compara o perfil estatístico dos dados atuais com o dos dados em que cada modelo foi treinado — usando o PSI (Population Stability Index), uma medida de divergência entre distribuições.

Reuso vs. Retreinamento

Se o padrão atual for suficientemente parecido com um padrão passado, o modelo antigo é reutilizado. Caso contrário, um novo modelo é treinado e adicionado ao repositório.

Concept Drift e Recorrência

O que é Concept Drift?

Concept drift é quando o comportamento dos dados muda ao longo do tempo, fazendo com que um modelo treinado no passado passe a errar mais no presente. Há dois tipos principais:

- **Drift real**: a relação entre as variáveis e o que se quer prever muda (ex.: o mesmo vento que antes indicava chuva agora não indica mais)
- **Drift virtual**: o perfil dos dados de entrada muda, mas a relação preditiva permanece a mesma

Por que a Recorrência Importa?

Em dados climáticos, padrões se repetem sazonalmente. O verão chuvoso de 2022 pode se parecer muito com o de 2024. Quando isso acontece, um modelo treinado em 2022 pode funcionar bem em 2024 — sem precisar retreinar.

Isso abre uma **terceira via**: em vez de apenas detectar mudança e retreinar, é possível identificar qual padrão passado está voltando e reutilizar o modelo correspondente.

Population Stability Index (PSI)

O PSI é uma medida estatística que quantifica o quanto a distribuição dos dados mudou entre dois períodos. No GerModel, ele é usado para comparar o perfil dos dados atuais com o perfil dos dados em que cada modelo histórico foi treinado.

Como é calculado?

$$PSI = \sum_{i=1}^{n} (P_i - Q_i) \ln\left(\frac{P_i}{Q_i}\right)$$

Onde P_i é a proporção observada no período de referência (treinamento) e Q_i é a proporção no período atual, calculadas por faixa de valores. Quanto maior o PSI, maior a diferença entre os dois perfis.

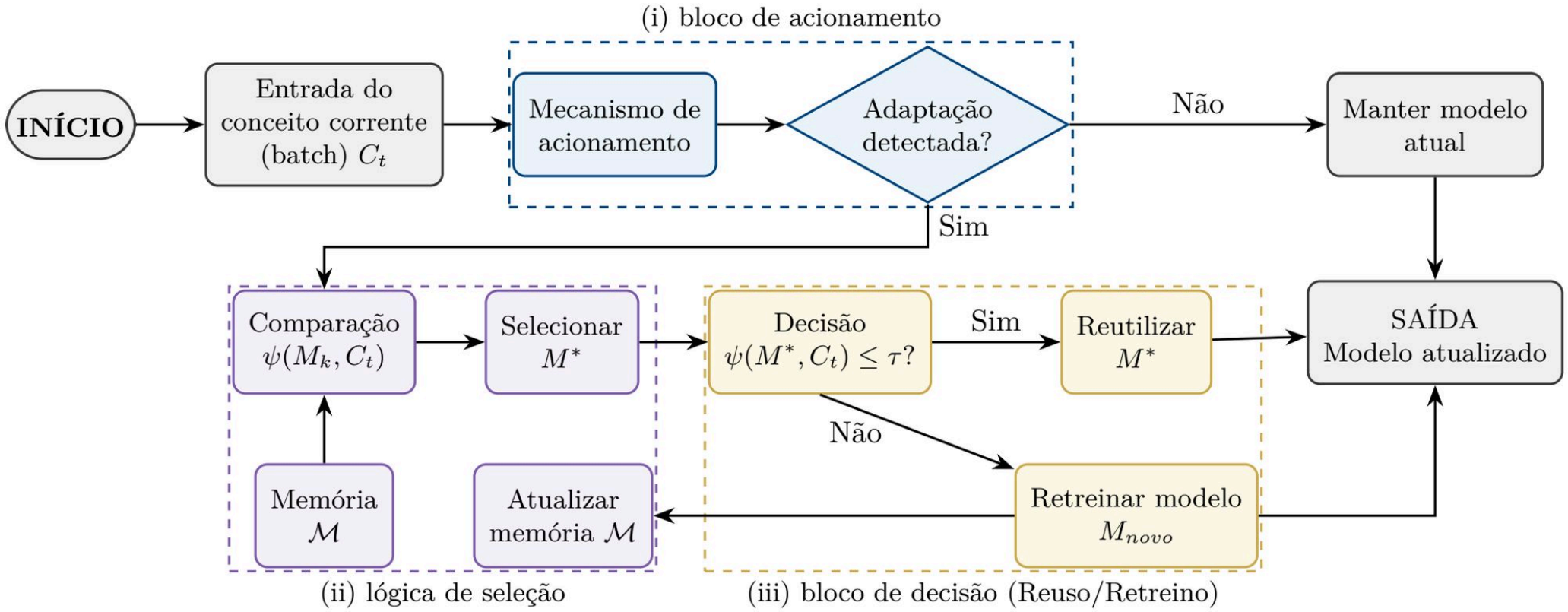
Como os dados são comparados?

- **Atributos numéricos** (ex.: temperatura): divididos em 10 faixas de igual frequência, comparando quantos dados caem em cada faixa
- **Atributos categóricos** (ex.: direção do vento): comparados pelas frequências relativas de cada categoria

⚠ **Limitação importante:** o PSI compara apenas o perfil das variáveis de entrada (X), não avalia diretamente se o modelo ainda acerta bem. Por isso, ele funciona como um filtro inicial — e não substitui a avaliação do desempenho preditivo.

Fluxo da GerModel

A arquitetura do GerModel organiza a adaptação em três blocos funcionais encadeados (Figura 1):



O bloco de acionamento (i) verifica se há necessidade de adaptação via DDM ou PASSIVE; a lógica de seleção (ii) calcula o PSI médio $\psi(M_k, C_t)$ contra modelos históricos e seleciona M^*; o bloco de decisão (iii) compara $\psi(M^*, C_t)$ com o limiar τ — se $\leq \tau$, reutiliza M^*; caso contrário, retreina M_{novo} e atualiza a memória $\mathcal{M}$.

Regra de Reuso

Com vários modelos históricos armazenados, o GerModel precisa decidir qual deles é o mais parecido com o momento atual — e se essa semelhança é suficiente para justificar o reuso.

Como encontrar o modelo mais similar?

$$\psi(M_k, C_t) = \frac{1}{p} \sum_{j=1}^{p} PSI_j(M_k, C_t)$$

$$M^* = \arg\min_k \ \psi(M_k, C_t)$$

O GerModel calcula o PSI médio entre os dados atuais e os dados de treinamento de cada modelo histórico. O modelo com menor PSI médio é o candidato ao reuso — ou seja, aquele cujo perfil de dados mais se assemelha ao momento atual.

Quando reutilizar ou retreinar?

Com o modelo candidato identificado, aplica-se um limiar τ:

- Se o PSI médio $\leq \tau$ → os dados atuais são suficientemente parecidos com os do modelo histórico → **reutiliza o modelo**
- Se o PSI médio $> \tau$ → a diferença é grande demais → **treina um modelo novo** e o adiciona ao repositório

No experimento, adotou-se $\tau = 0{,}25$, valor usualmente associado a mudanças substanciais nas distribuições.



Cenário Experimental

Fonte dos Dados

Dados horários do **Aeroporto Internacional de Miami**, cobrindo o período de 2001 a 2025.

Granularidade

Processamento **mensal**: cada mês é tratado como um conceito corrente independente.

Alvo Binário

Precipitação $\geq 0{,}01$ polegadas na hora seguinte — Miami foi escolhida pelo contraste entre períodos secos e chuvosos.

Preditores e Classificadores

Variáveis Preditoras

- Temperatura do ar (tmpf)
- Ponto de orvalho (dwpf)
- Umidade relativa (relh)
- Direção do vento (drct)
- Velocidade do vento (sknt)

SMOTE é aplicado para mitigar o desbalanceamento da classe positiva (precipitação).

Classificadores Avaliados

NB

Naive Bayes

DT

Árvore de Decisão

RF

Random Forest

MLP

Rede Neural

Implementação em **R** com os pacotes *daltoolbox* e *heimdall* (CRAN-R). Avaliação por F1-score médio, acurácia, precisão e recall. Análise estatística via teste de Wilcoxon pareado com ajuste de Bonferroni ($\alpha = 0{,}05$).

Acionamento e Baseline

Antes de decidir entre reutilizar ou retreinar, o sistema precisa saber quando agir. Dois mecanismos foram testados para disparar essa decisão:

DDM (Drift Detection Method)

O DDM monitora continuamente a taxa de erro do modelo. Quando os erros aumentam além de um limite estatístico, ele sinaliza que algo mudou e que é hora de adaptar. Parâmetros usados: `min_instances = 2000`, `warning = 3`, `out_control = 4`.

PASSIVE (Adaptação Sistemática)

Sem esperar por sinais de degradação, o sistema adapta o modelo a cada novo mês de dados — independentemente de ter detectado mudança ou não. Serve como contraste: o que acontece se sempre tentarmos adaptar?

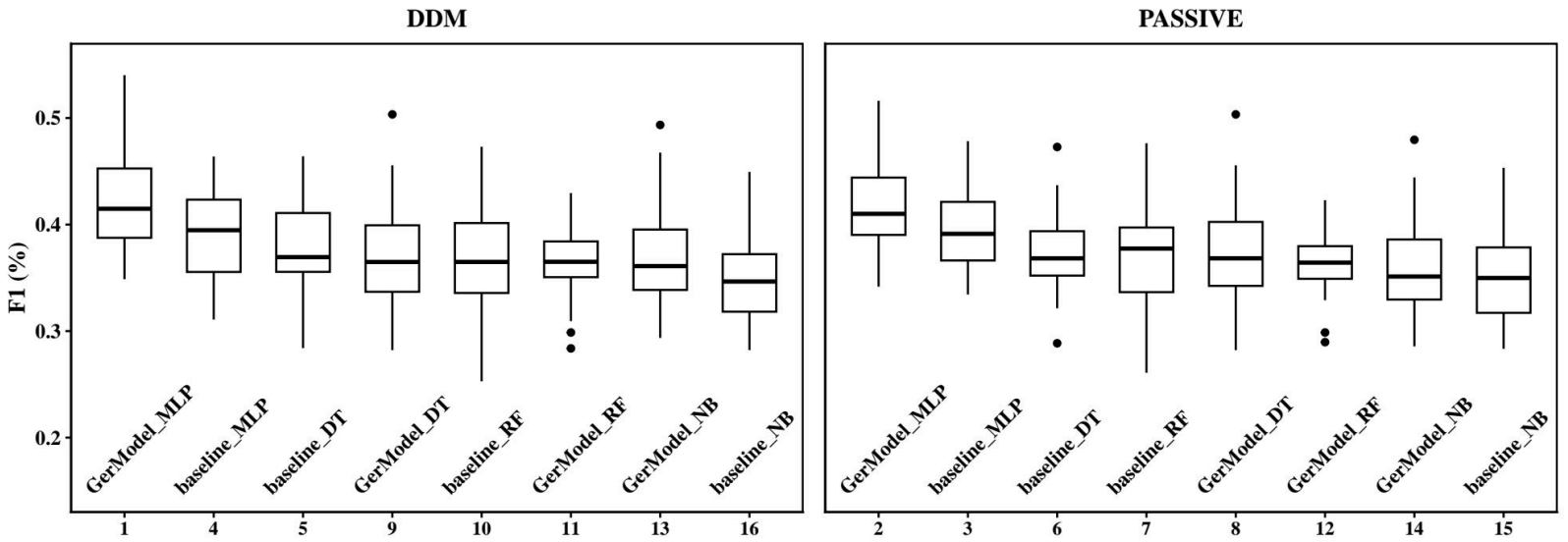
Baseline (Referência)

Em ambos os mecanismos, o baseline é simplesmente retreinar o modelo do zero sempre que a adaptação é acionada — sem tentar reutilizar nada. É o ponto de comparação para avaliar se o reuso do GerModel vale a pena.

Resultados Principais

O GerModel demonstrou ganhos consistentes de F1 e reduções expressivas de tempo de processamento em ambos os cenários de acionamento (Tabela 1 e Figura 2).

Tipo	Classif.	Mecanismo	F1(%)	Tproc(ms)
GerModel	MLP	DDM	42,4	48
baseline	MLP	DDM	39,3	188
GerModel	MLP	PASSIVE	41,4	34
baseline	MLP	PASSIVE	39,7	287



O ganho mais consistente é a **eficiência computacional** com F1 competitivo — DDM favorece o melhor F1 absoluto, enquanto PASSIVE oferece o melhor compromisso entre eficiência e reuso (menor tempo médio de processamento).

Reuso e Significância Estatística

Em quase todos os casos, o GerModel optou por reutilizar modelos históricos em vez de retreinar — com taxas acima de 98%. Isso significa que o custo computacional foi drasticamente reduzido. Mas será que o desempenho preditivo foi mantido?

Taxas de Reuso por Configuração

Mecanismo	Classificador	Reuso(%)
DDM	MLP	98,0
DDM	DT	98,1
DDM	RF	98,1
DDM	NB	98,0
PASSIVE	MLP	98,7
PASSIVE	DT	98,7
PASSIVE	RF	98,7
PASSIVE	NB	98,7

O ganho em F1 é real?

Para verificar se a melhora no F1 do GerModel com MLP é estatisticamente significativa (e não apenas variação aleatória), aplicou-se o teste de Wilcoxon pareado com correção de Bonferroni:

- **MLP + DDM**: $p = 0{,}0013$, efeito grande $(r = 1{,}76)$ ✓ Significativo
- **MLP + PASSIVE**: $p = 0{,}0406$, efeito grande $(r = 1{,}37)$ ✓ Significativo

Para os demais classificadores (DT, RF, NB), as diferenças entre GerModel e baseline não foram estatisticamente significativas — mas o reuso ainda reduz o tempo de processamento sem perda de desempenho.

Conclusão

O GerModel demonstra que o **reuso orientado por similaridade** pode ser uma alternativa viável ao retreinamento completo em fluxos com concept drift recorrente.

Desempenho Competitivo

Mantém F1 competitivo enquanto reduz drasticamente o custo de adaptação.

PSI como Triagem

O PSI funciona como critério operacional simples e eficaz para seleção de modelos históricos.

Trabalhos Futuros

Extensão para outros domínios e investigação de mecanismos de seleção mais sofisticados.

> Nem todo drift exige retreinamento completo — quando o conceito volta, o conhecimento histórico pode e deve ser reutilizado.



CEFET/RJ

Agradecimentos

FAPERJ

CAPES

CNPq
Conselho Nacional de Desenvolvimento Científico e Tecnológico

Rodrigo Machado · Lucas Tavares · Laís Baroni · Fabio Porto · Rafaelli Coutinho · Eduardo Ogasawara

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