Beyond Individual Input for Deep Anomaly Detection on Tabular Data

Autoria: Thimonier H. et al.

Apresentação: André Luiz Moreira Dutra

Índice

01

Introdução

Problema: Detecção de Anomalias

- Detecção de anomalias em dados tabulares
- Quais amostras/linhas de uma tabela fogem de uma noção de normalidade dos dados da tabela?

plant location	length of leaf (cm)		
	leaf 1	leaf 2	leaf 3
full shade	21	20	21
partial shade	42	40	41
no shade	61	72	62

anomaly

Motivações

Medicina

- Diagnóstico de doenças
- Diagnósticos antecipados

Cibersegurança

- Detecção de fraudes
- Detecção de tráfego intruso

Aprendizado de Máquina

- Limpeza de dados para remoção de outliers fora do padrão a ser detectado.

Peso	Altura
80kg	1.74m
3kg	0.5m
3kg	1.74m
400kg	7m

Dependências feature-feature

- Correspondem a padrões de dependência entre features (colunas) da tabela.
- Uma amostra anômala quanto a dependências feature-feature possui valores de features incompatíveis com suas próprias outras features.

Peso	Altura
80kg	1.74m
3kg	0.5m
3kg	1.74m
400kg	7m

Dependências amostra-amostra

- Correspondem a padrões de dependência entre amostras (linhas) da tabela.

- Uma amostra anômala quanto a dependências amostra-amostra possui valores de features incompatíveis com os valores das mesmas features em outras amostras da tabela.

02

Método

Abordagens

Abordagens Baseadas em Reconstrução

- Aborda a detecção de amostras anômalas por meio de autoencoders.
- Modelos recebem uma amostra e geram uma cópia.
- Anomalias são diferentes da cópia feita pelo modelo.

Técnicas de Masking

- Uso de máscaras nas amostras que ocultam algumas de suas features.
- Modelos tentam inferir os valores ocultos da amostra a partir de valores conhecidos da amostra.
- Anomalias têm valores ocultos reais diferentes dos valores preditos.

Abordagem do modelo

- Adicionar detecção de dependências amostra-amostra às técnicas de masking.
- Recebe um conjunto de dados de treino, conhecidamente não-anômalo, e um conjunto desconhecido.
- Aplica máscaras nas amostras do conjunto desconhecido
- O modelo tenta inferir as features ocultas a partir de ambos os conjuntos de dados.
- Anomalias têm valores reais diferentes dos valores preditos.

Objetivo Final do Modelo

- **x:** amostra, linha da tabela
- **m:** máscara aplicada à amostra. 1 sse o valor é oculto
- **phi_theta:** modelo de predição das features ocultas
- **theta:** parâmetros do modelo phi
- **D_train:** amostras de treino da tabela

$$\mathbf{x}^m = \{x_j : m_j = 1\} \qquad \mathbf{x}^o = \{x_j : m_j = 0\}$$

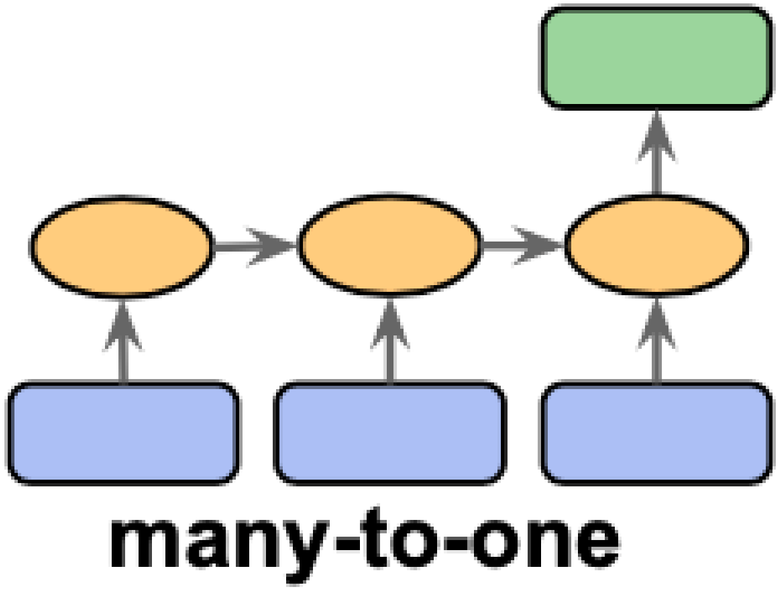
$$\min_{\theta \in \Theta} \sum_{\mathbf{x} \in \mathcal{D}_{train}} d\left(\mathbf{x}^m, \phi_\theta\left(\mathbf{x}^o \mid \mathbf{X}^O\right)\right).$$

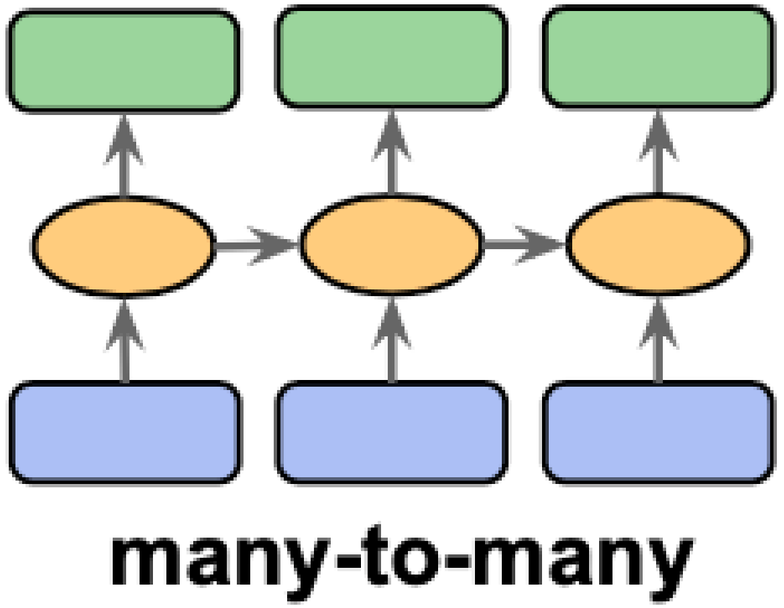
Modelo

03

Modelos de Processamento de sequências

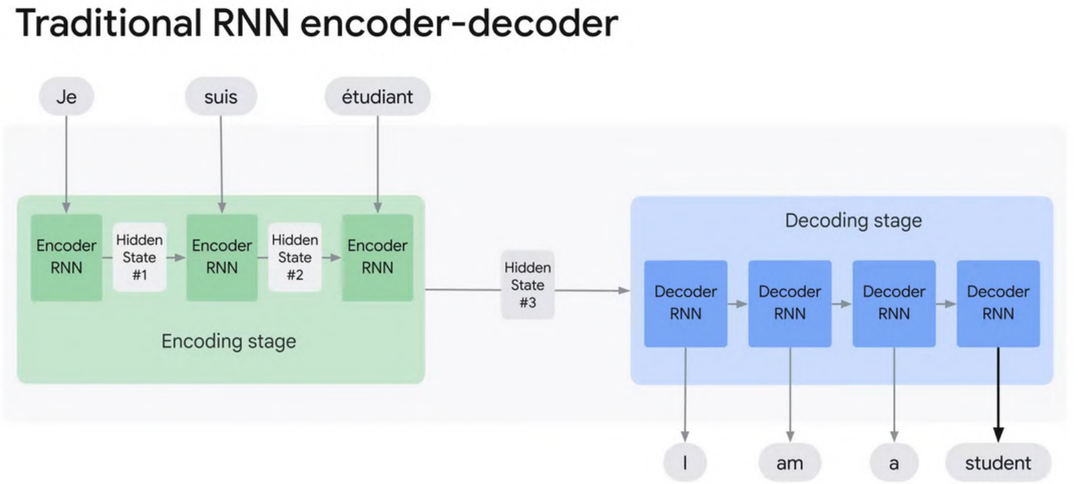
- Modelos genéricos tratam de dados de entrada e saída unitários.
- Modelos de processamento de sequências tratam de dados de entrada, saída ou ambos sequenciais:
  - Texto
  - Áudio
  - Vídeo
  - Macromoléculas sequenciais (DNA, proteínas)
- Nem sempre a sequência de entrada e de saída seguem a mesma ordem:
  - O CAVALO **Branco**
  - The **White** HORSE

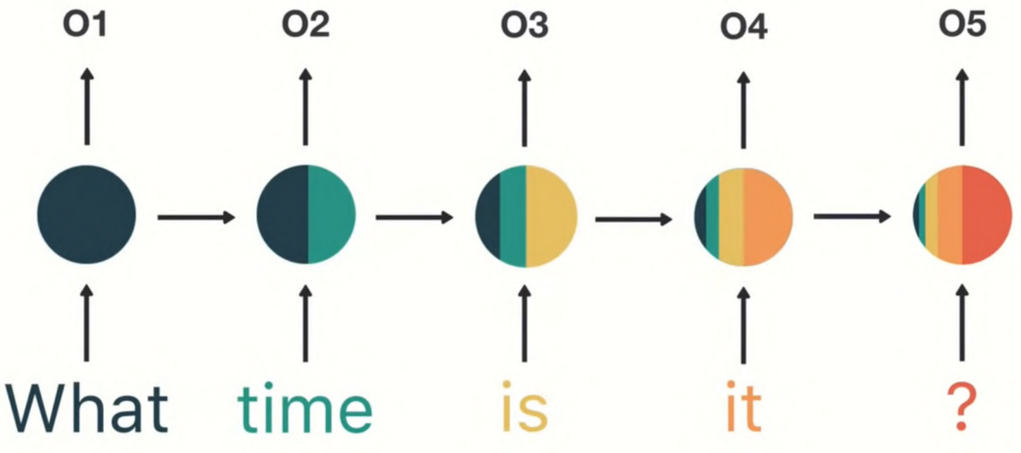


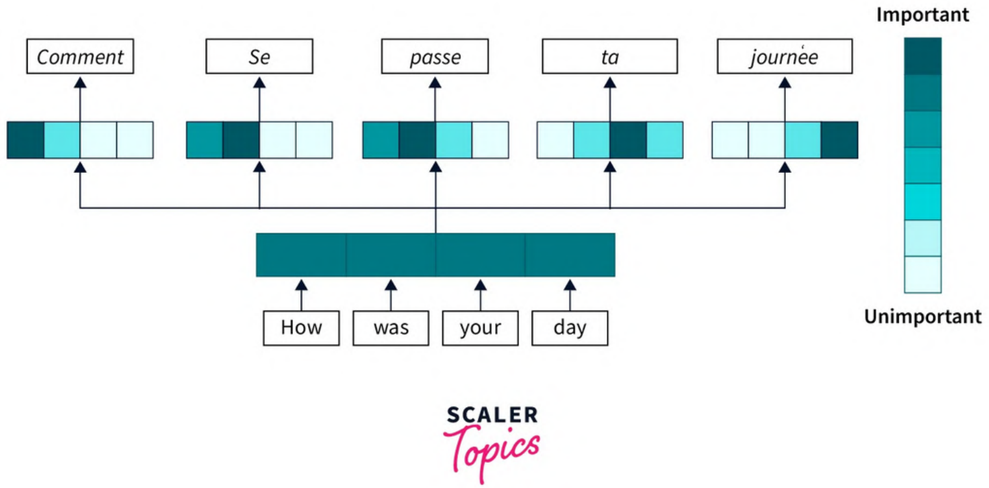


Soluções Anteriores: Redes Neurais Recorrentes

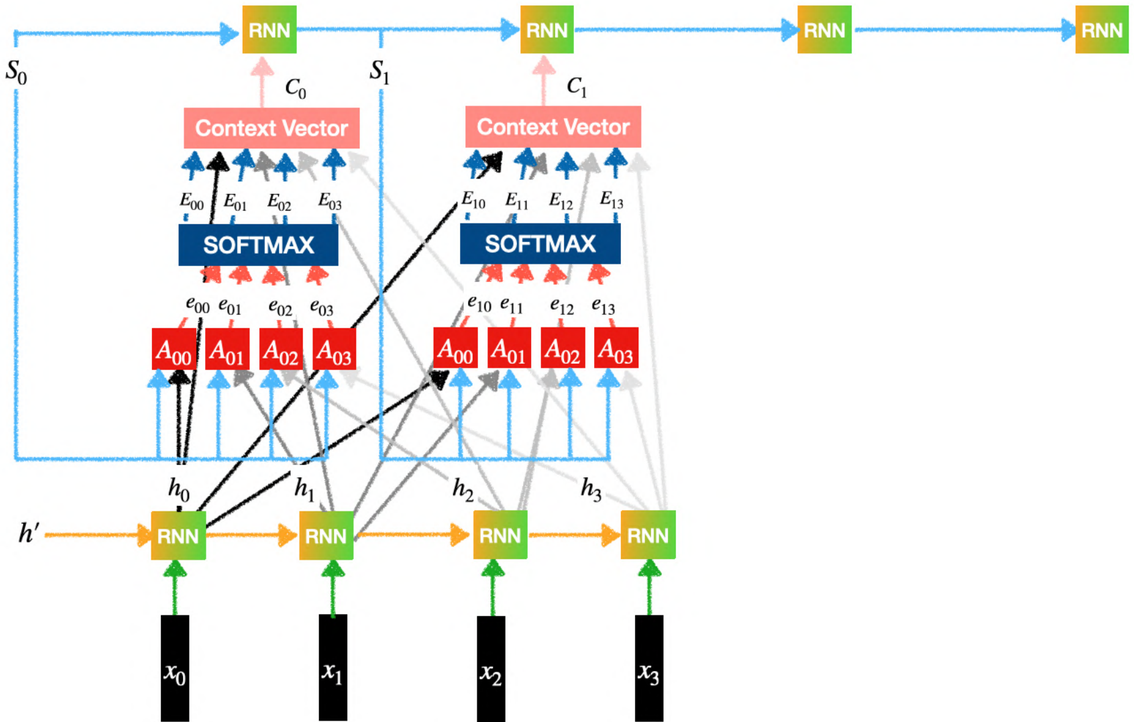
- Funciona como uma rede feed-forward, porém passada em ordem sobre em cada um dos dados da sequência.
- Possui um par de entradas e um par de saídas, de maneira unitária, como uma rede feed-forward
- Uma das entradas e uma das saídas correspondem a um estado oculto que representa a memória do modelo:
  - Ao passar pelo i-ésimo elemento, recebe como entrada a memória de quando passou pelo i-1
  - Ao passar pelo i-ésimo elemento, deixa a memória para quando passar pelo elemento i+1
- O significado da entrada e saída padrões depende da arquitetura.
- Problemas:
  - Vetor de contexto perde informação sobre as iterações mais antigas ao longo do tempo.
  - Tem desempenho ruim para sequências muito grandes.







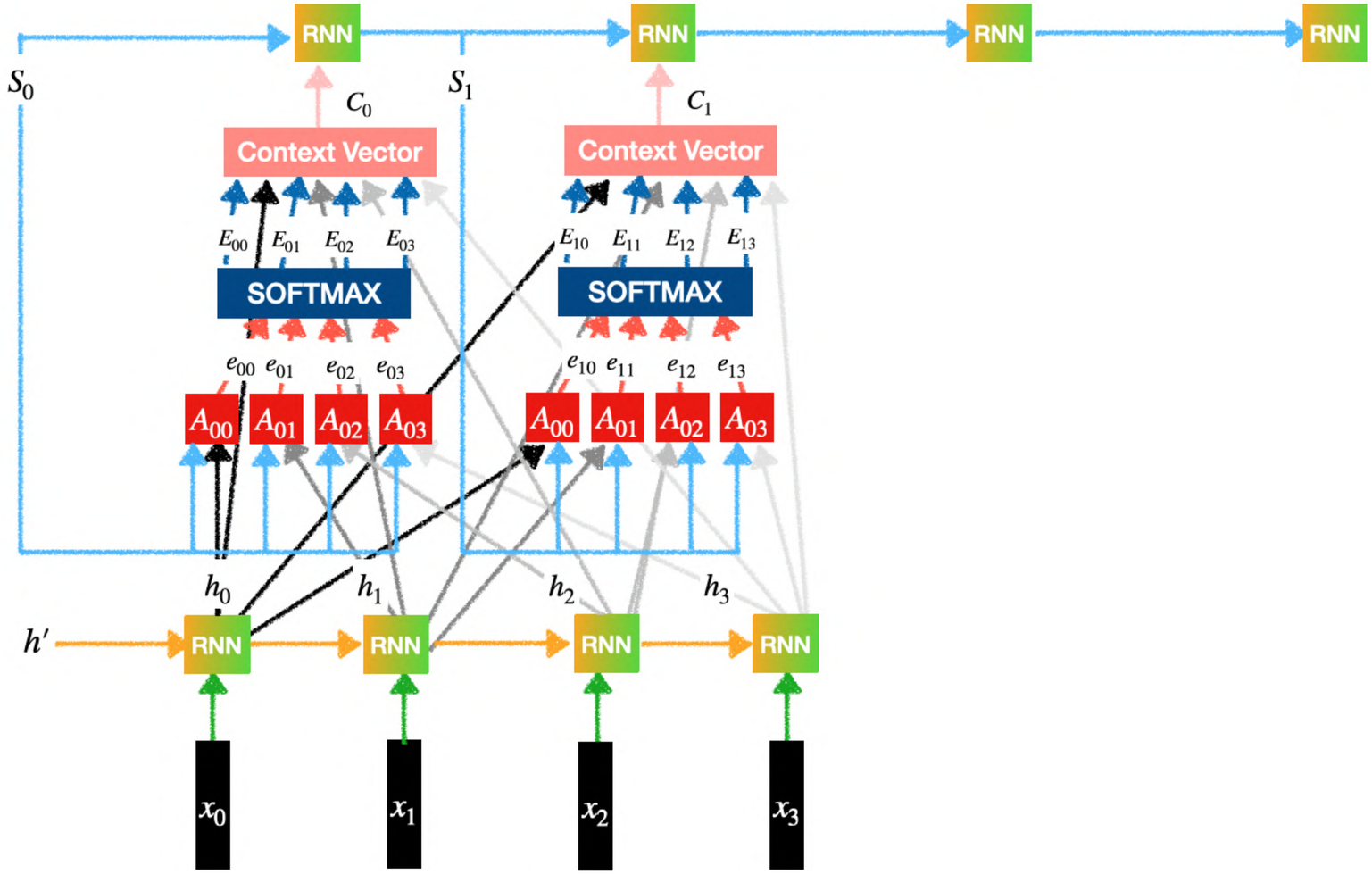
SCALER Topics



Transformer

- Uma versão melhorada da RNN
- Passa os vetores de contexto de todas as iterações do encoder para o decoder.
- Cada iteração do decoder decide quais vetores de contexto levar em consideração usando uma porcentagem de importância atribuída a cada vetor de contexto do encoder, denominada de mecanismo de atenção.
- Os mecanismos de atenção são definidos calculando a similaridade do vetor de contexto do decoder em uma iteração com os vetores de contexto de cada iteração do encoder e passando por uma softmax para retornar as porcentagens.
- A combinação dos vetores de contexto da entrada, ponderada pelas atenções, gera o vetor de contexto de saída do decoder em cada iteração.
- O vetor de contexto de saída, junto à saída da iteração anterior, geram a saída do decoder em cada iteração.

Transformer



Mecanismos de Atenção: Cálculo da Atenção

- Q[i]: Vetor Contextual da i-ésima saída
- K[i]: Vetor Contextual da i-ésima entrada
- V[i]: Valores da i-ésima entrada usados para o cálculo da atenção (usualmente igual a K[i]
- Attention: porcentagem de similaridade entre os vetores de K e cada vetor Q[i]
    - Vetores mais similares são mais influentes na saída Q[i]

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V}$$

Mecanismos de Atenção: Multi-Head Attention

- Múltiplos vetores de atenção diferentes são calculados e cada um divide um pedaço da influência.
- Resultados são combinados ao fim.
- Divisão da influência, combinação dos resultados são parâmetros de treino.

$$\text{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \underset{\text{axis=k}}{\text{concat}}(O_1, \ldots, O_k)W^O, \text{ where}$$

$$O_j = \text{Attention}(\mathbf{Q}W_j^Q, \mathbf{K}W_j^K, \mathbf{V}W_j^V)$$

Mecanismos de Atenção: Multi-Head Self-Attention

- Features de entrada e de saída são iguais.
- O Modelo aprende o padrão dos dados.

$$\mathrm{MHSelfAtt}(\mathbf{H}) = \mathrm{MultiHead}(\mathbf{Q} = \mathbf{H}, \mathbf{K} = \mathbf{H}, \mathbf{V} = \mathbf{H})$$

Mecanismos de Atenção: MHSA adaptada ao modelo

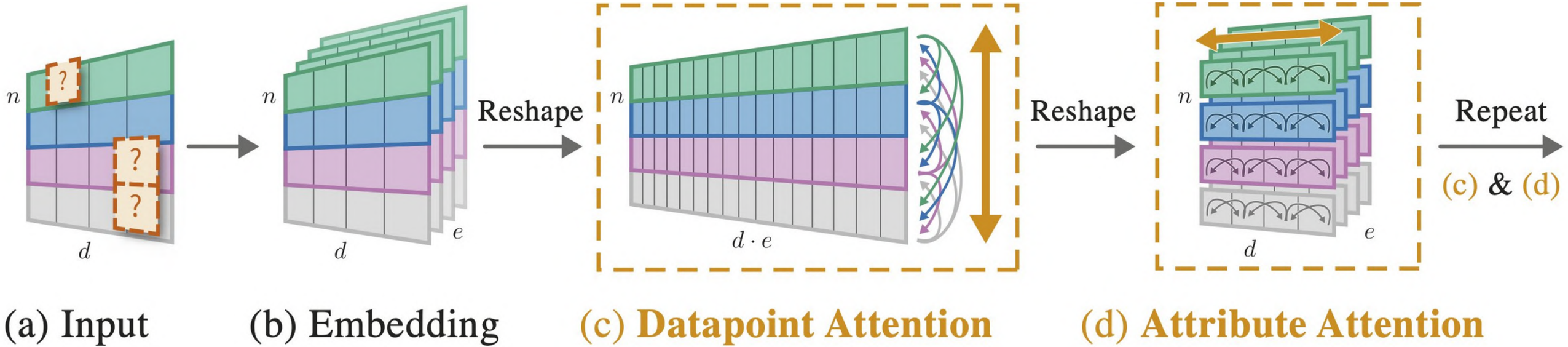
- Adaptações ao valor final da MHSA oriundas da definição e boas práticas dos modelos.
- Normalização de camada antes e depois do cálculo da atenção.
- Branch residual somado à atenção do modelo.
- Valor da predição anterior pela rede de saída (rFF) influencia a atenção final.

$$\mathbf{Res}(\mathbf{H}) = \mathbf{H}W^{\text{res}} + \mathbf{MHSelfAtt}\left(\mathbf{LN}(\mathbf{H})\right)$$

$$\mathbf{MHSA}(\mathbf{H}) = \mathbf{Res}(\mathbf{H}) + \mathbf{rFF}\left(\mathbf{LN}\left(\mathbf{Res}\left(\mathbf{H}\right)\right)\right) \in \mathbb{R}^{n \times h}$$

Non Parametric Transformer (NTP)

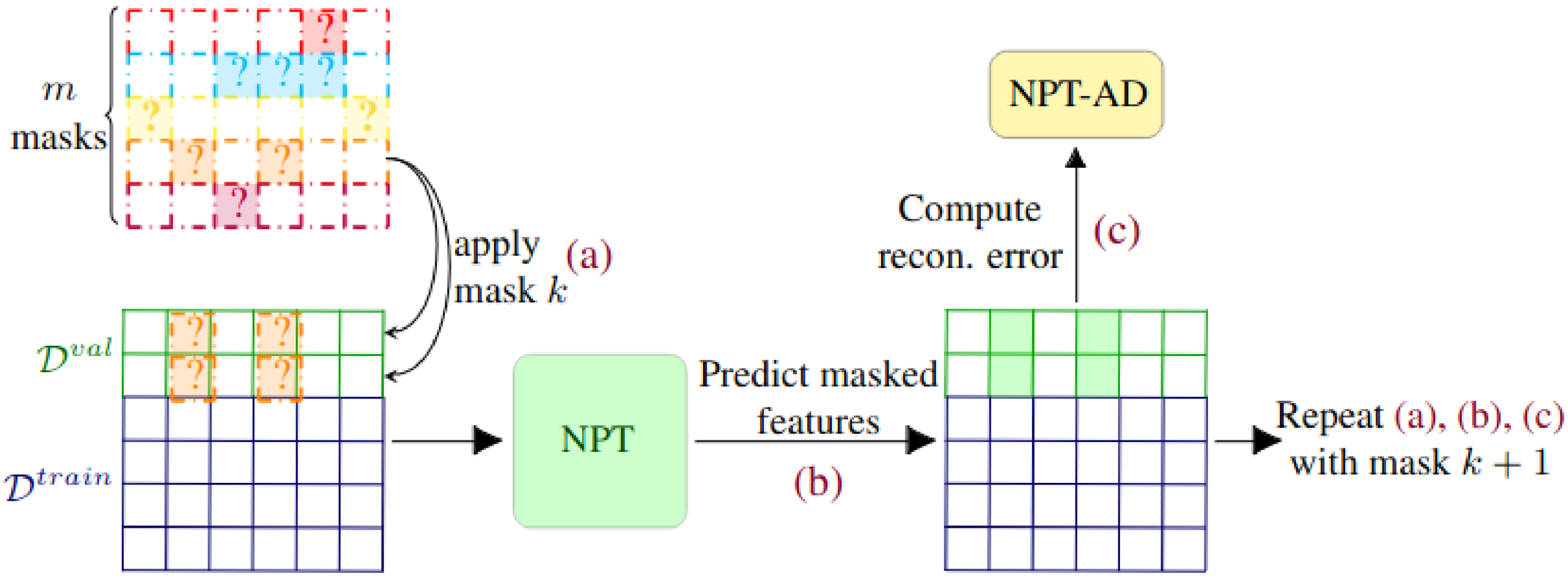
- Projetada para detectar padrões entre as amostras
- Utiliza uma arquitetura semelhante aos transformadores
  - Utiliza mecanismos de atenção MHSA
  - Faz embedding dos dados de entrada
- Ao invés de treinar a rede passando cada amostra, passa como entrada todas as amostras de uma vez.
- Em cada iteração, alterna entre usar a atenção entre as linhas (amostras) e as colunas (features) dos dados de entrada.



Modelo Final Proposto

- Para cada base de dados, é definido o treino e a validação:
  - O treino é constituído por metade das amostras originais
  - A validação possui a outra metade mais as amostras anômalas
- Para cada base, m máscaras são definidas deterministicamente:
  - É definido o número máximo r de features a serem mascaradas simultaneamente.
  - m é o total de combinações de features escolhendo de 1 a r.
  - Vetores unidimensionais com cada combinação possível de features dentro do limite estipulado são definidos como máscaras.
  - A mesma máscara é aplicada em todas as amostras da validação.
- A função de perda é a anomaly score da predição gerada:
  - Para cada amostra, seu valor é comparado com os resultados originais da tabela:
    - erro quadrado médio para valores numéricos
    - Cross-entropy para valores categóricos
- A classificação da amostra como anômala é feita com um T-teste do score da amostra contra os scores da tabela com limiar de 5%.

Modelo Final Proposto



Experimentos

04

Bases de dados utilizadas

Bases médicas

- 2 bases menores:
  - Arrythmia
  - Thyroid

Bases de cibersegurança

- 2 bases maiores, com informações de tráfego na internet:
  - KDD
  - KDDRev

Bases estruturadas para detecção de anomalias

- 28 bases selecionadas para problemas de detecção de anomalias.
  - Wine
  - Lympho
  - Vowels
  - ...

Method	DROCC (abalone)	GOAD (thyroid)	NeuTraL-AD (arrhy.)	Internal Cont.	NPT-AD
Wine	63.0±20.0	67.0±9.4	78.2±4.5	**90.0**±6.3	72.5±7.7
Lympho	65.0±5.0	68.3±13.0	20.0±18.7	86.7±6.0	**94.2**±7.9
Glass	14.5±11.1	12.7±3.9	9.0±4.4	**27.2**±10.6	**26.2**±10.9
Vertebral	**9.3**±6.1	**16.3**±9.6	3.8±1.2	**26.0**±7.7	20.3±4.8
Wbc	9.0±6.2	**66.2**±2.9	60.9±5.6	**67.6**±3.6	**67.3**±1.7
Ecoli	N/A	61.4±31.7	7.0±7.1	70.0±7.8	**77.7**±0.1
Ionosph.	76.9±2.8	83.4±2.6	90.6±2.4	**93.2**±1.3	**92.7**±0.6
Arrhyth.	37.1±6.8	52.0±2.3	59.5±2.6	**61.8**±1.8	60.4±1.4
Breastw	93.0±3.7	**96.0**±0.6	91.8±1.3	**96.1**±0.7	95.7±0.3
Pima	66.0±4.1	66.0±3.1	60.3±1.4	59.1±2.2	**68.8**±0.6
Vowels	66.2±8.8	31.1±4.2	10.0±6.2	**90.8**±1.6	88.7±1.6
Letter	55.6±3.6	20.7±1.7	5.7±0.8	62.8±2.4	**71.4**±1.9
Cardio	49.8±3.2	**78.6**±2.5	45.5±4.3	71.0±2.4	**78.1**±0.1
Seismic	19.1±0.9	24.1±1.0	11.8±4.3	20.7±1.9	**26.2**±0.7
Musk	99.4±1.5	**100.0**±0.0	99.0±0.0	**100.0**±0.0	**100.0**±0.0
Speech	4.3±2.0	4.8±2.3	4.7±1.4	5.2±1.2	**9.3**±0.8
Thyroid	72.7±3.1	72.5±2.8	69.4±1.4	76.8±1.2	**77.0**±0.6
Abalone	17.9±1.3	57.6±2.2	53.2±4.0	**68.7**±2.3	59.7±0.1
Optdigits	30.5±5.2	0.3±0.3	16.2±7.3	**66.3**±10.1	**62.0**±2.7
Satimage2	4.8±1.6	90.7±0.7	92.3±1.9	92.4±0.7	**94.8**±0.8
Satellite	52.2±1.5	64.2±0.8	71.6±0.6	73.2±1.6	**74.6**±0.7
Pendigits	11.0±2.6	40.1±5.0	69.8±8.7	82.3±4.5	**92.5**±1.3
Annthyr.	64.2±3.3	50.3±6.3	44.1±2.3	45.4±1.8	57.7±0.6
Mnist	N/A	66.9±1.3	84.8±0.5	85.9±0.0	**71.8**±0.3
Mammo.	32.6±2.1	33.7±6.1	19.2±2.4	29.4±1.4	**43.6**±0.5
Shuttle	N/A	73.5±5.1	97.9±0.2	**98.4**±0.1	**98.2**±0.3
Mullcross	N/A	99.7±0.8	96.3±10.5	**100.0**±0	**100.0**±0
Forest	N/A	0.1±0.2	51.6±8.2	44.0±4.1	**58.0**±10
Kdd	N/A	79.6±3.9	96.9±2.0	**99.4**±0.1	98.7±0.3
Kddrev	N/A	98.0±0.1	96.5±1.5	**99.2**±0.3	98.5±0.1
mean	33.6	55.9	53.9	69.7	**71.2**
mean std	4.6	4.2	3.9	2.9	**2.0**
mean rank	10.7	7.6	9.0	3.2	**3.0**

Resultados

- Cada modelo foi executado para cada base em 20 repetições (10 para as bases maiores como as KDD).
- O F1 e a AUROC (Area Under the Receiver Operating Characteristic curve) foram coletados para cada instância e a média e desvio padrão das repetições foi calculada.
- O F1 médio, o desvio médio e o ranking médio de cada modelo foram coletados (ranking n corresponde ao n-ésimo melhor modelo para uma base)
- NPT-AD teve melhor desempenho geral, e apresentou a melhor solução para quase todas as bases

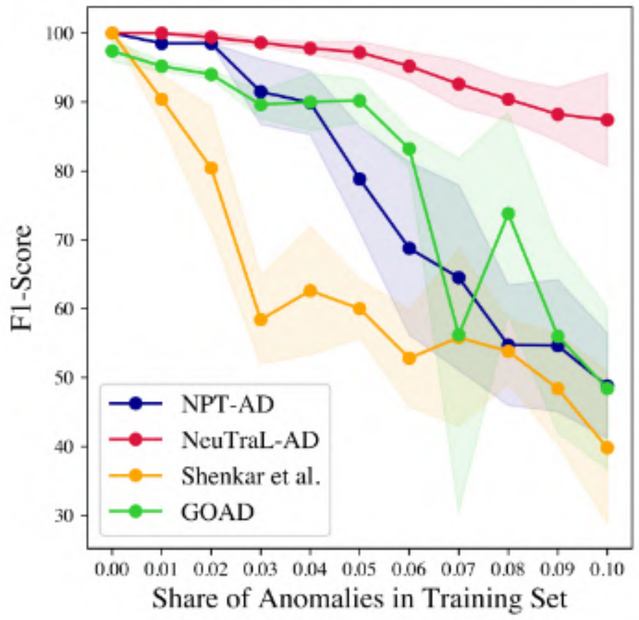
Method	COPOD	IForest	KNN	PIDForest	RRCF	NPT-AD
Wine	60.0±4.5	64.0±12.8	**94.0**±4.9	50.0±6.4	69.0±11.4	72.5±7.3
Lympho	85.0±5.0	71.7±7.6	80.0±11.7	70.0±0.0	36.7±18.0	**94.2**±7.9
Glass	11.1±0.0	11.1±0.0	11.1±9.7	8.9±6.0	15.6±13.3	**26.2**±10.9
Vertebral	1.7±1.7	13.0±3.8	10.0±4.5	12.0±5.2	8.0±4.8	**20.3**±4.8
Wbc	**71.4**±0.0	70.0±3.7	63.8±2.3	65.7±3.7	54.8±6.1	67.3±1.7
Ecoli	25.6±11.2	58.9±22.2	**77.8**±3.3	25.6±11.2	28.9±11.3	**77.7**±0.1
Ionosphere	70.8±1.8	80.8±2.1	88.6±1.6	67.1±3.9	72.0±1.8	**92.7**±0.6
Arrhythmia	58.2±1.4	60.9±3.3	**61.8**±2.2	22.7±2.5	50.6±3.3	60.4±1.4
Breastw	96.4±0.6	**97.2**±0.5	96.0±0.7	70.6±7.6	63.0±1.8	95.7±0.3
Pima	62.3±1.1	**69.6**±1.2	65.3±1.0	65.9±2.9	55.4±1.7	68.8±0.6
Vowels	4.8±1.0	25.8±4.7	64.4±3.7	23.2±3.2	18.0±4.6	**88.7**±1.6
Letter	12.9±0.7	15.6±3.3	45.0±2.6	14.2±2.3	17.4±2.2	**71.4**±1.9
Cardio	65.0±1.4	73.5±4.1	67.6±0.9	43.0±2.5	43.9±2.7	**78.1**±0.1
Seismic	29.2±1.3	**73.9**±1.5	30.6±1.4	29.2±1.6	24.1±3.2	26.2±0.7
Musk	49.6±1.2	52.0±15.3	**100.0**±0.0	35.4±0.0	38.4±6.5	**100**±0.0
Speech	3.3±0.0	4.9±1.9	5.1±1.0	2.0±1.9	3.9±2.8	**9.3**±0.8
Thyroid	30.8±0.5	**78.9**±2.7	57.3±1.3	72.0±3.2	31.9±4.7	77.0±0.6
Abalone	50.3±6.4	53.4±1.7	43.4±4.8	58.6±1.6	36.9±6.4	**59.7**±0.1
Optdigits	3.0±0.3	15.8±4.3	**90.0**±1.2	22.5±16.8	1.3±0.7	62.0±2.7
Satimage2	77.9±0.9	86.5±1.7	93.8±1.2	35.5±0.4	47.9±3.4	**94.8**±0.8
Satellite	56.7±0.2	69.6±0.5	**76.3**±0.4	46.9±3.7	55.4±1.3	74.6±0.7
Pendigits	34.9±0.6	52.1±6.4	91.0±1.4	44.6±5.3	16.3±2.6	**92.5**±1.3
Annthyroid	31.5±0.5	57.3±1.3	37.8±0.6	**65.4**±2.7	32.1±0.8	57.7±0.6
Mnist	38.5±0.4	51.2±2.5	69.4±0.9	32.6±5.7	33.5±1.7	**71.8**±0.3
Mammo.	**53.4**±0.9	39.0±3.3	38.8±1.5	28.1±4.3	27.1±1.9	43.6±0.5
Shuttle	96.0±0.0	96.4±0.8	97.3±0.2	70.7±1.0	32.0±2.2	**98.2**±0.3
Mullcross	66.0±0.1	99.1±0.5	**100.0**±0.0	67.4±2.1	**100.0**±0.0	**100.0**±0.0
Forest	18.2±0.2	11.1±1.6	**92.1**±0.3	8.1±2.8	9.9±1.5	58.0±10.0
Kdd	44.5±0.1	95.6±2.2	**98.9**±0.4	92.1±2.2	74.7±0.9	**98.7**±0.3
Kdd-rev	30.9±0.4	96.4±2.4	85.2±0.3	50.9±4.1	9.8±1.2	**98.5**±0.1
mean	44.7	58.2	67.7	43.4	37.0	**71.2**
mean std	**1.5**	4.0	2.2	3.9	4.2	2.0
mean rank	10.1	6.8	5.1	10.8	11.9	**3.0**

Resultados

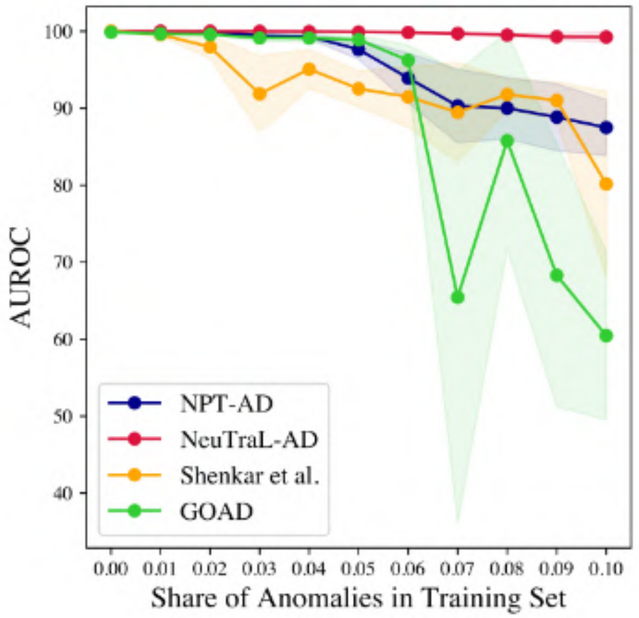
- Mesmos testes foram feitos para modelos não-baseados em deep-learning.
- O NPT-AD também teve um desempenho geral melhor que todas as demais soluções.
- Só perdeu em desvio padrão, mas a diferença no desvio é muito menor que a diferença positiva na média.

06

Discussões



(a) F1-Score (↑)



(b) AUROC (↑)

Dados de treino contaminados

- Definir uma base de treino robusta e livre de anomalias é difícil em um cenário real.
- Teste do modelo para dados de treino contaminados:
  - Uma base de treino limpa de 900 amostras foi definida, com 100 amostras anômalas separadas.
  - Amostras variando entre 1 e 10% de contaminação foram utilizadas para treino do modelo.
  - F1 e AUROC (area under the receiving operator characteristic curve) coletadas
- Quedas de performance significativas a partir de 2% de contaminação

Impacto de dependências amostra-amostra

- O objetivo final do modelo era ser capaz de detectar dependências amostra-amostra.
- Para isso, o modelo já treinado foi executado 20 vezes com colunas permutadas aleatoriamente (apenas dependência entre amostras impactaria o resultado).
- Queda na F1 e na AUROC em relação ao modelo original coletadas.
- Maiores quedas na F1: AUROC leva pouco em consideração dependências amostra-amostra

F1(T_random) - F1(T)

	Mammo.	Glass	BreastW	Pendigits
$\Delta F1$	-1.0	-9.6	-0.5	-2.8
ΔAUROC	-0.1	-0.1	-0.1	-0.1

AUROC(T_random) - AUROC(T)

07

Limitações e Conclusão

Limitações

- Modelo altamente complexo
- Requer poder computacional muito alto
  - Complexidade do NPT
  - Complexidade da função de score da anomalia
- Por esse motivo, pode não ser adequado para bases com muitas features

Conclusão

- O modelo proposto é o primeiro na literatura a utilizar dependências feature-feature e amostra-amostra para a detecção de anomalias.

- Utilizando uma extensa quantidade de dados, o modelo teve desempenho melhor que técnicas estado-da-arte para métricas de F1 e AUROC.

- Os experimentos demonstram a robustez do método para dados de treino contaminados

- Com este trabalho, é enfatizada a importância de considerar dependências amostra-amostra para a detecção de anomalias em tabelas.

Obrigado!

Bibliografia

Attention mechanism: Overview: https://www.youtube.com/watch?v=fjJOgb-E41w&ab_channel=GoogleCloudTech

Attention Mechanism in a Nutshell: https://www.youtube.com/watch?v=oMeIDqRguLY&ab_channel=HalflingWizard

Illustrated Guide to Recurrent Neural Networks: Understanding the Intuition: https://www.youtube.com/watch?v=LHXXI4-IEns&t=339s&ab_channel=TheA.I.Hacker-MichaelPhi

Kossen J et al., Self-Attention Between Datapoints: Going Beyond Individual Input-Output Pairs in Deep Learning

Thimonier H. et al, Beyond Individual Input for Deep Anomaly Detection on Tabular Data

Vaswani A. et al, Attention Is All You Need