COMPARATIVO ENTRE AS ABORDAGENS PARA ROTULAÇÃO DE DADOS - ACTIVE LEARNING X PASSIVE LEARNING

Raôny Magnago Traspadini[1] , Gabriel Tozatto Zago[2]
[1,2]Instituto Federal do Espírito Santo, Serra – ES, Brasil
E-mail: raonytraspadini@gmail.com

Resumo – O custo de aquisição de grandes bancos de dados de qualidade é alto, haja visto o tempo de processamento ou alocação de mão de obra qualificada para rotulagem. Diante disso, é apresentado um comparativo entre os métodos de rotulagem de dados – Aprendizagem Passiva e Ativa. O banco de dados utilizado trata de classificação de doença cardíaca, obtido através do *Kaggle*, e o modelo utilizado na análise foi o *Support Vector Classification*. O aprendizado ativo alcançou a acurácia de 82% com apenas 40 amostras, ao passo que a abordagem passiva atingiu a mesma marca com cerca de 110 amostras, mas apresentando comportamento instável e, por vezes, com déficit de acurácia. Dessa forma, a rotulagem de dados utilizando o aprendizado ativo é benéfica visto que o tempo de processamento é menor e a performance esperada é obtida, o que consequentemente traz redução de custos.

***Palavras-Chave* – Aprendizado Ativo, Rotulagem de dados, Machine Learning, Classificação.**

COMPARISON BETWEEN THE LABELING DATA APPROACHES – ACTIVE LEARNING X PASSIVE LEARNING

Abstract – The acquisition cost of large quality databases is high, given the processing time or qualified labor allocation for labelling. Therefore, the present work shows a comparison between the labelling data methods – Passive Learning vs Active Learning. The used database handles of Heart Failure Classification, obtained from Kaggle, and the utilized model was the Support Vector Classification. The active learning reached 82% of accuracy with only 40 samples, while the passive approach reached the same point with 110 samples, although showing an unstable behavior and, sometimes, with accuracy deficit. Thus, the labelling data by the active learning method is beneficial, due the lower processing time and the reached expected performance, consequently bringing cost reduction.

***Keywords* – Active Learning, labelling data, Machine Learning, Classification.**

I. INTRODUÇÃO

Obter grandes conjuntos de dados devidamente rotulados e organizados é uma tarefa que demanda tempo e, na maioria das vezes, alto custo. Como cita [1], a aquisição de um grande conjunto de dados rotulados e de qualidade irá exigir muita mão de obra, especialmente em áreas que demandam altos níveis de expertise e conhecimento, como reconhecimento de fala, extração de informações e imagens médicas.

Para a rotulagem desses dados podem ser citadas duas técnicas: *Passive Learning* e o *Active Learning*. A primeira abordagem, no entanto, procura apenas reunir a maior quantidade possível de instâncias rotuladas, não considerando nenhuma medida de informatividade destas [4], o que geralmente irá acarretar maior tempo e custo.

Já a segunda abordagem - *Active Learning*, conhecido como uma subárea do Machine Learning, conforme cita [2], trata-se de uma técnica seletiva, onde o algoritmo seleciona somente um subconjunto dos dados a serem rotulados ao invés de todo o conjunto.

Assim, busca-se apresentar um comparativo entre os modos de rotulagem de dados apresentados acima, utilizando como métrica a acurácia média obtida em cada técnica.

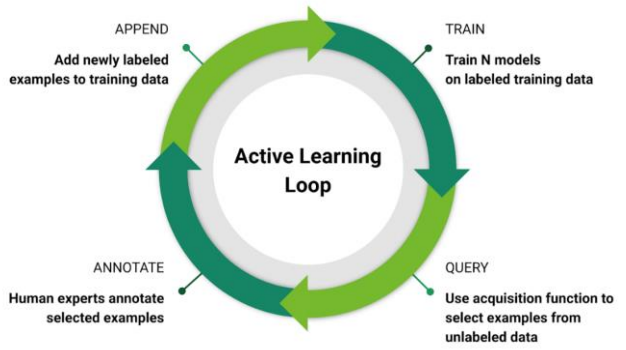
II.ACTIVE LEARNING

A fim de reduzir os custos com rotulagem de dados e conferir agilidade aos processos de Machine Learning, a técnica de aprendizado ativo pode ser utilizada nos problemas de classificação.

O algoritmo, diferentemente do aprendizado passivo, irá selecionar do banco de dados desconhecido um número menor de amostras para serem rotuladas. Porém, a seleção é baseada, por exemplo, em pontos próximos ao limiar de decisão do modelo, onde normalmente há dúvida para definir a qual classe aquele ponto pertence [2]. Em outras palavras, o algoritmo assume que as amostras possuem pesos diferentes para atualizar o treinamento do modelo e tenta selecionar aquelas que mais contribuirão para construir o conjunto de treinamento. Dessa forma, o *Active Learning* é um meio eficaz para garantir que o tempo não está sendo desperdiçado na tarefa de rotular dados não conhecidos e, por consequência, demandando custos menores [3].

Ainda citado por [3], escolher amostras para serem rotuladas de forma aleatória pode resultar em modelos com baixa acurácia na predição, como é feito nos métodos de aprendizagem passiva. A **Erro! Fonte de referência não encontrada.** mostra o ciclo de funcionamento do aprendizado ativo.

Fig. 1. Ciclo Active Learning



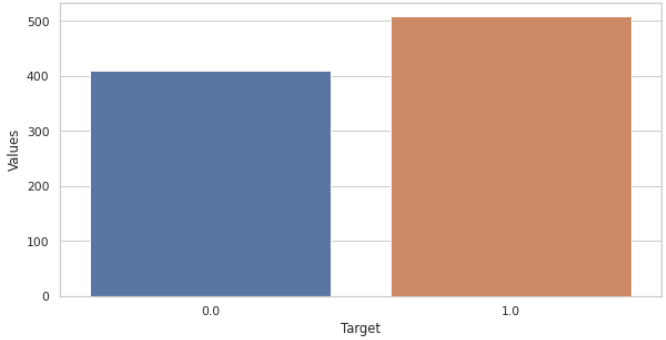
Fonte: Scalable Active Learning for Autonomous Driving: A Practical Implementation and A/B Test, NVIDIA AI 2019

- O treinamento se inicia a partir de um banco de dados onde pouquíssimas amostras são rotuladas;
- No *"Query"*, serão identificadas amostras que carregam maior dúvida sobre sua classificação;
- Essas amostras serão enviadas ao anotador, onde receberão rótulos;
- Os novos dados rotulados são adicionados ao banco de dados e usados pelo modelo para treinar novamente;
- O *loop* ocorre até que o desempenho desejado seja alcançado.

III. CONJUNTO DE DADOS

O banco de dados utilizado para comparar as abordagens é público e trata da predição de doença cardíaca, podendo ser encontrado na base de dados do *Kaggle* [5]**.** O problema de classificação binária baseia-se em características do estado de saúde e vida de vários indivíduos, como idade, sexo, se possuiu algum desconforto ou dor no peito, pressão do sangue, nível de colesterol, frequência cardíaca e entre outros. Os dados mostraram-se balanceados, conforme mostra a Fig. 2, totalizando 918 registros.

Fig. 2: Distribuição da Target no banco de dados



Fonte: Autor.

IV. METODOLOGIA

A abordagem utilizada para exemplificar a técnica de aprendizagem ativo é conhecida como *Uncertainty Sampling*, isto é, são identificados elementos próximos do limiar de decisão, onde não há exatidão sobre a classificação dos dados e, por consequência, são mais propensos a serem classificados erroneamente. O modelo então é retreinado utilizando esses dados [6].

**Processo no qual variáveis categóricas são convertidas em um formato que serão melhor interpretadas pelo algoritmo de ML*

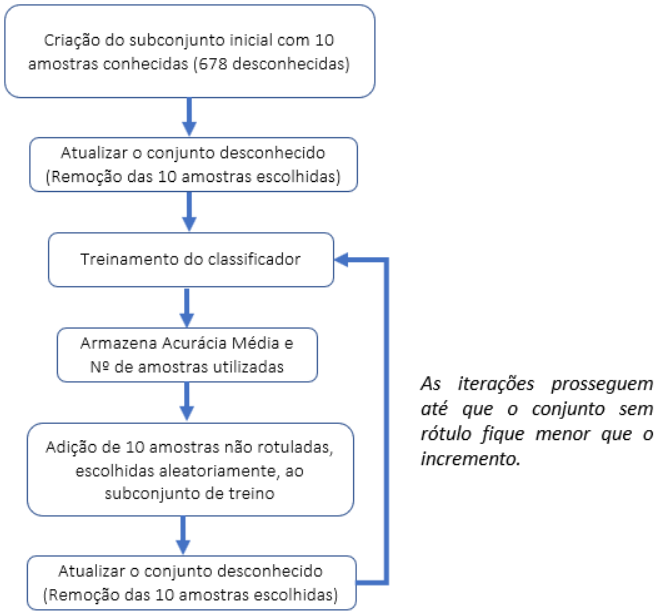
A. *Algoritmo Utilizado*

A tratativa dos dados deu-se, inicialmente, a partir da necessidade de transformar variáveis categóricas em numéricas utilizando *One Hot Encoding**, visto que alguns modelos de aprendizagem podem não suportar valores categóricos.
O segundo passo foi normalizar o banco de dados, a fim de introduzir alguma sequência lógica interpretável ao modelo.
Os dados foram divididos da seguinte maneira: 75% em treino (688 amostras) e 25% em teste (230 amostras), utilizando o balanceamento dos dados pela target a ser predita. Além disso, em ambas as abordagens o *Support Vector Classification* foi o modelo de classificação utilizado, inicializando o treinamento com um subconjunto de 10 amostras retiradas aleatoriamente dos dados de treino.
A cada iteração eram incrementadas ao subconjunto 10 amostras obtidas do conjunto de treino, nunca vistas antes pelo modelo (não rotuladas). Após cada etapa a acurácia era observada e os dados acrescentados ao subconjunto eram retirados do conjunto geral.
Cabe ressaltar que, apesar do *dataset* apresentar dados rotulados, a construção do algoritmo ignora essa informação e, consequentemente, os novos dados a serem identificados são desconhecidos (sem rótulos). O código em linguagem Python utilizado pode ser conferido através do link: https://github.com/RaonyTraspadini/-ActiveLearning-HeartFailure.

B. *Passive Learning*

O passo a passo geral utilizado é mostrado no diagrama da Fig. 3 abaixo.

Fig. 3. Fluxograma seguido pela abordagem passiva

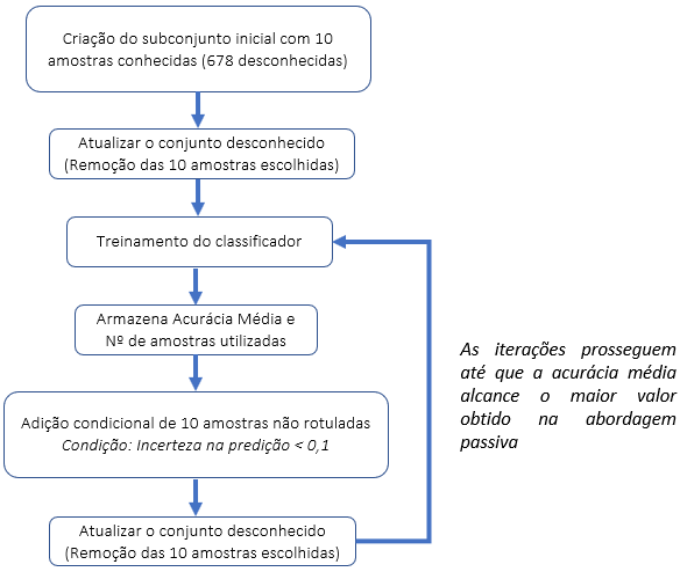


Fonte: Autor.

C. *Active Learning*

O passo a passo geral utilizado é mostrado no diagrama da Fig. 4 abaixo.

Fig. 4. Fluxograma seguido pela abordagem ativa



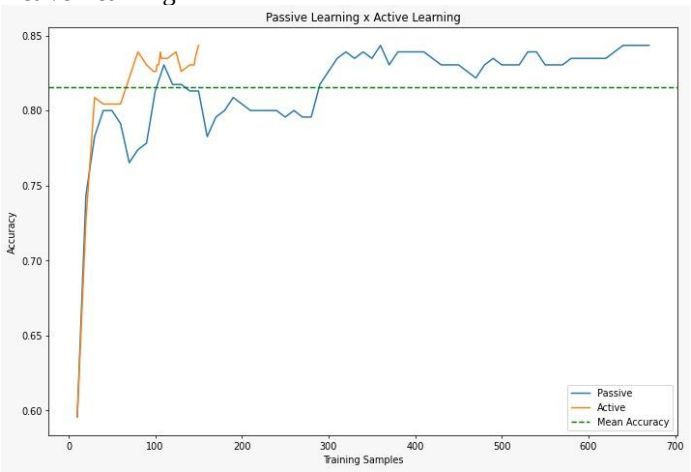
Fonte: Autor.

A condição para inserir 10 novas amostras ao subconjunto de treino é obtida através do *predict_proba*, um método presente em alguns modelos de classificação da biblioteca S*cikit Learn* do Python, que irá retornar a probabilidade das saídas para cada amostra. Dessa forma, serão selecionadas 10 amostras cuja probabilidade de predição estiver entre 0,4 e 0,6 (40% < p < 60%).

V. RESULTADOS

A acurácia media dos modelos utilizados está em torno de 82%, valor que foi alcançado pela abordagem ativa com 40 amostras (Fig. 5), cerca de 5% do banco de dados.
Observa-se que o modo passivo alcançou a acurácia média com cerca de 110 amostras. No entanto, essa acurácia se manteve instável e por vezes decaiu, o que não é observado no modo ativo, onde a acurácia possui a tendência de se manter acima da acurácia média.
Cabe ressaltar que, no início, as curvas se assemelham pois os dados iniciais tomados para treino são iguais, representando o primeiro passo do ciclo na Fig. 1, no qual as primeiras amostras rotuladas são integradas ao treinamento do modelo.

Fig. 5. Comparativo de performance do modelo - Passive x Active Learning



Fonte: Autor.

VI. CONCLUSÃO

O tempo de processamento utilizando a abordagem ativa é, consequentemente, muito menor em relação à abordagem passiva, visto que a acurácia máxima pode ser alcançada com o menor número de amostras, trazendo ganhos em produtividade às equipes de análise de dados.
Em estudos futuros, sugere-se analisar em bancos de dados maiores, permitindo um maior incremento de dados a cada iteração, o que possivelmente acarretará maiores discrepâncias entre as duas abordagens apresentadas.

REFERÊNCIAS

[1] REN, Pengzhen et al. A survey of deep active learning. ACM Computing Surveys (CSUR), v. 54, n. 9, p. 1-40, 2021.

[2] JANVEKAR A., Naveed**.** Active Learning and Its Benefits to Machine Learning Models. DATA TOPICS. Disponível em: https://www.dataversity.net/active-learning-and-its-benefits-to-machine-learning-models/. Acesso em: 16 de jun. 2022.

[3] FREDRIKSSON, Teodor et al. Data labeling: An empirical investigation into industrial challenges and mitigation strategies. In: International Conference on Product-Focused Software Process Improvement. Springer, Cham, 2020. p. 202-216.

[4] ACORDI, Marcelo. Seleção de amostras de dados menos representativas usando aprendizado ativo. 2021.

[5] KAGGLE. Heart Failure Prediction Dataset. Disponível em: https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction. Acesso em: 03 de jun. 2022.

[6] MOSQUEIRA-REY, Eduardo; ALONSO-RÍOS, David; BAAMONDE-LOZANO, Andrés. Integrating iterative machine teaching and active learning into the machine learning loop. **Procedia Computer Science**, v. 192, p. 553-562, 2021.