Desenvolvimento De Uma Programa Para Avaliar Os Resultados Referente A Utilização De Aprendizado Ativo Comparadas A Aprendizado Passivo

1st Gustavo Gregório Fernandes Santos
Dept. Engenharia de Controle e Automacão
Instituto Federal do Espírito Santo
Serra, Brasil
gustavo.gregorios@gmail.com

2nd Gabriel Tozatto Zago
Dept. Engenharia de Controle e Automacão
Instituto Federal do Espírito Santo
Serra, Brazil

***Resumo*—O aprendizado de máquina por muitas vezes pensado por ser um conceito novo, por sua vês tem cerca de 70 anos, uma ideia em que em seus primórdios a sua evolução era limitada devido a velocidade dos computadores da época e a pouca quantidade de dados, porém a tecnologia evoluiu muito em pouco tempo, e quantidade e a complexidade das informações também, com isso com isso, novos conceitos e ferramentas também foram implementados, criadas metodologias no qual tem a função de aprimorar o desempenho dos modelos de aprendizado de máquina, é nessa vertente que esse artigo se curva, nele é discorrido o desenvolvimento de um programa que utiliza um dos métodos mais elegantes de extração de informações, o aprendizado ativo, ou do inglês Active Learning, quando utilizado em um algoritmo de aprendizado de maquina de forma correta pode obter maior precisão com menos amostras, por conseguinte gastando menos tempo de processamento e evitando a aquisição de grandes quantidades de dados para treinamento.**

***Palavras Chave*—Aprendizado de máquina. Aprendizado passivo. Aprendizado Ativo.**

I. INTRODUÇÃO

O objetivo principal dos algoritmos de aprendizado de máquina é fazer com que os computadores possam extrair informações dos dados fornecidos e ser capazes de desenvolver modelos gerais que podem ser utilizados para preverem possiveis problemas. Os dados fornecidos podem representar diferentes tipos de problemas, e os algoritmos de aprendizado de maquina tentam extrair informações suficientes para criar um modelo matemático capaz de responder adequadamente aos estímulos de entrada [1].

O aprendizado de maquina consiste em induzir uma função f(x,Z, w) = fg(x), ou seja, obter um modelo matemático que melhor represente alguma função oráculo. A função oráculo é a função que representa a resposta de um problema dado algum estímulo x [1].

Os métodos de Aprendizado de Máquina geralmente são classificados nos seguintes grupos: Aprendizado Supervisionado, Aprendizado Não Supervisionado e Aprendizado Semi-supervisionado. Os algoritmos desses grupos podem ainda ser subclassificados em mais dois metodos de aprendizagem o aprendizado passivo e aprendizado Ativo [1].

Dentre os diferentes contextos e tipos de aprendizados, o aprendizado supervisionado é um dos mais comuns da área, tendo como principal enfoque a predição de um rótulo (variável dependente) com base em um conjunto de atributos (ou variáveis independentes) através de uma função [2].

Para uma boa performance de algoritmos de aprendizado supervisionado por muitas vezes está atrelado ao treinamento com um grande número de dados rotulados [2].

Porem quando há grande quantidade de dados não rotulados e o processo de rotulação tem um custo elevado, inviabilizando que todos os padrões sejam rotulados. Esse custo pode ser de natureza econômica, fadiga do especialista, tempo, dentre outros [2]

Tendo em vista a problemática mencionada acima, a utilização da metodologia de aprendizado ativo é extremamente indicado, visto que o aprendizado ativo se caracteriza basicamente por, a partir de diferentes métodos, escolher de maneira inteligente com algum tipo de heurística quais observações incluir no treinamento do modelo de forma que este tenha um bom desempenho com um conjunto de treinamento menor. Em outras palavras ao invés de sortearmos alguma nova observação para ser treinada (aprendizado passivo), escolhemos uma instância mais estratégica com base em diferentes medidas de informatividade para selecionar os dados a serem treinados [2]. Observe na figura 1 e 2 a estrutura basica desses dois tipos de metodologias.

Comparando a abordagem de aprendizagem passiva com aprendizagem ativa observamos que em contraste com o aprendizado passivo, a abordagem de aprendizado ativo treina iterativamente os dados e define novos pontos de medição com base no conjunto de treinamento anterior, o que pode melhorar significativamente a generalização do modelo. No entanto, este procedimento iterativo é computacionalmente mais caro devido ao processamento dos dados nos pontos de medição [3].

A ideia básica da aprendizagem ativa é criar uma interação entre a modelagem e o procedimento de avaliação dos dados. Essa estratégia se concentra em encontrar um meio termo entre a generalização aprimorada do modelo onde normamente para isso a quantidade de custo computacional geralmente aumenta com essa abordagem. O objetivo principal é construir um bom modelo com o menor número de dados [3].

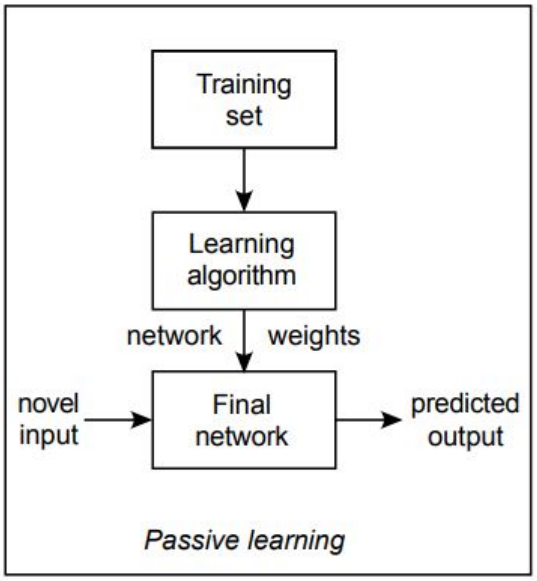

Fig. 1. Estrutura de um algoritmo de aprendizado de máquina utilizando o aprendizado passivo [3].

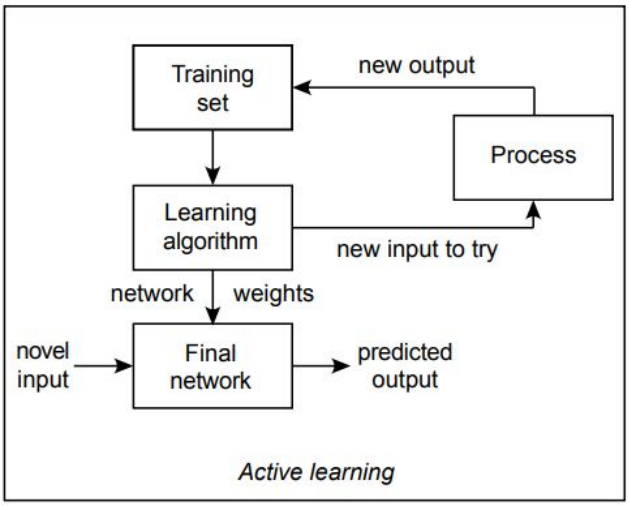

Fig. 2. Estrutura de um algoritmo de aprendizado de máquina utilizando o aprendizado Ativo [3].

Conforme pode ser observado nas figuras 1 e 2, a diferença entre aprendizagem ativa e passiva é que, no caso de aprendizagem ativa, o algoritmo de aprendizagem interage com o processo de forma que o conjunto de treinamento seja atualizado iterativamente. No caso de aprendizado passivo, o algoritmo de aprendizado obtém o conjunto de treinamento de uma só vez, em uma única etapa [3]

Dentro da abordagem de aprendizado ativo pode se dividir em:

- *Uncertainty Sampling.* Ela identifica elementos estão próximos ao limite de decisão e cuja o algoritmo não identificou exatamente em qual classe deve pertencer o elemento avaliado. Esses elementos são mais propensos a serem classificados erroneamente. O modelo é então treinado novamente com esses dados, com o objetivo de melhorá-lo, incluindo novos elementos das regiões de limiar em cada interação.
- *Diversity Sampling.* Inclui novos elementos no treinamento do modelo que não estão necessariamente próximos do limite de decisão, mas sim elementos novos e desconhecidos para o modelo [4].

II. METODOLOGIA

Para a elaboração deste trabalho foi utilizado um pequeno banco de dados contendo 918 registros, onde cada registro contém 11 features relacionados a saúde de um indivíduo e um target se esse indivíduo já teve alguma doença cardíaca ou não, dentro dos features temos variáveis numéricas e categóricas como pode se observar abaixo:

- Idade: 25 até 80 anos;
- Tipo de dor no peito: ASY, ATA, NAP, TA;
- Colesterol: 0 até 650;
- Dor no peito com exercício: SIM ou NÃO
- Glicose em jejum: 0 ou 1;
- Frequência cardíaca: 60 até 200;
- Old Peak: -3 até 7;
- Pressão Sanguínea em repouso: 0 até 200;
- ECG em repouso: LVH, Normal e ST;
- ST slope: Down, flat e UP;
- Sexo: F e M.

O dataset foi aquisitado pelo site Kaggle [5], sendo um bom conjunto para aplicar um estudo de aprendizado de máquinas visando a partir dos features prever se o indivíduo pode contrair alguma doença cardíaca ou não. O conjunto de dados é bem balanceado contendo 508 registros de indivíduos que tiveram doença cardíaca e 410 registros de indivíduos que não tiveram doença cardíaca.

Para o desenvolvimento do projeto foi utilizada a plataforma on-line Google Collaboratory. Esta escolha se deu pela facilidade das instalações de pacotes, como pandas, matplotlib e Scikit Learn.

O projeto foi desenvolvido em linguagem Python se dividindo em quatro partes principais: Tratamento do banco de dados, implementação do modelo de aprendizado de máquina utilizando aprendizado passivo, implementação do modelo de aprendizado de máquina utilizando aprendizado ativo e exibição de resultados que sera descorrido no tópico III *Resultados*.

A. Tratamento do banco de dados

Observando os tipos de variáveis encontradas no dataset, o foi necessário fazer a transformação das variáveis categóricas em variáveis numéricas, visto que o modelo utilizado de aprendizado de máquina não aceita variáveis categóricas como input. Para isso foi utilizado a função *One Hot Encoding*, do pandas, ela cria, por exemplo em uma variável categórica com 4 valores diferente, 4 novos features de valores binários. Com isso o dataset que continha 11 features passou a ter 20.

Após essa etapa foi realizado a normalização dos dados a fim de evitar que os features com valores maiores sejam considerados mais importantes para o modelo.

Posteriormente o dataset foi dividido nas proporções de 80% (734) para treino e 20% (184) para teste.

O modelo de aprendizado de máquina utilizado em ambos os métodos foi o *Support Vector Classification* do pacote sklearn.

B. Aprendizado Passivo

A utilização do método de aprendizagem passiva basicamente seleciona os dados de treino aleatoriamente para treino cada interação, iniciando com 104 amostras onde foi definido que em cada nova etapa de treinamento serão acrescidas 10 novas entradas ao subconjunto de treino até completar todo dataset e salvo a acurácia selecionada para exibição posterior.

Na figura 3 segue um fluxograma utilizado no funcionamento do treinamento do modelo utilizando aprendizado passivo.

Fig. 3. Fluxograma aprendizado passivo

C. Aprendizado Ativo

Durante o treinamento do modelo utilizando o aprendizado ativo, também é acrescido em cada interação 10 amostras até o fim do dataset, porém essas amostras não são selecionadas aleatoriamente mas sim de forma a atender o método mencionado anteriormente, o ”Uncertainty Sampling”, onde são identificados elementos próximos do limiar de decisão para serem adicionados nas próximas etapas de treinamento, para isso foi utilizado a função *predict_proba* presente na biblioteca Scikit Learn, ela retorna a probabilidade de saída de cada prova, com a porcentagem variando de 0 a 1 para cada dado, onde quando uma amostra fica muito próximo a 0,5 significa que esse dado está no limiar da indecisão, portanto tendo a possibilidade de ser classificado de forma errônea, esse tipo de dado que é importante adicionarmos novamente no loop para que o modelo possa ser retreinado.

O limiar de indecisão escolhido para o projeto retreinar as amostras avaliadas com o *predict_proba* foi de 0,5 até 0,53.

Na figura 4 segue um fluxograma utilizado no funcionamento do treinamento do modelo utilizando aprendizado ativo.

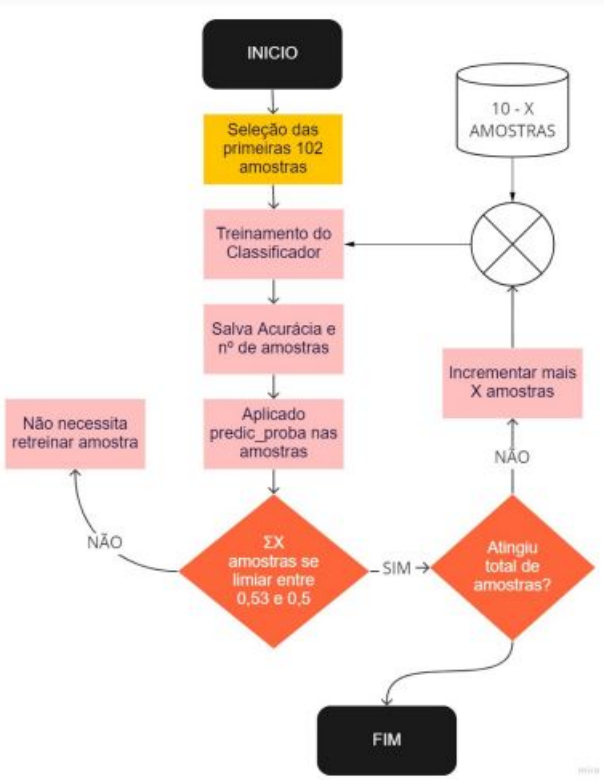

Fig. 4. Fluxograma aprendizado Ativo

O código utilizado para a elaboração do projeto pode ser encontrado no github do autor: https://github.com/GustavoGregorio/ActiveLearning

III. RESULTADOS

Como pode se observar na figura 5, é possível verificar que durante o início do processamento a acurácia do aprendizado ativo foi maior, alcançando uma acurácia máxima de 0,87 com 320 amostras, enquanto o aprendizado passivo iniciou com uma acurácia atrasada em relação ao método oposto, alcançando uma acurácia de mesmo valor apenas no final do treinamento com 680 amostras.

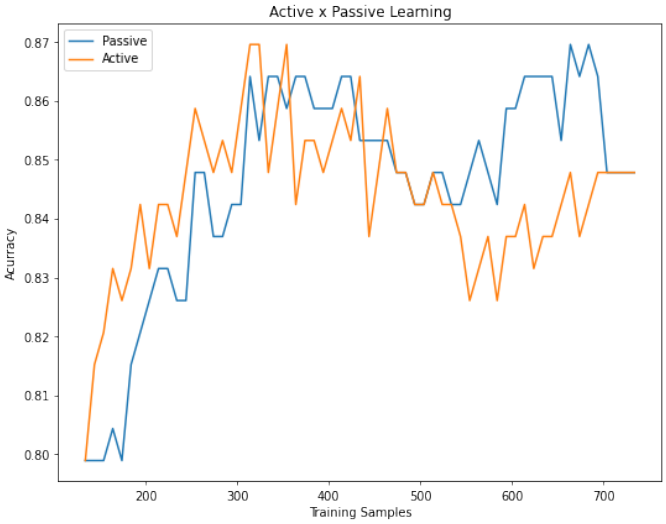

Fig. 5. Aprendizado ativo x aprendizado passivo

Após o aprendizado ativo atingir a acurácia máxima, a acurácia diminui com tendencia a se igualar ao aprendizado passivo no final do treinamento.

IV. CONCLUSÃO

Fica evidente que a abordagem de aprendizado ativo é mais eficiente que a de aprendizado passivo, visto que com uma quantidade menor de amostras foi alcançando uma acurácia máxima maior do que a abordagem passiva, polpando assim tempo de processamento e aquisição de mais dados rotulados que por algumas vezes pode ser custoso. Para os proximos passos sera utilizado um banco de dados.

Para os próximos passos será avaliado a mesma abordagem um dataset maior, permitindo verificar se as mesmas discrepâncias serão observadas, visto que com mais dados poderão ser realizados mais interações.

REFERÊNCIAS

[1] E. G. Horta e A. P. Braga, “Aplicação de Máquinas de Aprendizado Extremo ao Problema de Aprendizado Ativo”, in Anais do 11. Congresso Brasileiro de Inteligência Computacional, Porto de Galinhas, Pernambuco, mar. 2016, p. 1–6. doi: 10.21528/CBIC2013-069.

[2] L. M. C. Cabezas, ”Métodos de aprendizado ativo”.Universidade Federal de São Carlos - Centro de Ciencias Exatas e de Tecnologia Departamento de Estatistica

[3] B. Hartmann e O. Nelles, “Adaptive Test Planning for the Calibration of Combustion Engines – Methodology”, p. 17.

[4] E. Mosqueira-Rey, D. Alonso-Ríos, e A. Baamonde-Lozano, “Integrating Iterative Machine Teaching and Active Learning into the Machine Learning Loop”, Procedia Computer Science, vol. 192, p. 553–562, 2021, doi: 10.1016/j.procs.2021.08.057.

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