ANÁLISE DE CONFIABILIDADE DE ALGORITMOS QUÂNTICOS E CLÁSSICOS USANDO IA E BIG DATA PARA DIAGNÓSTICOS DE SAÚDE

MIANO, M. G. V.
Fatec Americana – "Ministro Ralph Biasi" - Análise e Desenvolvimento de Sistemas
mariana.miano@fatec.sp.gov.br

Reliability Analysis Of Quantum And Classical Algorithms Using AI and Big Data For Health Diagnosis

Eixo Tecnológico: Informação e Comunicação

Resumo

Os impactos da tecnologia e da transformação digital na área da saúde são significativos, principalmente com o desenvolvimento de técnicas de análises de dados via Big Data e IA, cujos sistemas computacionais dão o suporte para escolhas assertivas e melhoram a precisão das avaliações nas rotinas dos especialistas de saúde. E com o Big Data, revolucionando decisões administrativas, como investimentos, há a redução de custos e otimização de operações em centros médicos e hospitais. Para o diagnóstico de doenças, o processo de automatização impacta na redução do custo temporal e técnico para os profissionais de saúde. Assim, faz-se necessário uma ferramenta que auxilie esse profissional na tomada de decisão no momento do diagnóstico e que apresente elevada confiabilidade. A teoria clássica do aprendizado de máquina (ML) e a teoria da computação quântica estão entre as áreas científicas de desenvolvimento mais rápido em nossos dias. Nos últimos anos, pesquisadores investigaram se a computação quântica pode ajudar a melhorar os algoritmos clássicos de Machine Learning (ML), uma sub-área de IA. O Machine Learning Quântico (QML) inclui métodos híbridos que envolvem algoritmos clássicos e quânticos.Este trabalho tem o objetivo de analisar comparativamente a confiabilidade do algoritmo quântico QSVC e clássico SVM usando QML e Big Data para diagnósticos de saúde. Neste estudo, foi utilizado um dataset de câncer de mama.

Palavras-chave: Confiabilidade, Algoritmos Quânticos, Machine Learning, Big Data, Saúde.

Abstract

The impacts of technology and digital transformation in the health area are significant, especially with the development of data analysis techniques via Big Data and AI, whose computational systems support assertive choices and improve the accuracy of evaluations in specialists' routines of health. And with Big Data, revolutionizing administrative decisions, such as investments, there is the reduction of costs and optimization of operations in medical centers and hospitals. For the diagnosis of diseases, the process of automating the diagnosis has an impact on reducing the temporal and technical cost for health professionals. Thus, a tool is needed to help professionals make decisions at the time of diagnosis and to provide high reliability. Classical machine learning theory and quantum computing theory are among the most rapidly developing areas of science today. In recent years, researchers have investigated whether quantum computing can help improve classical Machine Learning (ML) algorithms, a sub-area of AI. Quantum Machine Learning (QML) includes hybrid methods involving classical and quantum algorithms. This work aims to comparatively analyze the reliability of the QSVC quantum algorithm and classical SVM using QML and Big Data for health diagnoses. In this study, a breast cancer dataset was used.

Keywords: *Reliability, Quantum Algorithms, Machine Learning, Big Data, Health.*

1. Introdução

A teoria clássica do aprendizado de máquina (ML) e a teoria da computação quântica estão entre as áreas científicas de desenvolvimento mais rápido em nossos dias. Nos últimos anos, pesquisadores investigaram se a computação quântica pode ajudar a melhorar os algoritmos

clássicos de Machine Learning (ML), uma sub-área de IA. O Machine Learning Quântico (QML) inclui métodos híbridos que envolvem algoritmos clássicos e quânticos. Abordagens quânticas podem ser usadas para analisar estados quânticos em vez de dados clássicos. Por outro lado, os algoritmos quânticos podem melhorar exponencialmente o algoritmo clássico da ciência de dados [1].

Segundo o relatório "*AI for Science*" do *Department of Energy* dos USA [2] novas técnicas de IA (nas quais se enquadram o QML) serão indispensáveis para suportar o crescimento contínuo e expansão da infraestrutura da Ciência através de sistemas *exascale*. Segundo [3], *Big Data* é um conjunto de três "Vs", sendo eles volume, velocidade e variedade. Atualmente, consideram-se cinco "Vs", incluindo veracidade e valor.

Os impactos da tecnologia e da transformação digital na área da saúde são significativos, principalmente com o desenvolvimento de técnicas de análises de dados via Bia Data e IA, cujos sistemas computacionais dão o suporte para escolhas assertivas e melhoram a precisão das avaliações nas rotinas dos especialistas de saúde. E com o Big Data, revolucionando decisões administrativas, como investimentos, há a redução de custos e otimização de operações em centros médicos e hospitais.

Um estudo feito pela IDC (International Data Corporation) aponta que até 2022 foram investidos US $ 1,931 milhões – equivalente a R $ 10 bilhões – no setor de saúde na América Latina, reforçando a evolução tecnológica e o foco para a IA e Big Data [4].

Diante do investimento e da importância do Big Data e IA na saúde, especialistas da área de saúde começam a vislumbrar um futuro em que os sistemas computacionais se tornarão essenciais para o setor medicinal.

Para o diagnóstico de doenças, o processo de automatização impacta na redução do custo temporal e técnico para os profissionais de saúde. A análise minuciosa dessas imagens passa pela expertise do profissional de saúde, assim como pela qualidade dos equipamentos utilizados para leitura e realização de laudo diagnóstico. Assim, faz-se necessária uma ferramenta que auxilie esse profissional na tomada de decisão no momento do diagnóstico e que apresente elevada confiabilidade [5].

Assim, esse trabalho tem o objetivo de analisar comparativamente a confiabilidade do algoritmo quântico QSVC e clássico SVM usando QML e Big Data para diagnósticos de saúde. Neste estudo, foi utilizado um dataset de câncer de mama.

2. Materiais e métodos

2.1. Materiais

Foram utilizadas as plataforma Jupyter Notebook [6] e IBM Quantum Lab [7], com o uso do Qiskit (uma estrutura de código aberto para computação quântica) [8] e Python [9], linguagem de programação multiparadigma que o Qiskit suporta e recomenda nas documentações. Foram utilizados os métodos Confusion Matrix, Classification Report [10] e curva ROC, de Big Data junto aos métodos SVM e QSVC, de QML, para o treinamento dos dados.

2.2. Metodologia

A metodologia transita entre exploratória e explicativa [11]. A abordagem resume-se como quantitativa, pois a pesquisa é teórico-prática comparativa, para avaliação de confiabilidade, utilizando métodos matemáticos-estatísticos aplicados em um dataset (BD), que passou por treinamento de métodos de QML

3. Resultados e Discussão

Para validação dos resultados, devemos saber como o modelo de Machine Learning está desempenhando as classificações e se os resultados são satisfatórios.

Ao utilizar a função .score() do Scikit-learn[10], o principal objetivo é saber quão bem o modelo irá generalizar, ou seja, se o modelo será efetivo ao receber um dado que ele desconhece. Para a execução dessa função, são inseridos dois principais parâmetros: os *features* de treino ou teste(X) e a label esperada(Y). A função verificará o número que seu modelo previu e fará a comparação com o valor esperado do conjunto de treino. Pode-se aplicar essa mesma função no conjunto de teste.

Como resultado dessa função, será apresentado um número na faixa de 0.0 a 1.0, onde quanto mais próximo de 1.0 melhor a aproximação do modelo. Porém é importante que o modelo não tenha um score muito alto no conjunto de treino, enquanto no conjunto de teste ele está bem abaixo. Se isso acontece, verifica-se um problema de *Overfitting*, que ocorre quando o modelo "adivinha" precisamente os dados que foram usados para treiná-lo, mas ele não consegue tratar dados desconhecidos. Ainda há os casos de *Underfitting*, onde o modelo não consegue prever a tendência de dados.

3.1 Confusion Matrix

A *Confusion Matrix* é um método que utiliza a matriz de confusão para avaliar a precisão de uma classificação. Por definição a matriz de confusão C é tal que C_{ij} é igual ao número de observações conhecidas por estarem no grupo i e previsto para estar no grupo j.

Assim, na classificação binária, a contagem de verdadeiro negativo (TN) é $C_{0,0}$, falso negativo (FN) $C_{1,0}$, verdadeiro positivo (TP) $C_{1,1}$,, e falso positivo (FP) $C_{0,1}$.

Os parâmetros recebidos por essa função são, y_test e model_test, sendo que o model_test é uma variável que recebe o model.predict(X_test) que seria a um modelo que pode ser criado e ajustado com dados treinados e usado para fazer uma previsão. Esses dados utilizados são do dataset que vem de uma matriz de subconjuntos aleatórios.

3.2 Classification Report

Esse método faz a construção do texto mostrando as principais métricas de classificação. Insere-se os valores de y_test e model_test e são retornados os valores de um resumo de precisão, recall e a pontuação de F1 para cada classe. A função retorna a *precision*, que é a razão do número de verdadeiros positivos dividido pela soma de número de falsos positivos mais os verdadeiros positivos.

"Precisão ou Confiança (como é chamado em DataMining) denota a proporção de casos positivos previstos que são corretamente Positivos Reais. É nisso que o Machine Learning, Data Mining e Information Retrieval se concentram, mas é totalmente ignorado na análise ROC. No entanto, pode ser analogamente chamado de Precisão do Positivo Verdadeiro (PPV), sendo uma

medida de precisão dos Positivos Previstos em contraste com a taxa de descoberta de Positivos Reais (PPR).

3.3 ROC curve

A Curva Característica de Operação do Receptor (Curva COR), ou, do inglês, Receiver Operating Characteristic Curve (ROC curve), ou, simplesmente, curva ROC, é uma representação gráfica que ilustra o desempenho (ou performance) de um sistema classificador binário à medida que o seu limiar de discriminação varia. A curva ROC é também conhecida como curva de característica de operação relativa.

A curva ROC é obtido pela representação da razão RPV = Positivos Verdadeiros / Positivos Totais versus a razão RPF = Positivos Falsos / Negativos Totais, para vários valores do limiar de classificação. O RPV é também conhecido como sensibilidade (ou taxa de verdadeiros positivos), e RPF = 1-especificidade ou taxa de falsos positivos. A especificidade é conhecida como taxa de verdadeiros negativos (RVN).

A análise ROC fornece ferramentas para selecionar modelos possivelmente ideais (modelos ótimos) e descartar modelos não tão ótimos, independentemente (e antes de especificar) o contexto de custos ou a distribuição de classe. A análise ROC está relacionada de forma direta e natural com a análise de custo/benefício do diagnóstico.

3.4 SVM e QSVC

O SVM é um algoritmo que busca uma linha de separação entre duas classes distintas analisando os dois pontos, um de cada grupo, mais próximos da outra classe. Isto é, o SVM escolhe a reta — também chamada de hiperplano em maiores dimensões— entre dois grupos que se distancia mais de cada um [12]. Estudos e aplicações do algoritmo SVM para a saúde, utilizado para classificar genes usando dados de expressão gênica aplicando técnicas de SVM são apresentados em [13]. O SVM foi escolhido pois existe sua versão quântica, o QSVC, que é apresentado no QISKIT, utilizando um kernel quântico.

3.5 Testes dos algoritmos SVM e QSVC utilizando o dataset de câncer de mama

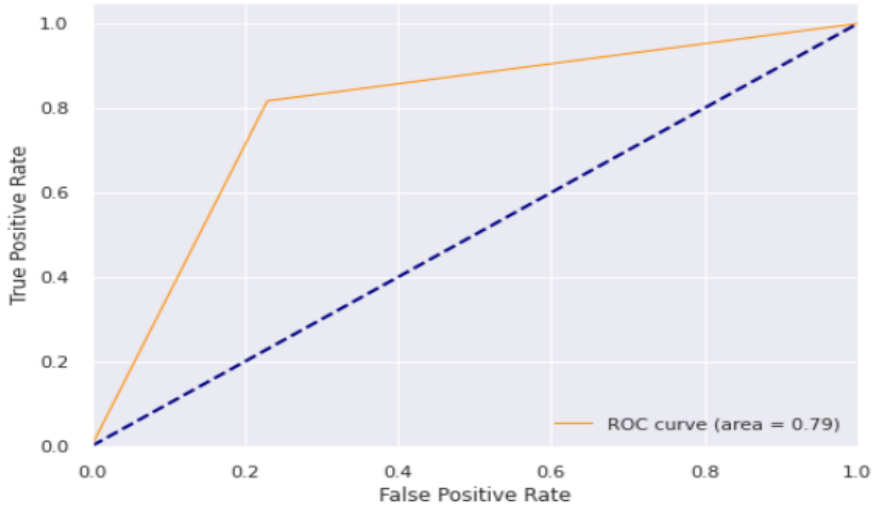
Nestes testes, para verificar a confiabilidade dos algoritmo SVM e QSVC, foi utilizado o *dataset* de câncer de mama, por ser um conjunto de dados relativamente grande e testar a capacidade da plataforma IBM. O *dataset* foi encontrado no *Kaggle* e contém 4024 linhas e 16 colunas. As figuras 1 a 3 mostram os resultados do algoritmo SVM e as figuras 4 a 6 do algoritmo QSVC nos quesitos precisão e acuracidade, com o *dataset* de câncer de mama.

Fig. 1 - Confusion Matrix e Classification Report do SVM utilizando o dataset de câncer de mama.

```
Confusion Matrix:
[[873 167]
 [260 745]]


Classification Report:
              precision    recall  f1-score   support

           0       0.77      0.84      0.80      1040
           1       0.82      0.74      0.78      1005

    accuracy                           0.79      2045
   macro avg       0.79      0.79      0.79      2045
weighted avg       0.79      0.79      0.79      2045
```

Fonte: Elaborado pela autora

Fig. 2 – ROC curve do SVM utilizando o dataset de câncer de mama.



Fonte: Elaborado pela autora

Fig. 3 – AUC e Precisão do SVM utilizando o dataset de câncer de mama.

```
Acuuracy Of the Model:  0.7767857142857143
```

```
score = svm.score(X_train, y_train.ravel())
print(f"Testing accuracy: {score}")
```

```
Testing accuracy: 0.8199233716475096
```

Fonte: Elaborada pela autora.

A Confusion Matrix da figura 1 mostra os valores que foram corretos e os valores que o modelo errou, que nesse teste foram os valores 873 e 745, que representam os valores verdadeiros negativos e verdadeiros positivos respectivamente. Os valores 167 e 260 representam os valores de falsos negativos e falsos positivos respectivamente. A partir desses valores, realizam-se os cálculos da Classification Report da figura 1, e elabora-se o ROC Curve da figura 2. Nesse teste, o modelo teve um resultado de AUC (*Accuracy of the model*) de "0.77" e uma precisão (*Testing accuracy*) de "0.81" como mostrado na figura 3. A AUC mostra o quanto o modelo pode distinguir entre duas classes enquanto a precisão é a quantidade Verdadeiros Positivos (*True Positives*) sobre o total de positivos do modelo (FP e TP).

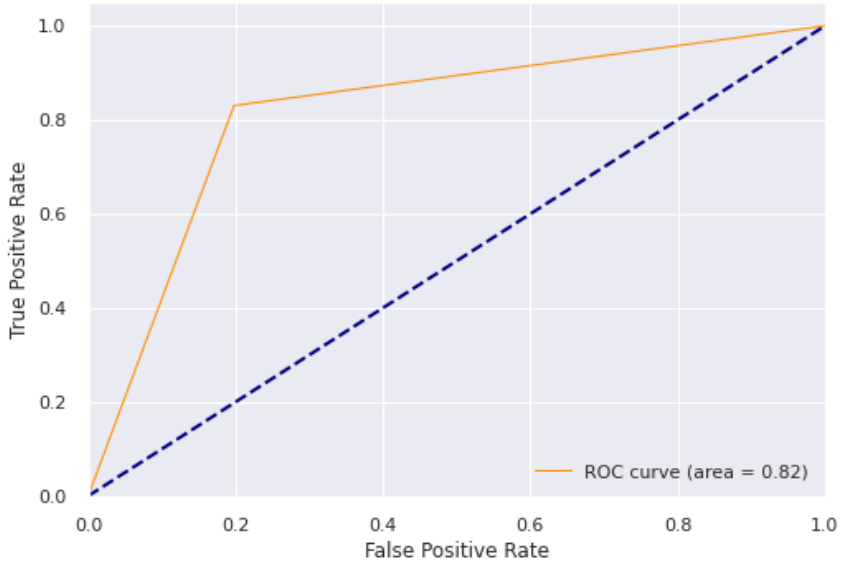
Na sequência, testou-se o algoritmo QSVC, utilizando o mesmo dataset de câncer de mama para fins de comparação. As figuras 4 a 6 mostram os resultados obtidos.

Fig. 4 – Confusion Matrix e Classification Report do QSVC utilizando o dataset de câncer de mama

```
Confusion Matrix:
[[879 161]
 [217 788]]


Classification Report:
              precision    recall  f1-score   support

           0       0.80      0.85      0.82      1040
           1       0.83      0.78      0.81      1005

    accuracy                           0.82      2045
   macro avg       0.82      0.81      0.81      2045
weighted avg       0.82      0.82      0.81      2045
```

Fonte: Elaborado pela autora.

Fig. 5 – ROC Curve do QSVC utilizando o dataset de câncer de mama.



Fonte: Elaborado pela autora

Fig. 6 – AUC e Precisão do QSVC utilizando o dataset de câncer de mama.

```
Acuuracy Of the Model:  0.8151589242053789
```

```
score = qsvc.score(X_train, y_train.ravel())
print(f"Testing accuracy: {score}")
```

```
Testing accuracy: 0.8310626702997275
```

Fonte: Elaborado pela autora

Nesse teste, o modelo teve um resultado de AUC (*Accuracy of the model*) de 0.81 e uma precisão (*Testing accuracy*) de 0.83 como mostrado na figura 6. A AUC mostra o quanto o modelo pode distinguir entre duas classes enquanto a precisão é a quantidade Verdadeiros Positivos (*True Positives*) sobre o total de positivos do modelo (FP e TP).

4. Considerações finais

Comparando os dois resultados dos testes do SVM e QSVC utilizando o dataset de câncer de mama, é possível observar que os resultados do QSVC (Fig.4) foram superiores aos resultados do SVM (Fig.1), em termos de acuracidade (AUC) e precisão (Testing accuracy), como apresentado nas figuras 4 e 1, respectivamente. Esses testes também apresentam os valores positivos e os valores falsos que os modelos previram por meio da *Confusion Matrix*, que contém o *Classification Report*, onde é possível visualizar as métricas de classificação dos modelos. Além disso o QSVC teve a AUC (Accuracy of the model) e a precisão (Testing accuracy) superior ao do SVM, como mostrado na Fig. 6. O QSVC além de ser melhor para

identificar as classes apresentou uma taxa superior de acerto. Embora as diferenças não tenham sido tão significativas entre os métodos Clássico (SVM) e Quântivo (QSVC), verificou-se que quanto maior o Banco de Dados utilizado, melhor a confiabilidade do método quântico, dentro da viabilidade do processamento do ambiente quântico utilizado.

Referências

[1] FASTOVETS, D. et al. Machine learning methods in quantum computing theory. In: **International Conference on Micro-and-Nano-Electronics**. Proceedings v. 11022. 2019. https://doi.org/10.1117/12.2522427.

[2] STEVENS R. et al. "***AI for Science***" Report on the Department of Energy of USA (DOE). 2019.

[3] AMARAL, F. **Introdução à ciência de dados: mineração de dados e big data**. Alta Books Ed. Rio de Janeiro. 2016.

[4] PIRES, R. "**Tecnologia aplicada à saúde segue avançando em 2022**". Medicina S/A. 03/01/2022. Disponível em : <https://medicinasa.com.br/tecnologia-saude-2022/>. Acesso em 21 de abr. 2022.

[5] SCHAEFER, O. et al. "**Precision Medicine and Big Data**". Asian Bioethics Review, 2019.

[6] JUPYTER. "**Jupyter Notebook**". Jupyter Team. 2015. Disponível em:< https://jupyter.org/>. Acesso em 02 ago. 2022

[7] IBM. "**IBM Quantum**". IBM. 29 de mar. de 2021. Disponível em: < https://www.ibm.com/quantum>. Acesso em 08 ago. 2022

[8]QISKIT. "**Open-Source Quantum Development**". Qiskit. 2021. Disponível em: https://qiskit.org/documentation/. Acesso em 15 ago. 2022

[9] PYTHON. "**Python**". Python Software Foundation. 2022. Disponível em: < https://www.python.org/>. Acesso em 25 de mar. 2022.

[10] PEDREGOSA, M. et al. **Scikit-learn: aprendizado de máquina em Python.** JMLR 12, pp. 2825-2830, 2011. Disponível em: < https://scikit-learn.org/stable/about.html#citing-scikit-learn>. Acesso em 10 de nov.2022.

[11]DIANA, J. **Pesquisa descritiva, exploratória e explicativa**. 2010. Disponível em: <https://www.diferenca.com/pesquisa-descritiva-exploratoria-explicativa/?utm_source=whatsapp&utm_medium=referral . Acesso em 11 de mar.2022.

[12] MISHRA, S. "**Breaking Down the Support Vector Machine (SVM) Algorithm**". Towards Data Science. 29 out. 2020. Disponível em: < https://towardsdatascience.com/breaking-down-the-support-vector-machine-svm-algorithm-d2c030d58d42>. Acesso em 16 de mai. 2022.

[13]HARERIMANA, G. et al. "**Health Big Data Analytics: A Technology Survey**". IEEE Access, 2018.