Relatório VII - Tópicos Especiais em Computação

Vandirleya Barbosa

[1]Universidade Federal de Piauí (UFPI)

vandirleya.barbosa@ufpi.edu.br

Resumo. *Este trabalho apresenta uma análise detalhada do conjunto de dados de câncer de mama, focando em como a normalização dos dados influência o desempenho dos modelos de aprendizado de máquina. O conjunto de dados, originário do UCI Machine Learning Repository, contém informações de exames de diagnóstico de câncer de mama, incluindo 30 características extraídas de imagens digitais de biópsias. Nosso objetivo principal foi conduzir uma análise comparativa utilizando o algoritmo ID3, aplicando validação cruzada com K=5 para avaliar a acurácia e o desvio padrão em dois cenários distintos: dados normalizados e dados não normalizados. Desse modo, este estudo propõe um resumo exploratório dos resultados obtidos, destacando como a normalização pode ser uma ferramenta eficaz para melhorar a qualidade das previsões em modelos de aprendizado de máquina. A análise comparativa fornece uma base sólida para compreender melhor as relações entre as características dos dados e o desempenho do modelo, oferecendo visões valiosas para a aplicação eficaz dessas técnicas em diversos domínios.*

Abstract. *This work presents a detailed analysis of the breast cancer dataset, focusing on how data normalization influences the performance of machine learning models. The dataset, originating from the UCI Machine Learning Repository, contains diagnostic exam information for breast cancer, including 30 features extracted from digital images of biopsies. Our main objective was to conduct a comparative analysis using the ID3 algorithm, applying 5-fold cross-validation to evaluate the accuracy and standard deviation in two distinct scenarios: normalized data and non-normalized data. Thus, this study proposes an exploratory summary of the obtained results, highlighting how normalization can be an effective tool to improve the prediction quality in machine learning models. The comparative analysis provides a solid foundation to better understand the relationships between data characteristics and model performance, offering valuable insights for the effective application of these techniques in various domains.*

1. Introdução

O câncer de mama é uma das principais causas de mortalidade entre as mulheres em todo o mundo. Caracterizado pelo crescimento descontrolado de células malignas nos tecidos mamários, ele pode se apresentar de diversas formas e exigir diferentes abordagens de tratamento. A detecção precoce é crucial para aumentar as chances de sucesso no tratamento e, por isso, métodos eficazes de diagnóstico são extremamente importantes.

Com a crescente incidência dessa doença, a pesquisa científica e médica tem se concentrado em desenvolver ferramentas e técnicas para melhorar a detecção, o diagnóstico e o tratamento do câncer de mama.

Nos últimos anos, os avanços tecnológicos têm transformado significativamente o setor de saúde. A introdução de tecnologias inovadoras, como a inteligência artificial, aprendizado de máquina, e a análise de big data, tem permitido a criação de sistemas mais precisos e eficientes para o diagnóstico e tratamento de diversas doenças, incluindo o câncer de mama. Hospitais e centros de pesquisa agora utilizam essas tecnologias para analisar grandes volumes de dados médicos, identificar padrões e tendências, e desenvolver novos métodos de tratamento personalizado. Esses avanços não só melhoram a qualidade do atendimento ao paciente, mas também abrem novas fronteiras para a pesquisa médica.

Especificamente no campo do câncer de mama, a análise de dados tem se mostrado uma ferramenta poderosa para melhorar a compreensão da doença e suas características. O uso de conjuntos de dados como o do UCI Machine Learning Repository, que contém informações detalhadas sobre exames de diagnóstico de câncer de mama, permite a aplicação de algoritmos de aprendizado de máquina para identificar padrões importantes. A normalização dos dados e a aplicação de técnicas como a validação cruzada e o uso de algoritmos de decisão, como o ID3, são essenciais para melhorar a precisão dos modelos preditivos. Este estudo busca explorar esses aspectos, comparando o desempenho de modelos treinados com dados normalizados e não normalizados, e destacando como a normalização pode ser uma ferramenta eficaz para melhorar a qualidade das previsões em modelos de aprendizado de máquina.

2. Desenvolvimento

Esta seção apresenta a metodologia para o desenvolvimento deste artigo, incluindo a base de dados selecionada, os algoritmos e métodos empregados. Além disso, as análises obtidas para os dados normalizados e não normalizados são apresentadas.

2.1. Metodologia

As etapas seguidas para a análise dos dados de câncer de mama utiliza o algoritmo de árvore de decisão (ID3) com validação cruzada K-Fold, comparando os resultados para dados normalizados e não normalizados. Primeiramente, os dados foram carregados a partir do conjunto de dados de câncer de mama disponível no *sklearn.datasets*. Este conjunto de dados contém informações sobre exames de diagnóstico de câncer de mama, incluindo 30 características extraídas de imagens digitais de biópsias. Os dados foram inicialmente analisados sem qualquer normalização. As características e o alvo foram separados em X_non_normalized e y, respectivamente. Para criar a versão normalizada dos dados, foi utilizado o *StandardScaler* da biblioteca *sklearn.preprocessing*. As Tabelas 1 e 2 apresentam os registros do dataset.

A normalização é um passo importante para garantir que todas as características estejam na mesma escala, o que pode influenciar o desempenho do modelo. A validação cruzada K-Fold foi implementada para avaliar a performance do modelo de forma mais robusta. Neste método, os dados são divididos em k subconjuntos (folds), e o modelo é treinado e testado k vezes, cada vez utilizando um fold diferente como conjunto de teste e

Tabela 1. Primeiros 14 Registros do Conjunto de Dados de Câncer de Mama

Registros	0	1	2	3	4
mean radius	17.99	20.57	19.69	11.42	20.29
mean texture	10.38	17.77	21.25	20.38	14.34
mean perimeter	122.80	132.90	130.00	77.58	135.10
mean area	1001.0	1326.0	1203.0	386.1	1297.0
mean smoothness	0.11840	0.08474	0.10960	0.14250	0.10030
mean compactness	0.27760	0.07864	0.15990	0.28390	0.13280
mean concavity	0.3001	0.0869	0.1974	0.2414	0.1980
mean concave points	0.14710	0.07017	0.12790	0.10520	0.10430
mean symmetry	0.2419	0.1812	0.2069	0.2597	0.1809
mean fractal dimension	0.07871	0.05667	0.05999	0.09744	0.05883
worst texture	17.33	23.41	25.53	26.50	16.67
worst perimeter	184.60	158.80	152.50	98.87	152.20
worst area	2019.0	1956.0	1709.0	567.7	1575.0
worst smoothness	0.1622	0.1238	0.1444	0.2098	0.1374

os outros como conjunto de treino. Para a classificação, foi utilizado o modelo de árvore de decisão com o critério de entropia, que implementa o algoritmo ID3. O modelo foi avaliado utilizando os dados normalizados e não normalizados com validação cruzada K-Fold. Por fim, foram calculadas a acurácia média e o desvio padrão das acurácias para entender o impacto da normalização dos dados no desempenho do modelo de aprendizado de máquina.

2.2. Dados Normalizados e Não Normalizados

A normalização dos dados é um passo crucial no pré-processamento de dados para muitos algoritmos de aprendizado de máquina, especialmente aqueles que são sensíveis à escala e distribuição dos dados. No contexto deste trabalho sobre análise de diagnósticos de câncer de mama, a normalização foi aplicada para garantir que todas as características dos dados estivessem em uma escala comparável, evitando que características com unidades diferentes dominassem o processo de aprendizado. No processo realizado, os dados foram divididos em duas partes principais: dados não normalizados e dados normalizados. Os dados não normalizados mantêm suas unidades originais e foram diretamente utilizados como entrada para o modelo de classificação. Já os dados normalizados passaram por um processo de escala utilizando o método *StandardScaler* do *scikit-learn*, o qual transforma os dados de forma que sua distribuição tenha média zero e desvio padrão igual a um.

Para avaliar o impacto da normalização nos modelos de aprendizado, foi utilizada a validação cruzada com K=5. Esse método divide o conjunto de dados em cinco partes iguais, onde cada parte é utilizada uma vez como conjunto de teste enquanto as outras partes são usadas como conjunto de treinamento. Esse processo foi repetido cinco vezes para garantir que todos os dados fossem usados tanto para treinamento quanto para teste, reduzindo o viés na avaliação do modelo. Os resultados obtidos demonstraram que a normalização dos dados resultou em uma leve melhoria na acurácia média do modelo de classificação de Decision Tree (ID3), conforme mostrado na Tabela 3.

A diferença na acurácia média entre os dados normalizados e não normalizados

Tabela 2. Os demais registros do Conjunto de Dados de Câncer de Mama.

Registros	0	1	2	3	4
mean radius	17.99	20.57	19.69	11.42	20.29
mean texture	10.38	17.77	21.25	20.38	14.34
mean perimeter	122.80	132.90	130.00	77.58	135.10
mean area	1001.0	1326.0	1203.0	386.1	1297.0
mean smoothness	0.11840	0.08474	0.10960	0.14250	0.10030
mean compactness	0.27760	0.07864	0.15990	0.28390	0.13280
mean concavity	0.3001	0.0869	0.1974	0.2414	0.1980
mean concave points	0.14710	0.07017	0.12790	0.10520	0.10430
mean symmetry	0.2419	0.1812	0.2069	0.2597	0.1809
mean fractal dimension	0.07871	0.05667	0.05999	0.09744	0.05883
worst texture	17.33	23.41	25.53	26.50	16.67
worst perimeter	184.60	158.80	152.50	98.87	152.20
worst area	2019.0	1956.0	1709.0	567.7	1575.0
worst smoothness	0.1622	0.1238	0.1444	0.2098	0.1374
worst compactness	0.6656	0.1866	0.4245	0.8663	0.2050
worst concavity	0.7119	0.2416	0.4504	0.6869	0.4000
worst concave points	0.2654	0.1860	0.2430	0.2575	0.1625
worst symmetry	0.4601	0.2750	0.3613	0.6638	0.2364
worst fractal dimension	0.11890	0.08902	0.08758	0.17300	0.07678

foi de aproximadamente 0.0018, com um desvio padrão um pouco maior nos dados normalizados. Do ponto de vista de desempenho, essa pequena diferença sugere que, apesar de haver uma ligeira melhoria na acurácia média com a normalização, ela pode não ser significativa o suficiente para justificar o custo adicional de processamento necessário para realizar a normalização. No entanto, a normalização ainda pode ser considerada benéfica para garantir uma melhor generalização e estabilidade do modelo, especialmente em casos onde as características dos dados possuem escalas muito diferentes.

Tabela 3. Comparação de Acurácia e Desvio Padrão para Dados Normalizados e Não Normalizados.

Tipo de Dados	Acurácia Média	Desvio Padrão
Não Normalizados	0.9437	0.0212
Normalizados	0.9455	0.0227

3. Conclusão

Neste estudo, exploramos os efeitos da normalização de dados no desempenho de modelos de aprendizado de máquina aplicados ao conjunto de dados de diagnóstico de câncer de mama. Inicialmente, verificamos que o conjunto de dados consiste em 30 características derivadas de imagens digitais de biópsias, cada uma representando aspectos morfológicos que podem ser indicativos de malignidade. A normalização foi aplicada para garantir que todas as características estivessem na mesma escala, minimizando o impacto de unidades diferentes nos resultados do modelo. Utilizamos o algoritmo ID3, uma árvore de decisão baseada na entropia, para classificar os dados. A avaliação foi realizada

através de validação cruzada com K=5, dividindo repetidamente o conjunto de dados em cinco partes para treinamento e teste. Isso nos permitiu calcular não apenas a acurácia média do modelo, mas também o desvio padrão, refletindo a consistência das previsões em diferentes divisões dos dados. Os resultados mostraram que tanto os dados normalizados quanto os não normalizados apresentaram altos níveis de acurácia média, com uma leve vantagem para os dados normalizados a Acurácia Média = 0.9455 e o Desvio Padrão = 0.0227 em comparação com os dados não normalizados com Acurácia Média = 0.9437 e Desvio Padrão = 0.0212. A diferença na acurácia média entre os dois métodos foi de aproximadamente 0.0018, indicando uma melhoria marginal com a normalização dos dados. Do ponto de vista de desempenho, a pequena diferença na acurácia média sugere que, embora a normalização possa proporcionar benefícios adicionais em termos de estabilidade e generalização do modelo, essa vantagem pode não ser significativa o suficiente para justificar o custo computacional adicional em todos os contextos.