Proceeding Series of the Brazilian Society of Computational and Applied Mathematics, v. 11, n. 1, 2025.

Trabalho apresentado no XLIII CNMAC, Centro de Convenções do Armação Resort - Porto de Galinhas - PE, 2024

Proceeding Series of the Brazilian Society of Computational and Applied Mathematics

Detecção de Câncer De Mama utilizando *Fuzzy* Unordered Rule Induction Algorithm

Josemary M. F. R. C. Rocha[1], Luiz H. Silva[2], Ronei M. Moraes[3]
PPGMDS/UFPB, João Pessoa, PB

Resumo. Estratégias para detecção precoce do câncer de mama incluem mamografia, ultrassom, ressonância magnética, biópsia, exame clínico das mamas e testes genéticos. No entanto, há lacunas quanto a interpretação de alguns exames e dificuldade quanto aos custos e procedimentos no paciente. A aplicação de Aprendizado de Máquina no setor da saúde tem sido cada vez mais frequente e desempenha um papel significativo. Objetivou-se diagnosticar o câncer de mama como benigno ou maligno com base na aplicação do algoritmo FURIA. A proposta do presente estudo foi utilizar algoritmo FURIA, por meio do *Software Waikato Environment for Knowledge Analysis*, utilizando o conjunto de dados *Breast Cancer Wisconsin* disponibilizado gratuitamente no *UCI Machine Learning*. Observou-se que o FURIA gerou resultados satisfatórios com uma Acurácia de 94,56%, dezenove regras e Coeficiente Kappa de 87,94%, identificou-se ainda um elevado Fator de Certeza das regras geradas pelo FURIA.

Palavras-chave. Aprendizado de Máquina, Câncer de mama, Inteligência Artificial, Lógica *Fuzzy*.

1 Introdução

O câncer é um problema de saúde global que resulta do crescimento celular anormal e ocasiona elevados índices de morbimortalidade. O corpo humano é constituído por milhões de células e quando há um crescimento desordenado, forma-se um tumor, denominado como tumor primário, que é uma das características que marcam o início do câncer [6].

Segundo a *World Health Organization*, o Câncer de Mama (CM) é a causa mais comum de óbito em mulheres, representa um em cada quatro casos de câncer e uma em cada seis mortes por câncer, além de ser o câncer mais comumente diagnosticado no mundo, com cerca de 2,3 milhões de novos casos e 685.000 óbitos em 2020 [1]. No Brasil, o câncer de mama também é um dos cânceres mais comuns em mulheres, para cada ano do triênio 2023-2025, foram estimados 73.610 novos casos [6].

Estratégias para detecção precoce do CM incluem mamografia, ultrassom, ressonância magnética, biópsia, exame clínico das mamas e testes genéticos. Os tumores podem ser classificados como benignos ou malignos, com escolha de tratamento com base no grau, estágio e subtipo molecular do CM, além do mais, as opções de tratamento são, por exemplo, cirurgia, radioterapia, quimioterapia e quimioterapia adjuvante [5]. Dessa maneira, a aplicação de Aprendizado de Máquina (AM) no setor da saúde tem sido cada vez mais frequente e desempenha um papel significativo devido ao seu desempenho na predição e diagnóstico.

O método *Fuzzy* Unordered Rule Induction Algorithm (FURIA) é um algoritmo de AM geralmente utilizado para solução de problemas incertos [7]. O FURIA é uma extensão e modificação do algoritmo RIPPER, onde o foco é o aprendizado a partir de regras *Fuzzy* não ordenadas, utilizando um mecanismo que expande as regras, para tratar exemplos descobertos [4].

[1]josemaryfreire.adv@gmail.com
[2]luizenf2014.2@gmail.com
[3]ronei@de.ufpb.br

As regras geradas pelo FURIA são mais globais do que as regras convencionais e apresentam diversas vantagens, como por exemplo, a capacidade de soluções de problemas complexos em diferentes áreas, como é o caso do presente estudo, que tem como problemática o diagnóstico do CM [4]. Logo, a aplicabilidade do FURIA pode transcender a compreensão acerca do diagnóstico do CM por métodos tradicionais, dado a modelagem flexível dos limites diagnósticos, com suas regras *Fuzzy*.

O presente estudo objetivou detectar se o CM é benigno ou maligno com base na aplicação do algoritmo FURIA, por meio do *Software Waikato Environment for Knowledge Analysis* (WEKA), utilizando o conjunto de dados *Breast Cancer Wisconsin* [10], disponibilizado gratuitamente no *UCI Machine Learning*.

2 Materiais e Métodos

O percurso metodológico para este estudo foi realizado partir de uma revisão de literatura sobre o algoritmo FURIA, suas qualidades, deficiências e aplicações em todas as áreas. O banco de dados utilizado foi o *Breast Cancer Wisconsin*, de forma secundária, sendo um banco de dados público onde se utilizou o algoritmo FURIA para analisar a eficiência desse algoritmo. Este banco de dados foi obtido no sítio eletrônico público *UCI Machine Learning*, onde os dados apresentam 699 instâncias, 9 atributos e uma classe [10].

2.1 Descrição da Base de Dados

A base de dados *Breast Cancer Wisconsin* com seus nove atributos é descrita na Tabela 1, são classificados em uma escala intervalar entre um a dez. Nesta base de dados, 241 (65,5%) registros são malignos e 458 (34,5%) os registros são benignos.

Tabela 1: Descrição dos dados do câncer de mama de Wisconsin.

Atributo	Descrição	Valores dos atributos	Média	Desvio padrão
1	Espessura do aglomerado	1-10	4,42	2,82
2	Uniformidade do tamanho	1-10	3,13	3,05
3	Uniformidade do formato	1-10	3,20	2,97
4	Adesão marginal	1-10	2,80	2,86
5	Tamanho das células epiteliais	1-10	3,21	2,21
6	Núcleo sem revestimento	1-10	3,46	3,64
7	Suavidade da cromatina	1-10	3,43	2,44
8	Nucléolos normais	1-10	2,87	3,05
9	Mitoses	1-10	1,59	1,71

A base de dados [10], foi criada a partir de amostras de fluídos coletados pelo *A System for Remote Cytological Diagnosis and Prognosis of Breast Cancer*(XCYT) de pacientes que possuiam massas mamárias sólidas. Em seguida, utilizando um método de ajuste de curva, calculou-se dez características para cada célula da amostra e, com seus respectivos: Valor Médio, Erro Padrão e Valor Extremo de cada característica.

2.2 O método FURIA

A inicialização do algoritmo ocorre quando um conjunto A com os antecedentes numéricos de regras não estiver vazio, seleciona-se o melhor antecedente a_{max}, de modo que itera-se sobre cada

DOI: 10.5540/03.2025.011.01.0377

antecedente A_i em A, o que resulta na melhor fuzzificação, dado a pureza do cálculo $pur_{A[i]}$. Em seguida, ocorre atualização do melhor antecedente se $pur_{A[i]}$ do antecedente atual for maior que a pureza máxima encontrada pur_{max}, as variáveis pur_{max} e a_{max} são atualizadas com os novos valores. Por fim, o antecedente a_{max} é removido de A e a regra r é atualizada com a_{max} [4]. As funções de pertinência geradas são trapezoidais usando a notação $[a, b, c, d]$, onde os valores $[a, d]$ denotam o suporte e $[b, c]$ denota o patamar do trapézio.

O algoritmo de fuzzificação dos antecedentes para uma única regra r pode ser observado a seguir.

Algorithm 1 Fuzzificação dos antecedentes para uma única regra r.

Input : Seja A um conjunto de antecedentes em r
Output: Regra r com antecedentes fuzzificados
while $\boldsymbol{A \neq \emptyset}$ **do**
  $a_{max} \leftarrow$ null ; // a_{max} denota antecedente com maior nível de Pureza
  $pur_{max} \leftarrow 0$; // pur_{max} denota maior valor de Pureza, até a presente iteração
  for $\boldsymbol{i \leftarrow 1}$ **to size**$\boldsymbol{(A)}$ **do**
    Calcular a melhor fuzzificação de $A[i]$ em termos de Pureza
    $pur_{A[i]} \leftarrow$ Pureza da melhor fuzzificação de $A[i]$
    if $\boldsymbol{pur_{A[i]} > pur_{max}}$ **then**
      $pur_{max} \leftarrow pur_{A[i]}$
      $a_{max} \leftarrow A[i]$
    end
  end
  $A \leftarrow A \setminus a_{max}$
  Atualizar r com a_{max}
end
Fonte: [4]

A adequação da regra gerada com base no conjunto de dados, é definida pelo *Fator de Certeza*, dado pela seguinte equação [4]:

$$CF\left(r_i^{(j)}\right) = \frac{2^{\frac{|D_T^{(j)}|}{D_T}} + \sum_{x \in D_T^{(j)}} \mu_{r_i^{(j)}(\mathbf{x})}}{2 + \sum_{x \in D_T} \mu_{r_i^{(j)}(\mathbf{x})}} \tag{1}$$

onde: CF é o *Fator de Certeza*; r_i a regra *Fuzzy* analisada atualmente; D_T representa os limites do suporte; e μ_{r_i} função de pertinência.

Para execução do algoritmo FURIA, o *Software* WEKA foi utilizado neste estudo para viabilizar as variáveis do algoritmo. As medidas utilizadas para este estudo do algoritmo FURIA e sua execução no WEKA foram: matriz de confusão, Coeficiente de Kappa e Acurácia. Diante da matriz de confusão, é possível se conhecer a qualidade de uma decisão e realizar a comparação com outra já usada e possibilitar o conhecimento de ser uma decisão correta ou desejada. Leva-se em conta as concordâncias ou discordâncias entre as fontes de informações apresentadas. A matriz de confusão permitirá através de uma tabela a visualização do desempenho do algoritmo [8].

2.3 Coeficiente Kappa

O coeficiente de Kappa indica a concordância das interpretações pela matriz de confusão, sua escala é definida de acordo com os intervalos: 0,01 a 0,20, levemente concordante; 0,21 a 0,40 razoavelmente concordantes; 0,41 a 0,60 moderadamente concordantes; 0,61 a 0,80 substancialmente concordantes e 0,81 a 0,99 quase perfeitamente concordantes [3].

DOI: 10.5540/03.2025.011.01.0377

O Coeficiente de Kappa é dado da seguinte forma:

$$Kappa = \frac{P(O) - P(E)}{1 - P(E)} \tag{2}$$

onde: $P(O)$ = proporção observada de concordâncias (soma das respostas concordantes dividida pelo total); $P(E)$ = proporção esperada de concordâncias (soma dos valores esperados das respostas concordantes dividida pelo total) [3].

O Coeficiente de Kappa possui críticas em função do seu pessimismo, porém é preferido, também em função dessa característica, pois quando se tem um bom resultado através do Coeficiente de Kappa, tem-se validade de metodologias [3].

O coeficiente de Kappa indica a concordância das interpretações pela matriz de confusão, sua escala é definida de acordo com os intervalos: 0,01 a 0,20 – levemente concordante; 0,21 a 0,40 razoavelmente concordantes; 0,41 a 0,60 moderadamente concordantes; 0,61 a 0,80 substancialmente concordantes e 0,81 a 0,99 quase perfeitamente concordantes [3].

Uma outra métrica utilizada para análise do modelo foi a acurácia, que é a taxa dos exemplos que são corretamente classificados pelo modelo de treinamento. Ou seja, medida que pode mascarar a eficiência do modelo, caso haja prevalência de uma classe [9].

3 Resultados e Discussões

O presente estudo objetivou diagnosticar o CM como benigno ou maligno com base na aplicação do algoritmo FURIA no *Software* WEKA, onde analisou-se 699 instâncias no total, com 9 variáveis, através da validação cruzada de 10 vezes. Obteve-se como resultados uma acurácia de 94,56%, com 661 instâncias classificadas corretamente e 38 (5,43%) incorretamente. Na Figura 1 é possível observar uma captura de tela do painel de resultados *Weka Explorer*, com trecho da descrição textual da saída gerada pelo algoritmo FURIA.

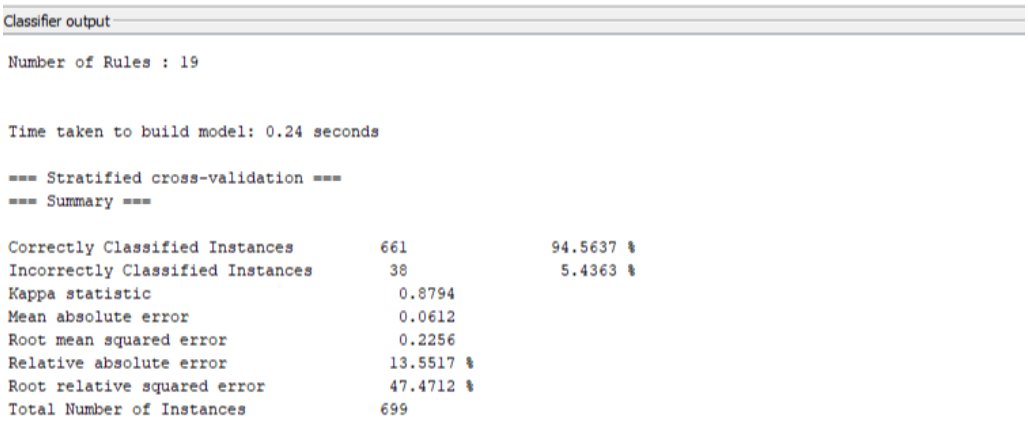

Figura 1: Resultados do FURIA com o banco de dados *Breast Cancer Wisconsin*. Fonte: dos autores, 2024.

Na validação cruzada de 10 vezes, o banco de dados foi fracionado em dez divisões. Na primeira execução, nove dessas foram utilizadas para treinar o modelo e a décima divisão foi utilizada para testar o modelo. Na execução seguinte, uma nova divisão dos dados foi utilizada como teste e o restante como treinamento. Assim, cada divisão feita no banco de dados foi apresentada como

teste e o processo é repetido dez vezes. Os resultados, ou seja, a acurácia e a estimativa do erro, são calculados como uma média dessas dez execuções. O WEKA utiliza o algoritmo FURIA para a décima primeira execução para produzir um modelo que possa ser usado para prever classes [2].

A matriz confusão com as frequências para cada classe do modelo pode ser observada na Tabela 2. Também chamada de tabela de contingência, no caso do presente estudo gerou-se uma matriz 2x2, classes Positiva e Negativa, ou seja, benigno ou maligno, respectivamente. Ressalta-se que o número de instâncias identificadas corretamente é a soma da diagonal principal da matriz.

Tabela 2: Matriz de confusão.

Banco de dados	Benigno	Maligno
Benigno	440	18
Maligno	20	221

A matriz de confusão obtida pelo algoritmo forneceu meios para a avaliação do modelo nas classes. As instâncias verificadas na matriz que foram corretamente identificadas com o diagnóstico investigado foram 661, com acurácia de 94,56% e as instâncias incorretas foram 38, com taxa de erro de 5,43%. Na área de saúde, a acurácia é uma das medidas mais utilizadas para avaliar o desempenho [9]. Neste estudo, através do algoritmo FURIA a acurácia demonstrou um desempenho satisfatório em relação ao total de instâncias corretas.

O valor do Coeficiente Kappa fornecido pela validação foi de 87,94%, ou seja, quase perfeitamente concordante. Dessa maneira, tenta-se assegurar a uniformidade do processo de avaliação de modo a controlar ou minimizar vieses nas conclusões e/ou análises subsequentes. Ressalta-se que o conjunto de dados *Breast Cancer Wisconsin* também foi analisado por [4].

A partir do FURIA, foram geradas dezenove regras *Fuzzy* utilizando funções de pertinências trapezoidais, algumas regras do modelo podem ser observadas a seguir:

- SE (Uniformidade do Tamanho pertence a [1, 1, 2, 3]) E (Núcleo sem Revestimento pertence a [1, 1, 2, 4]) E (Espessura do Aglomerado de Células pertence a [1, 1, 7, 8]) ENTÃO Classe = benigno (CF = 1,0);

- SE (Uniformidade do Formato pertence a [1, 1, 2, 3]) E (Espessura do Aglomerado de Células pertence a [1, 1, 5, 6]) E (Adesão Marginal pertence a [1, 1, 2, 3]) ENTÃO Classe = benigno (CF = 1,0);

- SE (Uniformidade do Formato pertence a [1, 3, 10, 10]) E (Suavidade da Cromatina pertence a [3, 4, 10, 10]) E (Núcleo sem Revestimento pertence a [8, 9, 10, 10]) ENTÃO Classe = maligno (CF = 0,99).

Uma generalização ou "ampliação" de uma regra é obtida pela exclusão de um ou mais de seus antecedentes. Assim, a generalização mínima de uma regra é obtida excluindo todos os antecedentes que não são satisfeitos pela consulta na iteração do algoritmo. Uma vez com as generalizações mínimas das regras, os autores propuseram [4] uma reavaliação de cada regra no conjunto de treinamento e, em seguida, ativa-se a regra com o maior valor de Pureza.

Com base nisso, as regras geradas pelo algoritmo FURIA podem ser utilizadas para prever a melhor decisão entre uma gama de possibilidades, e subsidiar a efetividade do diagnóstico do CM em diferentes cenários e também a possibilidade de explicar a decisão com base das informações de entrada.

4 Considerações Finais

Aplicou-se o FURIA no diagnóstico de dois possíveis CM (Benigno e Maligno), utilizando-se o *Software* WEKA. Observou-se que o FURIA gerou resultados satisfatórios com uma acurácia de 94,56%, dezenove regras a partir de funções de pertinência trapezoidais e coeficiente Kappa de 87,94%. Como proposta futura, é possível avaliar outros algoritmos mais recentes com técnica de seleção de atributos, por exemplo.

Ademais, identificou-se ainda um elevado Fator de Certeza das regras geradas pelo FURIA, o que demonstra a alta adequação do modelo ao conjunto de dados *Breast Cancer Wisconsin*. Como vantagem, evidencia-se a possibilidade de diagnosticar de maneira mais eficiente o CM com base nas informações de entrada.

De forma relevante, o estudo apresentou uma potencial metodologia para subsidiar o diagnóstico precoce do CM, bem como justificá-lo a partir das regras ativadas para cada caso. A detecção precoce pode ocasionar a diminuição de: procedimentos desnecessários, gastos ao serviço de saúde, e interpretações errôneas de exames de imagem. Como limitação, ressalta-se que neste artigo o foco principal foi apenas em identificar se o CM é maligno ou benigno, sem considerar variáveis externas da doença e o prognóstico pós-diagnóstico da patologia.

Referências

[1] M. Arnold, E. Morgan, H. Rumgay, A. Mafra, D. Singh, M. Laversanne, J. Vignat, J. R. Gralow, F. Cardoso, S. Siesling e I. Soerjomataram. "Current and future burden of breast cancer: Global statistics for 2020 and 2040". Em: **The Breast** 66 (2022), pp. 15–23. DOI: 10.1016/j.breast.2022.08.010.

[2] K. P. S. Attwal e A. S. Dhiman. "Exploring data mining tool-Weka and using Weka to build and evaluate predictive models". Em: **Advances and Applications in Mathematical Sciences** 19.6 (2020), pp. 451–469.

[3] J. L. Fleiss e J. Cohen. "The Equivalence of Weighted Kappa and the Intraclass Correlation Coefficient as Measures of Reliability". Em: **Educational and Psychological Measurement** 33 (1973), pp. 613–619. DOI: 10.1177/001316447303300309.

[4] J. Hühn e E. Hüllermeier. "FURIA: an algorithm for unordered fuzzy rule induction". Em: **Data Mining and Knowledge Discovery** 19 (2009), pp. 293–319. DOI: 10.1007/s10618-009-0131-8.

[5] S. Loibl, S. N. Han, V. G. Minckwitz, M. Bontenbal, A. Ring, J. Giermek, T. Fehm, V. K. Calsteren, S. C. Linn, B. Schlehe, M. M. Gziri, P. J. Westenend, V. Müller, L. Heyns, B. Rack, B. V. Calster, N. Harbeck, M. Lenhard, M. J. Halaska, M. Kaufmann, V. Nekljudova e F. Amant. "Treatment of breast cancer during pregnancy: an observational study". Em: **The lancet oncology** 13.9 (2012), pp. 887–896. DOI: 10.1016/S1470-2045(12)70261-9.

[6] M. O. Santos, F. C. S. Lima, L. F. L. Martins, J. F. P. Oliveira, L. M. Almeida e M. C. Cancela. "Estimativa de Incidência de Câncer no Brasil, 2023-2025". Em: **Revista Brasileira de Cancerologia** 69 (2023). DOI: 10.32635/2176-9745.RBC.2023v69n1.3700.

[7] E. A. M. G. Soares, L. C. L. Damascena, L. M. M. Lima e R. M. Moraes. "Analysis of the Fuzzy Unordered Rule Induction Algorithm as a Method for Classification". Em: **Quinto Congresso Brasileiro de Sistemas Fuzzy (V CBSF)**. 2018, pp. 17–28. ISBN: 978-85-8215-085-6.

[8] S. V. Stehman. “Selecting and interpreting measures of thematic classification accuracy”. Em: **Remote sensing of Environment** 62 (1997), pp. 77–89. DOI: 10.1016/S0034-4257(97)00083-7.

[9] J. A. Swets. “Measuring the Accuracy of Diagnostic Systems”. Em: **Science** 240 (1988), pp. 1285–1293. DOI: 10.1126/science.3287615.

[10] W. Wolberg. **Breast Cancer Wisconsin (Original)**. UCI Machine Learning Repository. 1992. DOI: 10.24432/C5HP4Z.

