Proceeding Series of the Brazilian Society of Computational and Applied Mathematics, v. 12, n. 1, 2026.

Trabalho apresentado no XLIV CNMAC, Fundação Getulio Vargas - Rio de Janeiro - RJ, 2025.

Proceeding Series of the Brazilian Society of Computational and Applied Mathematics

Uma Nova Abordagem Baseada em PSO e Naive Bayes para Análise de Base de Dados Desbalanceada de Doença de Chagas

André G. Coimbra[1], Matheus P. Libório[2]
PPGMCS/UNIMONTES, Montes Claros, MG
João B. Mendes[3], Marcos F. S. V. D'Angelo[4]
DCC/UNIMONTES, Montes Claros, MG
Reinaldo M. Palhares[5]
DELT/UFMG, Belo Horizonte, MG

Resumo. Este artigo propõe uma abordagem híbrida combinando Otimização por Enxame de Partículas (PSO) e Naive Bayes (NB) para melhorar a classificação de dados desbalanceados relacionados à sobrevida de pacientes com Doença de Chagas. O objetivo é otimizar a seleção de instâncias e variáveis, ajustando hiperparâmetros do NB para aumentar a acurácia e generalização do modelo. A metodologia emprega PSO na etapa de pré-processamento, seguida pela classificação com NB, avaliando desempenho por meio de métricas como Area Under the Receiver Operating Characteristic Curve (AUC-ROC), F1-Score e sensibilidade. Os resultados mostram que a abordagem PSO+NB supera métodos tradicionais, com melhorias estatisticamente significativas ($p < 0{,}05$) na classificação de casos minoritários. Os resultados evidenciam que a técnica proposta é eficaz para lidar com desbalanceamento de dados, oferecendo uma ferramenta promissora para auxiliar no diagnóstico em regiões com recursos limitados.

Palavras-chave. Particle Swarm Optimization, Naive Bayes, Dados Desbalanceados, Doença de Chagas

1 Introdução

A disponibilidade de grandes e complexos conjuntos de dados provenientes de pacientes e instalações médicas tem contribuído significativamente para a aplicação de métodos de aprendizagem de máquina no campo da saúde [14, 16]. Essas técnicas de aprendizagem de máquina podem analisar rapidamente grandes volumes de dados e gerenciar de forma eficaz as relações complexas dentro deles. Como resultado, elas demonstraram potencial para melhorar os indicadores de saúde e qualidade de vida, além de possibilitar que especialistas em saúde avancem na pesquisa clínica [15].

A principal contribuição deste trabalho é a proposta de uma nova abordagem que integra a Otimização por Enxame de Partículas (PSO) [7] com o classificador Naive Bayes [9] para enfrentar o desafio da classificação de dados desbalanceados, detalhado na Seção 2. Utilizando o PSO, conseguimos selecionar de maneira eficaz instâncias para balancear o conjunto de dados, garantindo

[1]andregcoimbra@gmail.com
[2]m4th32s@gmail.com
[3]joao.mendes@unimontes.br
[4]marcos.dangelo@unimontes.br
[5]rpalhares@ufmg.br

uma representação mais equitativa das classes. Posteriormente, empregamos o classificador Naive Bayes para a tarefa de classificação. Esta metodologia integrada foi especificamente aplicada para prever a mortalidade de pacientes dois anos antes do evento ocorrer em um conjunto de dados desbalanceado relacionado à doença de Chagas [12], demonstrando sua eficácia em melhorar o desempenho da classificação, conforme discutido na Seção 3. Nossos resultados sugerem que essa abordagem melhora significativamente a precisão e a confiabilidade do diagnóstico da doença de Chagas [5], oferecendo uma ferramenta valiosa para pesquisadores e profissionais da área médica que enfrentam desafios semelhantes de classificação.

2 Otimização por Enxame de Partículas (PSO) e Naive Bayes para Classificação de Dados Desbalanceados

Esta seção descreve uma nova abordagem para a classificação de dados desbalanceados usando o algoritmo de Otimização por Enxame de Partículas (PSO). O PSO é utilizado para selecionar as melhores instâncias, variáveis e parâmetros na construção de um modelo de classificação. Estudos anteriores exploraram o uso da Otimização por Enxame de Partículas (PSO) para lidar com dados desbalanceados, embora frequentemente focando em aspectos isolados. Por exemplo, Ping Cao et al. [1] propôs um método para classificar dados desbalanceados usando PSO para otimizar uma rede neural, melhorando o desempenho da classificação ao penalizar erros da classe minoritária e ajustar a importância das variáveis. Por outro lado, [6] sugeriu uma combinação de PSO com o classificador Naive Bayes (NB) para selecionar os melhores dados de treinamento, melhorando a eficiência do modelo. No entanto, o trabalho proposto realiza simultaneamente a seleção de instâncias, a seleção de variáveis e o ajuste de parâmetros. O principal objetivo é alcançar uma maior precisão e generalização, avaliando o F1-Score para cada classe, enquanto reduz a complexidade do modelo através da seleção adaptativa de variáveis.

A combinação da Otimização por Enxame de Partículas (PSO) com o Naive Bayes (NB) é particularmente promissora porque aproveita as forças complementares de cada técnica. A PSO é hábil em explorar o espaço de soluções ajustando eficientemente os parâmetros. Enquanto isso, o NB é um classificador probabilístico que assume independência condicional entre as variáveis e é notável por sua eficiência computacional, especialmente com grandes conjuntos de dados. No entanto, como observado por Zhang [17], o NB pode levar a estimativas imprecisas quando as suposições de independência não são atendidas. A sinergia dessas duas abordagens é poderosa porque a PSO pode otimizar os parâmetros do NB, melhorando a precisão da classificação ao encontrar configurações mais adequadas. Assim, esta combinação melhora a capacidade do modelo de lidar com dados do mundo real, onde as variáveis são frequentemente correlacionadas.

O estudo se desenvolve em duas etapas principais. A primeira etapa envolve a normalização de dados, abordada na Seção 2.1, crucial no pré-processamento, especialmente ao trabalhar com algoritmos de aprendizagem de máquina sensíveis às escalas das variáveis. A segunda etapa, detalhada na Seção 2.2, lida com a configuração do algoritmo PSO e descreve o desenvolvimento da função objetivo. Esta função avalia a qualidade das soluções obtidas pela PSO, que seleciona instâncias e variáveis com base em limiares estabelecidos, treina um modelo Naive Bayes com os dados selecionados e calcula o F1-Score ponderado das previsões. A função também incorpora penalização adaptativa para equilibrar a complexidade do modelo.

2.1 Padronização de Dados

A padronização de dados é uma técnica amplamente utilizada em estatística e aprendizagem de máquina para transformar diferentes variáveis em uma escala comum, facilitando a comparação e análise dos dados. Um dos métodos mais comuns de padronização é o uso do *Z-score* [8].

DOI: 10.5540/03.2026.012.01.0347

O *Z-score*, também conhecido como escore padronizado, é uma medida que descreve a posição de um valor em relação à média de um conjunto de dados, em unidades de desvio padrão. O *Z-score* de um valor x é calculado conforme equação (1).

$$z = \frac{x - \mu}{\sigma} \tag{1}$$

onde:

- x é o valor sendo padronizado.
- μ é a média do conjunto de dados.
- σ é o desvio padrão do conjunto de dados.

O *Z-score* indica quantos desvios padrão um valor está acima ou abaixo da média. Um *Z-score* positivo indica que o valor está acima da média, enquanto um *Z-score* negativo indica que o valor está abaixo da média. Por exemplo:

- Um *Z-score* igual a 0 significa que o valor é igual à média.
- Um *Z-score* igual a 1 significa que o valor é um desvio padrão acima da média.
- Um *Z-score* igual a -1 significa que o valor é um desvio padrão abaixo da média.

Padronizar dados usando o *Z-score* é particularmente útil ao comparar dados que estão em diferentes escalas ou unidades. Ao converter os dados em *Z-scores*, todas as variáveis terão uma média de zero e um desvio padrão de um, tornando-as diretamente comparáveis.

O *Z-score* é uma ferramenta poderosa para a padronização de dados, permitindo que variáveis de diferentes escalas sejam comparadas uniformemente. Sua aplicação é essencial em vários campos de estatística e aprendizagem de máquina, onde a comparação precisa de dados é crucial.

2.2 Otimização por Enxame de Partículas (PSO)

A PSO é um algoritmo de otimização baseado em população. Cada partícula, inicializada aleatoriamente, representa uma solução potencial e se move pelo espaço de busca influenciada pela sua melhor posição encontrada e pela melhor posição encontrada de seus vizinhos [2].

2.2.1 Parâmetros

A definição de parâmetros é de extrema relevância devido ao seu impacto na estratégia de exploração de possíveis soluções.

Neste trabalho, o critério usado para selecionar variáveis e instâncias da solução potencial fornecida pela PSO foi a configuração de um limiar de seleção. Ou seja, a variável ou instância é selecionada quando a sugestão, dentro dos limites do espaço de busca, excede o limiar pré-definido.

O peso de inércia é atualizado usando o método de resfriamento em cada iteração, o componente cognitivo foi progressivamente aumentado em cada iteração enquanto o componente social permaneceu estático. Para convergir rapidamente e eficientemente, foi adotada a exploração de vizinhança global, onde cada partícula pode se comunicar diretamente com todas as outras influenciadas pelo melhor resultado encontrado por qualquer partícula no enxame.

DOI: 10.5540/03.2026.012.01.0347

2.2.2 Função Objetivo

A função objetivo desempenha um papel central no processo de otimização e é responsável por avaliar, em cada iteração, a qualidade de cada solução candidata gerada pelo algoritmo de Otimização por Enxame de Partículas (PSO) — ou seja, uma combinação específica de variáveis e instâncias selecionadas. Neste trabalho, a função objetivo foi definida com os seguintes passos:

- Definição dos dados de treinamento a partir da seleção de variáveis e instâncias com base no limiar configurado.
- Execução do modelo de classificação Naive Bayes com as instâncias, variáveis e parâmetro de suavização selecionados.
- Previsão usando o modelo treinado no conjunto de teste.
- Cálculo do F1-Score ponderado das classes.
- Aplicação de penalização ao F1-Score com base no número de variáveis selecionadas para controlar a complexidade do modelo, conforme equação (2)

$$F_{objetivo} = (F1C_0 \cdot p_1) + (F1C_1 \cdot p_2) - P \cdot N^2 \tag{2}$$

onde:

- $F1C_0$ é o F1-Score para a classe 0.
- $F1C_1$ é o F1-Score para a classe 1.
- p_1 e p_2 são os pesos atribuídos a cada classe.
- P é o fator de penalização.
- N é o número de variáveis selecionadas.

3 Aplicação em uma Base de Dados Desbalanceada de Doença de Chagas

A Doença de Chagas (DC) é reconhecida como uma doença tropical negligenciada pela OMS e continua sendo um problema de saúde pública. Estima-se que 30% dos pacientes podem desenvolver anormalidades cardíacas que podem levar à morte [5]. Na América Latina, cerca de 5,7 milhões de pessoas estão infectadas, com uma taxa de mortalidade anual de 12.000 casos [5].

Estudos indicam que 80% dos infectados não têm acesso a diagnóstico e tratamento adequados, resultando em alta mortalidade e custos sociais significativos [10]. Nesse contexto, a Aprendizagem de Máquina tem se mostrado promissora para definir intervenções e reduzir o impacto da DC [5, 11]. Uma ferramenta que preveja o risco de morte com antecedência pode ajudar os profissionais de saúde, especialmente em regiões com acesso limitado a exames complexos.

Aplicamos nossa abordagem a um conjunto de dados sobre DC [5], com o objetivo de prever a mortalidade do paciente dois anos antes do evento. O conjunto de dados inclui variáveis de entrevista e exames complementares, totalizando 128 variáveis preditoras, além da classe 'morte' ou 'não morte' em dois anos. Os dados são da Coorte SaMi-Trop, com 551 pacientes com DC de 21 municípios em Minas Gerais, dos quais 134 (24,32%) morreram dentro de 2 anos, indicando desbalanceamento considerável.

DOI: 10.5540/03.2026.012.01.0347

Os dados foram coletados entre 2013 e 2016. Para validação do modelo, o conjunto de dados foi dividido de maneira estratificada, com 80% para treinamento e 20% para teste, garantindo a proporcionalidade da classe. A normalização de dados foi realizada usando a técnica de Z-score.

Para estabelecer uma linha de base, foram realizadas execuções de classificação usando apenas o algoritmo Naive Bayes. Os resultados, apresentados na Tabela 1, mostraram baixas precisões gerais e de classe, refletindo o desbalanceamento das classes.

Tabela 1: Classificador Naive Bayes.

Execução	Acc	R+	P+	F1+	F2	R-	P-	F1-	AUC-ROC	Variáveis	Instâncias de Treino
1	0.31	0.93	0.25	0.39	0.60	0.11	0.82	0.19	0.52	128	440
2	0.42	1.00	0.30	0.46	0.68	0.24	1.00	0.38	0.62	128	440
3	0.34	0.81	0.24	0.38	0.56	0.19	0.76	0.30	0.50	128	440
4	0.66	0.93	0.41	0.57	0.74	0.57	0.96	0.72	0.75	128	440
5	0.37	0.89	0.26	0.41	0.60	0.20	0.85	0.33	0.55	128	440
6	0.41	1.00	0.29	0.45	0.68	0.23	1.00	0.37	0.61	128	440
7	0.41	0.96	0.29	0.44	0.65	0.23	0.95	0.37	0.59	128	440
8	0.66	0.74	0.39	0.51	0.63	0.63	0.88	0.74	0.69	128	440
9	0.43	0.93	0.29	0.44	0.64	0.27	0.92	0.42	0.60	128	440
10	0.28	0.93	0.24	0.38	0.59	0.07	0.75	0.13	0.50	128	440

A Tabela 2 descreve os parâmetros utilizados na configuração de otimização do modelo PSO. O *Intervalo de Busca de Variáveis e Instâncias* foi definido com base na natureza do problema, onde o objetivo é decidir se deve-se selecionar uma variável ou instância específica. O *Intervalo de Busca de Suavização NB* foi determinado empiricamente. O valor do *Limiar de Variável* foi ajustado para aumentar a probabilidade de selecionar o menor número possível de variáveis, reduzindo assim a complexidade do modelo, o que é ainda mais aprimorado pela configuração da variável Penalidade (P). Esta penalidade foi calculada como o inverso do número total de variáveis preditoras, o que significa que selecionar mais variáveis incorre em uma penalidade maior. O *Limiar de Instância* foi configurado para balancear a probabilidade de selecionar uma instância. O Peso (p_1 e p_2) foi atribuído para garantir que o F1-Score de ambas as classes seja o mesmo. O *Peso de Inércia* (w) foi atualizado em cada execução usando o intervalo testado no estudo [13]. Para as variáveis *Cognitiva* (c_1) e *Social* (c_2), foi observada a valor 2.0 sugerido em [3]; no entanto, para cada execução, c_1 foi atualizado linearmente, aumentando gradualmente a ênfase nas contribuições individuais. A variável *Velocidade* foi ajustada para metade do espaço de busca para uma exploração controlada. As variáveis *Número de partículas* e *Iterações* foram escolhidas empiricamente.

Tabela 2: Parâmetros do PSO e Função Objetivo.

Descrição	Valor
Intervalo de Busca de Variáveis e Instâncias	[0, 1]
Intervalo de Busca de Suavização NB	[1, 1e-9]
Limiar de Variável	0.7
Limiar de Instância	0.5
Penalidade (P)	7.81E-03
Peso (p_1 e p_2)	0.5
Peso de Inércia (w)	0.9 a 0.4
Cognitiva (c_1)	1.6 a 2.0
Social (c_2)	2.0
Número de partículas	50
Velocidade	-0.5 a 0.5
Iterações	200
Inicialização	Aleatória
Exploração de Vizinhança	Global

Os resultados obtidos na execução do modelo PSO+NB mostraram um ganho substancial, tanto no desempenho geral quanto no desempenho de classes individuais, conforme observado na Tabela 3. Neste contexto, vale destacar que a média de AUC-ROC foi de 0.59 para as execuções de NB, 0.70 para GA+NB, 0.73 para MLP e 0.81 para PSO+NB. A AUC-ROC mede a capacidade de discriminação do classificador [4]. Valores mais próximos de 1 indicam melhor desempenho na classificação das classes. Os resultados claramente mostram que a abordagem do modelo PSO+NB proporcionou melhorias significativas na precisão geral e nas métricas de precisão, recall, AUC-ROC e F1 para ambas as classes. O modelo de classificação otimizado conseguiu encontrar um conjunto de variáveis e testes capazes de lidar com o desbalanceamento de dados, resultando em uma classificação mais equilibrada. Esses resultados superam os apresentados por [5], indicando que a metodologia proposta pode ser aplicada a diversos outros problemas no campo da saúde que sofrem com questões de desbalanceamento de classes.

Tabela 3: Otimização PSO + Classificador Naive Bayes.

Execução	Acc	R+	P+	F1+	F2	R-	P-	F1-	AUC-ROC	Variáveis	Instâncias de Treinamento
1	0.85	0.56	0.75	0.64	0.59	0.94	0.87	0.90	0.75	1	226
2	0.85	0.85	0.64	0.73	0.80	0.84	0.95	0.89	0.85	2	240
3	0.87	0.81	0.71	0.76	0.79	0.89	0.94	0.91	0.85	2	220
4	0.82	0.63	0.63	0.63	0.63	0.88	0.88	0.88	0.76	2	218
5	0.82	0.96	0.58	0.72	0.85	0.77	0.98	0.87	0.87	2	215
6	0.90	0.85	0.77	0.81	0.83	0.92	0.95	0.93	0.88	3	209
7	0.85	0.59	0.73	0.65	0.62	0.93	0.88	0.90	0.76	2	235
8	0.81	0.85	0.58	0.69	0.78	0.80	0.94	0.86	0.82	2	219
9	0.80	0.70	0.58	0.63	0.67	0.83	0.90	0.86	0.77	2	227
10	0.85	0.56	0.75	0.64	0.59	0.94	0.87	0.90	0.75	1	225

4 Considerações Finais

Os resultados apresentados nesse artigo indicam que nossa abordagem poderia melhorar significativamente a precisão e a confiabilidade do diagnóstico da doença de Chagas. Isso oferece uma ferramenta valiosa para pesquisadores e profissionais da área médica que enfrentam o desafio dos dados desbalanceados. Em cenários de saúde reais, falsos positivos podem causar ansiedade desnecessária e levar a testes e tratamentos infundados, enquanto falsos negativos podem resultar em diagnósticos perdidos e intervenções críticas atrasadas. Ao lidar efetivamente com dados desbalanceados, nossa metodologia permitiria prever o risco de mortalidade para indivíduos com condições de saúde raras, auxiliando os profissionais de saúde, especialmente em regiões com acesso limitado a procedimentos diagnósticos complexos. Essas implicações práticas destacam a relevância de nossa abordagem na melhoria dos resultados dos pacientes e na otimização dos recursos de saúde.

Agradecimentos

Os autores agradecem a FAPEMIG, CNPq e CAPES pelo suporte financeiro.

Referências

[1] P. Cao, D. Zhao e O. R. Zaïane. "A PSO-Based Cost-Sensitive Neural Network for Imbalanced Data Classification". Em: **Trends and Applications in Knowledge Discovery and Data**

Mining. PAKDD 2013. Vol. 7867. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, 2013, Cost–Sensitive Neural Network for Imbalanced Data Classification.

[2] L. N. de Castro. **Fundamentals of Natural Computing**. 1ª ed. Boca Raton: Taylor & Francis Group, 2007.

[3] M. Clerc e J. Kennedy. "The particle swarm-explosion, stability, and convergence in a multidimensional complex space". Em: **IEEE transactions on Evolutionary Computation** 6.1 (2002), pp. 58–73.

[4] T. Fawcett. "An introduction to ROC analysis". Em: **Pattern Recognition Letters** 27.8 (2006). ROC Analysis in Pattern Recognition, pp. 861–874. ISSN: 0167-8655. DOI: https://doi.org/10.1016/j.patrec.2005.10.010.

[5] A. M. Ferreira et al. "Two-year death prediction models among patients with Chagas Disease using machine learning-based methods". Em: **PLoS Negl Trop** 2 (2022). DOI: 10.1371/journal.pntd.0010356.

[6] N. K. Ghanad e S. Ahmadi. "Combination of PSO Algorithm and Naive Bayesian Classification for Parkinson Disease Diagnosis". Em: **Advances in Computer Science: An International Journal** 4.4 (2015), pp. 119–125.

[7] J. Kennedy e R. Eberhart. "Particle swarm optimization". Em: **Proceedings of the IEEE International Conference on Neural Networks** (1995), pp. 1942–1948. DOI: 10.1109/ICNN.1995.488968.

[8] D. S. Moore, G. P. McCabe e B. A. Craig. **Introduction to the Practice of Statistics**. 7th. New York: W.H. Freeman e Company, 2012. ISBN: 9781429240321.

[9] R. E. Neapolitan. **Learning Bayesian Networks**. Pearson Prentice Hall, 2003.

[10] M. C. P. Nunes et al. "Chagas cardiomyopathy: an update of current clinical knowledge and management: a scientific statement from the American Heart Association". Em: **Circulation** 138.12 (2018), e169–e209.

[11] C. C. R. Sady e A. L. P. Ribeiro. "Symbolic features and classification via support vector machine for predicting death in patients with Chagas disease". Em: **Computers in Biology and Medicine** 70 (2016), pp. 220–227.

[12] SaMi-Trop. **Chagas disease dataset**. Acesso em: 7 jul. 2024. 2022. URL: http://journals.plos.org/plosntds/article/asset?unique&id=info:doi/10.1371/journal.pntd.0010356.s003.

[13] Y. Shi e Eberhart. "Particle swarm optimization: developments, applications and resources". Em: **Proceedings of the 2001 congress on evolutionary computation (IEEE Cat. No. 01TH8546)**. Vol. 1. IEEE. 2001, pp. 81–86.

[14] H. H. Tseng et al. "Machine Learning and Imaging Informatics in Oncology". eng. Em: **Oncology** 98.6 (2020), pp. 344–362. ISSN: 0030-2414.

[15] J. Waring, C. Lindvall e R. Umeton. "Automated machine learning: Review of the state-of-the-art and opportunities for healthcare". Em: **Artificial Intelligence in Medicine** 104 (2020), p. 101822.

[16] J. Wiens e E. S. Shenoy. "Machine learning for healthcare: on the verge of a major shift in healthcare epidemiology". Em: **Clinical Infectious Diseases** 66.1 (2018), pp. 149–153.

[17] H. Zhang. "The optimality of naive Bayes". Em: **Proceedings of the Seventeenth International Florida Artificial Intelligence**. 2004.

