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

Usando Regressão Simbólica Através da Metaheurística BRKGP para Encontrar Modelos de Aplicação na Física com Auxílio da Análise Dimensional

Filipe Pessoa Sousa[1]
PPG-Ccomp/UERJ, Rio de Janeiro, RJ
Cristiane Oliveira de Faria[2]
IME/UERJ, Rio de Janeiro, RJ
Igor Machado Coelho[3]
IC/UFF, Niterói, RJ

Resumo. Este estudo propõe refinamentos na identificação representativa de funções usando o algoritmo BRKGP para fenômenos físicos. A implementação da Biblioteca GiNaC em C++ adiciona uma camada ao algoritmo, permitindo operações e visualização de grandezas físicas nas funções encontradas. O algoritmo BRKGP é empregado para validar os resultados por meio de abordagens diversificadas. Destaca-se a suscetibilidade do domínio das variáveis e parâmetros a erros de arredondamento. Além disso, a análise evidencia a inclinação do algoritmo em selecionar menos variáveis, aderindo ao princípio da navalha de Occam.

Palavras-chave. Otimização, BRKGP, Regressão Simbólica, Meta-heurística, Programação Genética

1 Introdução

O campo da Modelagem Matemática desempenha um protagonismo na interpretação e descrição dos fenômenos, principalmente no meio físico[4]. Entretanto, é perceptível que, geralmente, modelar relações funcionais e precisas entre as variáveis envolvidas (seja descrita por fórmulas ou por funções), especialmente em sistemas físicos e não lineares é um desafio.

Com o propósito de encontrar estas expressões matemáticas de forma simbólica, neste trabalho, é apresentado uma abordagem estruturada a partir da técnica de Regressão Simbólica combinada com o algoritmo de otimização BRKGP (*Biased Random Key Genetic Programming*) [3]. Através desta metodologia é possível explorar um amplo espaço de soluções, levando em consideração diferentes combinações de funções matemáticas e parâmetros [5]. Fazendo uso de Algoritmos Genéticos, a busca por soluções é guiada por critérios de otimização, como a minimização do erro entre os valores observados e os valores previstos pelo modelo [8].

No âmbito da pesquisa em Física, a aplicação da Regressão Simbólica, especificamente com o algoritmo BRKGP, ainda carece de um aprofundamento significativo. Embora outros métodos tenham sido empregados para abordar essa questão, destacam-se as vantagens observadas na identificação de relações não lineares e interações complexas entre variáveis, o que contribui substancialmente para uma compreensão mais refinada dos fenômenos físicos e fenômenos correlatos [9,

[1]Filipe.pessoa18@gmail.com
[2]cofaria@ime.uerj.br
[3]imcoelho@ic.uff.br

10]. Entretanto, é crucial salientar a importância de considerar o contexto específico do fenômeno físico em análise, juntamente com as limitações inerentes aos dados disponíveis. Além disso, a validação dos modelos obtidos é fundamental, sendo necessária a realização de testes independentes e a comparação dos resultados com dados experimentais [9, 10]. Em Sousa et. al [7], estudos utilizando a técnica de Programação Genética com Chaves Aleatórias Viciadas (do inglês, BRKGP) [3] visando melhorias no seu algoritmo foram feitos. O BRKGP é uma hiper-heurística [1] que estende a técnica do Algoritmo Genético com Chaves Aleatórias Viciadas (do inglês, *Biased Random Keys Genetic Algorithm* – BRKGA) [2] com ideias de Programação Genética.

Nesse trabalho, estendemos este estudo considerando a análise dimensional das variáveis, já que esta é essencial para garantir a consistência e validade das funções físicas, considerando as unidades das grandezas envolvidas. Diferentes operações matemáticas foram discutidas em relação à sua aplicação na análise dimensional, enfatizando a importância da compatibilidade das dimensões das grandezas envolvidas em cada operação. Parâmetros específicos do problema foram ajustados com base na propriedade das unidades para conduzir abordagens sustentáveis. Introduziu-se um novo parâmetro, *moreDiversity*, com resultados benéficos observados, especialmente para a função da Distância Euclidiana.

O presente artigo segue uma estrutura composta por seis seções, sendo esta introdução a primeira delas. Na Seção 2, realiza-se uma concisa revisão da literatura pertinente ao tema abordado. A metodologia proposta e os detalhes técnicos de sua implementação são abordados na Seção 3. Na sequência, na Seção 4, são apresentados os resultados obtidos a partir dos testes de validação. Estes resultados são organizados em tabelas e comparados com a função original, isto é, a função teoricamente proposta encontrada em textos que abordam o fenômeno físico em estudo. Por fim, na Seção 5, são delineadas as conclusões alcançadas, acompanhadas das perspectivas futuras para o aprimoramento tanto da metodologia quanto dos resultados obtidos.

2 Métodos

2.1 Regressão Simbólica

A regressão simbólica constitui uma técnica no domínio do aprendizado de máquina que visa identificar modelos matemáticos simbólicos capazes de descrever a relação existente entre um conjunto de variáveis de entrada e uma variável de saída específica. O procedimento engloba a construção de um conjunto de dados de treinamento, contendo valores previamente conhecidos das variáveis em questão. Posteriormente, emprega-se um algoritmo de regressão simbólica para identificar a função que melhor se ajusta aos dados fornecidos. O intento subjacente a esse processo é a minimização do erro entre as previsões geradas pelo modelo e os valores reais de saída, conforme discutido por [5].

2.2 Programação Genética

A Programação Genética (PG) emerge como uma abordagem amplamente adotada na resolução de problemas associados à regressão simbólica, dada a impraticabilidade de se encontrar soluções analíticas em muitos cenários. A PG, nesse contexto, inicia gerando uma população inicial de expressões simbólicas de forma aleatória. Essas expressões são, então, avaliadas quanto à sua adequação em relação ao conjunto de treinamento, e as mais promissoras são selecionadas para participar do processo de cruzamento e mutação. Esse ciclo é repetido iterativamente até que uma solução satisfatória seja alcançada, conforme explanado por [4].

Em uma ampliação do BRKGA [2], surge o algoritmo BRKGP. Este algoritmo representa uma nova abordagem para a Programação Genética, ao empregar uma representação baseada em

DOI: 10.5540/03.2025.011.01.0507

chaves aleatórias. O BRKGP demonstrou ser capaz de abordar problemas que anteriormente se mostravam impraticáveis para os algoritmos genéticos tradicionais [3].

3 Resultados e Discussão

Este trabalho se destina a implementar melhorias na metodologia proposta. Foi adicionado mais uma camada, onde é avaliado as grandezas dimensionais das funções geradas ao mesmo tempo que estas são construídas. Assim, sendo possível avaliar a eficácia da metodologia da Regressão Simbólica em conjunto com BRKGP, aplicando-a a algumas das funções de interesse já utilizadas em Sousa et al. [7]. Alinhado a essas premissas, foi decidido adotar uma calibração de desempenho que já havia sido previamente examinada com sucesso, conforme evidenciado por Sousa et al. [6].

O cálculo do erro adotado baseou-se no *Root Mean Square Error*, utilizando y_i para representar os valores observados e $\hat{y}_i$ para denotar os valores previstos: $RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$.

Neste estudo, as mesmas funções avaliadas em Sousa et al. [7] serão utilizadas. Uma abordagem distinta foi adotada para o banco de dados. Foram selecionadas 20 amostras da maior amostra (40) e, posteriormente, escolhidas 10 amostras da segunda maior amostra (20). Essa estratégia foi implementada para investigar o impacto que o banco de dados exerce no desenvolvimento do algoritmo. A Tabela 1 apresenta os problemas-testes que serão analisados. Todos os problemas foram retirados de [9].As constantes foram colocadas em negrito, não foi feita distinção entre constantes dimensionais e adimensionais.

Tabela 1: Descrição dos problemas-teste.

Problema	**Função**	**Variáveis e Constantes**
Distância Euclidiana	$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$	4, 0
Massa Relativística	$m = \frac{m_0}{\sqrt{\mathbf{1} - \frac{v^2}{\mathbf{c^2}}}}$	2, 2
Produto Interno	$\mathbf{A} = A_x \cdot B_x + A_y \cdot B_y + A_z \cdot B_z$	6, 0
Força Gravitacional	$\mathbf{F} = \frac{\mathbf{G} \cdot \mathbf{m}_1 \cdot \mathbf{m}_2}{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}$	6, 3
Lentes Delgadas	$f = \frac{\mathbf{1}}{\frac{\mathbf{1}}{d_1} + \frac{n}{d_2}}$	3, 1
Campo Elétrico	$E = \frac{q}{\mathbf{4\pi\varepsilon}_0 r^2}$	2, 3
Densidade de Probabilidade	$f = \frac{\mathbf{e}^{-\frac{\theta}{\mathbf{2}}}}{\sqrt{(\mathbf{2\pi})}}$	1, 3

3.1 Descrição dos Problemas-teste

Os parâmetros globais do BRKGP, conforme descritos no estudo de referência sobre Programação Genética, são os seguintes: - *populationLen* representa o tamanho da população, definido como 100; - *eliteSize* indica a porcentagem de indivíduos considerados elites na população, estabelecida em 25%; - *mutantSize* refere-se à porcentagem de indivíduos mutantes na população, definida como 10%; - *eliteBias* denota a probabilidade de propagação de genes considerados elites, com uma taxa de 85%.

Os valores das constantes utilizadas para cada função foram: - velocidade da luz no vácuo, c, é 3×10^8 metros por segundo; - constante gravitacional, G, que descreve a força da gravidade entre objetos, tem um valor de cerca de 6.67408×10^{-11} metros cúbicos por quilograma por segundo ao quadrado; A massa da Terra, m_1, é aproximadamente 5.972×10^{24} quilogramas; enquanto a massa da Lua, m_2, é de cerca de 7.34×10^{22} quilogramas; a permissividade do vácuo, "ε_0", é de aproximadamente 8.854×10^{-12} Faraday por metro. Os números fundamentais, como o número de Euler, e, com um valor de 2.71828, e o número Pi, π, com um valor de 3.14159.

DOI: 10.5540/03.2025.011.01.0507

O escopo deste estudo concentrou-se na validação do algoritmo empregado, com ênfase em cada função específica. Em cada execução, o algoritmo foi testado com funções unárias, que envolviam operações de potência (a), raiz (r) e exponenciação (e), bem como funções binárias, incorporando operações de soma, subtração, multiplicação e divisão. O processo de validação foi conduzido mediante a atribuição de diferentes sementes em cada execução, proporcionando robustez à exploração do espaço de soluções.

Adicionalmente, foram introduzidos novos parâmetros de entrada que impactam a calibração do algoritmo. Esses parâmetros influenciam aspectos cruciais da otimização, como o número de reinícios *(restartMax)*, o máximo de gerações sem melhora *(NoImprovement)*. A inclusão desses novos parâmetros visa expandir a flexibilidade do algoritmo, permitindo uma exploração mais abrangente do espaço de soluções. No Algoritmo **??** temos o pseudocódigo do BRKGP proposto com as modificações.

Algoritmo 1: Algoritmo BRKGP

Entrada: Parâmetros do BRKGP P, Conjunto de dados de entrada $Q = (Inp, Out)$, Critérios de parada $(restartMax, noImprovementMax)$

Saída: Melhor indivíduo I com custo f

$restart \leftarrow 0$;
$custoMelhorIndividuo \leftarrow \infty$;
$restart < restartMax$ $noImprovement \leftarrow 0$;
$populacao \leftarrow inicializarPopulacao(P)$;
$populacao \leftarrow decodificar(populacao, Q)$;
$populacao \leftarrow avaliarCusto(populacao)$;
$custoMelhorIndividuo \leftarrow obterMelhorCusto(populacao)$;
$noImprovement < noImprovementMax$
$populacaoMutante \leftarrow aplicarCrossover(P, populacao)$;
$populacaoMutante \leftarrow decodificar(populacaoMutante, Q)$;
$populacaoMutante \leftarrow avaliarCusto(populacaoMutante)$;
se $custoMelhor(populacaoMutante) < custoMelhorIndividuo$ **então**
    $custoMelhorIndividuo \leftarrow custoMelhor(populacaoMutante)$;
    $noImprovement \leftarrow 0$;
fim
senão
    $noImprovement \leftarrow noImprovement + 1$;
fim
$populacao \leftarrow populacaoMutante$;
$restart \leftarrow restart + 1$;
retorna $custoMelhorIndividuo$, I

4 Análise dos Resultados

Na Tabela 2, apresentamos os resultados obtidos por meio da aplicação do Algoritmo **??** com auxílio da análise dimensional para encontrar funções simbólicas que representam problemas físicos complexos. Nossa análise engloba uma comparação direta entre as funções originais (veja Tabela 1), que descrevem os fenômenos físicos, e as funções simbólicas derivadas pelo algoritmo. Esta comparação pretende avaliar a capacidade do algoritmo em identificar relações matemáticas precisas a partir de dados experimentais.

DOI: 10.5540/03.2025.011.01.0507

Tabela 2: Funções encontradas.

Problema	Equação Encontrada
Distância Euclidiana	$\sqrt{\sqrt{((x_3 - x_2)^2 + ((((1 - (1 * 1)) - (1 * 1))/1) * (x_1 - x_0))^2)}}^2$
Massa Relativística	$\frac{\sqrt{m_0}\sqrt{(m_0+(m_0+m_0))+\frac{m_0+m_0}{\frac{3\times10^8}{v}-1}\times1\times1\times v\times1}}{v+v}$
Lentes Delgadas	$n \times \frac{1}{\left(\frac{n}{d_2}+\frac{1}{d_1}\right)\sqrt{n^2}}$
Produto Interno	$x_2x_3 + 1\left(x_1x_0 + \left(x_5x_4\sqrt{\frac{(1\times1)^2}{1}}\right)\sqrt{1^2}\right)$
Força Gravitacional	$(\frac{\frac{1}{734}(((x_5(((x_0\times x_0)x_5)x_5))\times\frac{x_3}{0.667408+0.667408})\times0.667408^2)0.667408^2)}{10^{10}})^2$
Campo Elétrico	$\frac{\frac{q}{\frac{r}{r}}\times10^{10}}{\left(0.08854\times\left(4\times\frac{0.08854\times\left(\frac{3.14159\times\sqrt{r}}{0.08854}\right)}{\sqrt{r}}\right)\right)}$
Densidade de Probabilidade	$\frac{\frac{((\theta-\theta)^2+e^{\left(2-(\theta-\theta)^2\right)}}{\sqrt{2}}}{\sqrt{e^{((3.14159+(2+\theta^2)))}}}$

Em média, os resultados da execução apresentaram variações dependentes da função a ser descoberta. Em alguns casos, o algoritmo demonstrou desempenho satisfatório com 10 amostras (Distância Euclidiana e Lentes Delgadas), enquanto que em outras situações obteve melhores resultados com 20 amostras (Massa Relativística). Houve ainda casos em que 40 amostras (Produto Interno) resultaram em um desempenho superior. O critério utilizado para determinar a configuração ótima foi observar os casos em que o algoritmo conseguiu encontrar a função que represente os resultados da literatura.

Um aspecto digno de destaque reside na forma como o algoritmo reage de maneira distinta em relação ao número de variáveis que compõem os problemas, assim como em relação ao valor do domínio. Uma curiosidade sobre é, se o algoritmo busca realizar uma espécie de interpolação para preencher os pontos fornecidos. No entanto, permanece a necessidade de investigar por que ele não utiliza algumas das variáveis fornecidas no problema. Embora o estudo de Sousa et al. [7] não tenha abordado essa questão em detalhes, pois os resultados não foram suficientes para conclusões definitivas, é relevante revisitar essa problemática novamente neste trabalho. Quanto a uma segunda curiosidade, faz-se imprescindível uma compreensão mais aprofundada sobre como lidar com esses dados de entrada, sem comprometer os testes.

Entre os testes realizados, foi possível constatar melhorias significativas em diversas funções, resultado direto da inclusão de uma etapa de análise dimensional das grandezas físicas. Ao conduzir os testes com essa nova abordagem, foi possível observar melhorias substanciais em diversas funções. A inclusão da análise dimensional não apenas proporcionou uma avaliação mais precisa das funções, mas também permitiu identificar aquelas que não apenas se ajustavam aos dados disponíveis, mas também respeitavam as leis físicas subjacentes. É notável que funções como a da Força Gravitacional ou do Campo Elétrico apresentaram avanços significativos, produzindo resultados que não apenas refletiam os fenômenos observados, mas também eram consistentes com os princípios físicos estabelecidos. Especificamente, funções como a Distância Euclidiana, Lentes Delgadas e Produto Interno demonstraram resultados satisfatórios, reproduzindo não apenas os fenômenos físicos de interesse, mas também coincidindo com funções encontradas na literatura. Vale ressal-

tar que o algoritmo de teste adotado não simplifica os resultados, garantindo uma representação precisa dos fenômenos.

Destacamos também o êxito na obtenção de uma função adequada para a Densidade de Probabilidade, que não alcançou resultados satisfatórios no estudo anterior de Sousa et al. [7]. Os testes realizados resultaram em funções que representam adequadamente os fenômenos físicos estudados, seguindo a abordagem da navalha de Occam, com algumas funções excluindo variáveis de entrada, mas mantendo a consistência dimensional na saída.

Quanto à Massa Relativística, embora sua complexidade seja moderada em comparação com funções como a Força Gravitacional ou o Campo Elétrico, ainda não atingiu o nível ideal desejado. No entanto, houve melhorias promissoras desde o último estudo de Sousa et al.[7].

É relevante ressaltar que, apesar dos desafios mencionados, o algoritmo conseguiu produzir funções que obedecem aos resultados da análise dimensional da função em questão. Este aspecto destaca a capacidade do algoritmo em explorar o espaço de soluções e fornecer resultados que, embora não atendam plenamente aos critérios de tempo e erro estabelecidos, mantêm coerência com as propriedades fundamentais da função analisada.

5 Conclusão

Concluímos que três elementos desempenharam um papel significativo em nossos resultados. Em primeiro lugar, o erro numérico proveniente de arredondamentos teve um impacto crucial, especialmente ao lidar com amostras maiores devido à complexidade das operações envolvidas. O contexto físico, caracterizado por valores extremamente altos ou baixos, também influenciou substancialmente o desempenho computacional. Estratégias como a normalização dos dados podem ser empregadas para mitigar esses efeitos.

Além disso, observamos a ocorrência de erros específicos em execuções com amostras consideravelmente elevadas, destacando-se especialmente na função da Força Gravitacional. Esses erros, ausentes em amostras menores, demandam investigações mais aprofundadas para compreender suas origens.

Um aspecto relevante a ser destacado é que algumas das funções resultantes foram mais simplificadas em relação às originais, apresentando menos variáveis ou constantes. Isso sugere que o algoritmo encontrou soluções equivalentes, aderindo ao princípio da navalha de Occam, que favorece soluções mais simples sempre que possível.

Para futuros trabalhos, sugerimos a implementação de melhorias na mitigação de erros de arredondamento, buscando inspiração em abordagens semelhantes já empregadas na literatura. Técnicas avançadas de calibração, como o método IRACE, podem ser exploradas para aprimorar o desempenho do algoritmo. Outra direção promissora seria a expansão do algoritmo para lidar com uma variedade mais ampla de funções, diversificando assim os desafios matemáticos enfrentados. Isso não apenas ampliaria a aplicabilidade do algoritmo, mas também permitiria uma compreensão mais abrangente de sua eficácia em diferentes domínios matemáticos. Essas sugestões podem contribuir para o aprimoramento contínuo do algoritmo e sua aplicação em uma gama mais ampla de problemas.

Agradecimentos

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

Referências

[1] E. K. Burke, M. Hyde, G. Kendall, G. Ochoa, E. Özcan e J. R. Woodward. "A classification of hyper-heuristic approaches". Em: **Handbook of metaheuristics** (2010), pp. 449–468. DOI: 10.1007/978-1-4419-1665-5_15.

[2] J. F. Gonçalves e M. G. C. Resende. "Biased Random-Key Genetic Algorithms for combinatorial optimization". Em: **Journal of Heuristics** 17.5 (2011), pp. 487–525. DOI: 10.1007/s10732-010-9143-1.

[3] J. F. Gonçalves e M. G. C. Resende. "Biased Random-Key Genetic Progamming." Em: **Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science, Springer Proceedings in Mathematics and Statistics**. Ed. por R. Martí, P. Panos e M. G. C. Resende. Springer International Publishing, 2019. Cap. 23, pp. 1–16. DOI: 10.1007/978-3-319-07153-4_25-2.

[4] A. Grings. "Regressão simbólica via programação genética: um estudo de caso com modelagem geofísica". Dissertação de mestrado. UFU, 2006.

[5] P. V. Guimaraes, A. S. da S. Junior e I. M. Coelho. "Programacão Genética com Chaves Aleatórias Viciadas em Notação Pos-Fixada para o Problema de Regressão Simbólica". Em: **LII Simpósio Brasileiro de Pesquisa Operacional** (2020). DOI: 10.59254/sbpo-2020-122767.

[6] F. P. Sousa. "Modelos de Regressão Simbólica Através de Biased Random-Key Genetic Programming em Aplicações na Física". Dissertação de mestrado. UERJ, 2023.

[7] F. P. Sousa, C. O. de Faria e I. M. Coelho. "Encontrando modelos de regressão simbólica através da metaheurística BRKGP em aplicações na Física". Em: **Anais do Encontro Nacional de Modelagem Computacional e Encontro de Ciência e Tecnologia dos Materiais**. 2023. URL: https://www.even3.com.br/anais/xxvi-encontro-nacional-de-modelagem-computacional-xiv-encontro-de-ciencia-e-tecnologia-dos-materiais-338941/705122-ENCONTRANDO-MODELOS-DE-REGRESSAO-SIMBOLICA-ATRAVES-DA-METAHEURISTICA-BRKGP-EM-APLICACOES-NA-FISICA.

[8] M. J. F. Souza. **Inteligência Computacional para Otimização**. Ouro Preto, 2024. URL: http://www.decom.ufop.br/prof/marcone/Disciplinas/InteligenciaComputacional/InteligenciaComputacional.pdf.

[9] S.-M. Udrescu, A. Tan, J. Feng, O. Neto, T. Wu e M. Tegmark. "AI Feynman 2.0: Pareto-optimal symbolic regression exploiting graph modularity". Em: **Advances in Neural Information Processing Systems**. Ed. por H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan e H. Lin. Vol. 33. Curran Associates, Inc., 2020, pp. 4860–4871. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/33a854e247155d590883b93bca53848a-Paper.pdf.

[10] Silviu-Marian Udrescu e M. Tegmark. "AI Feynman: A physics-inspired method for symbolic regression". Em: **Science Advances** 6.16 (2020), eaay2631. DOI: 10.1126/sciadv.aay2631.

