GP4LR: Uma Ferramenta para Análise de Regressão Linear com Suporte de Programação Genética

Cícero Samuel R. Mendes[1], Guilherme Álvaro R. M. Esmeraldo[2]

[1]Centro de Informática – Universidade Federal de Pernambuco – Recife – PE – Brazil

[2]Laboratório de Sistemas Embarcados e Distribuídos – Instituto Federal do Ceará – Crato – CE – Brazil

csrm@cin.ufpe.br, guilhermealvaro@ifce.edu.br

Abstract. *Genetic Programming and Linear Regression have been used together in different applications, such as software/hardware projects, weather forecasting, drug experiments, among others. This combination introduced a new class of problems, making it necessary to explore it in order to establish its main characteristics and demands. This paper presents a new tool, which combines Genetic Programming and Linear Regression, in order to contribute to applied research, through statistical modeling and data prediction. Results show that the proposed tool offers great relevance to support statistical analysis applied in several fields of study.*

Resumo. *Programação Genética e Regressão Linear têm sido utilizadas conjuntamente em diferentes aplicações, tais como em projetos de software/hardware, previsão do tempo, experimentos com fármacos, entre outras. Essa combinação introduziu uma nova classe de problemas, sendo assim necessário explorá-la para estabelecer suas principais características e demandas. Este artigo apresenta uma nova ferramenta, que combina Programação Genética e Regressão Linear, com objetivo de contribuir com a pesquisa aplicada, por modelagem estatística e predição de dados. Resultados mostram que a ferramenta proposta oferece uma grande relevância para suporte a análises estatísticas aplicadas em diversas áreas de estudo.*

1. Introdução

Regressão simbólica é uma técnica que caracteriza, através de funções matemáticas, variáveis respostas com base nas variáveis de entrada [Dabhi and Vij 2011]. Para encontrar um Modelo de Regressão Simbólica (MRS) bem ajustado a um conjunto de dados, é comum utilizar uma técnica computacional chamada Programação Genética (PG) [Linden 2012]. PG é uma especialização de Algoritmos Genéticos (AG), os quais se baseiam em evolução biológica para encontrar funções preditivas. Nessa abordagem, cada indivíduo genético é avaliado através da sua função matemática para determinar como seu resultado se ajusta ao resultado desejado [Koza 1992]. De acordo com Koza (1995), para a maioria dos problemas, o ajuste é naturalmente medido pelo erro produzido entre o resultado obtido e o esperado. Entretanto, dependendo do domínio do problema, pode-se observar que as estimativas obtidas, a partir do MRS encontrado com PG, podem apresentar erros que afetam a precisão da função preditiva [Keijzer 2003] ou modelos resultantes muito complexos [Davidson, Savic, Walters 2003].

Para tratar esses problemas, estudos, como o apresentado em [Paterlini and Minerva 2010], substituem as funções preditivas, que basicamente são MRS, por Modelos de Regressão Linear (MRL) para compor os indivíduos genéticos. Regressão Linear (RL), segundo Weisberg (2005), é o estudo de como uma variável resposta varia em função da mudança de valores assumidos pelas preditoras. Da mesma forma que os demais tipos de análises estatísticas, seu objetivo é sumarizar, com simplicidade e

utilidade, os dados estudados. Os MRL podem ser utilizados, assim como os MRS, para modelar diferentes problemas e realizar predições, mas sua maior vantagem é a possibilidade de controlar os erros das estimativas. A abordagem de PG com MRL tem sido utilizada em diferentes aplicações, como em agendamento de tarefas [Cheng, Gen, Tsujimura 1996], previsão do tempo [Babovic and Keijzer 2002], predição de dosagens em experimentos com fármacos [Archetti et al. 2006], pesquisas e tratamentos de câncer [Worzel et al. 2009], projetos de sistemas embarcados [Esmeraldo and Barros 2010], problemas de classificação [Espejo, Ventura, Herrera 2010], avaliação de qualidade de alimentos [Arnaldo, Krawiec, O'Reilly 2014], entre outros.

Além de sua aplicação prática, a inclusão de MRL à técnica de PG introduziu uma nova classe de problemas, os quais consideram, em conjunto, suas particularidades. Esses problemas podem incluir: seleção de variáveis e complexidade do modelo linear [Dignum and Poli 2008]; identificação de *outliers* (pontos de influência) e *trade-off* entre complexidade e precisão do MRL[Chan, Kwong, Fogarty 2010]; avaliação de ajuste do MRL, seleção de variáveis e complexidade (número de termos do MRL) [Paterlini and Minerva 2010]; verificação das suposições sobre o modelo e sobre as distribuições estatísticas dos erros para avaliação do ajuste, uso de Modelos Lineares Generalizados e de variáveis qualitativas [Esmeraldo and Barros 2010]; *trade-off* entre a complexidade do MRL e desempenho de PG [Esmeraldo et al. 2020]; entre outros.

Percebe-se então que o uso de PG com MRL constitui-se de um ferramental teórico-prático que pode ser aplicado nos diferentes campos da ciência. Porém, como exposto, a abordagem necessita ser explorada, de forma a se tratar suas principais características, problemas e demandas. Tendo em vista a sua importância, este trabalho apresenta uma nova ferramenta denominada GP4LR (*Genetic Programming For Linear Regression*), que busca reunir as principais técnicas do estado da arte de RL e PG. Esta ferramenta tem como objetivo contribuir com a pesquisa aplicada através de análises estatísticas, tais como modelagem e predição de dados, e oferece grande relevância para aplicações práticas, não só em Sistemas de Informação, mas em diversas áreas de estudo.

2. Materiais e Métodos

Para o desenvolvimento da ferramenta proposta, o primeiro passo consistiu em realizar um aprofundamento teórico em PG e MRL, com o intuito de estabelecer todas as funcionalidades e características da ferramenta proposta. O passo seguinte consistiu no levantamento das tecnologias que seriam utilizadas para a codificação da aplicação, em que adotou-se a linguagem de programação Python 3.x, o framework gráfico Kivy, bem como a linguagem R para o processamento e análises estatísticas dos MRL. A fase de codificação se deu em duas etapas. Na primeira etapa, foram desenvolvidos os algoritmos de PG e de processamento estatístico, e, na segunda etapa, realizou-se a codificação da interface gráfica e sua integração com os demais algoritmos. Após a codificação da ferramenta, foram realizados estudos com a finalidade de validar o desempenho e a precisão dos resultados. Para isto, foi realizado um levantamento dos repositórios de bases de dados abertas disponíveis e optou-se pelo uso do *Machine Learning Repository* da University of California, Irvine [Dua and Graff 2019], devido à variedade de bases disponíveis e a sua popularidade na literatura de *machine learning* e áreas correlatas. Após realizada uma busca por bases utilizadas em problemas de regressão, decidiu-se pelo uso da *Bike Sharing Dataset* (BSD) para compor um estudo de caso (Seção 4).

3. A Ferramenta GP4LR

GP4LR é uma ferramenta que visa dar suporte a análises estatísticas e tem como principal objetivo a seleção de MRL com o apoio da abordagem de PG. A sequência de etapas para uma análise estatística utilizando a ferramenta proposta é dada em cinco

etapas: 1) inicialmente, insere-se a base de dados a qual deseja-se realizar o estudo; 2) em seguida, configura-se os parâmetros de PG; 3) e os parâmetros de RL; 4) logo após, há o processamento; e, por fim, 5) é gerado um relatório contendo resultados de análises estatísticas sobre o MRL obtido, através de PG, como a melhor solução para o conjunto de dados sob análise. Este processo é iterativo e pode ser repetido inúmeras vezes. As subseções a seguir detalham cada uma dessas etapas.

3.1. Inserção da Base de Dados

Nesta etapa, o analista necessita utilizar algum método externo para realizar a divisão (*split*) do conjunto de dados a ser analisado em conjuntos de Treinamento e Teste – os quais são utilizados para estimar os parâmetros e validar o MRL, respectivamente –, que deverão estar em formato de arquivo .csv.

3.2. Configuração dos Parâmetros do Algoritmo de PG

Os parâmetros de PG utilizados nesta ferramenta seguem os que são comumente adotados na literatura de AG e consistem em: 1) tamanho da população inicial – a quantidade de indivíduos genéticos na primeira geração ou, em outras palavras, a quantidade de soluções candidatas geradas aleatoriamente; 2) suporte à remoção de indivíduos genéticos duplicados; 3) suporte a elitismo; 4) fator de mutação; 5) critérios para que o algoritmo encerre a sua execução, que podem ser número máximo de execuções e melhor indivíduo por *n* gerações; e 6) método de seleção de indivíduos para operação de *crossover* (atualmente, a ferramenta proposta só suporta o método de *tournament*).

As subseções a seguir detalham as principais técnicas adotadas para representação dos indivíduos genéticos e das estratégias evolucionárias utilizadas para guiar o algoritmo na busca de um MRL ótimo (que se ajuste bem ao conjunto de dados sob análise) ou quase-ótimo, no espaço de busca do problema estudado.

3.2.1 Representação dos Indivíduos Genéticos

A representação do indivíduo – ou representação cromossômica –, consiste na maneira de traduzir as informações do problema de forma que seja possível ser tratada computacionalmente [Linden 2012]. A escolha de como os indivíduos genéticos são representados deve ser adequada ao problema para que se encontre boas soluções.

Um MRL expressa a relação entre a variável resposta (dependente) e as variáveis explicativas (independentes). A estrutura destes modelos segue o padrão $Y=\beta_0+\beta_1 X_1+\beta_2 X_2+...+\beta_n X_n+\varepsilon$, onde Y é a variável resposta, $X_1+X_2+...+X_n$ são as variáveis explicativas, $\beta_0+\beta_1+\beta_2+...+\beta_n$ são os parâmetros do modelo, onde estes devem ser estimados por meio de alguma técnica de estimação (e.g. Mínimos Quadrados) aplicada sobre os dados do problema [Esmeraldo et al. 2020], e ε representa os erros (ou resíduos). No algoritmo de PG proposto, os indivíduos genéticos, que seguem esse padrão, são codificados por meio da estrutura de dados lista encadeada [Cormen 2009], com nós para os termos de um MRL.

3.2.2 Geração da População Inicial e Estratégias Evolutivas

A população inicial é gerada de forma aleatória, assim como ocorre em outros AG, onde cada indivíduo genético terá sua estrutura criada a partir da seleção aleatória das variáveis explicativas do problema e/ou de suas interações [Esmeraldo et al. 2020]. Em seguida, cada indivíduo da população é submetido à função de avaliação para que seja mensurada a sua adequação e o algoritmo possa seguir com o processo evolucionário por meio da aplicação dos operadores evolucionários de seleção, *crossover* e mutação.

A seleção simula o processo de seleção natural, em que os pais mais aptos geram mais filhos, ao mesmo tempo que os menos aptos também podem gerar descendentes [Linden 2012]. A abordagem utilizada para o operador de seleção na ferramenta aqui

proposta consiste no *tournament*, onde k indivíduos são selecionados aleatoriamente da população e comparados uns aos outros, onde o mais bem avaliado destes, de acordo com a função de avaliação adotada, é escolhido para fazer parte do conjunto de pais. Caso haja empate na avaliação entre indivíduos, o que tiver a menor complexidade será escolhido e o processo se repete até que o conjunto de indivíduos-pais esteja completo [Esmeraldo et al. 2020]. Após a seleção dos pais, a geração dos novos indivíduos da próxima geração segue com a aplicação dos operadores genéticos *crossover* e mutação.

Crossover atua selecionando pontos de corte em um par de indivíduos-pais para que eles possam combinar o material genético e gerar os indivíduos-filhos. O método implementado nesta ferramenta consiste no *crossover* de dois pontos (são selecionados aleatoriamente dois pontos para corte em cada par de indivíduos-pais e as partes resultantes são permutadas para compor os filhos).

Após o *crossover*, uma determinada quantidade dos novos indivíduos serão submetidos ao operador de mutação. O operador de mutação permite que o processo evolucionário explore novas soluções potenciais para o problema no espaço de busca [Simon 2013]. Cada indivíduo possui uma probabilidade de sofrer uma mutação, que se dá através da seleção de um novo ponto de corte no indivíduo, em que uma das partes é descartada e em seu lugar é inserida uma nova estrutura gerada aleatoriamente.

3.2.3 Avaliação dos Indivíduos Genéticos

A função de avaliação é a maneira utilizada para determinar a qualidade dos indivíduos genéticos em uma população. A função calcula um valor numérico que reflete quão bons os parâmetros representados no indivíduo resolvem o problema [Linden 2012]. Na GP4LR, os indivíduos genéticos são avaliados por meio da estimação dos parâmetros, através do Método dos Mínimos Quadrados, bem como também pela função de ajuste que pode ser selecionada arbitrariamente ou de acordo com as características e particularidades do conjunto de dados para o qual se deseja encontrar o MRL. Entre as funções de ajuste disponíveis na ferramenta proposta, estão: *Root Mean Squared Error* (RMSE), *Root Means Squared Log Errors* (RMSLE), *Mean Absolute Error* (MAE), *Mean Absolute Percentage Error* (MAPE), *Mean Absolute Scaled Error* (MASE) e *Coefficient of Determination* (R^2) [Schelling and Robertson 2020]. Essas métricas possuem particularidades – tais como sensibilidade a pontos de influência e variável resposta com valores próximos de zero ou em diferentes escalas –, que fazem com que a sua escolha seja condicionada às características do conjunto de dados sob estudo.

3.2.4 Elitismo

O Elitismo consiste em uma técnica utilizada em AG para manter o indivíduo com a melhor avaliação entre gerações. A utilização do elitismo implica na seleção do melhor indivíduo da geração corrente e mantê-lo, sem nenhuma alteração, na geração seguinte.

3.3. Configuração dos Parâmetros de Regressão Linear

A terceira etapa do fluxo de execução consiste em informar os parâmetros utilizados para a avaliação da adequação dos MRL (indivíduos genéticos) ao conjunto de dados e, dessa maneira, guiar o algoritmo de PG na exploração do espaço de busca. Nesta etapa, os parâmetros que podem ser configurados são: 1) Função de transformação da variável resposta, em que pode-se optar (ou não) por utilizar a função de logaritmo; 2) Inserção de conjunto de dados para predições (opcional); 3) Critério de ajuste do MRL, no qual pode-se optar por *Root Mean Squared Error* (RMSE), *Root Means Square Log Errors* (RMSE), *Mean Absolute Error* (MAE), *Mean Absolute Percentage Error* (MAPE), *Mean Absolute Scaled Error* (MASE) ou *Coefficient of Determination* (COD - R^2). O critério de ajuste é utilizado como função de avaliação pelo algoritmo de PG, ou seja, todos os indivíduos genéticos de uma população são avaliados segundo o critério escolhido a fim de guiar o algoritmo a encontrar a solução ótima ou quase-ótima; 4) Testes de hipóteses para avaliar o ajuste do MRL ao conjunto de treinamento e

suposições sobre os resíduos, visando estabelecer a sua precisão. Para a verificação do ajuste do MRL aos dados do conjunto de treinamento, pode-se utilizar os testes Kolmogorov-Smirnov, Mann-Whitney-Wilcoxon e χ^2 (*Chi-Squared*). Já para a verificação das suposições sobre a natureza dos resíduos, pode-se utilizar os testes Kolmogorov-Smirnov, Anderson-Darling e Shapiro-Wilk para teste de normalidade; Breusch-Pagan para teste de homoscedasticidade (resíduos com variância desconhecida e constante); e Durbin-Watson para testes de independência dos resíduos [Esmeraldo et al. 2020]; e 5) Gráficos que visam apoiar a interpretação e análise da adequação do modelo e suposições sobre a natureza dos resíduos, pois apresentam uma abordagem menos formal em relação aos testes de hipóteses. Entre os gráficos disponíveis, estão: QQ-Plot Distribuição dos Resíduos vs Distribuição Teórica, Histograma dos Resíduos, Dispersão de Resíduos vs Valores Ajustados, Dispersão de Resíduos x Ordem dos Resíduos, Função de Distribuição Acumulada dos Resíduos, Bolhas com as Distâncias de Cook Proporcionais dos Resíduos e Box-Plot com as Distâncias de Cook em Escala Logarítmica dos Resíduos.

3.4. Processamento

A etapa de processamento envolve a execução do algoritmo de PG, de acordo com os parâmetros informados pelo analista (ver Subseção 3.2). Ao final da execução do algoritmo de PG, o indivíduo genético final, que é o MRL que melhor se ajustou ao conjunto de dados sob estudo, passará por uma série de avaliações, de acordo com os parâmetros definidos na Subseção 3.3.

3.5. Relatório

A última etapa consiste na apresentação de um relatório contendo estatísticas para a melhor solução encontrada pelo algoritmo de PG. O relatório inclui: o MRL encontrado e sua complexidade; o número da geração em que o algoritmo convergiu para a solução candidata; testes de hipóteses e gráficos para o ajuste do MRL e para análise dos resíduos para os conjuntos de treinamento e teste. Ressalta-se que o relatório incluirá apenas os resultados das análises selecionadas na etapa de "Configuração dos Parâmetros de Regressão Linear".

4. Estudo de Caso

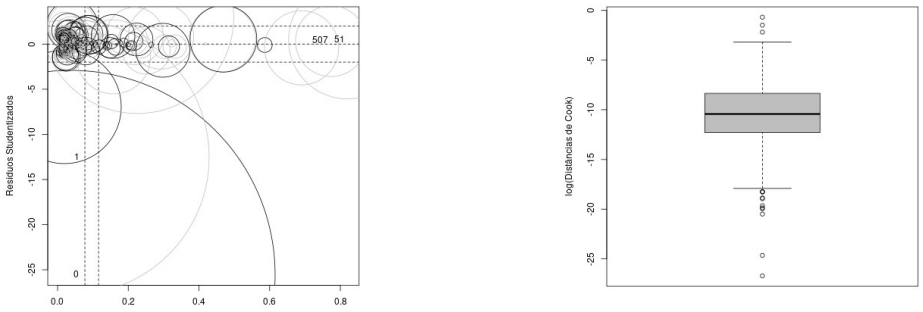
Para avaliar a ferramenta proposta, como estudo de caso, realizou-se um experimento de modelagem e predição com a utilização da base de dados aberta *Bike Sharing Dataset (BSD)* [Dua and Graff 2019], que consiste em dados coletados de um sistema de compartilhamento de bicicletas ao longo de dois anos na cidade de Washington, D. C., EUA. Esta base de dados contém 17.379 registros e 17 variáveis. O objetivo deste estudo de caso consistiu em estabelecer a relação do número total de aluguéis por usuários casuais e registrados (variável *cnt*) e as condições climáticas e de periodicidade (demais variáveis presentes no conjunto de dados), por meio de um MRL obtido através da utilização da ferramenta proposta. Assim, a variável *cnt* foi adotada como variável resposta e as demais compõem o conjunto de variáveis preditivas. A base de dados BSD foi dividida em conjunto de treinamento, que corresponde a 80% do tamanho total e é utilizado para estabelecer a relação entre as variáveis através do MRL obtido, e conjunto de teste, que é composto pelos 20% restantes dos registros e é utilizado para comparação com as estimativas fornecidas pelo MRL encontrado para validação.

Nas configurações do algoritmo de PG, a população inicial foi composta por 50 indivíduos. Para preservar o melhor indivíduo, a fim de que boas soluções possam se perpetuar ao longo das gerações, optou-se pela técnica do elitismo. Além disso, visando preservar a variabilidade genética entre indivíduos, optou-se por remover soluções candidatas duplicadas. O algoritmo utiliza dois critérios de parada, onde configurou-se para encerrar a execução após 100 gerações ou após um determinado indivíduo

permanecer como o melhor avaliado após 30 gerações consecutivas, onde será apresentado como solução candidata. Já para os parâmetros utilizados para o ajuste e avaliação da adequação dos MRL, devido à natureza do conjunto de dados em estudo, em que não há grande amplitude de valores assumidos pela variável resposta, optou-se pela não utilização de função de transformação. Já para a função de ajuste, utilizada para avaliação da adequação dos MRL, escolheu-se o *Coefficient of Determination* (R^2), pois, ao realizar a estimativa dos parâmetros de um MRL utilizando o software R, automaticamente já se calcula o R^2 para o modelo obtido, evitando-se assim *overhead* ao cálculo de uma das métricas. Além disso, utilizou-se todo o conjunto de testes de hipóteses disponíveis para a avaliação das suposições sobre o MRL e sobre a natureza dos resíduos. Por fim, também utilizou-se todo o conjunto de gráficos disponíveis para a análise da adequação do MRL, para caso ocorram falhas nos testes de hipóteses.

Foram realizadas 40 execuções da GP4LR, sendo obtidos, portanto, 40 MRL diferentes. Em média, os MRL apresentados como solução candidata obtiveram complexidade de 331.1. Vale destacar que, dentre as soluções encontradas, os MRL com menor e maior complexidade apresentaram 22 e 1122 termos, respectivamente. Já em relação ao algoritmo de PG, em média, foram alcançadas 30.9 gerações, com mínimo de 12 e máximo de 65. De acordo com o estudo apresentado em [Esmeraldo et al. 2020], há uma relação direta entre a complexidade dos MRL, o número de gerações e o desempenho de um algoritmo genético. Nesse sentido, visando estabelecer um *trade-off* entre a precisão dos MRL e o desempenho do algoritmo de PG, para este estudo de caso, dentre os 40 MRL, selecionou-se aquele com menor complexidade (22 termos).

Nos testes de hipóteses para o MRL escolhido, considerando que estabeleceu-se um nível de significância de 5%, obteve-se os seguintes resultados: 1) Ajuste ao conjunto de treinamento: Kolmogorov-Smirnov, Mann-Whitney-Wilcoxon e χ^2 (*Chi-Squared*) apresentaram resultados 1.0000e0, 9.972e-1 e 0.0000e0, respectivamente, demonstrando que o MRL se ajustou bem aos dados do conjunto de treinamento; e 2) Suposições sobre a natureza dos resíduos: os testes Kolmogorov-Smirnov, Anderson-Darling e Shapiro-Wilk (Normalidade), Breusch-Pagan (Homoscedasticidade) e Durbin-Watson (Independência) apresentaram resultados iguais à 0.0000e0, o que demonstra que todas os testes sobre a natureza dos resíduos falharam. Nesta situação, deve-se considerar que os testes de hipóteses são sensíveis a *outliers* (pontos de influência), que podem distorcer as estimativas dos parâmetros do MRL, prejudicando assim o ajuste aos dados [Weisberg 2005]. Para investigar a presença de *outliers* foram utilizados os gráficos de Bolhas e Boxplot, apresentados nas Figuras 1(a) e 1(b), respectivamente.



(a)Distâncias de Cook proporcionais dos resíduos. **(b)Distâncias de Cook em Escala Logarítmica.**

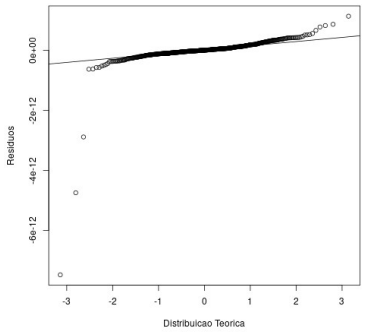
Figura 1. Gráficos para verificação da presença de *outliers*.

No gráfico apresentado na Figura 1(a), os círculos são plotados baseando-se proporcionalmente na Distância de Cook [Cook 1977], onde os círculos maiores

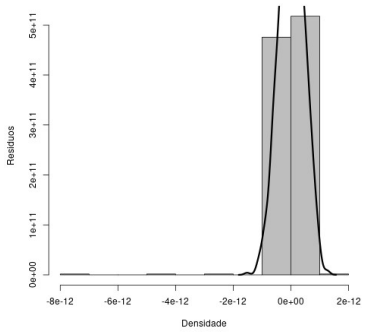
representam os *outliers*. Já no gráfico da Figura 1(b) é apresentado um BoxPlot com as distâncias de Cook em escala logarítmica, sendo possível perceber a existência de pontos fora da faixa de valores máximo e mínimo, indicando a presença de *outliers*. Devido à existência de *outliers*, pode-se concluir que os ajustes dos MRL tratados pelo algoritmo de PG podem ter sido prejudicados e, dessa forma, influenciado negativamente os resultados dos testes de hipóteses. Apesar de haver, na literatura, técnicas para tratar *outliers*, o foco deste estudo de caso consistiu na modelagem e predição de dados com MRL obtido de forma automática através de PG. Portanto, levando isso em consideração, optou-se pela preservação da BSD sem tratamento/remoção dos *outliers*.

Como uma alternativa aos testes de hipóteses, a análise gráfica pode ajudar a checar as suposições e avaliação da adequação de um MRL [Chatterjee and Hadi 2012]. Assim, a análise gráfica dos resíduos do MRL para a *BSD* é apresentada na Figura 2. Na Figura 2(a), é apresentado o gráfico QQ-plot dos resíduos confrontados com a distribuição teórica normal. Pode-se observar que a maioria dos pontos está concentrada sobre a reta. Além disso, pode-se perceber que o comportamento de cauda apresenta alguns pontos distantes da reta, sugerindo a presença de *outliers*.

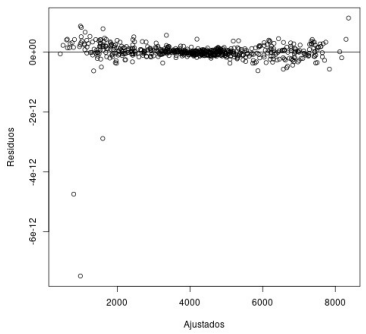
Na Figura 2(b) é apresentado um histograma dos resíduos, onde evidencia-se uma curva em comportamento de sino. De acordo com os dois gráficos, há evidências de que a suposição de normalidade dos resíduos pode ser satisfeita, desde que sejam aplicadas técnicas de tratamento de *outliers* presentes na literatura. As Figuras 2(c) e 2(d) apresentam os gráficos de dispersão de resíduos contra valores ajustados e de ordem dos resíduos, respectivamente. Nestes gráficos também fica explícita a presença de *outliers*. No entanto, percebe-se a presença de pontos distribuídos em torno da reta horizontal e próximos a zero. Desta forma, também há evidências que corroboram para que as suposições de homoscedasticidade e independência dos resíduos possam ser satisfeitas, desde que, como discutido anteriormente, haja tratamento dos *outliers*.



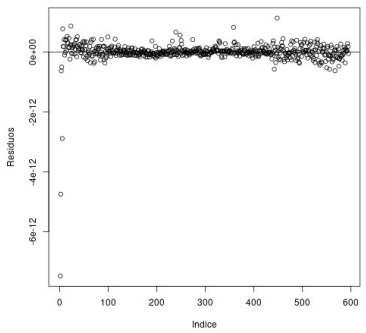
(a)QQ-Plot: Suposição de Normalidade.



(b)Histograma: Suposição de Normalidade.



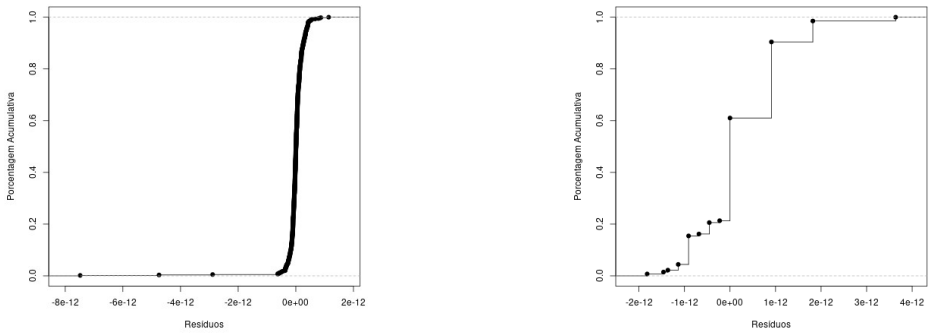
(c)Resíduos x Valores Ajustados: Suposição de Homoscedasticidade.



(d)Ordem dos Resíduos: Suposição de Independência.

Figura 2. Gráficos para análise residual.

Uma vez que há evidências suficientes de que as suposições sobre o MRL foram satisfeitas e, portanto, considerando-se que ele foi aprovado na etapa de validação, o próximo passo consiste na avaliação dos resíduos para o conjunto de Teste. As Figuras 3(a) e 3(b) apresentam os gráficos de Função Distribuição Acumulada – a qual descreve a porcentagem de uma determinada variável assumir um valor ou valores em diferentes faixas –, para os conjuntos de Treinamento e Teste, respectivamente.



(a)Conjunto de treinamento. **(b)Conjunto de testes.**

Figura 3. Gráficos de distribuição acumulada para os resíduos.

O gráfico da Figura 3(a) mostra que quase 100% dos resíduos acumulados do conjunto de Treinamento tendem a zero, o que mostra que o MRL encontrado apresentou estimativas muito precisas para a variável resposta de treinamento. Já no gráfico da Figura 3(b), é possível perceber que, em torno de 20% dos resíduos acumulados para o conjunto de Testes estão entre a faixa de valores -2e-12 e 0, e os outros 80% situam-se no intervalo de 0 à 4e-12, mostrando que o MRL encontrado também se ajustou bem ao segundo conjunto de dados.

5. Conclusões

Este trabalho apresentou GP4LR, uma ferramenta para suporte à modelagem e análises estatísticas, na qual objetiva selecionar MRL com o apoio de técnicas de PG. GP4LR dispõe de interface gráfica que possibilita configurar os parâmetros do algoritmo de PG e RL, onde pode-se optar pela execução de diversos testes estatísticos e a geração de um conjunto de gráficos que são apresentados em um relatório ao final de cada execução. De acordo com o estudo de caso, foi possível constatar que, com o uso da ferramenta proposta, pode-se encontrar MRL precisos para a modelagem e predição de dados.

Por fim, ressalta-se que esta é uma ferramenta experimental e que são necessárias melhorias e adição de novos recursos. Portanto, entre os trabalhos futuros estão elencados: estudo das populações do algoritmo de PG para identificar a variedade dos indivíduos e a convergência para a solução ótima, provimento de melhorias da diversidade das soluções no espaço de busca e favorecer a pressão evolutiva, guiando, desta forma, o algoritmo para encontrar soluções ótimas; inclusão de processamento paralelo, visando aumentar o desempenho da ferramenta, de funções de ajuste tolerantes à *outliers* e de Modelos Lineares Generalizados, visando aumentar o espectro de tipos de análises de regressão suportadas. Com isto, pretende-se prover um ambiente robusto e flexível que permita a sua aplicação prática em diversas áreas de estudo.

Referências

Archetti, F., Lanzeni, S., Messina, E., Vanneschi, L. (2006) “Genetic programming for human oral bioavailability of drugs”. In: Proceedings of the 8th annual conference on Genetic and evolutionary computation. p. 255-262.

Arnaldo, I., Krawiec, K., O'Reilly, U. M. (2014) “Multiple regression genetic

programming". In: Proceedings of the 2014 conference on Genetic and evolutionary computation. p. 879-886.

Babovic, V., Keijzer, M. (2002) "Rainfall runoff modelling based on genetic programming". In: Hydrology Research, 33(5). p. 331-346.

Chan, K. Y., Kwong, C. K., Fogarty, T. C. (2010) "Modeling manufacturing processes using a genetic programming-based fuzzy regression with detection of outliers". In: Information Sciences, 180(4), p. 506-518.

Chatterjee, S., Hadi, A. S. (2012) "Regression analysis by example". John Wiley & Sons.

Cheng, R., Gen, M., Tsujimura, Y. (1996) "A tutorial survey of job-shop scheduling problems using genetic algorithms". I. Representation. Computers & industrial engineering, 30(4). p. 983-997.

Cook, R. D. (1977) "Detection of Influential Observation in Linear Regression". In: Technometrics, 19(1), p. 15–18.

Cormen, T. H. (2009) "Introduction to algorithms". [S.l.]: MIT press.

Dabhi, V. K., Vij, S. K. (2011) "Empirical modeling using symbolic regression via postfix genetic programming". In: 2011 International Conference on Image Information Processing (ICIIP), p. 1-6.

Davidson, J., Savic, D. A., Walters, G. A. (2003) "Symbolic and numerical regression: experiments and applications". In: Information Sciences, Elsevier, v. 150, n. 1, p. 95–117.

Dignum, S., Poli, R. (2008) "Operator equalisation and bloat free GP". In: Genetic Programming, Springer Berlin Heidelberg. p. 110-121.

Dua, D., Graff, C. (2019) UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science. http://archive.ics.uci.edu/ml.

Esmeraldo, G., Barros, E. (2010) "A genetic programming based approach for efficiently exploring architectural communication design space of mpsocs". In: IEEE. Programmable Logic Conference (SPL), 2010 VI Southern. [S.l.]. p. 29–34.

Esmeraldo, G. A. R. M., Oliveira, C. C. F., Sales, M. M., Souza, F. A. (2020) "Uma Abordagem para Análise de Regressão com Suporte de Programação Genética". In: Guttenberg S. S. Ferreira; Régia T. S. Araújo. (Org.). ENSAIOS DE MATEMÁTICA: pesquisas em ensino e ciências aplicadas, Vo. 1. CRV. p. 51-72.

Espejo, P. G., Ventura, S., Herrera, F. (2010) "A survey on the application of genetic programming to classification". In: IEEE Transactions on Systems, Man, and Cybernetics, Part C, 40(2), p. 121-144.

Keijzer, M. "Improving Symbolic Regression with Interval Arithmetic and Linear Scaling. Genetic Programming". Heidelberg: Springer, p. 70-78, 2003.

Koza, J. R. (1992) "Genetic Programming: On the Programming of Computers by Means of Natural Selection", MIT Press.

Koza, J. R. (1995) "Survey of genetic algorithms and genetic programming". In: Wescon conference record. WESTERN PERIODICALS COMPANY, p. 589-594.

Linden, R. (2012) "Algoritmos Genéticos. 3.ed.". Ciência Moderna.

Schelling, X., Robertson, S. (2020) "A development framework for decision support systems in high-performance sport". In: International Journal of Computer Science in Sport, v. 19, n. 1, p. 1-23.

Simon, D. (2013) “Evolutionary optimization algorithms”. John Wiley & Sons.

Paterlini, S., Minerva, T. (2010) “Regression Model Selection Using Genetic Algorithms”. In: Proceedings of the 11th WSEAS International Conference on Recent Advances in Neural Networks, Fuzzy Systems & Evolutionary Computing.

Weisberg, S. (2005) “Applied linear regression”. Vol. 528. John Wiley & Sons.

Worzel, W. P.; Yu, J., Almal, A. A., Chinnaiyan, A. M. (2009) “Applications of genetic programming in cancer research”. In: The international journal of biochemistry & cell biology, 41(2), p. 405-413.