Um estudo bibliográfico sobre regressão linear com suporte de programação genética

A bibliographic study on linear regression with genetic programming support

DOI:10.34117/bjdv6n11-181

Recebimento dos originais: 07/10/2020
Aceitação para publicação: 10/11/2020

Hyago Sayomar Dias Ferreira
Bacharelando em Sistemas de Informação
Instituto Federal de Educação, Ciência e Tecnologia do Ceará *campus* Crato
Endereço: Rod. CE 292 Km 15 S/N. Gisélia Pinheiro. Crato – CE. Cep: 63.115-500
Email: hyagodiasf@gmail.com

Elder Cordeiro
Bacharelando em Sistemas de Informação
Instituto Federal de Educação, Ciência e Tecnologia do Ceará *campus* Crato
Endereço: Rod. CE 292 Km 15 S/N. Gisélia Pinheiro. Crato – CE. Cep: 63.115-500
Email: jnelderc@gmail.com

Cícero Samuel Rodrigues Mendes
Bacharelando em Sistemas de Informação
Instituto Federal de Educação, Ciência e Tecnologia do Ceará *campus* Crato
Endereço: Rod. CE 292 Km 15 S/N. Gisélia Pinheiro. Crato – CE. Cep: 63.115-500
Email: mr.samuelmendes@gmail.com

Guilherme Álvaro Rodrigues Maia Esmeraldo
Doutor em Ciência da Computação
Instituto Federal de Educação, Ciência e Tecnologia do Ceará *campus* Crato
Endereço: Rod. CE 292 Km 15 S/N. Gisélia Pinheiro. Crato – CE. Cep: 63.115-500

RESUMO
A combinação das técnicas de Programação Genética e de Regressão Linear tem sido utilizada em diferentes aplicações, como em agendamento de tarefas, projetos de software e de hardware, previsão das condições climáticas, experimentos com fármacos, tratamento de câncer, avaliação de qualidade de alimentos, entre outros. Contudo, percebe-se que, essa combinação introduziu uma nova classe de problemas, os quais consideram as particularidades das duas abordagens, sendo assim necessário explorá-la, para relacionar e classificar suas principais características e demandas. Este trabalho visa, através de estudo aprofundado da literatura em Programação Genética e Regressão Linear, examinar suas potencialidades e estabelecer o estado de arte dessa comunhão, além de disponibilizar uma ferramenta para sua aplicação prática.

Palavras-chave: Análise Estatística, Regressão Linear, Programação Genética.

ABSTRACT
The combination of Genetic Programming and Linear Regression techniques has been used in different applications, such as scheduling tasks, software and hardware projects, weather forecasting, drug experiments, cancer treatment, food quality assessment, among others. However, it is remarkable that this combination introduced a new class of problems, which consider the particularities of both approaches, so it is necessary to explore it to relate and classify its main characteristics and requirements. This work aims, through an in-depth study of the literature on Genetic Programming and Linear Regression, to examine their potentialities and establish the state of art of this communion, and provide a tool for its practical application.

Keywords: Statistical Analysis. Linear Regression. Genetic Programming.

1 INTRODUÇÃO

A Regressão Simbólica consiste na manipulação de termos matemáticos para descoberta de funções que descrevem um conjunto de dados multidimensionais, frequentemente desbalanceados, com pequenas ou grandes amostras e o uso para predição. Para encontrar um Modelo de Regressão Simbólica (MRS) bem ajustado ao conjunto de dados, é comum utilizar uma técnica computacional chamada Programação Genética (PG), que é uma especialização de Algoritmos Genéticos (LINDEN, 2008) 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). No entanto, 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, que basicamente são constituídas de modelos matemáticos determinísticos. Para solucionar este problema, tem-se buscado utilizar Modelos de Regressão Linear para compor os indivíduos genéticos. A Regressão Linear (RL), segundo Weisberg (2005), é o estudo da dependência ou, em outras palavras, de como uma variável resposta varia em função da mudança de valores assumidos pelos preditores. De mesma forma que os demais tipos de análises estatísticas, seu objetivo é sumarizar, com simplicidade e utilidade, os dados estudados, e sua maior vantagem é a possibilidade de controlar os erros das estimativas.

Percebe-se então que uso de PG com RL, para modelagem e predição, constitui-se de um acessório prático e teórico que pode ser aplicado nos diferentes campos da ciência. Entretanto, a abordagem necessita de mais exploração, de forma a estudar, relacionar e classificar suas principais características, problemas e demandas. Nesse sentido, este trabalho tem como objetivos estabelecer o estado da arte de análise de regressão com suporte de programação genética, bem como oferecer uma ferramenta de software para sua aplicação prática nos diversos campos científicos.

O artigo está dividido da seguinte maneira: Na Seção 2, apresenta-se a revisão da literatura. A Seção 3 apresenta sucintamente a metodologia utilizada neste trabalho. A Seção 4 apresenta resultados preliminares e a Seção 5 apresenta as conclusões e trabalhos futuros.

2 ESTADO DA ARTE

A abordagem de PG com RL tem sido utilizada em diferentes aplicações, como em agendamento de tarefas, previsão do tempo, predição de dosagens em experimentos com fármacos, pesquisas e tratamentos de câncer, otimização de desempenho de sistemas embarcados, problemas de classificação, avaliação de qualidade de alimentos, entre outros. Além de sua aplicação prática, a inclusão de MRL à técnica de PG introduziu uma nova classe de problemas, os quais consideram, conjuntamente, as particularidades dessas duas abordagens. Esses problemas podem incluir: Avaliação de ajuste do modelo de regressão; Verificação as suposições sobre o modelo (ESMERALDO et al., 2012) e sobre as distribuições estatísticas dos erros, para avaliação do ajuste; Utilização de diferentes modelos de regressão como indivíduos genéticos; Aplicação de transformações lineares; Estabelecimento de compromisso entre a complexidade e a precisão do modelo (CHAN; KWONG; FOGARTY, 2010); Utilização de variáveis qualitativas; Seleção de variáveis e complexidade do modelo linear (PATERLINI; MINERVA, 2010).

Nesta perspectiva, este trabalho vem buscando investigar, à luz de categorias e facetas, para estabelecer uma base de conhecimento acerca da totalidade de estudos e pesquisas em Programação Genética e Regressão Linear.

3 MATERIAIS E MÉTODOS

A metodologia utilizada incluiu inicialmente a coleta, classificação, leitura, interpretação, análise e organização de bibliografias relacionadas aos temas de programação genética e regressão linear. Em paralelo à revisão bibliográfica, realizava-se a definição dos requisitos e características da ferramenta de software para aplicação prática de RL com suporte de PG. Inicialmente, focou-se no desenvolvimento de um protótipo, utilizando a linguagem de programação Python - por ser uma linguagem produtiva, simplificada e ter suporte de bibliotecas estatísticas e de geração de gráficos. Entretanto, buscando otimizar o desempenho da ferramenta, além de simplificar a sua distribuição e utilização por público-alvo específico (analistas estatísticos), o software será portado para compor uma biblioteca da Linguagem R[1].

[1]R-Project. The R Project for Statistical Computing. Disponível em: <https://www.r-project.org/>. Acesso em: 17 set. 2017.

4 RESULTADOS PRELIMINARES

Atualmente, a pesquisa encontra-se em fase de síntese do estado da arte. Estão sendo analisados mais alguns trabalhos relacionados, publicados após o início desta pesquisa, que seguirá com a escrita de um artigo do tipo *survey*. Já a ferramenta proposta, atualmente, inclui um algoritmo básico de PG (com 595 linhas de código), com os seguintes recursos: modelos de regressão linear codificados como indivíduos genéticos, seleção de pais genéticos via *tournament*, operadores de crossover (com um e dois cortes), mutação e elitismo, ajuste dos indivíduos genéticos via a técnica de Mínimos Quadrados, aplicação de funções de transformação à variável resposta, seleção de distribuição estatística para os erros e realização de testes de hipóteses (e.g. Qui-quadrado e Kolmogorov-Smirnov) para validar soluções genéticas. Com esses recursos, é possível realizar análises básicas como modelagem estatística para determinados conjuntos de dados e, através dos modelos de RL obtidos com o algoritmo proposto de PG, realizar predições.

5 CONCLUSÕES

Na literatura, tem-se combinado a técnica de Programação Genética à de Regressão Linear, visando automatizar todo o esforço para obtenção dos modelos de regressão, e essa combinação tem sido aplicada em diferentes campos do saber. Este trabalho tem como objetivos contribuir, através de pesquisas bibliográficas e exploratórias, estabelecendo uma base de conhecimento acerca da totalidade de estudos e pesquisas em PG com suporte de RL e oferecer uma nova ferramenta para sua aplicação prática. Atualmente, está em desenvolvimento um protótipo, o qual já pode ser utilizado para modelagem estatística e predição, em determinadas bases de dados. Trabalhos futuros incluem a síntese do estado da arte, adição de mais recursos ao protótipo e porte da ferramenta para uma biblioteca na linguagem R.

REFERÊNCIAS

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

ESMERALDO, G.; FEITOSA, R.; ESMERALDO, D.; BARROS, E. Genetically Programmed Regression Linear Models for Non-Deterministic Estimates. In: Genetic Programming - New Approaches and Successful Applications, Dr. Sebastian Ventura Soto (Ed.), InTech, 2012.

KOZA J. R. Genetic Programming: On the Programming of Computers by Means of Natural Selection, MIT Press, 1992.

LINDEN, R. Algoritmos Genéticos, uma importante ferramenta da inteligência computacional. Rio de Janeiro: Brasport, 2008WEISBERG, S. Applied linear regression. Vol. 528. John Wiley & Sons, 2005.

PATERLINI, S.; MINERVA, T. Regression Model Selection Using Genetic Algorithms, Proceedings of the 11th WSEAS International Conference on RECENT Advances in Neural Networks, Fuzzy Systems & Evolutionary Computing, 2010.

WEISBERG, S. Applied linear regression. Vol. 528. John Wiley & Sons, 2005.