

Universidade de Pernambuco

Programa de Pós-Graduação em Engenharia da Computação (PPGEC)

Proposta de Dissertação de Mestrado

Área: Computação Inteligente

Título: Métodos de Regressão utilizando Big Data

Orientador – Roberta Andrade de A. Fagundes (roberta.fagundes@upe.br)

Coorientador – Byron Leite Dantas Bezerra (byronleite@ecomp.poli.br)

O cenário cresceu cada vez mais. Por isso, tornou-se necessário capturar dados de forma precisa, classificar, descobrir o que é significativo, enxergar padrões e extrair informações relevantes. Matemática e estatística tornam-se cada vez mais necessárias para manipular os dados, metodologia científica para ter embasamento na elaboração de hipóteses e em modelos preditivos de cenários cada vez mais complexos e até lógica de programação passa a ser valiosa para trabalhar de forma avançada com o grande volume de dados na forma digital.

Pela primeira vez na história[1], temos a capacidade de coletar e analisar quantidades massivas de informações de forma mais fácil e barata. O volume de dados disponível mais do que dobra a cada dois anos e os algoritmos aperfeiçoam-se rapidamente, ao passo que, em razão quase inversamente proporcional, os custos de armazenamento decrescem. Técnicas de análise de dados, antes acessíveis apenas às agências de espionagem, laboratórios de pesquisa e grandes conglomerados comerciais são, paulatinamente, democratizadas. Big Data [1] será uma parte imprescindível da solução para os problemas globais mais prementes como o enfrentamento das mudanças climáticas, a erradicação de doenças, além da promoção da boa governança e do desenvolvimento econômico. Analisar dados não é nem um pouco novo, mas está sendo definido um novo tipo de profissional nesse meio: o Data Scientist. Um profissional que consegue pensar em algoritmos, não é ludibriado com excesso de dados e trabalha com métodos de análise, fazendo suas descobertas dentro do Big Data [2]. Na estatística, essas técnicas podem ser aplicadas para descobrir estruturas ou associações em conjunto de dados, realizar previsões, etc. Dentre elas, destacam-se modelos de regressão [3], que são modelo matemáticos e tem como um dos objetivos prever o valor da variável dependente (Y) a partir das informações provenientes de um conjunto de variáveis independentes (X). A presente pesquisa tem como objetivo geral avançar no estudo de Big Data representados por métodos de regressão, abrangendo aspectos teóricos e práticos, a fim de fornecer subsídios aos usuários na decisão sobre qual sistema utilizar. Para isso, os seguintes objetivos específicos são necessários: Estudar e desenvolver métodos de regressão e algoritmos de otimização aplicados a Big Data; Avaliar os métodos de regressão e algoritmos de otimização utilizados nesses projetos através da análise de resíduos e implementar funções que manipulam Big Data representados por dados simbólicos através da utilização da linguagem R [3]. Durante o estudo um ambiente experimental para avaliação do modelo proposto será implementado, como também, serão utilizados dados (artificiais/reais) e o desempenho do modelo será medido pelo erro de predição estimado pelo método Monte Carlo no conjunto de teste utilizando a Linguagem R.

Referências Bibliográficas

[1] Mayer-schönberger, Viktor; Mayer-schönberger, Viktor; Cukier, Kenneth; Cukier, Kenneth, Big Data - Como Extrair Volume, Variedade, Velocidade e Valor da Avalanche de Informação Cotidiana, campus, 2013.

[2] Taurion, Cezar, Big Data, Brasport Livros e Multimídia Ltda, 2013.

[3] MONTGOMERY D.C.; PECK, E.A. e VINING, G. G. Introduction to Linear Regression Analysis, Wisley-Interscience, 2006.