Engenharia de Controle e Automação

Desenvolvimento de algoritmos de aprendizagem de máquina baseado em Curvas Principais

Victor Daniel Reis - 12 módulo de Engenharia de Controle e Automação, bolsista PIBIC/FAPEMIG.

Fernando Elias de Melo Borges - Coorientador, Doutorando em Engenharia Agrícola, UFLA.

Danton Diego Ferreira - Professor do Departamento de Automática, UFLA. - Orientador(a)

Resumo

Os dados são a base da inteligência artificial e da economia moderna. Entretanto, obtê-los com alta qualidade, baixo ruído e boa confiabilidade para representar a aplicação do mundo real torna-se caro e desafiador. Para compensar esta lacuna, gerar dados sintéticos pode ser uma forma ilimitada, barata e simples. Assim, Curvas Principais (PC) são um modelo de representação de dados em que um conjunto de dados multidimensional é representado por uma curva unidimensional. São uma generalização não linear da Análise de Componentes Principais (ACP) e têm a forma sugerida pelos dados que são representados pela curva. O objetivo deste trabalho é produzir um algoritmo capaz de gerar dados sintéticos utilizando Curvas Principais. A metodologia empregada neste trabalho pode ser dividida em quatro partes: a primeira parte consiste na escolha da base de dados para a geração, neste trabalho foram utilizadas quatro bases de dados sintéticas com formatos variados: anelar, espiral, lunar e espiral 3D; sendo as três primeiras bidimensionais e a quarta tridimensional. A segunda parte consiste na obtenção das curvas através do algoritmo k-segmentos onde são obtidos parâmetros importantes (como pontos projetados no segmento e distâncias quadráticas). A terceira parte inclui a aplicação do algoritmo produzido para gerar dados sintéticos, uma vez que o conjunto de dados real e dois hiperparâmetros são especificados pelo usuário: fator de dispersão e aumento percentual dos dados. A quarta e última parte consiste na validação da diferença entre os dados reais e os gerados sinteticamente através de testes estatísticos como qui-quadrado, teste t e p-valor. Os resultados indicaram que os dados sintéticos seguem as características dos dados originários, mas ao mesmo tempo não são cópias da base de dados utilizada na geração, o que é interessante para treinamento de algoritmos de machine learning.

Palavras-Chave: Curvas Principais, Aumento de Dados, Aprendizado de Máquina.
Instituição de Fomento: FAPEMIG

Link do pitch: https://youtu.be/7yoMZf5MlEk