Engenharia de Controle e Automação

Desenvolvimento de algoritmos de aprendizagem de máquina baseado em Curvas Principais

Victor Daniel Reis - 9° Módulo de Engenharia de Controle e Automação, bolsista PIBIC/FAPEMIG.

Luiz Fernando Alves Rodrigues - 13º Módulo de Engenharia de Controle e Automação.

Fernando Elias de Melo Borges - Coorientador, Mestrando do Programa de Pós-Graduação de Engenharia de Sistemas e Automação, UFLA.

Danton Diego Ferreira - Professor do Departamento de Automática, UFLA. –danton@ufla.br. - Orientador(a)

Resumo

Desenvolvimento de algoritmos de aprendizagem de máquina baseado em Curvas Principais Data is the foundation of artificial intelligence and modern economics. However, obtaining them with high quality, low noise and good reliability to represent the real-world application becomes expensive and challenging. In addition, sophisticated artificial intelligent techniques require great volume of data for training pattern recognition methods and are limited to balanced classes. In order to compensate this gap, to generate synthetic data may be an unlimited, cheap and simple way. Thus, the Principal Curves (PC) technique acts as data representation model in which a multidimensional data set is represented by a one-dimensional curve. They are a non-linear generalization of Principal Component Analysis (PCA) and have them shape suggested by the data that projects it. The objective of this work is to produce an algorithm capable of generating synthetic data using Principal Curves. The methodology employed in this work can be divided into four parts: the first consists in obtaining the curves through the k-segments algorithm where important parameters are obtained (such as points projected on the segment and quadratic distances). The second includes the application of the produced algorithm to generate synthetic data, since the real dataset and two user-specified hyperparameters: dispersion factor and the percentage of data increases. The third part consists of training, validation (using the k-fold method with 10 folds) and prediction for a classifier model. For the last step, a comparison is made with the SMOTE (Synthetic Minority Over-sampling Technique) technique was made. The results indicated that the classifier that was trained with the Curves-based synthetic data obtained 99.6% accuracy, while using SMOTE obtained 96% accuracy. It was concluded that the method using Principal Curves is effective in generating synthetic data.

Palavras-Chave: Principal Curves, Data Augmentation, Machine Learning.
Instituição de Fomento: FAPEMIG

Link do pitch: https://www.youtube.com/watch?v=jsxopxmNPPk