

Data preparation and cleaning

July 4, 2023

pre-processing steps

By digitizing the data from a graph obtained online we obtained about 40 data points which were supposed to represent a range of data of approximately 10 million. For that reason, the Gaussian Mixture Model which is an expectation-maximization model was used to learn the distribution after which more data points were generated to help improve the performance of the classification algorithm.