JPUR
JOURNAL OF PURDUE UNDERGRADUATE RESEARCH

STEM

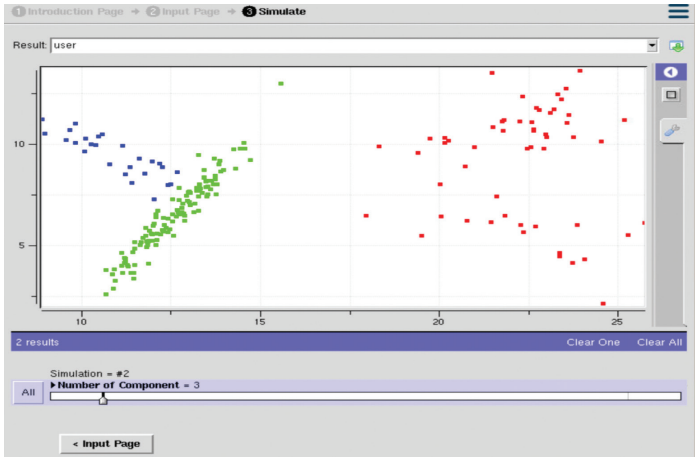
Model Selection Using Gaussian Mixture Models and Parallel Computing

Student researcher: Tian Qiu, Senior

Learning from data and analyzing data are the key factors that people want a machine to do in the machine learning field. A simple explanation for this process is that given the collected data, a machine can learn from it and predict the results for other, similar situations. More precisely in cluster analysis, machines are required to iteratively separate data into groups by their similar properties. The expectation-maximization algorithm and the Gaussian mixture model are common ways to cluster the dataset, according to Ryan M. Reynolds in his 2002 article titled "A Software Based Speaker Identification System Using Gaussian Mixture Model Classification." The generalized Gaussian distribution includes the Gaussian distribution as a particular case, and it can be parameterized in such a manner that its mean and variance coincide with the mean and variance of Gaussian distribution, writes Ruoxia Li, Vinay Prasad, and Biao Huang in their 2016 article titled "Gaussian Mixture Model-Based Ensemble Kalman Filtering for State and Parameter Estimation for a PMMA Process."

A model selection system is developed by using the finite multivariate generalized Gaussian mixture model, which organizes data points into clusters. Clustering basically assigns each dataset into different groups based on similarity. In this model, the expectation maximization method is used to calculate the distance from each point to the dummy center point, where the center point will change with the process of simulation to achieve the best-fitting results. Parallel computing is utilized to accelerate the simulation process. The performance of the developed model is studied through experimental evaluation with tens of thousands of data points and identification accuracy. Finally, the scalability of the system is determined by keeping the ratio of dataset points over the number of threats constant, which indicates a strong scalability.

Research advisor Guang Lin writes: "This work shows how the algorithm is scaled as we increase the data size and the number of threads."



This page will give the user results, which are calculated by this program. In this stage, the user can download the results and also see the details of each data point.

Qiu, T. (2017). Model selection using Gaussian mixture models and parallel computing. *Journal of Purdue Undergraduate Research*, *7*, 82–83. https://doi.org/10.5703/1288284316412