Gaussian mixture modeling for Smart Grid Stability Prediction

Nithin Sai Jalukuru
School of Engineering and Applied Sciences
University at Buffalo, Buffalo, NY, USA
njalukur@buffalo.edu

Abstract : In this present work, Classification of Smart grid stability is done using Gaussian mixture model with Expectation and Maximization Algorithm based on the various attributes of Electric Grid stability data.

Keywords : GMM, Clustering, Binary classification, Grid stability, EM algorithm

1.INTRODUCTION

The Gaussian Mixture Model (GMM) is a useful tool for modeling the clustering of data streams. The domains of signal and information processing make extensive use of it. GMM has well established modeling or estimate techniques where the number of Gaussian components k is known or constant.

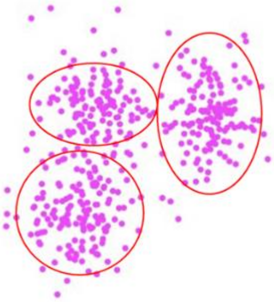


Fig1: Mixture of Gaussians (Source : CSE 474/574 Prof chen lecture slides

There are two methods for parameter estimation in GMMs: maximum likelihood (ML) and Bayesian approach. The Expectation Maximization (EM) technique is a popular solution to the ML problem.

Gaussian mixture model (GMM) clustering has been extensively studied due to its effectiveness and efficiency. Although it has shown promising performance in several applications, it is unable to handle the missing features in the data, which is a frequent occurrence in real-world applications.

1.1 Gaussian Mixture Model

The maximum likelihood estimation approach can be used to solve the model parameters given a data matrix, $X = \{\mathcal{X}_1, \mathcal{X}_2, \ldots\ldots, \mathcal{X}_m\}$, The objective formulation is represented as follows

$$\max_{\alpha,\mu,\Sigma} \sum_{j=1}^{n} \ln\left(\sum_{i=1}^{k} \alpha_i \frac{1}{2\prod^{n/2}|\Sigma_i|^{1/2}} e^{\frac{-1}{2}(x_j-\mu_j)^T \Sigma_i^{-1}(x_j-\mu_j)}\right)$$

where $\sum_i$ and μ_i are the parameters of the i^{th} Gaussian mixture component, and α_i is the corresponding mixture coefficient, which satisfies $\sum_{i=1}^{k} \alpha_i = 1$

Hence GMM formulation given posterior distribution is given by

$$p(X|Z,\mu) = \prod_{d=1}^{D}\prod_{k=1}^{k} \mathcal{N}(x_d|\mu_{dk})^{1(z=k)}$$

Thus, we now only have to learn the GMM parameter μ in terms of a univariate Gaussian for each dimension of d and k as follows

$$\mu = \begin{bmatrix} \mu_{11} & \cdots & \mu_{1K} \\ \vdots & \mu_{dk} & \vdots \\ \mu_{D1} & \cdots & \mu_{DK} \end{bmatrix}$$

Even though the current GMM clustering and its upgraded algorithms have had considerable success and shown promise in a variety of applications, they all use the assumption that the dataset is completely observable.

1.2 Expectation – Maximization algorithm

The EM (Expectation-Maximization) algorithm is a general-purpose method for estimating the maximum likelihood in a wide range of incomplete-data problems. The Expectation step, also known as the E-step, and the Maximization step, often known as the M-step, are the two steps that make up each iteration of the EM algorithm. As a result, the algorithm is known as the EM algorithm .

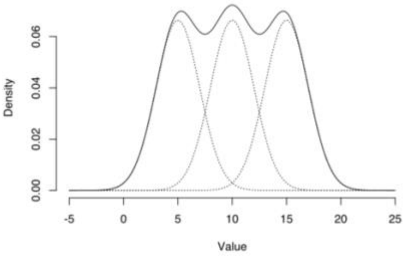


Fig2 : Gaussian Mixture Example (Source :Statistics how to)

To estimate latent variables, such as those derived from mixture distributions, one can utilize the EM algorithm (you know they came from a mixture, but not which specific distribution).

2. EXPERIMENTATION

1.1 Methodology

To implement the Gaussian mixture model, a data pipeline structure is constructed as illustrated in the below figure 4.

Data collection and loading into the environment are both parts of data extraction . After loading, the data has been labeled and cleaned so that the ML model can analyze it effectively. . In order to extract a few features or traits, feature extraction was also carried out, and the pertinent ones were chosen to train the model.

The proper ML model has been selected, trained, and validated in the model selection, training, and validation processes by maximizing the more suitable assessment criteria.

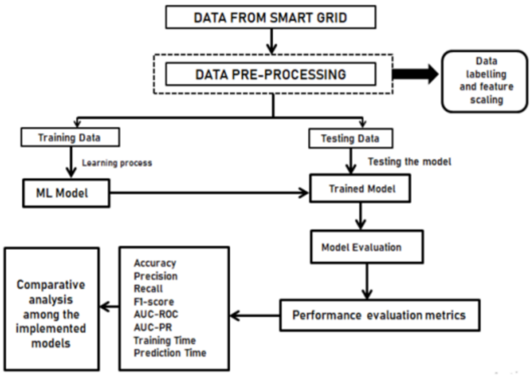


Fig4 : Data-pipeline

1.2 Data Extraction

The UCI Machine Learning Repository is where the Electrical Grid Stability Simulated Data dataset is sourced from. The dataset includes the grid stability simulation results for the reference 4-node star network depicted in Fig. 5.

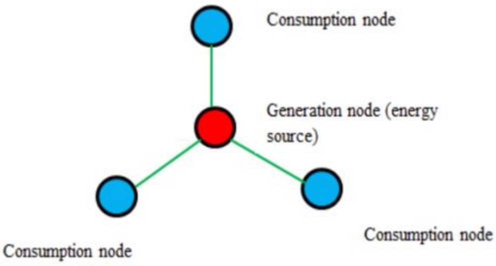


Fig5: Reference 4-node star network

This data contains 10000 rows and 14 predictive attributes. These are:

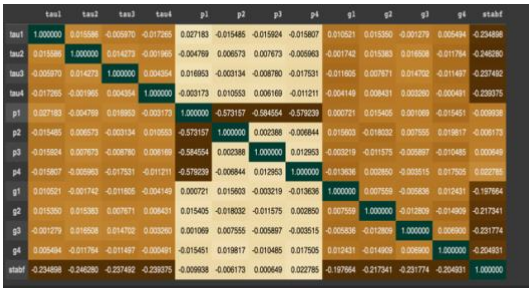
a) tau1 : Reaction time of Energy producer
b) tau2 : Reaction time- Consumer 1
c) tau3 : Reaction time- Consumer 2
d) tau4 : Reaction time - Consumer 3
e) p1 : Power balance - Energy producer
f) p2 : Power balance - Consumer 1
g) p3 : Power balance - Consumer 1
h) p4 : Power balance - Consumer 1
i) g1: Price elasticity co-efficient (gamma)- Energy producer
j) g2:Price elasticity co-efficient (gamma) - Consumer-1
k) g3:Price elasticity co-efficient (gamma) - Consumer-2
l) g4:Price elasticity co-efficient (gamma) - Consumer-3
m) stability target variable for regression. (Can be ignored)
n) The prediction is a categorical (binary) label
→ Stable
→Unstable

1.3 Data Cleaning and Feature Engineering

a) Removed unwanted columns , As The goal of the project is to classify the target , so regression target can be removed
b) To improve the performance of the ML model, Label Encoding has been applied to the target attribute. where 1 denotes instability and 0 indicates stability.
c) Performed feature standardization on the numerical data attributes , where the feature values are re-scaled to have a mean of $\mu = 0$ and standard deviation $\sigma = 1$.

3. EDA AND MODEL SELECTION

a) Correlation between features



	tau1	tau2	tau3	tau4	p1	p2	p3	p4	g1	g2	g3	g4	stabf
tau1	1.000000	0.015586	-0.005970	-0.017265	0.027183	-0.015485	-0.015924	-0.015807	0.010521	0.015350	-0.001279	0.005494	-0.234898
tau2	0.015586	1.000000	0.014273	-0.001965	-0.004769	0.006573	0.007673	-0.005963	-0.001742	0.015383	0.016508	-0.011764	-0.246280
tau3	-0.005970	0.014273	1.000000	0.004354	0.016953	-0.003134	-0.008780	-0.017531	-0.011605	0.007671	0.014702	-0.011497	-0.237492
tau4	-0.017265	-0.001965	0.004354	1.000000	-0.003173	0.010553	0.006169	-0.011211	-0.004149	0.008431	0.003260	-0.000491	-0.239375
p1	0.027183	-0.004769	0.016953	-0.003173	1.000000	-0.573157	-0.584554	-0.579239	0.000721	0.015405	0.001069	-0.015451	-0.009938
p2	-0.015485	0.006573	-0.003134	0.010553	-0.573157	1.000000	0.002388	-0.006844	0.015603	-0.018032	0.007555	0.019817	-0.006173
p3	-0.015924	0.007673	-0.008780	0.006169	-0.584554	0.002388	1.000000	0.012953	-0.003219	-0.011575	-0.005897	-0.010485	0.000649
p4	-0.015807	-0.005963	-0.017531	-0.011211	-0.579239	-0.006844	0.012953	1.000000	-0.013636	0.002850	-0.003515	0.017505	0.022785
g1	0.010521	-0.001742	-0.011605	-0.004149	0.000721	0.015603	-0.003219	-0.013636	1.000000	0.007559	-0.005836	0.012431	-0.197664
g2	0.015350	0.015383	0.007671	0.008431	0.015405	-0.018032	-0.011575	0.002850	0.007559	1.000000	-0.012809	-0.014909	-0.217341
g3	-0.001279	0.016508	0.014702	0.003260	0.001069	0.007555	-0.005897	-0.003515	-0.005836	-0.012809	1.000000	0.006900	-0.231774
g4	0.005494	-0.011764	-0.011497	-0.000491	-0.015451	0.019817	-0.010485	0.017505	0.012431	-0.014909	0.006900	1.000000	-0.204931
stabf	-0.234898	-0.246280	-0.237492	-0.239375	-0.009938	-0.006173	0.000649	0.022785	-0.197664	-0.217341	-0.231774	-0.204931	1.000000

Fig6 : Correlation plot

It is evident that p1 has high correlation with p2,p3,p4(As it p1 is absolute of p1+p2+p3). so removed p1(power balance- energy producer) because of collinearity.

b) Distribution of stability (numerical data)

The stability attribute clearly follows the gaussian distribution as shown in figure 7

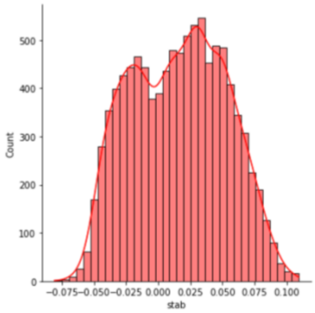


Fig7: Distribution plot

c) Stable and unstable Samples

Figure 8 demonstrates that samples for unstable and stable samples are nearly equivalent.

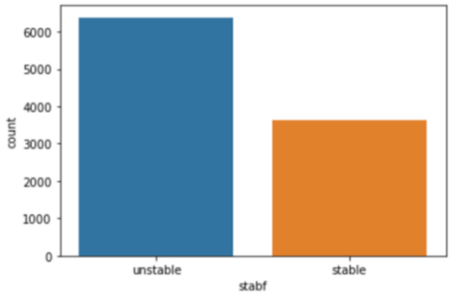
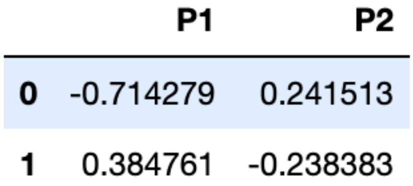


Fig8 : Count plot of the target variable

4. MODEL FITTING AND RESULTS

Given the large number of features in the data, it may be difficult to fit the GMM model and to see the results. Therefore, PCA (principal component analysis) has been developed to reduce the dimensionality while still containing all the data's information in order to lessen this issue.

Therefore, PC1 and PC2's is considered for further GMM analysis.



	P1	P2
0	-0.714279	0.241513
1	0.384761	-0.238383

Fig9: PC1 and PC2

Now that GMM is adjusted for PC1 and PC2, it can see the clustering that was done by GMM. In Fig10, results demonstrate how effectively the GMM model clusters the stability of the electric grid as stable and instable from actual data.

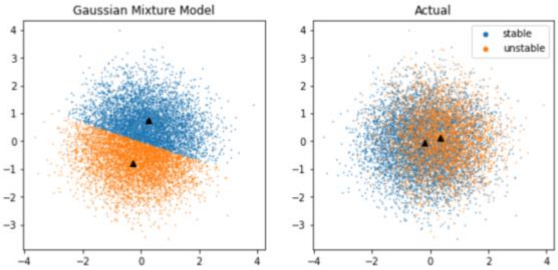


Fig10: Clustering plots

Let's also see how the GMM model's with many iterations of clustering look.

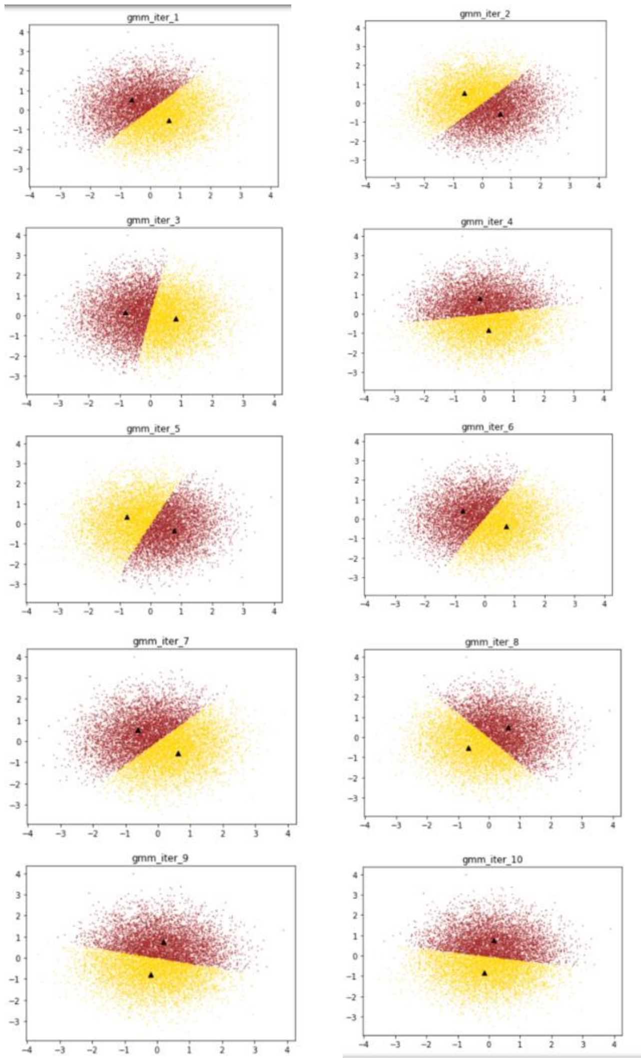
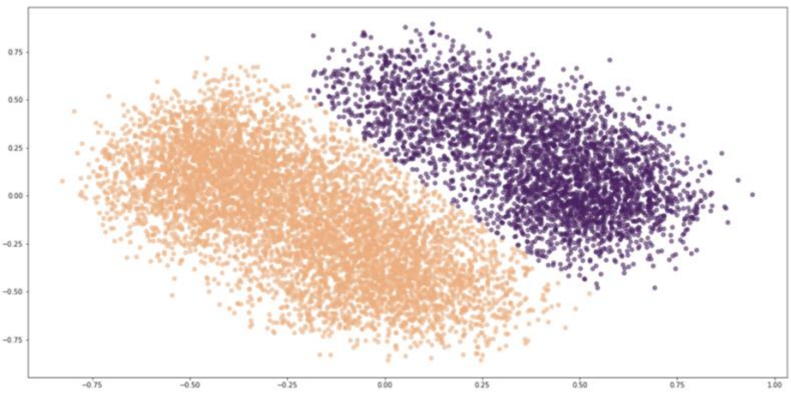


Fig11 : Clustering with respect to 10 iterations

4. CONCLUSION

The results above indicate that the data on grid stability have been successfully clustered using Gaussian Mixture modelling. Other Unsupervised learning techniques, such as K-means clustering, can further improve this.

Let's examine the clustering by GMM Model in its greater picture.



5. REFERENCES

1.CSE 474/574 lecture slides (Prof. Chen)

2. YI ZHANG at.El .”Gaussian Mixture Model Clustering with Incomplete Data”

3. GAO Ming-ming, Chang Tai-hua at.EL “Application of Gaussian Mixture Model Genetic Algorithm in Data Stream Clustering Analysis”

4. Kart-Leong Lim, Han Wang and Xiaozheng Mou ,”Learning Gaussian Mixture Model with a Maximization-Maximization Algorithm for Image Classification”