Regime Detection via Unsupervised Learning

Ashutosh Rabia
Department of Mathematics and Scientific Computing
Indian Institute of Technology Kanpur
arabia22@iitk.ac.in (+91-8949433821)

***Abstract*—This report summarizes an unsupervised learning approach aimed at detecting market regimes using high-frequency order book and trade data. The goal is to segment the market into different behavioral states along three dimensions: price dynamics (trending vs. mean-reverting), volatility (volatile vs. stable), and liquidity (liquid vs. illiquid). The analysis leverages two datasets: `depth20` (order book snapshots up to the top 20 levels) and `aggTrade` (aggregated trade volume data), with a series of hand-crafted features (e.g., bid-ask spread, imbalance, microprice, cumulative depth, rolling returns, and volatility).**

I. Feature Engineering

The initial steps involved loading the data, synchronizing timestamps into 1-second intervals using `pd.Grouper`, and engineering key features. The data was normalized (via StandardScaler) so that features could be combined effectively. A PCA was then applied to the normalized feature set (using 2 components) to facilitate visualization, although clustering was done on the full feature set.

II. Clustering and Regime Labeling

Multiple clustering methods (KMeans, Gaussian Mixture, and BIRCH) were evaluated using metrics such as Silhouette, Calinski-Harabasz, and Davies-Bouldin scores. Based on these metrics, the BIRCH algorithm produced the most distinct and compact clusters. The optimal number of clusters was determined as 7, implying 7 market regimes. Each timestamp was subsequently labeled with a regime number (0 to 6). Summary statistics for each regime (e.g., average volatility, bid-ask spread, mid-price return, and liquidity metrics) were computed to aid in the interpretation and naming of regimes. For instance, regimes with low volatility and high liquidity were labeled "Stable & Liquid," while those with high volatility and wide spreads were interpreted as "Volatile & Illiquid."

III. Model Architecture

A. Finding the Optimal Number of Clusters

To determine the optimal number of clusters, four different evaluation metrics were employed:

- **Elbow Method:** This method evaluates the within-cluster sum of squares (WCSS) as a function of the number of clusters (k). The optimal number of clusters is identified at the "elbow" point, where adding more clusters results in diminishing improvements.

$$WCSS = \sum_{i=1}^{k} \sum_{x \in C_i} ||x - \mu_i||^2 \tag{1}$$

  where C_i represents the i^{th} cluster, μ_i is the centroid of C_i, and x are the data points.
- **Silhouette Score:** This metric measures how similar a data point is to its own cluster compared to other clusters. It is defined as:

$$S(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \tag{2}$$

  where $a(i)$ is the average intra-cluster distance, and $b(i)$ is the average nearest-cluster distance for sample i.
- **Calinski-Harabasz Score:** Also known as the Variance Ratio Criterion, it is computed as:

$$CH = \frac{Tr(B_k)}{Tr(W_k)} \times \frac{n - k}{k - 1} \tag{3}$$

  where $Tr(B_k)$ is the trace of the between-cluster dispersion matrix, $Tr(W_k)$ is the trace of the within-cluster dispersion matrix, n is the number of samples, and k is the number of clusters.
- **Davies-Bouldin Score:** This index evaluates the average similarity between each cluster and its most similar cluster. It is given by:

$$DB = \frac{1}{k} \sum_{i=1}^{k} \max_{j \neq i} \left(\frac{s_i + s_j}{d_{ij}} \right) \tag{4}$$

  where s_i is the average distance between points in cluster i, and d_{ij} is the distance between cluster centroids i and j.

B. Clustering Algorithms Used

The following clustering algorithms were implemented:

- **K-Means:** A centroid-based clustering algorithm that minimizes intra-cluster variance:

$$J = \sum_{i=1}^{k} \sum_{x \in C_i} ||x - \mu_i||^2 \tag{5}$$

  where C_i is a cluster, and μ_i is its centroid.
- **Birch:** A scalable hierarchical clustering method that constructs a clustering feature (CF) tree and incrementally groups data.
- **Gaussian Mixture Model (GMM):** A probabilistic clustering method that models the data distribution as a mixture of multiple Gaussian distributions:

$$p(x) = \sum_{i=1}^{k} \pi_i \mathcal{N}(x | \mu_i, \Sigma_i) \tag{6}$$

Method	Silhouette Score	Calinski-Harabasz Score	Davies-Bouldin Score
KMeans	0.1995	47906.9184	1.2326
BIRCH	0.6138	23369.3622	0.8626
Gaussian Mixture	-0.0590	14233.3566	3.1914

TABLE I
CLUSTERING PERFORMANCE COMPARISON OF DIFFERENT ALGORITHMS WITH 7 CLUSTERS.

where π_i are the mixture weights, and $\mathcal{N}(x|\mu_i, \Sigma_i)$ represents the Gaussian distribution with mean μ_i and covariance Σ_i.

IV. Results and Discussion

The clustering analysis produced 7 regimes that capture distinct market behaviors. For instance:

- **Regime 0 – "Stable & Liquid":** Characterized by low volatility, narrow bid-ask spreads, and high cumulative quantities.
- **Regime 1 – "Volatile & Illiquid":** Exhibits high volatility with wide spreads and lower liquidity, suggesting market stress.
- **Regime 3 – "Positive Trending & Highly Liquid":** Indicates robust upward price movements with high liquidity and low volatility.

V. Conclusion

The analysis demonstrates that unsupervised learning techniques, coupled with thorough feature engineering and robust clustering validation, can effectively segment market conditions into interpretable regimes. The identification of 7 regimes provides actionable insights into market behavior, with regime transitions offering further potential for predictive modeling in trading strategies. Future work may include incorporating additional features, refining clustering methods, or integrating soft-clustering probabilities to enhance regime dynamics understanding.

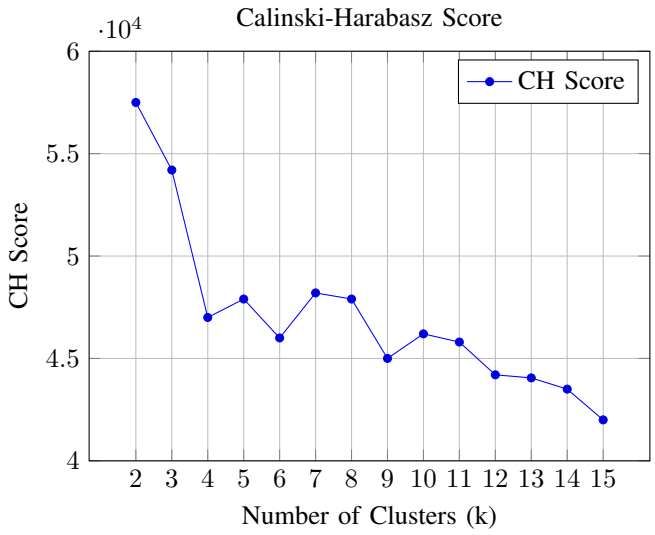

Fig. 1. Calinski-Harabasz Score for Different Clusters.

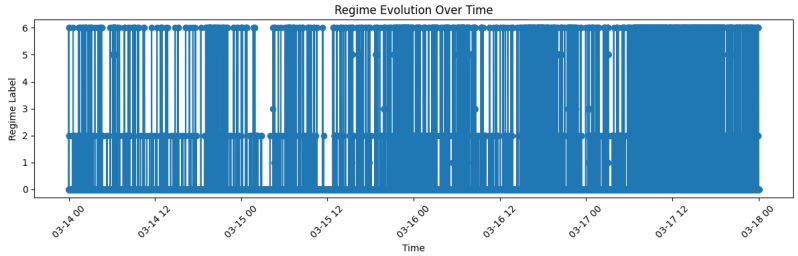

Fig. 2. Time-series plot showing regime evolution over time with mid-price overlay.

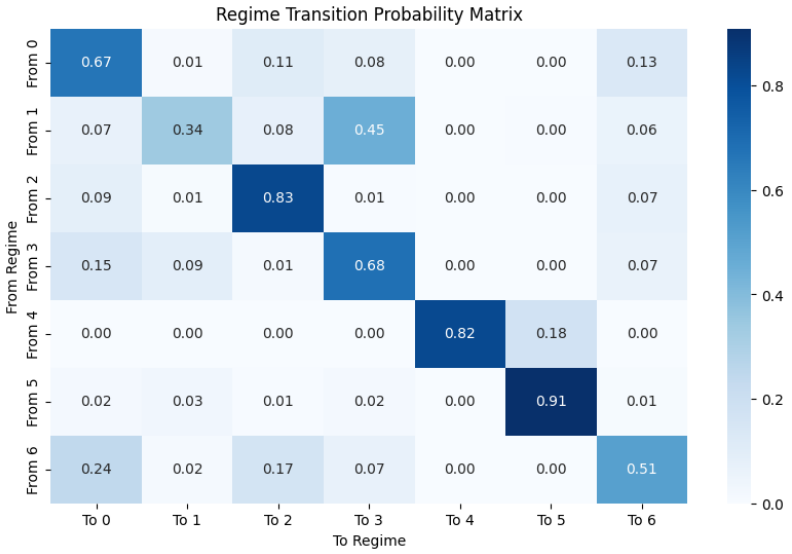

Fig. 3. Heatmap of the Regime Transition Probability Matrix.

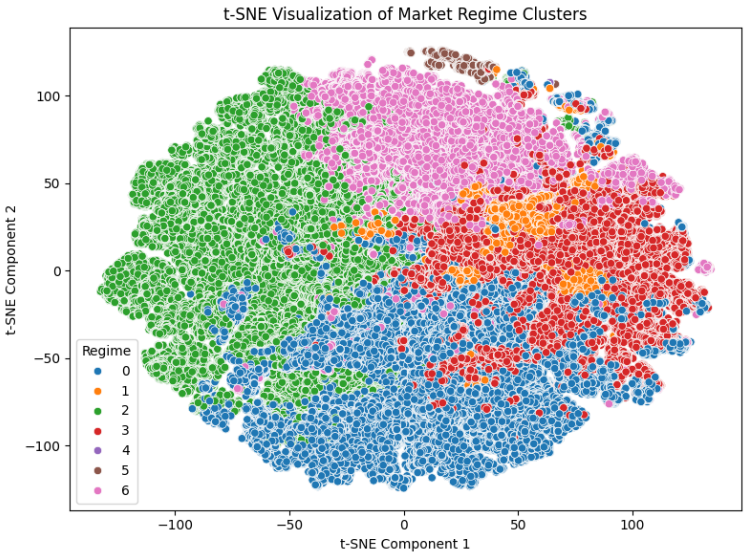

Fig. 4. TSNE visualization of Market regime clusters