CytoAutoCluster Overview

CytoAutoCluster is a deep learning-based solution designed for clustering cells based on identifiable characteristics. It uses semi-supervised learning techniques to improve clustering accuracy and computational efficiency, enabling meaningful insights from high-dimensional cytometry data.

Key Features

- Semi-Supervised Learning: Leverages both labelled and unlabelled data to enhance clustering accuracy.
- Efficient Grouping: Segments cells into distinct clusters for better interpretation of complex datasets.
- Interpretability: Generates clear visualizations of clusters for improved insights.

Problem Overview

Cytometry data is high-dimensional, which poses several challenges for analysis:

- High Dimensionality: Makes it difficult to interpret and visualize.
- Scarcity of Labelled Data: Limited labelled datasets hinder model performance.
- Noise and Variability: Biological noise can distort clustering accuracy.

CytoAutoCluster addresses these challenges by minimizing reliance on labelled data while maintaining high performance and interpretability.

Features Breakdown

- Dimensionality Reduction:
  - PCA: Reduces dimensions while preserving variance.
  - t-SNE: Preserves local structure in lower dimensions.
- Semi-Supervised Learning:
  - Consistency Regularization: Stabilizes predictions with input perturbations.
  - Entropy Minimization: Promotes confident predictions for unlabelled data.
  - Binary Masking: Focuses on relevant features.
- Advanced Models:
  - Autoencoder: Learns efficient data representations for clustering.
  - XGBoost: Enhances classification performance using both labelled and unlabelled data.
  - Logistic Regression: A baseline model to improve clustering accuracy with labelled data.
- Visualization:
  - t-SNE Visualization: Visualizes clusters using t-SNE for easy interpretation.

  - **Correlation Matrix: Identifies relationships between features.**
- **Performance Metrics:**
  - **Silhouette Score: Measures cluster quality.**
  - **Purity Score: Evaluates the purity of clusters.**
  - **Adjusted Rand Index: Assesses clustering accuracy.**
- **Gradio Interface: An interactive interface for real-time model exploration and visualization.**

Methodology

1. **Data Preparation: Clean high-dimensional cytometry data through exploratory analysis.**
2. **Dimensionality Reduction: Apply PCA and t-SNE for dimensionality reduction and visualization.**
3. **Semi-Supervised Learning: Use binary masking and consistency regularization for improved performance.**
4. **Model Training: Train both baseline and custom models to enhance clustering accuracy.**
5. **Visualization: Generate clear visualizations using Matplotlib and Seaborn.**

Technical Details

- **Key Techniques:**
  - **Kurtosis & Skewness Analysis: Identifies outliers and assesses data distribution.**
  - **Cluster Validation: Use Silhouette Score, Purity Score, and Adjusted Rand Index to evaluate clustering quality.**

Results

- **Improved clustering accuracy using semi-supervised learning techniques.**
- **Reduced reliance on labelled datasets with the help of consistency regularization and entropy minimization.**
- **Demonstrated robustness in handling noisy, high-dimensional data with minimal preprocessing.**
- **Provided clear and interpretable visualizations for the clustering results.**