CytoAutoCluster: Semi-Supervised Deep Approach for Cytometry Data Analysis

Project Overview

CytoAutoCluster is an innovative project aimed at advancing the detection and classification of cell populations through sophisticated deep clustering methodologies, specifically adapted for cytometry data. The core objective is to employ semi-supervised learning techniques that integrate both labeled and unlabeled data to improve the accuracy and interpretability of cell population classifications. This hybrid approach enables the creation of models that not only excel in cell classification but also provide meaningful biological insights into each cluster.

Data Cleaning and Pre-processing

Exclusion of Non-Informative Features

Several dataset columns, including Event, Time, Cell_length, and unique identifiers, were deemed irrelevant for this analysis and were therefore removed. These columns did not offer valuable information regarding cell characteristics, and including them might have introduced noise. By focusing only on the most relevant features, the clustering model can achieve better performance and interpretability.

Managing Missing Data

Missing values can adversely affect machine learning models. To handle this, we used mean imputation, where missing entries in each feature column are filled with that feature's mean value. This approach maintains the dataset’s overall

structure without introducing significant biases that might result from more complex imputation methods. Mean imputation is well-suited for our continuous dataset, preserving the data's distribution and integrity.

Data Standardization

To prepare the data for analysis, standardization was applied using StandardScaler, transforming each feature to have a mean of 0 and a standard deviation of 1. This scaling step is crucial for machine learning algorithms sensitive to input scales, ensuring all features are equally weighted. Standardization also facilitates dimensionality reduction methods like PCA (Principal Component Analysis) and t-SNE (t-Distributed Stochastic Neighbour Embedding), which are essential for visualizing clusters without bias from features with larger ranges.

Exploratory Data Analysis (EDA)

EDA is an essential step that provides a foundational understanding of the patterns and characteristics within the dataset.

Skewness and Kurtosis Evaluation

We conducted an analysis of skewness and kurtosis to better understand each feature's distribution. Skewness reflects the asymmetry of a feature's distribution, while kurtosis measures the "tailedness." By examining these metrics, we pinpointed features that might benefit from transformations to create a more normalized distribution. For instance, features with high skewness can be adjusted with log transformations, stabilizing variance and making the data more suitable for further analysis.

Correlation Analysis

A correlation matrix was created to evaluate the interdependencies and multicollinearity among features. This matrix offered insight into the relationships between features, helping us identify any potential redundancies. By understanding these correlations, we could make data-driven choices on feature retention or exclusion prior to applying dimensionality reduction methods like PCA. This step is critical for model interpretability and ensuring that clustering is based on the most informative features.

Dimensionality Reduction Techniques

Reducing the dimensionality of the dataset is essential for effective visualization and modelling, especially in high-dimensional spaces.

Principal Component Analysis (PCA):

PCA is a popular method for reducing dimensionality while preserving the original dataset's variance. After standardizing the features, we applied PCA and visualized the results using 2D and 3D scatter plots. These visualizations allow us to observe how data points group together in a lower-dimensional space and to identify the principal components that capture the most variance. Analyzing the PCA results provides insight into the data structure, guiding our choice of clustering methods.

t-Distributed Stochastic Neighbour Embedding (t-SNE):

t-SNE is a nonlinear dimensionality reduction technique that excels in visualizing high-dimensional data, particularly by preserving local relationships between data points. By applying t-SNE, we generated visualizations that reveal clustering and separation patterns within the dataset, highlighting potential structures among

cell populations. These insights help to better understand the relationships among cell groups and enhance the interpretability of clustering outcomes.

Data Augmentation Techniques

Binary Masking:

We used binary masking to simulate missing data by randomly hiding 30% of feature values. This approach allows us to evaluate the model's performance under incomplete data conditions, common in real-world datasets. Training on this augmented data improves model resilience and generalizability, preparing it to manage missing values effectively in real-world applications.

Column Shuffling for Feature Visibility:

To further diversify the dataset, we shuffled values within each feature column, increasing the data's complexity and adding controlled randomness. This augmentation technique enhances the clustering algorithms' ability to detect significant patterns in noisy data, improving the model's generalizability and reducing overfitting.

These detailed preprocessing, EDA, and augmentation steps lay a robust foundation for implementing semi-supervised learning models in CytoAutoCluster, enhancing both the accuracy and interpretability of cytometry data analysis. The project ultimately aims to advance our understanding of cell populations, with potential applications in fields such as immunology, oncology, and personalized medicine.