Report: Genetic Syndrome Classification

1. Introduction

This project aims to classify genetic syndromes based on image embeddings. The embeddings are 320-dimensional vectors from a pre-trained classification model. We used K-Nearest Neighbors (KNN) algorithms with different distance metrics to perform the syndrome classification.

2. Methodology

1. Data Loading and Preprocessing: The data was loaded from the file mini_gm_public_v0.1.p and transformed from a hierarchical structure (syndrome_id, subject_id, image_id) to embedding arrays and labels.

2. Exploratory Analysis: Descriptive statistics of the dataset were calculated, including total number of samples, number of different syndromes, and sample distribution per syndrome.

3. Visualization: The t-SNE technique was used to reduce embeddings from 320 dimensions to 2 dimensions, allowing visualization of the data in a two-dimensional space.

4. Classification: The K-Nearest Neighbors (KNN) algorithm was implemented with two different distance metrics (Euclidean and Cosine). For each metric, we tested k values from 1 to 15.

5. Cross-Validation: 10-fold cross-validation was used to evaluate model performance and determine the optimal k value.

6. Evaluation Metrics: AUC (Area Under the ROC Curve), F1-Score, Accuracy, and Top-k Accuracy metrics were calculated to compare the performance of the different configurations.

3. Results

Report: Genetic Syndrome Classification

3.1. Exploratory Analysis:

- Total number of samples: 1116

- Number of different syndromes: 10

- Average samples per syndrome: 111.60

- Standard deviation of samples per syndrome: 51.27

- Minimum samples per syndrome: 64

- Maximum samples per syndrome: 210

3.2. Classification:

Euclidean Distance - Best k: 15, AUC: 0.9504, F1: 0.7547, Accuracy: 0.7634

Cosine Distance - Best k: 15, AUC: 0.9630, F1: 0.7874, Accuracy: 0.7948

3.3. Top-k Accuracy:

Euclidean Distance (k=15):

Top-1 Accuracy: 0.7634, Top-3 Accuracy: 0.9247, Top-5 Accuracy: 0.9659

Cosine Distance (k=15):

Top-1 Accuracy: 0.7948, Top-3 Accuracy: 0.9418, Top-5 Accuracy: 0.9749

3.4. Conclusion:

The Cosine distance metric performed better with an AUC of 0.9630, indicating that the angular similarity between embedding vectors is more appropriate for this task than Euclidean distance.

4. Analysis and Interpretation

The superior performance of the Cosine distance metric compared to Euclidean suggests that directional similarity between embeddings is more important for genetic syndrome classification than absolute distance in the space.

Report: Genetic Syndrome Classification

The high AUC value (0.9630) for the Cosine metric indicates that the model is very effective at distinguishing between the different syndrome classes.

Exploratory analysis revealed class imbalance, with some syndromes having more samples than others (64 to 210 samples). This imbalance may affect model performance, especially for minority classes.

5. Challenges and Solutions

Challenges:

- Imbalanced data: Some classes had significantly more samples than others
- High dimensionality: 320-dimensional embeddings presented computational challenges
- Multiclass complexity: Classification among 10 different classes required robust methods

Solutions:

- Use of appropriate metrics such as AUC, F1-Score, and Accuracy for multiclass evaluation
- Cross-validation to obtain reliable performance estimates
- t-SNE visualization for understanding data clustering

6. Recommendations for Improvements

1. Increase dataset size for better representation of minority classes.
2. Explore other distance metrics and classification algorithms.
3. Apply balancing techniques such as SMOTE to handle class imbalance.
4. Perform feature engineering to extract additional information from embeddings.
5. Use ensemble techniques to improve model robustness.