Report on t-SNE and KNN Experiment

Introduction

The goal of this experiment was to visualize high-dimensional embeddings using t-SNE and to evaluate the performance of the K-Nearest Neighbors (KNN) algorithm with different distance metrics on the same dataset.

Data Loading and Extraction

The data was loaded from a file named `mini_gm_public_v0.1.p`. Syndrome IDs and embeddings were extracted using the `extract_data` function.

t-SNE Visualization

t-Distributed Stochastic Neighbor Embedding (t-SNE) was used to visualize the high-dimensional embeddings in 2D or 3D space. The embeddings were first normalized using the `normalize_data` function. The t-SNE algorithm was then applied with a perplexity of 35 and a maximum of 500 iterations. The results were visualized using `generate_tsne_plot` for 2D and `generate_tsne_plot_3d` for 3D.

The t-SNE visualization, as shown in Figure 1, exhibits a significant amount of overlap among clusters, regardless of the parameter settings chosen. This overlap could be indicative of the inherent complexity and structure of data. It is also possible that the embeddings contain noise or are not sufficiently informative to allow for clear separation.

KNN Performance Evaluation

The KNN algorithm's performance was evaluated using K-fold cross-validation with 10 splits. Two distance metrics were used: cosine and euclidean. The cross-validation results for both metrics were written to a PDF file using `write_pdf_comparison`.

Probability Predictions and ROC Curves

The KNN algorithm was then used to predict probabilities for the original embeddings and the t-SNE reduced embeddings using both distance metrics. The true labels and predicted probabilities were used to plot multiclass Receiver Operating Characteristic (ROC) curves using the `plot_multiclass_roc_curve` function.

The ROC curves generated in this experiment reveal that the performance of the KNN algorithm is better when using the originalembeddings compared to the t-SNE reduced embeddings. This outcome can be attributed to the loss of information that inherently occurs during the dimensionality reduction process. While t-SNE is effective for visualizing high-dimensional data in a lower-dimensional space, the technique may discard subtle nuances and relationships present in the original feature space that are crucial for classification tasks. The ROC curves suggest that the original embeddings contain discriminative features that are significant for the KNN algorithm to accurately classify instances, which are not as well-preserved in the t-SNE reduced space.

Reproduction Instructions

To reproduce this experiment, follow these steps:

1. Ensure that the `mini_gm_public_v0.1.p` file is placed in the `src/data/` directory.
2. Install all necessary Python packages. Use the command `poetry install` then `poetry shell`.
3. Run the `main.py` script, which will execute the entire workflow, from data loading to visualization and performance evaluation.

Make sure to adjust parameters such as `n_components`, `perplexity`, `n_splits`, and `n_neighbors` if needed to match the specific requirements of the dataset or the experiment's objectives.

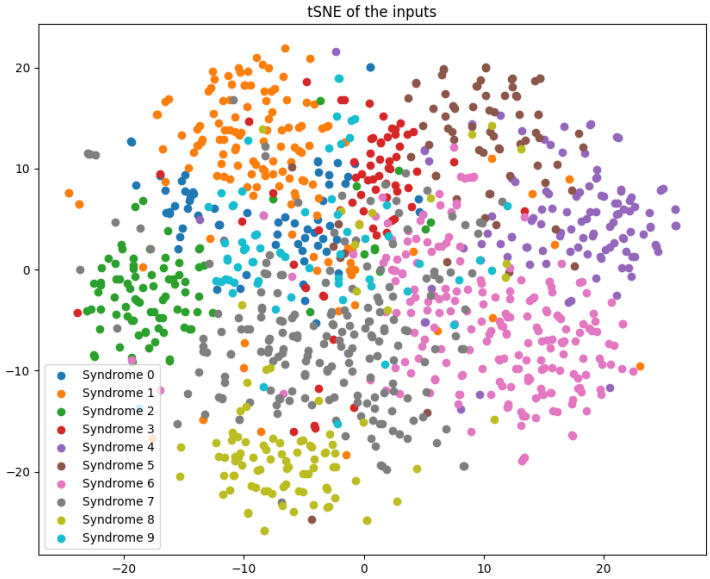

Figure 1: t-SNE visualization.

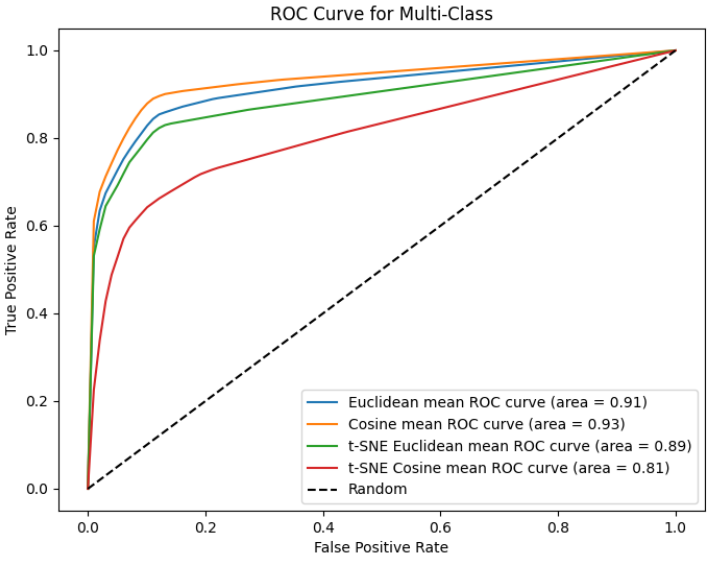

Figure 2: ROC Curves.