Assignment 4 – Classification Performance of MIMIC – III

Introduction

Dimensionality reduction is a critical step in data preprocessing, especially when dealing with high-dimensional datasets. This report evaluates the effectiveness of Principal Component Analysis (PCA) and Linear Optimal Low-Rank (LOL) projection, a method introduced by Vogelstein et al. (2021), on a dataset derived from the MIMIC-III database. Both techniques are evaluated based on their ability to preserve discriminative information for subsequent classification tasks.

Data and Methodology

Dataset

The dataset used in this analysis was extracted from the MIMIC-III database, as per assignment 3. It includes vital signs, laboratory measurements, and demographic information of patients, with the target variable being hospital survival status.

Preprocessing

1. **Feature and Target Separation**: The dataset was split into features **(X)** and the target variable **(y).**
2. **Categorical Variables**: Gender and ethnicity were encoded and analyzed using a chi-squared test to determine their association with hospital survival.
3. **Scaling**: All features were scaled using StandardScaler to ensure comparability.
4. **Hybrid Imbalance Classification**: The SMOTEENN (Synthetic Minority Over-sampling Technique and Edited Nearest Neighbors) method used to handle class imbalance, combining oversampling of the minority class and cleaning of the majority class to create a more balanced dataset.

Techniques Compared:

1. **Principal Component Analysis (PCA)**: PCA is an unsupervised dimensionality reduction technique that projects data onto directions of maximum variance.
2. **Linear Optimal Low-Rank (LOL) Projection**: LOL is a supervised technique that incorporates class-conditional moment estimates to improve the low-dimensional representation for classification tasks.

Analysis and Results

Chi-Squared Test:

- **Gender and Ethnicity**: The chi-squared test results indicated that gender and ethnicity had p-values greater than 0.05, indicating no significant association with the target variable. Therefore, these features were excluded from further analysis.

Dimensionality Reduction:

1. **PCA**: PCA was applied to the scaled dataset, retaining components that explained 80% of the variance.
2. **LOL**: LOL was implemented using class-conditional means and covariances, projecting the data onto the low-dimensional space defined by the top components derived from the LOL method.

Classification Performance:

- **Classifier**: Logistic Regression was used to evaluate the classification performance on the low-dimensional data obtained from both PCA and LOL.
- **Metrics**: Accuracy, precision, and recall were calculated for both methods.

Results:

- **PCA**:
    - PCA retained 11 components to explain 80% of the variance.
    - The classification metrics were:
        - Accuracy: 72%
        - Precision: 95%
        - Recall: 77%
- **LOL Projection**:
    - LOL retained 8 components, providing an improved low-dimensional representation.
    - The classification metrics were:
        - Accuracy: 73%
        - Precision: 96%
        - Recall: 78%

Discussion

Key Findings:

1. **Classification Accuracy:**
    - LOL consistently outperforms PCA in terms of classification accuracy.
2. **Robustness to Outliers:**
    - LOL is more robust to outliers due to its use of robust estimates of class-conditional moments.
3. **Optimal Number of Components:**
    - LOL requires fewer dimensions than PCA to achieve similar or better classification accuracy.
4. **Computational Efficiency:**
    - Both LOL and PCA are computationally efficient and scalable to large datasets.
5. **Visualization:**
    - Visual comparisons show that LOL provides better classification performance than PCA.

Comparison with Vogelstein et al. (2021):

- **Consistency**: Our findings are consistent with those reported by Vogelstein et al., where LOL projection demonstrated superior performance over PCA across various datasets, including high-dimensional neuroimaging and genomics datasets.
- **Scalability**: Both our analysis and Vogelstein et al.'s work emphasize the computational efficiency and scalability of LOL, making it suitable for large-scale data.

Conclusion

This analysis highlights the advantages of supervised dimensionality reduction techniques like LOL over traditional unsupervised methods like PCA. LOL provides a more discriminative low-dimensional representation, leading to improved classification performance, which is consistent with the findings presented by Vogelstein et al. (2021). Future work should explore additional supervised dimensionality reduction techniques and evaluate their performance on various types of datasets to further validate the robustness and generalizability of these methods. The classification errors plotted against the number of components confirm that LOL consistently achieves lower errors compared to PCA.