Report on

“Application of PCA for Improving Classifier Performance on the Fashion MNIST Dataset.”

Introduction

The Fashion MNIST dataset is a widely used benchmark dataset consisting of 70,000 grayscale images of fashion items, divided into 10 categories (such as T-shirts, trousers, bags, etc.). Each image is 28x28 pixels, resulting in 784 features per image. This high-dimensional dataset can benefit from dimensionality reduction techniques, such as Principal Component Analysis (PCA), which can help improve the performance of classifiers by reducing noise, computational complexity, and potential overfitting.

This report evaluates the performance of three classifiers-Decision Tree, Random Forest, and Support Vector Machine (SVM)-on the Fashion MNIST dataset. The evaluation is conducted both with and without applying PCA to the dataset. The objective is to assess whether PCA can enhance the accuracy and efficiency of these classifiers.

Methods:

1. **Loading the Dataset:**
    a. The Fashion MNIST dataset was loaded using the fetch_openml function from sklearn.datasets. The dataset contains 70,000 samples, each with 784 features (28x28 pixel images) and a corresponding label indicating one of the 10 categories.
2. **Data Preprocessing:**
    a. The dataset was split into training (80%) and testing (20%) sets to ensure an unbiased evaluation of model performance.
3. **Applying PCA:**
    a. PCA was applied to the training dataset to reduce its dimensionality while preserving 95% of the variance. The dimensionality reduction was performed using PCA from sklearn.decomposition.
    b. The transformed dataset was then used to train classifiers, and the PCA transformation was also applied to the test dataset for evaluation.
4. **Classifiers Used:**

   a. **Decision Tree Classifier:** A non-parametric supervised learning method used for classification and regression.
   b. **Random Forest Classifier:** An ensemble method that constructs multiple decision trees and outputs the mode of their predictions.
   c. **Support Vector Machine (SVM):** A supervised learning model that finds the optimal hyperplane for classification tasks.

5. **Training and Evaluation:**
   a. Each classifier was trained twice: once on the original dataset and once on the PCA-transformed dataset.
   b. Accuracy scores were computed for both scenarios to compare their performance.

Results:

The following are the accuracy scores for each classifier, both with and without PCA:

Classifier	**Accuracy Without PCA**	**Accuracy With PCA**
Decision Tree	**0.7968**	**0.7855**
Random Forest	**0.8764**	**0.8631**
SVM	**0.8912**	**0.8837**

Analysis:

1. **Decision Tree Classifier:**

   The accuracy of the Decision Tree classifier without PCA was slightly higher (0.7968) than with PCA (0.7855). This indicates that PCA did not significantly improve the performance of the Decision Tree classifier and may have even slightly reduced it. This outcome suggests that the Decision Tree classifier might already be capturing relevant patterns without the need for dimensionality reduction, or that the reduction in features may have led to a loss of some discriminative power.

2. **Random Forest Classifier:**

   For the Random Forest classifier, the accuracy without PCA (0.8764) was also slightly higher than with PCA (0.8631). Similar to the Decision Tree, the Random Forest classifier did not benefit from PCA. This might be due to the ensemble nature of Random Forests, which can inherently handle high-dimensional data by combining multiple decision trees and thus are less prone to overfitting compared to a single decision tree.

3. **Support Vector Machine (SVM):**

   The SVM classifier achieved the highest accuracy among the three classifiers, both with and without PCA. The accuracy without PCA (0.8912) was marginally higher than with PCA (0.8837). Since SVMs perform well in high-dimensional spaces, the marginal decrease in performance with PCA suggests that the original feature space was more suited to SVM, or that the dimensionality reduction led to a slight loss in separability.

Conclusion

Applying PCA to the Fashion MNIST dataset did not improve the performance of the Decision Tree, Random Forest, or SVM classifiers. A slight reduction in accuracy was observed for all classifiers with PCA, suggesting they can handle the high-dimensional data without dimensionality reduction.