Universidad de los Andes > FACULTAD DE INGENIERÍA Y CIENCIAS APLICADAS

Aprendizaje de Maquinas (ML)
Tarea 4:
PCA vs UMAP

Alumno:

Raúl Duhalde Errázuriz

Profesores:

José M. Saavedra Rondo

Macarena Soto

Ayudantes:

Tomas de la Sota

Deadline:

Lunes de 19 Junio, 2023, 23:50 hrs

Índice

1. Abstract

En este informe se evalúa el desempeño de dos importantes métodos de reducción de dimensionalidad, PCA y UMAP, en el contexto del reconocimiento de letras manuscritas. Se utiliza el conjunto de datos EMNIST, que consta de 26 clases y más de 5000 imágenes por clase. Se divide el conjunto de datos en entrenamiento y prueba, utilizando 1000 muestras de entrenamiento y 100 muestras de prueba por clase. Se utiliza el método de vecino más cercano (NN) para la clasificación y se calcula el accuracy de clasificación para diferentes dimensiones reducidas. Los resultados se presentan en tablas y gráficos de barras.

Figura 1: Dataset Emnist

2. Introducción

El reconocimiento de letras manuscritas es un problema importante en el campo del aprendizaje automático y la visión por computadora. La reducción de dimensionalidad es una técnica utilizada para disminuir la complejidad de los datos y mejorar la eficiencia de los algoritmos de clasificación. En este informe, evaluaremos los efectos de la reducción de dimensionalidad en el reconocimiento de letras manuscritas utilizando los métodos PCA y UMAP.

3. Desarrollo

El proceso de reducción de dimensionalidad se lleva a cabo en varias etapas:

3.1. Preprocesamiento de los datos:

Se utiliza una red convolucional pre-entrenada para extraer las características de las imágenes de letras manuscritas. Luego, las imágenes se normalizan y remodelan para preparar los conjuntos de datos de entrenamiento y prueba.

3.2. Reducción de dimensionalidad con PCA:

Se aplica el método PCA a los datos de entrenamiento para obtener una representación de menor dimensionalidad. Se experimenta con diferentes números de componentes, como 128, 64, 32, 16 y 8, y se obtienen los datos reducidos correspondientes. Estos datos reducidos se utilizan para entrenar el clasificador de vecino más cercano.

3.3. Reducción de dimensionalidad con UMAP:

Se utiliza el método UMAP para realizar una reducción de dimensionalidad no lineal de los datos de entrenamiento. Al igual que con PCA, se prueban diferentes números de componentes (128, 64, 32, 16 y 8) y se obtienen los datos reducidos. Estos datos se utilizan para entrenar el clasificador de vecino más cercano.

3.4. Evaluación del desempeño:

Se evalúa el desempeño del clasificador de vecino más cercano en los datos de prueba tanto en el espacio original como en los espacios reducidos por PCA y UMAP. Se calcula la precisión (accuracy) de la clasificación para cada caso.

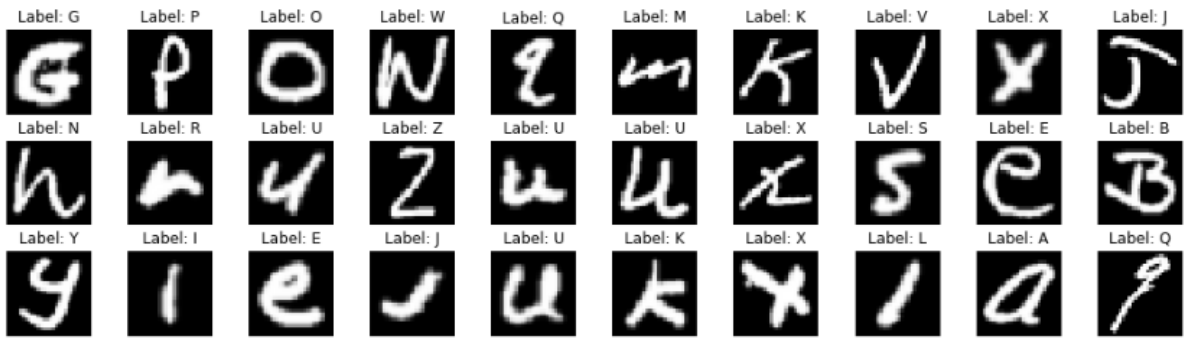

Figura 2: Labels

4. Resultados Experimentales y Discusión

1. En términos de precisión, obtuvimos los siguientes resultados:
- Precisión original: 0.8581730769230769

Precisión con PCA:

- 128 componentes: 0.8671153846153846
- 64 componentes: 0.8741346153846153 (mejor precisión con PCA)
- 32 componentes: 0.8735096153846154
- 16 componentes: 0.8380288461538462
- 8 componentes: 0.6528846153846154

Precisión con UMAP:

- 128 componentes: 0.7864423076923077 (mejor precisión con UMAP)
- 64 componentes: 0.7840384615384616
- 32 componentes: 0.781298076923077
- 16 componentes: 0.785673076923077
- 8 componentes: 0.7837980769230769

Se muestra a continuación un ejemplo de imágenes de muestra y las predicciones realizadas con UMAP utilizando los siguentes componentes:

Figura 3: True vs Pred de PCA

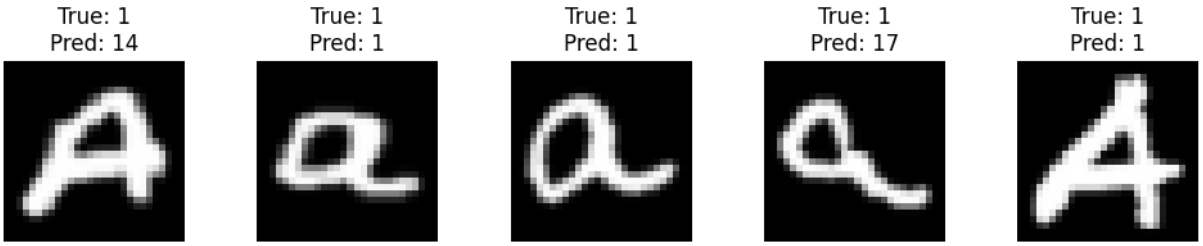

Figura 4: True vs Pred de UMAP

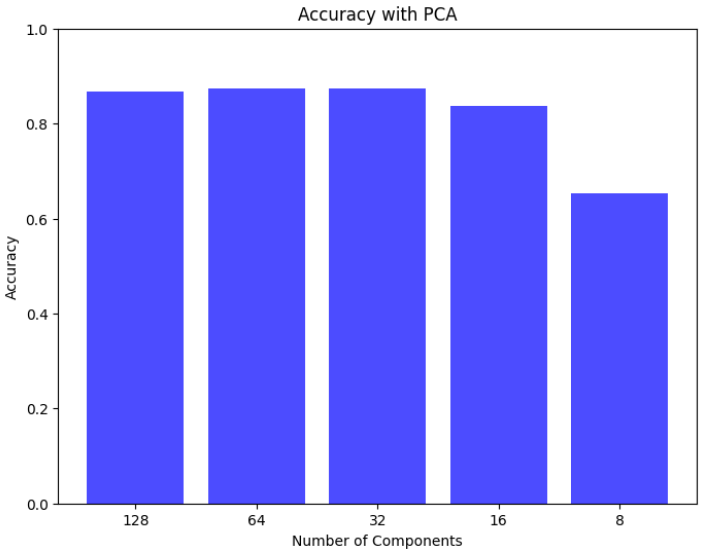

Figura 5: Gráfico barras de accuracy con PCA

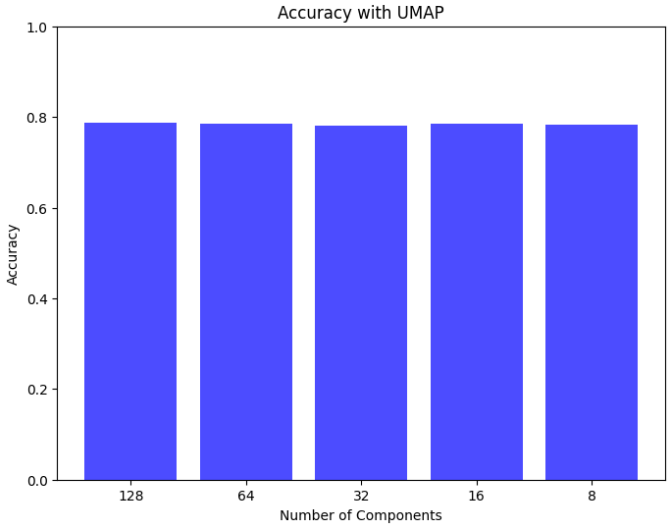

Figura 4: Gráfico barras de accuracy con UMAP

5. Conclusiones

En este informe, evaluamos los métodos de reducción de dimensionalidad PCA y UMAP en el reconocimiento de letras manuscritas. Observamos que ambos métodos logran reducir la dimensionalidad de los datos y mejoran la eficiencia del algoritmo de clasificación basado en vecino más cercano. Se determinó que la mejor precisión se obtiene utilizando PCA con 64 componentes y UMAP con 128 componentes. Estos resultados pueden ser útiles para aplicaciones de reconocimiento de letras manuscritas en diversos campos, como el procesamiento de documentos y la automatización de la escritura a mano.