Ejercicio Grupo E8FP

Jesús Gomez, Maxim Polyak

Resumen

En este trabajo se analiza un conjunto de datos de señales electromiográficas (EMG) correspondientes a cuatro gestos de mano distintos. Se incluyen los pasos de carga y preprocesado, un análisis exploratorio de datos (EDA) con estadísticas descriptivas y visualizaciones (histogramas, distribución de clases, proyección PCA), y la evaluación de tres clasificadores (Random Forest, LDA y QDA) sobre una partición 80 % entrenamiento – 20 % test. Finalmente, se discuten los resultados y se proponen recomendaciones para futuras mejoras.

1. Introducción

Las señales EMG reflejan la actividad eléctrica generada por la contracción muscular y se emplean ampliamente en interfaces cerebro-máquina, prótesis controladas por músculo y sistemas de reconocimiento de gestos. El objetivo de este estudio es comparar distintos modelos de clasificación automática para determinar cuál distingue mejor cuatro gestos básicos de mano.

2. Objetivos

1. Cargar y preprocesar los datos brutos de múltiples archivos CSV.
2. Realizar un EDA completo para entender la distribución, rango y correlaciones de las señales.
3. Entrenar y evaluar tres modelos de clasificación:
    a. Random Forest
    b. Linear Discriminant Analysis (LDA)
    c. Quadratic Discriminant Analysis (QDA)
4. Comparar la precisión, recall y F1-score de los modelos sobre un conjunto de test (20 %).
5. Extraer conclusiones y proponer mejoras.

3. Descripción de los datos

Origen: Cuatro archivos CSV (HandGesture0.csv … HandGesture3.csv) en /data.

Formato: Cada fila es una muestra temporal de EMG; las primeras columnas son valores numéricos de amplitud (uV) en distintos canales, y la última columna es la etiqueta de gesto (0–3).

Volumen: ≈ 3 000 muestras por clase (total ≈ 12 000 muestras). No existen nombres de columna en los CSV.

4. Metodología

1. Carga y concatenación
    a. Se leen todos los CSV sin cabeceras con pandas.read_csv(header=None) y se concatenan.
    b. Se separan X (características) y y (etiquetas).
2. Particionado
    a. División aleatoria en 80 % entrenamiento y 20 % test, manteniendo proporción de clases (stratify=y).
3. Análisis Exploratorio de Datos (EDA)
    a. Estadísticas descriptivas: media, desviación, cuartiles, mínimo, máximo y rango por canal.
    b. Histograma de amplitudes para un canal de ejemplo.
    c. Distribución de clases: gráfico de barras.
    d. Proyección PCA 2D para visualizar la separación entre gestos.
4. Modelado
    a. Se definen tres clasificadores con sus hiperparámetros por defecto:
        i. RandomForestClassifier(random_state=42)
        ii. LinearDiscriminantAnalysis()
        iii. QuadraticDiscriminantAnalysis()
    b. Cada modelo se entrena sobre X_train, y_train y se evalúa en X_test, y_test.

5. Métricas
    a. Accuracy global.
    b. Precision, recall y F1-score por clase.
    c. Matriz de confusión del modelo más preciso (QDA).

5. Análisis Exploratorio de Datos

1. Estadísticas descriptivas

    a. Media cercana a 0 uV, desviaciones típicas de ~20–30 uV y rangos de hasta ±100 uV.
    b. La variabilidad sugiere suficiente información para distinguir gestos.

2. Histograma canal 0

    a. Concentración principal de valores entre –5 uV y +5 uV.
    b. Colas largas en extremos: picos musculares transitorios.

3. Distribución de clases

    a. Aproximadamente 580 muestras por clase.
    b. Balance evita sesgos de entrenamiento.

4. Proyección PCA

    a. Gestos 0 y 1 aparecen más separados; gestos 2 y 3 se solapan moderadamente.
    b. Indica la necesidad de clasificadores con capacidad de modelar estructuras no lineales o covarianzas específicas.

6. Resultados de Modelado

Modelo	Accuracy	Precision medio	Recall medio	F1-score medio
Random Forest	0.9225	0.922	0.922	0.922
LDA	0.3390	0.350	0.340	0.340
QDA	0.9358	0.935	0.935	0.935

7. Discusión

- **QDA** obtiene la mejor precisión (≈ 93,6 %), gracias a su estimación de covarianza por clase.
- **Random Forest** (≈ 92,3 %) es casi tan eficaz y más robusto al ruido, pero no modela explícitamente la forma de las distribuciones.
- **LDA** falla (≈ 34 %) al imponer una covarianza común, demasiado restrictiva para este conjunto.
- El **gesto 3** presenta el recall más bajo (≈ 0.89 en QDA, ≈ 0.87 en RF), confirmando su solapamiento con el gesto 2 observado en la PCA.

8. Conclusiones

1. La variabilidad y el balance de clases en las señales EMG facilitan el entrenamiento de modelos fiables.
2. Para este problema, clasificadores con flexibilidad no lineal o con covarianzas por clase (QDA) son imprescindibles.
3. QDA se recomienda como primer candidato por su máxima precisión, siempre valorando su coste computacional.
4. Random Forest es una alternativa sólida en entornos productivos, debido a su tolerancia al ruido y a configuraciones sencillas.
5. Para mejorar el gesto 3:
    a. Incrementar el número de muestras o sesiones.
    b. Extraer características adicionales (por ejemplo, potencias en bandas de frecuencia EMG).
    c. Explorar métodos de selección de características y modelos avanzados (SVM con kernels, redes neuronales ligeras).