

Tarea 1 – Aprendizaje Automático

Reconocimiento de letras manuscritas usando SVM – RBF kernel.

Nícolas Farfán Cheneaux

Resumen

1. Objetivo del Experimento

El objetivo principal es evaluar la eficacia del modelo SVM (Support Vector Machine) con kernel RBF (Radial Basis Function) en la clasificación de letras manuscritas, empleando diferentes estrategias de preprocesamiento de los datos para optimizar el rendimiento de clasificación.

2. Conjunto de Datos

Se utiliza el conjunto de datos EMNIST Letters que consta de:

- **26 clases:** Correspondientes a cada letra del alfabeto.
- **145,600 imágenes:** Cada imagen de 28x28 píxeles en escala de grises.

3. División de Datos

Para el entrenamiento y las pruebas se siguieron los siguientes criterios:

- **Datos de entrenamiento:** 26,000 imágenes, con 1000 muestras seleccionadas aleatoriamente por cada clase.
- **Datos de prueba:** 2,600 imágenes, con 100 muestras aleatorias por cada clase que no se solapan con las muestras de entrenamiento.

4. Configuración del Modelo SVM

El modelo SVM se configura con un kernel RBF, utilizando inicialmente los parámetros por defecto. Posteriormente, se realizó una búsqueda en grilla aleatoria para encontrar los mejores valores para los hiperparámetros **C** y **gamma**, con el fin de optimizar el modelo para el conjunto de datos específico.

5. Estrategias de Preprocesamiento

Se experimentó con varias técnicas de preprocesamiento:

1. **Directa:** Las imágenes se convirtieron directamente en vectores de 784 elementos.
2. **PCA:** Se aplicó PCA para reducir la dimensionalidad de los vectores a 128

componentes, mejorando así la eficiencia computacional y posiblemente la capacidad de generalización del modelo.

3. **Características CNN:** Se extrajeron características utilizando una red convolucional simple, probando una aproximación más sofisticada para la extracción de características.

6. Métricas de Evaluación

Se emplearon las siguientes métricas para evaluar el rendimiento:

- **Accuracy por Categoría:** Medida de la precisión de clasificación para cada una de las 26 letras.
- **Accuracy Total:** Precisión global del modelo en el conjunto de datos de prueba.

7. Resultados Esperados y Análisis

- **Variabilidad del Rendimiento:** Se anticipa que las técnicas de preprocesamiento influirán significativamente en el rendimiento del modelo. Específicamente, se espera que la reducción de dimensiones y las características avanzadas extraídas permitan al SVM manejar mejor la variabilidad intraclase en las imágenes.
- **Optimización de Hiperparámetros:** La búsqueda en aleatoria es fundamental usando validación cruzada para afinar el modelo, esperándose que ajustes en **C** y **gamma** resulten en mejoras sustanciales en la precisión, particularmente en clases con formas de letras más ambiguas o variadas.

8. Visualización de Resultados

Los resultados se visualizarán mediante tablas que muestren la precisión por categoría y la precisión total, además de gráficos de barras para una representación clara y directa del rendimiento del modelo en cada una de las configuraciones de preprocesamiento.

Este resumen proporciona una vista integral del enfoque metodológico, los procedimientos técnicos y los objetivos analíticos para la evaluación del modelo SVM en el reconocimiento de letras manuscritas, apuntando a resultados que apoyen decisiones informadas sobre la selección de modelos y técnicas de procesamiento en aplicaciones de visión por computadora relacionadas con el procesamiento de texto manuscrito.

Introducción

En el campo de la visión por computadora, el reconocimiento de caracteres manuscritos es un área de investigación intensiva que presenta desafíos significativos debido a la variabilidad inherente en la escritura humana. El reconocimiento preciso de caracteres manuscritos es crucial para numerosas aplicaciones, incluyendo la automatización de la entrada de datos, la mejora de interfaces de usuario accesibles, y el procesamiento de documentos históricos digitalizados.

El conjunto de datos EMNIST Letters, derivado del bien conocido conjunto de datos MNIST, proporciona una oportunidad amplia para desarrollar y evaluar algoritmos de clasificación debido a su gran volumen de datos y diversidad representativa de las formas de letras en la escritura manuscrita. Con 26 clases que representan cada letra del alfabeto inglés y más de 145,600 imágenes, este conjunto de datos es ideal para entrenar modelos robustos y precisos.

Dentro de este contexto, el modelo de Máquinas de Vectores de Soporte (SVM) con el kernel de Función de Base Radial (RBF) es notable por su capacidad para manejar espacios de características de alta dimensión y su eficacia en la clasificación de datos complejos. Sin embargo, el rendimiento del SVM puede ser altamente dependiente de la selección adecuada de hiperparámetros y del preprocesamiento de los datos.

Este estudio se centra en evaluar sistemáticamente la eficacia del SVM en la tarea de clasificación de letras manuscritas, explorando diferentes estrategias de preprocesamiento para optimizar su rendimiento. Mediante la implementación de técnicas como la conversión directa de imágenes en vectores, la reducción de dimensionalidad a través del Análisis de Componentes Principales (PCA), y la extracción de características mediante redes convolucionales, se busca determinar el enfoque más efectivo para mejorar la precisión de la clasificación del SVM. Adicionalmente, se empleará una búsqueda en grilla para ajustar los hiperparámetros del modelo, buscando maximizar la precisión global y por categoría.

El resultado de este estudio proporcionará una visión integral sobre la aplicabilidad del SVM en el reconocimiento de letras manuscritas y ofrecerá directrices sobre cómo configurar eficazmente el modelo y sus parámetros para este tipo de tareas. Esto no solo avanzará en nuestro entendimiento del comportamiento del SVM en contextos específicos, sino que también servirá

como referencia para futuras investigaciones y desarrollos en tecnologías de procesamiento de imágenes manuscritas.

Desarrollo

1. Carga de Datos

El primer paso en el proceso experimental fue la carga eficiente de los datos del dataset EMNIST Letters. Este dataset está estructurado en un archivo .mat, que contiene imágenes de letras manuscritas distribuidas en 26 clases, cada una correspondiente a una letra del alfabeto. Utilizamos la biblioteca **scipy.io** para cargar estos datos desde el archivo .mat.

2. Preprocesamiento de Datos

El preprocesamiento de datos es crucial para la eficacia del modelo de aprendizaje automático. Para el SVM, es esencial que las imágenes se normalicen y, opcionalmente, se reduzcan en dimensionalidad. Implementamos lo siguiente:

- **Normalización:** Escalamos los datos para tener un rango de valores entre 0 y 1. Esto se realiza para mejorar la convergencia del modelo durante el entrenamiento.
- **Reducción de Dimensionalidad:** Utilizamos PCA (Análisis de Componentes Principales) para reducir la dimensionalidad de los datos, facilitando así un entrenamiento más rápido y posiblemente más efectivo.

3. Configuración y Entrenamiento del Modelo SVM

Donde en el primer experimento se obtiene un **accuracy** total de **85%**, sin ajustar hiperparámetros y en el segundo **87%** y el tercer experimento se obtiene **90%**.

Ajuste de hiperparámetros

Se realiza un ajuste de hiperparámetros usando **RandomizedGridSearch** para los hiperparámetros de “C” y “gamma”, se toman los primeros 3500 datos para acelerar el proceso de la validación cruzada.

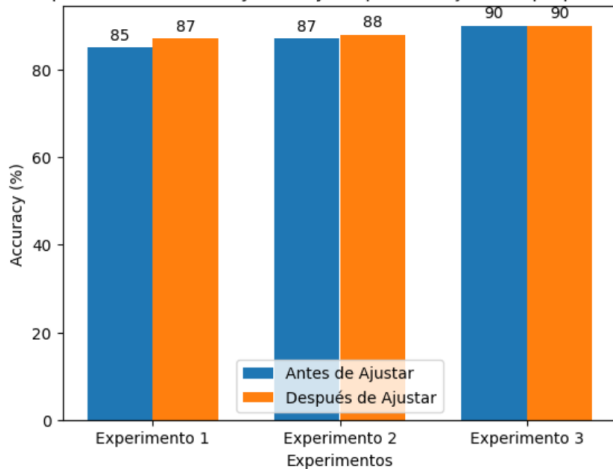
Se obtiene que el mejor modelo con los correspondientes hiperparámetros usando **best_estimator_**.

```
rnd_search_cv.best_estimator_  
SVC  
SVC(C=6.136205694069264, gamma=0.0050287915480361)
```

Se utilizan estos hiperparámetros para entrenar nuevamente el modelo, logrando una **accuracy** de **87%**.

Se realiza el mismo procedimiento para el segundo experimento, logrando una **accuracy** de **88%**.

Comparación de Accuracy antes y después de ajustar hiperparámetros



Tal como se muestra en el gráfico.

4. Diferencia entre experimentos

La teoría dice que un modelo **SVM** de menor dimensión, es decir menor largo de los vectores de características tendrá un mejor desempeño. Dado los experimentos hechos, ajustando parámetros para el primer experimento y segundo experimento, esto es así dado que aumentó en 2% el **accuracy** al aplicar Análisis de Componentes principales. Además, en términos de desempeño es más rápido entrenar un modelo con menos dimensiones pasando de 784 a 128. Por el lado de la red neuronal, se obtiene una mayor **accuracy** dado que los datos se encuentran mejor caracterizados ya que el vector de características no solamente son los valores de cada píxel sino una selección discriminante para cada imagen.

Comparación con referencias externas

Se consultan diversas implementaciones de personas usando el mismo **dataset** y se encuentra que obtienen valores de **accuracy** total aproximadamente rondando el 83% al 87%. Por lo que se considera que el enfoque de la presente solución usando el kernel RBF y ajustando los hiperparámetros en caso sea necesario dado que se tiene un vector de características muy extenso es correcto y válido.

Conclusión

El uso de SVM con kernel RBF es muy útil para conjuntos de datos donde la relación entre las clases no es lineal. Es crucial ajustar los parámetros como **C** y **gamma** para

encontrar el mejor modelo que equilibre bien entre el ajuste y la generalización. Scikit-learn ofrece una implementación robusta y fácil de usar que facilita la experimentación y optimización de estos modelos.

Fuentes externas

<https://www.kaggle.com/code/tepemok/emnist-different-approaches>

<https://www.kaggle.com/code/great23u5/emnist-using-svm>

https://dmkothari.github.io/Machine-Learning-Projects/SVM_with_MNIST.html