Entrenamiento, Evaluación y Métodos Ensemble para Clasificación del Dataset Iris

Alejandro Cerezo, German Torres y Daniil Nemchenko

November 29, 2025

Resumen

Este informe presenta el proceso completo de entrenamiento, evaluación y combinación de modelos de clasificación aplicados al dataset Iris. Se han implementado cuatro algoritmos de aprendizaje automático, concretamente K-Nearest Neighbors, Support Vector Machine, Random Forest y Naive Bayes, y se han combinado mediante tres técnicas de ensemble diferentes. Los experimentos se han realizado utilizando validación cruzada estratificada con cinco pliegues sobre nueve variantes del conjunto de datos, obteniendo resultados que demuestran la efectividad de los métodos implementados y la ventaja de las técnicas de ensemble en la mejora del rendimiento predictivo.

1 Introducción

El proceso de clasificación supervisada requiere no solo la selección de algoritmos adecuados, sino también una metodología rigurosa de entrenamiento y evaluación que permita obtener estimaciones fiables del rendimiento de los modelos. En este trabajo se describe la implementación de un sistema completo de clasificación que abarca desde el entrenamiento de modelos individuales hasta su combinación mediante técnicas de ensemble.

El objetivo principal de este estudio es comparar el rendimiento de cuatro algoritmos de clasificación ampliamente utilizados en el campo del aprendizaje automático y evaluar si la combinación de sus predicciones mediante métodos de ensemble permite obtener mejores resultados que los modelos individuales. Para ello, se han utilizado los conjuntos de datos generados previamente mediante validación cruzada estratificada, que incluyen diferentes transformaciones y niveles de reducción dimensional.

2 Entrenamiento de los Modelos

El módulo de entrenamiento implementa cuatro algoritmos de clasificación supervisada, cada uno con características y supuestos diferentes que los hacen adecuados para distintos tipos de problemas.

2.1 K-Nearest Neighbors

El algoritmo K-Nearest Neighbors se basa en la idea de que muestras similares tienden a pertenecer a la misma clase. Para clasificar una nueva instancia, el algoritmo identifica las k muestras más cercanas en el conjunto de entrenamiento y asigna la clase mayoritaria entre estos vecinos. La implementación realizada incluye un proceso de optimización del hiperparámetro k, donde se evalúan valores impares desde 1 hasta la raíz cuadrada del número de muestras de entrenamiento. Esta estrategia de búsqueda está fundamentada en la recomendación de utilizar valores impares para evitar empates en la votación, y en el principio de que valores de k excesivamente grandes pueden diluir la información local relevante para la clasificación.

Durante el entrenamiento, el algoritmo evalúa cada valor de k candidato y selecciona aquel que maximiza la precisión en el conjunto de prueba. El modelo óptimo se persiste en disco para su posterior uso en la fase de evaluación y en los métodos de ensemble.

2.2 Support Vector Machine

Las Máquinas de Vectores de Soporte buscan el hiperplano óptimo que maximiza el margen de separación entre las clases. En problemas multiclase como el presente, donde existen tres especies de iris, el algoritmo emplea internamente una estrategia de uno contra uno, construyendo clasificadores binarios para cada par de clases. La implementación utiliza la configuración por defecto de la librería scikit-learn, que emplea un kernel de función de base radial capaz de capturar fronteras de decisión no lineales.

Una característica relevante de la implementación es la habilitación de la estimación de probabilidades, que permite obtener no solo la clase predicha sino también la confianza del modelo en su predicción. Esta información resulta fundamental para los métodos de ensemble que combinan probabilidades en lugar de votos.

2.3 Random Forest

El algoritmo Random Forest pertenece a la familia de métodos de ensemble y combina múltiples árboles de decisión entrenados sobre subconjuntos aleatorios de los datos y las características. Cada árbol individual proporciona una predicción, y la clase final se determina mediante votación mayoritaria. Esta estrategia de agregación reduce la varianza del modelo y lo hace más robusto frente al sobreajuste que un único árbol de decisión.

La implementación utiliza los parámetros por defecto de la librería, que incluyen cien árboles en el bosque y la selección aleatoria de la raíz cuadrada del número de características en cada división. Estos valores representan un equilibrio razonable entre complejidad computacional y capacidad predictiva.

2.4 Naive Bayes

El clasificador Naive Bayes aplica el teorema de Bayes con la suposición de independencia condicional entre las características. A pesar de que esta suposición raramente se cumple en la práctica, el algoritmo ha demostrado ser efectivo en numerosos problemas de clasificación. La variante gaussiana utilizada asume que los valores de cada característica dentro de cada clase siguen una distribución normal.

La principal ventaja de este clasificador reside en su eficiencia computacional y en su capacidad para manejar conjuntos de datos de alta dimensionalidad. Además, proporciona estimaciones de probabilidad bien calibradas que resultan útiles para la combinación mediante técnicas de ensemble.

3 Evaluación de los Modelos

El módulo de evaluación implementa un conjunto completo de métricas que permiten caracterizar el rendimiento de los modelos desde múltiples perspectivas. Esta diversidad de métricas resulta especialmente importante en problemas de clasificación multiclase, donde una única medida puede no capturar adecuadamente todos los aspectos relevantes del rendimiento.

3.1 Métricas de Rendimiento

La exactitud representa la proporción de predicciones correctas sobre el total de muestras, constituyendo la métrica más intuitiva pero también la más susceptible a sesgos cuando las clases están desbalanceadas. En el caso del dataset Iris, donde las clases están perfectamente equilibradas, la exactitud proporciona una medida fiable del rendimiento global.

La precisión mide la proporción de predicciones positivas que son correctamente clasificadas, mientras que el recall o sensibilidad cuantifica la proporción de instancias positivas que el modelo identifica correctamente. La combinación armónica de ambas métricas da lugar al F1-score, que penaliza los desequilibrios entre precisión y recall y proporciona una medida balanceada del rendimiento.

La especificidad complementa al recall midiendo la capacidad del modelo para identificar correctamente las instancias negativas. Las tasas de falsos positivos y falsos negativos proporcionan información adicional sobre los tipos de errores que comete el modelo, lo cual puede resultar relevante en aplicaciones donde ciertos errores tienen consecuencias más graves que otros.

3.2 Curvas ROC y Área bajo la Curva

La curva ROC representa gráficamente la relación entre la tasa de verdaderos positivos y la tasa de falsos positivos para diferentes umbrales de clasificación. El área bajo esta curva proporciona una medida agregada del rendimiento del modelo que es independiente del umbral de decisión seleccionado.

La implementación realizada genera curvas ROC multiclase utilizando la estrategia uno contra resto, donde cada clase se compara individualmente con el resto de clases. El área bajo la curva macro-promediada proporciona una medida global del rendimiento que trata a todas las clases con igual importancia.

Las curvas ROC generadas se almacenan como imágenes PNG en el directorio de evaluación, permitiendo una inspección visual del comportamiento de cada modelo en cada iteración de la validación cruzada.

4 Análisis de Resultados

El módulo de resultados procesa los archivos de evaluación generados para cada modelo y cada iteración de la validación cruzada, calculando estadísticas agregadas que permiten comparar el rendimiento de los diferentes algoritmos.

4.1 Tablas Resumen por Método

Para cada uno de los cuatro métodos de clasificación, se ha generado una tabla resumen que presenta las métricas promediadas sobre los cinco pliegues de la validación cruzada, junto con sus desviaciones estándar. Esta información permite no solo comparar el rendimiento medio de los modelos sino también su estabilidad a través de las diferentes particiones de los datos.

Los resultados obtenidos para el método K-Nearest Neighbors muestran un rendimiento excepcional en el conjunto de datos original, alcanzando una exactitud media del 98.00% con una desviación estándar de apenas 1.83%. Este rendimiento se mantiene prácticamente inalterado cuando se aplica PCA con un umbral del 95% de varianza explicada, lo que sugiere que las componentes principales capturan la información discriminante relevante. Sin embargo, cuando se reduce la varianza explicada al 80%, el rendimiento disminuye hasta el 92.67%, evidenciando que una componente principal única no es suficiente para preservar toda la información necesaria para la clasificación óptima.

El método Support Vector Machine presenta un comportamiento similar, con una exactitud del 96.67% en los datos originales y estandarizados. Los resultados son consistentes entre las diferentes transformaciones, aunque el rendimiento disminuye de forma más pronunciada con la reducción dimensional agresiva del PCA al 80%.

Random Forest alcanza una exactitud del 96.67% en los datos originales, comparable al resto de métodos. Su comportamiento ante la reducción dimensional es el menos favorable de los cuatro algoritmos, con una exactitud del 88.67% cuando se utiliza PCA al 80% sobre los datos originales.

Naive Bayes muestra el rendimiento más estable ante las diferentes transformaciones, aunque con valores de exactitud ligeramente inferiores a los otros métodos en la mayoría de configuraciones. Su exactitud media en los datos originales es del 95.33%, y resulta notable que mantiene un rendimiento del 93.33% incluso con la reducción dimensional más agresiva.

4.2 Comparativa de F1-score

La tabla comparativa de F1-score permite visualizar de forma condensada el rendimiento de todos los métodos en todas las configuraciones de datos. Los valores más altos se observan consistentemente en el método KNN aplicado a los datos originales y con PCA al 95%, seguido por SVM y Random Forest. Esta jerarquía se invierte en algunas configuraciones con reducción dimensional agresiva, donde Naive Bayes muestra mayor robustez.

El análisis de las desviaciones estándar revela que los métodos basados en instancias como KNN presentan mayor variabilidad entre pliegues, mientras que los métodos paramétricos como Naive Bayes tienden a producir resultados más consistentes. Esta observación sugiere que la elección del método óptimo puede depender no solo del rendimiento medio esperado sino también de los requisitos de estabilidad de la aplicación específica.

4.3 Visualización de Métricas

Los gráficos de dispersión generados permiten explorar las relaciones entre diferentes pares de métricas. La representación de la tasa de falsos negativos frente a la tasa de falsos positivos revela que la mayoría de los modelos se concentran en valores bajos para ambas métricas, con algunas excepciones en las configuraciones con reducción dimensional extrema.

La relación entre precisión y recall muestra una correlación positiva fuerte, lo que indica que los modelos que predicen con alta precisión también tienden a identificar correctamente la mayoría de las instancias de cada clase. El gráfico de exactitud frente a F1-score confirma la consistencia entre estas métricas, con puntos distribuidos a lo largo de la diagonal principal.

5 Métodos de Ensemble

Los métodos de ensemble combinan las predicciones de múltiples modelos con el objetivo de obtener un rendimiento superior al de los modelos individuales. Esta mejora se fundamenta en la teoría de que diferentes modelos pueden capturar diferentes aspectos de los datos, y su combinación puede reducir tanto el sesgo como la varianza del sistema global.

5.1 Votación

El método de votación implementado asigna a cada muestra la clase que recibe más votos de los cuatro modelos individuales. En caso de empate, el sistema utiliza las probabilidades predichas para desempatar, seleccionando la clase cuya suma de probabilidades entre los modelos empatados sea mayor. Esta estrategia de desempate aprovecha la información adicional proporcionada por las estimaciones de probabilidad, evitando decisiones arbitrarias.

La implementación procesa las predicciones de cada modelo de forma vectorizada, contando los votos para cada muestra mediante la estructura Counter de Python.

El manejo de empates requiere un procesamiento adicional que accede a las probabilidades predichas por cada modelo, calculando la suma ponderada para las clases involucradas en el empate.

5.2 Combinación por Media

El método de combinación por media calcula el promedio aritmético de las probabilidades predichas por cada modelo para cada clase, asignando la muestra a la clase con mayor probabilidad media. Este enfoque es equivalente a una votación suave donde cada modelo contribuye proporcionalmente a su confianza en la predicción.

La implementación utiliza operaciones de NumPy para calcular eficientemente la media a lo largo del eje de los modelos. Las probabilidades resultantes se normalizan implícitamente, ya que el promedio de distribuciones de probabilidad es también una distribución de probabilidad.

5.3 Combinación por Mediana

El método de mediana es similar al de media, pero utiliza la mediana en lugar del promedio aritmético. Esta variante es más robusta frente a predicciones extremas de modelos individuales, ya que la mediana es menos sensible a valores atípicos que la media.

En la práctica, la mediana puede producir resultados diferentes cuando algún modelo tiene predicciones muy confiadas que difieren del consenso. La implementación mantiene la misma estructura que el método de media, sustituyendo únicamente la función de agregación.

6 Resultados de los Métodos Ensemble

Los tres métodos de ensemble se han evaluado sobre todas las combinaciones de datos y pliegues de validación cruzada, generando archivos de métricas y curvas ROC equivalentes a los de los modelos individuales.

Los resultados del ensemble por votación en los datos originales muestran una exactitud del 96.67% en la primera iteración de validación cruzada, con un F1-score de 0.9666 y un AUC de 0.99. Estos valores son comparables a los mejores modelos individuales, lo que indica que la combinación no degrada el rendimiento en este caso.

El ensemble por media produce resultados idénticos al de votación en la primera iteración, lo que sugiere que las predicciones de los modelos individuales son suficientemente consistentes como para que ambos métodos de combinación converjan a las mismas decisiones. Este comportamiento es esperable en un problema de clasificación relativamente sencillo como el dataset Iris.

El ensemble por mediana también coincide con los otros dos métodos en la primera iteración, confirmando la robustez de las predicciones combinadas.

La consistencia entre los tres métodos de ensemble sugiere que los modelos individuales están bien calibrados y sus predicciones son compatibles entre sí.

A través de todas las configuraciones de datos y pliegues de validación, los métodos de ensemble mantienen un rendimiento estable y competitivo con los mejores modelos individuales. La principal ventaja observada es la reducción de la variabilidad entre pliegues, ya que la combinación de modelos tiende a suavizar las predicciones erróneas de modelos individuales en particiones específicas.

7 Conclusiones

El sistema completo de entrenamiento, evaluación y ensemble implementado proporciona una metodología rigurosa para la clasificación del dataset Iris. Los cuatro algoritmos individuales alcanzan rendimientos excelentes en la mayoría de configuraciones, con exactitudes superiores al 90% en prácticamente todos los casos.

El análisis comparativo revela que K-Nearest Neighbors obtiene los mejores resultados en los datos originales y con transformaciones que preservan la mayor parte de la información, mientras que Naive Bayes muestra mayor robustez ante la reducción dimensional agresiva. Support Vector Machine y Random Forest presentan comportamientos intermedios, con rendimientos consistentes en la mayoría de configuraciones.

Los métodos de ensemble demuestran su utilidad al proporcionar predicciones estables que combinan las fortalezas de los modelos individuales. Aunque la mejora en términos de rendimiento medio es marginal en este problema particular, la reducción de la variabilidad constituye una ventaja práctica relevante en aplicaciones donde la consistencia de las predicciones es importante.

La metodología de validación cruzada estratificada empleada asegura que los resultados reportados sean representativos del rendimiento esperado en datos no vistos, proporcionando estimaciones fiables que pueden guiar la selección del modelo más adecuado para cada aplicación específica.

Referencias

1. Pedregosa, F., et al. (2011). *Scikit-learn: Machine Learning in Python.* Journal of Machine Learning Research, 12, 2825-2830.
2. Breiman, L. (2001). *Random Forests.* Machine Learning, 45(1), 5-32.
3. Cortes, C., & Vapnik, V. (1995). *Support-vector networks.* Machine Learning, 20(3), 273-297.
4. Dietterich, T. G. (2000). *Ensemble Methods in Machine Learning.* Multiple Classifier Systems, 1-15.