Classificação de Dados Utilizando Random Forest e k-Nearest Neighbors para Análise de Desempenho de Algoritmos de Machine Learning

Marcel Matsumoto
Universidade Federal de Viçosa - Campus Rio Paranaíba
Curso de Sistemas de Informação

12 de novembro de 2024

Rio Paranaíba - MG

Resumo

Neste trabalho, são investigados e comparados dois algoritmos de classificação amplamente utilizados em machine learning: Random Forest (RF) e k-Nearest Neighbors (kNN). O objetivo é analisar o desempenho de ambos os algoritmos em termos de acurácia e robustez na classificação de um conjunto de dados. Com base nos resultados obtidos, avalia-se a eficiência de cada método, utilizando métricas de avaliação como precisão, revocação e a matriz de confusão para uma análise mais detalhada dos acertos e erros de classificação.

1 Introdução

A classificação de dados em machine learning é uma área que visa categorizar dados de entrada em classes predefinidas com base em padrões identificáveis. Neste estudo, analisei dois algoritmos de aprendizado supervisionado: Random Forest (RF) e k-Nearest Neighbors (kNN), com foco em como cada um lida com a variação e complexidade dos dados.

A motivação para este trabalho surge do projeto da disciplina de Visão Computacional, que propõe realizar um estudo que visa comparar esses algoritmos através da análise das matrizes de confusão resultantes, ilustradas nas imagens `CM-RF` (Random Forest) e `CM-kNN` (k-Nearest Neighbors), respectivamente.

2 Metodologia

A metodologia adotada inclui a aplicação dos algoritmos RF e kNN sobre dados previamente processados, o conjunto de imagens "mpeg7". Ambos os algoritmos foram implementados utilizando a biblioteca `scikit-learn` em Python, com o objetivo de garantir a comparabilidade dos resultados em um ambiente controlado.

2.1 Random Forest

O algoritmo Random Forest é um método de aprendizado por conjunto que constrói múltiplas árvores de decisão durante o treinamento e gera a classe de saída que é a mais representativa entre as saídas de cada árvore. Esse método é particularmente eficiente para dados com alta dimensionalidade e oferece vantagens em termos de precisão, pois reduz o risco de overfitting.

2.2 k-Nearest Neighbors

O algoritmo kNN é um classificador baseado em instâncias que classifica uma amostra de acordo com a maioria das classes dos k vizinhos mais próximos. Este algoritmo é intuitivo e fácil de implementar, embora tenha um custo computacional elevado em termos de memória e processamento para grandes conjuntos de dados, pois precisa calcular distâncias para todas as amostras.

3 Resultados e Discussão

A comparação entre os algoritmos foi realizada com base nas métricas obtidas nas matrizes de confusão apresentadas. A matriz de confusão mostra a distribuição dos acertos e erros para cada classe prevista, permitindo uma análise detalhada do desempenho dos algoritmos. As figuras 1 e 2 exibem as matrizes de confusão dos algoritmos RF e kNN, respectivamente.

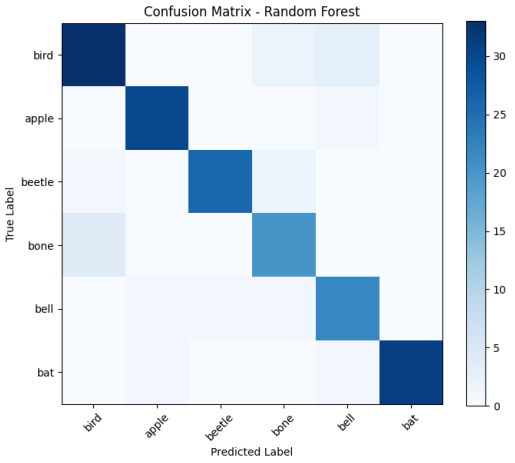

Figura 1: Matriz de Confusão para o Classificador Random Forest

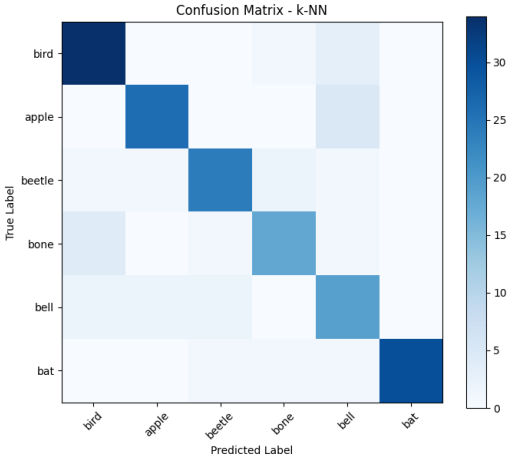

Figura 2: Matriz de Confusão para o Classificador k-Nearest Neighbors

3.1 Análise de Desempenho

Observa-se que o algoritmo Random Forest apresentou uma taxa de acurácia superior em comparação com o kNN, especialmente em classes onde há maior variabilidade de características. O resultado médio de precision, recall e f1-score para o kNN foi de 0.85, 0.84 e 0.84 respectivamente, com uma accuracy de oitenta e quatro por cento. Já o RF demonstrou os resultados 0.90, 0.90 e 0.90, com uma accuracy de noventa por cento. Isso indica que o RF é mais robusto ao lidar com dados de natureza variada, provavelmente devido à capacidade de agregar decisões de múltiplas árvores.

Por outro lado, o kNN obteve resultados satisfatórios em classes mais homogêneas, embora tenha se mostrado sensível a outliers. Isso pode ser explicado pelo fato de o kNN basear sua classificação na proximidade de instâncias, o que o torna suscetível a erros em dados com maior ruído ou que contenham valores atípicos.

3.2 Interpretação das Matrizes de Confusão

As matrizes de confusão indicam que o RF teve menos erros de classificação nas classes dominantes, enquanto o kNN apresentou maior dispersão nos erros, refletindo seu comportamento em relação a outliers. A matriz de confusão do RF, ilustrada na Figura 1, mostra um número maior de acertos ao longo da diagonal principal, sinalizando uma maior precisão do modelo. Em contraste, a matriz do kNN na Figura 2 apresenta uma distribuição de erros mais difusa, o que confirma sua vulnerabilidade a variações no conjunto de dados.

4 Conclusão

Os resultados indicam que o Random Forest é mais adequado para conjuntos de dados com classes variadas e onde a robustez contra overfitting é desejada. Já o kNN pode ser útil para conjuntos de dados homogêneos e de menor escala, onde o custo computacional de calcular as distâncias não é um problema significativo.

Este estudo evidencia a importância de escolher o algoritmo adequado com base nas características do conjunto de dados e nas necessidades específicas de aplicação.

Trabalhos futuros podem incluir a análise de outros hiperparâmetros que possam influenciar o desempenho dos classificadores, bem como a exploração de técnicas de pré-processamento de dados para reduzir o impacto de outliers.