Instituto brasileiro de Ensino, Desenvolvimento e Pesquisa - IDP

Relatório sobre otimização do algoritmo de Random Forest

Fábio Carvalho e Sara Pacheco
3º semestre Engenharia de Software

O algoritmo de Random Forest é uma técnica de aprendizado de máquina que utiliza um conjunto de árvores de decisão para realizar previsões. O processo de construção envolve a criação de várias árvores de forma independente, onde cada uma delas é treinada com uma amostra aleatória dos dados e com uma seleção aleatória de características para cada divisão de nó.

A principal ideia é que, ao combinar os resultados de várias árvores, o modelo consiga melhorar sua precisão e robustez. Para tarefas de classificação, o Random Forest toma a decisão com base na maioria dos votos das árvores. Já quando se trata da regressão, ele calcula a média das previsões de todas as árvores. Essa técnica é eficaz porque reduz o risco de overfitting (caso em que o modelo se apoia excessivamente nos dados de treinamento, oferecendo pouco desempenho quando se tratando de dados novos), que pode ocorrer em árvores de decisão individuais, garantindo assim maior generalização para novos dados.

O uso do Random Forest é muito usado em problemas que envolvem grandes conjuntos de dados com muitas variáveis, sendo eficaz tanto em tarefas de classificação quanto em regressão.

Neste trabalho, utilizamos um dataset de classificação de vinhos para avaliar o desempenho do algoritmo de Random Forest. Inicialmente, o modelo apresentou uma acurácia de 88%, o que já é considerado um resultado muito bom para essa tarefa. Nosso objetivo foi, então, investigar se técnicas de otimização poderiam melhorar ainda mais esse desempenho.

Para realizar a otimização, aplicamos três métodos distintos: *Randomized Search*, *Bayesian Search* e, por fim, *Grid Search*. Essas técnicas permitem explorar diferentes combinações de hiperparâmetros na tentativa de encontrar a configuração que maximiza o desempenho do modelo.

Os principais parâmetros mantidos fixos durante esse processo foram:

- n_estimators: Representa o número de árvores no Random Forest. Um valor maior pode aumentar a robustez do modelo, mas também eleva o custo computacional.

- max_features: Define o número máximo de características que serão consideradas em cada divisão do nó da árvore. Ajustar esse valor controla o balanceamento entre diversidade nas árvores e correlação entre elas.

- max_depth: Especifica a profundidade máxima das árvores. Limitar a profundidade pode evitar o overfitting, pois restringe a complexidade do modelo.

- min_samples_split: Determina o número mínimo de amostras necessárias para dividir um nó. Um valor maior força a criação de nós com mais dados, promovendo uma generalização mais forte.

No entanto, mesmo após a aplicação dessas técnicas de otimização, os melhores resultados obtidos inicialmente variaram entre 80% e 82% de acurácia, ou seja, abaixo do desempenho original de 88%, isso ocorreu por conta da base de dados desbalanceada que não foi levada em conta na realização dos primeiros testes, mas depois de um ajuste nos parâmetros foi obtido 89%. Esse resultado pode ser interpretado como uma demonstração de que, embora os algoritmos de otimização sejam poderosos para ajustes finos, nem sempre conseguem superar muito a acurácia inicial de um modelo, especialmente quando esta já é relativamente alta .

Em conclusão, o Random Forest aplicado ao dataset de vinhos obteve um resultado impressionante sem a necessidade de ajustes complexos e dos algoritmos de otimização utilizadas. As tentativas, embora metodologicamente válidas, não proporcionaram ganhos adicionais.