Trabalho Prático de Estrutura de Dados 2 - *Random Forest*

Guilherme Aguilar de Oliveira*, Tiago Gonçalves da Silva†, e Vitor Matheus Triches da Rosa‡
Departamento Acadêmico de Informática - DAINF
Universidade Tecnológica Federal do Paraná - UTFPR, Curitiba, Brasil
RA: *2127954, †2023644, ‡2040042

***Resumo*—A disciplina de Estrutura de dados 2 tem como trabalho final a aplicação e aprofundamento de conceitos trabalhados como árvores binárias de busca, árvores balanceadas, árvores B, filas de prioridades entre outros. Dessa forma, a aplicação escolhida foi implementar e descrever a estrutura *Random Forest* na qual provê um método eficaz de aprendizagem para classificação de dados. Neste trabalho, foi implementado na linguagem C a *Random Forest* e algumas funções auxiliares a fim de testar a eficácia da estrutura. Além disso um conjunto de dados foi utilizado e separado em amostras aleatórias de treino e teste em que verificamos aproximadamente uma taxa de acerto de 93.37% em média.**

***Palavras-chave*—*Random Forest*, Estruturas de Dados, Modelos de Classificação**

I. Introdução

Nesse contexto, este trabalho tem como objetivo principal o aprofundamento do que foi trabalhado durante as aulas, visando o trabalho em grupo e a aplicação de conceitos previamente vistos ao decorrer do curso. Conceitos estes, relacionados as estruturas de dados como árvore binárias de busca, árvore balanceadas, árvore B, fila de prioridades, entre outros.

Para realizar tal tarefa, foi escolhida a estrutura *Random Forest* que consiste em um conjunto de árvores de decisão geradas a partir de amostras aleatórias dos dados.

II. Descrição do Problema

O problema abordado é a classificação estatística, na qual é preciso identificar qual é a categoria de uma nova observação a partir de dados em que a categoria já é conhecida.

III. Solução

A ideia consiste em implementar o algoritmo *Random Forest*, criado por Tin Kam Ho em 1995 [1], cujo objetivo é conseguir por meio da amostra, prever com precisão, o interesse dos indivíduos no produto, sendo classificados em zero, aqueles que não tem interesse, e um, aqueles que tem interesse.

Visando solucionar tal problema, foram implementados na linguagem C dois algoritmos. O algoritmo de árvores de decisão, servindo de base para a implementação, e a *Random Forest* em si, com um número fixo de árvores de decisão com altura aleatória e treinadas com amostras aleatórias.

A implementação das árvores de decisão consiste em construir as melhores divisões dos dados para cada nó da árvore, que divide os dados em duas partes baseado no Coeficiente de Gini, este retorna um valor entre zero e um, onde zero significa a divisão perfeita e igual dos dados e um a divisão imperfeita.

Sendo assim, testou-se todas as divisões dos dados para obter o menor Coeficiente de Gini possível. Na sequência é calculado o valor esperado encontrando a classe que mais aparece naquela divisão dos dados e é registrado este valor no nó. O número de divisões dependerá da altura, que é um valor fixo dado como parâmetro par a construção da árvore de decisão, parando assim a subdivisão da árvore quando este limite é atingido. Sabendo disso, a *Random Forest* é um conjunto de árvores de decisão treinadas com uma amostra aleatória dos dados e com alturas aleatórias.

Ainda para testar a eficácia do algoritmo, os dados foram separados em 80% para treinar a *Random Forest* e 20% para testa a taxa de previsões corretas.

IV. Testes

O algoritmo foi testado com dados que representam o interesse dos indivíduos em um determinado produto, sendo classificados em zero, aqueles que não tem interesse, e um, aqueles que tem interesse. Baseado na idade e renda anual dos indivíduos, a *Random Forest* é construída para prever o interesse dos indivíduos no produto.

Foi possível para esses dados testar 30 *Random Forest*s treinadas com diferentes divisões de treino e teste dos dados que resultou em 93.37% de acerto em média e desvio padrão de 2.68%. Pela distribuição de t de Student com confiança de 95% temos que a média μ real das taxas de acerto da *Random Forest* está no intervalo $92.71\% \leq \mu \leq 94,71\%$. Assim, a taxa de acerto é bastante significativa o que pode ser muito útil para prever se determinado cliente vai ter interesse no produto ou para saber qual é o público-alvo.

Outros dados utilizados são de diagnóstico de câncer de mama em que diversos atributos de células são usados para identificar a classe dessas células [2]. Se a célula for benigna a classe é igual a 2 e se for maligna é 4. Com base nisso, separando os dados em treino e teste o algoritmo conseguiu uma taxa de acerto de aproximadamente 94% para os dados de teste, sendo igualmente uma predição significativa.

Agradecimentos

Agradecemos o professor Rodrigo Minetto pelas ótimas aulas e pela oportunidade de aprofundar nossos conhecimentos

neste trabalho. Nossos sinceros agradecimentos também vão para o monitor vigente da disciplina, o excelentíssimo Frank Bloemer por extensivos esclarecimentos.

REFERÊNCIAS

[1] T. K. Ho, "Random Decision Forests," *Proceedings of the 3rd International Conference on Document Analysis and Recognition*, pp. 278–282, August 1995.

[2] UCI, "Machine learning repository," 2021. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29