Projeto 2 - Leaf Database

Disciplina: Fundamentos de Sistemas Inteligentes 2019/1

Matheus Schmitz Oliveira 15/0018371
dept. Ciência da Computação (CIC)
Universidade de Brasília (UnB)
Brasília, Brazil

Pedro Aurélio Coelho de Almeida 14/0158103
dept. Ciência da Computação (CIC)
Universidade de Brasília (UnB)
Brasília, Brazil

***Abstract*—O seguinte projeto tem como objetivo explorar um conjunto de características extraídas de imagens de folhas pertencentes a diferentes tipos de plantas. A partir de diversas combinações de parâmetros no algoritmo de classificação supervisionado 'florestas randômicas' obteve-se acurácia final no conjunto de teste de 79%, para a melhor configuração.**

***Index Terms*—aprendizado de máquina, classificação, florestas 'randômicas', reconhecimento de padrões**

I. Introdução

Reconhecimento de padrões (RP) é uma tarefa extremamente importante. Em geral, consiste em mostrar ao computador de forma supervisionada, um conjunto de características ('features', em inglês) de forma que ele possa classificar um novo registro desconhecido. Dentre as diversas aplicações de RP, pode-se mencionar classificar dígitos manualmente escritos, informar qual a espécie de uma planta e avaliar se uma imagem médica contém ou não algum sinal de doença.

Modelos matemáticos como, por exemplo, árvores de decisão, análise de discriminante linear ('linear discriminant analysis (LDA)', em inglês) e regressão logística são frequentemente utilizados para problemas que envolvem RP.

Uma das formas de avaliar a qualidade do modelo é dividir os dados classificados em uma base de treino e outra de teste, calculando a taxa de acurácia nesta última. Se o conjunto de dados for muito pequeno, pode-se utilizar abordagens de validação cruzada como, por exemplo, o 'K-Fold', que consiste em dividir os dados em K grupos de tamanhos aproximadamente iguais e usar um deles como teste e outro como treino, repetindo o processo K vezes. A acurácia média considerando as K repetições é considerada como a acurácia média do modelo.

Obter classificações com uma alta capacidade de generalização é extremamente desejável ao se trabalhar com RP. Esse objetivo pode ser entendido como a capacidade de conseguir manter a taxa de acurácia elevada para dados diferentes daqueles usados no treinamento do classificador. Por um lado, se o modelo for mais complexo que a quantidade de dados disponíveis, ocorrerá o sobre-ajuste dos dados de treino, reduzindo a capacidade de generalização do problema. Por outro lado, se o modelo não tiver complexidade suficiente, baixos níveis de acurácia serão obtidos nas bases de treino e teste.

No presente trabalho, árvores de decisão serão construídas para classificar plantas de acordo com a espécie utilizando a base de dados 'Leaf' [1].

A. Base de Dados 'Leaf'

A base de dados é composta por 340 imagens distribuídas entre 40 classes de espécies de folhas diferentes. As observações foram coletadas a partir de uma câmera com resolução de 720x920 pixels, onde posteriormente foram extraídas 14 características para cada imagem. Além dessas, os autores também disponibilizaram a espécie e número do espécime [2].

Embora exista uma média de 10 observações por classe, o dataset é desbalanceado, no qual a classe 'Taxus bacatta' tem 5 exemplos, sendo a menor, e a classe 'Acer palmatum' tem 16 exemplos, sendo a maior.

B. Florestas Randômicas

O algoritmo de florestas 'randômicas' (*'Random Forest'* em inglês) é um dos principais métodos de aprendizado supervisionado. Pode ser utilizado tanto para regressão quanto para classificação. Em uma abordagem superficial, consiste na criação de conjunto de N árvores de decisão que classificarão os exemplos a partir da classe que aparecer de forma mais frequente entre as N árvores.

Os algoritmos baseados em árvores lidam bem com grandes quantidades de dados, sendo preferíveis em relação aos modelos clássicos lineares quando a relação entre as variáveis independentes for não-linear. O modelo de 'florestas randômicas' pode ser preferível a outros modelos clássicos não-lineares quando não há ortogonalidade entre os dados de entrada.

II. Metodologia

Um classificador do tipo de árvores de decisão, conhecido como florestas 'randômicas' (*'Random Forest'* em inglês), foi utilizado, tendo, como entradas, as características e, como saída, a espécie de cada folha.

Antes de construir o modelo, uma análise exploratória dos dados foi realizada com o objetivo de determinar se as características medidas eram ortogonais entre si. Após essa etapa, vários modelos de florestas 'randômicas' foram testadas com diferentes alturas máximas permitidas, a fim de tentar reduzir o sobre-ajuste e escolher a árvore menos complexa

possível que gerasse a classificação mais genérica. Além da profundidade máxima, o critério de impureza, número de árvores de decisão e número de entradas consideradas em cada divisão foram variados.

A. Análise Exploratória dos Dados

A etapa de análise exploratória dos dados tem como objetivo proporcionar uma melhor visualização das relações presentes entre os dados a serem trabalhados. A partir disso, é possível determinar se há a necessidade de utilizar técnicas de transformação aos mesmos.

A Figura 1 fornece a matriz de correlação entre as entradas da base de dados, onde as variáveis entre *Lobedness* e *Eccentricity* representam características de forma e as demais representam características de textura [2].

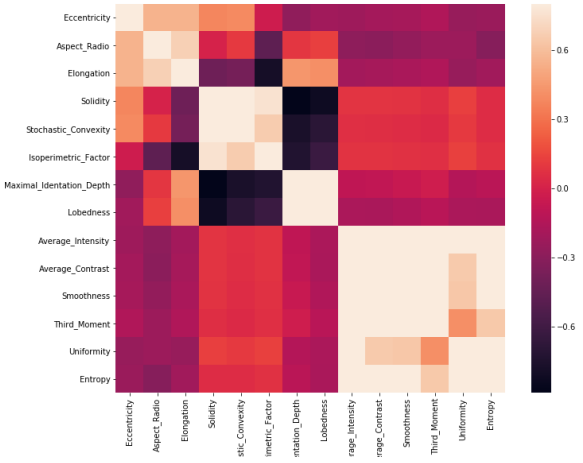

Figure 1. Matriz de Correlação entre as características da base de dados.

Percebe-se na Figura 1 uma forte correlação positiva no quarto quadrante da matriz (região inferior à direita), onde se encontram as variáveis relacionadas às médias de intensidade e contraste, suavidade e entropia das imagens coletadas. Tais variáveis são referentes à textura das observações coletadas.

O segundo quadrante (região superior à esquerda) possui as variáveis relacionadas à forma das observações coletadas. É interessante destacar uma forte correlação negativa presente em alguns pares dessas.

Os atributos de forma e textura pouco se correlacionam entre si, o que pode ser visto nos quadrantes restantes da matriz.

Nenhuma técnica de extração de características ou transformação dos dados originais adicional foi realizada, visto que poderia aumentar o grau de incerteza e complexidade na construção das árvores de decisão. Além disso, não foram aplicadas técnicas de normalização ou escala.

Uma divisão do conjunto de dados em 80% para treinamento e 20% para teste foi adotada. Uma vez que a base de dados completa é pequena, será realizada uma validação cruzada do tipo *'K-Fold'* com K=10.

B. Construção do Modelo

A fim de testar múltiplas combinações para o modelo de florestas randômicas utilizou-se da técnica '*Grid Search Cross-Validation*', passando valores distintos de parâmetros para o classificador, como mostrado na Tabela I.

Table I
PARÂMETROS AVALIADOS PELO '*Grid Search Cross-Validation*'

Parâmetro	Valores
n_estimators	[10, 15, 25, 50, 100, 150, 200, 250, 300]
criterion	['gini', 'entropy']
max_depth	[**None**, 2, 5, 8, 10, 15, 18, 20, 24, 30]
max_features	['sqrt', 'log2']

Os parâmetros presentes na Tabela I possuem o seguinte significado:

- **n_estimators**: Quantidade de árvores
- **criterion**: Métrica para medir a impureza das variáveis ao construir a árvore
- **max_depth**: Profundidade máxima permitida na árvore
- **max_features**: Número máximo de características utilizadas na busca da melhor divisão da árvore

O objetivo de tal escolha foi o de realizar diferentes variações para o algoritmo e analisar os resultados de acurácia para cada combinação. Assim, ao final do treinamento, foi possível escolher os melhores parâmetros.

C. Linguagem de Programação e Bibliotecas

A linguagem de programação Python ™ foi utilizada devido a sua versatilidade e facilidade em se desenvolver programas.

A biblioteca *Sklearn* fornece uma série de funções para o tratamento dos dados, construção de modelos de aprendizado de máquina e avaliação de resultados. Foi amplamente utilizada em decorrência dessas características. [3]

Para a visualização dos dados, as bibliotecas *Seaborn* [4] e *Matplotlib* foram adotadas. A importação dos dados e manipulações matemáticas nos mesmos foram realizadas por meio das bibliotecas *Pandas* e *Numpy*.

III. RESULTADOS

A partir dos parâmetros fornecidos, aplicou-se o 'Grid Search Cross Validation', obtendo-se, dentre outros, o melhor conjunto de parâmetros para o classificador, o tempo decorrido para encontrar o melhor conjunto de parâmetros e a melhor acurácia na base de teste.

As Tabelas II e III mostram, respectivamente, os 3 melhores parâmetros ordenados de acordo com a acurácia média nas bases de teste e os tempos médios para o ajuste nos dados e predição dos dados, além das acurácias médias para as bases de treino. As Tabelas IV e V mostram as mesmas variáveis para os 3 piores parâmetros.

Table II
3 MELHORES PARÂMETROS DE ACORDO COM A ACURÁCIA MÉDIA NA BASE DE TESTES.

Criterion	Max depth	Max features	N estimators
gini	18	log2	200
gini	24	log2	200
gini	18	sqrt	200

Table III
MÉDIAS DE DESEMPENHO DOS PARÂMETROS EXIBIDOS NA TABELA II

Fit time	Score time	Test accuracy	Train accuracy
0.4401	0.0297	0.7904	1.0000
0.4183	0.0237	0.7904	1.0000
0.4171	0.0259	0.7904	1.0000

Table IV
3 PIORES PARÂMETROS DE ACORDO COM A ACURÁCIA MÉDIA NA BASE DE TESTES.

Criterion	Max depth	Max features	N estimators
entropy	2	log2	10
entropy	2	sqrt	10
gini	2	sqrt	10

Table V
MÉDIAS DE DESEMPENHO DOS PARÂMETROS EXIBIDOS NA TABELA IV

Fit time	Score time	Test accuracy	Train accuracy
0.0258	0.0023	0.3419	0.4212
0.0307	0.0023	0.3419	0.4212
0.0425	0.0074	0.3419	0.4640

Uma vez que os parâmetros descritos na Tabela II possuem o mesmo valor médio de acurácia, quaisquer um deles poderia ser usado como configuração para o classificador. A Tabela VI mostra o conjunto de parâmetros escolhidos.

Table VI
PARÂMETROS ESCOLHIDOS PARA O CLASSIFICADOR

Parâmetro	Valores
n_estimators	200
criterion	**'gini'**
max_depth	18
max_features	**'sqrt'**

Sabendo que a profundidade máxima de cada árvore é um dos fatores essenciais com relação à qualidade da classificação, um gráfico da acurácia média na fase de treino ('Mean Train Accuracy') x profundidade máxima ('Max depth') foi gerado para as bases de teste e treino criadas durante a fase de validação cruzada. Ele é exibido na Figura 2.

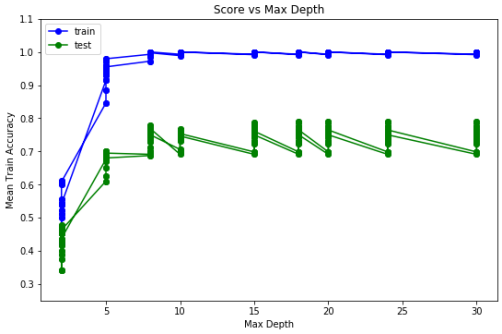

Figure 2. Acurácia média na fase de treino x profundidade máxima de cada árvore durante a fase de validação cruzada.

Também foi construído um gráfico relacionando acurácia média na fase de treino ('Mean Train Accuracy') x Número de árvores na floresta ('Num Trees'), parâmetro importante na construção das florestas 'randômicas', exibido na Figura 3.

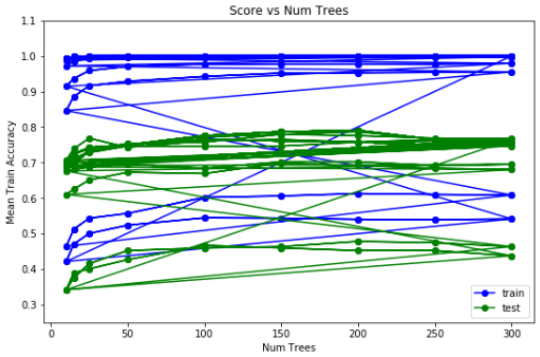

Figure 3. Acurácia média na fase de treino x número de árvores durante a fase de validação cruzada.

Utilizando os valores obtidos na Tabela VI, obteve-se uma acurácia de 100% para a base de treino e 77,94% para a base de teste.

A partir os melhores parâmetros obtidos na Tabela VI, gerou-se o gráfico correspondente à primeira árvore de decisão. Uma vez que a árvore é muito larga, escolheu-se uma fração desta, mostrada na Figura 4.

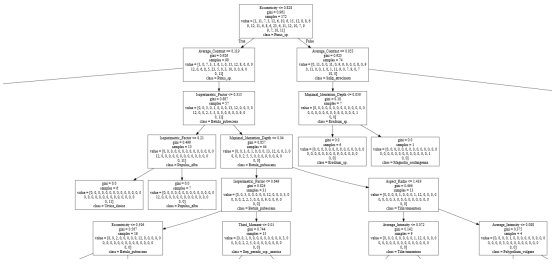

Figure 4. Fragmento da primeira árvore de decisão

IV. DISCUSSÃO

Problemas de classificação e RP podem ser desafiadores, especialmente quando existem muitos dados não ortogonais

entre si ou há poucos dados disponíveis para treino. Considerando o problema de reconhecimento de espécies de plantas dado 14 características de imagem [2], pode-se perceber pela Figura 1 que existe uma forte correlação (positiva ou negativa) entre diversos padrões utilizados. Além disso, a quantidade de dados relacionada à cada classe é relativamente baixa.

O uso do classificador conhecido como florestas 'randômicas' pode contornar a falta de ortogonalidade entre as entradas, além de auxiliar na redução de sobre-ajuste ao reduzir o número de parâmetros disponíveis em cada ramificação da árvore de decisão, criando árvores menos correlacionadas entre si [5].

Uma das causas da perda da capacidade de generalização desses classificadores se dá quando cada nó folha possui um pequeno número de exemplos [6], como o ocorrido no fragmento de uma das árvores mostrado na Figura 4. Nela, é possível observar que há nós folha com somente 1 amostra. Além disso, mesmo testando várias alturas diferentes para cada árvore, as árvores resultantes são em geral muito amplas, o que constitui como outro indício de alta incerteza nos dados, uma vez que várias ramificações estão sendo necessárias para classificar os dados corretamente.

Como o esperado, modelos muito simples não permitem que o classificador se ajuste adequadamente aos padrões presentes nos dados (Tabelas IV e V), enquanto que modelos mais complexos podem se adequar a todos os dados de entrada e, eventualmente, gerar sobre-ajuste (Tabelas II e III e Figura 2).

Devido à redução de parâmetros em cada ramificação e ao fato de que o algoritmo de florestas 'randômicas' escolhe a classificação mais frequente entre todas as árvores, é comum que várias árvores de decisão sejam utilizadas pelo melhor classificador, como mostrado pelo parâmetro 'n_estimators' da Tabela VI. Vale ressaltar que o número de árvores utilizadas não têm muita influência na acurácia (Figura 3) contanto que esse número seja maior que um limiar (aproximadamente 50 para o problema dado, como mostrado na Figura 3).

Há uma tendência de que a acurácia obtida na fase de treino tenha uma pequena queda ao se considerar a base de teste, sendo que ambas devem estar aproximadamente na mesma faixa. Isso foi observado, uma vez que a acurácia média de teste utilizando a validação cruzada na fase de treino foi de 79,04% (Tabela III), enquanto que se observou uma acurácia de 77,94% na fase de teste.

Uma das formas de melhorar os níveis de acurácia obtidos pode ser através da coleta de mais amostras de plantas para cada classe, permitindo assim que os nós-folha contenham mais observações. Além disso, é importante construir uma base com a quantidade de amostras equilibradas entre as classes. Outra maneira consiste na engenharia de características ('feature engineering', em inglês) com o intuito de se obter variáveis que contenham menores níveis de impureza por cada nó, possibilitando assim a criação de uma árvore com maior capacidade de generalização.

V. Conclusão

Em suma, a tarefa de classificar plantas a partir de características previamente calculadas de suas folhas [2] mostrou resultados satisfatórios considerando as limitações geradas pela falta de ortogonalidade de muitas variáveis de entrada (Figura 1) e a pequena quantidade de amostras para treinamento e teste.

A partir do exposto, é possível traçar novas abordagens para aumentar a acurácia e capacidade de generalização do modelo em trabalhos futuros. Dentre essas, pode-se mencionar a aquisição de novos dados para o treinamento e a extração de outras características ortogonais entre si que reduzam a impureza entre cada ramificação.

A comparação com outros modelos de classificação também deve ser realizada a fim de avaliar os resultados finais obtidos.

References

[1] Repositório contendo os dados de folhas, suas espécies e características calculadas: https://archive.ics.uci.edu/ml/machine-learning-databases/00288/ (Acessado em 13/05/2019)

[2] Pedro F. B. Silva, "Development of a System for Automatic Plant Species Recognition", 2013

[3] "Sklearn Library", https://scikit-learn.org/stable/index.html (Acessado em 13/05/2019)

[4] "Seaborn Library", http://seaborn.pydata.org/index.html (Acessado em 13/05/2019)

[5] Gareth James, "An Introduction to Statistical Learning", 2017.

[6] T. M. Mitchell, "Machine Learning", March 1997.