Classificação de Espécies de Plantas Utilizando Algoritmo de Floresta Aleatória

Khalil Carsten do Nascimento
Departamento de Ciência da Computação
Universidade de Brasília
Brasília, Brasil
khalilcarsten@gmail.com

Renato Avelar Nobre
Departamento de Ciência da Computação
Universidade de Brasília
Brasília, Brasil
rekanobre@gmail.com

***Index Terms*—florestas aleatórias, validação cruzada, algoritmos de classificação, aprendizagem de máquina, inteligência artificial, processamento de imagens**

I. Introdução

Problemas de classificação são inerentes ao nosso dia a dia, onde muitas vezes é preciso distinguir informações em diferentes aspectos e categorias. Considere o problema de classificar de que espécie de planta é uma determinada folha. Sabe-se que há milhares de espécies de plantas no mundo, pode se tornar um trabalho árduo para o ser humano tentar classificar uma folha, baseada em suas características, dentro destas espécies. Para realizar tal tarefa de classificação, é proposta uma abordagem na qual modelos de aprendizagem de máquina recebem imagens tratadas de folhas e classificam de acordo com seu conhecimento prévio das características das folhas.

Utilizou-se, neste trabalho, o algoritmos de classificação supervisionada de Florestas Aleatórias, para tratar o problema de reconhecimento de espécies de plantes com base nas características de suas folhas. O algoritmo utilizado é , uma abordagem não-paramétrica para classificações e regressões.

Escolheu-se para utilização um banco de imagens de folhas, desenvolvido por Pedro F. B. Silva, André R. S. Marçal, e Rubim Almeida da Silva [1]. O banco contém 40 espécies de plantas diferentes, cada qual com 15 características sobre os aspectos de suas folhas.

Este relatório é organizado do seguinte modo. A Seção II desenvolve conhecimentos teóricos importantes para o entendimento dos algoritmos de aprendizagem e da proposta implementada; a Seção III explica como a plataforma foi desenvolvida e a estrutura do banco de imagens; a Seção IV descreve as metodologias de tratamento de imagens implementadas na plataforma; a Seção V mostra os resultados experimentais da bateria de testes realizadas; por ultimo, a Seção V finaliza o relatório, sumarizando os resultados obtidos e sugerindo trabalhos futuros.

[1]Disponível em: https://archive.ics.uci.edu/ml/machine-learning-databases/00288/

II. Algoritmos de Classificação

A. Algoritmo de Florestas Aleatórias

Florestas Aleatórias é um algoritmo supervisionado capaz de ser utilizado tanto para classificação quanto para regressão. Como o nome sugere ele utiliza uma combinação de árvores de decisão para gerar de maneira aleatória vários modelos. Essa aleatoriedade resulta em modelos de maior acurácia e evita *overfittings*.

Uma árvore de decisão se utiliza de ganhos de informação, a partir da entropia de cada característica, e também o Gini index para percorrer os nós de características. Árvores de decisão muito profunda podem resultar *overfittings*, o que se torna mais raro de ocorrer em florestas aleatórias, pois gera árvores menores a partir de subconjuntos das características.

Resumidamente uma floresta randômica combina diversas árvores de decisão geradas a partir de subconjuntos aleatórios das características do problema. Na Figura 1 podemos ver duas árvores de decisão onde a decisão de classificação final advém do resultado das somas das probabilidades de ambas as árvores.

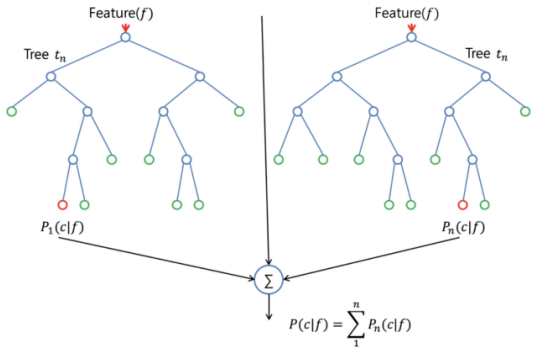

Figura 1. Diagrama de uma floresta aleatória de duas árvores

III. Plataforma de Classificação e Banco de Imagens

Para realizar a plataforma de classificação utilizou-se um conjunto de ferramentas e técnicas.

A. Especificações do Sistema

A estrutura do sistema foi desenvolvida para evitar acoplamento entre os algoritmos de aprendizagem, os algoritmos de tratamento de imagens, e os algoritmos de suporte. O sistema adaptador é responsável por conciliar o ambiente para respeitar as interfaces requeridas pelo usuário ao realizar entradas pela linha de comando.

Todo o sistema foi construído utilizando a linguagem de programação *Python 3* que é a linguagem usada em milhares de aplicações de negócio no mundo, incluindo diversos sistemas de larga escala [2]. Para o auxilio no desenvolvimento das funções de classificação, foi utilizada uma biblioteca de aprendizagem de máquina denominada *scikit-learn*. A biblioteca, acessível para todos, é um conjunto de métodos simples e eficientes para mineração e analise de dados [3]. Também foi utilizada a biblioteca *matplotlib* para criar gráficos e a matriz de confusão [4]. E a biblioteca *pandas* para fazer a leitura dos dados [5].

B. Banco de Imagens

Escolheu-se para utilização um banco de imagens de folhas, desenvolvido por Pedro F. B. Silva, André R. S. Marçal, e Rubim Almeida da Silva. O banco contém 40 espécies de plantas diferentes, cada qual com 15 características sobre os aspectos de suas folhas.

Dentro das 40 espécies de folhas do banco, 30 são de espécies de folhas consideradas simples, e outras 10 de folhas complexas. Para o propósito deste trabalho, foram utilizados somente dados de folhas simples, visto que os valores da características das folhas complexas não foram fornecidas pelo banco de imagens.

Cada amostra de folhas foi fotografada sobre um fundo colorido usando um dispositivo Apple iPad 2. As imagens gravadas têm uma resolução de 720 × 920 pixels e 24 bits. O banco fornece também, versões binárias para folhas simples.

As imagens de folha possuem suas características divididas em formato e textura, e são propriamente descritas no relatório do banco de imagens. As características de formato utilizam objetos de interesse da imagem, como diâmetro, distancia entre pontos de interesse, área e contorno. As seguintes foram levadas em consideração:

- Excentricidade
- Relação de Aspecto
- Alongamento
- Solidez
- Convexidade Estocástica

[2]Disponível em: https://www.python.org
[3]Disponível em: http://scikit-learn.org
[4]Disponível em: https://matplotlib.org
[5]Disponível em: http://pandas.pydata.org

Tabela I
Tabela contendo o nome cientifico das folhas utilizadas e a quantidade de ocorrências da mesma no banco, bem como seu código de identificação para referencia nos gráficos deste relatório.

Classe	Nome	Ocorrências
1	Quercus suber	12
2	Salix atrocinera	10
3	Populus nigra	10
4	Alnus sp.	8
5	Quercus robur	12
6	Crataegus monogyna	8
7	Ilex aquifolium	10
8	Nerium oleander	11
9	Betula pubescens	14
10	Tilia tomentosa	13
11	Acer palmatum	16
12	Celtis sp.	12
13	Corylus avellana	13
14	Castanea sativa	12
15	Populus alba	10
16	Primula vulgaris	12
17	Erodium sp.	11
18	Bougainvillea sp.	13
19	Arisarum vulgare	9
20	Euonymus japonicus	12
21	Ilex perado ssp. azorica	11
22	Magnolia soulangeana	12
23	Buxus sempervirens	12
24	Urtica dioica	12
25	Podocarpus sp.	11
26	Acca sellowiana	11
27	Hydrangea sp.	11
28	Pseudosasa japonica	11
29	Magnolia grandiflora	11
30	Geranium sp.	10

- Fator Isoperimétrico
- Profundidade Máxima de Indentação
- Lobedness

Os atributos de textura são baseados em propriedades estatísticas de histogramas de intensidade de transformações em escala de cinza das imagens originais. As seguintes características foram levadas em consideração:

- Intensidade Média
- Contraste Médio
- Suavidade
- Terceiro momento
- Uniformidade
- Entropia

IV. Proposta e Desenvolvimento

A. Avaliação de Modelo com Validação Cruzada

Visto o algoritmo de Florestas aleatórias e os bancos de imagens apresentado propomos uma experimentação em cima da base de dados utilizando florestas aleatórias. As predições serão feitas utilizando validação cruzada o que permite que o modelo seja testado com diversas permutações de teste melhorando a confiança dos resultados finais das predições . Para cada teste fizemos executamos o a predição 3 vezes.

As validações cruzadas foram executas dividindo a base de dados em 10 partes e executadas com umas quantidade de

Tabela II
RESULTADO DE VALIDAÇÃO CRUZADA COM 14 CARACTERÍSTICAS, NÚMERO DE ÁRVORES 10 COM TRÊS REPETIÇÕES DO EXPERIMENTO

Splits	1	2	3	4	5	6	7	8	9	10	Media
1	0.69	0.68	0.74	0.77	0.81	0.77	0.6	0.66	0.77	0.75	0.67
2	0.67	0.61	0.62	0.68	0.80	0.70	0.7	0.73	0.71	0.62	0.621
3	0.71	0.70	0.62	0.68	0.64	0.74	0.8	0.66	0.75	0.66	0.624
											63.83%

10, 100, 1000 e 10000 árvores na floresta aleatória com a intenção de checar o quanto que a quantidade de árvores altera na precisão.

B. Filtragem de parâmetros

Além disso entre as 14 características escolhemos as 5 com maior importância e refizemos os modelos com a intenção de otimizar o algoritmo diminuindo a quantidade de árvores sem perder precisão nas classificações. Após o treino do modelo somos capazes de identificar quais as características mais importantes mostrado na Figura 3. As características selecionadas foram, *Solidity, Aspect Ration, Elongation, Eccentricity, Isoperimetric Factor*.

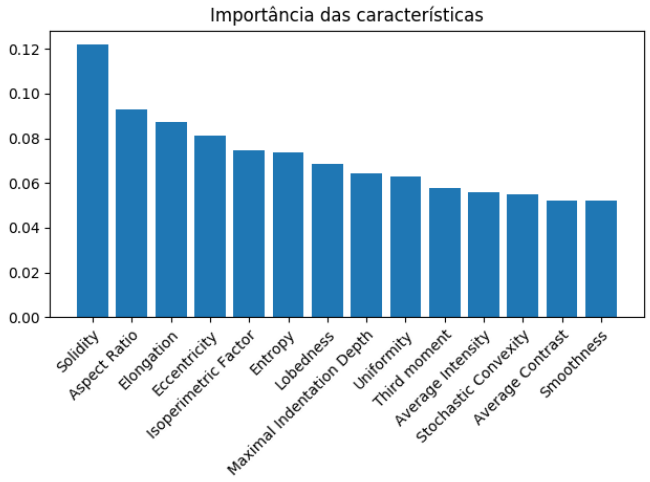

Figura 2. Gráfico de barras em ordem descrescente de importância das características

V. RESULTADOS EXPERIMENTAIS

Os melhores resultados em relação ao modelo com as 14 características foi com 1000 árvores geradas alcançando uma média de 79% de precisão. Supomos que isso ocorre devido a quantidade de características e a quantidade de observações de treino. Não necessariamente o número ideal para o modelo é 1000 árvores, o número ideal pode estar entre 1000 e 100 já que com 10000 árvores a precisão diminuiu.

Já em relação a tentativa de manter somente as 5 características mais importantes, os resultados não melhoraram. Supomos que isso se deve a proximidade das importâncias das características como um todo, como demonstrado na Figura 3.

Tabela III
RESULTADO DE VALIDAÇÃO CRUZADA COM 14 CARACTERÍSTICAS, NÚMERO DE ÁRVORES 100 COM TRÊS REPETIÇÕES DO EXPERIMENTO

Splits	1	2	3	4	5	6	7	8	9	10	Media
1	0.76	0.68	0.74	0.71	0.83	0.83	0.73	0.73	0.85	0.74	0.76
2	0.80	0.79	0.74	0.81	0.80	0.83	0.73	0.8	0.82	0.70	0.71
3	0.78	0.74	0.84	0.83	0.90	0.76	0.7	0.75	0.82	0.74	0.723
											73.1%

Tabela IV
RESULTADO DE VALIDAÇÃO CRUZADA COM 14 CARACTERÍSTICAS, NÚMERO DE ÁRVORES 1000 COM TRÊS REPETIÇÕES DO EXPERIMENTO

Splits	1	2	3	4	5	6	7	8	9	10	Media
1	0.78	0.75	0.74	0.81	0.83	0.83	0.76	0.76	0.89	0.77	0.792
2	0.78	0.72	0.74	0.78	0.83	0.93	0.76	0.76	0.85	0.74	0.789
3	0.80	0.75	0.74	0.81	0.83	0.87	0.73	0.73	0.89	0.77	0.792
											79,1%

Tabela V
RESULTADO DE VALIDAÇÃO CRUZADA COM 14 CARACTERÍSTICAS, NÚMERO DE ÁRVORES 10000 COM TRÊS REPETIÇÕES DO EXPERIMENTO

Splits	1	2	3	4	5	6	7	8	9	10	Media
1	0.82	0.72	0.74	0.81	0.83	0.90	0.76	0.76	0.89	0.77	0.80
2	0.82	0.72	0.74	0.81	0.83	0.90	0.73	0.76	0.89	0.74	0.794
3	0.80	0.72	0.74	0.81	0.83	0.87	0.76	0.8	0.89	0.77	0.727
											77,36%

Tabela VI
RESULTADO DE VALIDAÇÃO CRUZADA COM 5 CARACTERÍSTICAS MAIS IMPORTANTES, NÚMERO DE ÁRVORES 10 COM TRÊS REPETIÇÕES DO EXPERIMENTO

Splits	1	2	3	4	5	6	7	8	9	10	Media
1	0.67	0.52	0.65	0.56	0.77	0.74	0.53	0.53	0.71	0.66	0.634
2	0.69	0.61	0.57	0.46	0.58	0.64	0.63	0.53	0.64	0.55	0.59
3	0.57	0.65	0.45	0.53	0.61	0.51	0.46	0.6	0.75	0.66	0.58
											60.13%

Tabela VII
RESULTADO DE VALIDAÇÃO CRUZADA COM 5 CARACTERÍSTICAS MAIS IMPORTANTES, NÚMERO DE ÁRVORES 100 COM TRÊS REPETIÇÕES DO EXPERIMENTO

Splits	1	2	3	4	5	6	7	8	9	10	Media
1	0.71	0.61	0.57	0.65	0.70	0.67	0.60	0.63	0.71	0.66	0.651
2	0.71	0.65	0.65	0.59	0.70	0.70	0.63	0.60	0.71	0.59	0.653
3	0.69	0.61	0.62	0.59	0.74	0.70	0.63	0.60	0.78	0.74	0.670
											65.8%

Tabela VIII
RESULTADO DE VALIDAÇÃO CRUZADA COM 5 CARACTERÍSTICAS MAIS IMPORTANTES, NÚMERO DE ÁRVORES 1000 COM TRÊS REPETIÇÕES DO EXPERIMENTO

Splits	1	2	3	4	5	6	7	8	9	10	Media
1	0.73	0.63	0.62	0.56	0.70	0.70	0.63	0.63	0.71	0.70	0.661
2	0.73	0.63	0.6	0.56	0.70	0.70	0.63	0.66	0.71	0.66	0.664
3	0.73	0.63	0.62	0.70	0.70	0.63	0.63	0.75	0.70	0.56	0.665
											66.33%

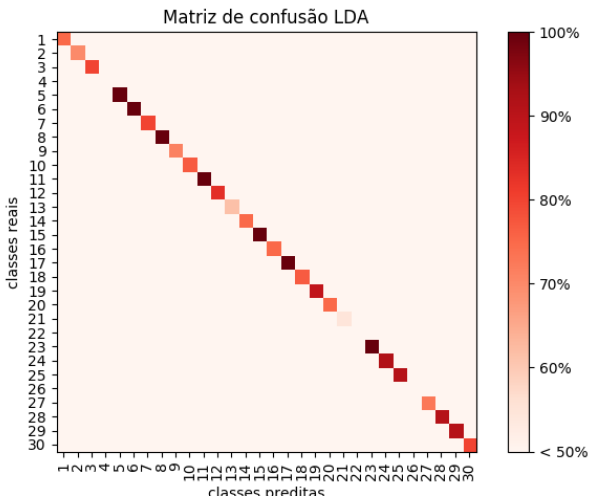

Figura 3. Matriz de confusão para 1000 árvores com as 14 características

Tabela IX
RESULTADO DE VALIDAÇÃO CRUZADA COM 5 CARACTERÍSTICAS MAIS IMPORTANTES, NÚMERO DE ÁRVORES 10000 COM TRÊS REPETIÇÕES DO EXPERIMENTO

Splits	1	2	3	4	5	6	7	8	9	10	Media
1	0.73	0.63	0.62	0.56	0.70	0.70	0.63	0.63	0.71	0.66	0.584
2	0.73	0.65	0.6	0.56	0.70	0.70	0.63	0.63	0.71	0.66	0.603
3	0.71	0.65	0.6	0.56	0.70	0.70	0.63	0.63	0.71	0.66	0.584
											59%

Com base nos problemas citados na Seção IV, e com os métodos implementados no sistema descritos na Seção III foram elaboradas as seguintes perguntas:

- É possível classificar as folhas com uma taxa de acerto por classe maior que 50% utilizando uma validação cruzada com $k = 10$?
- É possível filtras os parâmetros de forma a encontrar os melhores parâmetros para melhorar o resultado da pergunta anterior?
- Reduzir as características para as mais importantes acarreta em uma melhora da precisão

VI. CONCLUSÃO

Este trabalho implementa o algoritmo de floresta aleatória em características de imagens relativas a folhas de árvores, onde temos 30 classes de folhas analisadas. As características foram inseridas no algoritmo e testadas com 10, 100, 1000, 10000 árvores e com características reduzidas para as 5 mais importantes.

As implementações foram testadas por diversos experimentos. Seus resultados foram analisados:

- É possível classificar as folhas com uma taxa de acerto maior que 50% com validação cruzada com k = 10.
- Alterando o número de árvores máximo é possível melhorar a precisão do algoritmo.
- Utilizar somente as características mais importantes não necessariamente funciona já que depende da distancia de importância entre essas características.

REFERÊNCIAS

[1] MLA Kotsiantis, Sotiris B., I. Zaharakis, and P. Pintelas. “Supervised machine learning: A review of classification techniques.” Emerging artificial intelligence applications in computer engineering 160 (2007): 3-24.