Estudo comparativo de modelos de aprendizagem de máquina para o problema de classificação de grãos de pistache

Resumo

Os grãos de pistache são importantes para comercialização, possuindo preços variados, sabores e valores nutricionais diferentes que atendem a um vasto mercado. A eficiência do processo industrial pós-colheita dos grãos de pistache é muito importante para manter o seu valor econômico, ocupando uma posição importante na economia agrícola. Para fornecer essa eficiência, são necessários novos métodos e técnicas de separação e classificação de pistaches. Neste trabalho, fazemos um estudo comparativo de modelos de aprendizagem de máquina para o problema de classificação de grãos de pistache, utilizando conjuntos de dados de imagens e de características.

1. Introdução

O pistache é um produto agrícola nativo do Oriente Médio e Ásia Central e é bastante conhecido por seus nutrientes (proteínas, fibras, vitamina B entre outros) e uma variedade de minerais que auxiliam em uma dieta. Também possui alto valor energético. Por exemplo, 100g fornece cerca de 560 kcal (quilocalorias). Atualmente, os maiores produtores de pistache do mundo são Irã, Estados Unidos, Turquia e Síria [1].

Os grãos de pistache são importantes para comercialização, possuindo preços variados, sabores e valores nutricionais diferentes que atendem a um vasto mercado. A eficiência do processo industrial pós-colheita dos grãos de pistache é muito importante para manter o seu valor econômico, ocupando uma posição importante na economia agrícola. Para fornecer essa eficiência, são necessários novos métodos e técnicas de separação e classificação dos grãos de pistache.

Machine Learning, ou aprendizado de máquina, em português, é uma técnica em que computadores são capazes de aprender com base nas respostas esperadas, correlacionando dados diferentes, como por exemplo, imagens e números. Um detalhe que causa confusão em muitas pessoas é que os termos aprendizado de máquina e inteligência artificial são sinônimos, mas não. Na verdade, inteligência artificial é um conceito amplo que inclui o aprendizado de máquina como uma de suas características. Como pontuado por [2], a máquina toma decisões e realiza predições baseadas nos dados disponíveis. Além do que, dependendo da disponibilidade dos tipos e categorias dos dados de treino, uma das técnicas pode ser selecionada: Aprendizado Supervisionado, Aprendizado Nao-Supervisionado, Aprendizado Semi-Supervisionado e Aprendizado por Reforço.

O principal objetivo deste trabalho é avaliar o desempenho dos modelos de aprendizagem de máquina *K-Nearest Neighbors* (KNN), Regressão Logística e Rede Neural Convolucional (CNN) para o problema de classificação de grãos de pistaches. Além disso, comparar os resultados do algoritmo de agrupamento *k-means* com os resultados dos modelos de classificação anteriores, tentando corroborar os resultados obtidos pelos algoritmos de classificação.

2. Fundamentação teórica

Neste trabalho é feito um estudo comparativo de modelos de aprendizagem de máquina sobre um problema de classificação. Classificação é um problema de aprendizagem de máquina cujo objetivo é classificar um vetor de entrada X, atribuindo-o a uma das K classes de um espaço amostral Ck onde k = 1, ..., K [3]. Já o agrupamento, ou *clustering*, é um processo que busca separar dados em partições com um certo grau de similaridade entre seus membros [4].

Neste trabalho, foram utilizados três conjuntos de dados para avaliar o desempenho dos modelos de aprendizagem de máquina propostos. Um conjunto de imagens de grãos de pistache com 2.148 amostras e outros dois conjuntos de dados com 16 e 28 características extraídas do conjunto de imagens. Das 2.148 amostras, 1.232 pertencem à classe *Kirmizi* e 916 pertencem à classe *Siirt*. O conjunto de dados com 16 características possui 12 atributos morfológicos (área, perímetro, comprimento do eixo maior entre outros) e 4 outros atributos. O conjunto de dados com 28 características possui as mesmas 16 características do conjunto anterior com mais 12 atributos de cor [5, 6].

Em [5], um modelo *K-Nearest Neighbors* (KNN) enviesado de alta performance é aplicado ao problema de classificação de grãos de pistache, utilizando o mesmo conjunto de dados deste trabalho. Já [6] utiliza uma abordagem diferente, fazendo uso de múltiplas Redes Neurais Convolucionais (CNN) pré-treinadas com *AlexNet*, VGG16 e VGG19. Finalmente, [1] utiliza outro conjunto de dados com Redes Neurais Artificiais (ANNs), Processos Gaussianos (GP) e *Random Forest*.

3. Metodologia

Os modelos de aprendizagem de máquina foram escolhidos, majoritariamente, pela sua facilidade de implementação, popularidade, suporte e documentação. Os três primeiros modelos foram utilizados para fazer a classificação dos grãos de pistache. Já o último modelo foi utilizado para tentar agrupar os dados originais para comparar com os resultados da classificação dos modelos anteriores. As métricas utilizadas na avaliação dos modelos de classificação foram acurácia, precisão, revocação, *F1-score* e matriz de confusão. Já para o modelo de agrupamento, a métrica utilizada foi o índice de *Davies-Bouldin* e o erro de reconstrução.

Todos os modelos foram implementados na linguagem de programação *Python* e usando as bibliotecas *Scikit-learn* e *Keras*. Os gráficos foram gerados utilizando as bibliotecas *Matplotlib* e *Seaborn*. Para simplificar o processo de desenvolvimento, treino e teste dos modelos, foi utilizado o *Jupyter Notebooks*. Para cada um dos modelos de aprendizagem, geralmente, um notebook foi utilizado.

4. Experimentos

Nesta seção, são detalhados cada um dos modelos de aprendizagem propostos, além da configuração e outros detalhes relevantes para a execução dos experimentos. No final de cada subseção são discutidos os resultados.

K-Nearest Neighbors

K-Nearest Neighbors (KNN) é um dos modelos de aprendizagem de máquina mais populares para problemas de classificação [7].

Neste trabalho, foi utilizado o classificador KNN com os atributos padrões, exceto o número de vizinhos *(n neighbors)* e a métrica de distância *(metric)*, que foram selecionados via *Grid Search*. O processo a seguir foi executado em ambos os conjuntos de dados com 16 e 28 características, com o objetivo de comparar os resultados e ressaltar as diferenças de cada um.

Primeiramente, foi realizado o *One-Hot Encoding* na coluna Classe, para facilitar e otimizar a execução dos algoritmos. Os dados foram normalizados usando *MinMaxScaler*, para que todo ponto do espaço amostral ficasse dentro do intervalo fechado [0,1], ideal para o cálculo de distâncias quando as grandezas dos atributos são bem diferentes [8]. O conjunto de dados foi dividido em duas parcelas: 80% (1.718 amostras) para busca de hiperparâmetros e 20% (430 registros) para avaliação. Essa divisão levou em consideração a distribuição original dos dados. Devido ao número de características, foi aplicado também o método PCA para redução de dimensionalidade [9]. O número de componentes do PCA também foi selecionado na etapa de *Grid Search*.

O método *Grid Search* com validação cruzada em 10 *folds* foi executado sobre a primeira parcela dos dados para maximizar a métrica acurácia. Como previamente mencionado, foram variados os hiperparâmetros:

1. Número de vizinhos (de 3 até 45, passo 2);
2. Métrica de distância (Euclidiana, Minkowski e Manhattan) e
3. Número de componentes do PCA.

Os hiperparâmetros selecionados via *Grid Search* para ambos os conjuntos de dados são mostrados na tabela 1. Percebemos que cada conjunto de dados apresentou resultados bem diferentes. No conjunto de dados 1, foi selecionada a distância Euclidiana, enquanto no conjunto de dados 2 a distância de Manhattan foi selecionada. O número de vizinhos não mudou muito em cada, sendo 11 no conjunto de dados 1 e 15 no conjunto de dados 2. Já o número de componentes novamente variou bastante. Percebemos que a inclusão de novas características no conjunto de dados 2 melhorou a acurácia ao ponto de não precisar realizar redução de dimensionalidade. No conjunto de dados 1 foram 5 componentes, enquanto no conjunto de dados 2, 28, ou seja, todas as características.

Com os hiperparâmetros definidos, o procedimento foi executado com a segunda parcela de dados, previamente separada para avaliação. As métricas encontradas nessa etapa final estão na tabela 2. Na avaliação, o conjunto de dados 2 apresentou melhores resultados em todas as métricas comparado ao conjunto de dados 1, especialmente revocação e F1-score.

Adicionalmente, foram gerados dois gráficos para os resultados (?). O primeiro, na figura 2, representa a matriz de confusão das predições feitas na avaliação. Vemos que, novamente, o conjunto de dados de 28 características apresentou um desempenho melhor, com mais acertos em positivos e negativos. Já nas curvas ROC (figura 3), o mesmo acontece, com o conjunto de dados 2 apresentando taxa AUC de 0.89, comparado ao conjunto de dados 1 com 0.87.

Regressão Logística

A regressão logística é um método estatístico poderoso que pode modelar resultados binomiais com uma ou mais variáveis explicativas. Ele mede a relação entre uma variável dependente categórica e uma ou mais variáveis independentes estimando probabilidades usando uma função logística (ou seja, uma distribuição logística cumulativa).

Neste experimento, utilizamos regressão logística nos dois conjuntos de dados com 16 e 28 características. Primeiramente, foi realizado o *One-Hot Encoding* na coluna Classe, para facilitar e otimizar a execução dos algoritmos. Separamos os conjuntos de dados em conjunto de treino e conjunto de teste (respectivamente 80% e 20%) e, então, realizamos a predição das classes para cada amostra de ambos os conjuntos de dados. Os resultados foram bastante semelhantes em ambos os conjuntos de dados, como podemos verificar na figura 4.

Os resultados das métricas para os dois conjuntos de dados foram bem parecidos, sendo o conjunto de dados que possui 28 características, o que possui melhor precisão, revocação e *F1-score*, como mostrados nas tabelas 3 e 4. Por último, podemos notar que a área sob a curva para o conjunto de dados com 28 características é levemente menor que para o conjunto de dados com 16 características.

Rede Neural Convolucional

Rede Neural Convolucional (Convolutional Neural Network, CNN) tem sua arquitetura análoga ao padrão de conectividades por neurônios. Seu principal diferencial é a capacidade de, dada uma imagem como entrada, atribuir pesos a características daquela imagem, o que a torna capaz de diferenciar as classes passadas. Por conta de produzir suas próprias características para predição, o pré-processamento dessas arquiteturas se torna bem menor em comparação com outros métodos. Destaca-se também o seu bom desempenho com imagens.

Neste experimento, as imagens foram separadas em conjuntos de treino, teste e validação, mantendo a proporção das classes, como pode ser observado na figura 6. Os conjuntos de treino, teste e validação ficaram com 1.739, 215 e 194 imagens, respectivamente. Os valores de cada classe podem ser observados na Tabela 5.

Após a separação do conjunto de dados, as imagens passaram pela etapa de pré-processamento. Todas foram carregadas e normalizadas para *pixels* de valores RGB entre 0 e 1. Para evitar *overfitting* durante o treinamento, o conjunto de treino passou pela etapa de aumento de dados, tendo variações nas seguintes caracterısticas:

1. Rotações aleatórias de zero a 180 graus;
2. *Zoom* aleatório nas imagens de valores até 0.2;
3. Translado verticalmente ou horizontalmente e
4. Espelhamento na vertical ou horizontal.

A arquitetura da rede desenvolvida consta com entradas de imagens (224, 224, 3) com três
camadas convolucionais de funções de ativação ReLU, três camadas *MaxPool2D* e um *Dropout*. No final, encontram-se uma camada *Flatten* e duas camadas densas, uma com 128 neurônios e a outra com o número de classes. Essa arquitetura pode ser vista em mais detalhes na figura 7.

O modelo foi treinado em 400 épocas, otimizando a função de entropia cruzada de categoria utilizando *Adam*. Tendo como entrada o conjunto de treinamento pós aumento de dados e o de validação. Na figura 8, vemos as curvas de aprendizado observando acurácia e função de perda.

Após todas as épocas, o modelo foi salvo com o menor valor de perda. Como resultado, o modelo teve uma acurácia de 89% para os dados de teste. Também foi gerado uma avaliação de classe a classe das métricas de precisão, revocação e *F1-score* como pode ser visto na tabela 6. Para a classe *Kirmizi* foi obtido 91%, 90% e 90% de precisão, revocação e *F1-score,* respectivamente. *Siirt* obteve 87%, 88% e 88% para as mesmas métricas.

Para uma taxa ótima buscando o equilıbrio entre as taxas de falsos positivos e verdadeiros positivos pode ser escolhido o limiar de 0,1, como pode-se observar na figura 9.

Visualização dos dados originais com PCA

A Análise de Componentes Principais (PCA) é um método clássico usado, geralmente, para redução de dimensionalidade de conjuntos de dados com alta dimensão. A redução é feita, “espremendo” a informação ou variância dos dados originais em componentes principais. Dessa forma, conjuntos de dados de alta dimensão podem ser visualizados em duas ou três dimensões.

Usamos PCA nos conjuntos de dados com 16 e 28 características para reduzir a dimensionalidade dos dados originais e podermos visualizá-los. Para isso, analisamos a variância explicada acumulada. Para o conjunto com 16 características, mais de 90% da variância explicada se encontra nas três primeiras componentes. Já para o conjunto com 28 características, nas três primeiras componentes se encontra pouco mais de 65% da variância explicada. O resultado da projeção das três componentes principais dos dados originais pode ser visto no gráfico abaixo.

K-mean

O *k-means* é um algoritmo de agrupamento cujo objetivo é particionar um conjunto de dados em *k* grupos tal que cada elemento do conjunto pertença a um grupo com média

mais próxima a um centro de um grupo. Isso é feito minimizando as variações dentro dos grupos utilizando a distância Euclidiana ao quadrado. O resultado do algoritmo é o particionamento do espaço dos dados em células de Voronoi.

O k-means é muito sensível à escala dos dados, portanto, os conjuntos de dados com 16 e 28 características foram padronizados, ou seja, as características foram ajustadas para a mesma escala. Para determinar o melhor método de inicialização dos centróides, executamos o algoritmo com os valores *k-means++* e random e avaliamos os resultados usando a soma das distâncias ao quadrado. Os resultados mostraram que não há uma melhora significativa na utilização de um parâmetro ou de outro pelo algoritmo. Assim, optamos por usar o parâmetro *k-means++* como método de inicialização dos centros do grupos.

Para determinar o número ótimos de grupos, avaliamos o índice de Davies-Bouldin e o erro de reconstrução. Para os conjuntos de dados com 16 e 28 características, o algoritmo foi executado, variando-se o número de grupos de 2 à 16 e 2 à 28, respectivamente. O método de inicialização dos centros dos grupos usado foi *k-means++* e número de vezes que o algoritmo executará com sementes diferentes igual a 20. Os resultados das execuções para ambos os conjuntos de dados podem ser vistos no gráfico abaixo.

Para o conjunto de dados com 16 características, o número ótimo de grupos foi 11 com o valor do índice de *Davies-Bouldin* igual à 1,13184. Já para o conjunto de dados com 28 características, o número ótimo de grupos foi 18 com o valor do índice de Davies-Bouldin igual à 1,601393.

5. Conclusão

Neste trabalho, foram analisados os resultados de quatro modelos de aprendizagem de máquina sobre conjuntos de dados de grãos de pistache. Devido às peculiaridades de cada modelo, vários experimentos foram realizados e, em seguida, uma análise dos resultados foi feita, utilizando métricas específicas para cada modelo. Para os modelos de classificação, a Rede Neural Convolucional (CNN) apresentou os melhores resultados para as métricas de acurácia, precisão, revocação e *F1-score*. Para o modelo de agrupamento, o número ótimo de grupos para os dois conjuntos de dados foi bastante diferente do número real de classes, mostrando que o *k-means* é muito sensível a dados com muitos atributos.

Como trabalho futuro, propomos a realização de experimentos adicionais com redes neurais, especificamente redes convolucionais, que apresentaram os melhores resultados neste trabalho. Propomos também, a realização de experimentos com outros tipos de grãos de pistache, a fim de verificar se os resultados são tão bons quantos os encontrads neste trabalho.