Árvore de decisão e Naive Bayes: análises e experimentos

Luiz Felipe de Souza Silva[1]

[1]Departamento de Engenharia da Computação (DCA) – Universidade Federal do Rio Grande do Norte (UFRN)

lf06092004@gmail.com

Resumo. *Este relatório apresenta e compara resultados do treinamento de modelos de classificação para raças de gatos e cachorros (Egyptian Mau, Ragdoll, Pomeranian e Shiba Inu) utilizando os algoritmos Árvore de Decisão e Naive Bayes. Os modelos foram treinados com bases de dados contendo características extraídas pelos métodos de HOG e CNNs pré-treinadas (VGG16 e VGG 9). Nas 6 melhores bases de classificação obtidas com o k-NN, foram aplicadas a redução de dimensionalidade com o PCA, com objetivo de analisar se essa técnica favorece o desempenho dessas bases.*

1. Informações Gerais.

O objetivo da atividade foi realizar a classificação de raças de gatos e cachorros a partir de imagens, utilizando os classificadores Árvore de Decisão e Naive Bayes. As características das imagens foram extraídas por meio dos métodos HOG (Histogram of Oriented Gradients) e CNNs pré-treinadas (VGG16 e VGG19). Nas 6 melhores bases de acurácia obtidas com o classificador k-NN, foram aplicadas a redução de dimensionalidade com o PCA, com objetivo de analisar se o PCA é favorável para essas bases de dados. Os resultados obtidos usando o modelo de Árvore de Decisão e Naive Bayes foram utilizados para responder questões chaves da atividade.

1. Como se comportam os modelos de treinamento com as diferentes extrações HOG e CNN, e nas bases com aplicação do PCA?

2. Quais os melhores parâmetros e métricas para o treinamento desses modelos em suas bases de dados? e seu comportamento?

3. Quais vantagens e limitações de cada modelo para esse tipo de problema multiclasse?

Alguns dados podem ser vistos acessando o relatório anterior ou acessando o repositório do github deixado nos direcionamentos a seguir:

Relatório: Relatório - Checkpoint 1

Repositório GitHub: Classificador de Raças

E os resultados dos testes entre Árvore de Decisão e Naive Bayes podem ser vistos nessa tabela: Tabela_AD_NB_Checkpoint2

1.1 HOG ou CNN para AD e NB?

Para os algoritmos utilizados, a extração de características com CNN demonstrou ser a mais eficaz na classificação multiclasse das raças de gatos e cachorros. Tanto a Árvore de Decisão quanto o Naive Bayes conseguiram discriminar muito bem as classes ao

utilizar as características extraídas pelas CNNs. No entanto, o modelo Naive Bayes apresentou um desempenho superior, alcançando uma acurácia significativamente mais elevada em comparação à Árvore de Decisão. Enquanto a Árvore de Decisão obteve uma acurácia média de 90,66%, o Naive Bayes alcançou expressivos 95,15%, mas ainda assim reforça a CNN como o melhor extrator de características para este problema de classificação.

O HOG por sua vez, apresentou desempenho inferior de 50% em ambas os algoritmos testados, o que demonstra que o histograma orientado à gradiente tem dificuldade de conseguir extrair características mais complexas, como diferentes ângulos e textura, mesmo as imagens padronizadas na escala de cinza antes de sua aplicação como extrator. No entanto, o Naive demonstrou uma melhora média significativa de 10 pontos percentuais com relação aos treinamentos com modelos anteriores (k-NN e AD). Bases de HOG treinadas com Naive Bayes obtiveram acurácia média de 45,70%. Esse resultado reforça uma característica importante do Naive Bayes: seu desempenho costuma ser melhor quando o número de características é razoável ou elevado.

1.2 Melhores parâmetros para os modelos AD e NB nessa classificação

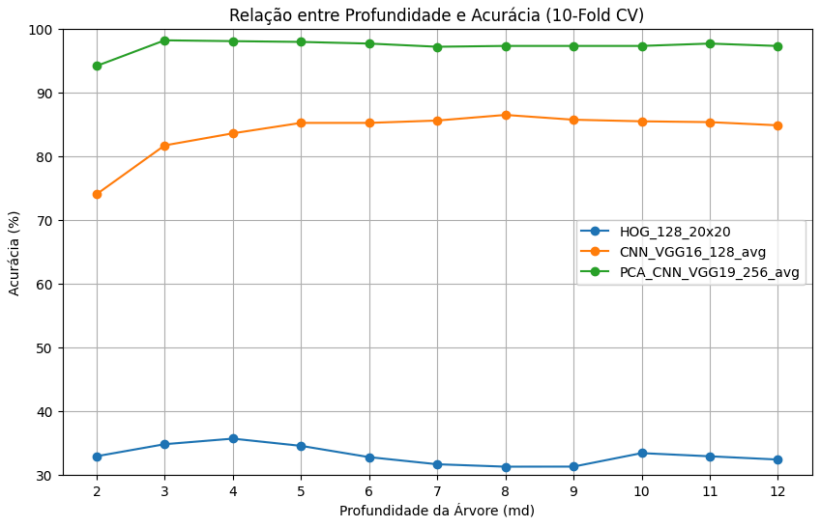
Para o algoritmo de Árvore de Decisão, o parâmetro de ajuste e análise considerado foi a profundidade da árvore, ou depth model (md). No entanto, realizar uma análise exclusivamente baseada na média geral de cada profundidade pode conduzir a interpretações distorcidas ou simplificadas dos dados. Isso ocorre porque a média tende a ser fortemente influenciada pelos valores extremos ou pela discrepância de desempenho entre diferentes tipos de extração de características, como observado entre os descritores HOG e CNN.

Diante disso, fez-se necessária uma análise mais criteriosa da profundidade ideal para a Árvore de Decisão. Assim, foi elaborada uma avaliação gráfica, por meio de um gráfico de linha, que relaciona a variação da acurácia em função da profundidade da árvore. Para essa análise, foram selecionadas três bases representativas:

- ➔ A melhor base com descritor HOG, a HOG_128_20x20;
- ➔ A melhor base extraída com CNN, a CNN_VGG16_256_avg;
- ➔ E a melhor base após a aplicação da PCA, a PCA_CNN_VGG19_256_avg.

A tabela nos leva à pensar que a melhor profundidade de árvore é *md* = 3 apenas analisando a média geral das bases em cada profundidade, no entanto, analisando individualmente é possível notar que *md*=3 é falso ponto de máximo, justamente pela discrepância dessa profundidade no HOG, dessa forma, é possível notar que a melhor profundidade para essa classificação em questão de acurácia e discriminalidade das classes está entre *md* = 7 e *md* = 9. São nesses valores onde ocorre estabilidade do algoritmo e após isso uma queda de acurácia nesses valores, além disso analisando à tabela com as bases de CNN é possível notar que o md = 9 demonstrou uma acurácia

média alta de suas bases, mostrando melhor profundidade e consistente com seu baixo desvio padrão com relação as demais.



Para o Naive Bayes a diferença de parâmetro foi a forma da probabilidade: Multinomial, Gaussiana e Complemento. Para esses algoritmos, o modelo Gaussiano apresentou um desempenho ligeiramente maior com relação aos outros parâmetros, mas isso não foi algo definitivo para a acurácia, com média geral de 78,54% e analisando à tabela com as CNNs alcançou 95,95%. O que realmente causou diferença no modelo foi o pipeline introduzido para a base de dados, em alguns casos as bases de dados sem introdução do PCA tem acurácia melhor nas probabilidades Multinomial e Complemento. No entanto, no modelo Gaussiano, CNN + PCA apresentaram melhor desempenho de acurácia na classificação. O que indica as distribuições das features de cada classe se aproximam de uma distribuição normal, favorecendo o Naive Bayes Gaussiano.

1.3 Limitações e Vantagens de cada modelo.

O Naive Bayes conseguiu aumentar a acurácia das bases com extração de HOG, mas mesmo assim, não é o ideal, pode-se presumir que para esse tipo de problema a extração realizada pelo HOG não seja o ideal. Ambos os modelos se saíram muito bem com bases extraídas com o CNN e principalmente aliadas a essas extrações, o PCA. No entanto, quando falamos em interpretabilidade talvez o algoritmo AD seja o mais adequado em questão de visualização e explicação.