Primeiro Trabalho de Inteligência Artificial

Filipe Gomes Arante de Souza

Maio 2023

Resumo

Este artigo consiste num estudo de comparação experimental entre determinadas técnicas de aprendizado de máquina aplicadas a um problema de classificação. O objetivo deste estudo é ver como diferentes classificadores se comportam perante o mesmo conjunto de dados, e compreender como a escolha do modelo pode impactar os resultados finais. Para visualizar tais diferenças, foram realizadas etapas de treino, validação e teste em cada um dos classificadores a fim de computar o valor de suas acurácias, analisar par a par de classificadores através de testes de hipótese e gerar gráficos boxplot para detectar possíveis outliers no processo de validação cruzada.

Palavras-chave: Aprendizado de máquina; Classificação; Comparação Experimental;

1 Introdução

O campo do aprendizado de máquina tem ganhado grande destaque recentemente devido à sua capacidade de processar grandes volumes de dados e extrair informações relevantes para tomada de decisões. No entanto, a eficácia do aprendizado depende muito da escolha adequada do classificador. Tendo em vista este contexto, será apresentada uma comparação experimental entre os seguintes tipos de classificadores:

ZeroR, Bagging, AdaBoost, RandomForest e Heterogeneous Pooling.

Essa comparação será feita utilizando um conjunto de dados previamente selecionado, visando imitar uma situação real perante um problema de classificação. Desse modo, o objetivo é detectar se existem ou não diferenças significativas entre as técnicas mencionadas acima utilizando algumas métricas a fim de fixar os conteúdos ministrados em sala de aula.

2 Base de Dados

A base de dados do trabalho foi obtida a partir de um projeto de pesquisa que usa informações de imagens de lâmpadas e luminárias de iluminação pública para atualização de cadastro das concessionárias de energia.

2.1 Descrição do Domínio

A partir do processamento das imagens dos dispositivos, são extraídas as características que compõem a base da dados, que consistem em:

- 10 descritores de Fourier (descrevem o contorno do objeto);
- 7 descritores de Hu (definem um conjunto de momentos invariantes da imagem);
- 6 descritores de Haralick (descrevem as texturas da imagem);

2.2 Definição das Classes e das Características

As classes das luminárias são determinadas por:

- Tipo: Mercury Vapor (MV), High Pressure Sodium (HPS) e Metal Halide (MH);
- Potência: 70W, 100W, 125W, 150W, 250W e 400W;

Desse modo, temos o seguinte conjunto de classes:

sodio70, sodio100, sodio150, sodio250, sodio400, metalica150, metalica250, metalica400, metalica400 e mercurio125

As características utilizadas para este trabalho estão definidas de acordo com número de matrícula do aluno, que neste caso é 2020100625. Conforme especificação, **serão selecionados para o dataset os descritores de Fourier e de Hu** para final de matrícula 5.

2.3 Número de Instâncias

A base completa possui 297 exemplos de imagens de lâmpadas e luminárias. Segue abaixo distribuição por classe:

Classe	Tipo	Potência (W)	Quantidade	(%)
sodio70	HPS	70	30	10.1%
sodio100	HPS	100	32	10.8%
sodio150	HPS	150	35	11.8%
sodio250	HPS	250	33	11.1%
sodio400	HPS	400	37	12.5%
metalica150	MH	150	23	7.7%
metalica250	MH	250	49	16.5%
metalica400	MH	400	37	12.5%
mercurio125	MV	125	21	7.1%
Total			297	100%

Table 1: Distribuição dos exemplos por classe

3 O método Heterogeneous Pooling

Um dos classificadores utilizados neste trabalho é o Heterogeneous Pooling **(HP)**, que foi implementado. Ele é um combinado de classificadores, que utiliza como classificadores base:

- Árvore de Decisão **(DT)**;
- Naive Bayes Gaussiano **(NB)**;
- K Vizinhos Mais Próximos **(KNN)**;

O HP possui um único hiperparâmetro: a quantidade de cada um dos classificadores citados acima que fará parte do combinado, que foi denominado **n_samples**.
Por exemplo, se **n_samples = 5**, haverão **5 DT's**, **5 NB's** e **5 KNN's** no combinado.

O critério para classificar uma determinada instância será **votação majoritária**, em outras palavras, a classe mais escolhida dentre os classificadores do combinado. Em caso de empate, a classe escolhida deve ser a de maior frequência na base de treino original dentre as que empataram na votação.

4 Descrição dos Experimentos Realizados e seus Resultados

Para cada um dos classificadores **(ZeroR, Bagging, AdaBoost, Random Forest e HP)** foram realizadas as seguintes etapas:

1. Treino e teste com 3 rodadas de validação cruzada estratificada de 10 folds para os classificadores sem hiperparâmetros (Neste caso, somente o **ZeroR**). Cada conjunto de treino teve seus dados normalizados com z-score;

2. Treino, validação e teste com 3 rodadas de ciclos aninhados de validação e teste, com o ciclo interno de validação contendo 4 folds e o externo de teste com 10 folds para os classificadores com hiperparâmetros. Cada conjunto de treino teve seus dados normalizados com z-score.

   O gridsearch do ciclo interno deve considerar os seguintes valores de hiperparâmetros para cada técnica de aprendizado:

   - **Bagging:** $n\_estimators = [3, 9, 15, 21]$
   - **AdaBoost:** $n\_estimators = [3, 9, 15, 21]$
   - **Random Forest:** $n\_estimators = [3, 9, 15, 21]$
   - **Heterogeneous Pooling:** $n\_estimators = [1, 3, 5, 7]$

Após execução deste processo, foram obtidas as seguintes informações dos modelos de aprendizado:

4.1 Estatísticas referentes a acurácia

Para cada método foram computadas a média, desvio padrão e intervalo de confiança a 95% das acurácias dos folds nas etapas de treino, teste e validação. As estratégias que obtiveram melhor média foram **AB** e **RF**, entretanto foram as que apresentaram maior desvio padrão.

Método	Média	Desvio Padrão	Limite Inferior	Limite Superior
ZR	0.165057	0.010883	0.161163	0.168952
BA	0.513180	0.102518	0.476495	0.549865
AB	0.266092	0.028426	0.255920	0.276264
RF	0.509923	0.093142	0.476593	0.543253
HP	0.461494	0.081464	0.432343	0.490645

Table 2: Média, desvio padrão e intervalo de confiança a 95% das acurácias nos folds para cada método de aprendizado.

Segue abaixo boxplot das acurácias para melhor visualização de sua distribuição. Note que alguns outliers foram detectados, especialmente no método AdaBoost.

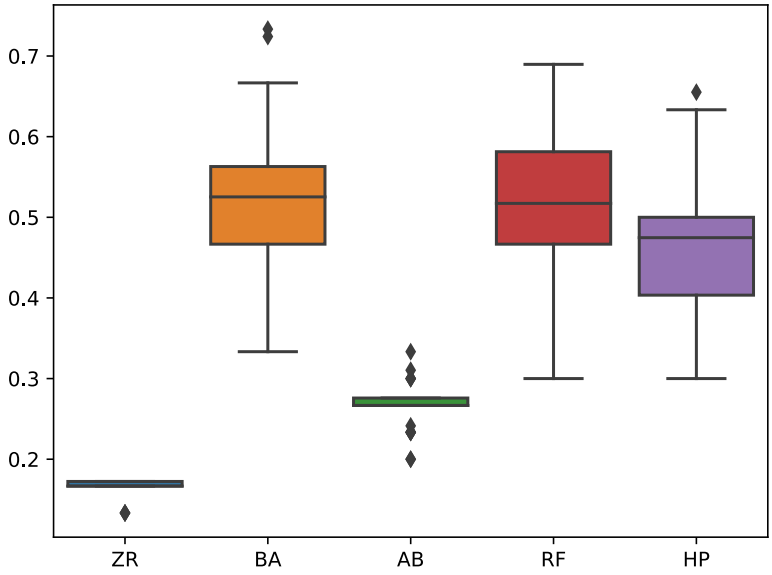

Figure 1: Boxplot das acurácias nos folds em cada classificador.

4.2 Testes de Hipótese (Wilcoxon e T Pareado)

Seguem abaixo resultados dos testes de hipótese para cada par de classificadores. Na parte triangular inferior encontram-se os valores do teste de Wilcoxon, enquanto na parte triangular superior estão os valores do Teste T Pareado. Os valores da tabela que rejeitaram a hipótese nula para um nível de significância de 95% foram colocados em negrito.
Veja que apenas os métodos AdaBoost e Random Forest não apresentaram diferença estatística significativa, em ambos os testes.

ZR	**0.0**	**0.0**	**0.0**	**0.0**
0.000002	**BA**	**0.0**	0.842035	**0.000252**
0.000001	**0.000002**	**AB**	**0.0**	**0.0**
0.000002	0.913692	**0.000002**	**RF**	**0.004076**
0.000002	**0.001473**	**0.000002**	**0.00434**	**HP**

Table 3: Testes de hipótese para cada par de classificadores.

5 Conclusões

5.1 Análise geral dos resultados

No geral, os classificadores divergiram bastante nos resultados, com exceção do Bagging e do Random Forest. Os que obtiveram melhor desempenho foram justamente estes dois, **BA** e **RF**, com acurácia média de cerca de 50%.

É interessante notar que os combinados do **BA**, **AB** e **RF** foram todos compostos por árvores de decisão, devido a serem utilizados os valores default da Sklearn. Isso justifica o fato do **BA** e **RF** terem desempenhos

praticamente iguais, pois ambos são treinados em paralelo com Bootstrap. Já o desempenho inferior do **AB** em relação aos outros dois é um indicativo de que seu treinamento em série não se adaptou bem ao conjunto de dados utilizado, além de se apresentar mais instável, pois é possível observar diversos outliers em suas acurácias no boxplot.

O Heterogeneous Pooling possui parte de seu combinado formado por árvores de decisão, contudo teve um desempenho de 46%, um pouco abaixo dos melhores resultados obtidos. Portanto, podemos perceber que os métodos de K Vizinhos Mais Próximos e Naive Bayes Gaussiano se adaptaram aos dados um pouco menos em relação às árvores de decisão.

Todos os métodos foram bem superiores ao baseline ZeroR, o que é um bom sinal. Este obteve baixa acurácia média de aproximadamente 16.5%, pois a base de dados era razoavelmente balanceada.

Com relação aos testes de hipótese, o teste de Wilcoxon e T Pareado apresentaram valores bem próximos para os mesmos pares de classificadores. A maioria dos resultados deu 0 ou perto disso, o que indica grande diferença estatística entre os métodos de aprendizado.
O único par de classificadores que não apresentou diferença estatística significativa foi o Bagging e Random Forest, cujo motivo de sua similaridade já foi explicada acima.

Assim sendo, podemos concluir que os melhores classificadores deste estudo foram os combinados de árvores de decisão. Todavia, é possível observar nos bloxpots que a distribuição das acurácias em todos os métodos de aprendizado foi bem baixa, não ultrapassando os 60%. Logo, utilizar apenas os descritores de Fourier e Hu não foram suficientes para determinar a classe da lâmpada com maior grau de confiança.

5.2 Contribuições do Trabalho

Para a tarefa de identificar o material e potência de uma lâmpada, vimos que as árvore de decisão treinadas com Bootstrap se adaptaram melhor. Assim, alguma pesquisa nesta área pode se basear nesta técnica para efetuar seus estudos utilizando mais informações das lâmpadas, como por exemplo, os descritores de Haralick.

5.3 Melhorias e trabalhos futuros

O próximo passo deste trabalho, a fim de melhorar seus resultados no futuro, é incluir mais tipos de classificadores. Por exemplo, os ensembles homogêneos como o Bagging e AdaBoost podem ser utilizados com outros classificadores como base de seu combinado, tais como o KNN e Naive Bayes Gaussiano. Assim, é possível analisar outras técnicas de aprendizado de forma independente. Utilizar Redes Neurais também é uma ótima ideia a ser aplicada.

Um outro ponto interessante a ser melhorado é explorar um conjunto de hiperparâmetros maior para cada classificador. O aumento do tamanho dos ensembles do gridsearch pode possibilitar o encontro de mais padrões fase de treinamento, por exemplo.

Por fim, outras métricas de avaliação podem ser incluídas, tais como matrizes de confusão, precision, recall, dentre outras.

Referências Bibliográficas

Os materiais utilizados para o desenvolvimento do artigo foram os slides e notebooks do professor Flávio Varejão mostrados em sala de aula. Já para o desenvolvimento do código presente no notebook enviado teve como referência a documentação dos classificadores da biblioteca *scikit-learn*.