Frederick Salvador Tavares Prates
Luiz Carlos da Silva Fernandes Junior

Relatório de Machine Learning II – Base Letter

1. Introdução

A área de aprendizado de máquina (Machine Learning) cresceu de maneira exponencial nos últimos anos, fornecendo a capacidade para criação de modelos capazes de aprender as relações entre os dados e com isso realizar classificações e predições. Nesse contexto, este relatório visa apresentar os resultados de vários experimentos com uma série de algoritmos de aprendizado de máquina, com o intuito de comparar seus desempenhos em cenários com e sem a etapa de pré-processamento dos dados.

Para isso, foi utilizada a base de dados produzida por [1], que possui atributos de letras maiúsculas em formato de imagem. As imagens das letras foram avaliadas em 16 atributos numéricos que variam num intervalo de 0 até 15 em seu valor para compor a base.

Em uma primeira etapa, os classificadores foram submetidos ao *Grid-Search* para avaliação dos parâmetros que geravam as classificações com melhor acurácia. Esses melhores parâmetros obtidos foram aproveitados para os modelos serem analisados por meio de 2 métodos de seleção de atributos: Filter e Wrapper.

Nesse sentido, os resultados evidenciam a capacidade do pré-processamento de melhorar em diversos aspectos os resultados dos modelos, impactando diretamente dimensões como acurácia, R2-score e demais métricas de avaliação.

2. Conceitos Básicos

Antes de tudo, aplicamos o pré-processamento por meio da seleção de atributos com base nas abordagens de Filter e Wrapper. Esse pré-processamento foi escolhido em vista da quantidade de atributos. Abaixo são apresentados os métodos e algoritmos utilizados.

2.1 Pré-processamento - Seleção de Atributos

2.1.1 Filter

A seleção por filtro é uma técnica que utiliza métricas estatísticas para avaliar a relevância dos atributos com a variável de interesse. São selecionados aqueles que apresentam maior grau de correlação ou dependência. Além disso, os seletores permitem a seleção de um elemento com base na presença de um atributo sozinho ou em várias correspondências diferentes com o valor do atributo.

Para o relatório, aplicou-se para avaliação o critério da variância. O limiar da variância é uma técnica usada para eliminar atributos que têm pouca ou nenhuma variação nos dados. A justificativa é que atributos com baixa variância não contribuem significativamente para a distinção entre diferentes classes, e, portanto, podem ser

removidos para simplificar o modelo e melhorar seu desempenho. Essa técnica foi direcionada a selecionar quantidades pré-estabelecidas de atributos, sendo estas12 (75% do total de atributos), 8 (50%) e 4 (25%).

2.1.2 Wrapper

O método Wrapper avalia a performance dos modelos sobre os todos possíveis subconjuntos de atributos com o intuito de escolher os melhores destes. É um método guloso que escolhe os melhores atributos com base na pontuação da validação cruzada dos estimadores para formar o melhor subconjunto de características.

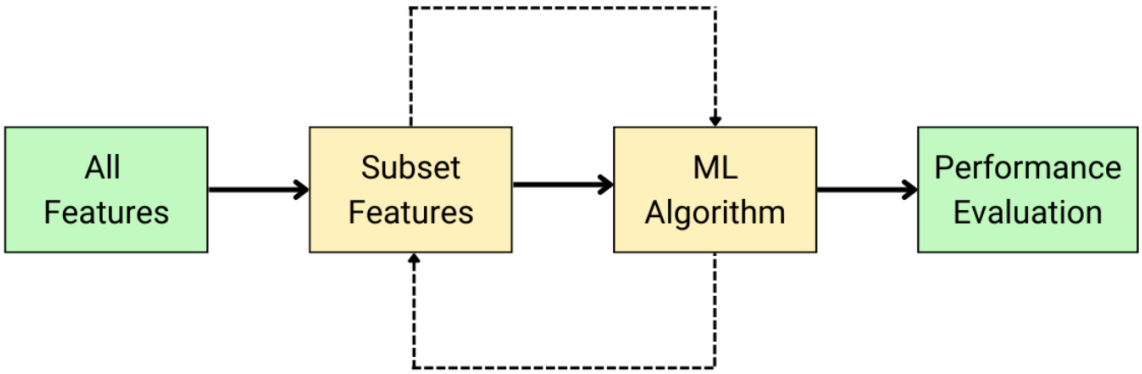

Figura 1. Wrapper - Retirado de [2].

Nos experimentos realizados, foi utilizado o método SequentialFeatureSelector, que possui variações com o processo de adição (*forward selection*) ou remoção (*backward selection*) de atributos para formar o melhor subconjunto que atende ao critério mínimo de quantidade de características fornecido inicialmente. Para os experimentos, foi utilizada a abordagem *forward selection*, pelo fato desta performar mais rapidamente, junto com a quantidade mínima de atributos assumindo os valores de 4, 8 e 12.

2.2 Algoritmos de aprendizagem de máquina

2.2.1 K-Nearest Neighbors (KNN)

É um algoritmo de aprendizado supervisionado que avalia a distância entre os vizinhos mais próximos para fazer a clusterização de elementos em grupos de K vizinhos e partir deles determinar a classificação das classes alvo.

2.2.2 Árvore de Decisão

É um algoritmo que de aprendizado supervisionado que modela um problema de classificação em forma de árvore.

2.2.3 Floresta Randômica

A Floresta Randômica é um conjunto de Árvores de Decisão que funciona combinando para várias árvores com o intuito de melhorar os dados, sendo cada árvore construída por meio de uma amostra aleatória de dados e características.

2.2.4 Regressão Logística

É um algoritmo estatístico que utiliza probabilidade para prever a classe de uma instância.

2.2.5 Naive Bayes

Também é um classificador probabilístico como a Regressão Logística, mas que avalia a probabilidade de classificação com base na distribuição do conjunto.

2.2.6 Perceptron e Multi-Layer Perceptron (MLP)

Utilizam o princípio das redes neurais. Um perceptron segue o modelo de um neurônio onde as entradas são os atributos a serem avaliados juntamente com um bias de ponderação. Esses atributos passam pela função de ativação e então a saída é a classificação. O MLP utiliza várias camadas de perceptrons para alcançar modelagens mais complexas entre os dados e refinar o resultado.

2.2.7 Support Vector Machine (SVM)

É um algoritmo que classifica os dados encontrando uma linha ou um hiperplano ideal que maximiza a distância entre cada classe em um espaço N-dimensional.

3. Metodologia dos Experimentos

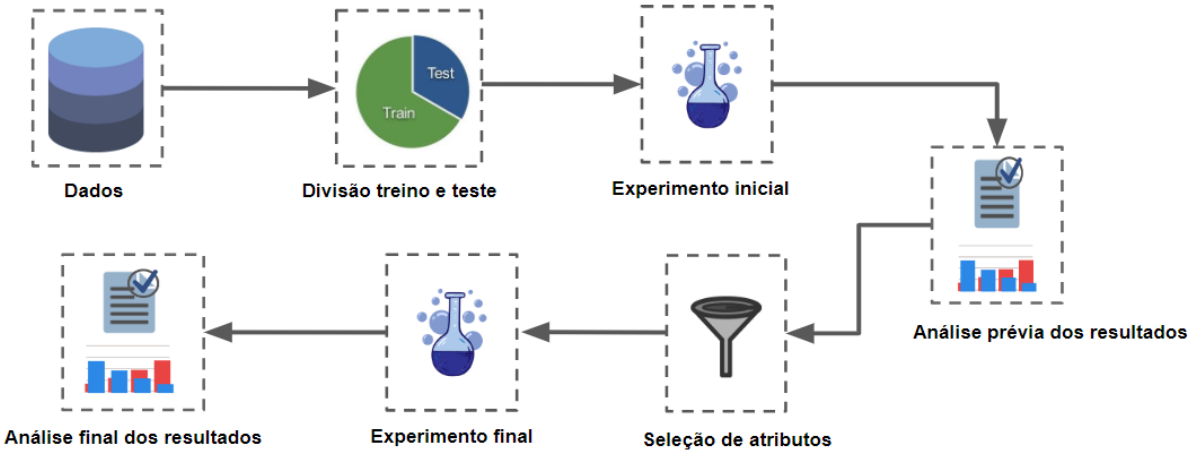

Figura 2. Metodologia.

Foi utilizado o Jupyter Notebook para analisar a base de dados, em conjunto com o pacote Scikit-Learn da linguagem Python para gerar os modelos e métricas, assim como o pacote Scipy para ler o dado no formato ARFF em conjunto com a biblioteca Pandas.

A base de dados foi carregada em um dataframe e em seguida foram efetuadas transformações para facilitar a visualização das classes. Após isso, o dataframe foi dividido a partir de suas colunas, separando as colunas de atributos selecionados pelos algoritmos Filter e Wrapper, da coluna de classificação. Em seguida, os dados foram separados e em um conjunto de treinamento e teste, com 80% das linhas destinadas ao treinamento do modelo e 20% para validar o modelo gerado.

A figura abaixo apresenta os modelos avaliados e seus parâmetros. Baseado nos resultados obtidos na análise com o Grid-search em experimentos prévios, os modelos configurados com os parâmetros abaixo obtiveram o melhor desempenho para classificação dentre os parâmetros avaliados para cada modelo.

```
models_functions = [
    KNeighborsClassifier(n_neighbors=1,metric='euclidean'),
    DecisionTreeClassifier(criterion='log_loss',max_depth=100,splitter='best'),
    RandomForestClassifier(criterion='entropy',max_depth=1000,n_estimators=10),
    LogisticRegression(penalty='l2',C=0.5,solver='newton-cg'),
    GaussianNB(),
    MLPClassifier(activation='tanh', hidden_layer_sizes= (16, 26), learning_rate= 'invscaling'),
    SVC(C= 20, decision_function_shape= 'ovo', kernel= 'rbf')
]
```

Figura 3. Modelos utilizados.

Para avaliação da seleção de parâmetros, executaremos os modelos acima variando as colunas selecionadas. As métricas de R2 score, erro médio quadrático e acurácia serão avaliadas para cada iteração.

3.1 Banco de Dados

O banco de dados possui 26 categorias de letras que ficam na coluna “class”. Além da coluna target “class”, há também 16 colunas de features, que foram utilizadas para gerar os modelos classificadores a serem avaliados no relatório.

A base é bem balanceada como pode ser visto no gráfico abaixo:

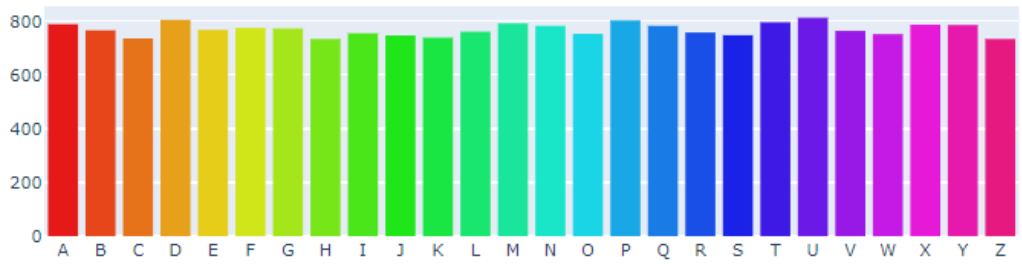

Figura 4. Distribuição dos dados por classe.

Também não há a ocorrência de valores faltantes em nenhuma das 17 colunas.

Variable Name	Type	Description	Units	Missing Values
class	Categorical	capital letter		no
x-box	Integer	horizontal position of box		no
y-box	Integer	vertical position of box		no
width	Integer	width of box		no
high	Integer	height of box		no
onpix	Integer	total # on pixels		no
x-bar	Integer	mean x of on pixels in box		no
y-bar	Integer	mean y of on pixels in box		no
x2bar	Integer	mean x variance		no
y2bar	Integer	mean y variance		no
xybar	Integer	mean x y correlation		no
x2ybr	Integer	mean of x * x * y		no
xy2br	Integer	mean of x * y * y		no
x-ege	Integer	mean edge count left to right		no
xegvy	Integer	correlation of x-ege with y		no
y-ege	Integer	mean edge count bottom to top		no
yegvx	Integer	correlation of y-ege with x		no

A tabela acima descreve cada coluna da base com seu tipo, significado e se possui valores faltantes. Em todas as colunas do tipo inteiro, o range dos valores vai de 0 até 15.

3.2 Métricas

As 3 principais métricas utilizadas foram R2-score, Erro Médio Quadrático e Acurácia.

3.2.1 R2-score

É pronunciado como R ao quadrado e é conhecido como coeficiente de determinação. Ele funciona medindo a variação nas previsões e o resultado da base de teste.

3.2.2 Erro Médio Quadrático

É uma métrica comumente usada para verificar a acurácia de modelos e dá um maior peso aos maiores erros, já que, ao ser calculado, cada erro é elevado ao quadrado individualmente e, após isso, a média desses erros quadráticos é calculada.

3.2.3 Acurácia

É a taxa de acerto do modelo gerado em relação ao todo.

4. Resultados

Na tabela abaixo é possível observar os resultados dos primeiros experimentos sem a etapa de pré-processamento:

	R2 score	MSE	Acurácia
KNN	0.92	4.18	0.95
Árvore de Decisão	0.75	13.96	0.88
Floresta Randômica	0.87	7.20	0.93
Regressão Logística	0.50	27.79	0.77
Gaussian Naive Bayes	0.26	40.95	0.64
MLP	0.71	16.18	0.86
SVM	0.94	3.15	**0.96**

4.1 Resultados do Filter

Filter com a variância de 4.10 (seleção de 12 colunas):

	R2 score	MSE	Acurácia
KNN	0.88	6.90	**0.94**
Árvore de Decisão	0.72	15.82	0.86
Floresta Randômica	0.83	9.28	0.92
Regressão Logística	0.36	35.83	0.72

Gaussian Naive Bayes	0.17	46.16	0.62
MLP	0.62	21.48	0.83
SVM	0.89	6.25	**0.94**

Colunas selecionadas: 'y-box', 'high', 'onpix', 'x-bar', 'y-bar', 'x2bar', 'y2bar', 'xybar', 'x2ybr', 'xy2br', 'x-ege' e 'y-ege'.

Filter com a variância de 5.25 (seleção de 8 colunas):

	R2 score	MSE	Acurácia
KNN	0.78	12.26	**0.90**
Árvore de Decisão	0.68	18.99	0.85
Floresta Randômica	0.76	13.70	0.89
Regressão Logística	0.25	42.04	0.61
Gaussian Naive Bayes	0.13	48.91	0.56
MLP	0.51	27.04	0.78
SVM	0.76	13.39	0.89

Colunas selecionadas: 'y-box', 'y-bar', 'x2bar', 'y2bar', 'xybar', 'x2ybr', 'x-ege' e 'y-ege'

Filter com a variância de 6.5 (seleção de 4 colunas):

	R2 score	MSE	Acurácia
KNN	-0.07	60.01	0.49
Árvore de Decisão	0.02	54.71	**0.55**
Floresta Randômica	-0.02	54.74	**0.55**
Regressão Logística	-0.53	85.71	0.30
Gaussian Naive Bayes	-0.56	87.04	0.27
MLP	-0.16	64.88	0.47
SVM	-0.03	57.61	0.53

Colunas selecionadas: 'y-box', 'x2bar', 'x2ybr' e 'y-ege'

4.2 Resultados do Wrapper

Nas tabelas abaixo é possível observar os resultados da aplicação do método Wrapper para os modelos selecionados, com a quantidade mínima de atributos sendo 12, 8 e 4, respectivamente.

	R2 score	MSE	Acurácia
KNN	0.93	3.84	**0.96**
Árvore de Decisão	0.76	13.15	0.89
Floresta Randômica	0.86	7.47	0.93
Regressão Logística	0.45	30.36	0.74
Gaussian Naive Bayes	0.29	39.48	0.65
MLP	0.67	18.01	0.84
SVM	0.91	4.48	**0.96**

Colunas selecionadas do melhor algoritmo com 12 atributos: 'width', 'x-bar', 'y-bar', 'x2bar', 'y2bar', 'xybar', 'x2ybr', 'xy2br', 'x-ege', 'xegvy', 'y-ege', 'yegvx'.

	R2 score	MSE	Acurácia
KNN	0.87	6.97	**0.93**
Árvore de Decisão	0.76	13.27	0.87
Floresta Randômica	0.86	7.60	0.92
Regressão Logística	0.34	36.56	0.68
Gaussian Naive Bayes	0.33	37.41	0.65
MLP	0.61	21.40	0.81
SVM	0.85	8.07	0.91

Colunas selecionadas do melhor algoritmo com 8 atributos: 'x2bar','y2bar', 'xybar','x2ybr', 'xy2br', 'x-ege', 'y-ege', 'yegvx'.

	R2 score	MSE	Acurácia
KNN	0.34	36.77	0.67

Árvore de Decisão	0.39	33.87	**0.71**
Floresta Randômica	0.41	32.50	**0.71**
Regressão Logística	-0.08	60.50	0.47
Gaussian Naive Bayes	0.03	53.74	0.49
MLP	0.26	40.86	0.63
SVM	0.38	34.42	0.67

Colunas selecionadas do melhor algoritmo com 4 atributos: 'x2bar','x2ybr', 'x-ege', 'y-ege'.

5. Conclusões

Dois métodos de seleção de parâmetros foram avaliados neste relatório: o método Filter baseado no critério de variância e o Wrapper de seleção sequencial usando a abordagem de adição atributos.

Os resultados do primeiro relatório foram também integrados para servir como base para comparação dos resultados obtidos, já que foi executado com o total de atributos que a base possui.

Em maior parte dos casos avaliados houve decréscimo da acurácia em relação a execução com todas as colunas ao remover atributos, mas a acurácia do método Wrapper com 12 atributos se destacou pela proximidade com o resultado com 16 atributos, e nos casos dos classificadores KNN, Árvore de Decisão, Regressão Logística e Naive Bayes Gaussiano houve melhora na acurácia. Para o SVM e Floresta Randômica, a execução com 12 atributos alcançou resultados equivalentes de acurácia, enquanto o MLP com 12 características foi inferior a execução com 16 atributos.

Essa melhora abre a possibilidade de se avaliar a relação precisão e tempo de processamento, que em alguns casos é muito importante, principalmente para necessidades de retreinamento do modelo.

Como proposta de trabalhos futuros, indicamos a análise da acurácia dos classificadores utilizando a combinação de modelos.

6. Referências

[1] P. W. Frey and D. J. Slate. "Letter Recognition Using Holland-style Adaptive Classifiers". Machine Learning 6(2), 1991.

[2] Gupta, Shashank. “Feature Selection Method in Machine Learning”. **Enjoy Algorithms.** Disponível em <https://www.enjoyalgorithms.com/blog/feature-selection-techniques>. Acesso em: 24 de agosto de 24.