Validação dos Resultados

1

1. Material e Métodos

1.1. Bases de Dados

O *Iris dataset* consiste de um conjunto de dados onde são descritas a altura e a largura das sépalas e pétalas de 4 tipos de flores com um total de 150 exemplos. A base de dados *Statlog (Image Segmentation)* é composta por segmentações feitas manualmente de 7 imagens que permitem a classificação pixel a pixel. Cada instância em uma região 3x3 de uma imagem, são 2310 exemplos e 19 atributos de dados contínuos. O *Waveform Database Generator* contém dados relacionados a geração de três tipos de onda. São 21 atributos de dados contínuos que variam de 0 a 6 e 500 mil instâncias.

1.2. Pré-processamento

O pré-processamento utilizado levou em consideração ao menos uma técnica. O objetivo principal foi melhor preparar os dados para a classificação utilizando o algoritmo KNN, WKNN, Decision Tree e Multinomial Naïve Bayes (NB). Os valores de K para os algoritmos KNN e W-KNN foram escolhidos a partir dos melhores resultados obtidos nas atividades anteriores. Além disso, foi utilizado o *multilayer perceptron MLP classifier* com backpropagation usando os seguintes parâmetros: 15 neurônios na camada escondida, 10000 interações, warm start e a função igual a lbfgs. Os valores de K igual a 7 representando o melhor valore de k para a base de dados Iris, o valor 3 que é o melhor valor para k para o dataset *Statlog (Image Segmentation)* e finalmente o valor 15 para o dataset *Waveform Database Generator*. O algoritmo W-KNN foi ponderado pela distância. Para o *Iris dataset* foi realizada a normalização dos atributos entre os valores de 0 e 1 de todas as instâncias. Sendo assim, a soma de todos os atributos são iguais a 1. Diferentemente da base de dados original, agora, teremos a proporção de cada atributo o que pode revelar a importância dele para classe em específico. Além disso, dados reais entre [0,1] ajudam a evitar possíveis distorções em relação aos atributos. Para as bases de dados *Statlog (Image Segmentation)* e *Waveform Database Generator* além da normalização dos dados feita na base de dados do *Iris dataset* foi utilizada uma análise do principal componente (*PCA*) . Isso permite que possamos reduzir a quantidade de atributos, extraindo as informações mais relevante. Foram selecionados 4 componentes principais de cada conjunto de atributos relacionado as instâncias da base de dados.

2. Resultados Experimentais

Os experimentos consistiram na avaliação do algoritmo KNN usando diferentes valores para os k-vizinhos mais próximos a serem considerados pelo classificador. Nesse experimento foram considerado os valores de k = 3, 5, 7, 9 e 15. For realizado um teste usando *cross validation* para cada base de dados com 10 subconjuntos de dados escolhidos de maneira aleatória. 33% do conjunto de dados será separado para teste e o restante para treinamento. Cada teste foi repetido 5 vezes gerando 50 experimentos.

2.1. Iris Dataset

O primeiro teste foi feito usando a base de dados bruta. Nele os classificadores conseguem obter um bom resultado apesar de não ter sido realizado o pré-processamento. Em uma análise empírica fica claro que o algoritmo KNN e W-KNN com k igual a 7 obtêm os melhores resultados. Não foi possível obter resultados de teste de hipótese.

Dados Brutos Classificador	Média	Mediana	Desvio Padrão	Tempo Médio
DecisionTree	0.938	0.94	0.034582	0.002225
KNN-15	0.960	0.96	0.036140	0.004321
KNN-3	0.958	0.96	0.022946	0.004287
KNN-7	0.974	0.98	0.023990	0.004402
MLPClassifier	0.610	0.60	0.364132	0.038844
MultinomialNB	0.768	0.79	0.182991	0.002587
W-KNN-15	0.968	0.97	0.030237	0.003000
W-KNN-3	0.958	0.96	0.022946	0.003162
W-KNN-7	0.972	0.98	0.024244	0.003154

No segundo experimento, foi realizado o experimento usando as base de dados pré-processadas. Nesse experimento, fica claro uma melhora significativa na média de desempenho do classificador KNN e W-KNN para diferentes valores de k tendo um resultado semelhante entre si. Nesse caso, todos os classificadores obtiveram resultados melhores do que no experimento anterior exceto o MultinomialNB que obteve uma piora muito significativa nos resultados. O teste de hipótese realizado mostra o mesmo do experimento anterior. O algoritmo KNN obtém resultados similares para todos os valores de K com os melhores resultados e os demais aparecem atrás na comparação. Em contrapartida o algoritmo MLP tem o pior resultado.

Base Pré-processada classifier_name	Média	Mediana	Desvio Padrão	Tempo Médio
DecisionTree	0.944	0.95	0.034047	0.002311
KNN-15	0.976	0.98	0.023561	0.004483
KNN-3	0.974	0.98	0.015779	0.004301
KNN-7	0.972	0.98	0.020603	0.004163
MLPClassifier	0.260	0.27	0.042378	0.005709
MultinomialNB	0.544	0.59	0.133034	0.002623
W-KNN-15	0.978	0.98	0.022946	0.003031
W-KNN-3	0.972	0.98	0.018516	0.002978
W-KNN-7	0.974	0.98	0.022223	0.003085

2.2. Statlog

Em relação ao experimento usando a base de dados bruta, uma análise preliminar indica que o classificador Decision Tree obteve um melhor resultado juntamento com o W-KNN para k igual a 3. Os algoritmos têm uma acurácia média maior, e apresentam resultados mais estáveis, sendo o W-KNN um pouco melhor nesse quesito tendo o menor desvio

Base Bruta classifier_name	Média	Mediana	Desvio Padrão	Tempo Médio
DecisionTree	0.953893	0.952163	0.008104	0.016430
KNN-15	0.894758	0.897117	0.011235	0.047987
KNN-3	0.943644	0.943644	0.006067	0.042364
KNN-7	0.920970	0.920708	0.010084	0.044885
MLPClassifier	0.935387	0.935125	0.010101	5.781562
MultinomialNB	0.000000	0.000000	0.000000	0.002285
W-KNN-15	0.932503	0.930537	0.009441	0.029206
W-KNN-3	0.953080	0.952163	0.006371	0.022724
W-KNN-7	0.945085	0.947575	0.008650	0.025153

padrão. O resultado do algoritmo MLP é melhor do que no teste com o dataset iris. Porém esse não é o melhor resultado e ele gasta o maior tempo entre os concorrentes. Por outro lado o Decision Tree gasta menos tempo. Um dos possíveis motivos pelo qual o modelo MultinomialNB obteve um resultado, é o fato da entrada de dados bruta tem valores negativos. Decision Tree apresentar um desempenho superior também no teste de hipótese seguido em segundo e terceiro lugar por KNN com k igual a 3 e 5 respectivamente.

Base Pré-processada classifier_name	Média	Mediana	Desvio Padrão	Tempo Médio
DecisionTree	0.816750	0.816514	0.008745	0.008394
KNN-15	0.784404	0.789646	0.020699	0.028444
KNN-3	0.843644	0.847969	0.014743	0.025940
KNN-7	0.819528	0.823067	0.016430	0.026868
MLPClassifier	0.627654	0.648100	0.129098	2.307945
MultinomialNB	0.415334	0.422674	0.040080	0.003060
W-KNN-15	0.840236	0.834207	0.017317	0.009246
W-KNN-3	0.860026	0.865007	0.014403	0.006246
W-KNN-7	0.855177	0.852556	0.014722	0.007278

Em relação ao experimento usando a base de dados pré-processada, uma análise preliminar indica que o classificador com k igual 3 e 7 para o W-KNN tem um desempenho superior aos demais classificadores. Além de terem um uma acurácia média maior, eles apresentam um dos resultados mais estáveis, tendo um menor desvio padrão. Além disso, eles gastam menos tempo. Portanto, com k igual 3 ou 7 o classificador obtém um resultado superior. Todos os algoritmos apresentaram um resultado inferior ao teste anterior exceto o MultinomialNB. Foi possível obter um resultado para o algoritmo MultinomialNB devido a normalização dos dados entre 0 e 1. O teste de hipótese mostra que o algoritmo KNN para k igual a 3 obtém o melhor resultado.

2.3. Waveform

Em uma análise empírica usando a base de dados brutos, é possível observar que nenhum dos classificadores apresentou um bom desempenho. Os melhores resultados são obtidos

com k 15 tanto para o KNN como para W-KNN. O teste de hipótese confirma a análise inicial em que o KNN tem um resultado melhor com k igual a 15 e o algoritmo MLP.

Base bruta classifier_name	Média	Mediana	Desvio Padrão	Tempo Médio
DecisionTree	0.438061	0.441212	0.012365	0.019341
KNN-15	0.493515	0.490303	0.009672	0.065842
KNN-3	0.463091	0.463636	0.007639	0.056769
KNN-7	0.479091	0.479091	0.009213	0.060671
MLPClassifier	0.502121	0.502727	0.013313	3.438021
MultinomialNB	0.000000	0.000000	0.000000	0.002401
W-KNN-15	0.498788	0.496061	0.011677	0.023832
W-KNN-3	0.465212	0.465152	0.006818	0.013978
W-KNN-7	0.484545	0.483333	0.010898	0.017850

Em um análise posterior, usando a base de dados pré-processada é possível ver que o algoritmo de classificação piora de maneira substância para todos os valores de k e também para os algoritmo Decision Tree e MultinomialNB. Além disso, dessa vez o melhor valor é apresentado pelo modelo MultinomialNB. Mais uma vez os valores foram normalizados entre 0 e 1 para tornar possível a utilização dos dados por esse classificador. Essa análise se confirma pelo teste de hipótese realizado onde algoritmo KNN e W-KNN para k igual a 15 têm os melhores resultados. É interessante ver que o MLPclassifier anteriormente tinha um dos melhores resultados e agora tem um dos piores.

Pré-proessado classifier_name	Média	Mediana	Desvio Padrão	Tempo Médio
DecisionTree	0.809430	0.807879	0.007990	0.016071
KNN-15	0.847636	0.847879	0.005918	0.062543
KNN-3	0.828909	0.828485	0.004338	0.055878
KNN-7	0.838000	0.837879	0.004321	0.058215
MLPClassifier	0.338303	0.341818	0.008679	0.017511
MultinomialNB	0.730242	0.739697	0.027249	0.003267
W-KNN-15	0.845939	0.847576	0.005556	0.020821
W-KNN-3	0.826970	0.826970	0.004261	0.012952
W-KNN-7	0.834970	0.835152	0.003643	0.015821

3. Conclusão

Nesse trabalho, foi usado o algoritmo conhecido como k-Vizinhos mais próximos, (*KNeighbors Classifier*) ou KNN Decision Tree e MultinomialNB, MLP e KNN-W para classificar um conjunto de três bases de dados *1. Iris*, *2. Statlog (Image Segmentation)*, *3. Waveform Data Generator (Version 1)*. Os resultados estarão dispostos em tabelas e gráficos ilustrativos nas seções a seguir. Em geral, o desempenho do algoritmo é melhor após o pré-processamento além de uma tendencia de melhor estabilização dos resultados com o menor desvio padrão. Além disso, em apenas um dos casos os resultados usando dados pré-processados foi pior do que usando a base de dados bruta.